مقدمة حول التحليل الكمي
بهدف وضع سياسات لبرامج الحماية
الاجتماعية قائمة على الأدلة

الأمم المتحدة
الاسكوا
ESCWA

ازدهارُ البلدان **كرامةُ الإنسان**

رؤيتنا

طاقاتٌ وابتكار، ومنطقتُنا استقرارٌ وعدلٌ وازدهار

رسالتنا

بشغَف وعزْم وعَمَل: نبتكِر، ننتج المعرفة، نقدِّمُ المشورة،
نبني التوافق، نواكب المنطقةَ العربية على مسار خطة عام 2030.
يداً بيد، نبني غداً مشرقاً لكلِّ إنسان.

E/ESCWA/CL2.GPID/2022/TP.9

اللجنة الاقتصادية والاجتماعية لغربي آسيا

مقدمة حول التحليل الكمي بهدف وضع سياسات لبرامج الحماية الاجتماعية قائمة على الأدلة

الأمم المتحدة
بيروت

النتائج والتفسيرات والاستنتاجات الواردة في هذه المطبوعة هي للمؤلفين، ولا تمثل بالضرورة الأمم المتحدة أو موظفيها أو الدول الأعضاء فيها، ولا ترتب أي مسؤولية عليها.

ليس في التسميات المستخدمة في هذه المطبوعة، ولا في طريقة عرض مادتها، ما يتضمن التعبير عن أي رأي كان من جانب الأمم المتحدة بشأن المركز القانوني لأي بلد أو إقليم أو مدينة أو منطقة أو لسلطات أي منها، أو بشأن تعيين حدودها أو تخومها.

الهدف من الروابط الإلكترونية الواردة في هذه المطبوعة تسهيل وصول القارئ إلى المعلومات وهي صحيحة في وقت استخدامها. ولا تتحمل الأمم المتحدة أي مسؤولية عن دقة هذه المعلومات مع مرور الوقت أو عن مضمون أي من المواقع الإلكترونية الخارجية المشار إليها.

جرى تدقيق المراجع حيثما أمكن.

لا يعني ذكر أسماء شركات أو منتجات تجارية أن الأمم المتحدة تدعمها.

المقصود بالدولار دولار الولايات المتحدة الأمريكية ما لم يُذكر غير ذلك.

تتألف رموز وثائق الأمم المتحدة من حروف وأرقام باللغة الإنكليزية، والمقصود بذكر أي من هذه الرموز الإشارة إلى وثيقة من وثائق الأمم المتحدة.

مطبوعات للأمم المتحدة تصدر عن الإسكوا، بيت الأمم المتحدة، ساحة رياض الصلح،
صندوق بريد: 8575-11، بيروت، لبنان.

الموقع الإلكتروني: www.unescwa.org.

22-00244

الرسائل الرئيسية

- تم تصميم دليل "مقدمة إلى تحليل البيانات لدعم القرارات المستندة إلى الأدلة" للخبراء التقنيين المسؤولين عن المستفيدين من المساعدة الاجتماعية وقواعد بيانات المتقدمين (أي السجلات الاجتماعية)، والاستهداف وتقييم البرنامج.

- يهدف الدليل إلى بناء قدرة المتدربين على فهم وإجراء تحليل كمي منتظم ومستمر لقاعدة بيانات البرنامج. سيسمح لهم ذلك بالتوصل إلى توصيات سياسية قائمة على البيانات تعمل على تحسين كفاءة وفعالية برامج الحماية الاجتماعية.

المحتويات

مقدمة

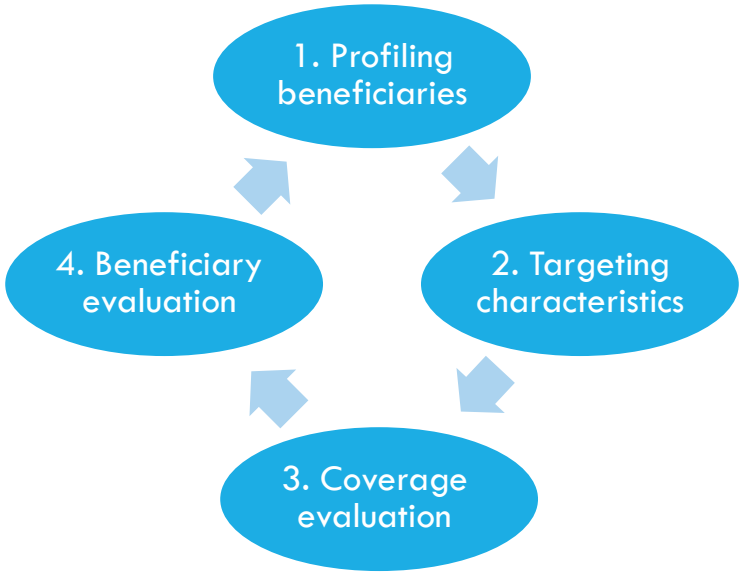
في ظل الأزمات الاجتماعية والاقتصادية والبيئية الحالية، يواجه مواطنو العالم تحديات متزايدة، وفي كل عام يقع ملايين الأشخاص ضحية الفقر والظروف المعيشية السيئة. لهذا السبب، أصبحت برامج الحماية الاجتماعية جزءاً أساسياً من الخطط الإنمائية، وتهدف هذه البرامج إلى دعم أولئك الذين يواجهون صعوبات في تأمين سُبُل العيش الكريم، بسبب الأحداث غير المتوقعة مثل الأحداث الطبيعية والأزمات الاقتصادية المفاجئة، والذين، إن لم يتلقوا الدعم اللازم، قد تؤثّر عليهم تداعيات الصدمة على مدى سنوات عديدة. وتستهدف برامج الحماية الاجتماعية الفئات التي تواجه تحديات هيكلية في المجتمع (مثل الأمهات العازبات، والأشخاص ذوي الإعاقة، والسكان المسنّين) للحد من مشاكلهم أو على الأقل تخفيفها، وذلك عبر تزويدهم بما يكفي من الأدوات لتطوير حياتهم المهنية والشخصية. وتتضمن البرامج الحالية مجموعة واسعة من الاستراتيجيات لتجنّب الفقر أو الحد من أثره، ومعالجة تداعياته والمشاكل الهيكلية التي تسببه.

وفي فترة تتّصف بندرة الموارد، وازدياد احتياجات الناس، تحتاج الحكومات إلى وضع استراتيجيات تسمح للفئات التي هي بأمسّ الحاجة إلى الموارد بالاستفادة منها. ومن خلال هذه الاستراتيجيات، يمكن لصناع السياسات تحسين فرصهم في اختيار البرامج المناسبة لاستخدام الأموال بكفاءة، وبالتالي تغطية احتياجات السكان قدر الإمكان. لا يوجد وصفة سحرية لنجاح هذه الاستراتيجيات، لأن واقع كل بلد وثقافته، وبيئته، ومؤسساته وتاريخه يجعل كل حالة فريدة من نوعها من نواح عديدة. ولهذا السبب، ليس من المعقول الادعاء بوجود حلّ موحّد يناسب جميع البلدان. ومع ذلك، فمن المهم الاعتراف بأن بعض العناصر الهيكلية لهذه البرامج لها أنماط مماثلة. فعلى سبيل المثال، ونظراً لمحدودية الموارد، تحتاج برامج الحماية الاجتماعية كافة إلى آلية لجمع المعلومات من الأفراد، وتحديد مدى تمتّعهم بالخصائص اللازمة لكي يصبحوا مستفيدين. ولذلك، وفي حين أنه ليس من الواقعي وضع دليل موحّد يحدد المبادئ التوجيهية الصحيحة بشأن كيفية تنظيم برامج الحماية الاجتماعية، فمن الممكن وضع دليل لتوجيه صنّاع السياسات بشأن كيفية تنظيم البيانات الحالية بطريقة تغذي البرامج، وتحدّد المجالات القابلة للتحسين. وبعد إجراء هذه التقييمات الأوّلية، سيكون لدى صناع السياسات المعلومات الكافية عن بعض النقاط التي تحتاج إلى التحسين للتمكن من المضي قدماً.

إطار التقييم السريع لبرامج الحماية الاجتماعية (SPP-RAF)

تُعتبر أدوات تقييم الحماية الاجتماعية شائعة في الأدبيات الدولية، وهي تحدد النقاط الرئيسية التي ينبغي على برامج الحماية الاجتماعية النظر فيها من جوانب مختلفة. وبعض هذه الأطر مفاهيمية للغاية ومُصمَّمة خصيصاً لتوجيه القارئ لتحديد الاستراتيجية الشاملة للبرنامج، في حين أن البعض الآخر يركز على الرصد المستمر لأهداف البرنامج من أجل تحديد كفاءتها وفعاليتها، ويهتمّ باستراتيجية الدعوة وكيفية التعامل مع مختلف أصحاب المصلحة لضمان التزامهم بتنفيذ الخطط. وتقوم أطر أخرى بتطوير أدوات تحليلية منطقية لتحديد القدرات والتحديات التي يمكن أن يواجهها البرنامج. (UNICEF, 2019)، (OECD, 2019)، (GIZ, 2017).

ويهدف إطار التقييم السريع لبرامج الحماية الاجتماعية SPP-RAF، انطلاقاً من الأطر والأدوات المتاحة، إلى تقديم تقييمات عملية منخفضة التكلفة يمكن تنفيذها بانتظام كجزء من البرامج، وتزويد صناع السياسات بالنقاط الرئيسية التي تؤثر على نجاحها. لهذا السبب، يتم إنشاء الإطار وفق أربع خطوات محاذية مع الرسم التخطيطي المعروض في الشكل أدناه.



بينما يتم شرح كل خطوة في الفصول التالية، يمكن تحديد المسار المنطقي العام على الشكل التالي:

(أ) **تحديد المستفيدين (Profiling beneficiaries):** يستند الجزء الأول إلى البيانات التي يتم جمعها من المستفيدين المصنفين وفقاً لسماتهم في هذه المرحلة بالاعتماد على مجموعة من الأدوات الإحصائية. وتتيح هذه الملامح لصناع السياسات تحديد المستفيدين من البرنامج، واحتياجاتهم، وضمان تعافيهم الفوري أو التخفيف الفعال من حدّة ظروفهم الصعبة؛

(ب) **خصائص الاستهداف (Targeting characteristics):** يستند الجزء الثاني إلى عملية اختيار المستفيدين. وتُحدَّد هذه الخطوة، من خلال مراجعة معايير الاختيار المعتمدة ومقارنة سمات المستفيدين الحاليين بالخصائص التي تستند إليها عملية الاختيار. توفر هذه الخطوة مدخلات لإجراء مناقشة رفيعة المستوى يمكن فيها لصنّاع القرار التفكير في المتغيرات الحالية التي تحدد المستفيدين من البرنامج، ومناقشة أهميتها بالنظر إلى البرامج الحالية والمستقبلية، والتحقق من مدى الحاجة إلى متغيرات إضافية أو إلغاء، أو إعادة صياغة المتغيرات الحالية؛

(ج) **تقييم التغطية (Coverage evaluation):** يستند الجزء الثالث إلى بيانات المستفيدين الحاليين والمعلومات التي تجمعها مصادر إدارية أخرى عن غير المستفيدين. من خلال تحديد السكان غير المستحقين ضمن السكان المستفيدين الحاليين، وتقدير الخصائص المفقودة لغير المستفيدين، يقدر هذا القسم الأفراد المحتملين الذين يستحقون أن يتم اختيارهم كمستفيدين ولكنهم غير مُدرجين حالياً ضمن

البرنامج. كما يحدد المستفيدين الذين لا تكون خصائص اختيارهم نمطية، والحالات التي لم يتم فيها تسجيل بعض المتغيرات ذات الصلة. وعلى الرغم من أن هذه المعلومات لا تكفي لتحديد الأشخاص الذين يساء تصنيفهم، إلاّ أنّها توجه صناع القرار لتحديد السمات الاجتماعية - الديمغرافية الرئيسية وحلّ المشاكل العالقة والمضي قدماً وفقاً لذلك. مثلاً، يمكن لهذه الخطوة توجيه صناع السياسات لتحديد المحافظات التي يمكن أن يكون لديها سجل فرعي منهجي للمستفيدين، وتسمح لهم بمناقشة أفضل السُبُل لتحسين استراتيجيات تقديم الخدمات ذات الصلة؛

(د) **تقييم المستفيدين (Beneficiary evaluation):** تستند الخطوة الأخيرة إلى القدرة على تتبع الأفراد أثناء مشاركتهم في البرامج وفهم كيفية تغيّر سماتهم. وأُنشئت برامج الحماية الاجتماعية لتحسين ظروف المستفيدين منها؛ ولذلك، يمكن طرح التساؤل التالي حول قدرة البرامج على تحسين هذه الظروف، ومن خلال تحديد المتغيرات الرئيسية للتتبع ومتابعة الأفراد أثناء مشاركتهم في البرنامج، من الممكن التحقق مما إذا كانوا يشاركون بفعالية بطريقة تسمح بتحسين سُبُل عيشهم. تحذير: على نقيض الخطوات الأخرى، تتطلب هذه الخطوة تحليلاً مفصلاً لتحديد أجهزة التعقب وتقييم الاتجاه (القدرة على التمييز بين الفوائد الناتجة عن البرنامج والفوائد الناتجة عن التحسّن العام للمجتمع)، والقدرة على معرفة ما إذا كان الفرد يستفيد فعلياً من البرنامج أم أنّ مشاركته صوريّة. وبما أن هذه العناصر تتطلب مجموعة كبيرة من التقنيات بعيداً عن الخطوات الثلاث الأخرى، فإن هذا الدليل سيغطي الخطوات الثلاث الأولى بالتفصيل ويترك هذه الخطوة الأخيرة ليتمّ تناولها في دليل لاحق.

ولتحقيق أقصى قدر من التطبيق العملي لإطار SPP-RAF، فقد تم بناؤه استناداً إلى أربع ركائز:

(أ) **التنظيم:** تتسم التقنيات الإحصائية المستخدمة لتنفيذ إطار SPP-RAF بأنها: قابلة للتكرار (وبالتالي تكون النتائج مستقلة عن الفرد الذي يقوم بعملية الحساب)، قابلة للقياس (يمكن استخدام نفس الأدوات حتى لو زادت البرامج من عدد المستفيدين منها أو من نطاق تغطيتها)، قابلة للتوحيد (يمكن تطبيق نفس الإجراء التحليلي على برامج مختلفة)، وقابلة للتكيّف (مع تطور أولويات البرنامج، تعكس الأدوات التحليلية هذه التغييرات وتُعلم صناع السياسات بها)؛

(ب) **التحليل الكمي:** تُستخدم البيانات الإدارية، بما يُتيح جمع بيانات منخفضة التكلفة والقيام بتحديثات منتظمة. مثلاً، يمكن استخدام بيانات الاستبيانات التي يُقدّمها الناس للاستفادة من برنامج الحماية الاجتماعية، والبيانات الإضافية التي يقومون بتعبئتها عند الإبلاغ عن ضرائبهم أو المدفوعات التي يتكبدونها مقابل الخدمات العامة. وعلاوة على ذلك، فقد صُمم هذا النظام بحيث لا يتطلب الكثير من البيانات في البداية، بل يتطور مع نضوج نظام الحماية الاجتماعية. وأخيراً، تعكس الاستنتاجات المستخلصة من التحليل الحالة المحليّة، كما يتضح من البيانات التي جُمعت من أجل البرامج. وبالتالي، يمكن أن تساعد الأدوات في توجيه صناع القرار في وضع السياسات، وكذلك مديري البيانات لتحسين نظام المعلومات باستمرار؛

(ج) **القرارات القائمة على الأدلة**: تدعم الأدلة الاستنتاجات المستخلصة. وبالتالي، تعتبر هذه العملية شفافة وتعزز من مساءلة المنظمة وشرعيتها. فمثلاً، يمكن للسياسات المستمدة من هذه الاستنتاجات أن تشكل قاعدة المناقشات العامة، كونها تشير بوضوح إلى البيانات التي تدعم هذه السياسات؛

(د) **الحالة الراهنة والتطور:** من خلال المراقبة الدورية للنظام، تساعد هذه العملية على تحديد:

(1) السكان الذين يواجهون صعوبات في الوصول إلى البرامج. كالأشخاص الذين يواجهون معوّقات للوصول إلى المؤسسات العامة لتقديم طلب المشاركة في البرنامج.

(2) البرامج التي لم تعد أساليب الاختيار فيها مناسبة في الوقت الحالي كما كانت من قبل. مثلاً، البرامج التي كانت معتمَدة في الماضي من أجل تحديد الطابع الرسمي للوظائف، ولكن بعد تغيّر البرنامج المذكور مع مرور الوقت، لم تعد هناك حاجة إليها.

(3) البرامج التي تغيَّر عدد السكان المستهدفين فيها. على سبيل المثال، ازدياد عدد الشابات المستعدّات للانضمام إلى سوق العمل وحاجتهنّ إلى سياسات جديدة لدعمهن.

هيكلية الدليل والملاحظات الختامية

لا يُشكّل هذا الدليل مادة للتعلم الذاتي، بل هو أداة تعليمية أنتجتها اللجنة الاقتصادية والاجتماعية لغربي آسيا (الإسكوا) لاستكمال منهج التدريب على التحليل الكمّي لبرامج الحماية الاجتماعية. ويمكن استخدامه لتنشيط ذاكرة المتدربين. وللحصول على أقصى استفادة منه، فإنه يأتي مرفقاً بالملفات التي تحتوي على الرموز ومجموعات البيانات المستخدمة لشرح المحتويات على اختلافها.

ويقسم الدليل إلى جزأين. يهدف الجزء الأول، الذي يتضمن الفصول الخمسة الأولى، تقديم عرض شامل للبرمجيات الإحصائية المستخدمة لتنفيذ إطار SPP-RAF. والبرنامج الإحصائي المختار هو R. ونظراً لمرونته الكبيرة وخصائصه، وكونه منصة مفتوحة المصدر، أصبح البرنامج الإحصائي R من البرامج الأكثر شعبية في تطوير التحليلات الكمية التي تتطلب عادة توحيد مجموعات البيانات الكبيرة وتقوم بوضع تقارير ضخمة. ولهذه الأسباب، يستند برنامج التدريب الحالي إلى هذه المجموعة الإحصائية. وفي حين أن المتدربين المتمكّنين من هذا البرنامج يمكن أن يتجهوا مباشرة إلى الجزء الثاني، للاطلاع على الأدوات الإحصائية المطلوبة لتنفيذ إطار SPP-RAF، ويفضّل أن يلقوا نظرة ولو سريعة على الفصول الأخيرة من الجزء الأول (ولا سيما الفصلين الرابع و الخامس) لأنها تقدم سُبُلاً لربط البرنامج مع (إكسيل) Excel بطرق تسهّل إنشاء التقارير. يقدم الجزء الأول من الدليل لمحة عن البرنامج الإلكتروني، فيما يناقش الجزء الثاني منه، الذي يتضمن الفصول من السادس إلى الثامن، الخطوات المختلفة لإطارSPP-RAF على الصعيد المفاهيمي ثم يعرض الأدوات والرموز ذات الصلة اللازمة لتنفيذه.

وأخيرا، ولتحسين الصلة بين المفاهيم والتطبيق الفعلي، تستند جميع الأمثلة المقدمة لتنفيذ إطار SPP-RAF - إلى بلد افتراضي لديه برنامج للحماية الاجتماعية بحاجة للتقييم. تبدأ دراسة الحالة في مرحلة مبكرة من الفصل الثاني، حيث يتم توضيح وظائف البرنامج الاحصائي R من خلال وصف بيانات البلد المعني. لذلك، يوصى المتدرّبون الراغبون في الاطلاع مباشرة على أدوات التنفيذ بقراءة الفصول من الثاني إلى الخامس بما أنها تصف دراسة الحالة لأجل فهم أفضل للأمثلة التوضيحية المقدمة في الجزء الثاني.

الجزء الأول. توحيد العمل في إطار البرنامج الاحصائي R

الفصل الأول. مقدمة حول البرنامج الاحصائي R

يهدف هذا الفصل إلى تقديم لمحة عن البرنامج الاحصائي R للمتدربين الذين لم يتسنّ لهم الاطلاع على هذا البرنامج، وذلك عبر وصف وظائفه الأساسية. ويبدأ الفصل بتعريف البرنامج الاحصائي R وتحديد طريقة تحميله، ثم يصف مختلف المكونات التي يستخدمها البرنامج، وينتهي بشرح كيفية إنشاء الرموز الأساسية[1].

1. ما هو البرنامج الاحصائي R؟

(أ) البرنامج الاحصائي R هو عبارة عن لغة وبيئة للحوسبة الإحصائية والرسومات البيانية؛

(ب) وهو يوفر مجموعة من التقنيات الإحصائية والرسومية (النمذجة الخطية وغير الخطية، والاختبارات الإحصائية الكلاسيكية، وتحليل السلسلة الزمنية، والتصنيف، والتكتلات، ...). تستخدم هذه التقنيات في تنظيم التحليل الكمي لبرامج الحماية الاجتماعية لتوفير الدعم في عمليات اتخاذ القرارات القائمة على الأدلة؛

(ج) لا يوجد لدى R واجهة سهلة الاستخدام تسهل تصور المدخلات والمخرجات التي يستخدمها. لذلك، للتمكن من استخدام R بطريقة مريحة، يجب تحميل الواجهة البينية (آر ستوديو) Rstudio التي تحسن طريقة اتصال المستخدم بالبرنامج. ويعرض القسم التالي الخطوات حول كيفية تحميل كل من R وR Studio. للحصول على شرح تفصيلي حول عملية التحميل، يرجى الاطلاع على:
https://courses.edx.org/courses/UTAustinX/UT.7.01x/3T2014/56c5437b88fa43cf828bff5.371c6a924

2. تحميل برنامج R والواجهة البينية Studio R

الخطوة 1: قم/قومي بزيارة الرابط https://www.r-project.org ثم انقر/ي على رابط CRAN الموجود على يسار صفحة الويب.

---

1 وللمتعلمين الراغبين بتوسيع نطاق معرفتهم بالبرنامج الاحصائي R، يوصى بقراءة "الإحصاءات التطبيقية مع برنامج R" من قبل ديفيد دالبياز "Applied Statistics with R"، الذي يتم تحديثه بانتظام ويمكن العثور عليه على الرابط (https://daviddalpiaz.github.io/appliedstats/applied_statistics.pdf). ويستند إلى الفصول 1-7 من هذا الكتاب.

[Home]

Download

CRAN

R Project

About R
Logo
Contributors
What's New?
Reporting Bugs
Conferences
Search
Get Involved: Mailing Lists
Developer Pages
R Blog

The R Project for Statistical Computing

Getting Started

R is a free software environment for statistical computing and graphics. It compiles and runs on a wide variety of UNIX platforms, Windows and MacOS. To **download R**, please choose your preferred CRAN mirror.

If you have questions about R like how to download and install the software, or what the license terms are, please read our answers to frequently asked questions before you send an email.

News

- **R version 4.1.0 (Camp Pontanezen)** has been released on 2021-05-18.
- **R version 4.0.5 (Shake and Throw)** was released on 2021-03-31.
- Thanks to the organisers of useR! 2020 for a successful online conference. Recorded tutorials and talks from the conference are available on the R Consortium YouTube channel.
- You can support the R Foundation with a renewable subscription as a supporting member

News via Twitter

الخطوة 2: اضغط/ي على أي رابط للمتابعة، من القائمة التي تبدأ بـ 0-Cloud.

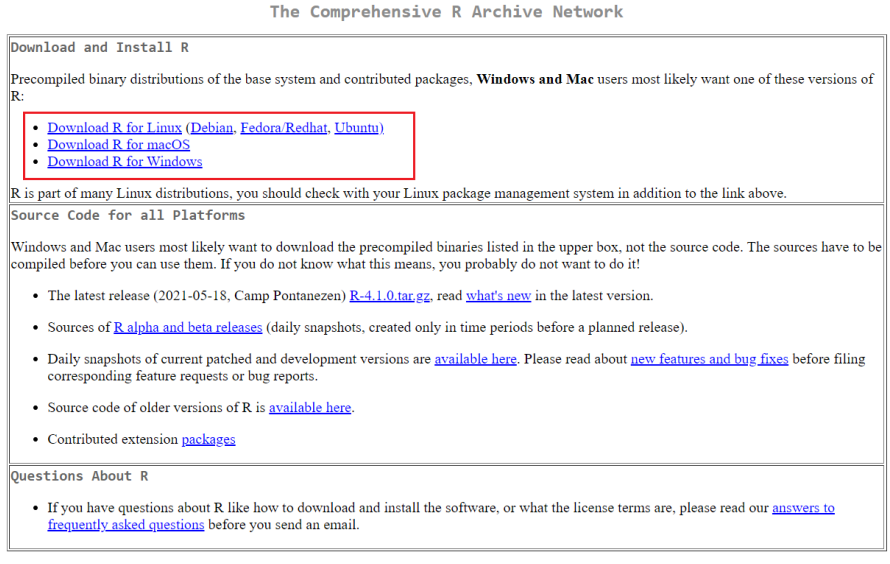
CRAN Mirrors

The Comprehensive R Archive Network is available at the following URLs, please choose a location close to you. Some statistics on the status of the mirrors can be found here: main page, windows release, windows old release.

If you want to host a new mirror at your institution, please have a look at the CRAN Mirror HOWTO.

0-Cloud	
https://cloud.r-project.org/	Automatic redirection to servers worldwide, currently sponsored by Rstudio
Algeria	
https://cran.usthb.dz/	University of Science and Technology Houari Boumediene
Argentina	
http://mirror.fcaglp.unlp.edu.ar/CRAN/	Universidad Nacional de La Plata
Australia	
https://cran.csiro.au/	CSIRO
https://mirror.aarnet.edu.au/pub/CRAN/	AARNET
https://cran.ms.unimelb.edu.au/	School of Mathematics and Statistics, University of Melbourne
https://cran.curtin.edu.au/	Curtin University
Austria	
https://cran.wu.ac.at/	Wirtschaftsuniversität Wien
Belgium	
https://www.freestatistics.org/cran/	Patrick Wessa
https://ftp.belnet.be/mirror/CRAN/	Belnet, the Belgian research and education network
Brazil	
https://nbcgib.uesc.br/mirrors/cran/	Computational Biology Center at Universidade Estadual de Santa Cruz
https://cran-r.c3sl.ufpr.br/	Universidade Federal do Parana
https://cran.fiocruz.br/	Oswaldo Cruz Foundation, Rio de Janeiro
https://vps.fmvz.usp.br/CRAN/	University of Sao Paulo, Sao Paulo
https://brieger.esalq.usp.br/CRAN/	University of Sao Paulo, Piracicaba
Bulgaria	
https://ftp.uni-sofia.bg/CRAN/	Sofia University

الخطوة 3: انقر/ي على Download R أي تحميل R وينطبق ذلك على كل من ويندوز Windows وماك Mac (يعتمد ذلك على جهازك).



The Comprehensive R Archive Network

Download and Install R

Precompiled binary distributions of the base system and contributed packages, **Windows and Mac** users most likely want one of these versions of R:

- Download R for Linux (Debian, Fedora/Redhat, Ubuntu)
- Download R for macOS
- Download R for Windows

R is part of many Linux distributions, you should check with your Linux package management system in addition to the link above.

Source Code for all Platforms

Windows and Mac users most likely want to download the precompiled binaries listed in the upper box, not the source code. The sources have to be compiled before you can use them. If you do not know what this means, you probably do not want to do it!

- The latest release (2021-05-18, Camp Pontanezen) R-4.1.0.tar.gz, read what's new in the latest version.
- Sources of R alpha and beta releases (daily snapshots, created only in time periods before a planned release).
- Daily snapshots of current patched and development versions are available here. Please read about new features and bug fixes before filing corresponding feature requests or bug reports.
- Source code of older versions of R is available here.
- Contributed extension packages

Questions About R

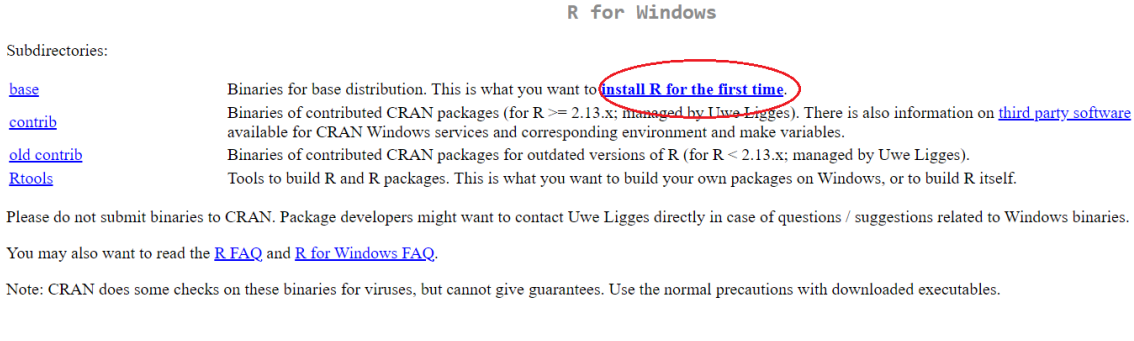
- If you have questions about R like how to download and install the software, or what the license terms are, please read our answers to frequently asked questions before you send an email.

What are R and CRAN?

R is 'GNU S', a freely available language and environment for statistical computing and graphics which provides a wide variety of statistical and graphical techniques: linear and nonlinear modelling, statistical tests, time series analysis, classification, clustering, etc. Please consult the R project homepage for further information.

CRAN is a network of ftp and web servers around the world that store identical, up-to-date, versions of code and documentation for R. Please use the CRAN mirror nearest to you to minimize network load.

الخطوة 4: انقر/ي على "install R for the first time" أي "تثبيت R للمرة الأولى".



R for Windows

Subdirectories:

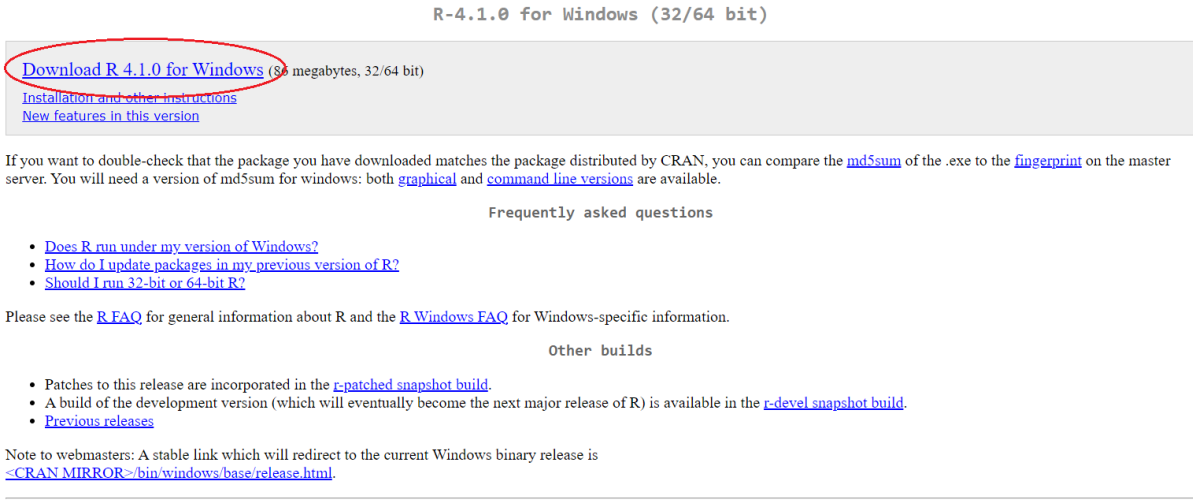
base	Binaries for base distribution. This is what you want to **install R for the first time**.
contrib	Binaries of contributed CRAN packages (for R >= 2.13.x; managed by Uwe Ligges). There is also information on third party software available for CRAN Windows services and corresponding environment and make variables.
old contrib	Binaries of contributed CRAN packages for outdated versions of R (for R < 2.13.x; managed by Uwe Ligges).
Rtools	Tools to build R and R packages. This is what you want to build your own packages on Windows, or to build R itself.

Please do not submit binaries to CRAN. Package developers might want to contact Uwe Ligges directly in case of questions / suggestions related to Windows binaries.

You may also want to read the R FAQ and R for Windows FAQ.

Note: CRAN does some checks on these binaries for viruses, but cannot give guarantees. Use the normal precautions with downloaded executables.

الخطوة 5: ثم انقر/ي فوق "Download R for Windows" "تحميل R لنظام التشغيل ويندوز" (يستند الرسم التوضيحي على جهاز يعتمد Windows، وينطبق أيضاً على نظام التشغيل Mac).



R-4.1.0 for Windows (32/64 bit)

Download R 4.1.0 for Windows (86 megabytes, 32/64 bit)
Installation and other instructions
New features in this version

If you want to double-check that the package you have downloaded matches the package distributed by CRAN, you can compare the md5sum of the .exe to the fingerprint on the master server. You will need a version of md5sum for windows: both graphical and command line versions are available.

Frequently asked questions

- Does R run under my version of Windows?
- How do I update packages in my previous version of R?
- Should I run 32-bit or 64-bit R?

Please see the R FAQ for general information about R and the R Windows FAQ for Windows-specific information.

Other builds

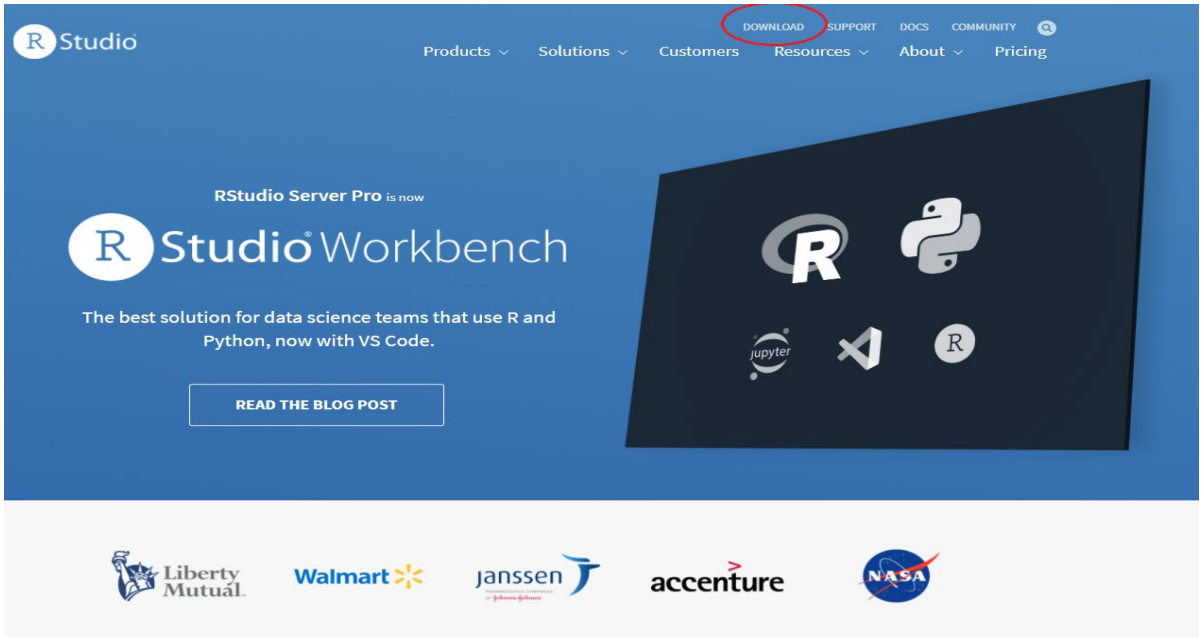
- Patches to this release are incorporated in the r-patched snapshot build.
- A build of the development version (which will eventually become the next major release of R) is available in the r-devel snapshot build.
- Previous releases

Note to webmasters: A stable link which will redirect to the current Windows binary release is <CRAN MIRROR>/bin/windows/base/release.html.

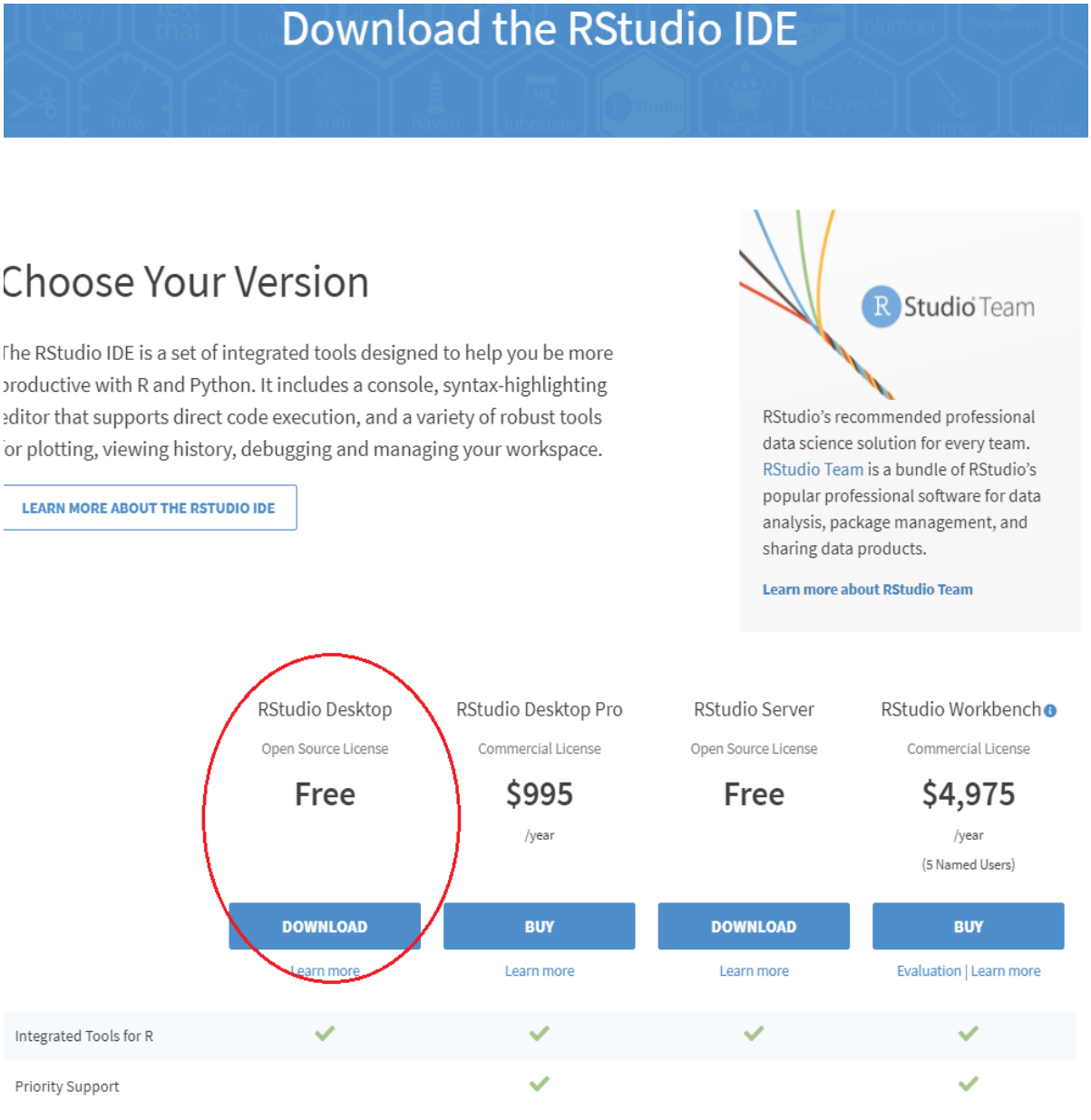
الخطوة 6: قم/قومي بتشغيل ملف التثبيت على جهازك واتبع/ي الإرشادات الموصى بها في برنامج التثبيت.

بعد تثبيت برنامج "R" بنجاح، يمكن المضي قُدماً في تحميل "Rstudio".

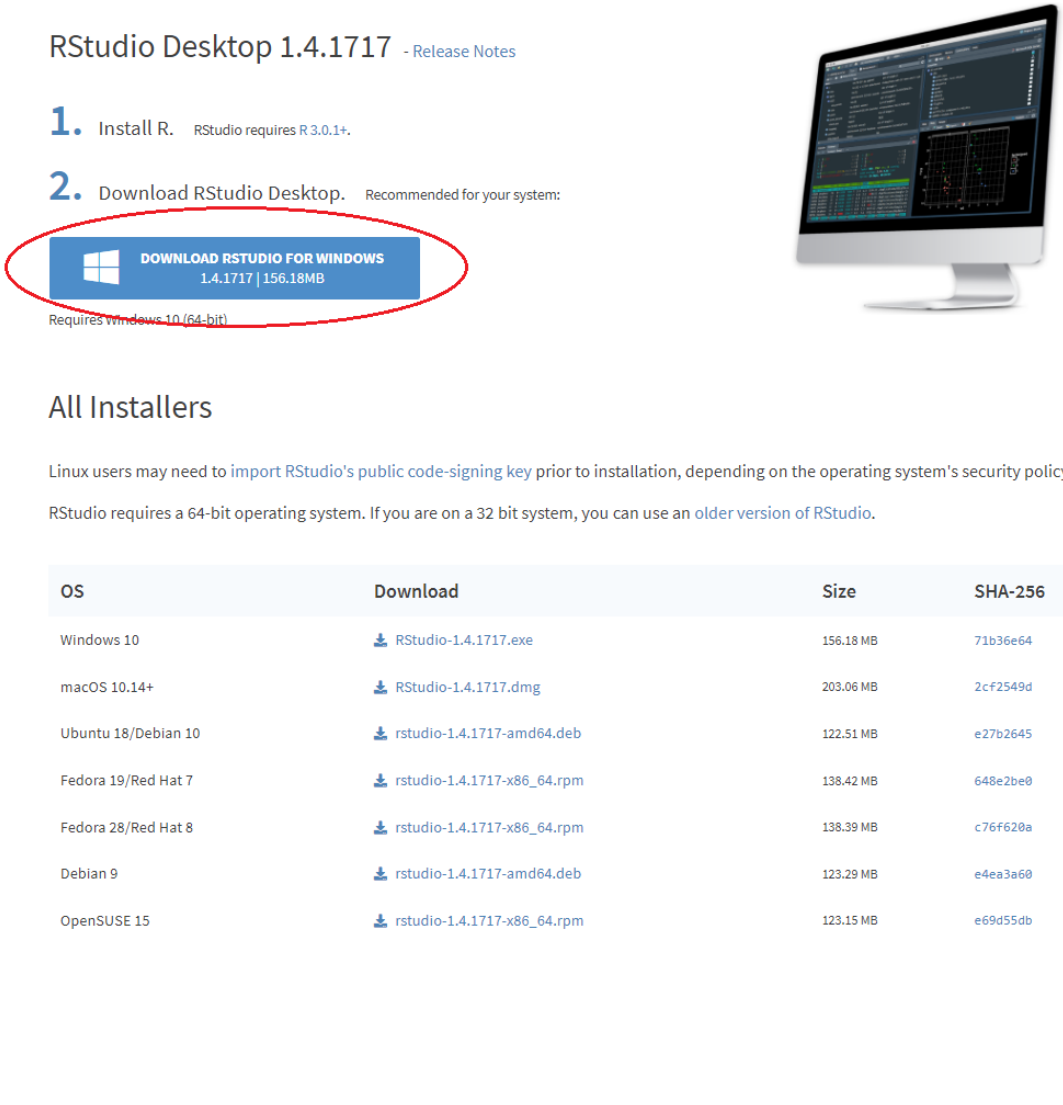
الخطوة 1: قم/قومي بزيارة موقع www.rstudio.com، وانقر/ي على "Download" أو "تحميل".



الخطوة 2: انقر/ي على النسخة "المجانية" من Rstudio.

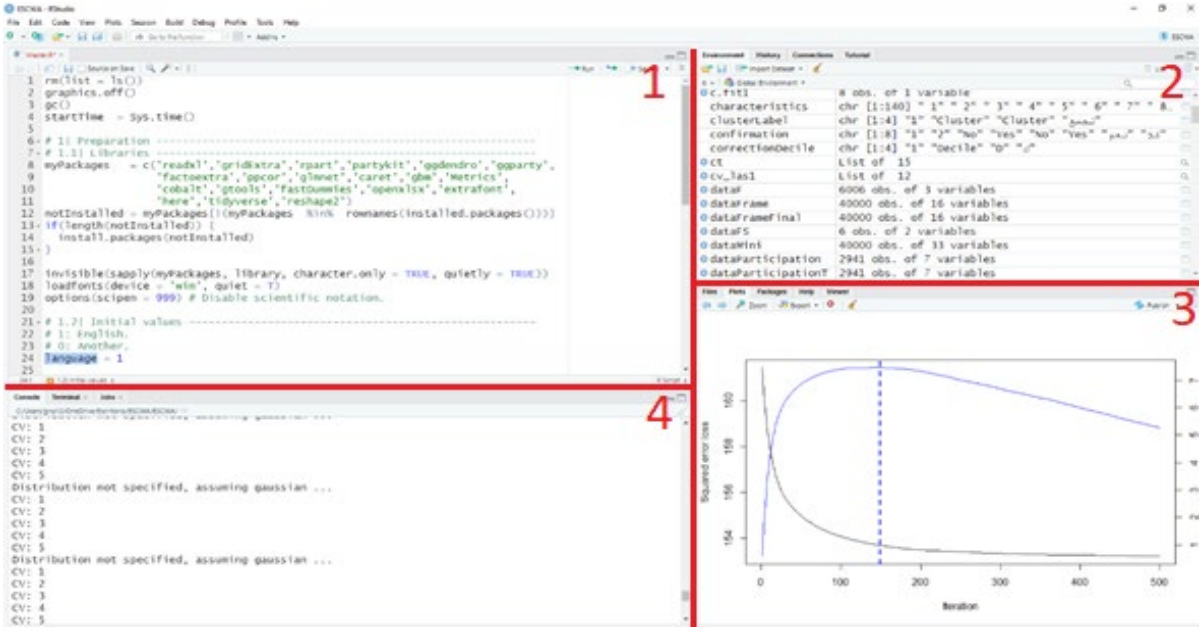


الخطوة 3: انقر/ي على "Download R Studio" أي "تحميل Rstudio". بمجرد تنزيل ملف التثبيت، افتحه واتبع الإرشادات التي تظهر على الشاشة.

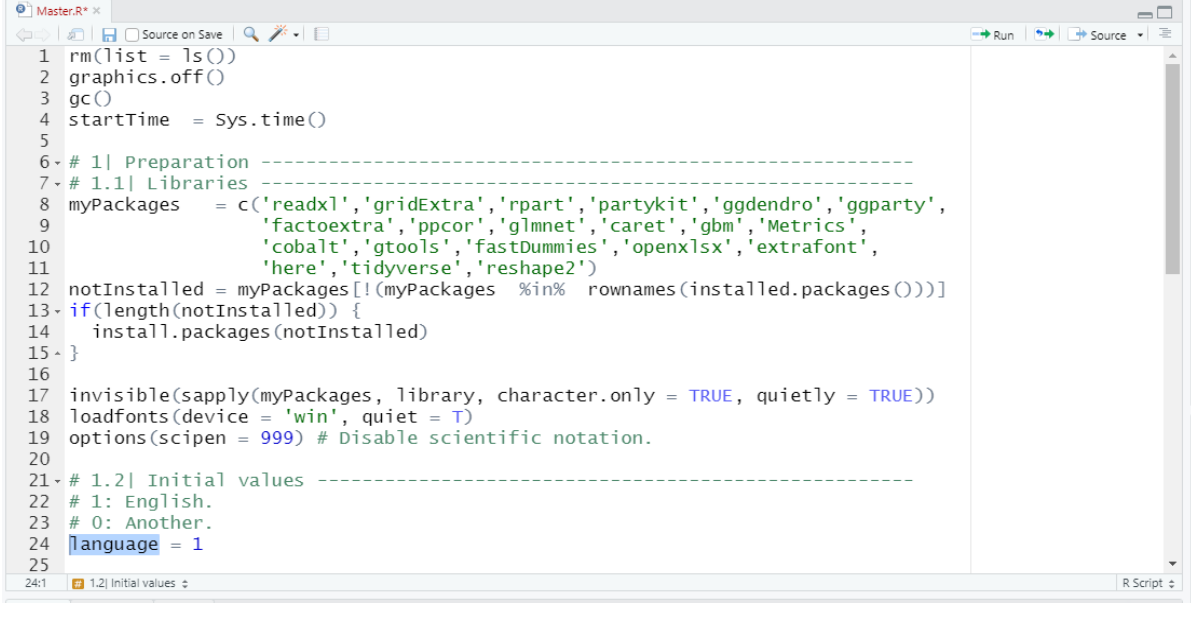


3. الواجهة البينية RStudio

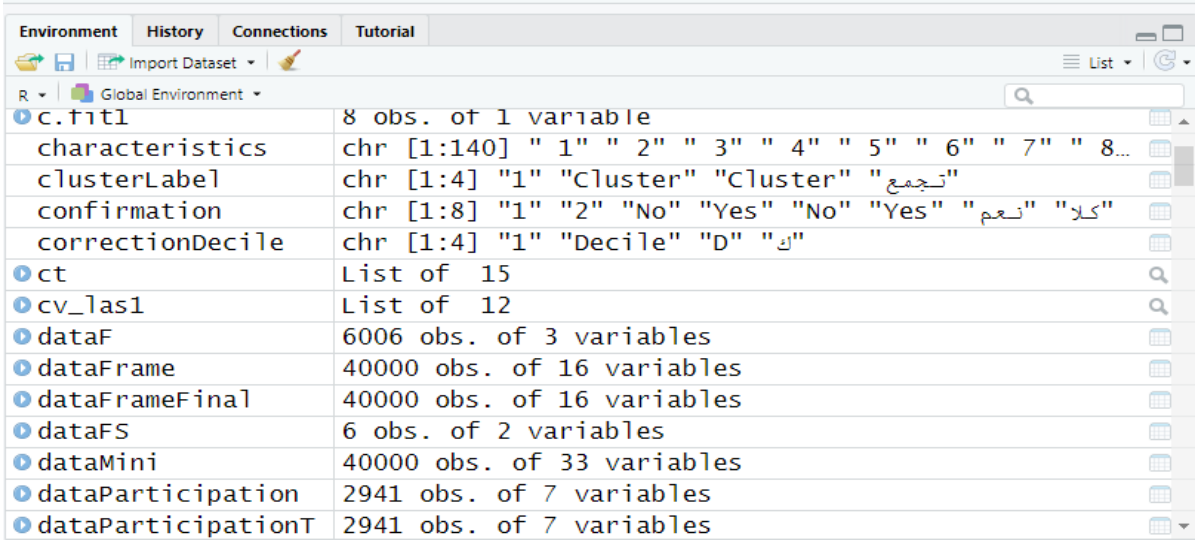
عند تشغيل RStudio على جهاز الحاسوب الخاص بك، ستلحظ/ي أربع مجموعات من واجهة: Rstudio.



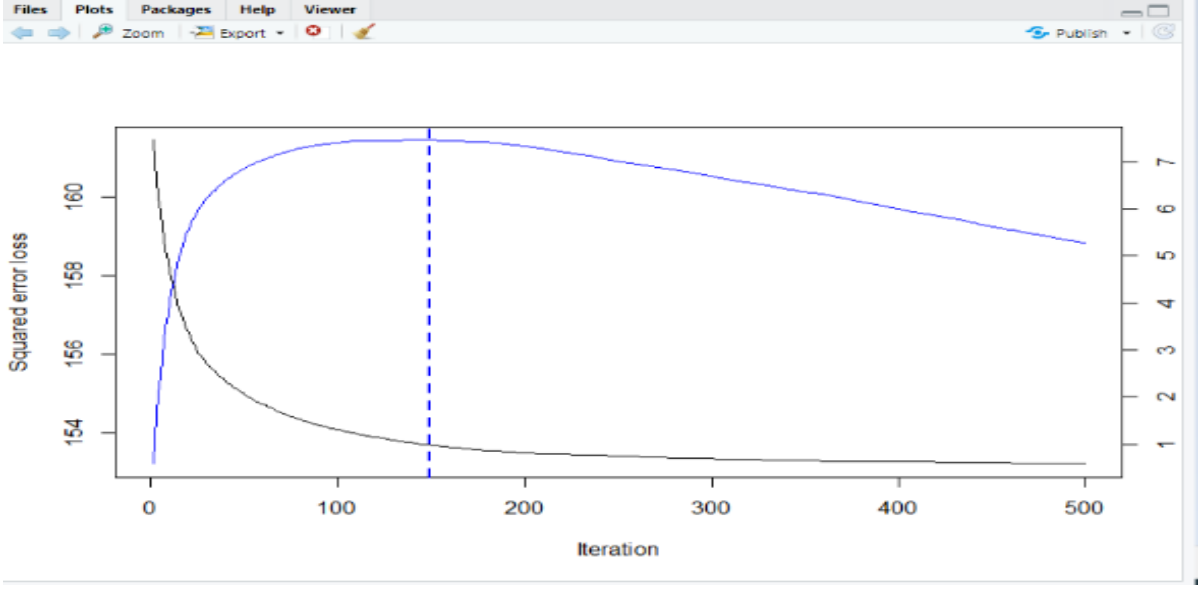
ويسمى الربع الأول بالمصدر؛ حيث يمكن كتابة نصوص التعليمات البرمجية أو الرموز، وحفظها، ومراجعة قواعد البيانات، وتحديد خطوط التعليمات البرمجية التي تريد تنفيذها.



