Quick R per l'incontro del 5 Febbraio

Elisabetta Rocchetti

2/4/2022

Prima di tutto: impostiamo la working directory. Questo passaggio è comodo perché, in seguito, dovremo inserire dei PATH (vedremo dopo cosa sono). Impostato questa working directory, ci risparmiamo di scrivere tutto il PATH assoluto (dalla root del file system) ogni volta.

```
#ognuno ne avrà una diversa
setwd("~/Desktop/Universita/tirocinio/cheat sheet/")
```

Importare i dati

Per importare i dati è necessario sapere dove è collocato il dataset che ci serve e l'estensione del file che vogliamo importare. Infatti, ci sono molti modi per salvare dei dati, dobbiamo dire a R che strumento utilizzare per estrarre i dati.
Ogni funzione mostrata in seguito può essere utilizzata per importare un tipo di file in particolare; per una spiegazione più accurata di come utilizzare questi metodi e quali argomenti poter impostare è necessario andare a cercare in documentazione.
Inoltre, ogni volta che vedete scritto "file.qualcosa", dovete inserire il vero PATH del file che volete importare. Il PATH è il percorso univoco che il File System in un Sistema Operativo utilizza per identificare e localizzare una risorsa (file).

.txt

Importo i dati da un file *.txt*: la funzione read_delim() legge qualsiasi file .txt nel quale i dati sono separati da un certo carattere. In questo esempio, se il parser vede il carattere "|", sa che ciò che leggerà dopo appartiene a un'altra riga o cella

```
read_delim("file.txt", delim="|")
```

.csv

Importo i dati da un file *.csv*. Importare i dati da questo tipo di file è molto più frequente. CSV vuol dire "Comma Separated Values", ma non fatevi ingannare: a volte i valori possono essere separati anche da un carattere diverso dalla virgola, come ad esempio il "semi-colon" (cioè il punto e virgola).Fate attenzione anche al modo in cui i numeri sono stati salvati: nei paesi anglosassoni, la virgola del decimale è in verità un punto (avrete infatti notato che alcune calcolatrici usano il punto anziché la virgola). Il formato stardard della funzione read_csv per indicare il decimale è il punto; se nel vostro dataset non è così, dovete specificarlo con un apposito parametro della funzione.

```
read_csv("file.csv")
#eseguire il codice qua sotto se la virgola è usata per separare la parte intera
#da quella decimale
read_csv("file.csv", decimal = ",")
```

.xlsx

Un altro modo per salvare i dati è utilizzando il formato excel .xlsx (probabilmente ciò che conoscete maggiormente).

```
#Per usare questa funzione è necessario importare la libreria readxl.
#Non ce l'hai? Esegui il comando "install.packages('readxl')".
library(readxl)
read_excel("file.xlsx")
```

Se vuoi saperne di più su come importare i file, ti raccomando di dare una sbirciata al cheat sheet dedicato sulla repository GitHub (oltre a cercare su Google). ## Esempio Importiamo il nostro dataset .xlsx.

```
library(readxl)
data <- read_excel("Data Science for Citizens_cleaned.xlsx")
```

Capire i dati

Come fai a fare una torta senza farina? Non puoi. E come reagiresti se hai impiegato 4 ore del tuo tempo al supermercato (di cui 2 per fare la coda alla cassa), dimenticandoti di comprare proprio la farina? Male. Per evitare di triggerarsi per farine dimenticate sugli scaffali e per torte non fatte, sarebbe necessario controllare bene che nel carrello ci sia tutto ciò che ti serve.
Collezionare i dati adatti a fare un'analisi è simile a fare la spesa: ci vuole tempo e non puoi ottenere i risultati desiderati se non hai i dati giusti. In alcuni casi, non solo non otterresti il risultato giusto, ma non riusciresti neanche a eseguire la statistica che vuoi (ad esempio, non puoi ottenere la media tra variabili categoriche).
Per questo motivo, appena abbiamo a disposizione dei dati, bisogna analizzarli per bene e controllare che siano giusti giusti per la nostra analisi: ci servono degli strumenti che ci aiutano a fare ciò.
Per prima cosa, otteniamo una visione generale sui nostri dati.

```
#al posto di "dataset" metti io nome della variabile che contiene il tuo dataset
View(dataset)
summary(dataset)
```

Esempio

```
View(data)
summary(data)
```

```
##   Indirizzo            Genere           Statistica        stat_software
##  Length:22          Length:22          Length:22          Length:22
##  Class :character   Class :character   Class :character   Class :character
##  Mode  :character   Mode  :character   Mode  :character   Mode  :character
##
##
##
##
##  Stat_importante Stat_difficile  Motivazione          Smartphone
##  Min.   :3.000   Min.   :3.000   Length:22          Min.   :1.00
##  1st Qu.:3.250   1st Qu.:3.000   Class :character   1st Qu.:3.00
##