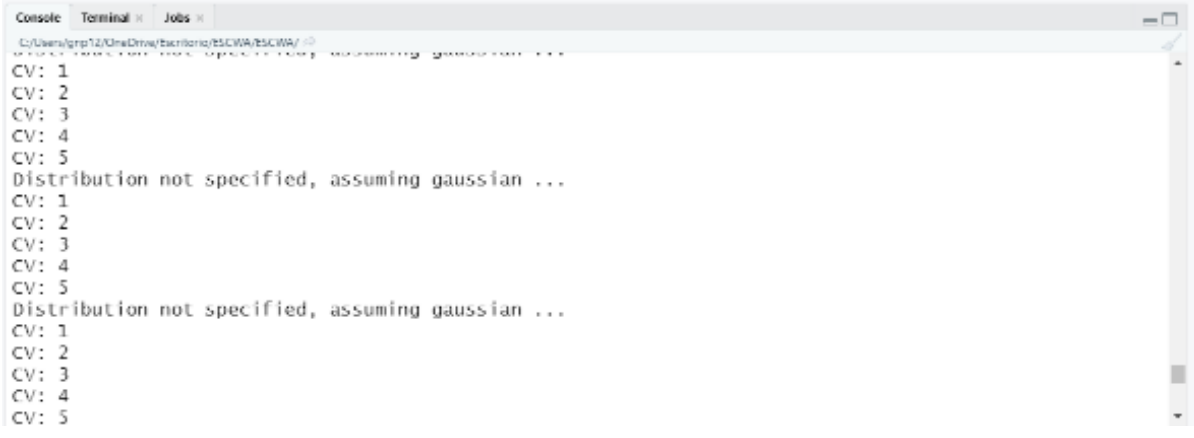
ويسمى الربع الثاني بالبيئة والتاريخ، ويستعرض جميع المتغيرات التي قمت بتحميلها. من بينها ستجد القوائم والوظائف ومجموعات البيانات وغيرها من المتغيرات التي ستفيد في تطوير التمارين ذات الصلة. يمكن التأكّد من المحفوظات، والاتصالات، والاطلاع على البرامج التعليمية، ولكن نوافذ التبويب هذه ليست ذات صلة بالدليل الحالي.



والربع الثالث المتعدد الأغراض، هو الأكثر فائدة لأنه يسمح بالتحقق من الملفات الخاصة بك، ومعاينة ما قمت بإنشائه، والتحقق من الحزم التي يتم تحميلها في النظام (موضّحة في القسم التالي)، والاطّلاع على دليل المساعدة في البرنامج.



والربع الأخير هو وحدة التحكم، حيث سترى فيه نتائج التعليمات البرمجية التي تقوم بتشغيلها في قسم المصدر، ويمكنك أيضا كتابة الرموز البرمجية الأساسية، ولكن لا يُنصح بذلك لأنها لا تحتفظ بتعقب واضح كما في قسم المصدر.



4. الحزم والمكتبات

وتمثل الحزم مجموعات من وظائف R والرموز التي يستخدمها البرنامج للقيام بالتحليل الإحصائي حيث يوجد مجموعة من الحزم الأساسية التي تستخدم كجزء من الشيفرة المصدرية لـ R ويمكن الوصول إليها مباشرة كجزء من تثبيت R، دون الحاجة لتحميلها على حدة. تحتوي هذه الحزم على الوظائف الرئيسية التي تسمح لـ R بالعمل، وتؤدي وظائف إحصائية ورسومية نموذجية.

يتم الاحتفاظ بالحزم في "المكتبات" "libraries". وإذا رغبت باستخدامها عليك أن تُعلم R بذلك، من خلال استخدام الدالة (library).

```
1  library(MASS)
```

بمجرد كتابة النص في قسم المصدر، يُطرح تساؤل حول عملية التنفيذ (أي إعلام R بأنك تطلب منه أن ينفذ). وهناك خياران للقيام بذلك.

"Ctrl + Enter" على لوحة المفاتيح

(الخيار الموصى به).

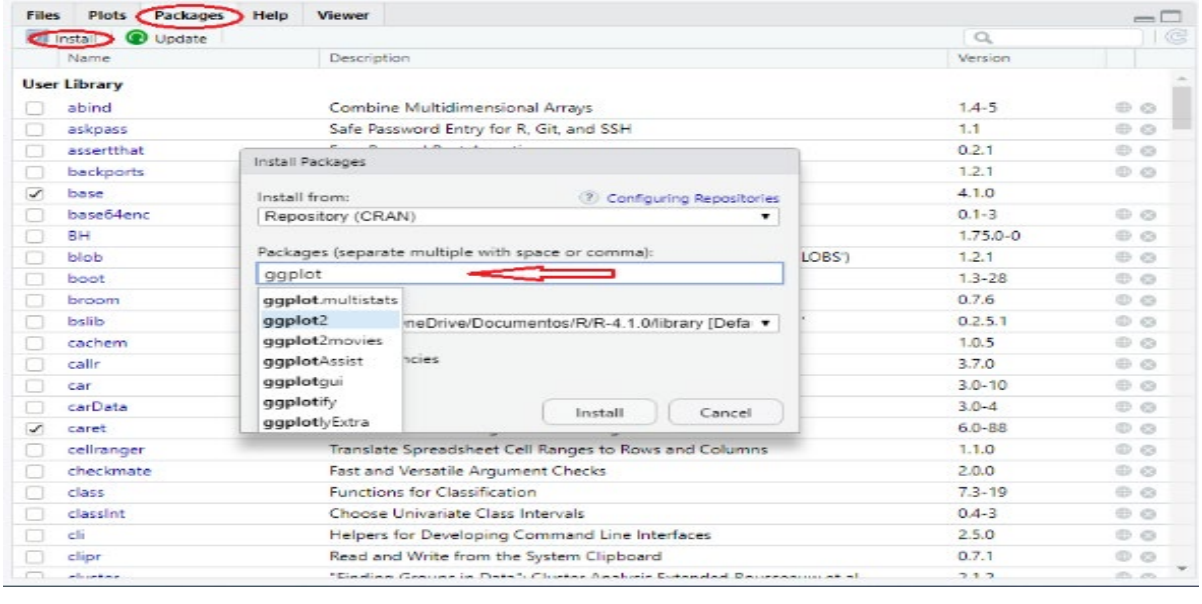
```
1  library(MASS)
```

الخيار الثاني، عبر تحديد سطر الأوامر ثمّ الضغط على الزر: Run "تشغيل".

5. حزم التثبيت

R هو برنامج مفتوح، حيث تسهل الحزم الجديدة التي يتم إنشاؤها كل يوم من قبل المستخدمين مهامهم التحليلية. ومن أجل استخدامها، تحتاج هذه الحزم إلى تحميل وتثبيت. ويمكن إجراء عملية الإعداد بطريقتين:

تستخدم الطريقة الأولى علامات التبويب الحالية في الربع الثالث المتعدد الأغراض. ثمّ انقر/ي على علامة التبويب "الحزم" "Packages" ثم اختر "تثبيت" أو "Install"، ويمكن كتابة اسم الحزمة المطلوب تحميلها وتثبيتها. لكن يلاحظ عند كتابة اسم هذه الحزمة التي تريد تثبيتها (مثلاً، نريد تثبيت "ggplot2")، ستقدم القائمة اقتراحات الحزم المتاحة. مما سيساعدك على تجنب الأخطاء الإملائية.



ويمكن أيضا تثبيت حزمة عبر وظيفة (install.packages) في سطر الأوامر. إلاّ أنّ استخدام هذا الخيار يحتاج لمعرفة الاسم الدقيق للحزمة ممّا يساعد في الحفاظ على رموزك واضحة وكاملة. وتحسين العمل الجماعي وتجنب مشاكل التنسيق. مثلاً، إذا قمت بمشاركة الرموز أو التعليمات البرمجية مع شخص آخر، واحتاجت هذه التعليمات البرمجية إلى حزمة، لن تتمكن من تشغيلها. أما إذا قمت بإضافة هذا السطر إلى التعليمات البرمجية، يتم تلقائياً تثبيت الحزمة في الحاسوب، وستتمكن من تشغيل التعليمات البرمجية.

```
1  install.packages("ggplot2")
```

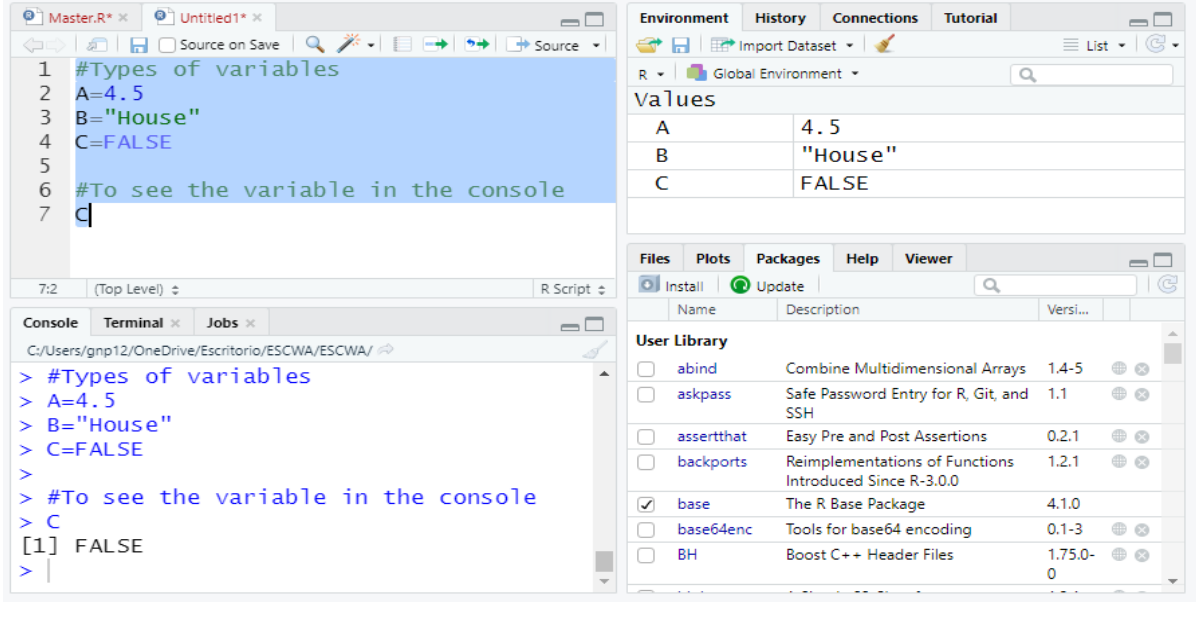
ملاحظة أخيرة، يختلف تحميل الحزمة عامة عن تحميلها في برنامج R ويتمّ تحميل كافة الحزم التي تحتاجها باستخدام الدالة المكتبة Library.

6. عناصر R

بعد تغطية معظم العناصر الهيكلية للبرنامج، يستعرض هذا القسم أنواع المتغيرات التي يستخدمها برنامج R لتطوير عمله.

7. المتغيّرات

يتم تقييم المتغيرات بواسطة النظام. حيث يحتوي R على أنواع مختلفة منها: متغيرات رقمية وأحرف ومتغيرات منطقية. ولإنشاء أي من هذه المتغيرات يجب فقط تسميتها، وعبر علامة "=" تقوم بتعريفها بالطريقة المناسبة.



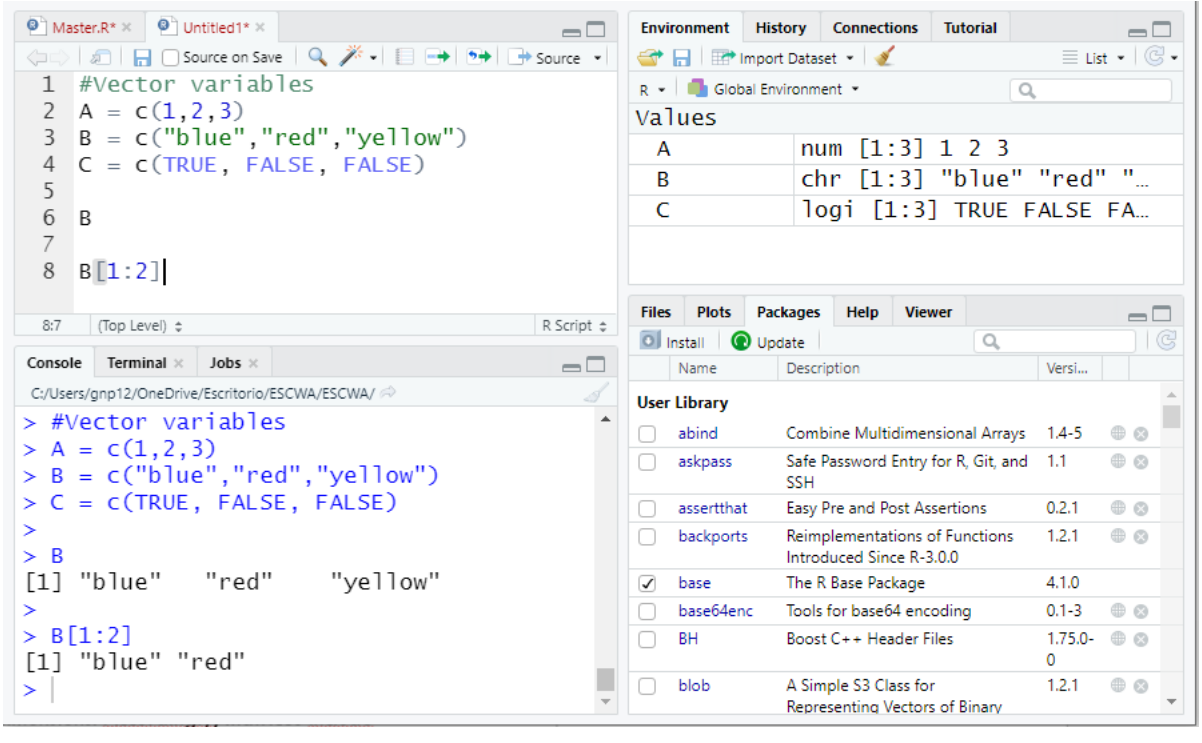
يُرجى ملاحظة أمرين:

(أ) بمجرد تشغيل خطوط الأوامر (السطور من 1 إلى 4)، سيبلّغ رباعي البيئة والتاريخ عن المتغيرات التي قمت بتعريفها. إلاّ أنّها لن تظهر على وحدة التحكم. وإذا أردتَ رؤيتها، يجب تشغيل الأمر فقط عبر اسم المتغير، كما هو الحال في المثال C؛

(ب) هناك بعض خطوط الأوامر التي تبدأ بـ "#". وقد تم تصميم برنامج R لإيقاف قراءة السطر بعد كتابة هذه العلامة. كما يمكن استخدامه لإضافة تعليقات إلى تعليماتك البرمجية ليسهل على الآخرين فهمك.

8. المتجهات

عنصر أساسي آخر في R هو المتجه الذي يجمع المتغيرات ذات النوع الواحد: مثلاً، جميع المتغيرات لنفس المتجه هي: متغيرات رقمية، أحرف أو متغيرات منطقية.



ويرجى الملاحظة:

(أ) لإنشاء المتجهات، يتم استخدام دالة (c) التي تشير إلى سلسلة، وهي تضع جميع العناصر داخل قائمة؛

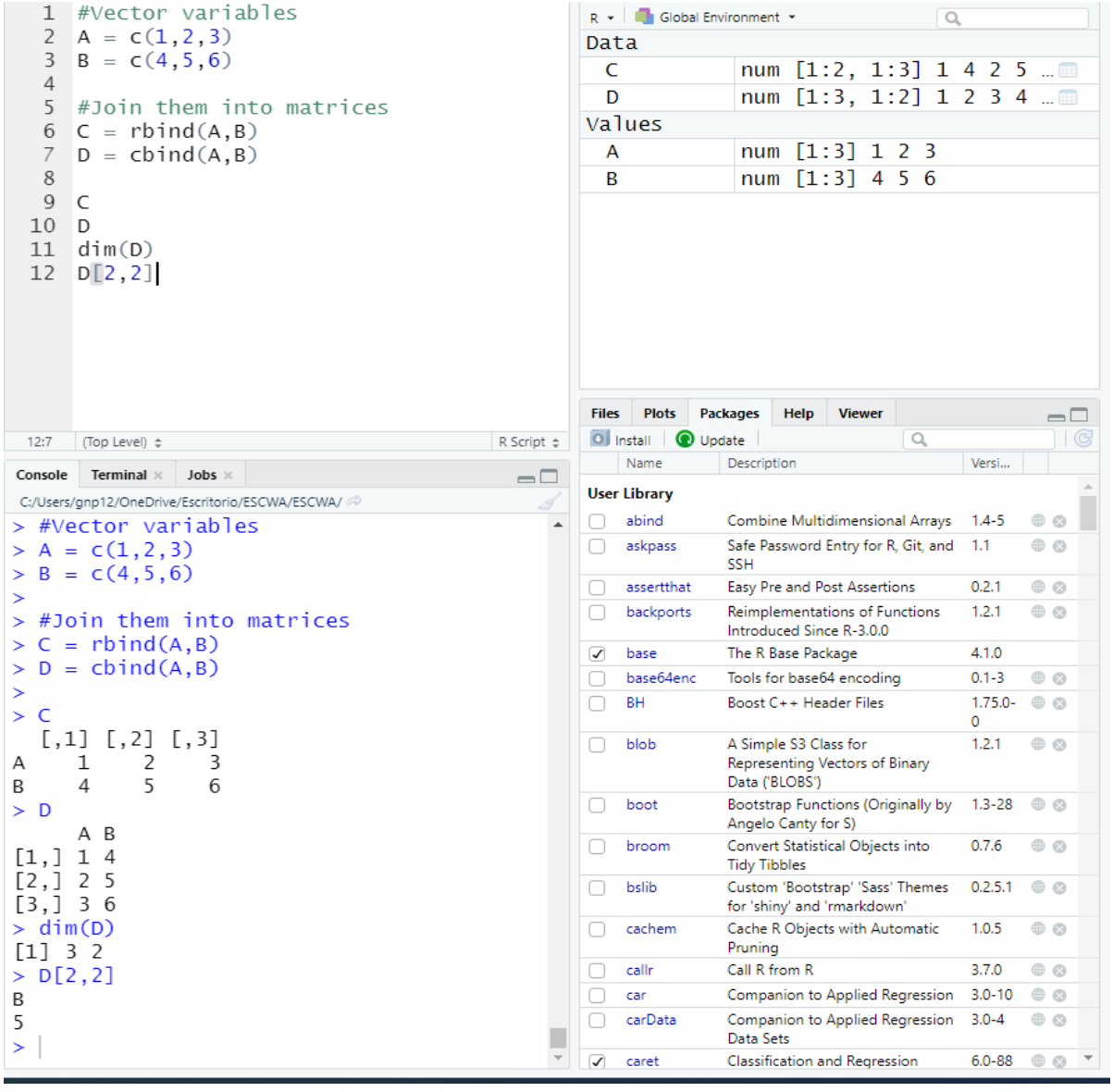
(ب) ويمكن ملاحظة التغييرات الطفيفة في وصف المتغير في الربع الثاني. مثلاً، يشار إلى المتغير A، بـ num بما يفيد أنه متغير رقمي، ويظهر [1:3] ممّا يفيد بوجود ثلاثة إدخالات مفهرسة من الرقم 1 إلى الرقم 3؛

(ج) عند استدعاء المتجه، كالمثال B (سطر 6)، تعرض وحدة التحكم كافة العناصر؛

(د) إذا أردت تحديد عنصر معين، يمكنك استخدام الأقواس المربعة وتحديد الإدخالات المطلوبة. في المثال، يطلب خط التعليمات البرمجية الإدخالات 1 و2 فقط.

9. المصفوفات

المصفوفات مشابهة للمتجهات، ولكنها ذات أبعاد أكثر. ويوضح المثال التالي كيفية ظهور المصفوفات، وكيفية إنشائها.

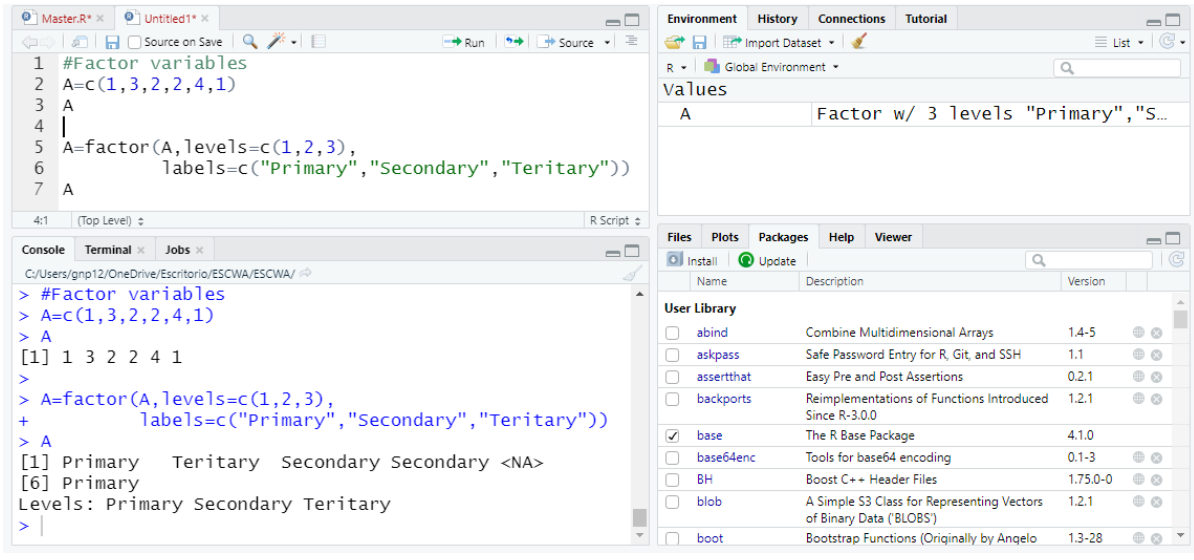


تتبع التعليمات البرمجية السابقة خطوات محددة. أولاً، فهي تخلق اثنين من المتجهات، A و B، وباستخدام وظائف (rbind) و (cbind)، تلصقهما "paste" لتشكيل مصفوفة، حيث تربط الوظيفة الأولى بينهما عبر صفوف والثانية عبر أعمدة. على نقيض المتجهات، فإنّ للمصفوفة بعدين، لذلك ترى في الربع الثاني بنية مزدوجة تذكر أولاً عدد الصفوف وثانياً عدد الأعمدة. وهناك طريقة سريعة لمعرفة الأبعاد تتمثل باستخدام وظيفة (dim). كما نرى في المثال، إخراج هذه الدالة هو متجه بقيم 3 (للصفوف الثلاثة في D) و2 (للعامودين في D).

وأخيراً، يمكن في حال المتجهات استخدام أقواس مربعة (brackets) لاستدعاء إدخالات معينة، وبما أن لها بعدين، يجب تحديد القيم. في هذا المثال، يظهر لنا استدعاء D[2,2] القيمة في الصف الثاني (الإدخال الأول في القوس) والعمود الثاني (الإدخال الثاني في القوس).

10. العوامل

تُستخدم العوامل لتحديد الفئات ذات المتغيرات. مثلاً، عندما يطلب استطلاع مستوى التعليم، فإنك ترغب في رؤية قيم مثل "الابتدائي" أو "الثانوي" أو "الثالثي"، ولكن وفقاً للطريقة التي تعطى بها مجموعة البيانات تظهر قيماً مثل 1 أو 2 أو 3 (تمثل كل مستوى تعليمي). لذلك يجب إبلاغ البرنامج أن هذه الأرقام ليست مجرد متغيرات رقمية، بل تمثل فئات محددة. على أنّ العوامل هي الوسيلة التي تسمح بالقيام بهذا التمرين.

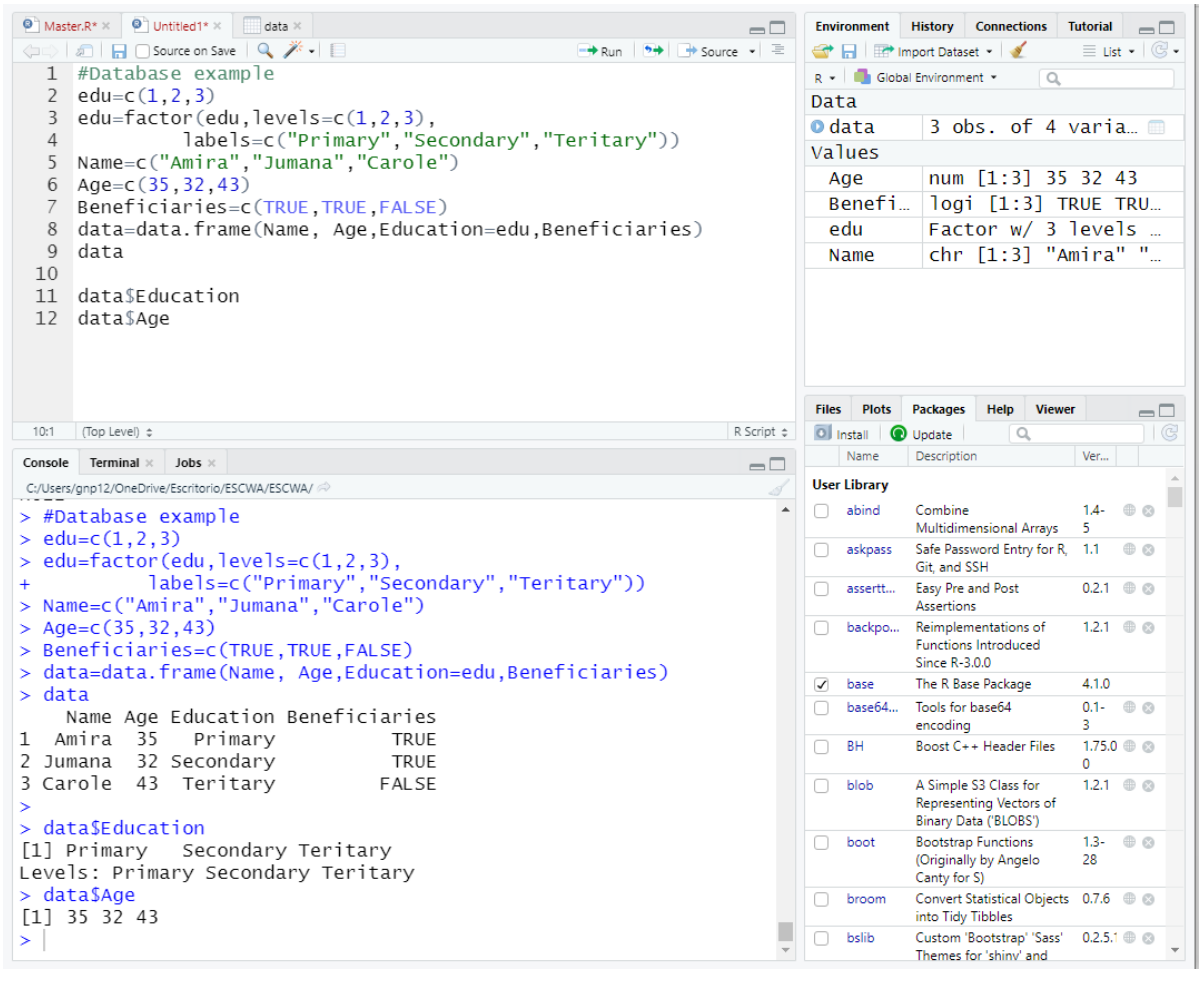


كما في التعليمات البرمجية السابقة، يحتوي المتجه الأصلي على قيم رقمية، وعند تشغيله (السطران 2 و3) سترى الأرقام فقط. وباستخدام عامل (الدالة) (factor)، سيفهم البرنامج أن هذه القيم تمثل أشياء محددة. لكن سنلاحظ أن ذلك لن يتمّ على الفور، ولهذه الوظيفة، على نقيض تلك السابقة ثلاثة عناصر. أولاً، تضع المتجه الذي تريد تحديده. ثانياً، تحدد المستويات التي تمثل فئات. وأخيراً، وبنفس ترتيب المستويات، تكتب التسميات التي تريدها لهذه الفئات.

الآن، سنلاحظ أنه عندما تنظر إلى محتوى A بعد الانتهاء من تمرين العامل، لا ترى أي رقم، ولكن ترى الفئات. ومع ذلك، ترى <NA> (قيمة مفقودة) في الموضع الخامس والسبب أنّ هناك ثلاثة مستويات محددة فقط (1، 2، 3)، ولكن المتجه الأصلي قيمته "4". ونظراً لعدم تحديد هذا المستوى، يضع البرنامج عليه علامة <NA> كقيمة مفقودة لأنه ربما يكون ناتجاً عن خطأ في الكتابة في مرحلة إعداد الاستطلاع.

11. أُطر البيانات

على نقيض المصفوفات، يمكن أن تحتوي أُطر البيانات على متغيرات من أنواع مختلفة. مما يسمح بالعمل بالفئات والأرقام والأحرف والمتغيرات المنطقية ضمن إطار واحد.



وبتسليط الضوء على بعض عناصر التعليمات البرمجية أو الرموز المشار إليها نلاحظ: أولاً، تنشئ التعليمات البرمجية أربعة متغيرات: الاسم (Name)، ومستوى التعليم (Edu)، والعمر (Age)، والمستفيدون (Beneficiaries)، في إشارة الى ثلاث سيدات يشاركن في برنامج الحماية الاجتماعية. ولكل متغير نوع مختلف،

فهناك أرقام في متغير العمر (Age) وفئات في مستوى التعليم (Edu) وأحرف في الأسماء (Names) ومتغيرات منطقية في تحديد المستفيدين (Beneficiaries). لذا ستحتاج للعمل في كل منها على حدة، واستخدام وظيفة أطر البيانات (data.frame) لجمعها، حيث يعتبر استخدامها سهلاً، إذ يمكنك إضافة المتجهات التي تريدها، وهي سترتب المتغيرات تلقائياً. إلاّ أنّه، وفيما خصّ التعليم، يمكن ملاحظة وجود فرق. وباستخدام الخدعة المعروضة في التعليمات البرمجية (Education = edu)، يمكنك ضبط الأسماء على مجموعة البيانات بحيث يسجل إطار البيانات تماماً الاسم الذي تريده. على خلاف المصفوفات، إذا رغبتَ في استدعاء قيمة من إطار البيانات تستخدم الرمز "$" واسم المتغير.

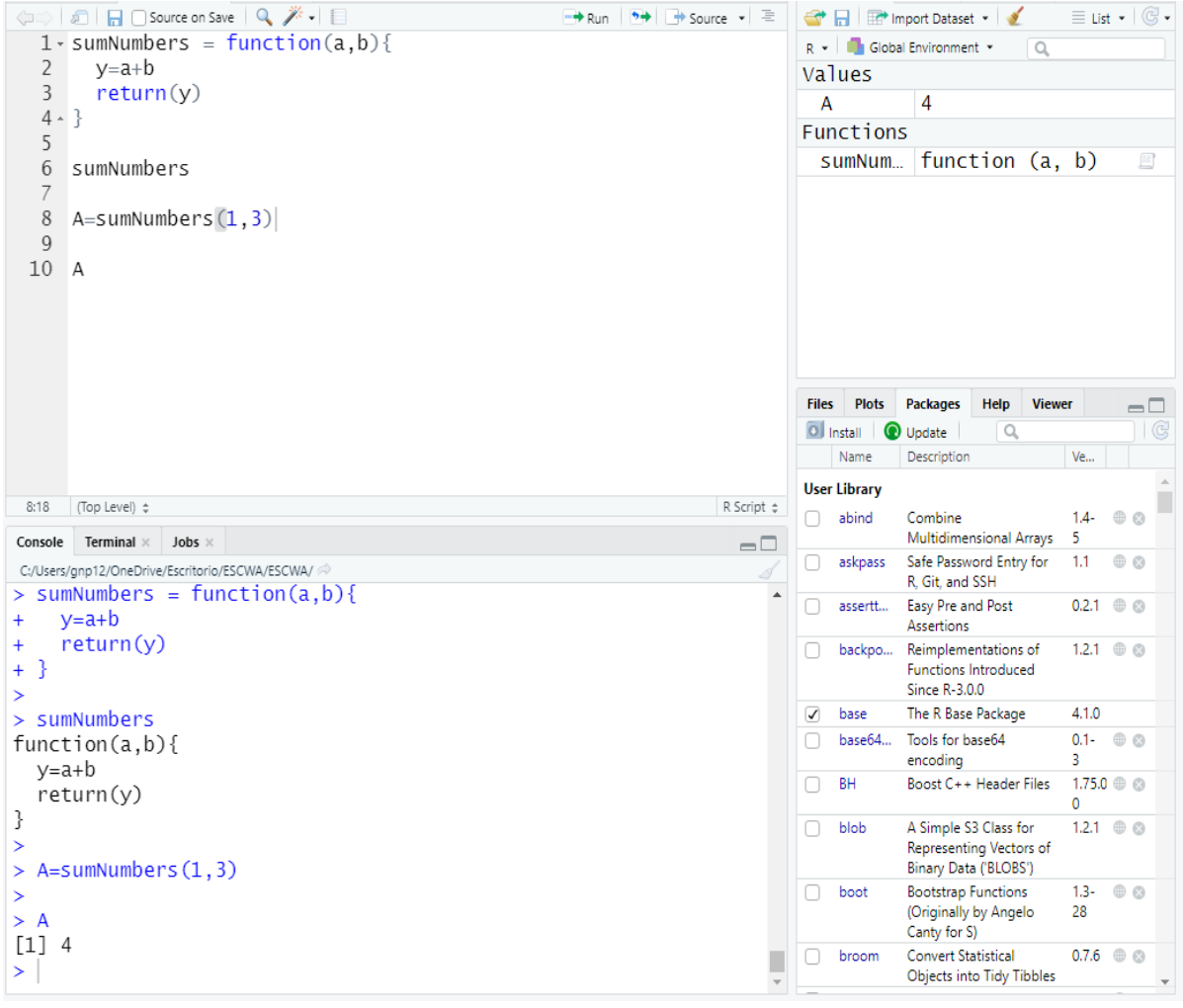
12. الدالة

الدالة هي رمز لغة R أو مجموعة من الإرشادات تهدف لتنفيذ مهمة معينة. يمكن أن تتلقى بعض قيم الإدخال ثم العمل معها للحصول على النتيجة المطلوبة ثم تعطي الرموز أو التعليمات البرمجية النتيجة للمستخدم. ويتمتع برنامج R بالعديد من الوظائف المدمجة، يسمح أيضاً للمستخدمين بإنشاء الوظائف الخاصة بهم وقد مروا على العديد منها، مثل rbind أو factor، إلاّ أنّ هذا القسم يقدّم المزيد من التفاصيل ذات الصلة.

يتم تمثيل دالة بما يلي:

```
Function-name = function (arg_1, arg_2, …) {Expression}
```

(أ) اسم **الدالة** (Function-name): اسم الدالة المخزنة في R؛
(ب) الوسيطات (Arguments): هي المدخلات التي تتطلب دالة، وهي اختيارية ويمكن أن يكون لها قيم افتراضية، إلاّ أنّه يجب تعريفها في حالة الحاجة إليها؛
(ج) التعبير (Expression): وهي مجموعة من العبارات التي تعرّف ما تقوم به الدالة.



في المثال السابق، تم استخدام البنية المشار إليها لإنشاء دالة. أولا، تبلّغ R أن الدالة تحتاج لإدخالين "a" و "b" ثم تشرح أنها تضيفهما في "y" وتبلّغ القيمة. وتشغيل الدالة فقط (السطور 1-4) لا يظهر النتائج على وحدة التحكم لأنها مجرّد تعليمات وليست عملية فعلية. ومع ذلك، سيظهر في الربع الثاني تعليق يشير إلى إنشاء هذه الدالة. التي إن قمت باستدعائها (سطر 6)، لن ترى سوى البرنامج النصي الخاص بها. ومن أجل استخدام الدالة، عليك أن تستدعيها عبر توفير اثنين من المدخلات على الأقل. لذلك، كما ترى في السطر 8، تُستدعى الدالة لإضافة 1 و3، وتحفظ ذلك في A، حيث أنّ قيمة A هي 4.

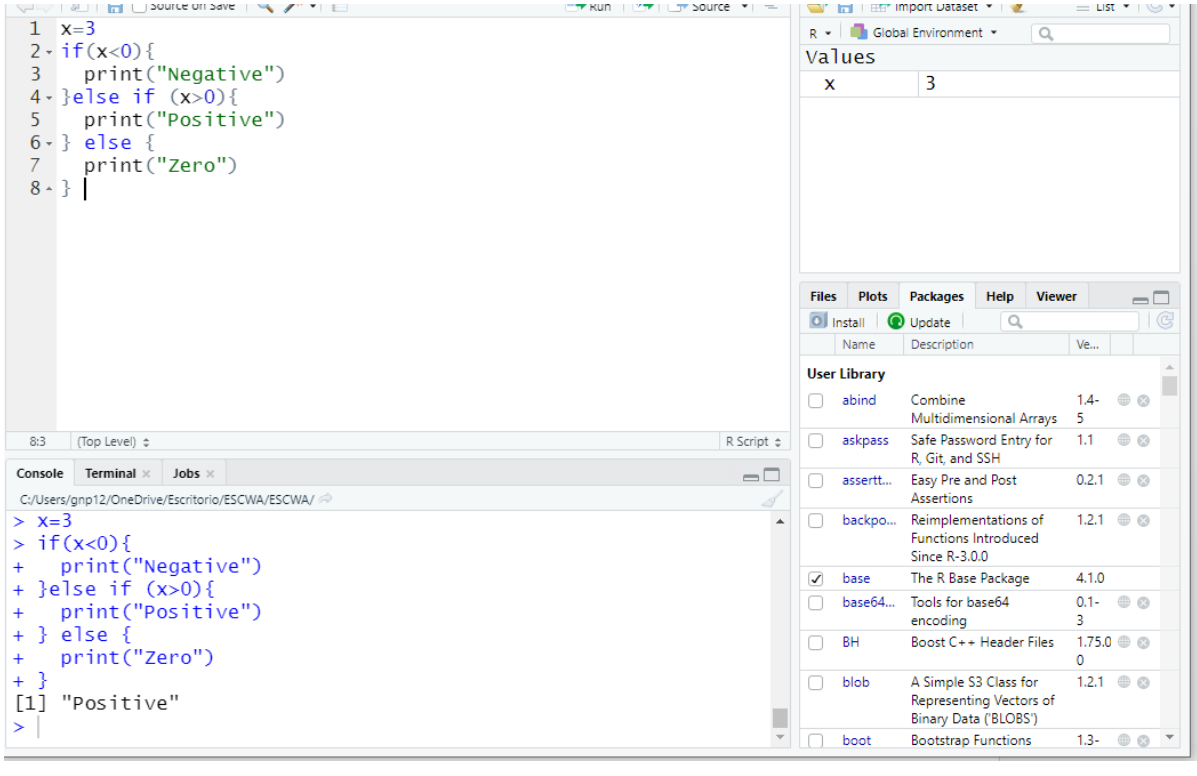
يجب أن تكون أفكارك منظمة، ولكن الدالة يمكن أن تساعد في الحصول على تعليمات برمجية أكثر دقة ووضوح. ويمكن ملاحظة أنه على الرغم من أن الدالة تستخدم متغير "y"، إلا أن هذا لا يظهر في النتيجة النهائية، باعتباره عملية داخلية.

13. عبارات شرطية

يمكن استخدام if/else لتقييم حالة، واعتماداً على ذلك، يُعمل على توجيه التعليمات البرمجية إلى مخرجين مختلفين. بعد تقييم الحالة (البيان 1 "statement1" في المثال أدناه)، إذا كانت القيمة تشير إلى عبارة "صحيح" أو "TRUE" يتم تنفيذ الإعلان الأول (البيان 2 "statement2"في المثال أدناه)؛ في المقابل، إذا كان القيمة "خطأ" أو "FALSE"، يتم تنفيذ الإعلان الثاني "Else" "آخر" (البيان 3 "statement3" في المثال أدناه). التركيبة الرئيسية هي:

```
If (statement1) {statement2} else {statement3}
```

في المثال أدناه، يقوم "if" (إذا) بتنفيذ التقييم التالي: إذا كانت قيمة x أقل من صفر، ستظهر علامة سالب "Negative" على الشاشة؛ إذا كانت قيمته أكبر من صفر، فستظهرعلامة موجب "Positive"؛ وإلا فإنه ستظهر علامة صفر "zero".



يظهر نموذج هذا المثال أن "else" لا شروط له لأنه ستتمّ طباعة كل ما يتلقاه طالما أنه لم يتم تلبية أي من التعليمات السابقة.

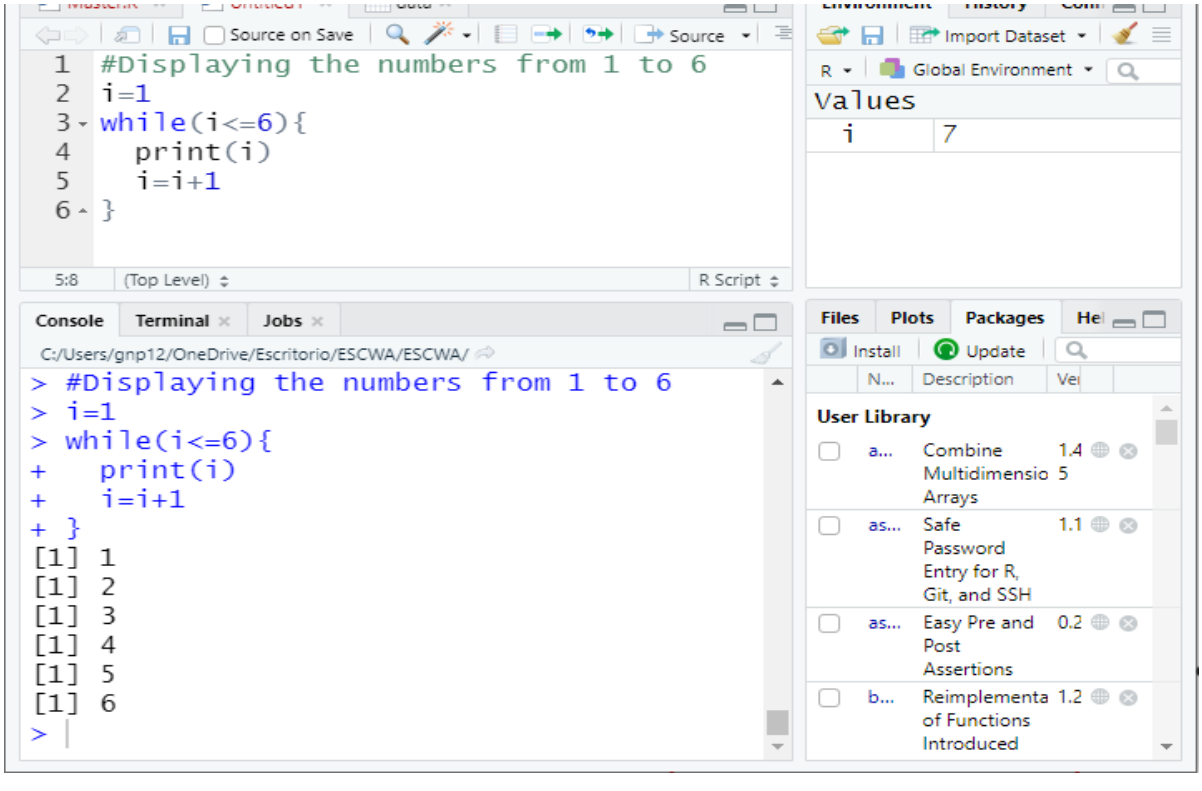
14. الحلقات

تحتاج بعض الرموز إلى تعليمات للتمكن من تكرارها عدة مرات. لهذه الحالة، يقدم R طريقتين للقيام بذلك بسهولة، وهي:

حلقة while

while (test_expression) {statement}

في هذه الحالة، يجب إنشاء شرط إيقاف وبمجرد استيفائه، تتوقّف التعليمات البرمجية داخل while عن التكرار.



في المثال، تعرض التعليمات البرمجية أرقاماً من 1 إلى 6. ومع ذلك، هناك عناصر هامة لا بد من تسليط الضوء عليها:

(أ) تبدأ تهيئة العداد في 1؛

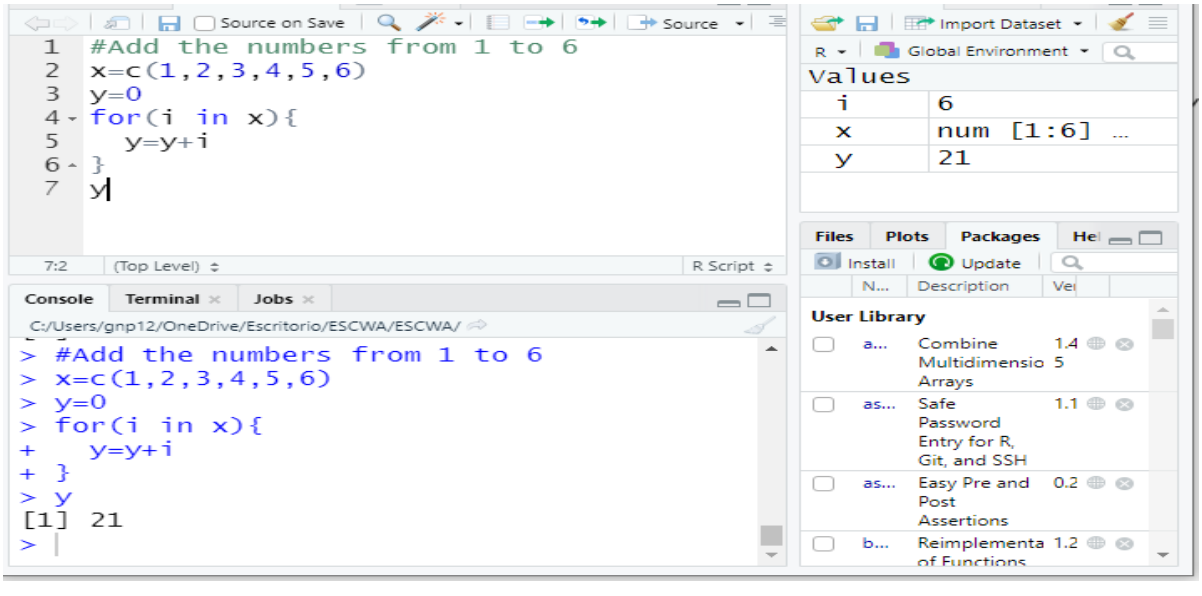
(ب) عند دخول حلقة while، وباستيفاء الشرط (أي i <= 6)، سيتم طباعة قيمة "i" وسوف تزداد القيمة. أمّا إذا لم تقم بإضافة شرط يمكن الوصول إليه، سيتم تشغيل الحاسوب من دون توقف؛

(ج) يظهر الربع الثاني النتيجة الأخيرة فقط بعد إضافة 1+، والتي تختلف عن النتيجة الأخيرة في وحدة التحكم التي تُظهر الرقم فقط قبل إضافة هذه القيمة (1 +).

15. حلقة For

for (element in a sequence) {statement}

لا تتمتّع الحلقة For بشرط إيقاف، ولكن يجب تحديد القيم الدقيقة التي تريد تقييمها.



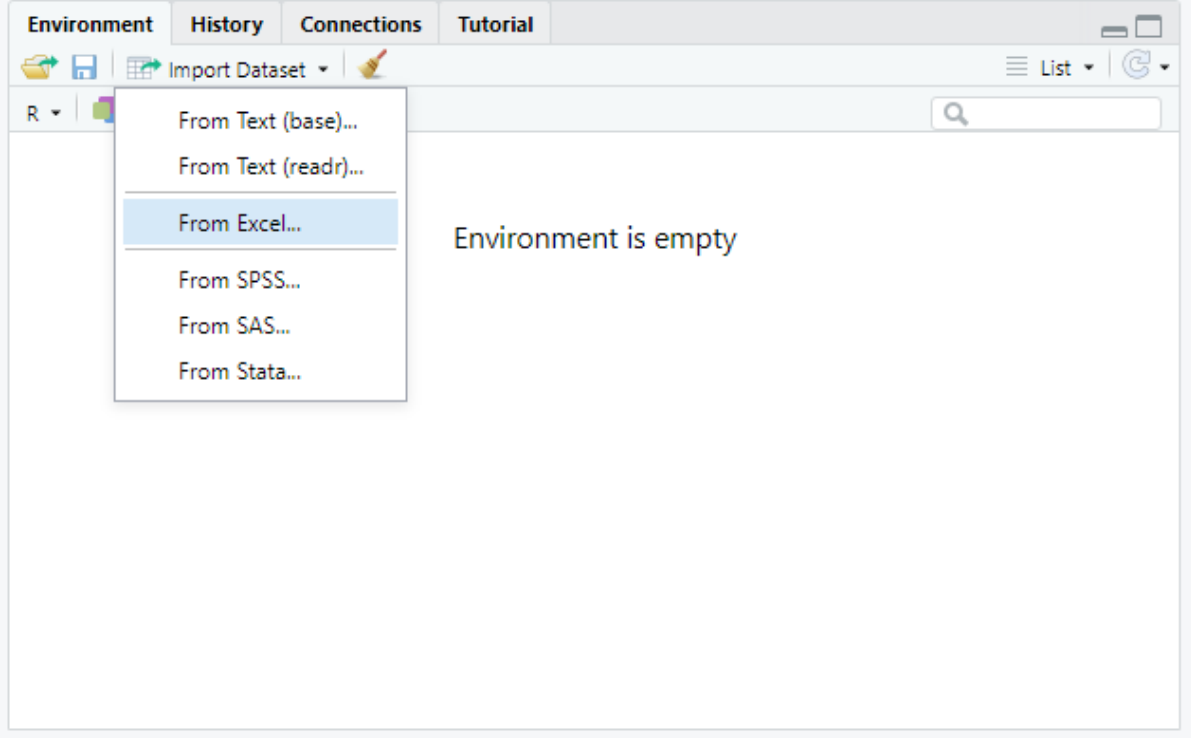
يلاحظ من التمرين السابق، أنه للقيام بالحلقة يجب تحديد متجه x مع جميع القيم ذات الصلة. ثم تبدأ بـ 0 كقيمة لـy، وفي كل خطوة تقوم بزيادة الإضافة على القيمة الأصلية y في كل من إدخالات x، حتى تحصل على النتيجة 21 أي مجموع: 0+1+2+3+4+5+6.

الفصل الثاني. إدارة البيانات

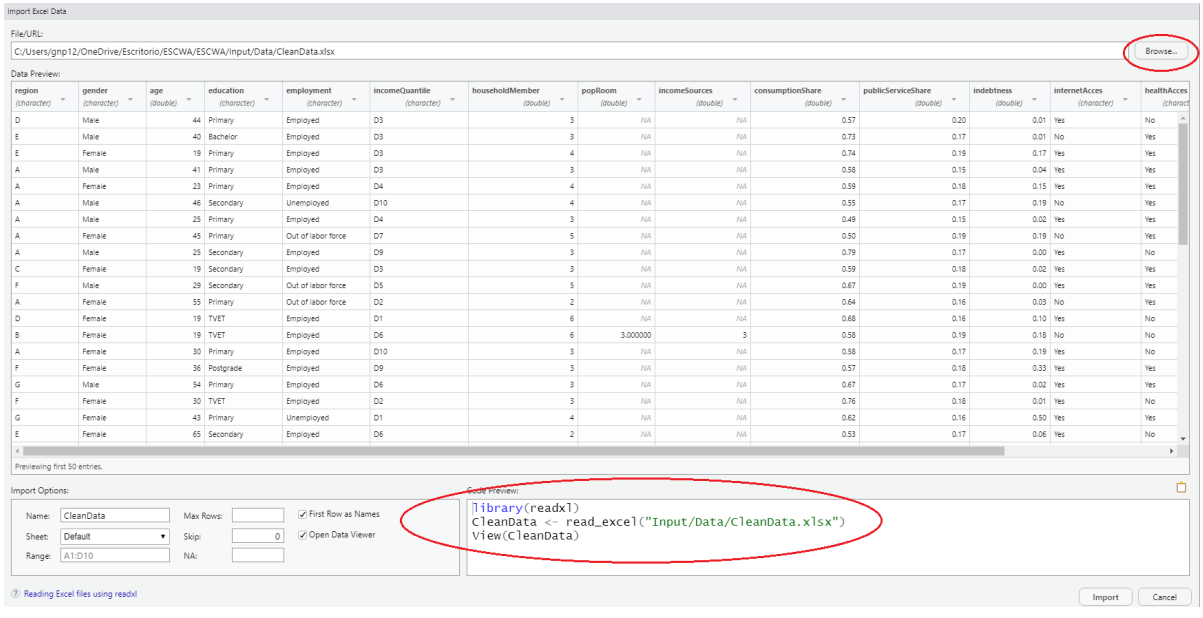
يهدف هذا القسم إلى العمل على كيفية إدارة مجموعات البيانات. لذلك، تم إنشاء حالة محاكاة لمساعدتك على فهم أفضل للوظائف الأكثر صلة بإدارة البيانات. خلال كل من التدريبات التالية يمكنك أن تطّلع على منظورات مختلفة من مجموعة البيانات التي ستفيد لاحقاً في تحليل برامج الحماية الاجتماعية.

1. تحميل البيانات

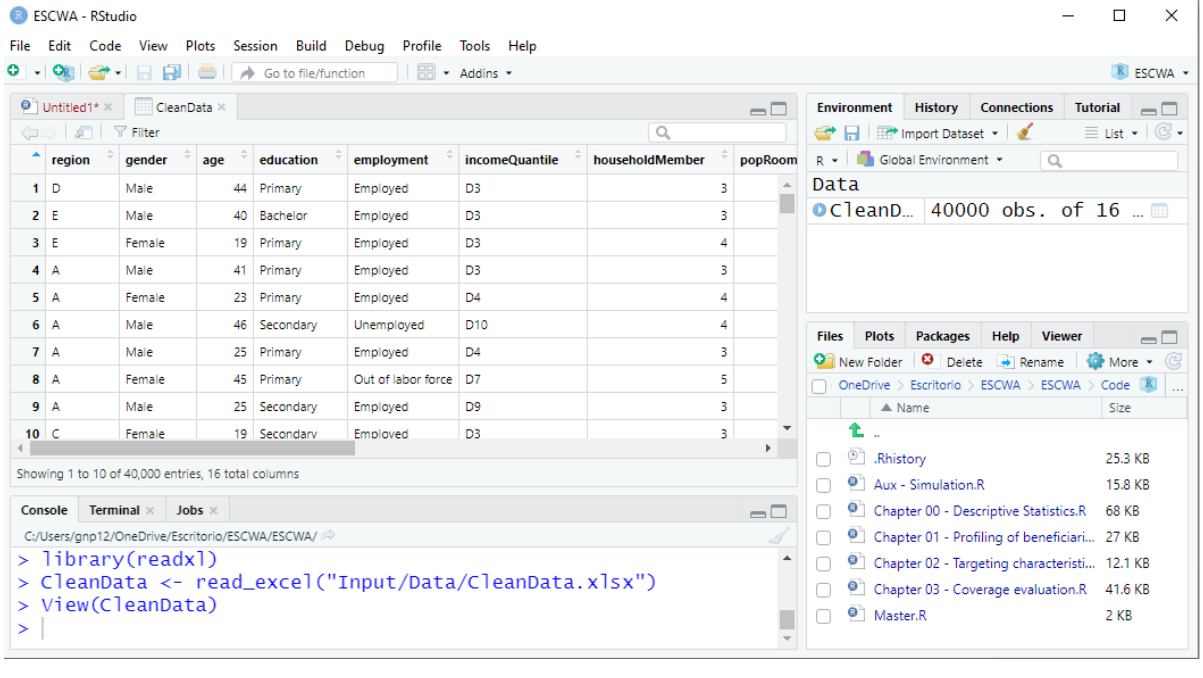
للبدء بتحليل مجموعة البيانات، تكمن الخطوة الأولى في معرفة كيفية تحميلها في الملف. لهذا التمرين، ستستخدم ملف "CleanData.xlsx" المتواجد داخل مجلد البيانات Data folder الموجود بدوره داخل مجلد المدخلات Input folder. بينما يستند هذا المثال إلى العمل في ملف (ايكسيل) Excel، نشير لإمكانية إجراء عملية مماثلة عن طريق تحميل البيانات من تنسيقات مختلفة.



لتحميل مجموعة البيانات، انتقل إلى الربع الثاني وانقر على "استيراد مجموعة البيانات" Import Dataset كما يظهر هذا المثال، ثم انقر على "from Excel...".



في هذه النافذة الناشئة، تظهر خاصية "تصفح" "Browse" حيث يمكنك العثور على مجموعة البيانات ذات الصلة وتحميلها. وبعد تحميلها والتحقق من جودة المعلومات، يمكنك النقر على "استيراد" "import". ومع ذلك، فمن المستحسن أن تطّلع على التعليمات البرمجية التي إن قمت بنسخها أو بنسخ الرموز في البرنامج النصي الرئيسي الخاص بك، فلن تحتاج إلى تشغيل كل شيء من الصفر مجدداً، ويمكنك تشغيله مباشرة من الربع الأول.



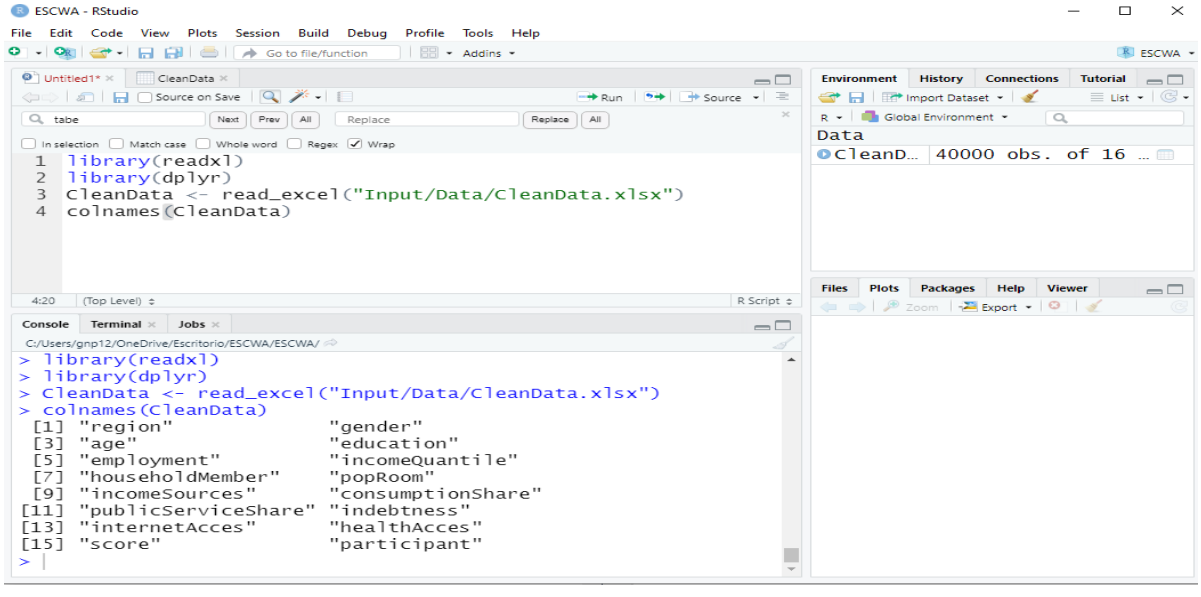
بإمكانك الآن تصور مجموعة البيانات الخاصة بك في R، وفي الربع الثاني ستجد ملخصاً سريعاً يشير لوجود 40000 ملاحظة من 16 متغير.

2. إدارة البيانات

لفهم مجموعة البيانات بشكل أفضل، تحتاج إلى العمل بشكل أكبر على محتواها. ولهذا الغرض يعتمد هذا القسم على الحزمة "dplyr". ومن أجل استخدام الحزمة المطلوبة، تحتاج لتحميل المكتبة "library" أولاً، لذا يرجى التأكد من تضمين هذا السطر الأول في أي رمز للغة R تقوم بإنشائه:

```
1  library(dplyr)
```

تُعتبر library dplyr مفيدة جدا لإدارة البيانات، لأنها تسمح بتنظيم العمليات المعقدة بواسطة تعليمات مجزأة بسيطة. ومن أجل القيام بذلك بطريقة صحيحة، لا بدّ من معرفة المتغيرات في مجموعة البيانات، وأفضل نقطة انطلاق هي معرفة المتغيرات التي تتضمنها.



وتجدر الإشارة إلى أننا نسخنا في سطر الأوامر هذا جزءاً من التعليمات البرمجية المستخدمة لتحميل البيانات لكي نتمكن من الآن فصاعداً من القيام بذلك تلقائياً، لكننا أضفنا سطر حزمة dplyr ليُصار إلى تحميلها تلقائياً. الرجاء ملاحظة أن مسار الملفات يعتمد على الحاسوب، بحيث يمكن أن يتغيّر النص الأخضر وفقاً للحاسوب المستخدم.

الدالة الجديدة هنا هي (colnames)، والتي تسمح بفهم أسماء المتغيرات داخل الملف بشكل أفضل. اعتماداً على مجموعة البيانات، تكون الأسماء مربكة أحياناً، لذا يوصى بتوضيح أي شك بالعودة إلى مصدر البيانات لفهم كل منها، ويصبح وصف مجموعة البيانات كالتالي:

وصف مجموعة البيانات

تم جمع مجموعة البيانات من برنامج الحماية الاجتماعية في مملكة X. وهي تشمل معلومات حول 40 000 فرد، و16 متغيرا. المتغيّرات المتعلّقة بالأفراد المشاركين في البرنامج مصدرها بيانات البرنامج، أمّا تلك التي تخصّ الأفراد سواء كانوا مشاركين أو غير مشاركين فمصدرها مختلف (مثلاً: الوزارة)

المنطقة: region تنقسم المملكة إلى 6 مناطق، تمثلها الحروف A وB وC وD وE وF وG.

الجنس: gender يمثل هذا المتغير جنس الأفراد "الذكور" "male" أو "الإناث" "female".

العمر: age متغير عددي مع العمر في حياة الفرد.

التعليم: education يصف الحد الأقصى لمستوى التعليم الذي يحققه الفرد. والفئات هي "الابتدائية" "primary"، "الثانوية" "secondary"، "بكالوريوس" "bachelor"، "التعليم والتدريب في المجال التقني والمهني" "TVET"، "الدراسات العليا "Postgrade"(نعم، يتخلل البيانات خطأ إملائي في "postgrade" سنقوم بتصحيحه في الخطوات التالية).

التوظيف: "employment" يُعنى بوصف حالة العمالة للفرد (في هذه الحالة، لا تشتمل التعليمات الفئات لذا سنتحقق منها بأنفسنا).

الدخل: "income quantile" وهو معدّل الدخل للفرد الواحد، يتراوح من D1 (الأفقر) إلى D10 (الأغنى).

أعضاء الأسرة المعيشية: "householdMember" ويمثّل عدد الأفراد في الأسرة المعيشية الواحدة.

المشاركين في السكن: "popRoom" هو عدد الأشخاص لكل غرفة. وهو متغيّر مخصّص فقط للأفراد الذين يشاركون في برنامج الحماية الاجتماعية.

مصادر الدخل: "incomeSources" وهو عدد مصادر الدخل في الأسرة. وهو متغيّر مخصّص فقط للأفراد الذين يشاركون في برنامج الحماية الاجتماعية.

حصّة الاستهلاك: "consumptionShare" وهو حصص النفقات المخصصة لاستهلاك الأغذية. تسجل المملكة هذا المتغير من مصدر مختلف، لذلك فهو يشمل جميع الأفراد، حتى لو لم يكونوا مسجّلين في البرنامج.

حصّة الخدمات العامة المشتركة: "publicServiceShare" وهي حصص النفقات المخصصة للخدمات العامة. تسجل المملكة هذا المتغير من مصدر مختلف، لذلك فهو يشمل جميع الأفراد، حتى لو لم يكونوا مسجّلين في البرنامج.

المديونية: "indebtness" يتمّ احتساب مؤشر المديونية من قبل القطاع المالي. تسجل المملكة هذا المتغير من مصدر مختلف، لذلك فهو يشمل جميع الأفراد، حتى لو لم يكونوا مسجّلين في البرنامج.

خدمة الانترنت: "internetAccess" "نعم" "yes" إذا كان للفرد امكانية الوصول إلى الإنترنت، "لا" no إذ لم يكن لديه هذه الامكانية. تسجل المملكة هذا المتغير من مصدر مختلف، لذلك فهو يشمل جميع الأفراد، حتى لو لم يكونوا مسجّلين في البرنامج.

النتيجة: باستخدام المتغيرات السبعة السابقة، تقوم الحكومة باحتساب النقاط، واستناداً إلى النتيجة يمكن للناس الوصول إلى البرنامج، وهذه القيمة متاحة فقط للمستفيدين.

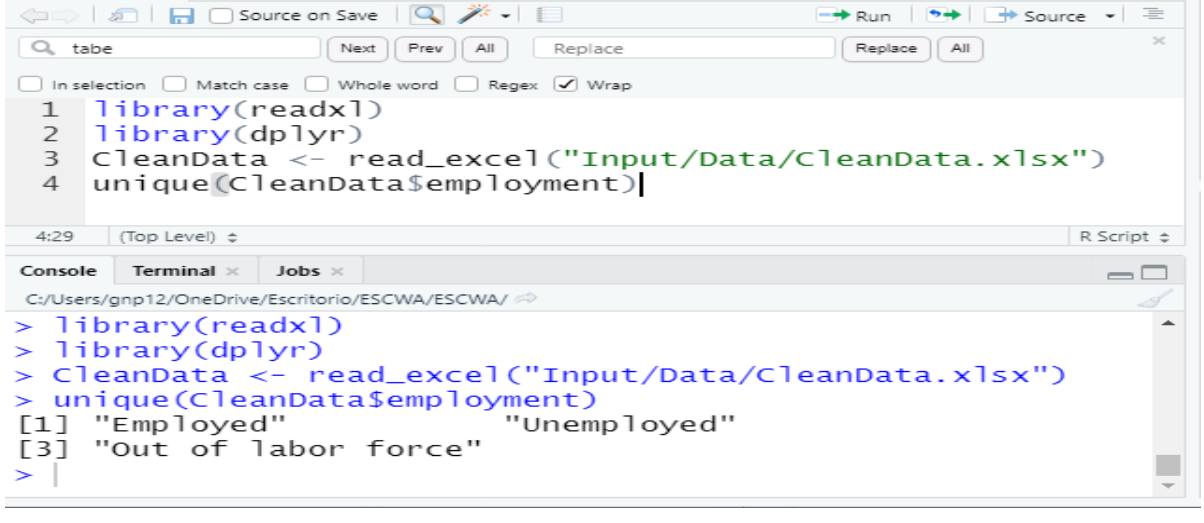
المشارك: "نعم" "yes" إذا كان الفرد يشارك في البرنامج، "لا" no إذا كان الفرد لا يشارك في البرنامج.

أثناء فحص قائمة المتغيرات، يتضح سبب ظهور بعض المتغيّرات على أنها غير محددة أو NA في مجموعة البيانات.

	region	gender	age	education	employment	incomeQuantile	householdMember	popRoom	ir
26324	F	Female	42	Primary	Employed	D1	3	NA	
26325	A	Male	31	Primary	Employed	D9	3	NA	
26326	D	Male	19	Secondary	Unemployed	D8	2	0.6666667	
26327	B	Male	36	Secondary	Employed	D9	2	NA	
26328	B	Male	19	Secondary	Employed	D10	2	NA	
26329	D	Male	23	TVET	Employed	D3	5	NA	
26330	F	Female	57	Primary	Employed	D6	5	NA	
26331	B	Male	28	Secondary	Employed	D6	2	NA	
26332	A	Male	40	TVET	Employed	D7	3	NA	
26333	F	Female	50	Secondary	Out of labor force	D2	2	NA	
26334	D	Male	19	Secondary	Employed	D7	1	NA	
26335	G	Male	28	TVET	Employed	D6	3	0.7500000	

مثلاً، إن الملاحظة 26325 غير محددة NA بالنسبة للمشاركين في السكن popRoom، ولكن الملاحظة 26326 لها قيمة محددة. ممّا يعني أن الأولى تعود لشخص غير مستفيد من البرنامج، بينما تعود الثانية لمستفيد منه

كما سمح لنا وصف مجموعة البيانات برصد العناصر التي نرغب بتصحيحها كالخطأ الإملائي في التعليم education، وزيادة بعض المعلومات الإضافية التي نود جمعها، كفئات التوظيف. وكما وضّحنا في الفصل السابق، يمكن استخدام بعض الأوامر commands من إطار البيانات مباشرة كما لو كانت تشير إلى متجهات، طالما يتم ذكرها بشكل صحيح.



ولحل مشكلة المعلومات غير المحددة، تُستخدم دالة المتجه (unique) أي (فريد) عبر تطبيق هذا الأمر على متجه التوظيف، فتُظهر وحدة التحكم أن هذا المتغير يحتوي على ثلاث فئات فقط.

ويمكن تصحيح الاسم باستخدام هذه الوظائف الأساسية، ويُنصح في هذه الحالة بالتحديد باستخدام حزمة dplyr التي من المؤكد أنها تجعل الأمور أسهل. ولكن قبل استخدامها، من المهم إدخال الرمز "%>%" الذي يعرف بـ "pipeline". وبمجرد استخدامه يصبح بالإمكان استخدام كل ما هو في يسار الرمز كمدخل إلى الدالة الموجودة على اليمين.

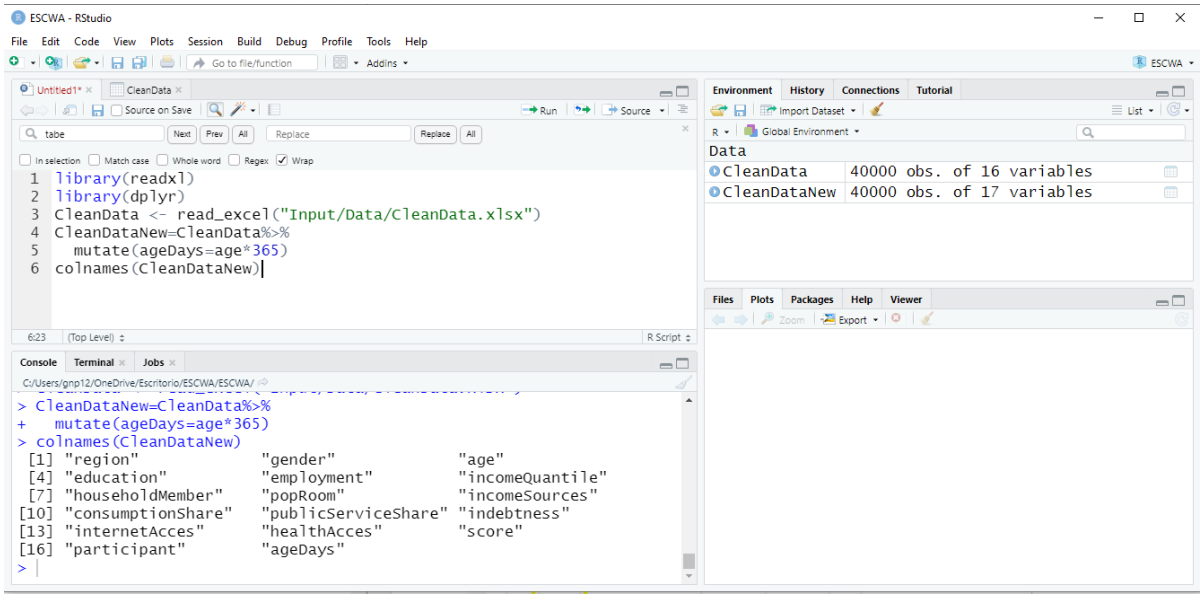
```
1  library(readxl)
2  library(dplyr)
3  CleanData <- read_excel("Input/Data/CleanData.xlsx")
4  unique(CleanData$education)
5  CleanDataNew=CleanData%>%
6    mutate(education=ifelse(education%in%c("Postgrade"),"Postgrad",education))
7  unique(CleanDataNew$education)
```

```
7:31  (Top Level)                                                    R Script
Console  Terminal  Jobs
C:/Users/gnp12/OneDrive/Escritorio/ESCWA/ESCWA/
> library(readxl)
> library(dplyr)
> CleanData <- read_excel("Input/Data/CleanData.xlsx")
> unique(CleanData$education)
[1] "Primary"   "Bachelor"  "Secondary" "TVET"      "Postgrade"
> CleanDataNew=CleanData%>%
+   mutate(education=ifelse(education%in%c("Postgrade"),"Postgrad",education))
> unique(CleanDataNew$education)
[1] "Primary"   "Bachelor"  "Secondary" "TVET"      "Postgrad"
>
```

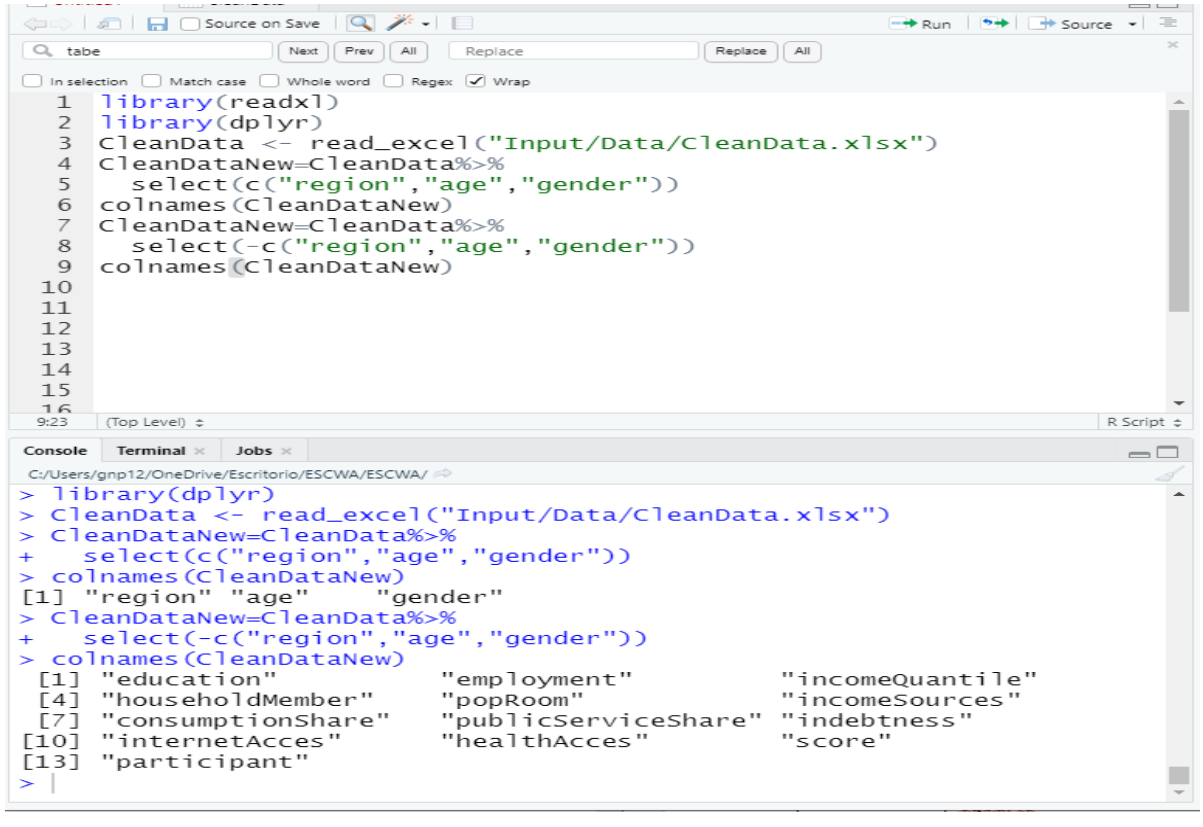
في هذا المثال، يتم التركيز على السطرين 5 و6. في السطر 5، يتم إنشاء إطار بيانات جديد يسمى CleanDataNew المماثل لـ CleanData، ولكن مع إضافة شرائح وتعليمات جديدة في السطر السادس. هذه التعليمات هي الدالة (mutate)، والتي تُتيح تعديل المتغيرات الموجودة أو إنشاء أخرى وفقاً لمعايير معينة. وسيكون لها في هذه الحالة بحدّ ذاتها وظيفة تسمى (ifelse)، وكما يوحي الاسم، تحتوي على ثلاثة أجزاء: الجزء الأول، هو الحالة condition، التي تنطوي على القيم في التعليم لاسيما فئة "الدراسات العليا" Postgrade. الجزء الثاني، هو ما يحدث إذا ما كانت الحالة صحيحة، هنا نريد تغيير التهجئة إلى "postgrad". أمّا في الجزء الثالث، إذا لم يتحقق الأمر، ستعمل على أن يكون لها قيمة مماثلة لما كانت عليه في الأصل (أي التعليم). بهذه الطريقة، مع مجموعة من الأوامر البديهية تم تصحيح الخطأ الإملائي.

3. وضع متغيّرات جديدة

الدالة mutate هي خاصية متاحة دائماً للمستخدم(ة) وتستعمل لأكثر من غرض، وتسمح أيضا بخلق متغيرات جديدة. في المثال التالي يُستخدم الأمر لإنشاء متغير جديد يقيس العمر بالأيام. ولهذا الغرض، تم إنشاء متغير جديد باسم "ageDays". حيث تقوم الدالة (mutate) بتنفيذ هذه المهمة تلقائياً وتضع هذا المتغير الجديد في نهاية مجموعة البيانات. $age \times 365$.



أحياناً، تكون مجموعات البيانات كبيرة جداً وتتضمن متغيرات متعددة غير ذات صلة بالتمرين. في هذه الحالات، تساعد بعض الدالات على تحديد المتغيرات المطلوبة أو إزالة غير المطلوب منها.



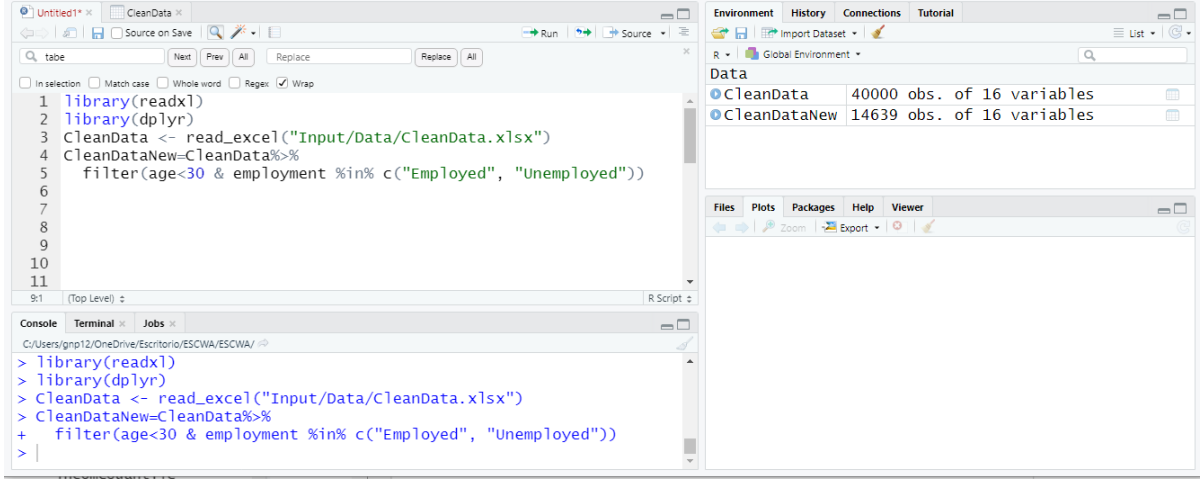
4. تحديد المتغيرات

وكما يُوضّح المثال السابق، فالوظيفة الرئيسية هي (select). إلاّ أنّه يمكن استخدامها بشكل مختلف. في الحالة الأولى، يتم تحديد المتغيرات الثلاثة التي تريد إظهارها فقط (ولاحظ أنها تظهر بنفس الترتيب الذي حددته لها). أمّا في الحالة الثانية، يضيف رمز لغة R علامة "-" قبل المتجه، وهذا يسمح للبرنامج بمعرفة المطلوب وهو إزالة هذه المتغيرات من مجموعة البيانات.

5. تصفية المتغيرات

على غرار فائض الأعمدة، يوجد أحياناً ملاحظات أكثر من تلك المطلوبة في مجموعة البيانات، ويصبح في هذه الحالة عامل التصفية (filter) أداة مفيدة لتحديد الملاحظات التي تفي بالشروط المطلوبة فقط. والتعليمات المنطقية الأكثر فائدة هنا هي:

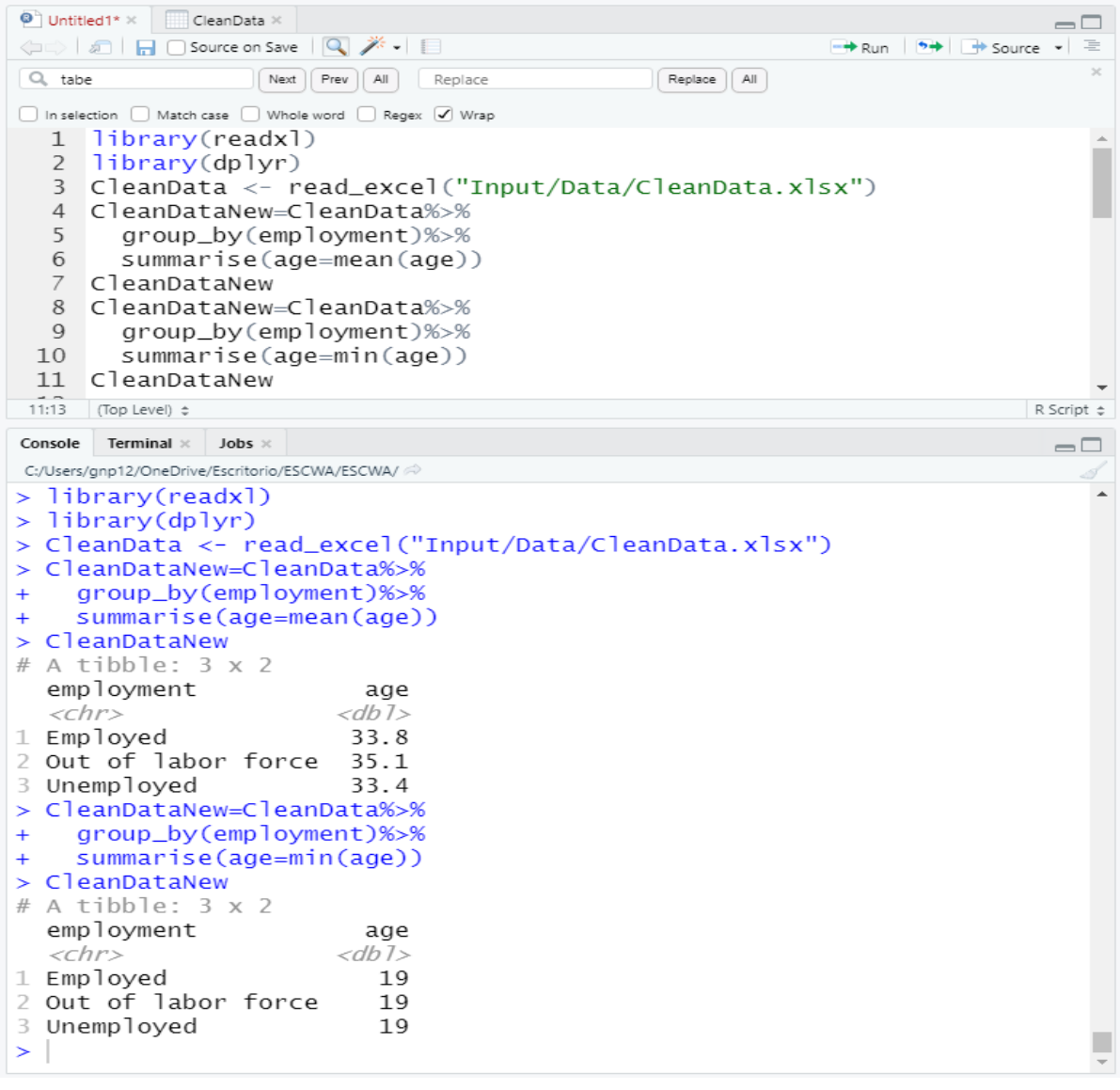
<	smaller than "أصغر"
>	greater than "أكبر"
= =	equal to "يساوي"
< =	less or equal to "أقل أو يساوي"
> =	greater or equal to "أكبر أو يساوي"
"%in%في"	القيمة في المتجه استخدمنا هذا الأمر من قبل في (ifelse)
&	and "و"
|	or "أو"



وبناءً عليه، تمّ تصفية البيانات ولم يتم اختيار سوى الأفراد الذين يقل سنهم عن 30 عاماً سواء كانوا عاملين أو عاطلين عن العمل. بينما يمكنك التحقق من البيانات ورؤية كيفية تغيرها، من الجيد ملاحظة أن مجموعة البيانات الجديدة تحتوي على 14,639 ملاحظة فقط بدلا من 40,000.

6. تلخيص المتغيرات

يعتبر تلخيص المتغيرات من الممارسات الشائعة شرط أتباعها لمجموعة معايير محددة. على أنّه في هذه الحالة يمكن استخدام اثنتين من الوظائف. الأولى هي (group_by) التي، كما يوحي الاسم، تُخبر R أن كل أمر بعد ذلك يتم بواسطة مجموعات. أمّا الثانية فهي (summarise)، التي تساعد على تصغير البيانات وفقاً للمعايير المحددة.

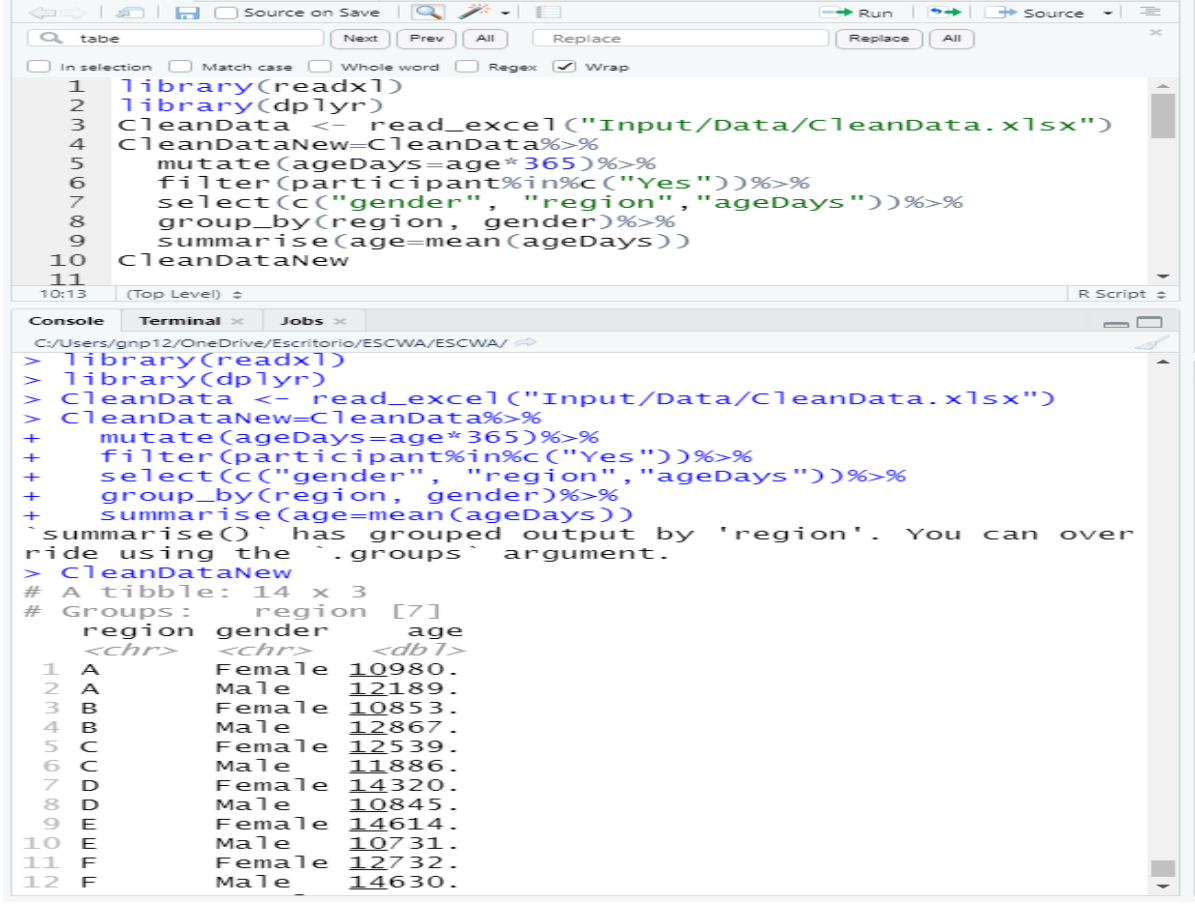


يستخدم المثال السابق هاتين الدالتين بطريقتين مختلفتين. تقوم الطريقة الأولى على تجميع البيانات وفقاً لحالة التوظيف واحتساب متوسط عمر كل مجموعة. أما الثانية فتحستب عمر أصغر فرد في كل مجموعة. ويستكشف هذا الأمر عنصرين جديدين:

(أ) يؤدي أمر التلخيص "summarise " إلى تغييرات جذرية في بنية البيانات. وبعد تطبيقه، ستظهر ملاحظة واحدة فقط لكل عضو في المجموعة (3)؛
(ب) يتم إضافة خطين متسلسلين (pipeline) للتمكن من تطوير التمرين.

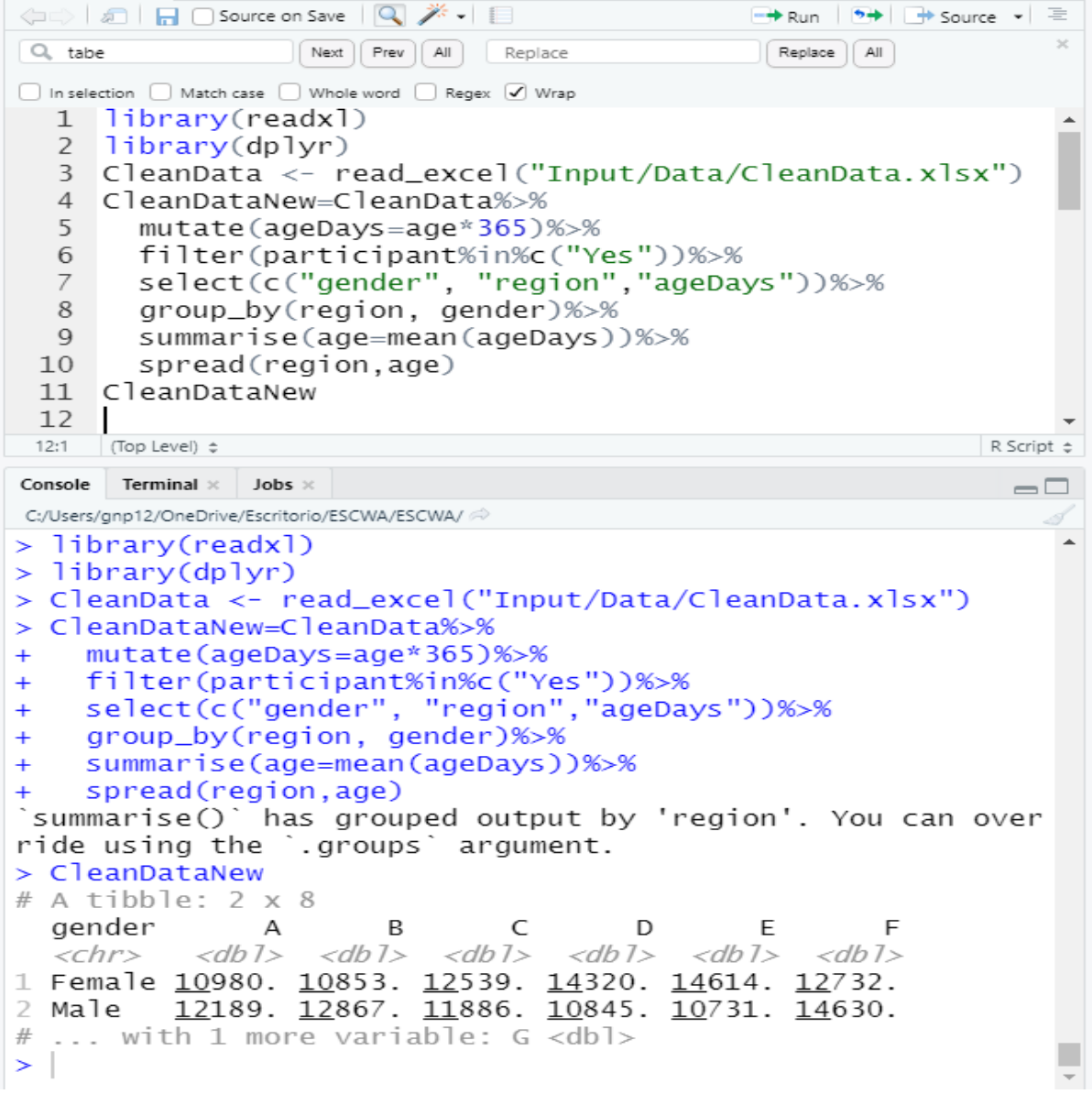
وباستكشاف الملاحظة الأخيرة، يمكن مشاهدة القيم، مع الأخذ بعين الاعتبار التعليمات التالية:

(أ) إنشاء متغير العمر وفقاً للأيام؛
(ب) اختر حصراً الأشخاص المستفيدين؛
(ج) اختر متغيرات العمر age (بالأيام) والجنس gender والمنطقة region فقط؛
(د) احتسب متوسط عمر المستفيدين (بالأيام) وفقاً لنوع الجنس والمنطقة.



في هذه الحالة، يوضح المثال السابق كيفية تنفيذ الإرشادات الأربعة في سطر أوامر واحد.

يُفيد هذا الجدول في عمليات الحوسبة المشار إليها في الفصل التالي، إلاّ أنّ التنسيق غير مريح إلى حد ما بالنسبة للقارئ. لذلك، يفضل العديد من المستخدمين عرض الجدول بتنسيق واسع (على عكس التنسيق الطويل الذي يتم إنتاجه حاليا). وتدعم الدالة (spread) هذا المطلب.

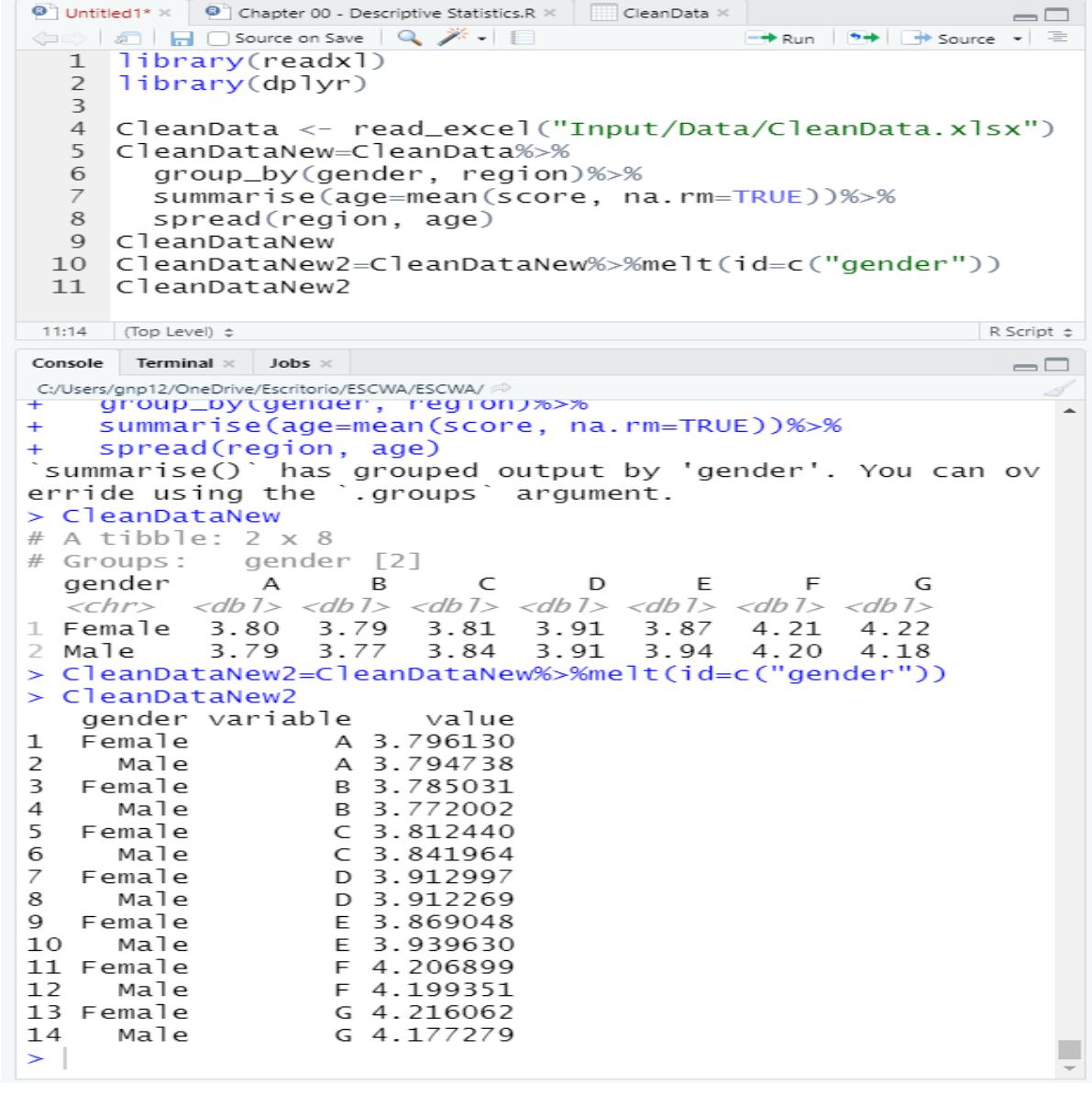


كما يلاحظ من التمرين السابق، فإن هذا التنسيق مريح أكثر للقارئ. الرجاء ملاحظة أمرين:

تستخدم دالة (spread) وسيطتين، الأولى هي المتغير الذي تستخدمه لتوسيع التنسيق (في كل عامود)، والثانية هي القيمة التي تعمل على محتوى الجدول (في هذه الحالة العمر).

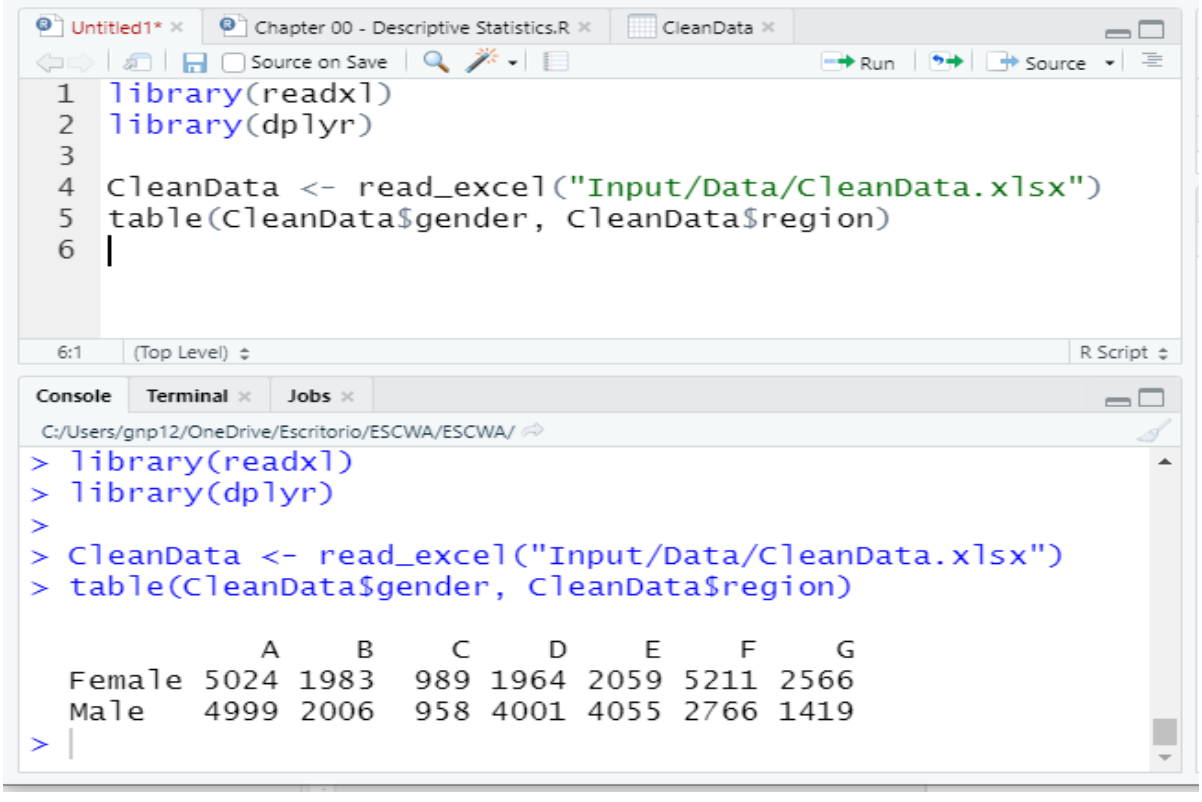
أحياناً، لا تعرض وحدة التحكم في R بعض الأجزاء في إطار البيانات ممّا لا يعني أن العامود أو الصف قد اختفيا. فبرنامج R يستخدمهما بشكل داخلي، كي لا يشكل ذلك ازعاجاً نظراً إلى حجم الشاشة وتفضيل القارئ عدم رؤيتهما. ملاحظتان أخيرتان:

أولاً، الدالة المعاكسة لـ (spread) هي (melt).



وكما يُلاحظ في المثال، تعيد الدالة (melt) هيكلة البيانات في تنسيق طويل، ولكن خلال تلك العملية يتغير اسم المتغيرات، لذا كن حذراً.

ثانياً، تظهر أحياناً الحاجة إلى إنشاء جداول، وبدلاً من اتخاذ مسار طويل من خلال تحويلات تلخص القيم، خُصّصت بعض الوظائف المحددة مسبقاً للقيام بذلك.



في هذا المثال وعلى نقيض الحالات السابقة، يساعد استخدام وظيفة مباشرة دون dplyr في الحصول بشكل أسرع على الجدول المطلوب مع الملاحظات الخاصة بكل فئة.

عموماً، يتم اختيار الوظائف المقدمة في هذا الفصل استناداً إلى فائدة الرموز ذات الصلة بـ SPP-RAF، ويمكن للعديد من الوظائف الأخرى التي يمكنها مساعدة القارئ على تحسين معرفته في إدارة قواعد البيانات. بالنسبة لهؤلاء، يوصى بنقاط انطلاق للتعلم ذي المستوى المتقدم:

(أ) https://cran.r-project.org/web/packages/dplyr/vignettes/dplyr.html؛

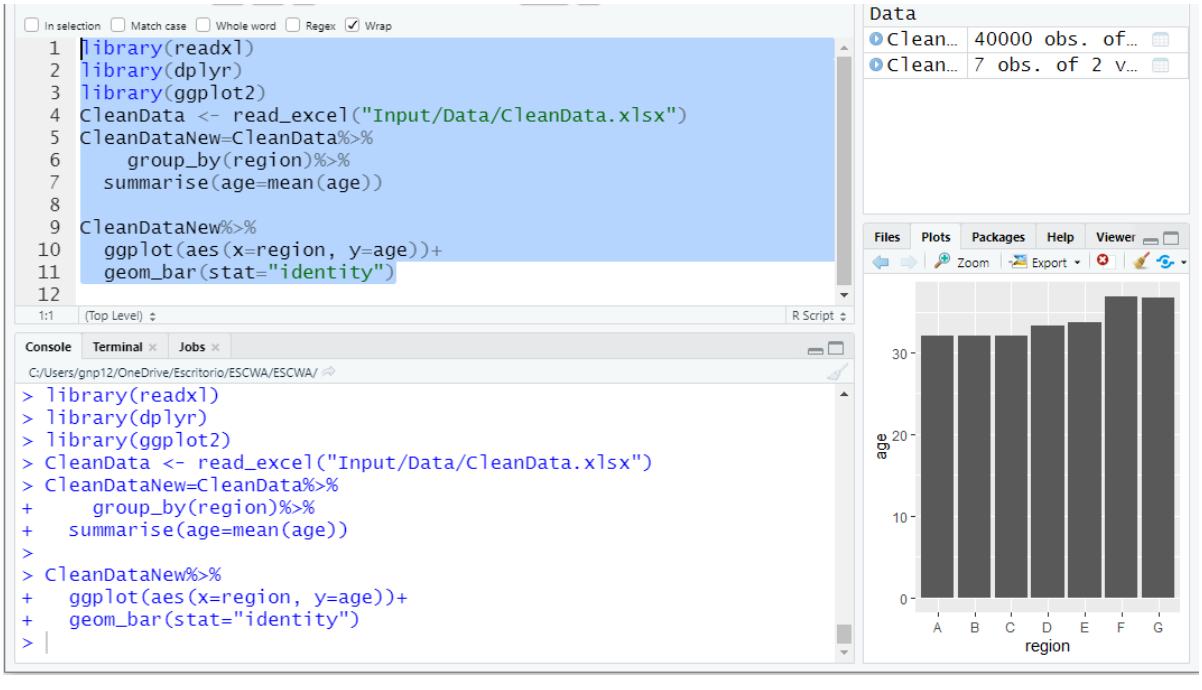
(ب) https://datacarpentry.org/R-ecology-lesson/03-dplyr.html.

الفصل الثالث. الرسوم البيانية

يقال بأن الصورة تساوي ألف كلمة وذلك للدلالة على أهميّتها. تعتبر عملية تمثيل المعلومات في الرسوم البيانية والأشكال فنّاً، وعلى المتدرب أن يدرك أنه اعتماداً على نوع البيانات والقيم التي يريد تسليط الضوء عليها، تؤدي بعض الرسوم البيانية دوراً أفضل من غيرها. وتسمح حزمة "ggplot2" بالوصول إلى مجموعة واسعة من الرسوم البيانية مع العديد من الخيارات المرنة التي تمكنك من توحيد عمليات الإبلاغ مع ايلاء اهمية للتصميم. هناك عدد كبير من التصاميم للرسوم البيانية وتظهر في كل يوم رسوم بيانية جديدة. ويشرح هذا القسم العناصر الأساسية لحزمة ggplot2 ثم يرشد القارئ حول كيفية العمل مع الرسوم البيانية. ونظراً لمرونتها، تقوم البرامج المختلفة بتدريس هذه الحزمة بطرق مختلفة، ولكن يتوقّع أن تؤدي الطريقة المعتمدة هنا تفسيراً أسهل للمدونات وتسهيل العملية على المتدرّبين الجدد.

1. العناصر الأساسية

لفهم العناصر الأساسية لـ ggplot2، دعونا نبدأ باستعراض متوسط عمر الأفراد على أساس منطقتهم في رسم شريطي.



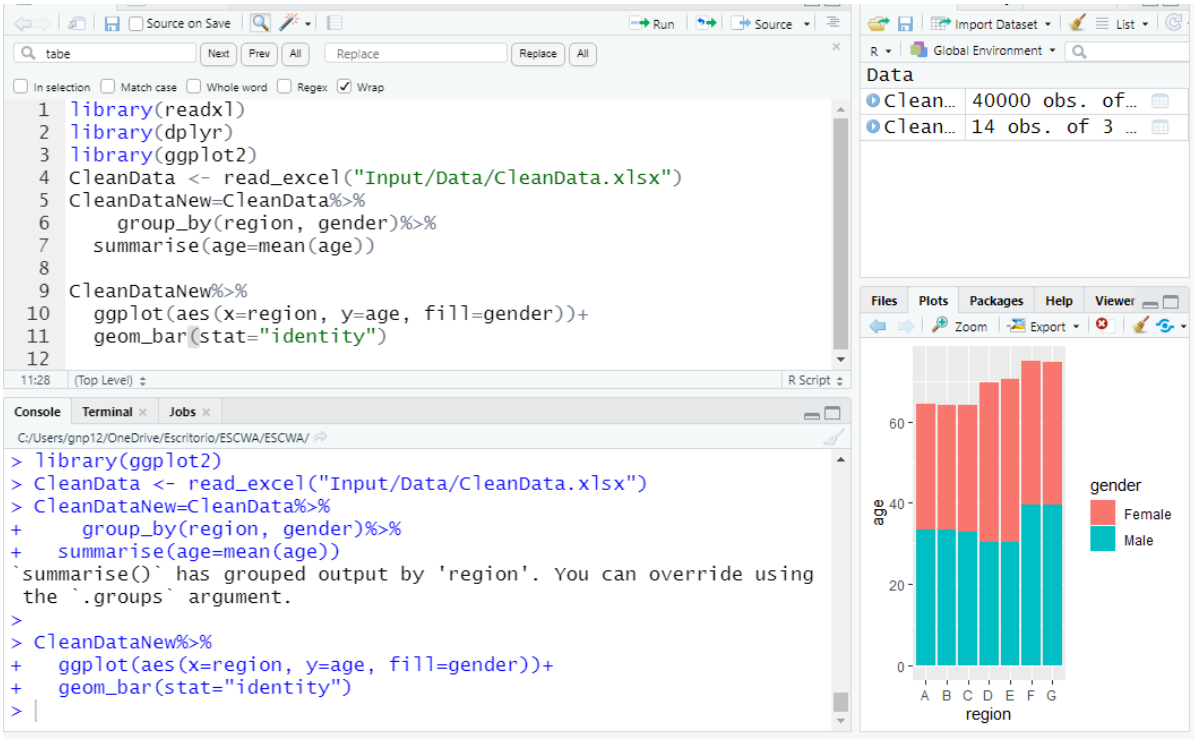
لإنشاء الرسم البياني، يمكن القيام بثلاث خطوات:

(أ) خط يربط البيانات التي ستستخدم في الرسم مع الدالة (ggplot)؛

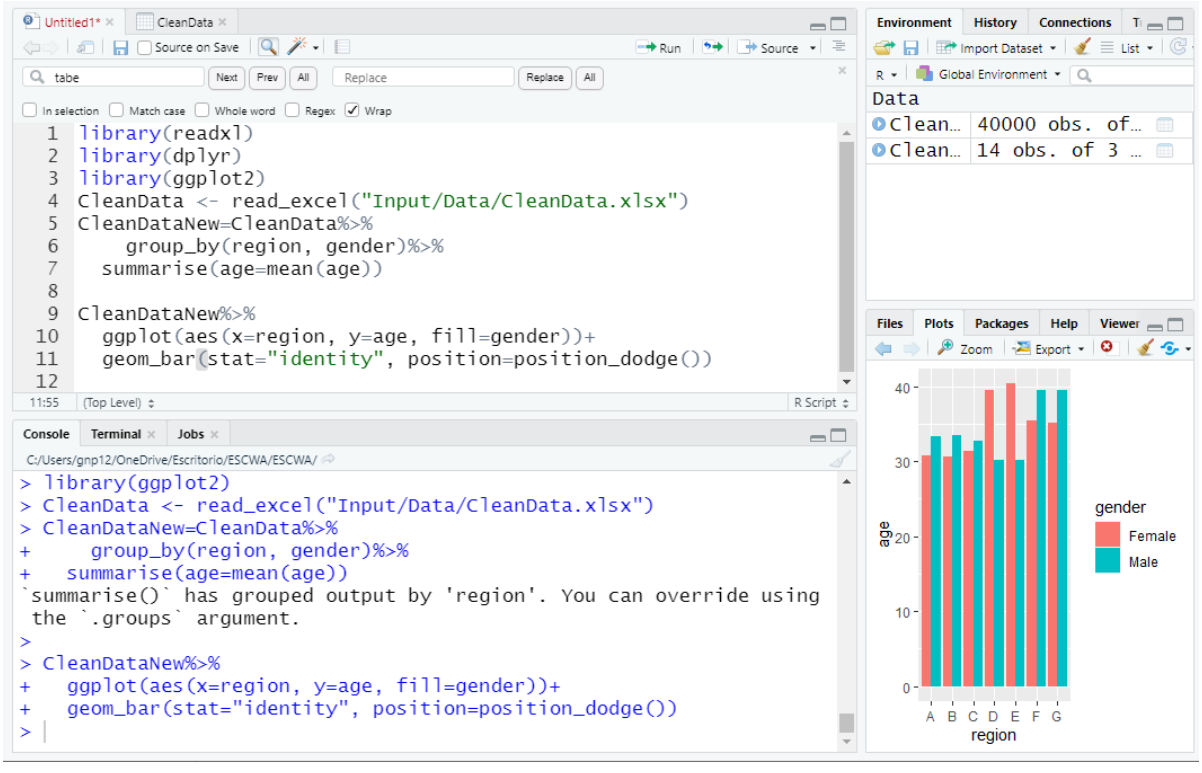
(ب) داخل دالة (ggplot) توجد دالة أخرى اسمها (aes) تسمح للمستخدم بتحديد أسماء المتغيّرات في المحورين X وY؛

(ج) ينتهي سطر الأوامر بعلامة "+" التي تعمل كخط اتصال بالتعليمات التي تليها في الرسم البياني؛ في هذه الحالة، تحدّد التعليمات شكل الرسم البياني والذي هو في هذه الحالة رسماً شريطياً، وبالتالي فإن الدالة التي تتولّى ذلك هي (geom_bar).

لرفع مستوى التحدي، فلنفترض أن المهمة تنحصر بإنتاج رسم شريطي يعرض متوسط عمر الأفراد استناداً إلى منطقتهم وجنسهم.



هناك تغيير واحد فقط بين هذا الرسم البياني والرسم السابق. حيث يقترن عامل إضافي في دالة (aes) بهذا المتغير الجديد. الرسوم البيانية المختلفة لها سمات مختلفة، سيتم عرضها في الأمثلة التالية، ولكن تسمح حزمة ggplot2 بالقيام بذلك بطريقة سهلة بحيث أنها تجمع كل الخصائص في دالة (aes) ليعرف المستخدم أين وضعها. لذلك، لا يمكن اعتبار الرسم البياني الحالي غير جيد، فمن غير المنطقي مراكمة أعمار النساء والرجال، بما يجعل الرسم البياني مربكاً وغير ذي فائدة.



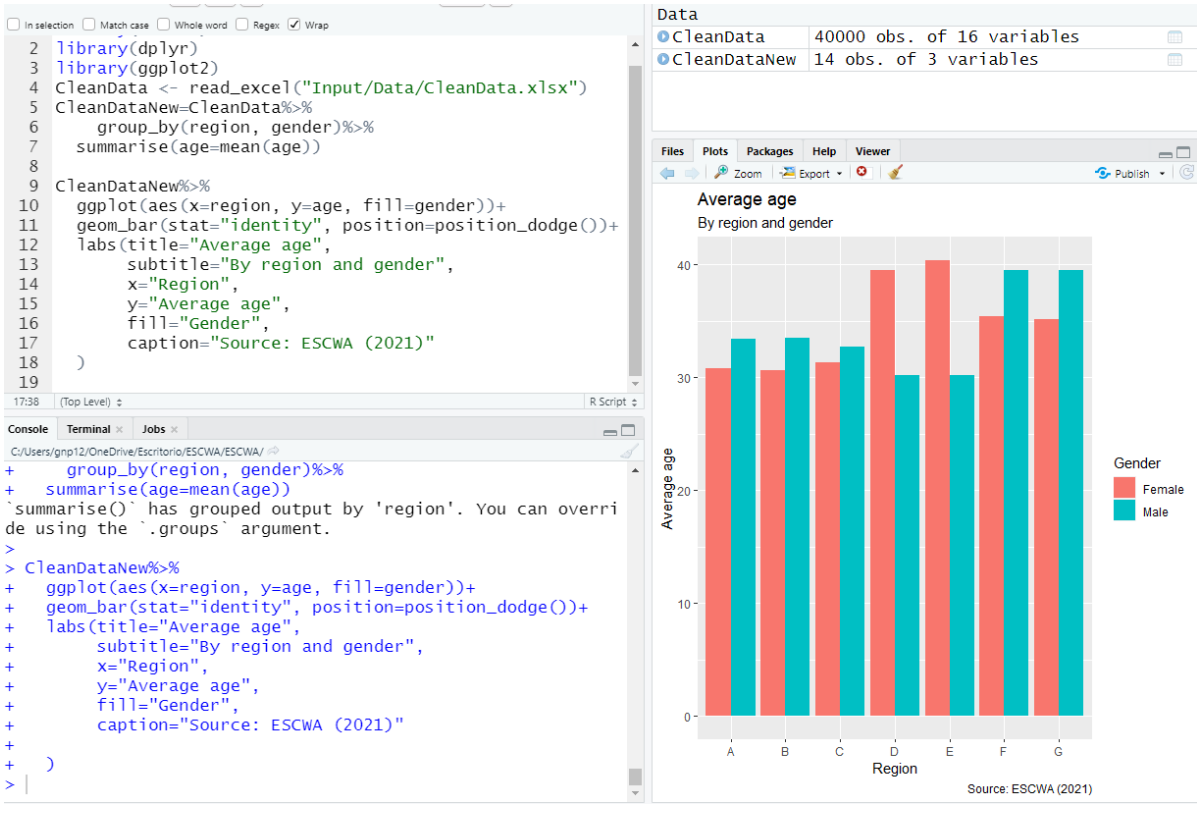
من خلال تحديد الوضع على أنه "position_dodge" داخل الدالة، يصبح الرسم البياني أفضل ويتم تفسير القيم بسهولة، مما يدل على أن المنطقتين D وE تميلان لأن يكون لديهما عدد أكبر من الإناث، في حين أن في المنطقتين F وG عدداً أكبر من الذكور. وبالمثل، ترتفع الفجوة بين المنطقتين ألف وباء بالنسبة للرجال، ولكن الفرق ليس كبيرا.

2. التنسيق

بعد تصميم البنية الأولية للرسم البياني، لا بدّ من تنسيقه ليبدو جذاباً ويحتوي على المعلومات ذات الصلة. تستند بعض النصائح الموضحة هنا إلى المبادئ التوجيهية التحريرية للبنك الدولي التي تتيح (بعد تعديلات قليلة) إنشاء رسوم بيانية احترافية يمكن دمجها بسهولة في تقرير (2016).

3. التسميات

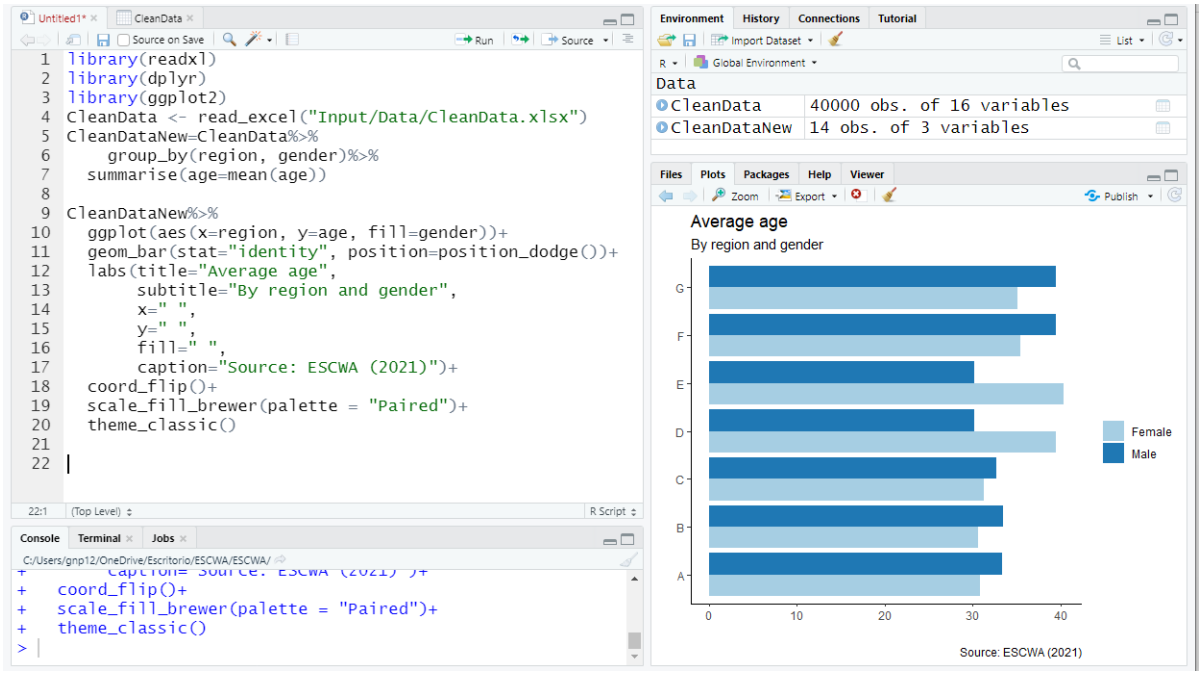
النقطة الأولى هي التسميات التي تظهر في الرسم البياني. في هذه الحالة، يوفر ggplot2 الدالة (labs) التي تستخدم للتحكم بشكل كبير في العوامل (parameters).



في الواقع، تزيد بعض القيم عن الحاجة. مثلاً، نظراً لوجود العمر في العنوان، فلا داعي لوضعه في المحور y. وبالمثل، إذا اتّضح وفقاً للسياق، أن من A إلى G هي المناطق، فلا داعي لذكرها في المحور x. ومع ذلك، يتم تقديم هذا الشكل من أجل تعلم كيفية دمج هذه القيم كجزء من دالة (labs). وقد تمت إزالتها من الرسوم البيانية التالية بحيث يصبح التصميم أكثر أناقة.

4. الموضع وخطوط الشبكة والألوان

في حين أن هذه العناصر تعتمد على الشخص الذي يقدم التقرير، يُستحسن تغيير التنسيق وقلب المحاور إذا كان ذلك يساعد على تصور التسميات بشكل أفضل، وإزالة خطوط الشبكة الداخلية، واستخدام الألوان الرصينة التي توفر تدرجاً لطيفاً ينسحب على طول التقرير.



في هذه الحالة، ولتحقيق هذه النتيجة، يُضاف ثلاثة أوامر إلى سطر الأوامر، هي:

(أ) (Coord_flip): وهو مسؤول عن تغيير المحور؛

(ب) (Scale_fill_brewer): وهو مسؤول عن تغيير لوحة الألوان. في هذه الحالة، لوح الألوان هو "paired" ولكنه ليس الوحيد المتوفر. للقراء الراغبين بالحصول على المزيد من المعلومات، يرجى زيارة الموقع https://www.r-graph-gallery.com/38-rcolorbrewers-palettes.html للعثور على عشرات التصاميم الأخرى والاطّلاع على كيفية إنشاء اللوحات؛

(ج) (Theme_classic): وهو يزيل خطوط الشبكة ويبقي الخلفية البيضاء.

5. المحاور والمسافات

يلاحظ في التمرين السابق، وجود فجوة بين 0 والمحور مما يضر بالميزة الجمالية للرسم البياني. مع الأخذ بعين الاعتبار امكانية إضافة الأرقام العشرية (2) إلى محور y (الذي أصبح الآن المحور x بسبب قلب المحاور).

يقوم هذا المثال الجديد بتغيير عنصرين في التعليمات البرمجية. أولاً، لاحظ إضافة مكتبة library جديدة. تساعد حزمة "scales" في العمل على المحور المخطط (plot axis). ثانياً، أضِف أمر جديد إلى الرسم البياني (السطران 22 و23). للدالة الجديدة (scale_y_continuous) وسيطتان: تتعلق الأولى بالتوسيع "expand" عبر إزالة الفجوة 0 من الرسم البياني، بينما تتعلّق الثانية بالتسميات "labels" التي تحدد دقة المحور وحجمه. في هذه الحالة وكما تسجل الدقة 0.01، يحتوي المحور على 2 عشري. وكالحالة السابقة، تأتي جداول الحزمة بمجموعة واسعة من الخيارات، ويمكن للراغبين بالمزيد من المعلومات استكشاف العديد من الروابط، مثل https://scales.r-lib.org/. وفي هذا الصدد لاحظ التشابه في السطرين 20 و22. فالدالة ggplot مصممة للحفاظ على هذه الأنماط وتسهيل الأمر على المستخدمين عبر تذكيرهم بكيفية كتابة خطوط الأوامر.

6. معايير تنسيق أخرى

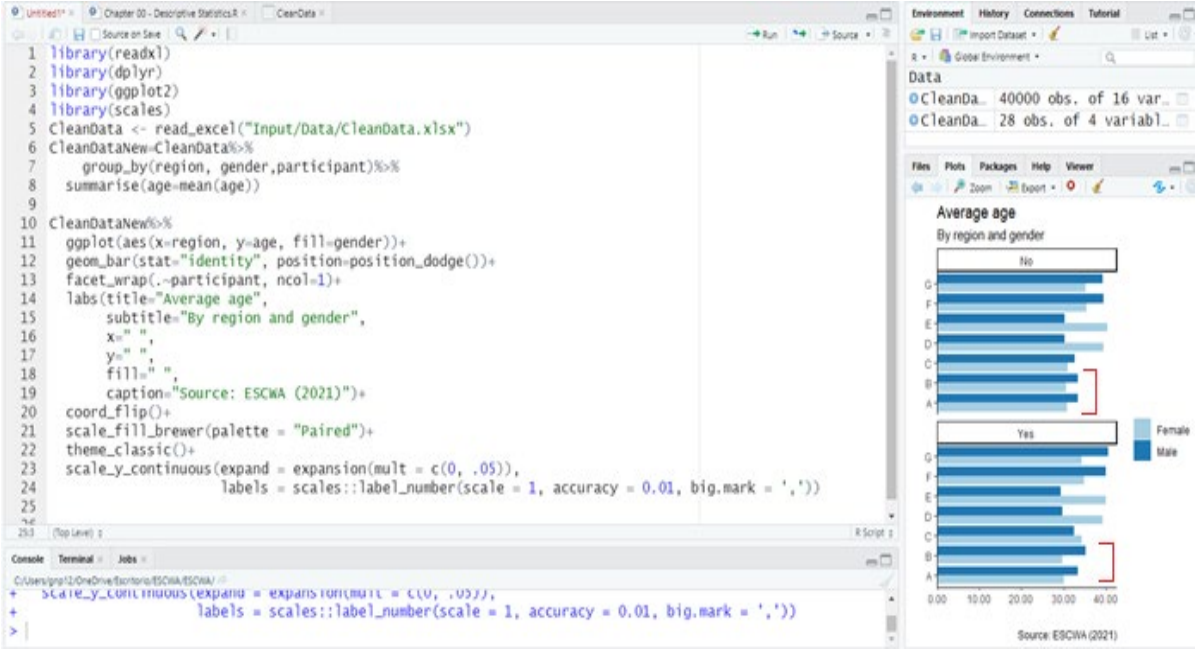
في حين سيتم إضافة العديد من معايير التنسيق الأخرى إلى رموز التنسيق التي سيتم تناولها بالتفصيل في الفصول التالية، يُستحسن ممارسة تشغيلها وإيقافها لفهم آثارها على الرسوم البيانية. ويمكن الاطلاع على بعض الإرشادات التفصيلية حول هذه المواضيع عبر المواقع التالية:

.https://ggplot2.tidyverse.org/reference/ggtheme.html

.http://www.sthda.com/english/wiki/ggplot2-themes-and-background-colors-the-3-elements

7. تجميع الرسوم البيانية

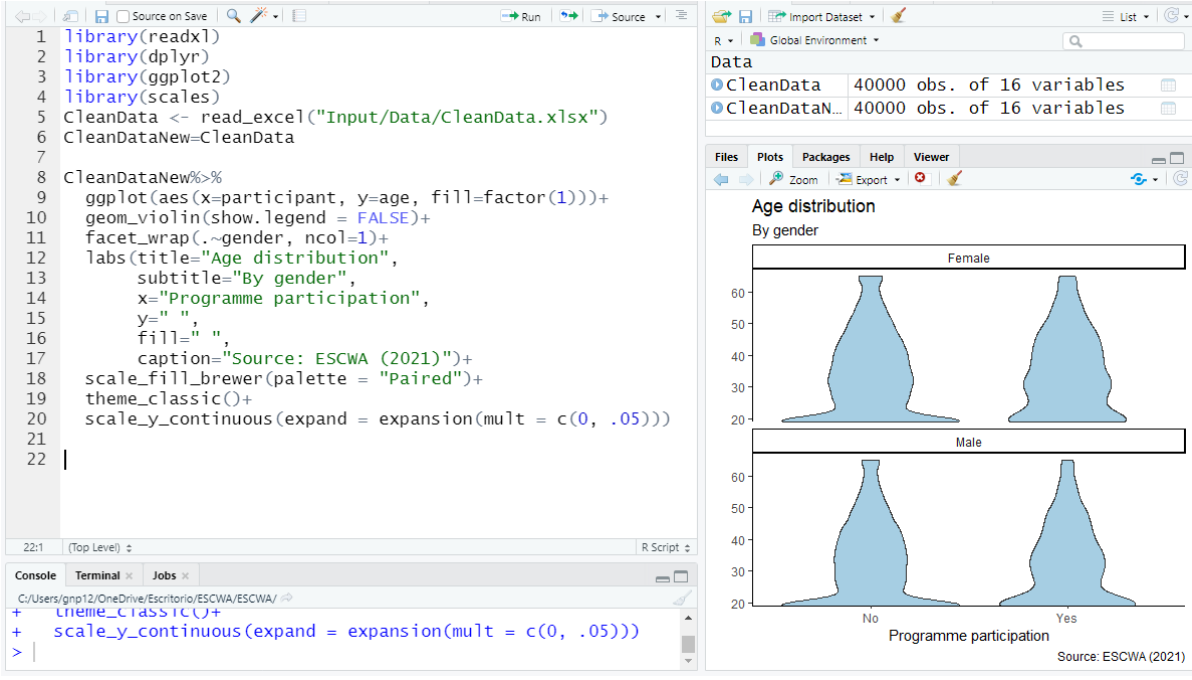
تكون أحياناً كمية المعلومات مفرطة وتظهر الحاجة إلى إنشاء رسوم بيانية فرعية للتمكن من فهم المضمون. مثلاً، افترض أن المهمة هي إنتاج رسم شريطي يعرض متوسط عمر الأفراد استناداً إلى منطقتهم وجنسهم وما إذا كانوا من المستفيدين.



في هذا الرمز الجديد، يسمح استخدام الدالة (facet_wrap) بإنشاء رسوم بيانية فرعية داخل نفس الإطار. ويتم استخدامncol لتحديد العدد المطلوب للأعمدة للرسوم البيانية الفرعية. وبالتالي، ويلاحظ أنه حتى مع هذه الأوامر القليلة، يمكن تكوين رؤية واضحة للبيانات. وفي الوقت الراهن، يمكن رؤية مناطق تتضمّن فجوات عمرية بين الذكور والإناث، ومن المثير للاهتمام، بالنسبة للفئتين A وB، هناك فجوة صغيرة بالنسبة لغير المستفيدين، ولكنها تزداد بالنسبة للمستفيدين مما يعني أن البرنامج يميل إلى اختيار الرجال الأكبر سناً من النساء في هذه المنطقة. وتسمح المهام التالية بفهم منشأ هذه الاختلافات. ومع ذلك، لا تزال بحاجة إلى معرفة المزيد من العناصر الرسومية والرمزية ذات الصلة قبل وصولك إلى تلك المرحلة.

8. أنواع الرسوم البيانية

إنّ اختيار نوع الرسم البياني هو فنّ، واعتماداً على الرسالة التي نريد توجيهها، يمكن تفضيل رسم بياني معيّن على آخر. فكّر بالباحث المهتم بتوزيع العمر مثلاً.

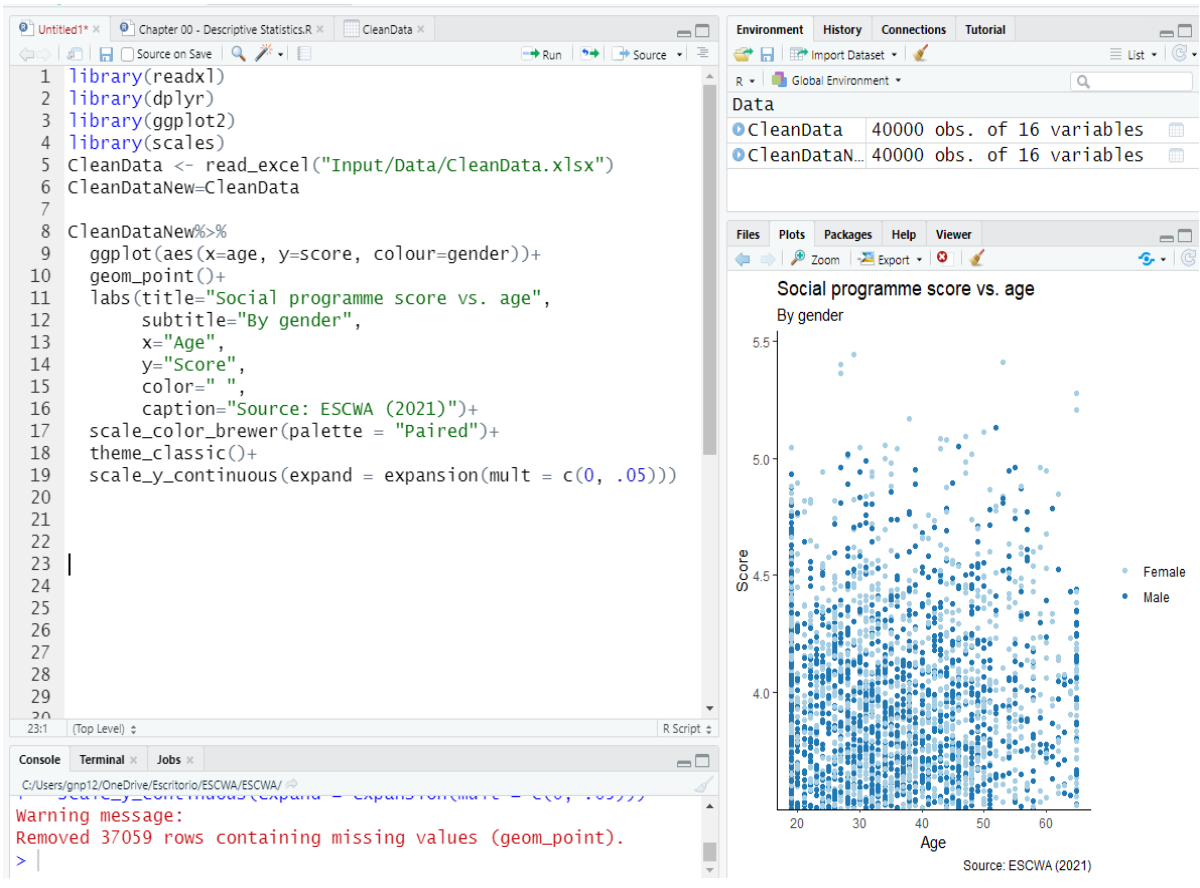


يمكن تسليط الضوء على ثلاثة عناصر هامة في إطار التعليمات البرمجية:

(أ) يتغير السطر 10 وبدلاً من تحديد تخطيط شريطي، فإنه يحدد الآن تخطيط كماني violin plot؛
(ب) على نقيض التخطيط الشريطي، يمكن تطبيق التخطيط الكماني مباشرة على مجموعة البيانات الكاملة، دون ان يتطلب أي معالجة مسبقة؛
(ج) أضيف عامل ولم يُستخدم، وأزيل بواسطة وسيلة الإيضاح. يمكن استخدام هذه الخدعة عندما لا يكون هناك تعبئة،fill للتأكد من أن الرسوم البيانية لها لوحة ألوان مصممة وليست رمادية كالرسوم البيانية الأولى في الفصل.

على خلاف الرسوم البيانية الشريطية السابقة، يساعدك التخطيط الكماني على ملاحظة أن الأفراد الأصغر سناً هم أقل تمثيلاً في المجموعة المستفيدة من غير المستفيدة. ما يؤشّر أنّ لدى الأفراد الأكبر سنا فرصاً أكبر ليصبحوا مستفيدين من البرنامج.

بمجرد وجود هذه الفرضية يمكنك إنشاء رسم بياني آخر يساعدك على دعم اكبر لهذا الادعاء. في هذه الحالة، يمكن إنشاء تخطيط مبعثر يوضح درجة المستفيدين مقابل اعمارهم.



يجب إبراز عناصر عديدة في هذا المثال:

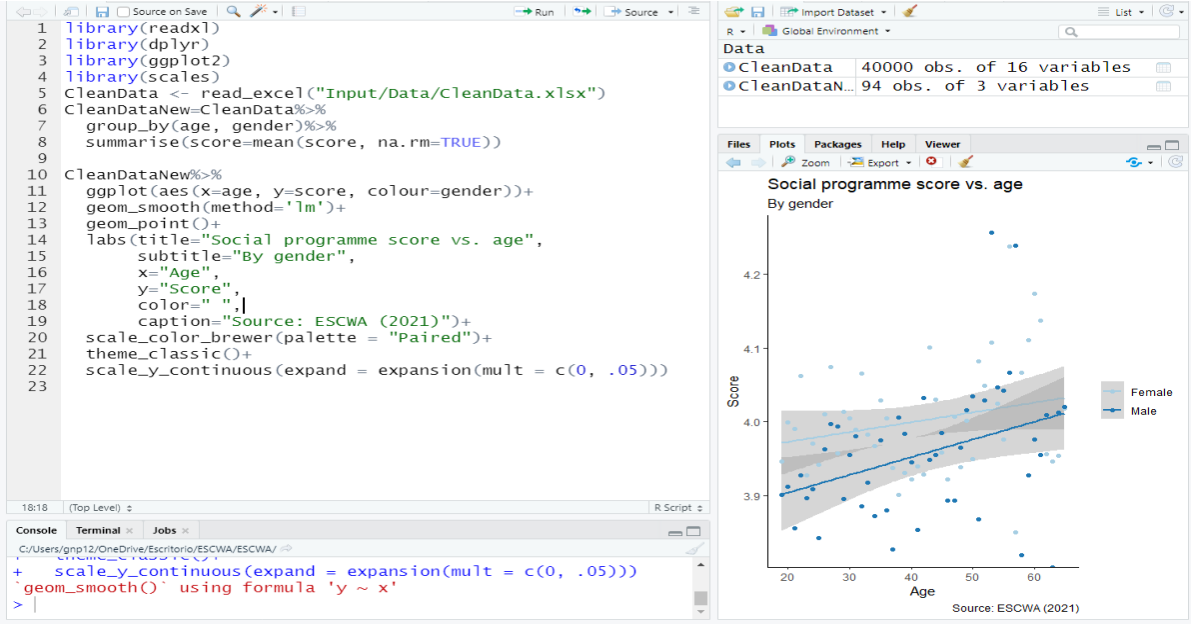
(أ) يتغير السطر 10 للحصول على التصميم الجديد للرسم البياني؛

(ب) في حين أن المتغيّر الجمالي في الحالة السابقة كان "fill"، فإن للنقاط متغير "colour" وله ثلاثة آثار، الأول في السطر 9، والثاني في السطر 15، والثالث في السطر 17؛

(ج) أحياناً، لا تكون الرسوم البيانية مفيدة، كأن يبدو الرسم البياني جميلاً، ولكن يصعب استنتاج أي شيء منه؛

(د) يوجد تحذير باللون الأحمر في وحدة التحكم. والتحذيرات ليست أخطاء، بل طرق يستخدمها البرنامج للإشارة إلى حدوث شيء غريب. في هذه الحالة، عند استخدام كل مجموعة البيانات، وعندما لا يكون للأفراد غير المستفيدين درجات، يوجد الكثير من البيانات التي لا يمكن تلوينها، ممّا لا يدل على وجود مشكلة. ومع ذلك، تقدّم التحذيرات أحياناً معلومات مفيدة.

في مثل هذه الحالة، ربما يكون من الأفضل تغيير الاستراتيجية الرسومية، وبدلاً من تخطيط مبعثر، يمكن استخدام رسم خط المسح، على الشكل المشار إليه في المثال التالي.



لا بد من إبراز عناصر عديدة في هذا المثال:

(أ) في حين أنه يبدو كرسم بياني واحد، نجد في الواقع اثنين من الرسوم البيانية المتراكبة، ونقاط وخطوط انحدار خطي، تنعكس في الخطين 11 و12؛

(ب) في هذه الحالة، تحتاج إلى إعداد البيانات قبل الرسم البياني. ومع ذلك، يمكن أن تجد "na.rm = TRUE" في الملخص. كما هو موضح في الحالة السابقة، يوجد ملاحظات غير محددة NA في متغير النقاط وإذا تم تضمينها ضمن المتوسط، فستكون النتيجة كلها NA وهذا يعتبر خطأ. لذلك، بإضافة هذا السطر الذي يشير إلى إزالة NA يفهم R أن الإجراء يتم فقط للقيم غير المرفقة بإشارة NA؛

(ج) في حين لا تزال الضوضاء موجودة في البيانات، يظهر الرسم البياني اتجاهاً واضحاً يسلط الضوء على أن الأفراد الأكبر سنا يميلون للحصول على درجات أعلى، أي أن لديهم المزيد من الاحتياجات. ومع ذلك، فإن الفرق في النتيجة بالنسبة للعمر أعلى لدى الذكور مقارنة بالإناث.

وبهذه الطريقة، سيساعد فهم المشكلة وتغيير تصميم الرسم البياني في الحصول على رؤى أفضل حول البرنامج.

خلال التدريبات السابقة، تم استعراض أنواع مختلفة من الرسوم البيانية، وسلط الضوء على الأنماط الشائعة للرموز. حالياً يمكن الوصول إلى بوابات متعددة تقدم أفكاراً مذهلة حول كيفية عرض المعلومات. يوصى للراغبين بتوسيع آفاق معلوماتهم بزيارة الموقع https://www.r-graph-gallery.com/ggplot2-package.html الذي يقدم تصاميم متعددة باستخدام أنماط الرموز نفسها. بطبيعة الحال، تقدم الصفحات الإلكترونية رموزاً مختلفة قليلاً، ولكن باستخدام التوصيات المقدمة في هذا القسم، يمكن للمتدرب التعرف بسرعة على الرموز التي تحتاج إلى تعديل والبدء بتصميم رسوم بيانية جديدة بنجاح.

الفصل الرابع. مزامنة اكسيل Excel

يعتبر (إكسيل) Excel أداة قوية. ومع ذلك، يصعب تنظيم التمارين المتعلقة بهذه الأداة. مثلاً، إذا احتجت لإنشاء مائة رسم بياني دون عنوان، وفجأةً طلب منك المشرف إضافة عنوان عام إلى جميع هذه الرسوم البيانية، ستحتاج قدراً كبيراً من الوقت (حوالي 30 ثانية لكل رسم بياني إذا كنت تعرف العناوين، أي 50 دقيقة دون توقف). وماذا لو قرر المشرف الخاص بك، بعد رؤية هذا الرسم، أنها لم تكن فكرة جيدة وأنه من الأفضل إزالة العناوين. ثم فقط عن طريق القيام بهذه المهمة، التي ليس لها قيمة مضافة، تخسر ما لا يقل عن 100 دقيقة من وقتك. يُركّز هذا الفصل والفصل التالي على تنسيق Excel (إكسل) و R للتأكد من أن هذه العملية يمكن أن تكون آلية، مماثلة لتلك السابقة، ويمكن إجراء هذه التغييرات فقط عن طريق تعديل بعض الرموز.

لبدء الفصل، لا بد من تقديم حزمة جديدة: "openxlsx" التي تجعل العمل مع (إكسيل) Excel (لصق الجداول والرسوم البيانية) أسرع.

1. الأوامر الرئيسية

على نقيض الأقسام الأخرى، يعرض القسم الحالي الرموز، وبمجرد شرح أجزائها، فإنه يقوم بعرض النتائج.

```
library(readxl)
library(dplyr)
library(ggplot2)
library(scales)
library(openxlsx)

wb = openxlsx::createWorkbook(creator = 'ESCWA')
newSheet = addWorksheet(wb, sheetName = "Age vs. Score")

CleanData <- read_excel("Input/Data/CleanData.xlsx")
CleanDataNew=CleanData%>%
  group_by(age, gender)%>%
  summarise(score=mean(score, na.rm=TRUE))

newPlot=CleanDataNew%>%
  ggplot(aes(x=age, y=score, colour=gender))+
  geom_smooth(method='lm')+
  geom_point()+
  labs(title="Social programme score vs. age",
       subtitle="By gender",
       x="Age",
       y="Score",
       color=" ",
       caption="Source: ESCWA (2021)")+
  scale_color_brewer(palette = "Paired")+
  theme_classic()+
  scale_y_continuous(expand = expansion(mult = c(0, .05)))

fileName="Output/Part1 Examples/Scatterplot age vs score.png"
cleanTable=CleanDataNew%>%spread(gender,score)
ggsave(fileName, plot = newPlot, width = 6 , height = 4 , scale = 1)
insertImage(wb, file = fileName, sheet = newSheet, startRow = 1, startCol = 1, width = 6, height = 4)
writeDataTable(wb, sheet = newSheet, x = cleanTable, startRow = 1, startCol = 10)

saveWorkbook(wb,
             file = "Output/Part1 Examples/Excel report.xlsx",
             overwrite = TRUE)
```

ويستند التمرين إلى الرسم البياني الأخير المُنجز في الفصل 3. في هذه الحالة، تتم إضافة مكتبة جديدة في السطر 5.

ثم، في السطر 7 و8 يظهر أمران commands جديدان:

(أ) OpenxlsX: تطلب دالة (createWorkbook) من R إنشاء ملف Excel الذي، سيظهر في هذه الحالة على أنه أُنشىء بواسطة "ESCWA". في هذه المرحلة لن تتمكّن من رؤية الملف، ولكن في سطر الأوامر الأخير، يأخذ R جميع التعليمات المنتجة ويُنشىء (إكسيل) ملف الـ Excel النهائي المتضمن جميع الصفات المطلوبة؛

(ب) addWorsheet: هذه الدالة من شأنها إعلام ملف إكسيل الذي بدأت بتصميمه (لهذا السبب الوسيطة الأولى الخاصة به هي ملف Excel -في هذه الحالة تظهر -wb)، والوسيطة الثانية هي اسم جدول البيانات، ولكن كن حذراً فلملف إكسيل بعض القيود على هذه الأسماء التي إن لم تتبعها، ستحصل على خطأ.

بعد هذه اللمحة، أصبح بالإمكان القيام بتمرين حول إنشاء الرسوم البيانية. ومع ذلك، يوجد تغيير دقيق في التعليمات البرمجية، ففي السطر 15 يتم تعيين كافة الرسوم البيانية كمتغير. في هذه الحالة، قد لا يتم عرضها في الواجهة، ولكن يتم تخزينها في القرص لأغراض أخرى. وإذا رغبتَ بمشاهدتها، يمكنك إنشاء سطر جديد فقط باسم المتغير.

يعرض السطر 29 متغير اسمه filename والذي يخبر R بمكان طباعة الرسم البياني ونوع التنسيق. وتتم الطباعة بتنسيق png.

يضبط السطر 30 مجموعة البيانات المستخدمة لإنتاج الرسم البياني في جدول باستخدام وظيفة spread.

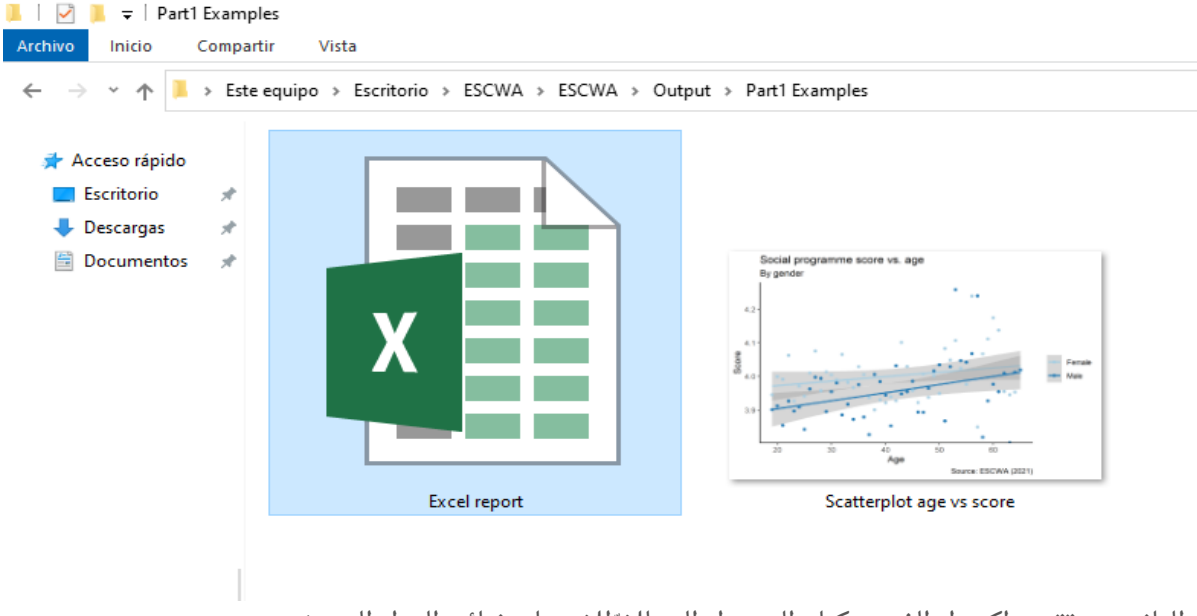
يحفظ السطر 31 الرسوم البيانية التي يتم تناولها بواسطة filename. وتقوم الدالة (ggsave) بذلك. يمكن إضافة الرسم البياني بشكل صريح في التخطيط، مع تحديد التفاصيل كالعرض والارتفاع والمقياس المرتبطة بالرسم البياني المطبوع.

يخبر السطر 32 R بإضافة هذا الرسم البياني إلى ملف اكسيل. وتتولى الدالة (insertImage) ذلك. حيث تجد فيها العديد من العناصر التي تحتاج إلى تحديد. أولاً، تحتاج إلى إدراج ملف اكسيل الذي تقوم بتصميمه، ثم إبلاغ الملف عن الصورة التي تم تعريفها بواسطة (fileName)، وعن الصفحة sheet التي تريد لصقها، ثم عليك تحديد الصفوف والأعمدة التي تريد لصقها وأبعاد الرسم البياني.

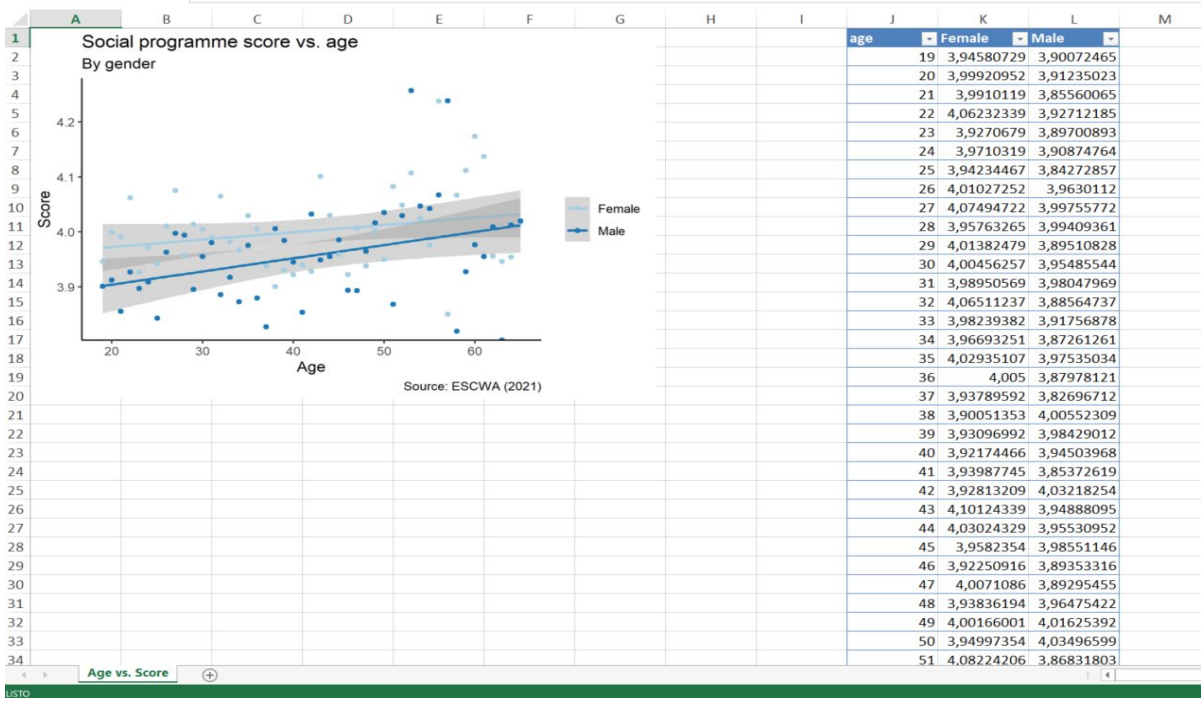
يخبر السطر 32 R بإضافة هذا الرسم البياني إلى ملف إكسيل. وتتولى الدالة (writeDataTable) ذلك. على غرار الحالة السابقة، تحتاج إلى تحديد ملف إكسيل الذي تقوم بتصميمه، ثمّ الصفحة sheet التي تريد لصقها. ويلي ذلك وضع المتغير الذي يحفظ الجدول (في السطر 30)، ثم تحديد الصفوف والأعمدة حيث تريد لصقه.

وأخيراً، تسمح السطور من 35 إلى 37 لبرنامج R، من خلال الدالة (saveWorkbook)، بإنشاء ملف اكسيل في المسار الذي قمت بتحديده. كما تسمح لك بالكتابة فوقه في حال وجود ملفات سابقة تحمل نفس الاسم.

نتيجة لهذه العملية ستجد ملفين جديدين في جهاز الحاسوب الخاص بك. الملف الأول هو الصورة التي تم إنشاؤها، وأصبحت جاهزة ليتم لصقها في أي تقرير تريده.



والثاني هو تقرير إكسيل الذي يمكنك الوصول إليه للاطّلاع على نتائج العمل المنجز.



age	Female	Male
19	3,94580729	3,90072465
20	3,99920952	3,91235023
21	3,9910119	3,85560065
22	4,06232339	3,92712185
23	3,9270679	3,89700893
24	3,9710319	3,90874764
25	3,94234467	3,84272857
26	4,01027252	3,9630112
27	4,07494722	3,99755772
28	3,95763265	3,99409361
29	4,01382479	3,89510828
30	4,00456257	3,95485544
31	3,98950569	3,98047969
32	4,06511237	3,88564737
33	3,98239382	3,91756878
34	3,96693251	3,87261261
35	4,02935107	3,97535034
36	4,005	3,87978121
37	3,93789592	3,82696712
38	3,90051353	4,00552309
39	3,93096992	3,98429012
40	3,92174466	3,94503968
41	3,93987745	3,85372619
42	3,92813209	4,03218254
43	4,10124339	3,94888095
44	4,03024329	3,95530952
45	3,9582354	3,98551146
46	3,92250916	3,89353316
47	4,0071086	3,89295455
48	3,93836194	3,96475422
49	4,00166001	4,01625392
50	3,94997354	4,03496599
51	4,08224206	3,86831803

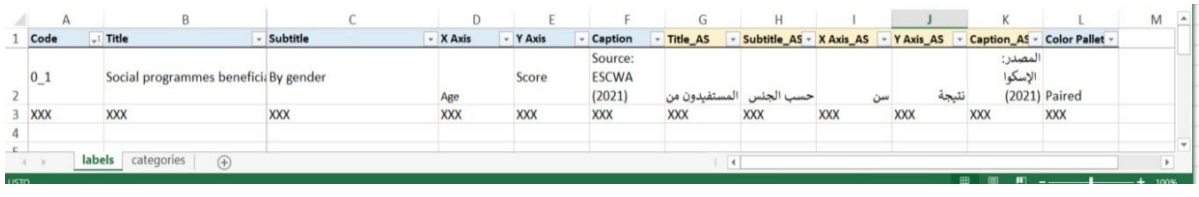
أمّا النقاط ذات الصلة التي يجب تسليط الضوء عليها، فهي:

(أ) لاحظ أن الخانات والجدول في بداية الرسم البياني تطابق تلك التي عُيّنت بواسطة التعليمات البرمجية؛
(ب) لاحظ اسم جدول البيانات الذي يطابق ذاك الذي تم تعريفه بواسطة التعليمات البرمجية؛
(ج) إذا احتاجت هذه البيانات لوضعها ضمن تقرير، يصبح من السهل إلى حد ما نسخها ولصقها من ملف اكسيل.

2. القواميس

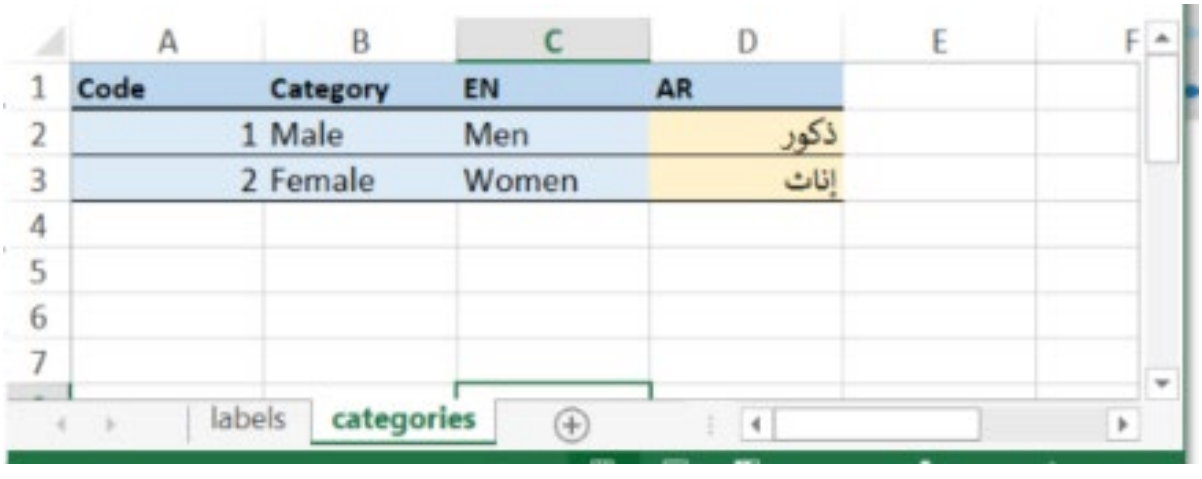
يعرض القسم السابق العناصر الرئيسية اللازمة لربط ملف اكسيل ببرنامج R. ويستخدم هذا القسم كافة المعارف التي تم التوصل إليها لإنشاء نظام تقرير وظيفي تلقائي، بالتفاعل السريع بين كل من اكسيل وR. ومع ذلك، يُطلب لهذا التمرين إنشاء ملف اكسيل جديد (يسمى قاموس)، ستجده في مجلد الإدخال، تحت اسم: "Dictionary Reduce.xlsx".

يتضمّن ملف إكسيل هذا صحيفتين مرفقتين بجدولين، يحتوي الأول على جميع عناصر الرسم البياني:



	Code	Title	Subtitle	X Axis	Y Axis	Caption	Title_AS	Subtitle_AS	X Axis_AS	Y Axis_AS	Caption_AS	Color Pallet
2	0_1	Social programmes benefici	By gender	Age	Score	Source: ESCWA (2021)	المستفيدون من	حسب الجنس	سن	نتيجة	المصدر: الإسكوا (2021)	Paired
3	XXX	XXX	XXX	XXX	XXX	XXX	XXX	XXX	XXX	XXX	XXX	XXX

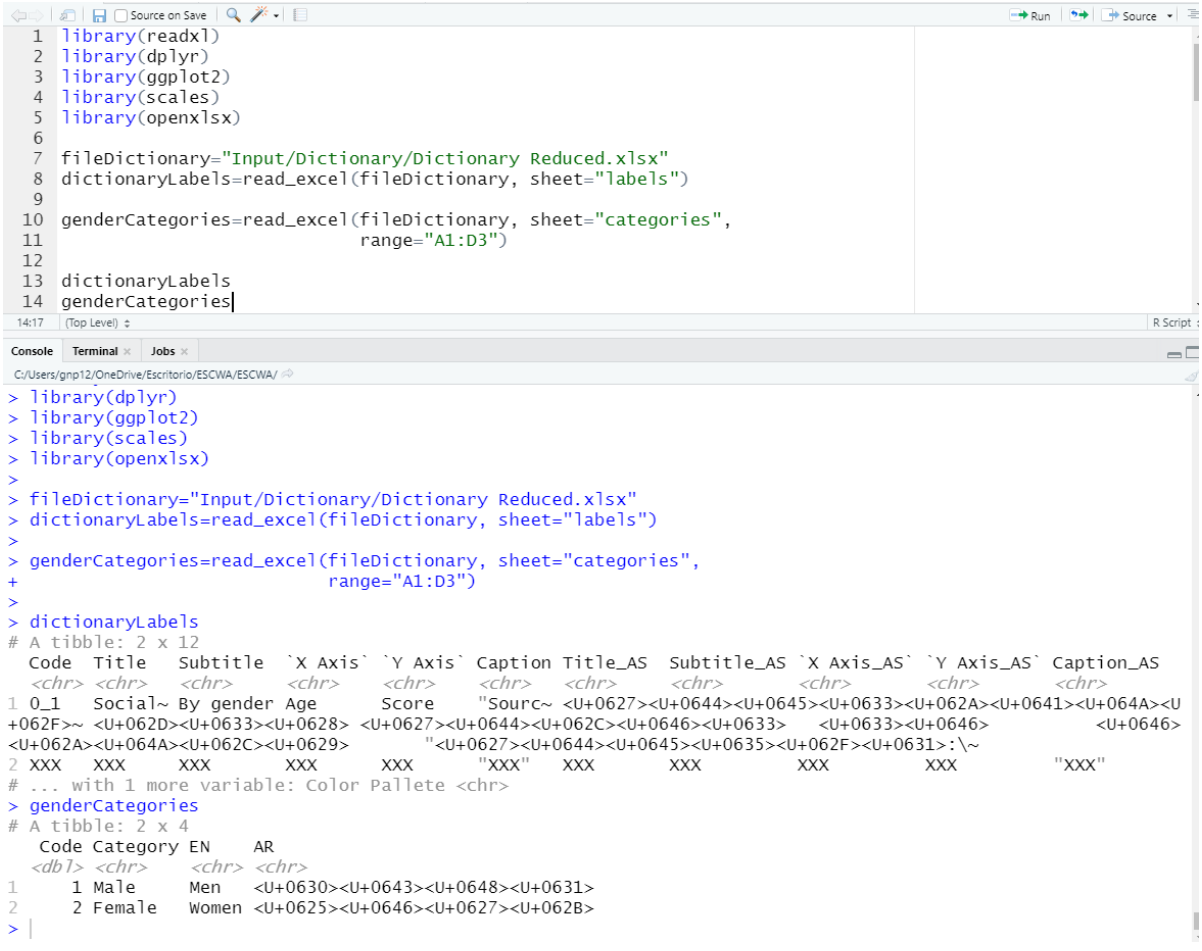
ويحتوي الثاني على جميع عناصر المتغير:



	Code	Category	EN	AR
2	1	Male	Men	ذكور
3	2	Female	Women	إناث

في هذه المرحلة، لا يمكن فهم سبب القيام بهذا العمل، ولكن عندما نتقدم في التعليمات البرمجية، تصبح فائدته واضحة.

وبالتالي، فإن المهمة الأولى هي تحميل هذه الملفات في برنامج R.



```
library(readxl)
library(dplyr)
library(ggplot2)
library(scales)
library(openxlsx)

fileDictionary="Input/Dictionary/Dictionary Reduced.xlsx"
dictionaryLabels=read_excel(fileDictionary, sheet="labels")

genderCategories=read_excel(fileDictionary, sheet="categories",
                            range="A1:D3")

dictionaryLabels
genderCategories
```

```
> library(dplyr)
> library(ggplot2)
> library(scales)
> library(openxlsx)
>
> fileDictionary="Input/Dictionary/Dictionary Reduced.xlsx"
> dictionaryLabels=read_excel(fileDictionary, sheet="labels")
>
> genderCategories=read_excel(fileDictionary, sheet="categories",
+                             range="A1:D3")
>
> dictionaryLabels
# A tibble: 2 x 12
  Code  Title   Subtitle  `X Axis` `Y Axis` Caption Title_AS  Subtitle_AS `X Axis_AS` `Y Axis_AS` Caption_AS
  <chr> <chr>   <chr>     <chr>    <chr>    <chr>   <chr>     <chr>       <chr>       <chr>       <chr>
1 0_1   Social~ By gender Age      Score    "Sourc~ <U+0627><U+0644><U+0645><U+0633><U+062A><U+0641><U+064A><U
+062F>~ <U+062D><U+0633><U+0628> <U+0627><U+0644><U+062C><U+0646><U+0633>   <U+0633><U+0646>          <U+0646>
<U+062A><U+064A><U+062C><U+0629>       "<U+0627><U+0644><U+0645><U+0635><U+062F><U+0631>:\~
2 XXX   XXX     XXX       XXX      XXX      "XXX"   XXX       XXX         XXX         XXX         "XXX"
# ... with 1 more variable: Color Pallete <chr>
> genderCategories
# A tibble: 2 x 4
   Code Category EN    AR
  <dbl> <chr>    <chr> <chr>
1     1 Male     Men   <U+0630><U+0643><U+0648><U+0631>
2     2 Female   Women <U+0625><U+0646><U+0627><U+062B>
>
```

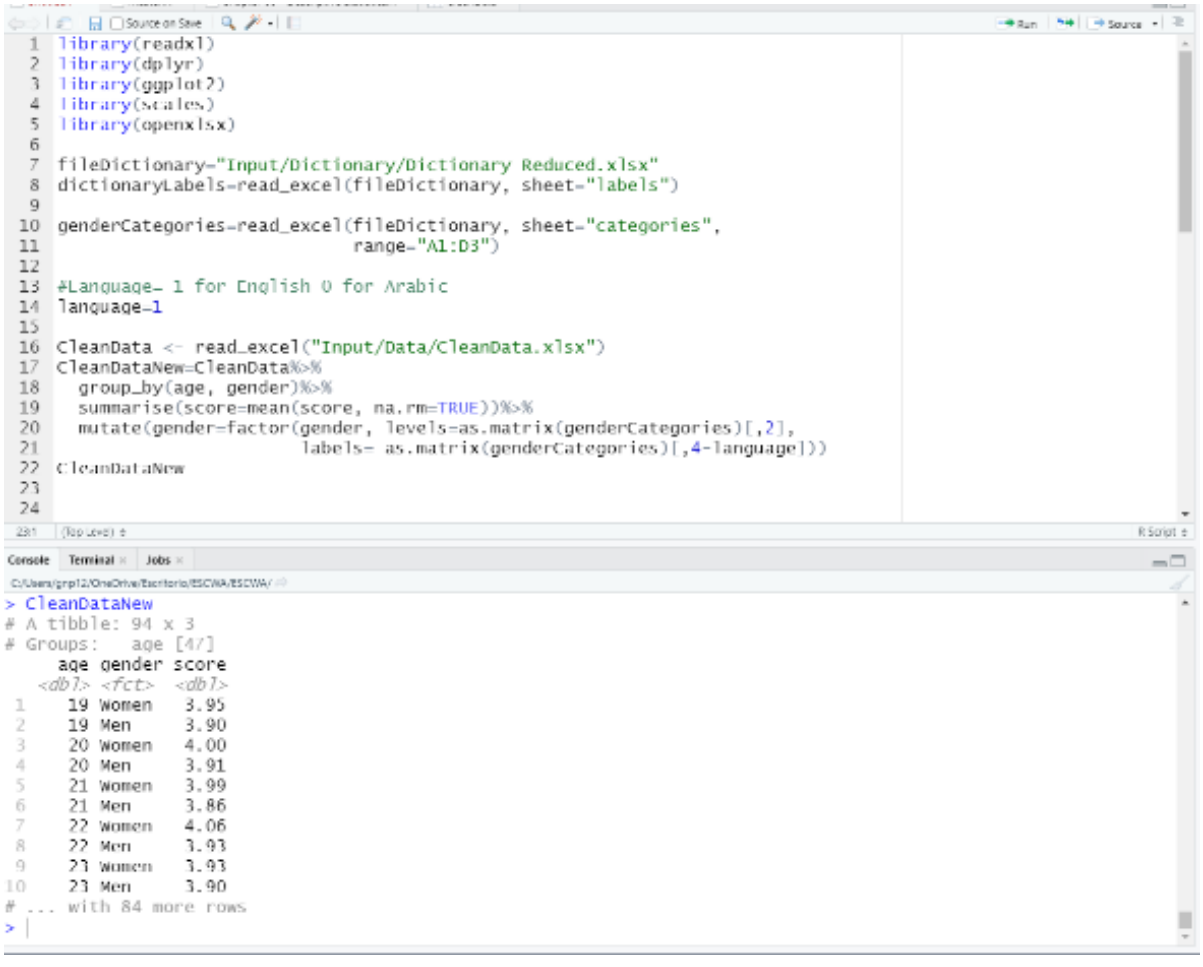
لاحظ أنه في هذا الرمز، تم توسيع الدالة التي استُخدمت سابقاً (read_excel) كما في إحدى الحالات حيث قمنا بتحديد الصفيحة، وفي الحالة الأخرى قمنا بتحديد الصفيحة والنطاق. أحيانا، يكون R "ذكياً" ويمكنه تحديد الجدول الذي تريده، ولكن بوجود العديد من الجداول، يحصل اللغط وبالتالي، فمن الأفضل تحديده.

الرجاء ملاحظة عنصرين آخرين:

(أ) يحتوي الجدول الأول في سطره الأخيرعلى صف من XXXXXXXXXXX. يساعد هذا الاقتراح العملي على تجنب الأخطاء التي قد يرتكبها البرنامج عندما تحتوي الجداول على أعمدة تتضمّن العديد من القيم المفقودة؛

(ب) لا يظهر النص العربي، بل رموزٌ غريبة تمثّل الطرق التي يعرض فيها R الأحرف، وبمجرد الانتهاء من الرموز، ستتضمن المخرجات المحتوى الصحيح.

تتمثّل الخطوة الثانية بإنشاء متغير يساعد في تحديد لغة الرسوم البيانية والجداول. بمجرد الانتهاء منه، يجب تعديل جميع المتغيرات ذات الفئات عبر عوامل، على أن تكون التسميات باللغة الصحيحة.



```
library(readxl)
library(dplyr)
library(ggplot2)
library(scales)
library(openxlsx)

fileDictionary="Input/Dictionary/Dictionary Reduced.xlsx"
dictionaryLabels=read_excel(fileDictionary, sheet="labels")

genderCategories=read_excel(fileDictionary, sheet="categories",
                            range="A1:D3")

#Language= 1 for English 0 for Arabic
language=1

CleanData <- read_excel("Input/Data/CleanData.xlsx")
CleanDataNew=CleanData%>%
  group_by(age, gender)%>%
  summarise(score=mean(score, na.rm=TRUE))%>%
  mutate(gender=factor(gender, levels=as.matrix(genderCategories)[,2],
                       labels= as.matrix(genderCategories)[,4-language]))
CleanDataNew
```

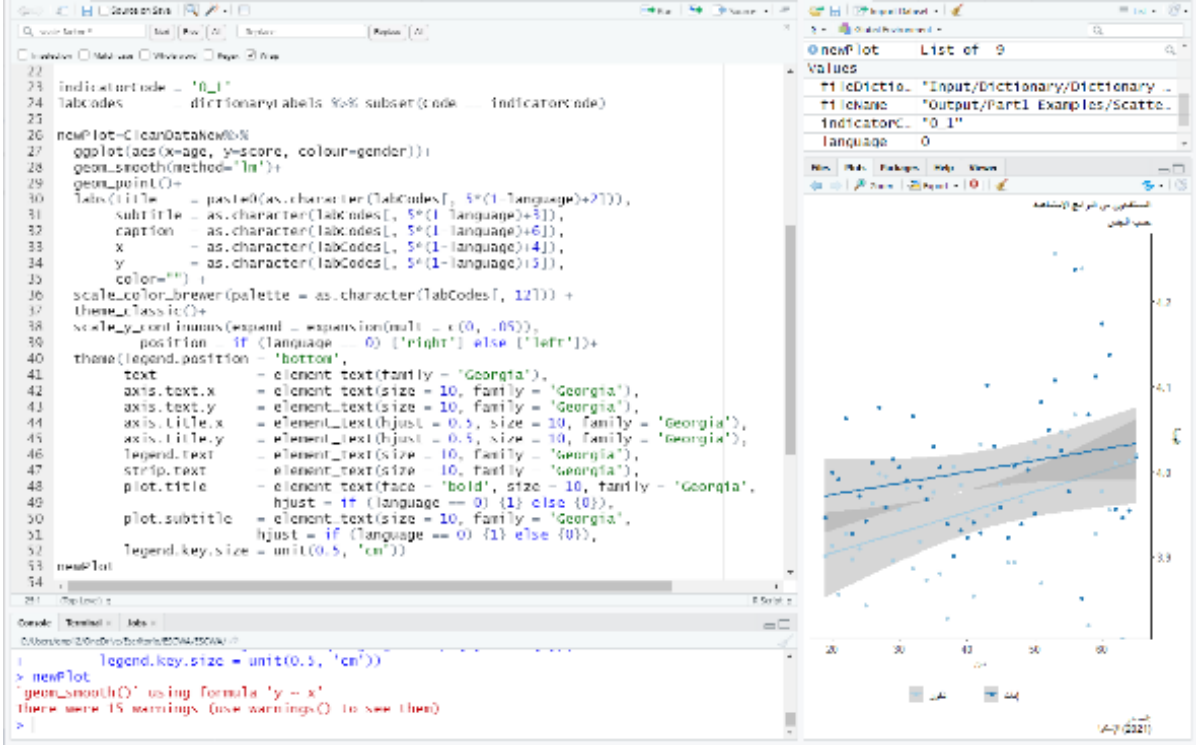
```
> CleanDataNew
# A tibble: 94 x 3
# Groups:   age [47]
     age gender score
   <dbl> <fct>  <dbl>
 1    19 Women   3.95
 2    19 Men     3.90
 3    20 Women   4.00
 4    20 Men     3.91
 5    21 Women   3.99
 6    21 Men     3.86
 7    22 Women   4.06
 8    22 Men     3.93
 9    23 Women   3.93
10    23 Men     3.90
# ... with 84 more rows
>
```

في هذا المثال، يمكن ملاحظة أن البيانات التي يتم تحميلها ومعالجتها، وفي المرحلة الأخيرة، يتم تحويل متغير الجنس فيها إلى عامل. في هذه الحالة، تستخدم دالة as.matrix(genderCategories)[,2] لإبلاغ R أن المستويات هي تلك المكتوبة في العامود الثاني من جدول اكسيل المقترن بالفئات. ثم، تستخدم دالة as.matrix(genderCategories)[,4-language] لتبلغ البرنامج حول: If language = 1 then 4-language = 3.

اذهب إلى العامود الثالث، الذي يحتوي على تسميات اللغة الإنكليزية الجديدة. بذلك، أصبحت البيانات تشير إلى الرجال والنساء بدلاً من الذكور والإناث. if then اذهب إلى العامود الرابع، الذي يحتوي على تسميات اللغة العربية الجديدة. وبالرغم من سهولة هذه الحيلة، إلا أنها فعّالة كونها تسمح بتغيير الفئات بسرعة إلى لغات مختلفة. وفي حال كانت الترجمة خاطئة، أو برزت حاجة لتغيير مرادف، يمكن تغيير التسمية بسرعة في قاموس اكسيل، وتشغيل التعليمات البرمجية في R ليتم تعديل جميع النتائج المتعلقة بهذا المتغير تلقائياً.

وتجدر الإشارة إلى أنه لا ينبغي أبدا لمس العامود 2 لأنه يحتوي على القيم ذات الإملاء الدقيق الموجود في مجموعة البيانات الأصلية.

باستخدام الفكرة نفسها مع العامل، يتم ربط برنامج البيانات الخاص بمايكروسوفت (excel) مع رسومه البيانية بعوامل (parameters) الرسم البياني.



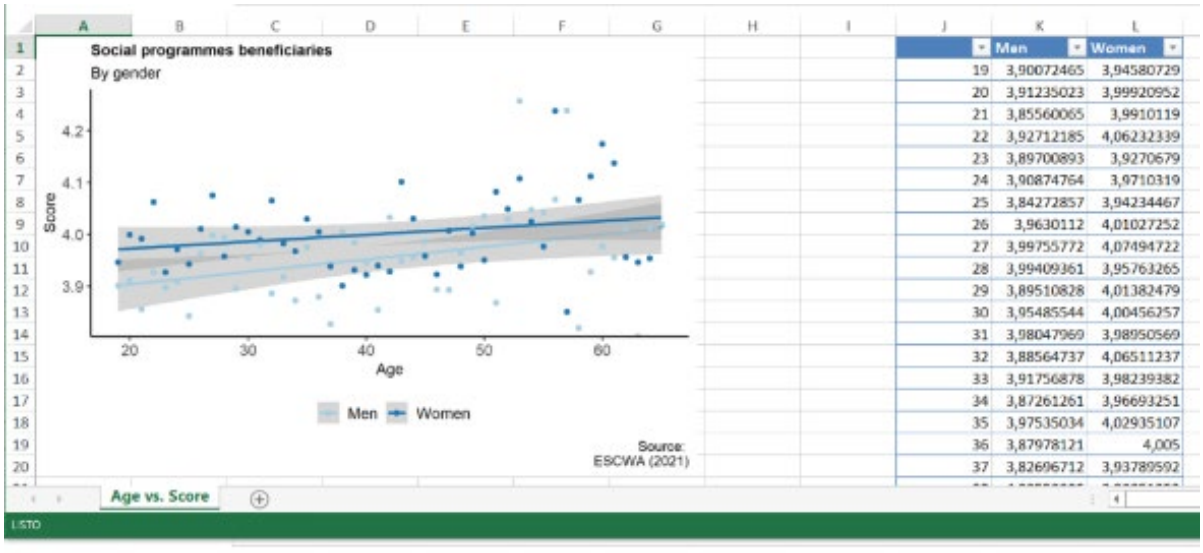
وبهدف زيادة الوضوح، يبدأ هذا الرمز مباشرة بعد انتهاء التعليمات البرمجية للمثال السابق. هناك بعض الاختلافات الصغيرة بين هذه التعليمات البرمجية أو الرموز وتلك المستخدمة في المقطع السابق. أولاً، يظهر الخطّان 23 و24 ودورهما هو تعيين رمز للرسم البياني (في هذه الحالة "0_1")، ومن ثم العثور على رمز لغة R في القاموس لتحديد الرسم البياني الذي ينبغي استخدامه، ويتم ذلك في إطار الدالة (subset).

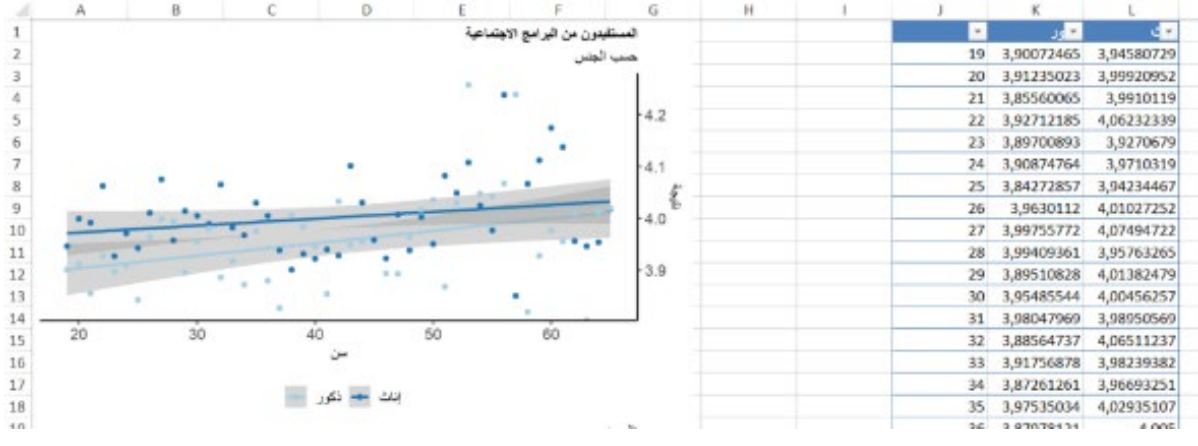
وفيما يتعلق بالرسم البياني، تجد في سطور الألوان والتسميات (31-35) خدعة كتلك المستخدمة للعامل، تستخدم الإصدار الصحيح من جدول إكسيل اعتماداً على متغير اللغة (لهذا التمرين حدد المتغير بـ 0 للإصدار العربي).

هناك أيضا إضافة إلى الدالة (scale_y_continuous). والغرض منها هو الاعتراف بتغيّر موقع المحور y في اللغة العربية. إذ يحتوي على if else كما في المثال الذي ورد في الفصل الأول، حيث يتغير المحور وفقا للّغة التي يستخدمها الرسم البياني.

وأخيراً، تضمن مجموعة كبيرة من التعليمات المتعلقة بـ (theme) تنسيقا أفضل للرسم البياني. كجزء من عملية التعلم، يوصى بأن يتعامل القارئ مع هذه السطور لفهم أفضل لكيفية تأثيرها على الرسم البياني.

يتوقّع حرص القراء على إضافة التعليمات البرمجية اللازمة لإنتاج ملفات اكسيل، كما ذكر بالتفصيل في الجزء والقسم السابقين، حيث تُعرض النتيجتان ويكون للمتدرب فكرة واضحة عن إمكانيات هذه التقنية.





في ختام الفصل، لاحظ أنه من خلال إنتاج قاموس منظم، يمكن لـ R أتمتة توليد التقارير في لغات متعددة. وعلاوة على ذلك، وكما وعدنا القارئ، إذا احتجت للقيام بتغييرات في عدد كبير من الرسوم البيانية، عليك فقط تغيير القواميس، وبعد تشغيل الرموز ستدمج كل التغييرات تلقائياً في التقرير. ممّا يوفر الوقت ويساعد على تجنب الأخطاء، لأنه بمجرد تصحيح القاموس ستنتفي الحاجة للتحقق من الرسوم البيانية والجداول لتحديد مكان التغييرات.

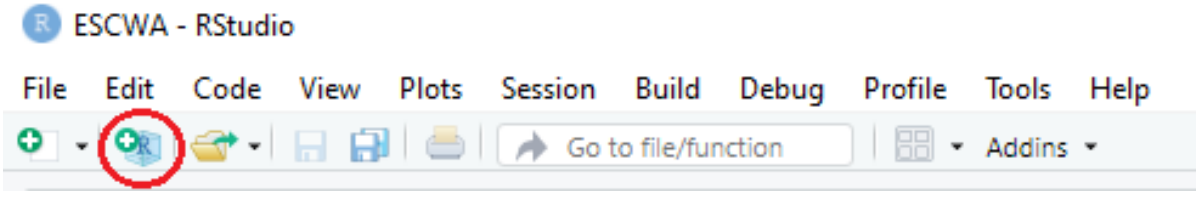
الفصل الخامس. أتمتة العمليات الكبيرة

تناول الفصل السابق تعليمات حول كيفية ربط اكسيل و R لإنتاج تقارير سريعة. يتولى هذا الفصل تنفيذ تلك التقنيات على نطاق واسع في إطار التقييم السريع لبرامج الحماية الاجتماعية (SPP-RAF). قد يتساءل القراء عن سبب غياب هذا الفصل عن الجزء الثاني. والإجابة هي أن الجزء الثاني متروك للتقنيات الإحصائية، ويعتبر هذا الفصل بمثابة مقدمة لهذه التقنيات، بما في ذلك إنتاج الإحصاءات الوصفية ذات الصلة. وبناءً عليه، يمكن أن يسمى هذا الفصل، عند وضعه في تقرير ما، الفصل 0 أو الفصل التمهيدي، حيث يتم وصف مجموعة البيانات عبر الرسوم البيانية والجداول، والسماح بتأمين البيئة المؤاتية للتحليل. لهذا الفصل وللفصول القادمة، فإنّ ملفات اكسيل التي سيتم استخدامها هي: RawData.xlsx (موجودة في Inputs/Data) وDictionary.xlsx (موجود في Inputs/Dictionary).

1. المشروع والملف الرئيسي

قبل بدء أي مشروع يوصى بتنظيم كافة الملفات وفق ترتيب محدد.

والترتيب المقترح هو ذاك الذي تجده في مواد المساعدة التعليمية، حيث المجلدات منفصلة عن المدخلات والرموز والمخرجات. بمجرد تنظيم كل شيء، يمكنك إنشاء مشروع R وحفظه في المجلد الأساسي (يمكن العثور على هذا المشروع تحت اسم الإسكوا (ESCWA)). وبالنسبة للحالة العامة، يمكنك حفظ مشروع باستخدام الرمز المميز في الرسم التوضيحي.



عملياً، يعتبر المشروع كمجلد، حيث يمكن وضع العديد من الرموز معاً. والتي تعرف بتفاعلها ضمن ملفات الحاسوب نفسها. ويمكنك العمل مع جميع البرامج النصية، كما فعلنا حتى الآن. ولكن عندما تكون أمام مشاريع كبيرة، فإن الممارسة الجيدة تفترض تقسيم التعليمات البرمجية إلى أجزاء يسهل مراجعتها بشكل منفصل (كفصول). ممّا يؤدّي إلى تسهيل عمليات التصحيح ووضوح الرموز. في هذا السياق، تسمح البنية القوية للمشروع بتكامل هذه الملفات المنفصلة عبر ملف رئيسي متزامن، عن طريق استدعاء الملف الرئيسي من خلال المشروع الذي يجمع المعلومات حول بنية المجلد ويحسن تنظيم النتائج. ترد التعليمات البرمجية الدقيقة للملف الرئيسي في رمز المجلد الخاص بالمساعدات التعليمية، وفي هذا القسم، ستطلع على مكوناته. على نقيض الأقسام الأخرى، لم تتم تغطية العديد من الوظائف هنا إلا سطحياً لأنها لا تضيف قيمة كبيرة للفهم العام القارئ ويمكن العثور على وظائفها بشكل دقيق على شبكة الانترنت.

```
1  rm(list = ls())
2  graphics.off()
3  gc()
4  startTime  = Sys.time()
```

تضمن السطور الأربعة الأولى من الملف الرئيسي، نظافة البرنامج وخلوّه من رسوم بيانية أو متغيرات محفوظة من التدريبات السابقة. ممّا يساعد الحسابات على الانتقال بشكل أسرع وتجنب الأخطاء التي قد تأتي من البيانات المخزنة مسبقاً. السطر الأخير هو للتحكم بالوقت حيث يخبر الحاسوب عند بدء تشغيل التعليمات البرمجية. لاحقاً، في نهاية الملف الرئيسي، يُبلّغ الخط التكميلي الحاسوب عند توقف تشغيل التعليمات البرمجية كما يُبلِّغ المستخدم حول مدة العملية. ممّا يجعله مفيداً للغاية حيث يمكن للمستخدم تقدير أوقات التشغيل وتخطيط عمله وفقاً لذلك.

```
6  # 1| Preparation ----------------------------------------------------------
7  # 1.1| Libraries ----------------------------------------------------------
8  myPackages   = c('readxl','gridExtra','rpart','partykit','ggdendro','ggparty',
9                   'factoextra','ppcor','glmnet','caret','gbm','Metrics',
10                  'cobalt','gtools','fastDummies','openxlsx','extrafont',
11                  'here','tidyverse','reshape2', 'StatMatch')
12 notInstalled = myPackages[!(myPackages  %in%  rownames(installed.packages()))]
13 if(length(notInstalled)) {
14   install.packages(notInstalled)
15 }
16
17 invisible(sapply(myPackages, library, character.only = TRUE, quietly = TRUE))
18 loadfonts(device = 'win', quiet = T)
19 options(scipen = 999) # Disable scientific notation.
```

إنها ممارسة جيدة حيث يتم كتابة المدونات والتعليق عليها بدقة، وهذا ما يساعد المستخدمين الآخرين على فهم ما تمت كتابته في البرنامج النصي. في هذا الجزء، تسأل التعليمات البرمجية الحاسوب عن تثبيت الحزم المذكورة في السطور 8 و11، وإذا لم تكن مثبتة من قبل تقوم التعليمات البرمجية بتثبيتها. كما تسمح بالتأكد من تحميل كل المكتبات libraries بحيث يمكن تشغيل جميع الرموز الأخرى بسلاسة، دون الحاجة لتحقّق المستخدم مما إذا كان قد تم تحميل المكتبة أم لا.

```r
# 1.2| Initial values ----------------------------------------------------
# 1: English.
# 0: Arabic
language = 0

userLocation   = enc2native(here()) # Replace by your own path.
if(!file.exists("Output")){
  dir.create("Output")
}
if(!file.exists("Output/Chapter 00")){
  dir.create("Output/Chapter 00")
}
if(!file.exists("Output/Chapter 01")){
  dir.create("Output/Chapter 01")
}
if(!file.exists("Output/Chapter 02")){
  dir.create("Output/Chapter 02")
}
if(!file.exists("Output/Chapter 03")){
  dir.create("Output/Chapter 03")
}

scriptLocation = paste0(userLocation, '/Code/')
inputLocation  = paste0(userLocation, '/Input/')

```

يقوم المقطع التالي بتهيئة مجلدات الإخراج وتعريف لغة التقرير الكامل، لاحظ الدالة في السطر 26. فكما ذُكر سابقاً، عندما تريد استدعاء ملف إكسيل أو ملف آخر تحتاج إلى استخدام مسارات كبيرة جداً. ولكن مع الدالة here، يبحث R عن موقع المشروع ويحدد كافة المجلدات المتعلقة به. مثلاً، إذا كنت تريد التحقق من وجود ملف يسمى output، يمكنك إعداده (باستخدام السطور من 27 إلى 29)، ولن تحتاج إلى التحديد الكامل للملف، عليك فقط ذكر اسم المجلد في المشروع. وفي السطرين 43 و44، بما أنّ المطلوب هو المسار الكامل، تم استخراجه بالفعل في السطر 26، فلا حاجة لكتابته بشكل صريح، إذ يقوم الحاسوب بهذه المهمة تلقائيا.

```r
# 2| Chapters ----------------------------------------------------------
source(paste0(scriptLocation, 'Chapter 00 - Descriptive Statistics.R'), encoding = 'UTF-8')
gc()
source(paste0(scriptLocation, 'Chapter 01 - Profiling of beneficiaries.R'), encoding = 'UTF-8')
gc()
source(paste0(scriptLocation, 'Chapter 02 - Targeting characteristics.R'), encoding = 'UTF-8')
gc()
source(paste0(scriptLocation, 'Chapter 03 - Coverage evaluation.R'), encoding = 'UTF-8')

endTime = Sys.time()
endTime - startTime
```

يقوم الجزء الأخير بتشغيل كل البرامج النصية ذات الصلة المقترنة بفصول التقرير، وبتوقيف عداد الوقت. وبهذه الطريقة، يمكن احتساب الوقت الذي استغرقه الحاسوب لمعالجة كل التقرير.

لا يغطي هذا الفصل من الدليل سوى الفصل 00 من التقرير، على أن يُغطّي الجزء الثاني كل فصل من الفصول الأخرى.

2. إنتاج إحصاءات وصفية

وبطريقة مماثلة لعرض الملف الرئيسي، يُسلّط الفصل 00 الضوء على العناصر الرئيسية للمدونة فقط، ويترك الباقي لتمارين المتدربين.

```
outputLocation = paste0(userLocation, '/Output/Chapter 00/')

# 1.1| Figure labels and frame-------------------------------------------------------------
dataPlots <- read_excel(paste0(inputLocation, 'Dictionary/Dictionary.xlsx'),
                        sheet = 'Labels')
dataPlots[is.na(dataPlots)] <- ''

dataFrameFinal <- read_excel(paste0(inputLocation, 'Data/RawData.xlsx'))

# 1.2| Dictionaries ----------------------------------------------------
Region         <- as.matrix(read_excel(paste0(inputLocation, 'Dictionary/Dictionary.xlsx'), sheet = 'categories', range = 'A1:D8'))
quantiles1     <- as.matrix(read_excel(paste0(inputLocation, 'Dictionary/Dictionary.xlsx'), sheet = 'categories', range = 'F1:I11'))
response1      <- as.matrix(read_excel(paste0(inputLocation, 'Dictionary/Dictionary.xlsx'), sheet = 'categories', range = 'K1:N3'))
characteristics <- as.matrix(read_excel(paste0(inputLocation, 'Dictionary/Dictionary.xlsx'), sheet = 'categories', range = 'P1:S36'))
studies        <- as.matrix(read_excel(paste0(inputLocation, 'Dictionary/Dictionary.xlsx'), sheet = 'categories', range = 'AE1:AH7'))
labor          <- as.matrix(read_excel(paste0(inputLocation, 'Dictionary/Dictionary.xlsx'), sheet = 'categories', range = 'AJ1:AM4'))
gender         <- as.matrix(read_excel(paste0(inputLocation, 'Dictionary/Dictionary.xlsx'), sheet = 'categories', range = 'AO1:AR3'))
deciles        <- as.matrix(read_excel(paste0(inputLocation, 'Dictionary/Dictionary.xlsx'), sheet = 'categories', range = 'AT1:AW11'))
confirmation   <- as.matrix(read_excel(paste0(inputLocation, 'Dictionary/Dictionary.xlsx'), sheet = 'categories', range = 'K5:N7'))
translationVar <- as.matrix(read_excel(paste0(inputLocation, 'Dictionary/Dictionary.xlsx'), sheet = 'categories', range = 'CH1:CK10'))

```

يُنشىء الجزء الأول من التعليمات البرمجية متغير لكافة المخرجات التي تم إنشاؤها في هذا البرنامج النصي، ويقوم بتحميل كافة الجداول المقترنة بالبيانات الخام (السطر 8) مع القواميس (السطور 4-7 والخطوط 11-20). يضمن القيام بهذه العملية في البداية تنظيم التعليمات البرمجية، ورصد أي تغيير في القاموس قد يتطلب تغييراً فيها (على سبيل المثال، ظهور فئة جديدة). أخيراً، يجب التنويه إلى عدم حاجة هذه التعليمات البرمجية إلى مكتبات بما أنه سيتم تشغيلها بواسطة الملف الرئيسي. ومن الجدير بالذكر أيضاً عدم المقدرة على تشغيل هذه التعليمات البرمجية وحدها. وإذا أردت ذلك، فعليك أولاً القيام بتشغيل الرموز الأولى من الملف الرئيسي (حتى القسم 2)، ثم تشغيل هذه التعليمات البرمجية.

```r
# 1.3| Factors ------------------------------------------------------------
dataFrame = dataFrameFinal
dataFrame$region = factor(dataFrame$region,
                          levels = Region[,2],
                          labels = Region[,4-language])
dataFrame$internetAcces = factor(dataFrame$internetAcces,
                                 levels = response1[,2],
                                 labels = response1[,4-language])
dataFrame$healthAcces = factor(dataFrame$healthAcces,
                               levels = response1[,2],
                               labels = response1[,4-language])
dataFrame$gender  = factor(dataFrame$gender,
                           levels = gender[,2],
                           labels = gender[,4-language])
dataFrame$education  = factor(dataFrame$education,
                              levels = studies[,2],
                              labels = studies[,4-language])
dataFrame$employment  = factor(dataFrame$employment,
                               levels = labor[,2],
                               labels = labor[,4-language])
dataFrame$incomeQuantile  = factor(dataFrame$incomeQuantile,
                                   levels = deciles[,2],
                                   labels = deciles[,4-language])
dataFrame$participant = factor(dataFrame$participant,
                               levels = confirmation[,2],
                               labels = confirmation[,4-language])


dataParticipation = dataFrame

# 2| Analysis ------------------------------------------------------------
wb = openxlsx::createWorkbook(creator = 'ESCWA')

```

يُترجم المقطع التالي من التعليمات البرمجية كافة المتغيرات الموجودة في مجموعة البيانات. ويوصى بالقيام بذلك في مرحلة مبكرة، ويفضل أن يكون من مجموعة البيانات الأصلية بحيث تُدمج تلقائياً بأي سطر من التعليمات البرمجية، ليتم بعدها إنشاء تصميم لملف اكسيل، والبدء بإنتاج الرسوم البيانية والجداول.

```r
# 2.1| Descriptive statistics ----------------------------------------
# '0_1'
rowLine     = 1
jumpOfRows = 20
colPlots    = 1
colTables   = 12

indicatorCode <- '0 1'
part          <- ''
tabCodes      <- dataPlots %>% subset(Code == paste0(indicatorCode, part))
newSheet      <- addWorksheet(wb, sheetName = indicatorCode)

# To export:
base_exp    = 1
heightExp   = 1.146
widthExp    = 1
scale_factor = base_exp/widthExp

newPlot = dataFrame %>%
  ggplot(aes(x=gender, y=age, fill=region)) +
  geom_violin(show.legend = F) +
  facet_wrap(~region, ncol=2)+
  labs(title    = paste0(as.character(labCodes[, 5*(1-language)+2])),
       subtitle = as.character(labCodes[, 5*(1-language)+3]),
       caption  = as.character(labCodes[, 5*(1-language)+6]),
       x        = as.character(labCodes[, 5*(1-language)+4]),
       y        = as.character(labCodes[, 5*(1-language)+5])) +
  scale_fill_brewer(palette = as.character(labCodes[, 12])) +
  theme_classic() +
  scale_y_continuous(expand = expansion(mult = c(0, .05)),
                     limits = c(0, NA),
                     position = if (language == 0) 'right' else 'left') +
  theme(legend.position = 'bottom',
        text              = element_text(family = 'Georgia'),
        axis.text.x       = element_text(size = scale_factor * 10, family = 'Georgia'),
        axis.text.y       = element_text(size = scale_factor * 10, family = 'Georgia'),
        axis.title.x      = element_text(hjust = 0.5, size = scale_factor * 10, family = 'Georgia'),
        axis.title.y      = element_text(hjust = 0.5, size = scale_factor * 10, family = 'Georgia'),
        legend.text       = element_text(size = scale_factor * 10, family = 'Georgia'),
        strip.text        = element_text(size = scale_factor * 10, family = 'Georgia'),
        plot.title        = element_text(face = 'bold', size = 10, family = 'Georgia', hjust = if (language == 0) (1) else (0)),
        plot.subtitle     = element_text(size = 10, family = 'Georgia', hjust = if (language == 0) (1) else (0)),
        legend.key.size   = unit(0.5 * scale_factor, 'cm')) +
  guides(fill  = guide_legend(title = ''))

# To excel:
fileName   = paste0(fileName   paste0(outputLocation, indicatorCode, part,
                                      if(language == 1) {' (English)'} else {' (Arab)'}, '.png'))
ggsave(fileName, plot = newPlot, width = 6 * widthExp, height = 4 * heightExp * widthExp, scale = scale_factor)
insertImage(wb, file = fileName, sheet = indicatorCode, startRow = rowLine, startCol = colPlots, width = 6 * widthExp, height = 4 * heightE
rowLine = rowLine + jumpOfRows
```

يتم استخدام كل هذا النص لإنشاء متغير واحد. وقد تم تغطية معظمه في الفصل الرابع، ولكن من المهم تسليط الضوء على ثلاثة عناصر:

(أ) تستخدم الخطوط 66-71 لتعديل بعض نسب الرسم البياني بحيث تبدو أفضل عند طباعتها؛

(ب) أحياناً، قد تظهر العديد من الرسوم البيانية واحدة تحت الأخرى في نفس الورقة من ملف إكسيل (وخاصة في الحالات التي يكون فيها الموضوع مماثلاً). ويساعد السطر 105 على إخبار R بوجود رسم بياني جديد. وفي هذه الحالة، يجب أن لا يظهر في الصف الأول من ورقة اكسيل، بل سيتراجع بعض الصفوف إلى الأسفل، بحيث يظهر الرسم البياني الجديد بعد السابق مباشرة؛

(ج) لاحظ أن اسم الملف مشفر بطريقة يظهر بها ملف png مع رمز المؤشر ويشير إلى لغة الرسم البياني بحيث يمكن للمحلل معرفة الرسم البياني ولغته بسرعة بمجرد التحقق من اسم الملف.

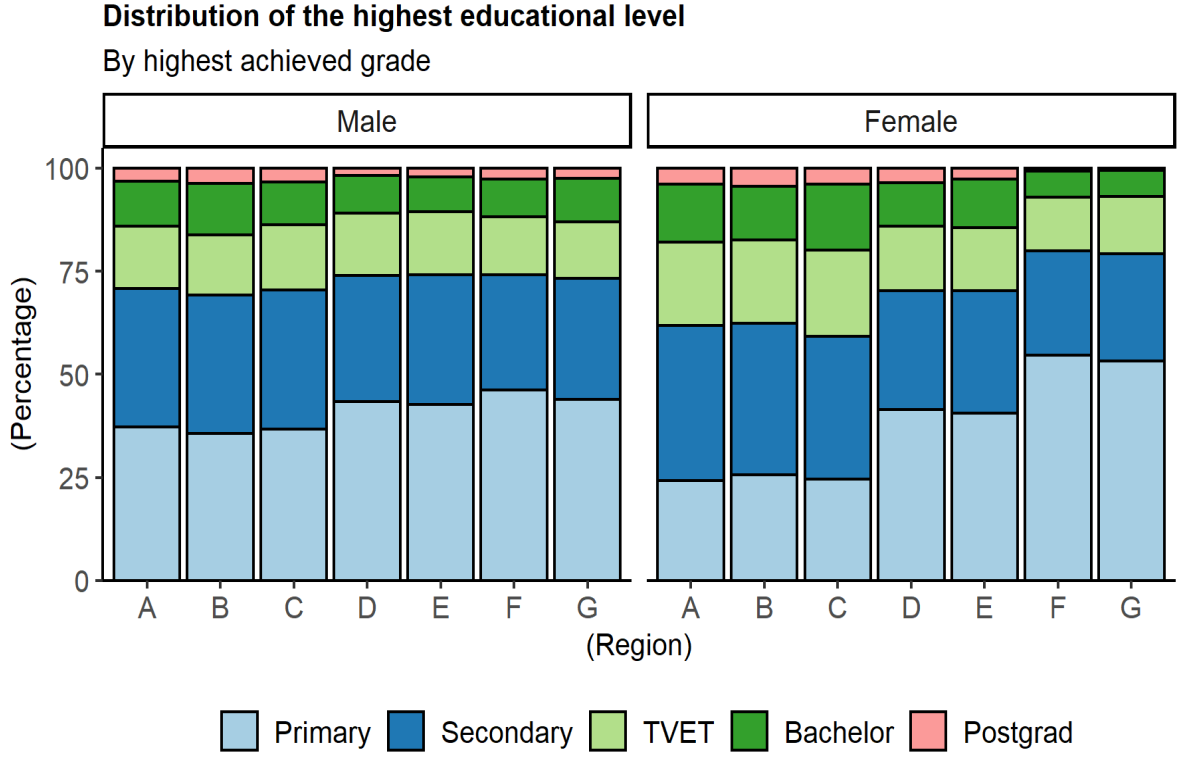
وبقية البرنامج النصي هي مجرد تكرار طويل من مؤشرات مختلفة، حتى نصل إلى السطور الأخيرة، التي تشير إلى طباعة ملف اكسيل.

```r
# final step
saveWorkbook(wb,
             file = paste0(outputLocation, if(language == 1) {'CH 0 EN.xlsx'} else {'CH 0 AR.xlsx'}),
             overwrite = TRUE)
```

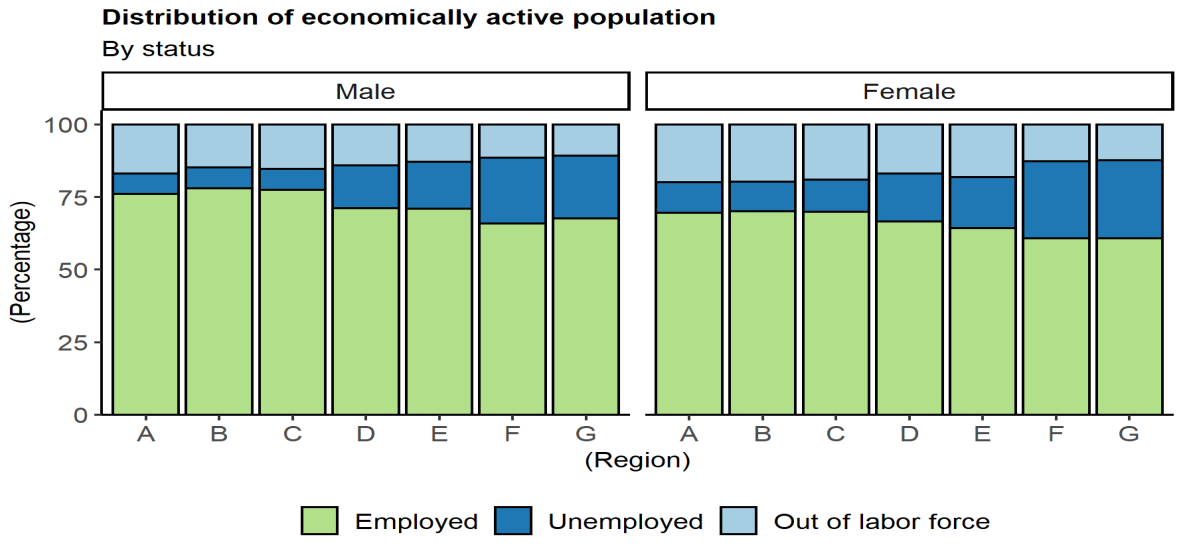
كملاحظة أخيرة، لاحظ أن البرنامج النصي مرمَز بطريقة تحول القاموس dictionary إلى إطار للبيانات. لا يعتبر القاموس مفيداً لكتابة التقرير فحسب، بل يُمثّل قائمة منظمة بجميع المؤشرات التي تم إنتاجها، ووضعها في رسم بياني، أو جدول. وهكذا، يصبح أداة موجزة مفيدة تلخص محتوى التقرير.

3. تحليل النتائج

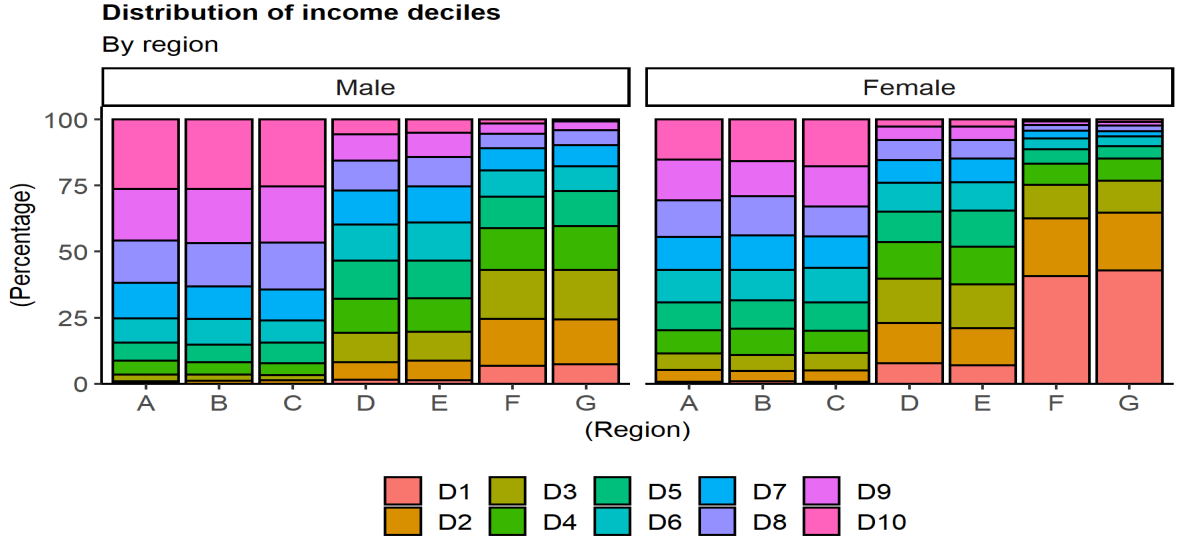
كل ما أُنجز إلى الآن يمكّن من الوصول الى نقطة تسمح بإنتاج الرسوم البيانية، وتقييم النتائج. مع كل ما قد انجزناه، دعونا نختار بعض المتغيرات التي يمكن أن تساعدك على فهم أفضل لواقع مملكة X. لاحظ أن بعض الرسوم البيانية غير موضحة في الفصل 3، ولكن من خلال التحقق من التعليمات البرمجية، من المتوقع أن يعرف المتدرب أين يضيف خطوط الأوامر للحصول على النتيجة المحددة.



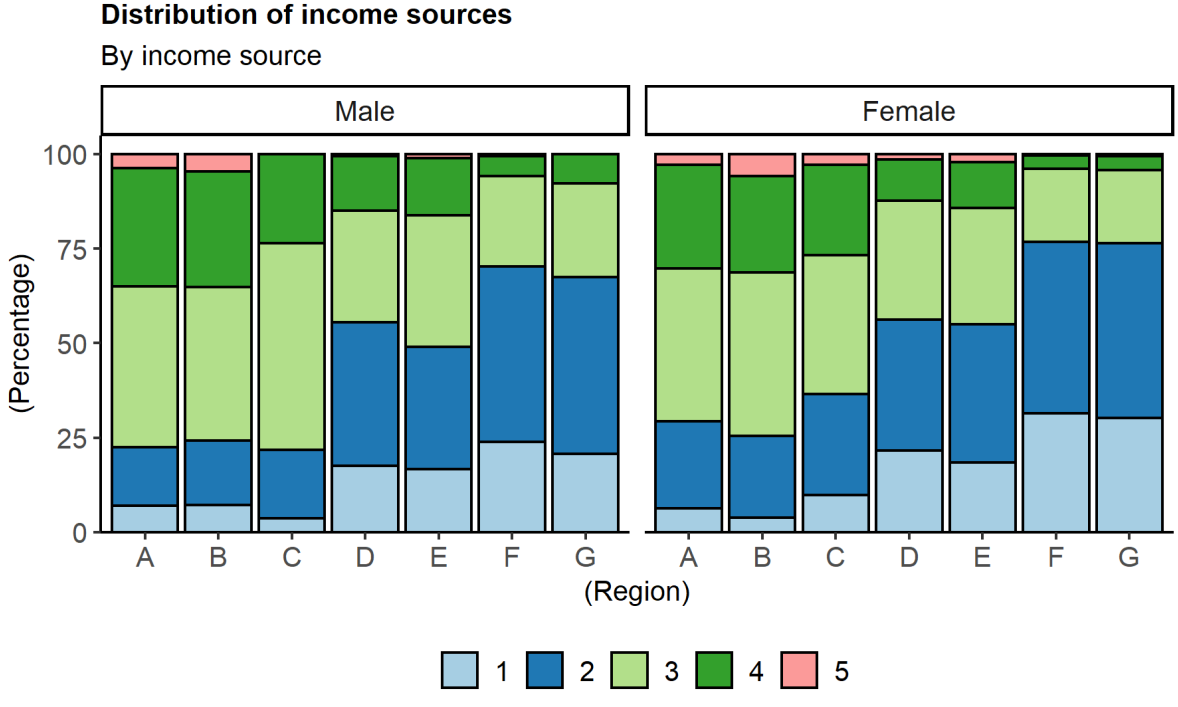
دعونا نبدأ في مراجعة مستويات التعليم للسكان. وفي حين أن هذه المستويات متشابهة نسبياً بين المناطق بالنسبة للرجال، تبرز اختلافات واضحة بالنسبة للنساء، حيث يوجد في المنطقتين A-C أكثر النساء تعليماً، وفي F-G أقلهنّ تعليماً.



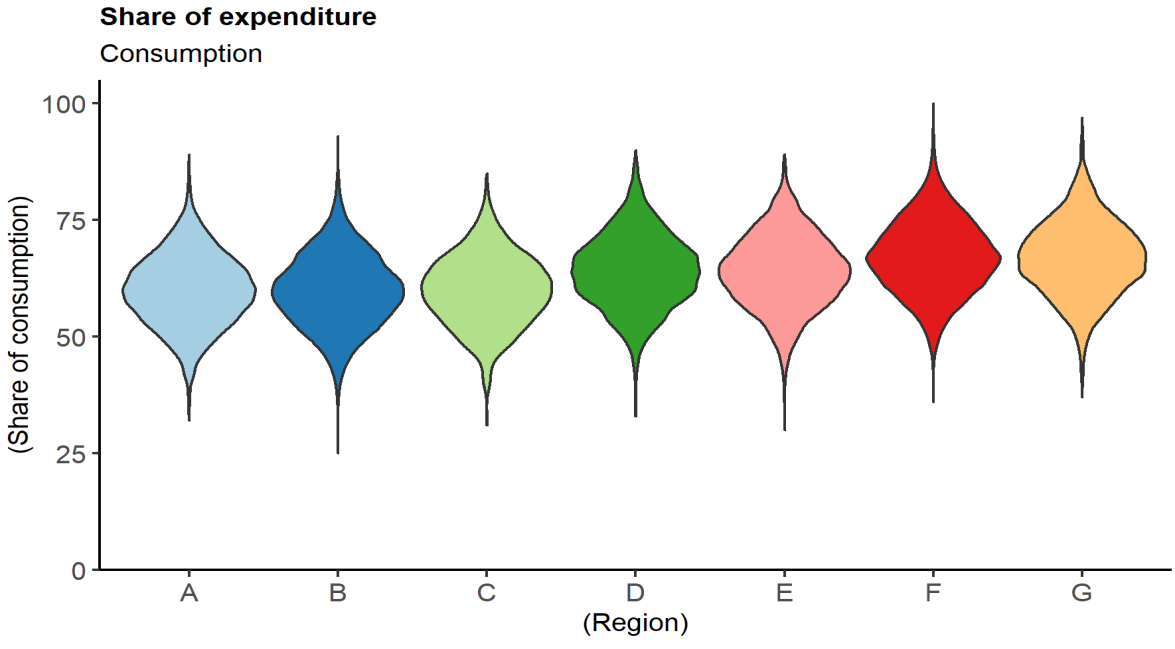
وبوجه عام، تشارك المرأة في القوّة العاملة بنسبة أقل من الرجل، وبالنسبة لكل من المنطقتين D وE، وحتى في F وG، فترتفع معدلات البطالة بين النساء. لاحظ أن هناك فرق بين أن تكون خارج القوى العاملة (غير مستعد للعمل) وبين أن تكون عاطلاً عن العمل (على استعداد للعمل، ولكن غير قادر على تأمين وظيفة). ومن ثم، وبفحص تاريخ مملكة X، يلاحظ المحلل أن المنطقتين F وG فقيرتان جداً، بالرغم من عدم اندماج النساء في القوى العاملة، فإن الحاجة إلى المال دفعت العديد منهنّ في هاتين المنطقتين إلى العمل. ومع ذلك، وبما أن المناطق فقيرة، تواجه النساء فيها تحديات تحول دون تمكنهن من الحصول على الوظائف.



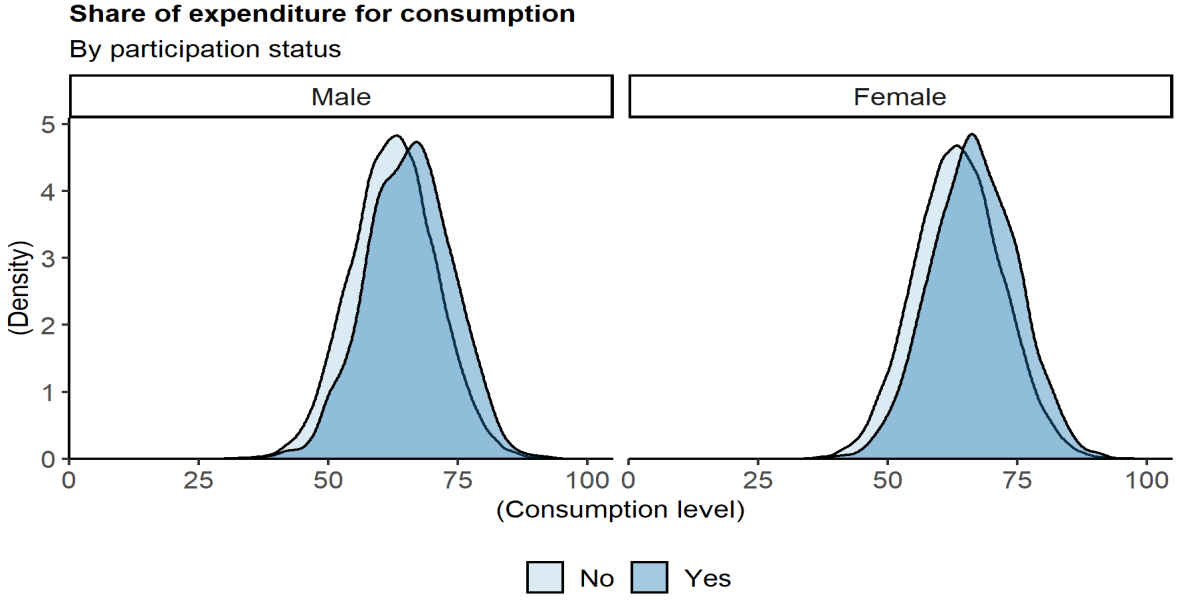
واستكمالاً لما سبق، لاحظ أن توزيع فئات الدخل العشرية قد سلط الضوء على كيفية تركُّز النساء الفقيرات في F وG. وهذا ينطبق على الرجال؛ ولو بنسبة أقل. في هذه الحالات، يتحقق المحلل من تاريخ البلاد مرة أخرى ويدرك أن F وG ليستا فقيرتين فحسب بل عانتا من الحرب مما أدى إلى تراجع أعداد الرجال. وهكذا، ترأس النساء معظم الأسر المعيشية المتبقية في تلك المناطق. وعلى نقيض ذلك، تزدهر المنطقتان A و C ويتركز فيهما أغنى أفراد البلد.



وتعقيباً على القصة السابقة، فإن لأفراد المناطق الأولى مصادر دخل متعددة، في حين أن لديهم في المناطق الأخيرة مصادر دخل أقل.

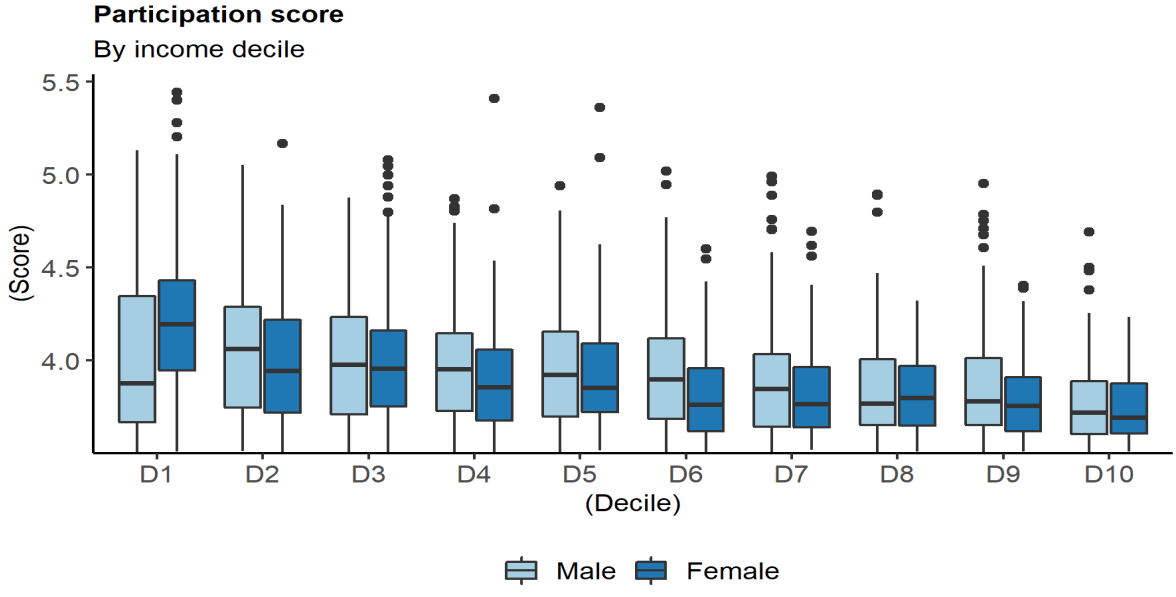


وتماشياً مع النتائج السابقة، يميل الأفراد في المناطق الفقيرة إلى تخصيص حصص أكبر من دخلهم لدعم احتياجاتهم الأساسية.



وفي هذا الشأن، من المثير للاهتمام أن نرى كيف يأخذ البرنامج الاجتماعي ذلك في الاعتبار، وبصفة عامة يمكنك أن ترى تخصيص الأفراد المشاركين في البرنامج، بشكل عام، حصصاً أكبر من نفقاتهم لتغطية

الاحتياجات الأساسية. وهذا إيجابي، لأنه يؤشّر أن البرنامج يستهدف السكان الأكثر حاجة لمزيد من الدعم. ومع ذلك، فكون التوزيعات قريبة من بعضها البعض يشير إلى عدم تركيز أولويات المستفيدين على هذا المتغير، وربما يبالغ من قيمة المتغيرات الأخرى.



وأخيراً، من خلال تحليل توزيع الدرجات مقارنه بالمدخول العشري، يُلاحظ أن (كما هو متوقع) الدرجة (التي تمثل حاجة الأفراد) تنخفض مع انخفاض الدخل. ومن المهم تسليط الضوء على عنصرين: أولاً، وجود فجوة غريبة في العشر D6. ثانياً، من المثير للاهتمام ملاحظة استفادة بعض الأشخاص من العشر الأغنى من البرنامج.

ويتيح الاستعراض الوصفي الحصول على صورة أفضل عن الاختلافات الإقليمية في البلاد، وبعض النقاط التي قد تكون ذات صلة لتسليط الضوء عليها من خلال SPP-RAF. الآن، يُترك للمتدرب التحقق من الرسوم البيانية المتبقية وتطوير قصص تكميلية لتلك المعروضة في هذا الفصل.

4. الخلاصة

في هذه المرحلة، اكتسب المتدرب الأدوات الأساسية اللازمة للعمل في مجال البحث، اصبح قادراً على إعداد تقارير تنتج إحصاءات وصفية قد توجه صانع السياسات في فهم الحالة العامة لبرنامج الحماية الاجتماعية. وبعد توضيح هذه الأفكار الأساسية، يذهب الجزء الثاني من الدليل إلى وصف الأدوات التحليلية وكيف يمكن لها أن توفر رؤى أعمق للتقييم.

الجزء الثاني. SPP-RAF

الفصل السادس. تحديد سمات المستفيدين

تركز الفصول الثلاثة التالية على الخطوات الثلاث الأولى من إطار SPP-RAF. وتستخدم هذه الفصول المثال الذي أُعطي سابقاً كمبدأ توجيهي في الفصل الثاني، كما تتضمن المعارف التي تم الحصول عليها بشأن سياق هذا البلد لتوضيح التحليل.

1. الغرض من تحديد سمات المستفيدين

وتهدف المرحلة الأولى من إطار SPP-RAF إلى تحديد الخصائص الاجتماعية - الديمغرافية للمستفيدين من البرنامج. وتسمح لصناع السياسات بجمع الأفراد ذوي الملامح المماثلة ضمن مجموعات، وتصور العوامل الهيكلية التي تجعل هذه الفئات السكانية عرضة للخطر. وبمجرد تحديد هذه الخصائص المشتركة، يمكن تصميم سياسات لمعالجة مصادر الضعف وتحسين رفاه هؤلاء السكان.

2. المخرجات

(أ) بناء مجموعات من المستفيدين وفق الخصائص الاجتماعية - الديمغرافية التي يتقاسمونها؛
(ب) فهم أفضل لطبيعة الفقر والضعف، فضلاً عن التباينات بحسب المناطق الإقليمية؛
(ج) طريقة الفحص السريع لتحديد الأفراد الذين ينتمون إلى كل مجموعة.

3. الوثيقة الختامية

(أ) التحديد الاستراتيجي لأسس الضعف الهيكلي المستمدة من الخصائص المشتركة للمستفيدين؛
(ب) الحصول على أدلة تسمح بابتكار خدمات تستهدف تلبية احتياجات مجموعات محددة من المستفيدين والحد من ضعفهم على المدى الطويل.

4. متطلبات البيانات

تتطلب هذه المرحلة إنشاء مجموعة بيانات تضم جميع الأفراد المستفيدين من برنامج معين خلال فترة محددة، وخصائصهم المشتركة. من الناحية المثالية، ينبغي أن تكون وحدة الدراسة "فردية"، ولكن يمكن استخدام المعلومات على مستوى الأسرة المعيشية اعتماداً على توفّر البيانات.

وبالنسبة لكل فرد، يجب إدراج القوائم التالية للخصائص الاجتماعية - الديمغرافية:

(أ) الجنس، العمر، مستوى التعليم، الموقع، الحالة الاجتماعية، مستوى الدخل، نوع العقد، المهنة، الصناعة (إذا كان الفرد يعمل) والجنسية؛
(ب) قيم الأسر المعيشية لمتغيرات الاستهداف التي يستخدمها برنامج الحماية الاجتماعية لتحديد أهلية المتقدمين (أو أكبر عدد ممكن)؛
(ج) المتغيرات الأخرى التي تهم صانعي السياسات.

5. التجمعات الاحصائية

يشرح هذا القسم العناصر المفاهيمية للتجمعات الإحصائية. وفي حين أنه يتناول المفاهيم الأساسية للتجمعات ويركز على تلك الموصى بها في الإطار SPP-RAF، يمكن للراغبين الاستفادة من كتاب Zelterman كمصدر جيد للتعمق في هذه المواضيع؛ لا سيما في الفصلين 11 و10 (في هذا الترتيب) (2015).

من السهل نسبياً عرض مشكلة التجمعات الاحصائية. يوجد عدد من الأفراد ذوي الصفات المختلفة ويوجب التحدي تجميعهم في مجموعات قليلة حيث تتشابه الخصائص داخل كل مجموعة وبين المجموعات تكون مختلفة. ومع ذلك، تجلب هذه الجملة بحد ذاتها تحديين:

(أ) ما هي المعايير التي يمكن استخدامها لتحديد مدى تشابه شخصين أو اختلافهما (الخطوة 1)؟
(ب) كيف يمكننا أن نقرر الاختلاف الكفيل بوضع شخصين في مجموعتين مختلفتين (الخطوة 2)؟

يترافق هذان السؤالان مع تحديات وبدائل متعددة سيتم استكشافها فيما يلي. ومع ذلك، وكما هو الحال في المواضيع السابقة، لا توجد لهذه المسائل إجابة واحدة فريدة. في الواقع، تثبت نظرية تسمى "البطة القبيحة" أنه بالنسبة للتجمعات الاحصائية، يوجد درجة من المعايير الذاتية دائماً، وعند هذه النقطة يحتاج المحلل إلى معرفة الخلفية التي تسمح له بفهم أفضل للبيانات على ضوء السياق (Watanabe, 1969).

الخطوة 1: المسافات

تكمن الخطوة الأولى في تحديد التجمعات الاحصائية بتعريف المسافة بين الأفراد. للتوضيح، يمكن تقديم ثلاثة أمثلة:

مقياس الإقليدية: وهو يقيس المسافة بين نقطتين، p وq، وهي محددة بطول الجزء الذي يصل بين النقطتين. ويتم احتساب ذلك من خلال نظرية فيثاغورس، كما هو يوضّح المثال التالي.

مثال: يبلغ أحمد 40 عاماً ولديه 20 سنة خبرة في العمل. فيما يبلغ محمد عمره 37.5 سنة ولديه 15 سنة خبرة في العمل. في هذه الحالة المسافة هي:

$$\sqrt{(40-37.5)^2+(20-15)^2}=5.59$$

مقياس مانهاتن: يحدد هذا المقياس المسافة بين نقطتين بمجموع الفرق المطلق بين كل قياس. ويعتبر أقوى من مقياس الإقليدية لأنه لا يربّع المسافة.

مثال: يبلغ أحمد 40 عاماً ولديه 20 سنة خبرة في العمل. فيما يبلغ محمد 37.5 سنة ولديه 15 سنة خبرة في العمل. في هذه الحالة المسافة هي:

$$|40-37.5|+|20-15|=7.5$$

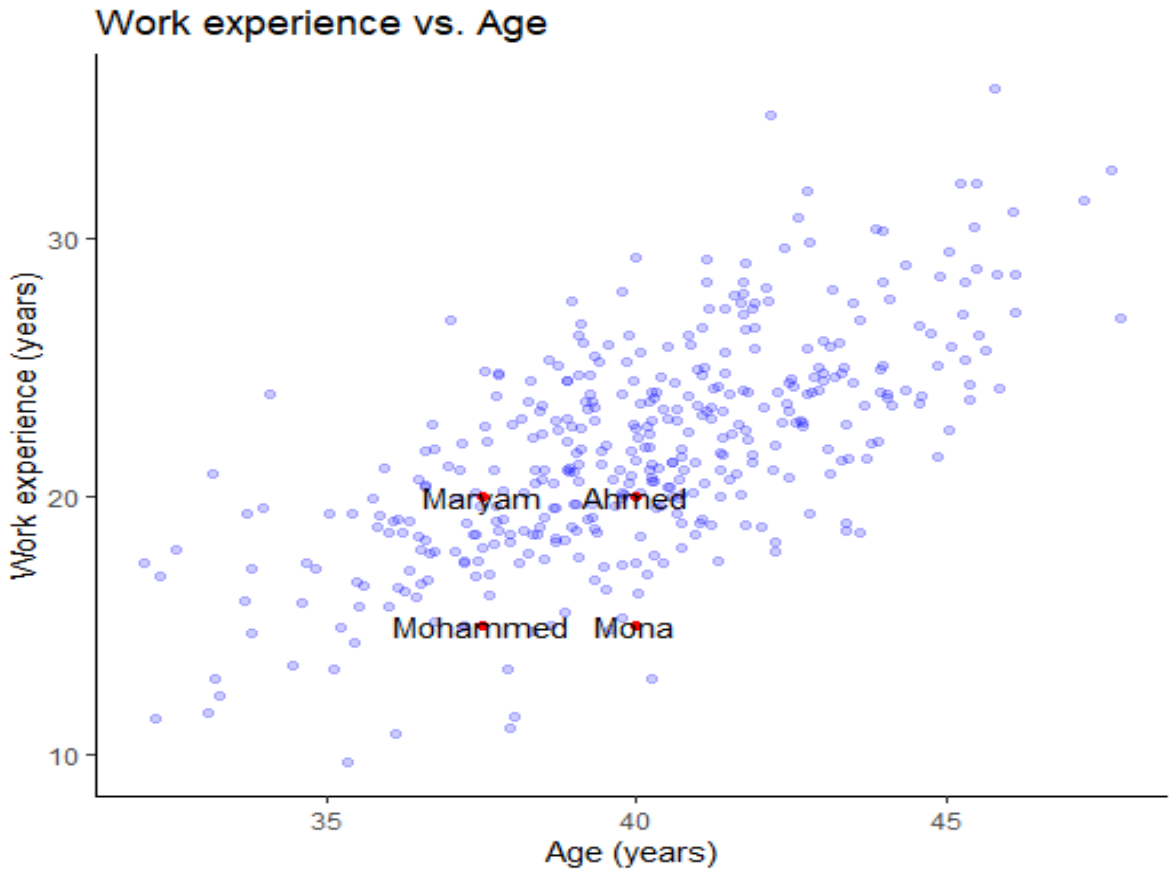
مؤشر جاكارد: يتم تطبيق هذا المقياس لتحديد التشابه بين الفئات، ثم لتحويله إلى مسافة، يمكنك طرح المجموع من قيمة 1.

مثال: لدى أحمد إنترنت وتلفاز، وليس لديه سيارة. في المقابل، ليس لدى محمد إنترنت أو سيارة، لكن لديه تلفاز. لاحظ أن المثالين يتقاسمان خيارين مشتركين من أصل ثلاثة. وبالتالي، فإن المسافة هي:

$$1-\frac{2}{3}=0.3333$$

كما ترى هناك اختلاف، واعتماداً على السياق، يمكن تفضيل أحدهما على الآخر. مثلاً، إذا كانت المسافة تتطلب قياس تحركات السيارات، يفضّل عادة مقياس "مانهاتن"، كما أنه يحاكي الحاجة إلى قياس عدد السيارات على طول شبكة من الشوارع بدلاً من اللجوء إلى أقطار قد تقطع المباني. وبالمثل، يعتبر مقياس (جاكّارد) Jaccard فعالاً جداً لبعض الحالات التي تكون فيها المتغيرات قاطعة. ومع ذلك، سيتركّز العمل في هذا القسم على المقياس الإقليدي لأنه قد يتضمن فئات، وبإدخال التعديلات الصحيحة، يمكن أن يتضمن ايضاً هيكل البيانات. ولكن، دعونا نسلط الضوء على مشاكل المقياس الإقليدي أولا.

دعونا نضيف إلى المثال شخصين آخرين: منى، 40 عاماً و15 سنة من الخبرة ومريم 37.5 عاماً و20 سنة من الخبرة. بتطبيق نفس الإجراء كما في المثال السابق، سيجد المتدرب أن المسافة بين مريم ومنى هي مرة أخرى 5.59. ومع ذلك، إذا قررنا قياس العمر بالأيام بدلاً من السنوات، ستكون المسافتان 16,303 للرجلين و15,572 للسيّدتين. ولا يعتبر ذلك تغييراً في الحجم، بل هو تغيير هيكلي في القياس. فمثلاً، كانت النسبة بين مسافتي الرجلين والسيّدتين في السابق 1 (أي أنها كانت متشابهة) ولكنها أصبحت 1.05. تؤثر هذه التغييرات في النسب على الخطوة التالية من عملية التجميع ممّا يجعلها غير مرغوب فيها. ومع ذلك، يُطرح تحدٍّ مفاهيميّ آخر في الشكل أدناه.



بطبيعة الحال يرتبط العمر والخبرة في العمل ارتباطاً وثيقاً، كما يوضّح الجزء العلوي من الرسم البياني. وعلاوةً على ذلك، وكما نلاحظ من موقع الأفراد الأربعة في المثال، هم هندسياً على نفس المسافة، ولكن الحال يختلف من الناحية المفاهيمية. في حين أن محمد وأحمد متماشيان مع هذا الاتجاه، إلا أن مريم ومنى تتعاكسان معه، لذلك من الشائع رؤية نمط أحمد ومحمد بدلاً من نمط مريم ومنى. لحسن الحظ، إذا تم تحويل المتغيرات بشكل كافٍ، عبر تعديلها من خلال علاقة التباين المشترك، ستصبح المسافة الإقليدية كافية. عندما يتم هذا التحول، لن تبقى المسافة تحت مسمى "الاقليدية" ولكنها تصبح مسافة (ماهالانوبيس) Mahalanobis، والتي تتميّز بأنها ثابتة من حيث الحجم وتشمل هيكل التباين المشترك. في المقطع اللاحق، وأثناء مناقشة التعليمات البرمجية، يشرح الدليل كيفية إجراء هذا التعديل على المسافة الإقليدية في R.(Mahalanobis, 1936)

الخطوة 2: التجميع

لأغراض التوضيح، قمنا بوضع سيناريو جديد حيث يتشارك كل من منى ومريم وأحمد ومحمد في مصفوفة المسافة الممثلة في الجدول أدناه ويتم ذلك عن طريق معيار المسافة (ما يهمّ في هذه الخطوة هو المصفوفة، وليس الطريقة المعتمدة).

	منى	مريم	أحمد	محمد	فاتح	بلسم
منى	0	2	3	9	4	5
مريم	2	0	1	3	10	1
أحمد	3	1	0	8	5	3
محمد	9	3	8	0	10	5
فاتح	4	10	5	10	0	2
بلسم	5	1	3	5	2	0

وبطبيعة الحال، فإن قيمة القطر هي 0 لأن المسافة مع النفس منتفية. الآن، كيف نبدأ بتجميع الأفراد؟ بطبيعة الحال، يحتاج أحمد ومريم إلى أن يكونا معاً، لأنهما على مسافة 1 (الأقرب). لكن الآن، ماذا عن بلسم؟ وهي قريبة جداً من مريم (المسافة 1)، ولكنها على مسافة 3 من أحمد. وعلاوة على ذلك، فإن بلسم قريبة من فاتح، لكن الأخير بعيد جداً عن مريم. وهناك حالياً عدة خيارات للقيام بهذه العملية، ثلاثة منها ممثلة في الشكل أدناه.

الربط الوحيد: المسافة بين أفراد المجموعة (أ) وأفراد المجموعة (ب) هي الحد الأدنى للمسافة بين شخص من المجموعة (أ) وآخر من المجموعة (ب).

الربط الكامل: المسافة بين أفراد المجموعة (أ) وأفراد المجموعة (ب) هي الحد الأقصى للمسافة بين شخص من المجموعة (أ) وشخص من المجموعة (ب).

الربط الوارد: تحسب المسافة بين المجموعات بطريقة تدمج التباين بينها (Ward, 1963).

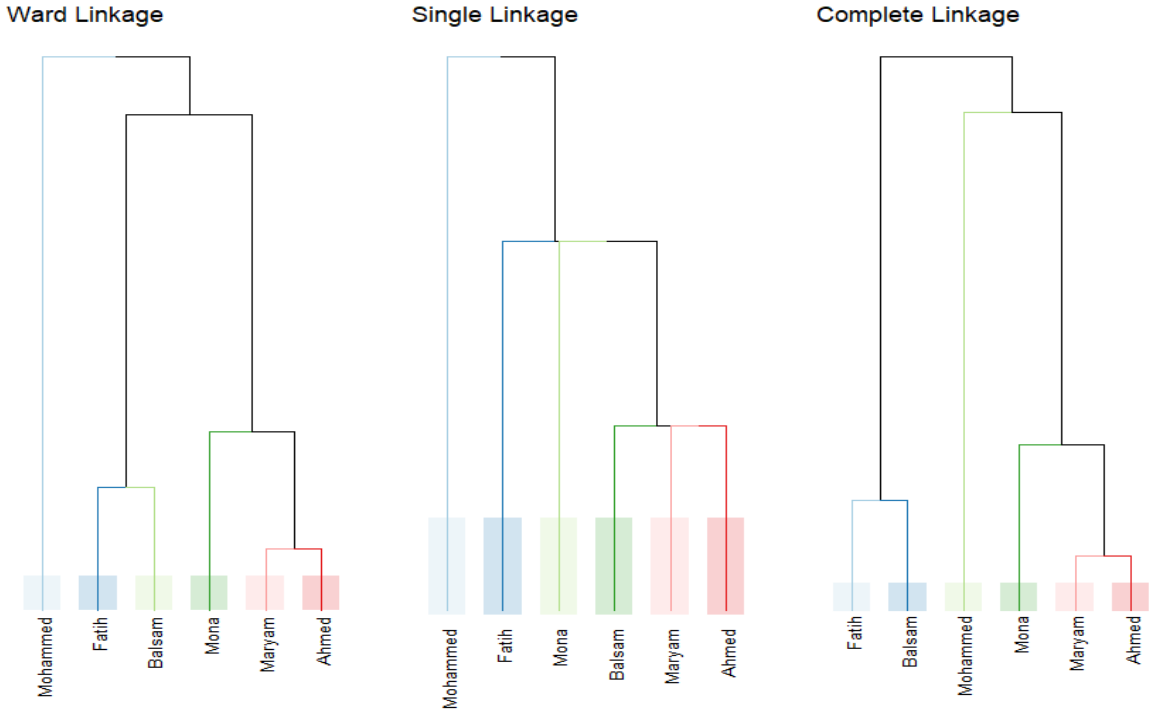
كما ترون في النتائج، فإنّ طريقة تجميع الأفراد مختلفة. مثلاً، أحمد ومريم دائماً معاً، ولكن في الربط الكامل، فإن التقارب بين بلسم وفاتح يجعلهما بعيدين عن مريم وأحمد، ولكن في الربط الوحيد، تزداد قرباً من مريم. وبوجه عام، لا يوجد دليل توجيهي واضح بشأن الطريقة التي يمكن اختيارها، تتمثّل التوصية باختبار مختلف الروابط لمعرفة تلك التي تعكس البيانات بشكل أفضل. ومع ذلك، إذا كان الهدف إنشاء مجموعة متوازنة، يميل مقياس ward للعمل بشكل أفضل لأن الطريقتين الأخريين تنتجان مخططات تسلسل متطرفة بسبب إدارتهما لمعياري الحد الأدنى والحد الأقصى من المعايير التي يديرونها.

بمجرد أن يكون لديك مخطط تسلسل القرارات، يصبح التحدي التالي، على مستوى السياسة، في تحديد عدد المجموعات التي تريدها. مثلاً:

(أ) الربط الوارد للشكل 3، على افتراض استعداد الحكومة لوضع ثلاث استراتيجيات استهداف، فإن التوصية المبنية على هذا التجميع هي وضع استراتيجية استهداف لمنى وأحمد ومريم، وأخرى لبلسم وفاتح، وواحدة لمحمد؛

(ب) الربط الوحيد؛ حيث سيوصي بوضع بلسم ومريم وأحمد في مجموعة واحدة، ومنى وفاتح في مجموعة اخرى، ومحمد وحيداً في مجموعة أخيرة.

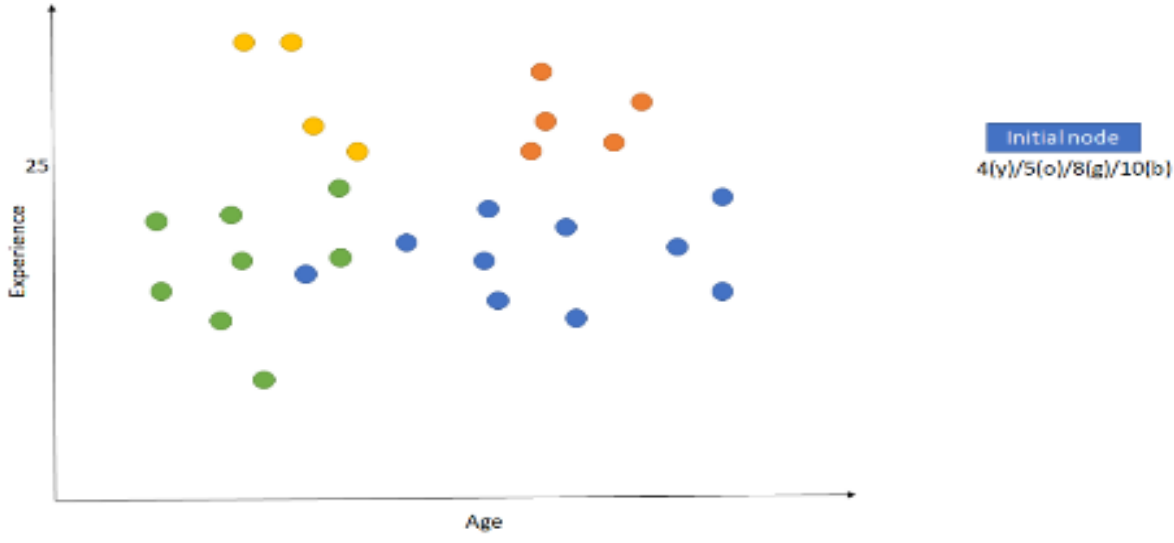
بذلك، تنتج التقنيات المختلفة مجموعات مختلفة ولكن ذات قواسم مشتركة تساعد صانعي السياسات على فهم الحالة. مثلاً، يختلف محمد كثيراً عن البقية، في مختلف التقنيات، ويمكن دائماً جمع أحمد ومريم معاً.



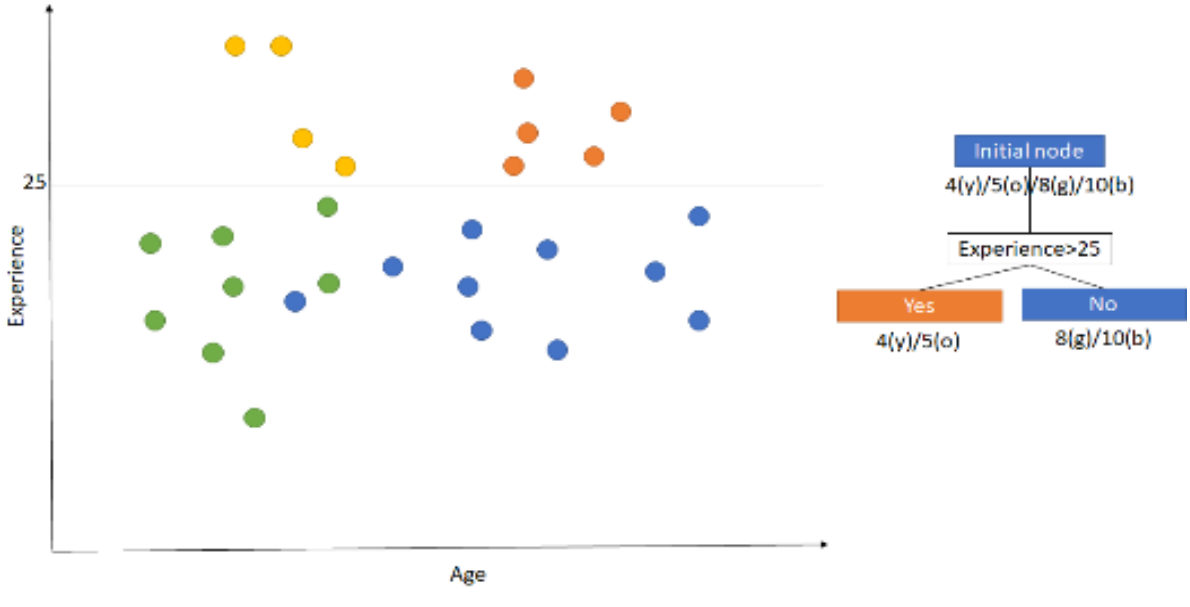
6. خطط تسلسل التصنيف

في حين أن التجميع هو عملية بصرية تساعدك على تجميع الأفراد في مجموعات، إلا أنه لا يقدم لك فهماً عن سبب تجميع هؤلاء وعن ماهية المتغيرات التي تقود القواسم المشتركة. لحل هذه المشكلة، تمّ تطوير تقنيتين، إحداهما وصفية للغاية، وبالتالي يُشرح مفهومها مباشرة في المثال، والأخرى تقنية أكثر، وقد خُصص هذا القسم للتوسع في شرحها.

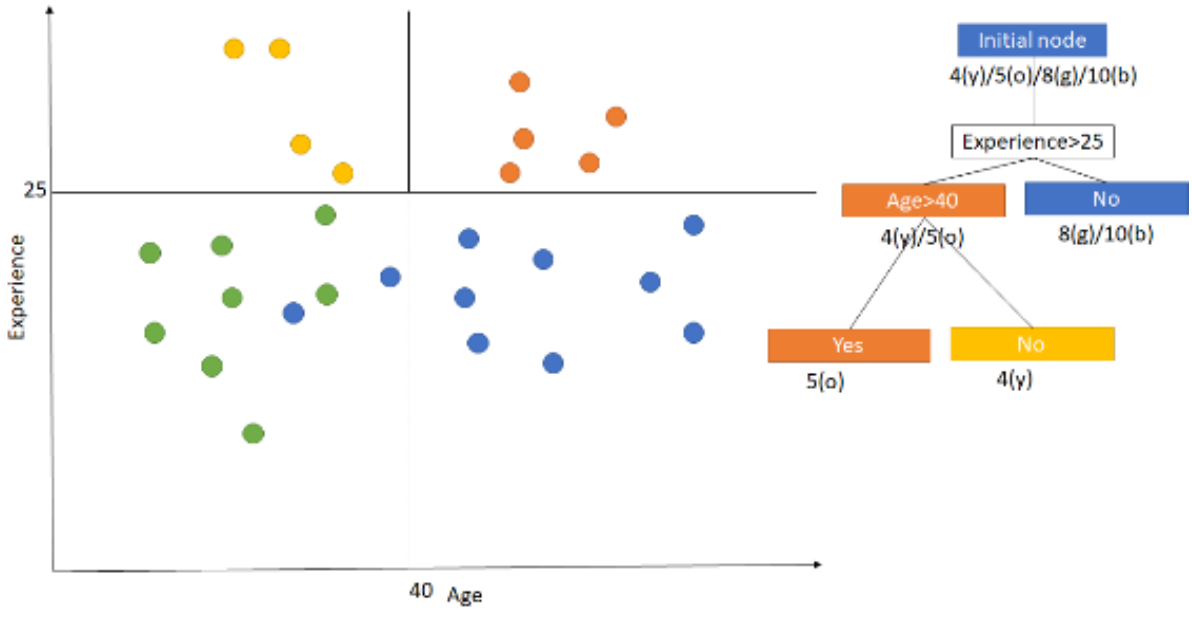
تعرف الطريقة التقنية باسم "التصنيفات وخطط التسلسل" أو "شجرة الانحدار" CART وهي إحدى أساسيات التعلم الآلي. ونظراً لكون هذه التقنية تتطلب خلفية معينة لدى القارئ، يقدم هذا الفصل لمحة مبسطة حول كيفية عملها، يمكن للقرّاء الراغبين بالحصول على المزيد من التفاصيل قراءة Breiman التي يتوسع في شرح وظيفتها(1984).



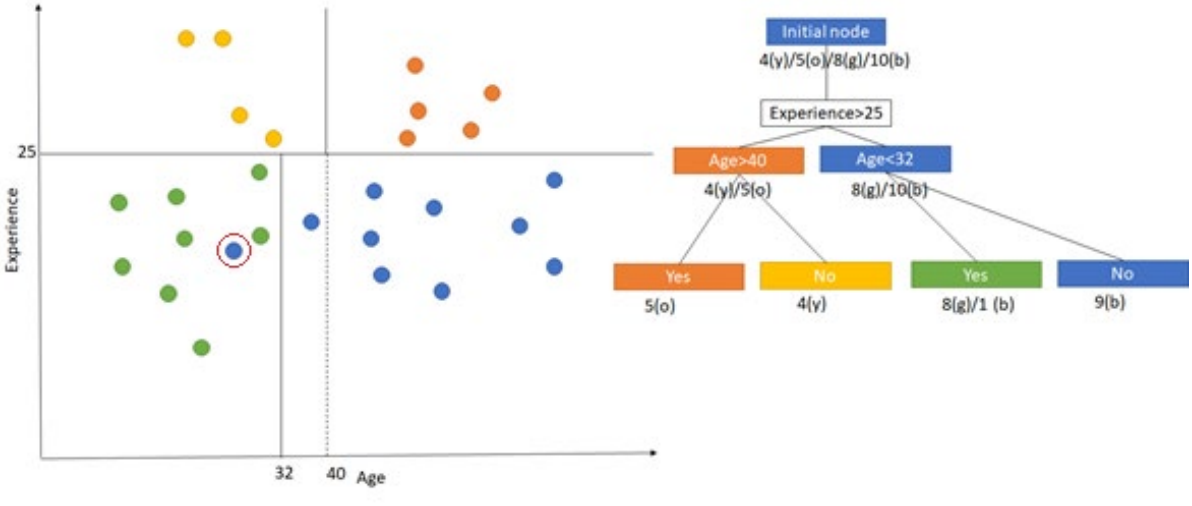
في الأصل، تبدأ الشجرة في الجذور، وهذا يعني جميع السكان. في هذه الحالة المجموعة الأكبر هي تلك الزرقاء مع 10 ملاحظات، لذلك يتم تعيين اللون الأزرق إلى الجذور. وبعبارة أخرى، إذا لم يكن لديك معلومات إضافية، يفضّل اختيار شخص عشوائيا، من القسم الملون بالأزرق.



وتعتبر التجربة من أهم المعايير لأنها قد تضع تقسيماً منصفاً وواضحاً بين المجموعات. وإذا كان الشخص يتمتع بأكثر من 25 عاماً من الخبرة يوضع في الجزء البرتقالي (مجموعة الأغلبية) وإلا يبقى في الجزء الأزرق.



يرتكز التصنيف التالي على الأشخاص الذين تتعدّى أعمارهم الـ 40 عاماً من ضمن المجموعة التي لدى أفرادها خبرة في العمل تتعدى 25 عاماً. من هنا نلاحظ أن الأولوية تُعطى للأسئلة المرتبطة بتوصيف الأفراد لفهم خصائص كل مجموعة. في هذه الحالة بالنسبة للبرتقالي والأصفر، فإنّ العنصر الأول هو التجربة والثاني هو العمر.



لاحقاً، يتم القيام بتقسيم مماثل للأشخاص ذوي الخبرة المنخفضة، إلاّ أنّ العمر يصبح 32 بدلاً من 40. لاحظ أن التصنيف ليس مثالياً (ولا يحتاج إلى أن يكون كذلك). في هذه الحالة، يمكن ملاحظة حصول أحد الأشخاص على تصنيف خاطئ ووضعه في الجزء الأخضر بدلاً من الأزرق إذا ما كنا اتّبعنا معيار الشجرة. من وجهة النظر

الإحصائية، يمكن زيادة نمو الشجرة حتى تتمتع كل ورقة (عقدة نهاية) بلون مختلف عن الأخرى. من خلال الزيادة المفرطة للأعداد، يمكن أن يخطىء العداد في الكتابة وتتمّ إضافة شخص، والعمل باعتباره ضمن التركيبة الهيكلية. من ناحية أخرى، تعتبر الشجرة وسيلة سريعة للتوصيف، لذلك إذا أردت ضمّ أي فرد جديد واضطررت لتعريف المجموعة التي ينتمي إليها بسرعة، عليك طرح سؤالين. بالطبع، فإنّ طرح المزيد من الأسئلة يمنحك إجابات أكثر دقة، ولكن الأسئلة مكلفة مادّياً، وبالتالي كلما قل عدد الأسئلة قلّت التكاليف.

وبعد توضيح العنصرين النظريّين الرئيسيّين في هذا القسم، يتناول القسم التالي الرموز، بطريقة مماثلة للفصل الخامس، ويسلط الضوء على كافة خطوات تحليل التوصيف.

7. تحليل التوصيف

يستمر هذا القسم باستخدام التعليمات البرمجيّة المعرّفة بـ chapter 01 (لأن التوصيف مقترن بـ chapter 1- من SPP-RAF). لاحظ أن التعليمات البرمجية تبدأ بنفس الطريقة كما في Chapter 00 لأنه يقوم بتحميل البيانات وتنسيق القواميس والفئات ذات اللغة الواحدة، وعلى النقيض من التعليمات البرمجية السابقة، نلحظ تغييراً في السطور من 37 إلى 39 للدلالة على أن التوصيف ينطبق فقط على الأشخاص المستفيدين من البرنامج.

```
37 # 1.4| Subset ------------------------------------------------------------
38 # Dataset related to those individuals that participate in social assistance programs.
39 dataParticipation = dataFrame %>% filter(participant == 'Yes')
```

ويتناول الجزء التالي إنشاء متغيرات وهمية، وهي عبارة عن متغيرات رقمية تمثل تلك الفئوية. مثلاً في نوع الجنس، يمثل 0 الإناث و1 يمثل الذكور. في مجموعة البيانات هذه، هناك العديد من المتغيرات الفئوية. ولتسهيل الأمور عند احتساب النقاط، تحوّل هذه الفئات إلى متغيرات وهمية، عبر استخدام دالة dummy_cols.

```r
# 2.1| Machine learning techniques ------------------------------------------
# 2.1.1| Hierarchical clustering ------------------------------------------
# - Incremental dummies:
incrementalDummy = dataFrame %>%
  dummy_cols(select_columns = c('region','gender','education',
                                'employment','incomeQuantile',
                                'internetAcces','healthAcces'),
             remove_first_dummy = T)
dataParticipation = incrementalDummy %>% filter(participant == 'Yes')
dataParticipation = dataParticipation %>%
  dplyr::select(-c('region', 'gender', 'education', 'employment',
            'incomeQuantile', 'participant','internetAcces', 'healthAcces'))

dataParticipation = dataParticipation %>% dplyr::select(-c('score'))

rownames(dataParticipation)=paste0('Case', 1:dim(dataParticipation)[1])

# - Factores for the original dataset:
dataFrame$gender  = factor(dataFrame$gender,
                           levels = gender[,2],
                           labels = gender[,4-language])
dataFrame$education  = factor(dataFrame$education,
                              levels = studies[,2],
                              labels = studies[,4-language])
dataFrame$employment  = factor(dataFrame$employment,
                               levels = labor[,2],
                               labels = labor[,4-language])
dataFrame$incomeQuantile  = factor(dataFrame$incomeQuantile,
                                   levels = deciles[,2],
                                   labels = deciles[,4-language])
dataFrame$participant = factor(dataFrame$participant,
                               levels = confirmation[,2],
                               labels = confirmation[,4-language])
```

العنصر المهم الآخر هنا هو إدراج dplyr قبل الدالة (select). وأحياناً تتضمن الحزم المتعدّدة التي يستخدمها R اسماً مماثلاً لوظائف مختلفة. في هذه الحالة، لتجنب الأخطاء، يفضّل استخدام "::" لتحديد الحزمة التي تأتي منها الدالة بشكل صريح.

بمجرد الانتهاء من هذا التحويل، تبدأ عملية احتساب المسافة.

```r
# '1_1'| Distance matrix ------------------------------------------------
#Variance of dataset variables

aux1=apply(dataParticipation, 2, var)

#Select variable whose variance isn't 0 and scaling them

dataParticipation=dataParticipation[, aux1!=0 ]
dd <- mahalanobis.dist(dataParticipation)
dd =as.dist(dd)
```

ومن بين هذه الخطوط، تكمن العناصر الرئيسية في تعريف مسافة ماهالانوبيس (Mahalanobis)، المستخدمة في السطر 86 من خلال دالة (mahalanobis.dist)، وتحويله إلى أداة للمسافة بحيث يقوم R بالعملية الحسابية عن طريق (as.dist).

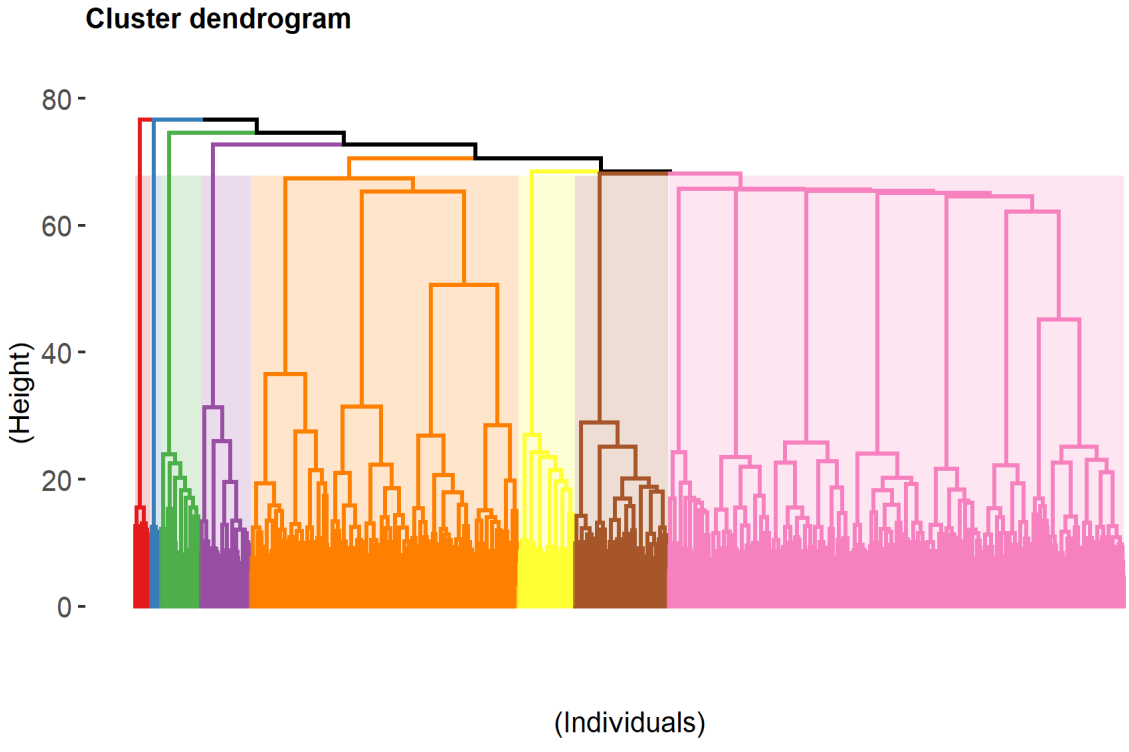
ويظهر المقطع التالي من التعليمات البرمجية، حتى السطر 137، الطريقة التي يستعرض فيها R مصفوفة المسافة. يغيب هذا الجزء من التعليمات من هنا لأنه يتّبع نفس المنطق المعروض في الفصل 3. ومع ذلك، تُعرض مصفوفة المسافة لتقييم عناصرها الرئيسية.

يكشف هذا التمثيل البياني لمصفوفة المسافة الأفراد القريبين أو البعيدين من بعضهم البعض. ويمثّل كل عامود (وصف) فرداً واحداً، ويعكس لون الخلية المسافة بينه وبين الآخرين، حيث الأزرق هو الحد الأدنى للمسافة والأحمر هو حدّها الأقصى. والخط القُطري هو الأزرق. وإذا ما تمعنت النظر، على طول الخط القُطري ستجد كتل ضوء تمثل هؤلاء الأفراد الذين تفصلهم مسافات صغيرة نسبياً. في المقابل، يمكنك ملاحظة مربعات صغيرة لونها أخضر فاتح في الزاوية اليسرى السفلى. في حين أنها متشابهة من الداخل، فهي مختلفة جداً عن أي فرد آخر (تنعكس في تلك الصفوف وحولها الأعمدة الحمراء). ويريح هذا التصور النظر، لكنّه لا يسمح لك بتحديد ترتيب عملية التجميع، لهذا السبب، وعلى غرار التفسير المفاهيمي، تبرز الحاجة للّجوء إلى الرسم الشجري أو الرسم التخطيطي التفرعي.

```
hc           = hclust(dd, method = 'ward.D2')
otter.dendro = as.dendrogram(hc)

dendro.plot  = ggdendrogram(data = otter.dendro, rotate = TRUE)+
  scale_x_discrete(expand = expansion(mult = c(0,0))) +
  scale_y_continuous(expand = expansion(mult = c(0,0)))+
  theme(axis.title.y = element_blank(),
        axis.ticks.y = element_blank())

cutNumber=8

treeOrdering  = hc$labels[hc$order]
sub_grp       = cutree(hc, k = cutNumber)
dataParticipationClusters = factor(sub_grp)

newPlot = fviz_dend(hc, k = cutNumber, color_labels_by_k = T, palette = 'Set1', show_labels = F,
                    rect = T, lower_rect = -0.02, rect_border = 'Set1', rect_fill = T) +
  labs(title    = paste0(as.character(labCodes[, 5*(1-language)+2])),
       subtitle = as.character(labCodes[, 5*(1-language)+3]),
       caption  = as.character(labCodes[, 5*(1-language)+6]),
       x        = as.character(labCodes[, 5*(1-language)+4]),
       y        = as.character(labCodes[, 5*(1-language)+5])) +
  scale_y_continuous(position = if (language == 0) {'right'} else {'left'}) +
  theme_classic() +
  theme(legend.position = 'bottom',
        text              = element_text(family = 'Georgia'),
        axis.line.x       = element_blank(),
        axis.line.y       = element_blank(),
        axis.text.x       = element_blank(),
        axis.ticks.x      = element_blank(),
        axis.text.y       = element_text(size = scale_factor * 10, family = 'Georgia'),
        axis.title.x      = element_text(hjust = 0.5, size = scale_factor * 10, family = 'Georgia'),
        axis.title.y      = element_text(hjust = 0.5, size = scale_factor * 10, family = 'Georgia'),
        legend.text       = element_text(size = scale_factor * 10, family = 'Georgia'),
        strip.text        = element_text(size = scale_factor * 10, family = 'Georgia'),
        plot.title        = element_text(face = 'bold', size = 10, family = 'Georgia', hjust = if (language == 0) {1} else {0}),
        plot.subtitle     = element_text(size = 10, family = 'Georgia', hjust = if (language == 0) {1} else {0}),
        legend.key.size = unit(0.5 * scale_factor, 'cm'))
```

الرسم الشجري هو مخطط يوضح نتائج نظام التجميع الهرمي. بالنسبة لهذا الرسم، يكمن الرمز المفتاح بين السطرين 156 و193. فيستخدم السطر 156 مصفوفة المسافة المحسوبة سابقاً في السطر 86 لينتج التجميع باستخدام روابط واردة عبر دالة hclust. وإذا كان من الضروري استخدام الروابط الوحيدة أو الكاملة، فيجب أن يتمّ التغيير بادخاله في هذا السطر. ويفيد السطر التالي في عملية التفسير التي يضطلع بها برنامج R. ثم، بين الخطين 159 و163، هناك رسم شجري صغير تم إنشاؤه، ولكن لن يُستخدم في هذه المرحلة. ويحدد السطر 165 عدد المجموعات التي نريد تحليلها، يتمّ حفظ المعلومات بين السطور 167-169 بحيث يمكن استخدامها للتحليلات لاحقاً. وأخيراً، تُنشىء السطور من 171 إلى 193 رسماً شجرياً يوفر معلومات مهمة حول هيكلية المستفيدين. لاحظ هنا ايضاً تُستخدم دالة ggplot، مع وجود فرق دقيق في البداية، ويتم تشجيع القارئ للتعامل مع هذه العوامل (Parameters) والتمرّن لفهم كيفية إنتاج رسوم بيانية مختلفة.

ويمكن اعتبار الرسم الشجري النهائي جذاب جداً. وهناك مجموعتان صغيرتان جداً من الأفراد (الأحمر والأزرق) يبدو أن لديهما اختلافات شديدة مع جميع المجموعات. كما نرى بعض المجموعات ذات الحجم المعقول، وأخيراً تشكّل المجموعتان الكبيرتان، الممثلتان باللونين البرتقالي والوردي، مجموعة كبيرة ذات سمات مماثلة. ومن خلال هذه المعلومات، يمكن لصناع السياسات البدء بتحديد الأماكن التي قد يكون فيها الاستهداف أكثر كفاءة. فمن ناحية، إذا استهدفت السياسة المجموعات المتوسطة والصغرى، فقد يكون من الممكن تحقيق نتائج سريعة، لأن المجموعات محددة جداً، ومع ذلك يكون التأثير على جميع السكان محدود بطبيعة الحال. ولكن، إذا كان صانع السياسة يستهدف المجموعة الكبيرة، قد تكون النتائج أقل وضوحاً بسبب التنوع الكبير داخلها (لاحظ مثلاً أنه يمكن تقسيم الوردي أيضا إلى سبع مجموعات فرعية استناداً إلى التجميع)، ومع ذلك ستكون هذه النتائج قادرة على تغطية المزيد من الناس. وتعتمد هذه القرارات على توفر النتائج، والميزانية، والأفضليات السياسية لصانع السياسات. ولكنّ الأداة بحدّ ذاتها توفر التوجيه بشأن البدائل الممكنة.

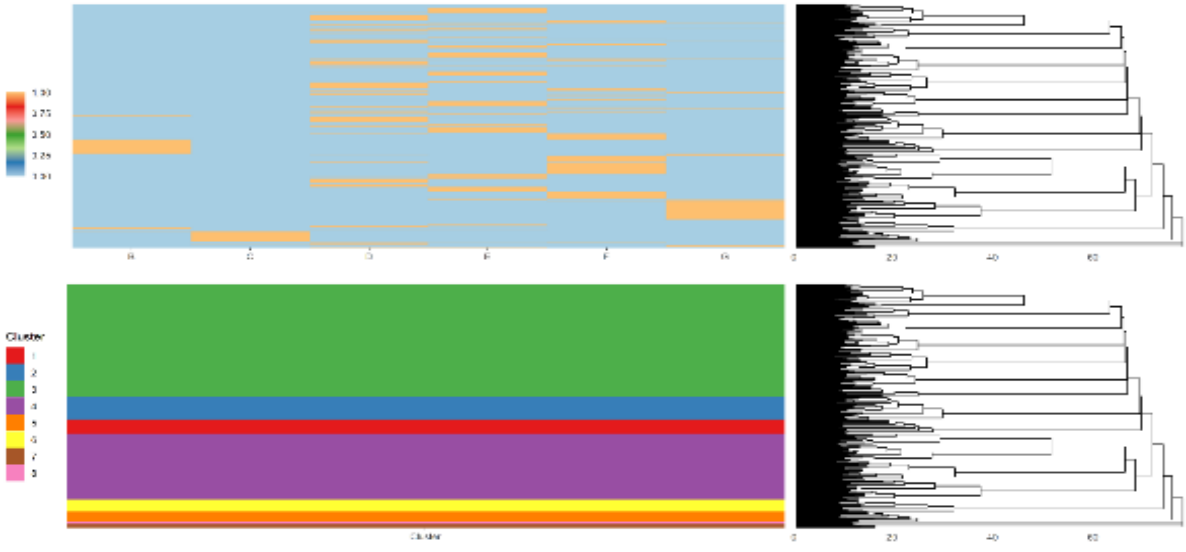


ومع ذلك، وكما ذكر سابقاً، لا يكفي أن نرى المجموعات، فالعنصر الرئيسي هو فهم كيفية تصنيفها. التحليل الأول هو وصفي بحت ويتضمن رسماً شجرياً يسمح لنا بالتوصل إلى استنتاجات حول المتغيرات ذات الصلة.

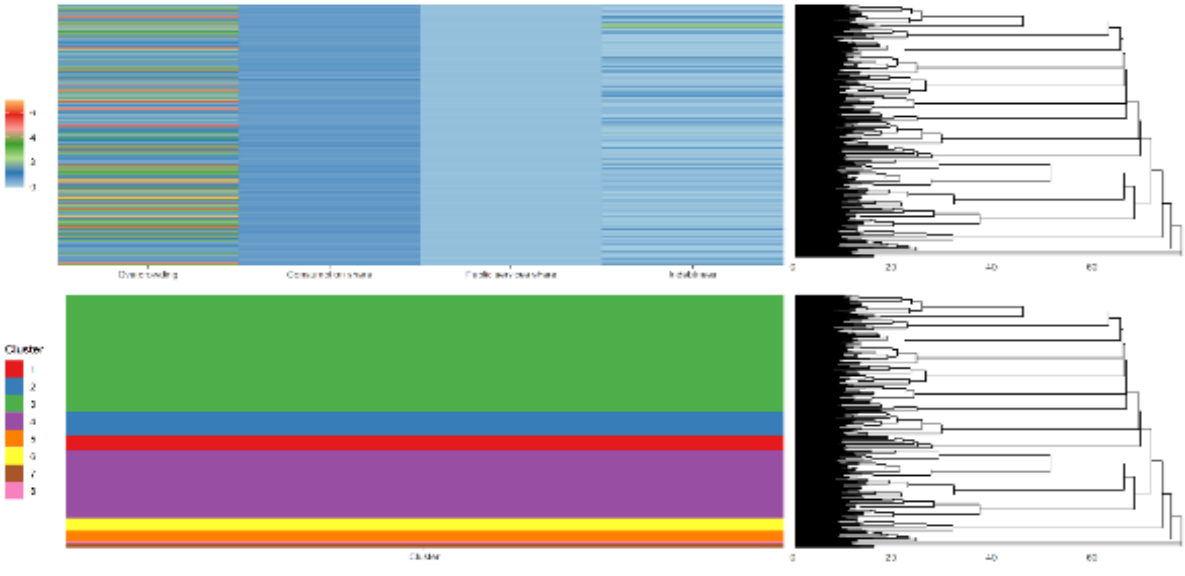
في هذه الحالة، تكون سطور التعليمات البرمجية الرئيسية مماثلة لـ 274

```
273 # Export:
274 newPlot = grid.arrange(heatmap.plot, dendro.plot, heatmap.plot2, dendro.plot, ncol = 2, widths = c(2, 1))
```

ينشئ هذا الخط مخططاً عن طريق وضع أربع مخططات معاً، في إطار يحتوي على عامودين، ويمثل عرض العامود الأول ضعفي عرض العامود الثاني.



لأغراض التوضيح، خذ الرسم البياني الأول الذي ينشأ عبر التعليمات البرمجية. تبرز الرسوم البيانية السفلية الرسم الشجري والمجموعة المواضيعية. وبالتالي، تأتي المجموعة الثالثة، المحدّدة باللون الأخضر في الجزء الأول من الرسم الشجري. في المقابل، تتمثل المجموعة الأولى، المحدّدة باللون الأحمر، بخط صغير في منتصف الرسم الشجري. من خلال مطابقة تلك المسافات في الشكل العلوي، يمكن التحقق من محتوى كل مجموعة. مثلاً، لاحظ الإطار البرتقالي في C (الرسم البياني العلوي) الذي يطابق المجموعة الخامسة (باللون البرتقالي). وهذا يؤشر أن المجموعة الخامسة تضم في الغالب الأشخاص من C. وبالمثل، تضمّ المجموعة الأولى الأشخاص من المنطقة B.



في حين أن المتغيرات ليست كلها ذات فائدة للتوصل إلى استنتاجات، لكنّ لبعضها أنماطاً مميزة جداً. مثلاً، عند التحقق من الازدحام، ترتبط القيم العليا بالمجموعة الرابعة. وفوقها مباشرة تكون القيم منخفضة جداً، في المجموعة الأولى. في الدليل عُرضت كافة المتغيرات بواسطة مجموعات، ولكن بالنسبة للتقارير، واستناداً إلى المعايير التي يريد المحلل تسليط الضوء عليها، يمكن إجراء الرسوم البيانية بعدد مختلف من المتغيرات.

```
#1.3.4 Tabulate results
indicatorCode <- '1_35'
part          <- ''
newSheet      <- addWorksheet(wb, sheetName = indicatorCode)

presentData=dataParticipation%>%dplyr::select(c("Individual","Clusters"))%>%
  right_join(Aux)%>%
  group_by(variable, Clusters)%>%
  summarise(value=mean(value))%>%
  spread(variable,value)

writeDataTable(wb, sheet = indicatorCode, x = presentData, startRow = 1, startCol = 1)
```

يُنتج الجزء الأخير من التعليمات البرمجية جدولاً يحتوي على ملخص عن كافة الوسائل التي تم احتسابها. وهذا يمكن أن يساعد في مراجعة بعض العناصر التي لم تتضح عبر الرسوم البيانية. مثلاً، لاحظ أن المجموعة الثامنة هي الوحيدة التي تضم طلاب الدراسات العليا. وبالمثل، تتركز الفئتان F وG، وهما أفقر مجموعتين، في المجموعتين الرابعة والسابعة، وهما أيضاً، كما هو متوقع، الفئتان اللتان تحتويان أعلى معدلات البطالة. وبهذه الطريقة، مثلاً، يمكن البدء بتحديد المجموعتين الرابعة والسابعة على أنهما تضمّان الأفراد الذين هم الأكثر عرضة للمخاطر.

	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R
1	Clusters	Age	Household	Overcrowd	Consumpt	Public serv	Indebtnes	B	C	D	E	F	G	Primary	Secondary	TVET	Postgrade	Unemploy
2	1	31,6287425	3,33532934	1,55796113	0,62532934	0,1708982	0,26694611	1	0	0	0	0	0	0,26946108	0,37724551	0,18562874	0	0,15568862
3	2	34,0395683	4,01079137	2,40935252	0,6707554	0,17010791	0,28892086	0	0	0,32014388	0,29496403	0,25899281	0	0,39928058	0,23741007	0,13669065	0	0,21223022
4	3	32,5402214	3,59926199	1,79766854	0,65056089	0,16859041	0,30923247	0,00811808	0,00295203	0,28413284	0,26420664	0,09815498	0,01697417	0,26273063	0,30479705	0,12841328	0	0,22435424
5	4	35,5219573	4,39774153	2,92029187	0,69397742	0,16971142	0,26200753	0	0	0,33174404	0,12797992	0,38895859	0,32622334	0,39523212	0,15683814	0,11919699	0	0,3174404
6	5	33,4453782	3,1512605	1,45853341	0,62663866	0,16705882	0,31201681	0	1	0	0	0	0	0,22689076	0,35294118	0,1512605	0	0,16806723
7	6	32,5743243	2,85135135	1,10119316	0,63040541	0,16743243	0,30871622	0,21621622	0	0,12837838	0,12837838	0	0	0,20945946	0,34459459	0,12162162	0	0,19594595
8	7	36,212766	5,0212766	3,55851064	0,69170213	0,1693617	0,73340426	0,0212766	0	0,25531915	0,25531915	0,23404255	0,10638298	0,46808511	0,19148936	0,08510638	0	0,31914894
9	8	38,4666667	2,66666667	1,08746032	0,64333333	0,165	0,338	0,06666667	0,1	0,1	0,23333333	0,16666667	0,03333333	0	0	0	1	0,13333333

الجزء الأخير من العملية هو إنشاء شجرة القرار، وهي طويلة ولكن من السهل نسبياً احتسابها.

```
# '1_4'| Decision tree ------------------------------------------------------
# Tuning ------------------------------------------------------------------
# Separation of samples:
set.seed(123)
dataParticipation=dataFrameFinal%>%filter(participant %in% as.character(confirmation[2,3:4]))%>%dplyr::select(-c('score', 'partici
lang=1

dataParticipation$region = factor(dataParticipation$region,
                                  levels = Region[,2],
                                  labels = Region[,4-lang])
dataParticipation$internetAcces = factor(dataParticipation$internetAcces,
                                         levels = response1[,2],
                                         labels = response1[,4-lang])
dataParticipation$healthAcces = factor(dataParticipation$healthAcces,
                                       levels = response1[,2],
                                       labels = response1[,4-lang])
dataParticipation$gender  = factor(dataParticipation$gender,
                                   levels = gender[,2],
                                   labels = gender[,4-lang])
dataParticipation$education  = factor(dataParticipation$education,
                                      levels = studies[,2],
                                      labels = studies[,4-lang])
dataParticipation$employment  = factor(dataParticipation$employment,
                                       levels = labor[,2],
                                       labels = labor[,4-lang])
dataParticipation$incomeQuantile  = factor(dataParticipation$incomeQuantile,
                                           levels = deciles[,2],
                                           labels = deciles[,4-lang])

dataParticipation$Clusters=factor(sub_grp)
```

يأخذ الجزء الأول إطار البيانات الأصلية، يختزل الأفراد الذين كانوا جزءاً من عملية التجميع، مستنداً إلى العوامل ذات الصلة، يضع اللمسات الأخيرة (السطر 459)، ويضيف إليها المجموعات التي تم احتسابها سابقاً.

في هذه المرحلة يبرز نقاش تقني يتعلق بالرسم الشجري. وكما تم شرحه في المناقشة المفاهيمية سابقاً، يعطي طول الشجرة ملاءمة أفضل، لكنه يأتي على حساب الإفراط في البيانات، وهذا يعني أن النموذج لم يعد قادراً على التصنيف تحقيق الهدف المقصود منه. لذلك، يجب أن يكون المحلل متوازناً وعليه تحديد الجانب الأيمن. إلاّ أنّ، هذا ليس علماً دقيقاً. وللقرّاء الراغبين بمزيد من المعلومات يمكن الاطلاع على كتاب غاريث وآخرون Gareth et al الذي يقدم تحليلاً عميقاً لهذه المسألة. أمّا بالنسبة للقرّاء المهتمين بالمفهوم، يوفر الرابط التالي تفسيرات عملية حول هذه المسألة: https://towardsdatascience.com/how-to-find-decision-tree-depth-via-cross-validation-2bf143f0f3d6.

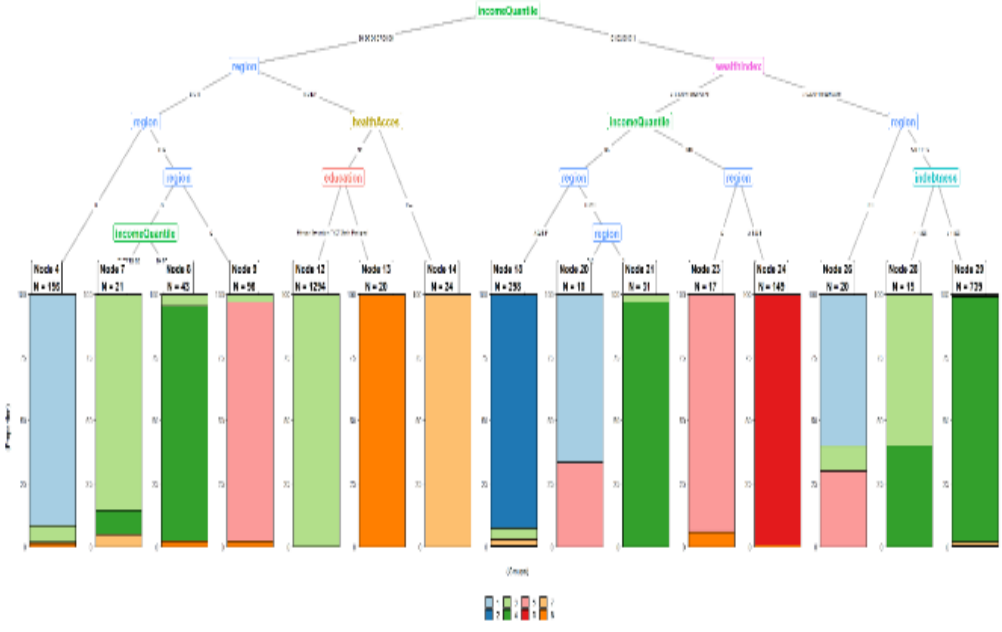
بالإضافة إلى التعامل مع جانبي النقاش، يستخدم الرمز (الخوارزمي) التحسين التلقائي (السطور 415-518) حيث تتولّى تقنية التحقق المتبادل فحص صحة الأشجار (decision tree) وفقاً لتركيبها المناسب أو المفرط. يوصى بالقيام ببعض التجارب اليدوية لأنها توفّر فهماً أفضل للمشكلة. وهكذا، فإن رمز لغة R المذكور أعلاه سيكون موجوداً في الدليل ويمكن الرجوع الى تفسيره من المرجع المذكور أعلاه.

```
# Estimation ------------------------------------------------------------
#In case optimal parameters want to be used, change the line 524 to control =bestModel$control
ct = rpart(Clusters ~ .,
           data    = dataParticipation,
           method  = 'class',
           control = rpart.control(minsplit=30, minbucket=15, cp=0.001))
plotcp(ct)
```

في المقابل، من السطر 521 فصاعداً، يتم إنشاء شجرة يدوية بواسطة دالة (rpart). يطلب الجزء الأول " ~ clusters"، من البرنامج اختيار أفضل طريقة لتحديد المجموعات على أساس جميع المتغيرات. ومع ذلك، إذا أردت القيام بعملية التجميع بالاستناد إلى متغيرات معينة عليك استبدال "." بقائمة متغيرات مفصولة بـ " + ". مثلاً، تستخدم دالة "Clusters ~ age + region" (المجموعات ~ العمر + المنطقة)، العمر والمنطقة فقط كمعيارين في تحديد أفراد المجموعة. يعرّف السطر التالي البيانات ثم يتبعه شرح للأسلوب المتبع. كما يوضّح الفصل الثامن، يمكن أيضاً استخدام CART للمتغيرات المستمرة، لذلك، وعند استخدام دالة "class" عليك تحديد استخدام متغيرات قاطعة. في خط التحكم، يمكنك تحديد عناصر، ذات صلة بهذا الدليل وهي minsplit وminbucket. يخبر الأول التعليمات البرمجية عن الحد الأدنى لعدد العناصر في المجموعة لكي تحدّد الشجرة مدى حاجتها للقطع الإضافي. ويحدد الثاني الحد الأدنى لحجم عقدة النهاية (أو ورقة) في الشجرة، لذلك تأكّد في هذه الحالة من أن لكل مجموعة خمس عشرة ملاحظة على الأقل.

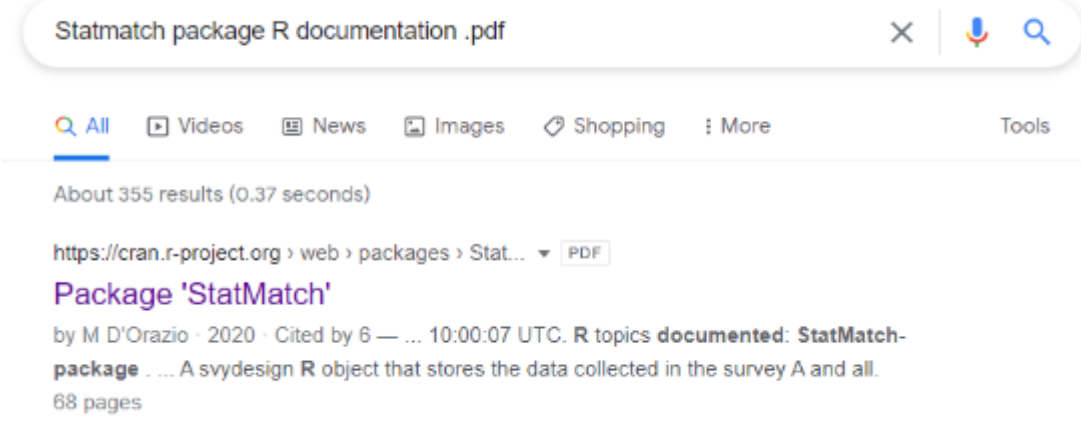
تمثّل الخطوط المتبقية العوامل الرسومية، وفي هذه الحالة سوف تبرز إمكانات ggplot. لاحظ أن الرسم البياني هو خليط من مجموعة خطوط شريطيّة، والعديد من التسميات الملونة المتضمنة على معلومات داخلية وبعض النصوص التي تتخلّل السهام. في حين أن التفاصيل غير ذات صلة بهذا الدليل، يتم تشجيع القارئ على تحليل الهيكل الكامل لقدرته على توسيع مهاراته عند العمل مع الرسوم البيانية في R.

كما وضحنا في القسم المفاهيمي تبدأ الشجرة بتقسيم الأفراد حسب المتغيرات ذات الصلة ومن ثم، تقوم بتحديد الخصائص الرئيسية التي تصف كل مجموعة. مثلاً، تمثّل المجموعة الثامنة (البرتقالية) أشخاصاً من الفئات العشرية الأعلى، معظمهم من المناطق ADEF، ولديهم درجات في الدراسات العليا ولكن ليس لديهم إمكانية الوصول إلى الخدمات الصحية. وعلى نقيض ذلك، تمثّل المجموعة السادسة (الحمراء) الأفراد ذوي الدخل المرتفع (أعلى عشر) من جميع المناطق باستثناء C. فمن ناحية، يسمح هذا بإنشاء مبادئ توجيهية سريعة للتوصيف، من ناحية أخرى تثير هذه الأداة أسئلة هامة تتعلق بالسياسات، مثلاً، لماذا تستبعد المنطقة C؟ ولذلك، يمكن أن نفهم الآن كيف لتقنيات بسيطة نسبياً، وبمجرد توحيدها، أن توفر رؤى سريعة عن نوع المستفيدين من برامج الحماية الاجتماعية، ومن خلال تطوير أدوات التوصيف الدقيقة، يمكن لصناع السياسات تحديد احتياجات هذه المجموعات ووضع استراتيجيات حازمة لتحسين سُبُل العيش.



8. ملاحظات ختامية

في هذا الفصل، كالفصول القادمة، تستخدم العديد من الحزم التي لم تُذكر من قبل. تم تحديد كافة هذه الحزم في الملف الرئيسي لذلك لا نحتاج لشرحها مجدّداً. للقرّاء الراغبين بالاطّلاع على المزيد من التفاصيل، يوصى بالتحقق من الحزم في الملف الرئيسي ووضعها في Google كما في المثال دالة "StatMatch".



مع استثناءات قليلة جداً، تأتي حزم R مرفقة بملفات ووثائق طويلة، ويتبع جميعها تنسيقاً مماثلاً لذلك مذكور أدناه، وتأتي مع تفسيرات مفاهيمية وتقنية، ومع أمثلة قد تساعدك في تحسين مهاراتك.

Package 'StatMatch'

July 16, 2020

Version 1.4.0

Title Statistical Matching or Data Fusion

Author Marcello D'Orazio

Maintainer Marcello D'Orazio <mdo.statmatch@gmail.com>

Depends R (>= 2.7.0), proxy, survey, lpSolve, ggplot2

Suggests Hmisc, MASS, mipfp, R.rsp, clue, RANN,

Description Integration of two data sources referred to the same target population which share a number of variables. Some functions can also be used to impute missing values in data sets through hot deck imputation methods. Methods to perform statistical matching when dealing with data from complex sample surveys are available too.

License GPL (>= 2)

VignetteBuilder R.rsp

NeedsCompilation no

Repository CRAN

Date/Publication 2020-07-16 10:00:07 UTC

الفصل السابع. خصائص الاستهداف

بمجرد الانتهاء من تحديد سمات المستفيدين، تتمثّل الخطوة التالية في إطار SPP-RAF في تقييم المعايير المستخدمة لاختيار الشخص الذي سيشارك في البرنامج وذاك الذي لن يشارك فيه.

1. الغرض من خصائص الاستهداف

تهدف المرحلة الثانية من الإطار إلى تحديد أهمية معايير اختيار المتغيرات المتصلة بالمجموعات المحدّدة سابقاً. وخلال هذه المرحلة، تستعرض الدرجات التي حصل عليها المستفيدون في كل معيار من معايير الاختيار لفهم قدرتها في التأثير على مجموع الدرجات وعملية اختيار هؤلاء.

2. المخرجات

(أ) مجموعة القياسات التي تصنّف المتغيرات المختارة وفقاً لأهميتها بين المستفيدين الحاليين وتأثيرها على درجة الأهلية/عدم الأهلية؛
(ب) خلق تباينات بين مقاييس عملية الغربلة ونتائج المتغيرات لتقييم المطابقة بين المستفيدين المستهدفين والمستفيدين الفعليين.

3. وثيقة ختامية

هي مجموعة من المؤشرات الرسومية والرقمية التي توجه صناع السياسات وتتيح تحديث منهجية الاستهداف وتحسينها باستمرار.

4. متطلبات البيانات

تتطلب هذه المرحلة إنشاء مجموعة بيانات تضم جميع الأفراد المستفيدين من برنامج معين خلال فترة محددة، وخصائصهم المشتركة. لكل أسرة، يُطلب تأمين المتغيرات التالية:

(أ) بالنسبة للمستفيدين، جميع المتغيرات المستخدمة كجزء من معايير الاختيار؛
(ب) بالنسبة للبرنامج الذي سيتم تقييمه، من الضروري معرفة المقاييس المستخدمة لتحديد درجات الأفراد.

5. استراتيجية الاستهداف

الموارد المتاحة لبرامج الحماية الاجتماعية محدودة، وفي حالات نادرة، يقل حجم الموارد عن عدد السكان المحتاجين إلى الدعم. ولهذا السبب، تحتاج الحكومات إلى وضع معايير لاختيار من سيستفيد أكثر من غيره من البرامج. ويمكن التمييز بين التصميم المفاهيمي للاستراتيجية وتطبيقها العملي. مثلاً، فكر في برنامج يهدف لتأمين هاتف محمول لكل الناس وفي منطقة يحوز كل شخص فيها على هاتف محمول. ثم، تنتفي الحاجة إلى

تحديد الأولويات. من ناحية أخرى، فكر في برنامج يعطي قيمة متساوية للأسر التي لديها أطفال، وتلك التي لديها أفراد يذهبون إلى المدرسة الابتدائية. وبما أن هذه المتغيرات ترتبط ارتباطاً وثيقاً، فقد يكون الموضوع المشترك (الأطفال في المدارس) مرجحاً أكثر من المتغيرات الأخرى التي قد تكون ذات صلة أيضاً باختيار المستفيدين. ولهذا السبب، يطوّر هذا الفصل تقنيتين متكاملتين لفهم مدى أهمية معايير القرار الحالية لاختيار المستفيدين.

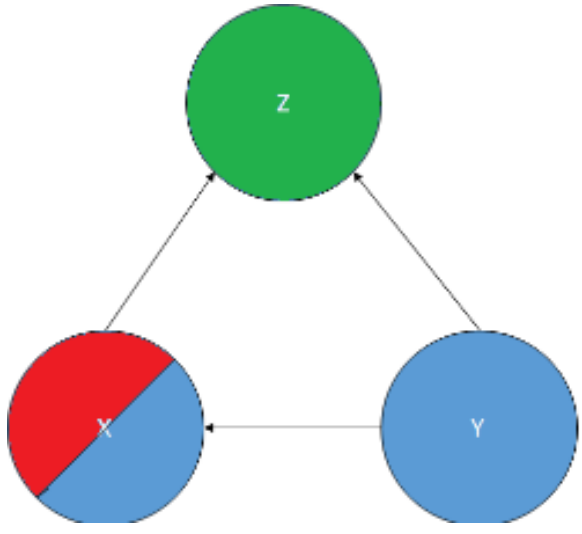
6. تقنية تحليل التباين

في حين أنّ صناع القرار السياسي قد حددوا تأثير كل من المتغيرات على معايير القرار، إلا أنه بمجرد أن تواجه النظرية البيانات، لن تجد ترتيباً محدداً. مثلاً، لنفترض سيناريو حيث يتم استخدام متغيرين، X وY لتطوير درجة تحدد مشاركة الفرد في البرنامج من عدمه. ولتجنب المضاعفات، الصيغة التالية هي:

$$score = 1 \times X + 1 \times Y$$

وبينما يبدو أن X وY متساويان في التمثيل، دعونا نأخذ بالاعتبار الفروق في المتغيرات. لنفترض أن للأفراد إجابات متشابهة جداً في X ومختلفة جدا في Y. وأن X وY مستقلان ويستخدمان المبادئ الإحصائية الأساسية، فإن تباين النتيجة (مدى انتشار قيم النقاط) هو 5. لذلك، تقوم Y بشرح 80 في المائة من الاختلاف في النتيجة بينما تشرح X 20 في المائة فقط. وعند وضع المعلومات في سياقها، ماذا لو كان الوصول لـ X صعباً جداً (ومكلفاً للحصول عليه). في هذه الحالة يطرح التساؤل هل الـ 20 في المائة من التباين تستحق العناء؟ $Var(X) = 1, Var(Y) = 4$.

في هذه الحالة، تستند التقنية المستخدمة في هذا الدليل إلى ارتباطات شبه جزئية. كما شرحنا بإيجاز في الشكل أدناه، يمكن اعتبار أن X وY قد قاما بشرح المتغير Z، ولكن جزءاً صغيراً من X قد تم شرحه من قبل Y ايضاً. لذا، لفهم أفضل لماهية الجزء الذي يفسر Z في X دون الخلط مع Y، علينا أولا حذف تأثير Y على X، ثم يمكن أن نرى كيف يؤثر ما تبقى من متغير في X على Z. ويشير بعض الناس أن هذه التقنية تحلل R^2، ولكن من الناحية الفنية، لا تؤثر بنية ارتباط المتغيرات على إضافة R^2 إلى مجموع الارتباطات شبه الجزئية. ومع ذلك، فإن R^2 مفيد (بعد تربيع الارتباط شبه الجزئي) وعادة ما تحتسب القيم استناداً إليه. لرؤية عرض مختلف لهذا الموضوع، فإن الرابط https://www.statisticshowto.com/partial-correlation. يقدم حالة توضيحية $R^2 R^2 R^2$ (Kim, 2015).



7. الارتباط شبه الجزئي

وختاماً، وكما سنرى في القسم الأخير، سيقدم البرنامج إرشادات بشأن الأهمية النسبية لكل متغير في عملية اختيار المستفيدين، ممّا سيوجه صانع السياسات بشأن أهمية هذه المتغيرات.

8. انحدار لاسو

ركز القسم السابق على قوة تفسير المتغيرات، ولكن يمكن إجراء دراسة مختلفة على معاملات المتغيرات.
لنفترض أن معادلة تحديد النتيجة هي:

$$score = 0.1 \times X + 10 \times Y$$

حيث يعتبر بصرياً أن وزن 0.1 هو أصغر بكثير من 10، ومع ذلك، هل هو صغير بما يكفي ليتم إهماله؟ لحل هذه المشكلة، يمكنك استخدام أسلوب إحصائي يعرف باسم انحدار لاسو. ولشرح هذا المفهوم يمكن افتراض انحدار نموذجي مع اثنين من المتغيرات وخطأ.(Tibshirani, 1996)

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \epsilon$$

توجب ممارسة تمرين الانحدار الخطي باحتساب القيم لتقليل مجموع الأخطاء التربيعية:$\beta_0, \beta_1, \beta_2$

$$\min_{\beta_0, \beta_1, \text{and } \beta_2} (Y - \beta_0 - \beta_1 X_1 - \beta_2 X_2)^2$$

الآن، افترض أن اختلاف قيم β عن 0 ستكون مكلفة. لذا تريد تضمين ذلك في عملية التقليل:β

$$\min_{\beta_0,\beta_1,\text{and }\beta_2} (Y - \beta_0 - \beta_1 X_1 - \beta_2 X_2)^2 + \lambda(|\beta_1| + |\beta_2|)$$

وبالتالي، إذا، λ=0 تصبح العملية نموذجية؛ في المقابل، إذا، λ=∞ أي أنّ β ستعادل 0 بسبب ارتفاع التكاليف المرتبطة بهما. وهكذا، من خلال الزيادة البطيئة لـ λ، تبدأ المعادلة بتحديد أهم المتغيرات، وتقوم بعض المعايير التقنية، بشرح أفضل λ في تفسير Y وباستخدام أقل عدد من المتغيرات. لذلك، لا يسمح هذا التمرين فقط بإزالة المتغيرات غير المفيدة، ولكن يعيد أيضاً ضبط الأوزان، وباستخدام متغيرات أقل يمكّن احتساب النقاط بشكلٍ صحيح. ولنتذكر مثال البرنامج الذي يعطي وزناً متساوياً للأسر التي لديها أطفال، وتلك التي لديها أفراد يذهبون إلى المدرسة الابتدائية. والنتيجة المحتملة باستخدام انحدار لاسو هي إزالة إحدى هذه المتغيرات وتكرار عامل العنصر الثاني لإظهار أن لهذا العنصر وزناً أعلى بكثير.

9. تحليل الاستهداف

يبدأ الرمز Chapter 02 كما في Chapter 00 وChapter 01 بتحميل البيانات والقواميس وموضوعات التنسيق ذات الصلة. لاحظ أنه على غرار التدريبات في Chapter 02، يوجد حاجة لإنشاء متغيرات وهمية ذات صلة بحيث يمكن إجراء العمليات. وتجدر الإشارة إلى وجود سطر مهم يجب التنبه إليه.

```
38 dataParticipation$score=dataParticipation$score+rnorm(n = length(dataParticipation$score), mean = 0, sd = 0.01)
```

على نقيض الانحدار الخطي حيث يوجد دائماً متغير خاطئ، في هذا التمرين، نظرياً، يجب أن تكون النتيجة احتساباً مباشراً للمتغيرات المستخدمة في الاستهداف. ولكن واقعياً، قد تقع أخطاء في الإبلاغ، خاصة إذا اعتمدت بعض الأسئلة على منظور المرشح المستفيد من البرنامج. وهكذا، يتم إضافة مصطلح "خطأ صغير" إلى النتيجة مما يعكس إمكانية وقوع بعض الأخطاء في البيانات.

وبمجرد الانتهاء من ذلك، يتم احتساب الارتباطات شبه الجزئية.

```r
# '2_1' Influence plot ------------------------------------------------
rowLine    = 1
jumpOfRows = 20
colPlots   = 1
colTables  = 12

indicatorCode <- '2_1'
part          <- ''
labCodes      <- dataPlots %>% subset(Code == paste0(indicatorCode, part))
newSheet      <- addWorksheet(wb, sheetName = indicatorCode)

# Data:
sp = spcor(dataParticipation, method = c('pearson')) # Semi-partial correlation.

# Weights of each variable on the R2:
weights           = sp$estimate^2
weights           = data.frame(Values = weights[dim(weights)[1], dim(weights)[2]])
weights$names     = dataParticipation %>% dplyr::select(-c('score')) %>% names() %>% .[1:nrow(weights)]
colnames(weights) = c('Values','Variable')
#Factors
weights$Variable = factor(weights$Variable,
                          levels = influence[,2],
                          labels = influence[,4-language])
R2 = summary(lm(score~., dataParticipation))$r.squared
```

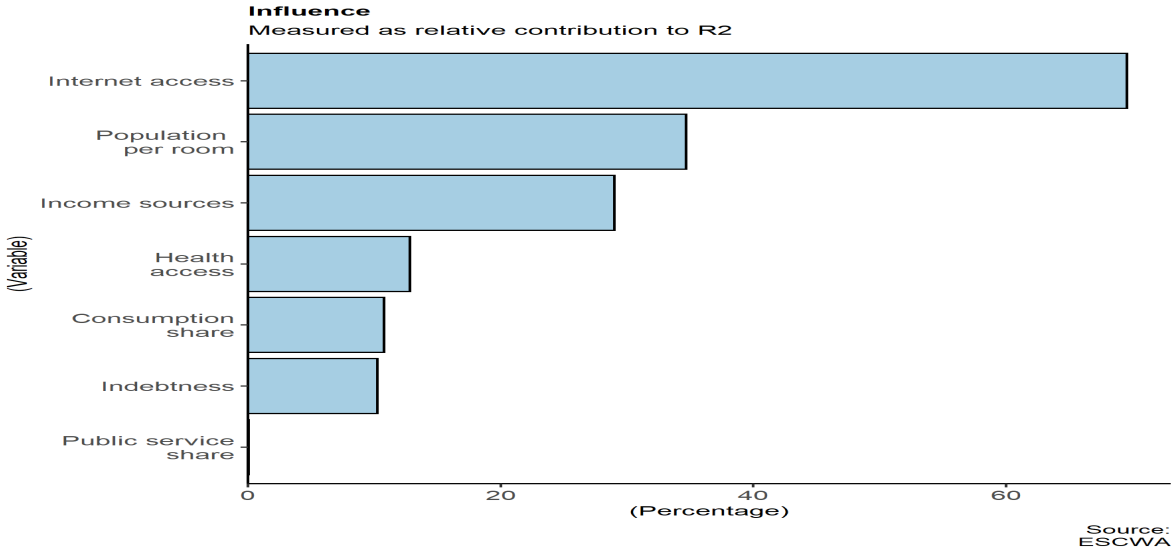
في هذا الجزء من التعليمات البرمجية، لا بد من تسليط الضوء على عدّة سطور ذات صلة:

(أ) يحسب السطر 54 الارتباطات شبه الجزئية بين كافة المتغيرات، ويحدّد السطران 58 و59 الخطوط ذات الصلة في المصفوفة؛

(ب) يحسب السطر 65 الانحدار الخطي، وسيحدد القارئ نمطاً مماثلاً للنمط المستخدم في (rpart) المعني بمخطط تسلسل القرارات للحصول على R^2.

```r
newPlot = weights %>%
  mutate(Values   = Values/R2,
         Variable = fct_reorder(Variable, Values)) %>%
  ggplot(aes(x = Variable, y = Values, fill = factor(1)))+
  geom_bar(colour = "black", stat = "identity", position = position_dodge(), show.legend = F) +
  labs(title    = paste0(as.character(labCodes[, 5*(1-language)+2])),
       subtitle = as.character(labCodes[, 5*(1-language)+3]),
       caption  = as.character(labCodes[, 5*(1-language)+6]),
       x        = as.character(labCodes[, 5*(1-language)+4]),
       y        = as.character(labCodes[, 5*(1-language)+5])) +
  scale_y_continuous(expand = expansion(mult = c(0,.05)),
                     labels = scales::percent_format(scale = 100, suffix = ''))+
  scale_fill_brewer(palette = as.character(labCodes[, 12])) +
  coord_flip() +
  scale_x_discrete(position = if (language == 0) ['top'] else ['bottom']) +
  theme_classic() +
  theme(legend.position = 'bottom',
        text             = element_text(family = 'Georgia'),
        axis.text.x      = element_text(size = scale_factor * 10, family = 'Georgia'),
        axis.text.y      = element_text(size = scale_factor * 10, family = 'Georgia'),
        axis.title.x     = element_text(hjust = 0.5, size = scale_factor * 10, family = 'Georgia'),
        axis.title.y     = element_text(hjust = 0.5, size = scale_factor * 10, family = 'Georgia'),
        legend.text      = element_text(size = scale_factor * 10, family = 'Georgia'),
        strip.text       = element_text(size = scale_factor * 10, family = 'Georgia'),
        plot.title       = element_text(face = 'bold', size = 10, family = 'Georgia', hjust = if (language == 0) {1} else {0}),
        plot.subtitle    = element_text(size = 10, family = 'Georgia', hjust = if (language == 0) {1} else {0}),
        legend.key.size = unit(0.5 * scale_factor, 'cm'))

# Table:
newData =  weights %>% mutate(Values = (Values/R2)*100) %>% arrange(Values)
```

وأخيراً الرسم هو بياني شريطي، لاحظ أنه في السطر 73 من الرسم البياني، و102 من الجدول، قد قام R^2 بتقسيم البيانات، بحيث يمكن أن يكون ذلك معياراً.



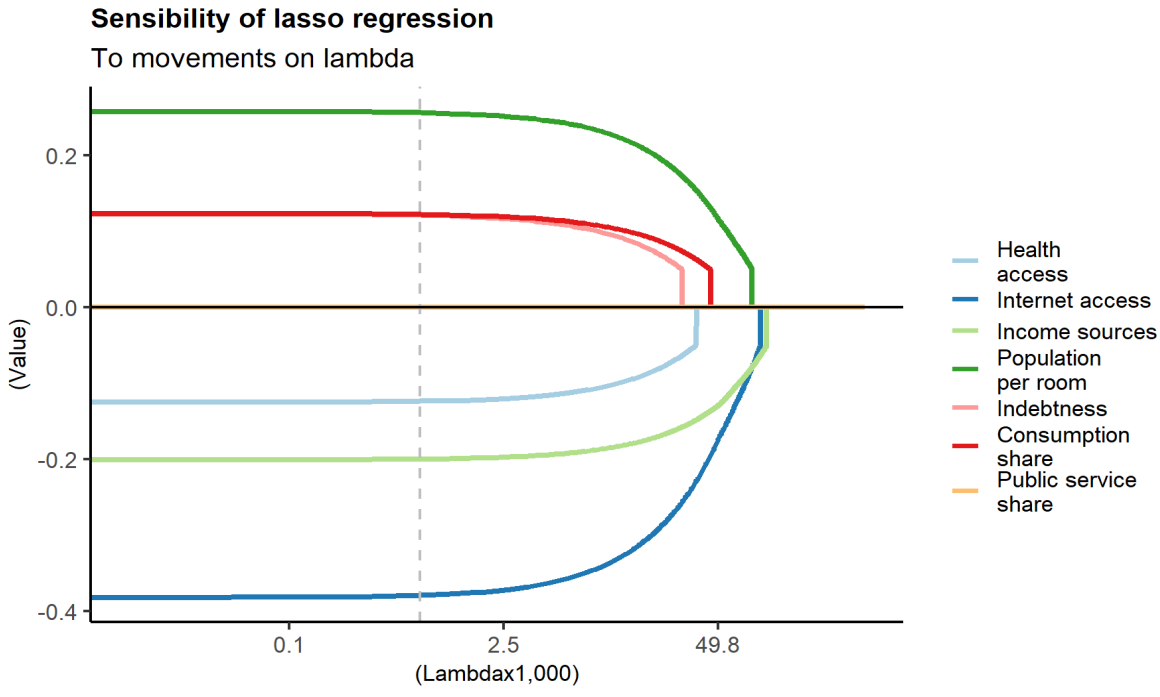
واستناداً إلى التحليل السابق، وباستثناء حصة الخدمة العامة، يمكن أن نرى أن جميع المتغيرات هي ذات صلة، حيث أن المتغيرات الأكثر صلة هي الوصول إلى الإنترنت، وعدد السكان لكل غرفة، ومصادر الدخل.

```
# Redundant variables ----------------------------------------------------
dataParticipationT=as.data.frame(scale(dataParticipation))
X <- model.matrix(score~.,dataParticipationT)
y <- dataParticipation$score


lb <- rep(-Inf,length(colnames(X)))
ub <- rep(Inf,length(colnames(X)))

# Estimation of the optimum lambda:
cv_las1 = cv.glmnet(x = X, y = y,
                    lower.limits = lb,
                    upper.limits = ub)
lambda = cv_las1$lambda.min

# Lasso regression with the calculated value:
las1 = glmnet(x = X, y = y,
              lower.limits = lb,
              upper.limits = ub,
              lambda       = lambda)

# Delete variables which don't contribute:
c.fit1           = coef(las1) %>% as.matrix() %>% as.data.frame()
colnames(c.fit1) = c('Coefficient')
namesVar         = rownames(c.fit1)
c.fit1$Coefficient[abs(c.fit1$Coefficient)<0.05]=0

RMSE=sum((y-predict(las1, newx = X))^2/(length(y)-(dim(X)[2]-1)))
result=as.data.frame(c(RMSE, c.fit1$Coefficient))
namesChosen=c("RMSE",rownames(c.fit1))
result$Variable=namesChosen
result$Lambda=lambda
colnames(result)[1]="Values"
result$Variable[2]="Intercept"
result=result[!result$Variable%in%result$Variable[3],]
```

إنّ الرمز المرتبط بانحدار لاسو معقد قليلاً. تشير السطور من 115 إلى 117 إلى النموذج والسطور، فيما يشير السطران 120-121 إلى تنظيم نطاق البحث عن β. حالياً، لا يوجد أي قيود، ولكن، مثلاً، من البديهي أن الأوزان ذات قيمة إيجابية، أي يمكن تعديل السطر 120 وفقاً لذلك. تقوم الكتلة التالية باحتساب القيمة الأمثل (السطور 123-127) وتستخدمها لتحديد المتغيرات التي يمكن إزالتها من المعادلة.

```
157 # Estimation ------------------------------------------------------------
158 # Lasso regression with the calculated for values around the optimum:
159 lMin=0.01*lambda
160 lMax=500*lambda
161
162 lambdaList=seq(from=lMin, to=lMax, length.out=1000)
163 for(i in lambdaList){
164 las1 = glmnet(x = X, y = y,
165               lower.limits = lb,
166               upper.limits = ub,
167               lambda       = i)
168
169 # Delete variables which don't contribute:
170 c.fit1           = coef(las1) %>% as.matrix() %>% as.data.frame()
171 colnames(c.fit1) = c('Coefficient')
172 namesVar         = rownames(c.fit1)
173 c.fit1$Coefficient[abs(c.fit1$Coefficient)<0.05]=0
174
175 RMSE=sum((y-predict(las1, newx = X))^2/(length(y)-(dim(X)[2]-1)))
176 resultT=as.data.frame(c(RMSE, c.fit1$Coefficient))
177 namesChosen=c("RMSE",rownames(c.fit1))
178 resultT$Variable=namesChosen
179 resultT$Lambda=i
180 colnames(resultT)[1]="Values"
181 resultT$Variable[2]="Intercept"
182 resultT=resultT[!resultT$Variable%in%resultT$Variable[3],]
183 result=rbind(result,resultT)
184 }
185
186 dataF=result
187
```

تقوم هذه الكتلة الثانية من الرموز باحتساب لاسو للمتغيرات الأخرى بشكل شبه مثالي، في السطرين 159 و160. ثم تستخدم التعليمات البرمجية a كحلقة حيث يتم تقييم s مع مراعاة منظور ديناميات المعاملات حول القيمة المثلى. أمّا بقية الرموز فهي الخطوات الرسومية المقترنة بالإجراء، ولا توفر أي عناصر جديدة للدليل.



النتيجة المثيرة للاهتمام هي خط التحليل. في هذا الرسم البياني أُلغيت حصة الخدمات العامة منذ البداية، وبزيادة القيود من قِبل عامل كبير فقط يمكن حذف متغيرات الدين، والوصول إلى الصحة، وتقاسم الاستهلاك، تُظهر هذه العملية أن الأشخاص الذين يستطيعون الوصول إلى الإنترنت، وإلى الخدمات الصحية، ومصادر الدخل يحوزون على درجات أقل (كما هو متوقع) في حين أن الأفراد ذوي الدرجات العالية هم اولئك الذين يعيشون في المنازل المكتظة، ويخصصون حصصاً عالية من نفقاتهم للاستهلاك، ولديهم مستويات عالية من الدين.

في هذا الإطار، لا بد من تسليط الضوء على مسألتين: تتعلق الأولى بدور الخدمات العامة، نظراً لأنها غير ذات صلة، لا بد من التمعّن في مدى تكلفة الحصول على هذا المتغير الأخير. والثانية هي التحقق مما إذا كانت هذه الأوزان تتماشى مع تصور البرنامج، ولهذا الغرض، يجب فهم الصيغة التي استُخدمت في تحديد المستفيدين.

لاحظ أن وزن المديونية هو 0.25 بينما حصة الاستهلاك هي 1.5. ومع ذلك، يؤشّر لاسو أن كلي الوزنين متماثلان نسبياً بحوالي 0.1 (كلاهما أقل من المتوقّع، ومع ذلك، فإن نسبة التخفيض بالنسبة للاستهلاك هي أكثر دراماتيكية). وثم، يمكن لصانعي السياسات، من خلال إجراء هذه الأنواع التحليلية، فهم الأوزان التي يجري تطبيقها للفحص على نحو أفضل وتعديل البرنامج وفقاً لذلك.

صيغة المستفيدين:

وفقاً لمملكة X. الصيغة لاختيار المستفيدين هي:

$$score = \frac{popRoom}{7} + \frac{6 - incomeSource}{5} + \frac{3 \times consumptionShare}{2} + \frac{1 \times publicServiceShare}{2} + \frac{indebtness}{4} + (1 - healthAcces) + (1 - internetAcces)$$

ويشارك في البرنامج الأفراد الذين تزيد درجاتهم عن 3.5 في حين لا يشارك فيها الأفراد الذين حصلوا على درجات أقل.

الفصل الثامن. تقييم التغطية

يتناول الفصل الأخير من الدليل الخطوة الثالثة من إطار SPP-RAF، كما ذُكر سابقاً، فإنّ للخطوة الرابعة بيانات ومتطلبات تقنية أعلى، وبالتالي فهي تجعل الطريقة اليدوية الحالية واسعة النطاق. ومع ذلك، يغطي القسم القادم، متطلبات هاتين المرحلتين بحيث يمكن للقارئ تكوين فكرة أوضح وأكثر شمولية حول الإطار.

1. الغرض من تقييم التغطية

تهدف المرحلة الثالثة من الإطار إلى تحديد أخطاء التضمين والاستبعاد التي تقع فيها البرامج. فمن ناحية، تستعرض هذه المرحلة التحليلية خصائص المستفيدين الحاليين لتحديد أولئك الذين قد تغيّرت حالتهم وربما استبعدوا من البرنامج أو نقلوا إلى برنامج مختلف. ومن ناحية أخرى، تحدد الأفراد الذين يُحتمل استيفاؤهم معايير المشاركة في البرامج (أو من المرجح أن يستوفوها قريبا) لكنهم غير مسجلين حالياً. ومن خلال ذلك، يتيح الإطار لصناع السياسات اتخاذ قرارات تزيد من كفاءة استخدام الموارد إلى حدّها الأقصى، ويحدد مدى الحاجة إلى نشر خطط التوعية لجعل جهود التسجيل أكثر فعالية.

2. المخرجات

الكشف عن المؤشرات التي قد تساعد على تحديد المستفيدين المحتملين الذين حذفهم البرنامج في منطقة جغرافية محددة، وكذلك احتمال انتمائهم إلى مجموعة المستفيدين أو غير المستفيدين، بخلاف تلك التي عيّنوا فيها.

3. وثيقة نهائية

(أ) الحدّ من أخطاء التضمين والاستبعاد؛
(ب) تحديد المناطق الجغرافية التي قد تُعطى الأولوية فيها لتنفيذ البرامج للحدّ من أخطاء الاستبعاد.

4. متطلبات البيانات

وتتطلب هذه المرحلة إنشاء مجموعة بيانات تحتوي على مجموعة المتغيرات نفسها التي تضمّنتها المرحلة الأولى، مع توسيع نطاقها لتشمل الأفراد الذين لا يشاركون حالياً في هذا البرنامج.

من المتوقع أن تكون بيانات غير المستفيدين محدودة جداً، ولمواجهة هذا التحدي، سيتم تطوير تقنيات مرنة للتعلم الآلي لتذليل الصعوبات المطروحة بسبب المتغيرات المفقودة. ومع ذلك، يعتمد نجاح التحليل على قدرة الحصول على أكبر مجموعة ممكنة من المتغيرات للمستفيدين وغير المستفيدين.

5. الغرض من تقييم المستفيد

وتهدف المرحلة الرابعة من الإطار إلى قياس أثر برنامج الحماية الاجتماعية على المستفيدين منه وتحديد الخصائص التي تمكن الأفراد من التغلب على ظروفهم الصعبة والخروج بنجاح من البرنامج. ولذلك، من المهم فهم قدرة البرامج على استخلاص النتائج المرجوة على صعيد سُبُل عيش المستفيدين، ومدى توفّر خصائص فردية تدعم نجاح الأفراد في حياتهم الشخصية (وينبغي تشجيعها)، وخصائص تتعلق بحوافز ذات أثر ضار غير مرغوب فيها تتطلب مناقشات بشأن تصميم البرامج وتطبيقها.

6. المخرجات

(أ) تحديد مدى تطور مؤشرات الحماية الاجتماعية مع الوقت وعبر مختلف مجالات السياسة العامة؛
(ب) قياس الحوافز ذات الأثر الضارّ المحتمل والتغيّرات السلوكية بين المستفيدين التي تؤثّر على احتمال تخرجهم من البرنامج.

7. وثيقة ختامية

(أ) إنشاء آلية للتقييم السريع لقياس أثر تدابير الحماية الاجتماعية؛
(ب) تحديد الخصائص والإجراءات التي تحسّن سُبُل عيش الفرد إلى درجة تغنيه عن المشاركة في برامج المساعدة.

8. متطلبات البيانات

(أ) وتتطلب هذه المرحلة إنشاء مجموعة بيانات تحتوي على مجموعة المتغيرات المذكورة نفسها في المرحلة الأولى، ولكنها تمتد عبر فترات مختلفة. لذلك، ينبغي تتبع نفس الشخص بتواتر معين (شهرياً مثلاً) لتقييم التغيّرات في الخصائص البرنامجية الاستراتيجية (وخاصة متغيرات الاستهداف، أو أكبر عدد ممكن من المتغيرات)؛
(ب) وينبغي أن تتجاوز متابعة الأشخاص فترة مشاركتهم في البرنامج، ورصد المتغيرات لعدة فترات بعد توقفهم عن المشاركة في البرامج.

9. تحدي البيانات المفقودة

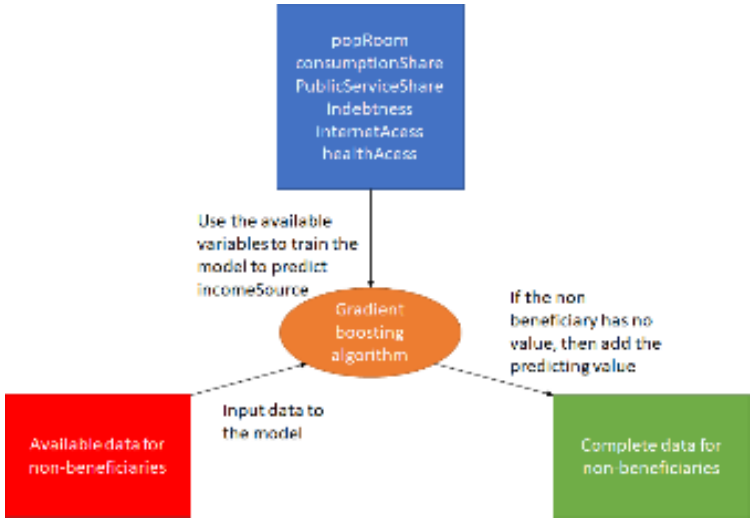
بالنسبة للفصل الثالث من التقرير، تبرز الحاجة لمقارنة بيانات المستفيدين مع تلك الواردة من غير المستفيدين. لاحظ أنه إذا جمعت كافة البيانات ذات الصلة باختيار المستفيدين من كل شخص في البلاد، فلا ينبغي وقوع أي سوء في توزيع الأفراد (باستثناء القيم الخاطئة التي تضلّل عملية التوزيع، وهذا ما سيغطيه القسم التالي). ومع ذلك، فإنّ الحال يختلف، لأنّ جمع البيانات مكلف ويتوقع أن يكون لدى المستفيدين فقط جميع المتغيرات ذات الصلة (يتوقّع أن يكون لديهم جميع المتغيرات، وإلا كيف سيحتسب النظام درجاتهم؟) وهنا تبدأ العديد من الحقائق المؤسفة بالظهور. مثلاً، ماذا لو لم تتمكن فرق البرنامج الاجتماعي من الوصول إلى مجموعة من

السكان الذين يصعب الوصول إليهم؟ في هذه الحالة، فإن هؤلاء الذين يحتاجون فعلاً إلى البرنامج لن يعرفوا فيه ولن يسجلوا أنفسهم؛ ولن تعرف الحكومة بالأمر باعتبار ان المتغيرات ذات الصلة غير مدرجة في النظام. وماذا لو كان الفرد، بسبب الفقر، مشغولاً جداً بالعمل وغير قادر على الذهاب إلى مكان التسجيل؟ أو ماذا لو أثّر البعض سلباً كما يتعرض له بعض أفراد المجتمع من تمييز إذا ما شاركوا في البرامج الحكومية؟ لهذه الأسباب، من المهم وضع تقنية ترشدنا إلى درجة الأشخاص الذين لا تتم تغطيتهم. والطريقة التقليدية لذلك هي من خلال إجراء الدراسات الاستقصائية لاستجواب الأفراد في مناطق محدّدة عن جميع المتغيرات ذات الصلة وتقييم عدد الأشخاص الذين تمت تغطيتهم من بين الذين يستوفون المعايير. للأسف فإنّ الدراسات الاستقصائية مكلفة وتستغرق وقتاً طويلاً، محلية (إذا تمّت في البلدية أ، ولكن البلدية ب هي التي تعاني من المشكلة، لا أحد سيعرف بذلك). حالياً، من الصعب جداً أن يكون الفرد خارج النظام تماماً (أمّا من هم كذلك، فستعرف الحكومة بذلك في نهاية المطاف). لذلك يوجد بعض السجلات الإدارية المتاحة للجميع، ثم، ستستخدم المنهجية المعروضة في هذا الفصل هذه البيانات لتقدير القيم المفقودة واحتساب الدرجة المحتملة للفرد. وهذا لا يُعتبر قياساً دقيقاً لإعلان أن شخصاً ما قد أسيء تصنيفه، ولكنه يعطي فكرة محتملة عن المناطق الجغرافية والقطاعات الاجتماعية والاقتصادية، ذات التمثيل الناقص ـ وبالتالي تصميم استراتيجيات توعية موجهة.

ويمكن أن تكون العلاقة بين المتغيرات معقدة للغاية، بالتالي فإن وجود نهج موحد، كالانحدار الخطي لن يكون كافٍ. من الناحية الأخرى، إذا كان عدد المتغيرات ذات الصلة المستخدمة في اختيار المستفيدين مرتفعاً، فإن وضع نموذج مصمم خصيصاً لكل حالة هو مسعى مرهق وممل، يحتاج إلى تحديث منتظم، وسيصعب ضمان استخدام النموذج الصحيح. لحسن الحظ، كانت خوارزميات التعلُّم الآلي متقدمة، وعلى غرار المثال الذي طوّرناه في الفصل 6 مع CART، يمكن لخوارزميات أخرى أن تعطي بشكل مستقل، أفضل مؤشر. غير أنّ هذه الطرق لا تستخدم عادة لتحليل البيانات لأن تتبع الروابط بين المتغيرات ليس بسيطاً كالحال في الانحدار أو في CART، ومع ذلك، لأغراض التنبؤ، فهي تؤدي دورها بشكل جيد للغاية. في هذه الحالة، يقترح الدليل خوارزمية تسمى التعزيز المتدرج Gradient Boosting، التي تقوم بتوسيع نطاق مخطط تسلسل القرارات وتضع العديد من المخططات بشكل متوازي، وتقارن نتائجها، وتعدّ شجرة أكبر استناداً إلى مجموعة النتائج التي تمّ التوصل إليها. ولكن لن نتوسّع في شرح الخوارزمية في الدليل باعتبار أنها لن تضفي قيمة مضافة تخدم أهدافه. بالنسبة للقراء الراغبين بالاطلاع على المزيد من المعلومات، يمكن العثور على مقدمة موصى بها عبر الرابط https://machinelearningmastery.com/gentle-introduction-gradient-boosting-algorithm-machine-learning. ويكمن السبب في شرح مخطط تسلسل القرارات هو أن عرضه يتيح لنا فهم طرق أفضل لتصميم المبادئ التوجيهية. بالمقابل، بالنسبة لهذا القسم، فإن القيمة الوحيدة للخوارزمية هي معرفة عدم حاجتها إلى خبير لإعطاء تنبؤ جيد استناداً إلى متغيّر، وهذا ما يسمح لنا بأتمتة العملية بسرعة نسبية. وإحدى مزايا هذه الطريقة، التي لوحظت أيضاً بالنسبة لمخطط تسلسل القرارات، هي أنها ستظل في حالة المعلومات المحدودة تنتج تخميناً مستنيراً لقيمة متغير، مما يسمح لنا بتنفيذ هذه العمليات حتى في السيناريوهات ذات الموارد المحدودة.(Hastie, Tibshirani, & Friedman, 2009).

وبعد تقديم هذه الملاحظات، أصبحت العملية في هذه المرحلة واضحة وصائبة. ولنتأمل مثلاً في حالة مملكة X. حيث تطبق فيها سبعة معايير للمستفيدين، وبعضها، مثلاً كمصادر الدخل، متاح فقط لهم. وبالتالي، تكمن الخطوة الأولى في تطوير خوارزمية التعزيز التدرجي للتنبؤ بمصادر الدخل على أساس الأفراد الذين يعيشون

في غرفة واحدة، وحصة الاستهلاك، والمتغيرات الأربعة الأخرى. ثم نستخدم هذا النموذج للتنبؤ بقيم مصادر الدخل للمستفيدين ثم بتغيير القيمة المفقودة أو استبدالها، كما هو موضح في الرسم التخطيطي التالي.



10. عملية التنبؤ بالأرقام للبيانات المفقودة

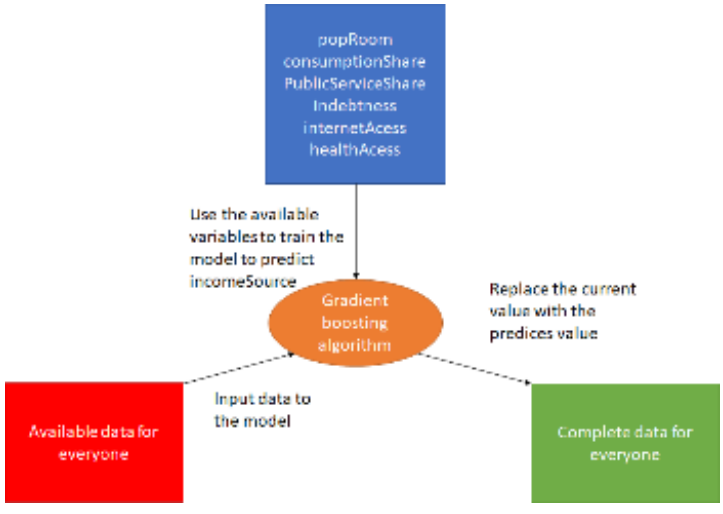
وبعد تنفيذ العملية السابقة، سيحصل جميع الأفراد، المستفيدين أو غير المستفيدين، على درجة، وباستخدام المعايير التي تحدد التأهل، يمكن تحديد الأفراد القادرين على أن يكونوا في البرنامج دون أن يكونوا مشاركين فيه. وبالتالي، سيكون لدى صانع السياسات، عبر تقاطع هذا المتغير الجديد بالمتغيرات الأخرى، بعض التوجيه بشأن المجالات التي يمكن أن تحدث فيها أخطاء استبعادية.

11. تحدي القيمة الخارجة

يرتبط التحليل النهائي بأخطاء التضمين. لسوء الحظ، عندما ترتبط المعلومات التي يقدمها الفرد بفرصة الحصول على بعض الفوائد، ستبرز حوافز ذات تأثير ضار تجعل الأفراد قادرين على تقديم معلومات خاطئة. في حين يمكن حلّ التحدي السابق (البيانات المفقودة) بدقة كبيرة من خلال إجراء الدراسات الاستقصائية، لا يمكن حل المشكلة الثانية مباشرةً لأنّ الفرد قد ربط الاستطلاع بإمكانية فقدان الفوائد، أو كسب المزيد من المال، وقد يعاني المسح أيضاً من التحيز. ومع ذلك، يمكن إعادة هيكلة الفكرة السابقة لاستكمال البيانات وتحديد الأفراد ذوي نمط الإجابات الشائعة ما يتطلّب دراسات أكثر تفصيلاً. كما يمكن أن تساعد هذه الفكرة في تحديد المتغيرات ذات القيم المتطرفة الأكثر شيوعاً.

ولنتأمل مثلاً في حالة مملكة X. وبناءً على الخطوة السابقة يوجد فعلاً نموذج للتنبؤ بمصادر الدخل على أساس جميع المتغيرات الأخرى. وعوضاً عن التنبؤ بهذه القيمة لأولئك الذين لم يقدموا هذه المعلومة، يمكن استخدام النموذج للتنبؤ بهذه القيمة للجميع. الآن، بالنسبة للمستفيدين، لدينا البيانات التي قدموها وتلك التي كان من المتوقع تقديمها. وباستخدام هذه المعلومات، يمكننا تحديد المستفيدين الذين لم يتوقّع انتسابهم إلى البرنامج،

وإيلاء المزيد من الاهتمام والبحث في هكذا حالات. كما سنقوم بتحليل مدى اختلاف التنبؤ النموذجي مع القيمة الحقيقية للبيانات لكل متغير وفهم المتغيرات ذات أنماط التنبؤ الواضحة ويمكن أن تسهل التفتيش، وتلك التي لديها إمكانية تنبؤ أقل وينبغي مراجعتها بعناية.



12. تحليل التغطية

التعليمات البرمجية ذات الصلة بهذا القسم هي تحت اسم chapter 03. كالحالات السابقة، تقوم السطور الـ 55 الأولى بتحميل البيانات والقواميس.

```
ppopRoom <- gbm(formula = popRoom ~ .,
                data = dataMini,
                n.trees = 500,
                cv.folds = 5,
                n.cores = 1)

ntree_opt_oob <- gbm.perf(ppopRoom, method = "OOB", oobag.curve = TRUE)     # n.trees I
ntree_opt_cv <- gbm.perf(ppopRoom, method = "cv")   # n.trees by cross-validation.

ppopRoom <- gbm(formula = popRoom ~ .,
                data = dataMini ,
                n.trees = ntree_opt_oob,
                cv.folds = ntree_opt_cv,
                n.cores = 1)
pincomeSources <- rpart(formula = factor(incomeSources) ~ .,
                        data = dataMini,
                        method="class")
pconsumptionShare <- gbm(formula = consumptionShare ~ .,
                         data = dataMini ,
                         n.trees = ntree_opt_oob,
                         cv.folds = ntree_opt_cv,
                         n.cores = 1)
ppublicServiceShare <- gbm(formula = publicServiceShare ~ .,
                           data = dataMini,
                           n.trees = ntree_opt_oob,
                           cv.folds = ntree_opt_cv,
                           n.cores = 1)
pindebtness <-  gbm(formula = indebtness ~ .,
                    data = dataMini%>%filter(indebtness>0)%>%
                      mutate(indebtness=log(indebtness)),
                    n.trees = ntree_opt_oob,
                    cv.folds = ntree_opt_cv,
                    n.cores = 1)
pinternetAcces <- rpart(formula = factor(internetAcces_1) ~ .,
                        data = dataMini ,
                        method="class")
phealthAcces <- rpart(formula = factor(healthAcces_1) ~ .,
                      data = dataMini ,
                      method="class")
```

ثم، من السطور 56 إلى 94، يتم تشغيل متغيرات خوارزمية التعزيز المتدرج المحدد بواسطة الدالة (gbm) مقابل كافة المتغيرات الأخرى ("."~") لإنشاء نماذج التنبؤ هذه. ويوفر السطران 62 و63 بعض البارامترات المثلى للتنبؤ وفقاً لبرنامج R. على غرار حالة الأشجار، إذا كان لدى المستخدم عوامل (Parameters) مفضلة، فيمكن وضعه يدوياً. في كل نموذج تنبؤ، هناك عوامل (Parameters) أخرى مرتبطة بعدد الأشجار التي يتوقع اعدادها. على وجه الخصوص، باعتبار أن مصادر الدخل، والوصول إلى الصحة، والإنترنت تندرج ضمن مسمى الفئات، يتم العمل لكي تعرض على الشجرة وفقاً لذلك. وأخيراً، يقدم متغير المديونية مثالاً على تحوّل المتغيّر. أحياناً، يعمل النموذج المتوقع بشكل أفضل عندما يكون المتغير في اللوغاريتم أو الأشكال الأساسية (لا يغطي الدليل هذا الموضوع للقرّاء المهتمين به، الرابط التالي هو نقطة انطلاق جيدة

http://sciences.usca.edu/biology/zelmer/305/trans/#:~:text=For%20regression%2C%20it%20is%20the,and%20meet%20the%20linearity%20assumption.&text=If%20transformation%20succeeds%20in%20producing,the%20dependent%20variable%20(Y).

كما يمكن الملاحظة في السطرين 84 و85 يمكن تحويل البيانات في هذه المرحلة بحيث قد يتضمن نموذج التنبؤ ذلك.

```
96  # Prediction of the missing values:
97  temp$participant='No'
98  temp$participant[!is.na(temp$popRoom)]='Yes'
99  temp$participant=factor(temp$participant)
100 temp2=temp
101 temp$popRoom[is.na(temp2$popRoom)]  = predict(ppopRoom, newdata = temp2%>%filter(participant=='No'), n.trees = ntree_opt_oob)
102 temp$incomeSources[is.na(temp2$incomeSources)] =
103   as.numeric(as.character(predict(pincomeSources, newdata = temp2%>%filter(participant=='No'), "class")))
```

تعمل السطور التالية على إنشاء متغير من شأنه تسجيل كل الأفراد المشاركين وغير المشاركين والأفراد الذين يعانون من نقص في بياناتهم عبر popRoom و incomeSources (لغير المستفيدين) واستخدام النماذج للتنبؤ بها. لهذه العملية، فإنّ الدالة الرئيسية هي (predict) التي تأخذ النموذج والبيانات كمدخلات، لاستخدامها في التنبؤ بالقيمة التي يجب وضعها في الملاحظة.

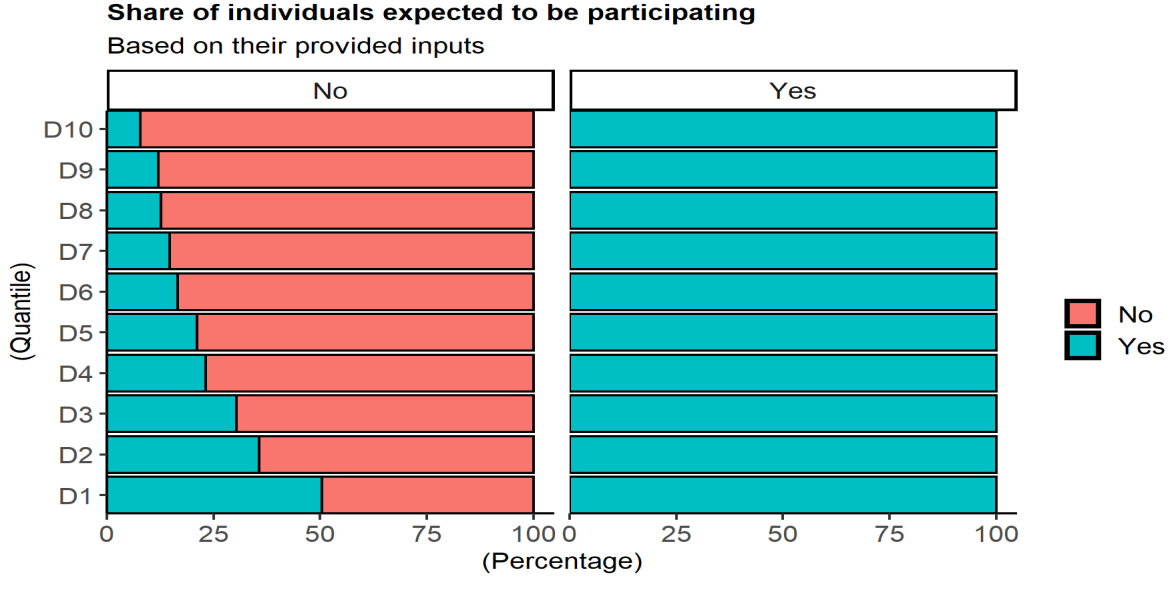
```
105 # 3.1| Calculation of score
106
107 temp=temp%>%mutate(adjpopRoom=popRoom/7)%>%
108   mutate(adjincomeSources=(6-incomeSources)/5)%>%
109   mutate(adjconsumptionShare=consumptionShare*1.5)%>%
110   mutate(adjpublicServiceShare=publicServiceShare*0.5)%>%
111   mutate(adjindebtness=indebtness/4)%>%
112   mutate(adjinternetAcces=1-internetAcces_1)%>%
113   mutate(adjhealthAcces=1-healthAcces_1)%>%
114   mutate(score=adjpopRoom+adjincomeSources+adjconsumptionShare+adjpublicServiceShare+
115            adjindebtness+adjinternetAcces+adjhealthAcces)%>%
116   dplyr::select(-c('adjpopRoom', 'adjincomeSources', 'adjconsumptionShare',
117                    'adjpublicServiceShare', 'adjindebtness', 'adjinternetAcces','adjhealthAcces'))%>%
118   mutate(Expected_Participant=case_when(score>3.5~ 1,
119                                         score<=3.5~ 0))%>%
120   mutate(Expected_Participant=factor(Expected_Participant,levels=c(0,1),labels=confirmation[,4-language]))
```

بمجرد الانتهاء من مجموعة البيانات، يستخدم الجزء التالي التعليمات الواردة لاحتساب درجة جميع الأفراد وتحديد أولئك الذين يُتوقّع أن يكونوا ضمن المشاركين. هذه السطور هي مجرد متابعة لتلك التعليمات وتقوم باحتساب المتغير الجديد Expected_Participant.

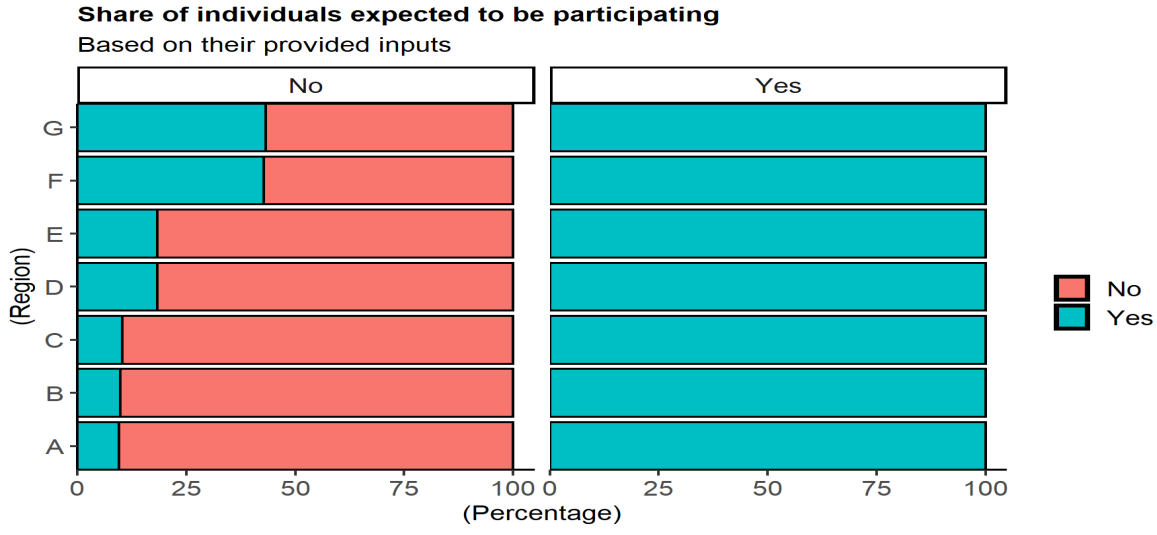
وتتناول السطور التالية الجداول والرسوم البيانية للقيم والتعليمات البرمجية التي غطتها الفصول السابقة، ويعتبر الجزء ذو الصلة هو ذاك الذي قام بتفسيرها.

المشارك المتوقع	المشاركة	الوتيرة
لا	لا	28827
نعم	لا	8232
لا	نعم	0
نعم	نعم	2941

في هذه الحالة، ونظراً لعدم المساس بقيم المشاركين، لا يزال كل من يشارك من المتوقع أن يشارك (الأرقام هي نفسها). ومع ذلك، 22 في المائة من بين الأفراد الذين لا يشاركون كان يتوقّع مشاركتهم لكنّهم لم يشاركوا فعلياً.



يقدم هذا الرسم البياني النتائج السابقة وفقاً لعشريات الدخل ويذكر أنّ الفئات العشر الأكثر فقراً، قد تتضمن أخطاء استبعاد أعلى من الأخطاء الخاصة بالفئات العشر الأكثر ثراء.



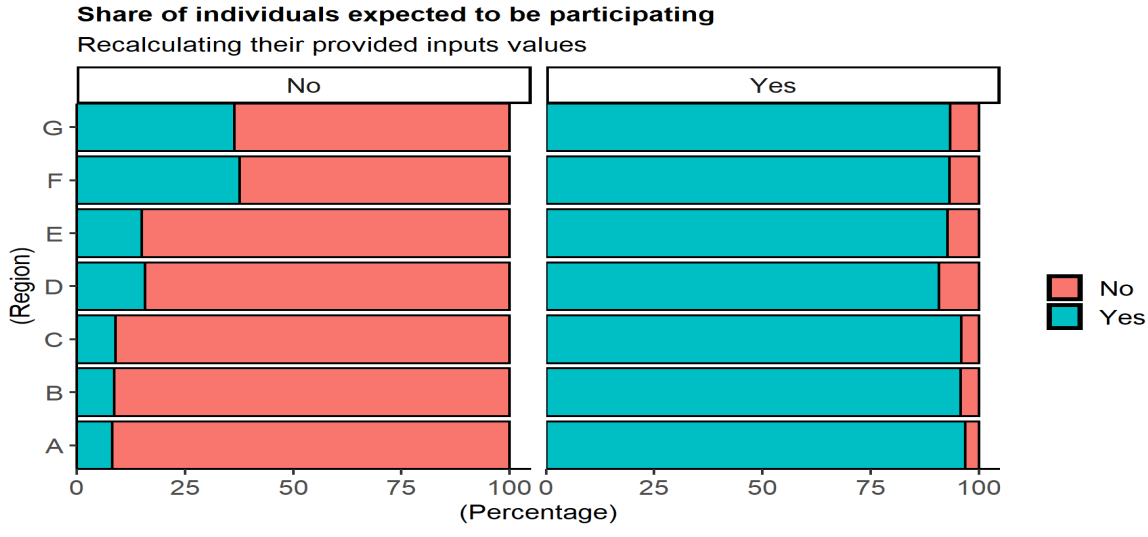
ويوفّر تكرار العمل حسب المنطقة نتيجة واضحة تبرز وقوع معظم أخطاء الاستبعاد في أفقر المناطق. ومع التذكير بأن هذه المناطق عانت من الصراع كان من الصعب على الحكومة الوصول إليها، ممّا يجعل هذه النتيجة واقعية. وهكذا، يمكن لصناع السياسات البدء بالتفكير في سياسات نشر تساعد على استهداف المناطق الصعبة حيث يحتاج الناس إلى الدعم ويكافحون من أجل الحصول عليه.

يكمن الجزء الرئيسي التالي من التعليمات البرمجية في السطور من 277 إلى 308 حيث تبدأ بتحديد الخطوط العريضة.

```
277 # 4) Outlier identification
278
279 temp$popDens          = temp1$popDens
280 temp$incomeSources0    = temp1$incomeSources
281 temp$consumptionShare0 = temp1$consumptionShare
282 temp$publicServicesShare0 = temp1$publicServicesShare
283 temp$indebtness0       = temp1$indebtness
284 temp$internetAcces_10   = temp1$internetAcces_1
285 temp$healthAcces_10     = temp1$healthAcces_1
286 temp1$yhat2
287 temp1$yhat2[temp2$participant  "Yes"]                predict(popRanm, newdata   temp1%>%filter(participant  "Yes"), n.trees   ntree_opt_ob)
288 temp1$incomeSources[temp2$participant=="Yes"]       = as.numeric(as.character(predict(pincomeSources      , newdata = temp1%>%filter(participant=="Yes"), "class")))
289 temp1$consumptionShare[temp2$participant=="Yes"]    = predict(pconsumptionShare , newdata = temp1%>%filter(participant=="Yes"), n.trees = ntree_opt_ob)
290 temp1$publicServicesShare[temp2$participant  "Yes"]  predict(ppublicServicesShare, newdata   temp1%>%filter(participant  "Yes"), n.trees   ntree_opt_ob)
291 temp1$indebtness[temp2$participant=="yes"]        = exp(predict(pindebtness, newdata = temp1%>%filter(participant=="Yes"), n.trees = ntree_opt_ob))
292 temp1$internetAcces_1[temp2$participant=="Yes"]   = as.numeric(as.character(predict(pinternetAcces      , newdata = temp1%>%filter(participant=="Yes"), "class")))
293 temp1$healthAcces_1[temp2$participant  "Yes"]     as.numeric(as.character(predict(phealthAcces      , newdata   temp1%>%filter(participant  "Yes"), "class")))
294
295 temp=temp%>%mutate(adjpopRanm=popRanm/7)%>%
296   mutate(adjincomeSources  (6 incomeSources)/5)%>%
297   mutate(adjconsumptionShare=consumptionShare*1.5)%>%
298   mutate(adjpublicServicesShare=publicServicesShare+0.5)%>%
299   mutate(adjindebtness indebtness/4)%>%
300   mutate(adjinternetAcces=2-internetAcces_1)%>%
301   mutate(adjhealthAcces=2-healthAcces_1)%>%
302   mutate(score adjpopRanm+adjincomeSources+adjconsumptionShare+
303            adjpublicServicesShare adjindebtness+adjinternetAcces+adjhealthAcces)%>%
304   dplyr::select(-c('adjpopRanm', 'adjincomeSources', 'adjconsumptionShare',
305                    'adjpublicServicesShare', 'adjindebtness', 'adjinternetAcces', 'adjhealthAcces'))%>%
306   mutate(Expected Participant(case_when(score>8.5~ 1,
307                                         score<3.5~ 0))%>%
308   mutate(Expected Participant factor(Expected Participant,levels c(0,1),labels c("Not Participating","Participating")))
```

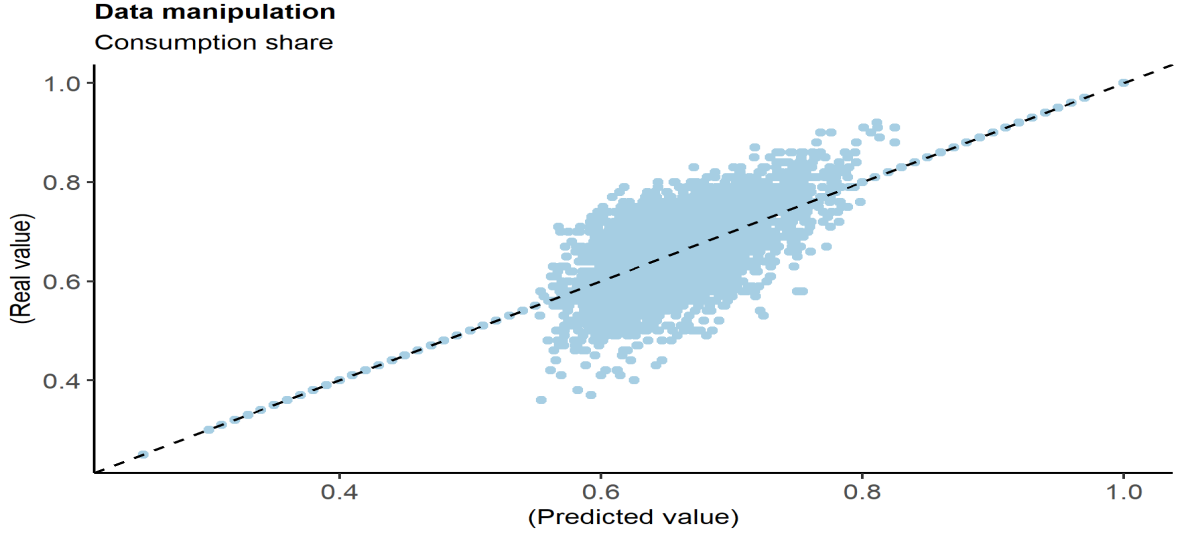
في هذه السطور، تكمن الخطوة الأولى في إنشاء بعض المتغيرات الزمنية المقترنة بقيم البيانات الأصلية. على أن تستخدم لكي تعكس مدى قوة النموذج في التنبؤ. ثم يتم التنبؤ بجميع المتغيرات (السطور 287-293) والتي تستخدم أيضاً للتنبؤ بالمشاركة المتوقعة (السطور 295-38). بعدها نقع على السطور المخصصة للرموز أو التعليمات البرمجية الاضافية التي تتولّى تتبّع النتائج. لاحظ أن التحديث ينحصر فقط بالمستفيدين. وذلك لأنه

من غير المتوقع أن يقدّم الأشخاص غير المستفيدين معلومات غير صحيحة لتجنب المشاركة في البرنامج. أحياناً لا يبلِّغ غير المستفيدين عن أي شيء.

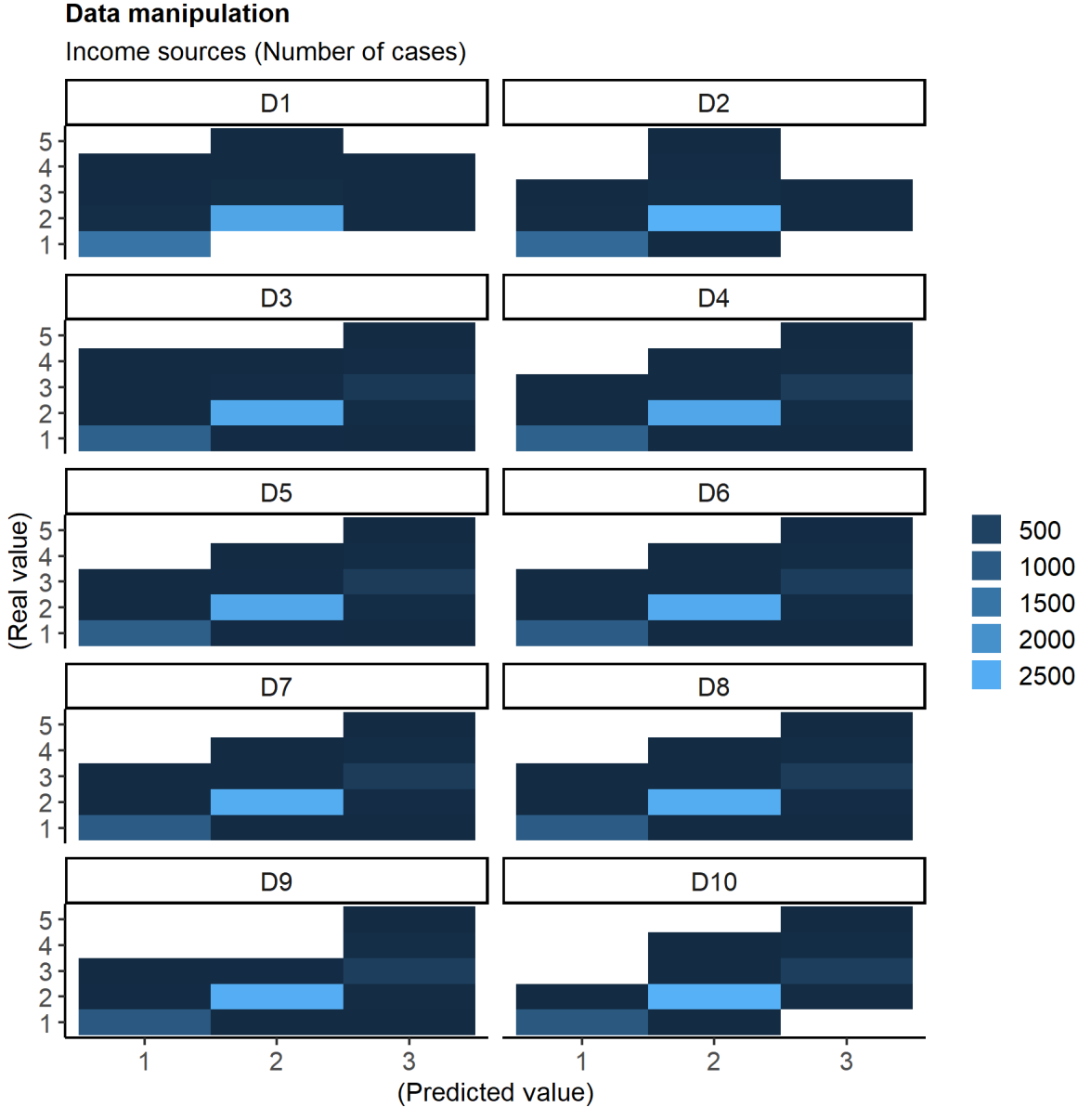


وعلى نقيض الرسم البياني السابق، تركز هذه التقنية على جانب الرسم البياني الخاص بالمستفيدين الحاليين. وفي هذه الحالة، يمكن ملاحظة شذوذ مرتبط بالمنطقة D حيث توجد أكبر نسبة من الأفراد الذين يقدّمون قيّماً غير نمطية، وبالتالي، يكون من الحكمة فهم ما يحدث هناك بشكل أفضل.

وفي هذا السياق، من المهم مراجعة قدرة النماذج على التنبؤ.

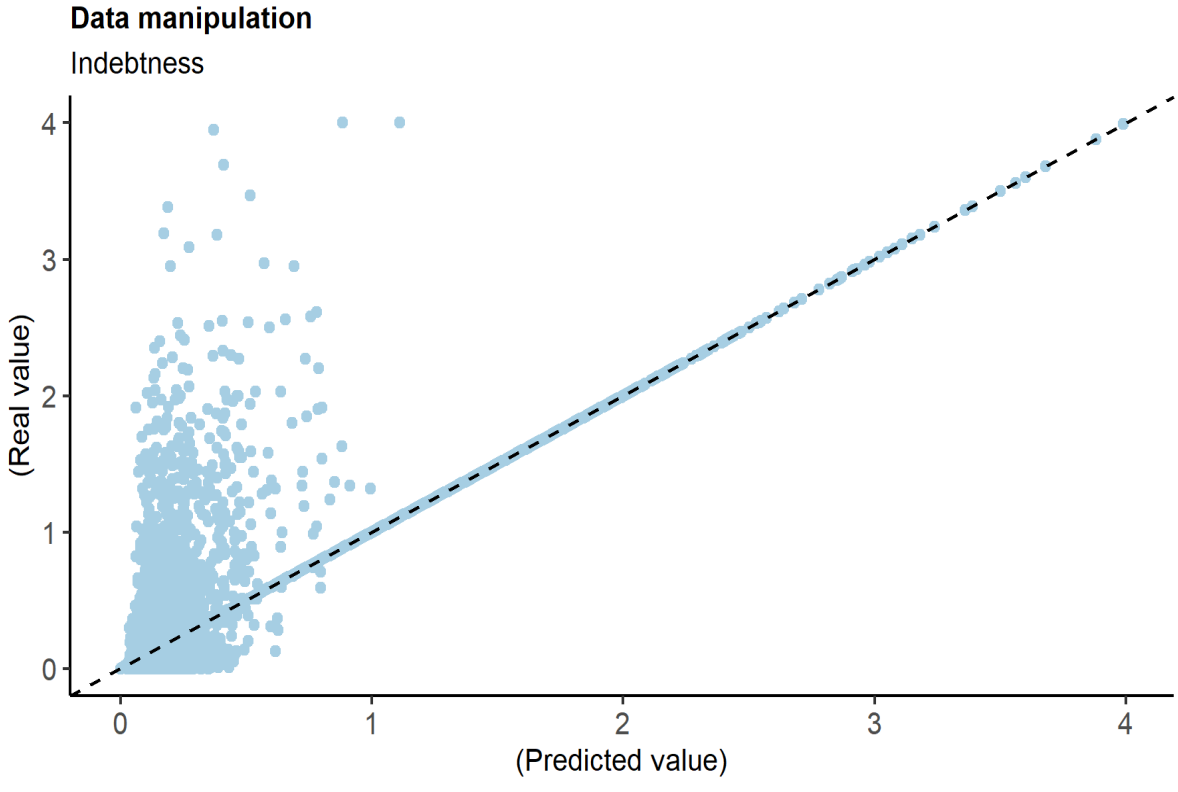


كما يمكنك الملاحظة، يشكّل هذا المخطط مثالاً عما هو متوقع من نموذج التنبؤ. التنبؤات محاذية بشكل أو بآخر للقيم الحقيقية، ويسجّل بعض التشويش الطبيعي حول هذا التنبؤ (لاحظ أن النقاط على الخط القُطري تعود لغير المستفيدين، كما لم يتم استبدال قيمها).



يُستخدم هذا الرسم البياني لاستعراض الفئات. لاحظ أن هذا النموذج يعرض مشاكل التنبؤ لأنه لا يتنبأ بأكثر من ثلاثة مصادر، جميعها إما 1، 2، أو 3. واستناداً إلى تدرج اللون، عادة ما يكون التنبؤ من مصادر 2 دقيقة جداً،

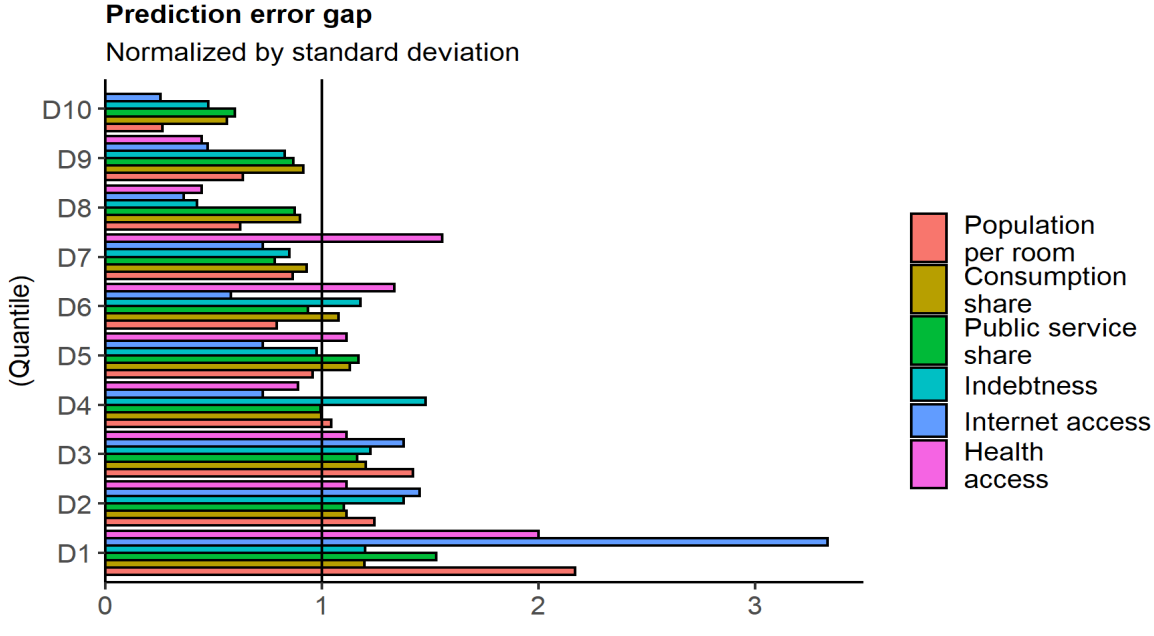
ولكن فيما خص 1 و3 لا تعتبر مصادر التنبؤ موثوقة للغاية. وهذا يشير إلى أنه بسبب التشويش داخل هذا المتغير، يفضل إنشاء آليات للتحقق من صحة المعلومات التي قد تكون مصدراً للخطأ.



وأخيراً، لاحظ أن للمديونية اتجاه غريب حيث كان يُتوقّع أن تحرز مجموعة كبيرة من الأفراد قيّماً أقل بكثير من تلك التي ادعوها. ويمكن وضع فرضية عمل عند الأخذ في الاعتبار أيضاً أن المنطقة D، وهي منطقة من الطبقة المتوسطة، قد أُشير إلى أنّها مصدراً محتملاً للمعلومات الخاطئة. مثلاً، ونظراً لكون الطبقة المتوسطة قريبة من مسألة خفض الاستحقاقات، سيبلّغ الناس أرقاماً أعلى من الواقع فيما يتعلّق بمصادر الدخل للتأكد من أنّ أرقام المديونية ستمكنهم من الوصول إلى البرنامج. وفي هذه الحالة، يجب أن تضع هيئة تنظيم السياسات المزيد من ضوابط الرصد في هذه المجموعة، وفي المنطقة D. وتتسق هذه الفرضية مع القصة المذكورة في الرسوم البيانية الأربعة السابقة. بالرغم من استحالة ضمان أن هذا ما يحدث فعلاً، لكنّ هذه الفرضية قد تساعد صانعي السياسات على إنشاء آليات رصد أفضل وأكثر تركيزاً من شأنها تحسين تخصيص الموارد للأشخاص الأكثر حاجة إليها. في هذا القسم، لم يتضمن الدليل بعض النتائج الأخرى، ولكن يمكن للمتدرب مراجعتها وتحديد غيرها من الفرضيات المحتملة المتعلقة بتغطية البرنامج.

```
# 6| Outliers by quantile ------------------------------------------------

temp$error_popRoom= (temp$popRoom-temp$popRoom0)^2
temp$error_popRoom=temp$error_popRoom/mean(temp$error_popRoom)

temp$error_consumptionShare= (temp$consumptionShare-temp$consumptionShare0)^2
temp$error_consumptionShare=temp$error_consumptionShare/mean(temp$error_consumptionShare)

temp$error_publicServiceShare= (temp$publicServiceShare-temp$publicServiceShare0)^2
temp$error_publicServiceShare=temp$error_publicServiceShare/mean(temp$error_publicServiceShare)

temp$error_indebtness= (temp$indebtness-temp$indebtness0)^2
temp$error_indebtness=temp$error_indebtness/mean(temp$error_indebtness)

temp$error_internetAcces_1= (temp$internetAcces_1-temp$internetAcces_10)^2
temp$error_internetAcces_1=temp$error_internetAcces_1/mean(temp$error_internetAcces_1)

temp$error_healthAcces_1= (temp$healthAcces_1-temp$healthAcces_10)^2
temp$error_healthAcces_1=temp$error_healthAcces_1/mean(temp$error_healthAcces_1)

temp=temp%>%dplyr::select(starts_with('error'), 'incomeQuantile')%>%
  melt(id='incomeQuantile')%>%group_by(variable, incomeQuantile)%>%
  summarise(value=mean(value))
```

لإنهاء التحليل، تقيس السطور 711-733 الخطأ في التنبؤ. في هذه الحالة يتم احتساب الخطأ في التنبؤ على شكل الخطأ المربع (اي الفرضية المتوقعة ناقص الواقع المسجل، جميعها مربعة)، ومن ثم توحّد النتيجة بمتوسط الخطأ. عند القيام بهذا الإجراء، يمكن أن يكون الخطأ دائماً بمثابة معيار حول القيمة 1.



بشكل عام، للشارات العالية قوة تنبوئية أكثر من تلك المنخفضة. وهذا منطقي بسبب ميل الأفراد الأغنياء للحصول على قيَم ايجابية أكثر ثباتا. مثلاً، يكون معدل الأشخاص في الغرفة الواحدة أقل، وحصص الاستهلاك أقل. وعلى النقيض من ذلك، يكبر النطاق بالنسبة للفقراء حيث يمكن أن يتجلى الفقر بطرق عديدة. ومع ذلك،

هناك متغيران يستحقان تسليط الضوء عليهما. الأول هو الصحة التي برز لديها ذروة غير نمطية في العشر السابع، والمديونية التي سجلت ذروة غير نمطية في العشر الرابع. بالنسبة للحالة الأخيرة، تدعم هذه النتيجة الفرضية المعلنة سابقاً حول وجود بعض الديناميكيات الجارية في الطبقة المتوسطة والتي يمكن أن تسبب خطأ في تصنيف الأفراد. من خلال ملاحظة تلك المتغيرات ذات الأخطاء الأكبر، يمكن لصناع السياسات تحديد تلك التي تحتاج أكبر قدر من السيطرة من جانب السلطات بسبب تباياتها الكبيرة التي تصعّب عملية التحقق منها إحصائياً. في هذه الحالة، مثلاً، فإنّ المتغير المرتبط بعدد الأفراد في الغرفة الواحدة، أو الإنترنت لأفقر عشر أُسر، يعتبر عشوائي. لذلك يصعب إنشاء خوارزمية تخمن بدقة صحّة المعلومات التي يقدّمها الفرد. ثم، من الضروري تنفيذ ضوابط إضافية لدعم رصد المستفيدين لتجنب سوء توزيع المعونة.

13. بعض الملاحظات حول تقييم المستفيدين

في حين أن هذا الموضوع لن يناقش رسمياً في الدليل الحالي، يقدم هذا القسم بعض الأفكار التي توجه المتدرب حول كيفية المضي قدماً. وتهدف هذه الخطوة الأخيرة إلى تتبع الأفراد عبر الزمن. ولهذا السبب، يكمن العنصر الأول في تحديد المتغيرات الرئيسية للتحقق من تطور المستفيد. مثلاً، إذا ركز البرنامج الاجتماعي على تحسين قابلية الفرد للتوظيف، فإن المتغير الطبيعي للتتبع هو الوقت الذي يمضيه الفرد في البرنامج. ومن الأمثلة الأخرى البرامج المتصلة بالصحة، حيث يكون متغير التتبع هو صحة المستفيد.

ومع ذلك، لا بد من تسليط الضوء على ثلاثة تحديات. ويتمثل التحدي الأول في تحسّن الحالة أو ازديادها سوءاً أحياناً بمعزل عن البرنامج. مثلاً، بالنسبة لبرامج مساعدات العاطلين عن العمل، إذا تحسّن أداء الاقتصاد، يتوقع أن يحصل الناس على وظائف أسرع، بمعزل عن مشاركتهم في البرنامج. وبالتالي، تحتاج الاستراتيجية الإحصائية إلى التمعّن في تطور المستفيد وتناقضه مع تطور الأشخاص غير المستفيدين (مجموعة المراقبة).

ويتعلق التحدي الثاني بمشاركة الأفراد في البرنامج. وتتجاوز عدة برامج خدمات تحويل الدخل، وتقوم بحلقات عمل وفعاليات وأنشطة تهدف إلى دعم المستفيدين. وينبغي معرفة ما إذا كان بالإمكان ربط هذا التحسن بالبرنامج إذا ما كان الفرد يشارك في هذه الأنشطة. ومع ذلك، فإن التوقيع على قائمة المساعدة لا يعني بالضرورة المشاركة في هذه الأحداث. وبالتالي، تحتاج البرامج إلى تطوير آليات رصد يمكنها تحديد مدى المشاركة الفعلية للأفراد.

ويرتبط التحدي الأخير بالمشاركة في برامج متعددة. وفقاً للبلد، يميل الأفراد لأن يكونوا مستفيدين من أكثر من برنامج، ولذلك، فإن أي تحسن يحتاج إلى تحليل ما إذا نتج عن المشاركة في برنامج واحد فقط، أو كنتيجة التآزر بين عدة برامج. وبالتالي، وعلى نقيض الخطوات السابقة، من المهم الاطّلاع على مختلف البرامج التي يشارك فيها الفرد.

ومما لا شك فيه أن المتطلبات الكبيرة من البيانات تجعل هذه الخطوة مستحيلة التنفيذ على العديد من البلدان بالنظر إلى حالة نُظُم المعلومات لديها. ومع ذلك، من المهم أخذ هذه الأفكار في الاعتبار كي يتطور النظام ويشتمل على المزيد من هذه المتغيرات، وبهذه الطريقة، يمكن أن يتحسن تحليل البرامج باستمرار.

14. ملاحظات ختامية

الدليل الحالي هو دليل تدريبي أعدته الإسكوا لدعم مجموعة من أنشطة بناء القدرات الموجهة للبلدان التي ترغب في تحسين طريقة استخدامها للمعلومات لرفع كفاءة برامج الحماية الاجتماعية وفعاليتها. وكما ذكر في الدليل، فلكل بلد مميزاته الخاصة، وبالتالي لن يكتمل التحليل الإحصائي دون معرفة معمقة للمؤسسات، الثقافة، البيئة، الممارسات الاجتماعية والاقتصادية والسياق السياسي لبلد ما. ومع ذلك، يمكن استخدام أدوات تحليل البيانات لإنشاء تقييم سريع لمعالجة الجوانب التي يمكن أن تضع معوّقات أو تحديات أو حتى فرصاً للبرامج.
وفي هذا الشأن، يمكن النظر إلى هذا الدليل على أنه يخدم غرضين. فمن ناحية، يعد الدليل أداة مفاهيمية، وعبر التفسيرات النظرية في الجزء 2، يمكن لصانعي السياسات الاطّلاع على مجموعة من الأدوات المفيدة، التي يتم تنظيمها في إطار تحليلي منطقي (SPP-RAF)). ومن جهة أخرى، يشكّل الدليل أداة تطبيقية توجه التحليل المتعلق بأفضل الممارسات لاستخدام البرنامج الإحصائي R وكيفية استخدامه لتحسين أتمتة التقارير وإضفاء الطابع المهني عليها. وثم، من المتوقع أن يكون للدليل، من خلال خدمة هذين الجمهورين، أثراً إيجابيّاً على برامج الحماية الاجتماعية للبلدان، ومن خلالهما، على تحسين سُبُل عيش الناس.

المراجع

Breiman, L. (1984). Classification and regression trees. *The Wadsworth and Brooks-Cole statisticsprobability series.* Chapman & Hall.

Gareth, J., Witten, D., Hastie, T., & Tibshirani, R. (2017). *An Introduction to Statistical Learning: with Applications in R.* New York: Springer.

GIZ. (2017). *Vulnerability and capacity assessment with regard to social protection.* Berlin: BMZ.

Hastie, T., Tibshirani, R., & Friedman, J. H. (2009). *The Elements of Statistical Learning (2nd ed.).* New York: Springer.

ILO. (2016). *Social protection assessment-based national dialogue: A global guide.* Geneva: ILO.

Kim, Seongho (2015). ppcor: An R package for a fast calculation to semi-partial correlation coefficients. *Commun Stat Appl Methods,* vol. 22, No. 6, pp. 665-674.

Mahalanobis, P. C. (1936). On the generalized distance in statistics. *Proceedings of the National Institute of Sciences of India*, 49-55.

OECD. (2019). *Monitoring and evaluating social protection systems.* Paris: OECD.

Tibshirani, R. (1996). Regression Shrinkage and Selection via the lasso. *Journal of the Royal Statistical Society. Series B (methodological)*, 267-288.

UNICEF. (2019). *UNICEF's Global Social Protection Programme Framework.* New York: UNICEF.

Ward, J. H. (1963). Hierarchical Grouping to Optimize an Objective Function. *Journal of the American Statistical Association*, 236-244.

Watanabe, S. (1969). *Knowing and guessing; a quantitative study of inference and information.* New York: Wiley.

World Bank (2016) مجموعة البنك الدولي Editorial Style Guide. Washington D.C.

Zelterman, D. (2015). *Applied Multivariate Statisticas with R.* New Haven: Springer.

يعمل دليل التدريب كأداة تعليمية أنتجتها اللجنة الاقتصادية والاجتماعية لغربي آسيا (الإسكوا) لاستكمال منهج التدريب على التحليل الكمي لبرامج الحماية الاجتماعية (حيث لا يعتبر بمثابة مادة للتعلُّم الذاتي).

يأتي الدليل مرفقاً بالملفات التي تحتوي على الرموز ومجموعات البيانات المستخدَمة لشرح المحتويات على اختلافها. ويمكن استخدامه أيضا لتنشيط ذاكرة المتدربين.

يقدم هذا الدليل إطار التقييم السريع لبرنامج الحماية الاجتماعية (SPP-RAF) الذي يهدف إلى تقديم تقييمات عملية منخفضة التكلفة ويمكن تنفيذها بانتظام كجزء من البرامج ويجري اطلاع صناع السياسات على النقاط الرئيسية التي تؤثر على نجاح البرامج.

ويقسم الدليل إلى جزأين. يهدف الجزء الأول إلى تقديم عرض شامل للبرمجيات الإحصائية المستخدَمة لتنفيذ إطار SPP-RAF. وقد تمّ اختيار البرنامج الإحصائي R للقيام بذلك. ونظراً لمرونته الكبيرة وخصائصه، وكونه منصة مفتوحة المصدر، أصبح البرنامج الإحصائي R أكثر شعبية في تطوير التحليلات الكمية التي تتطلب عادة توحيد مجموعات البيانات الكبيرة وتقوم بوضع تقارير ضخمة. ولهذه الأسباب، يستند برنامج التدريب الحالي إلى هذه المجموعة الإحصائية.

يقدم الجزء الأول من الدليل لمحة عن البرنامج الإلكتروني، فيما يناقش الجزء الثاني منه ، الخطوات المختلفة لإطار SPP-RAF وذلك على الصعيد المفاهيمي ثم يعرض الأدوات والرموز ذات الصلة اللازمة لتنفيذه.

وأخيرا، ولتحسين الصلة بين المفاهيم والتطبيق الفعلي، تستند جميع الأمثلة المقدمة لتنفيذ إطار SPP-RAF - إلى بلد افتراضي لديه برنامج للحماية الاجتماعية بحاجة للتقييم.

22-00244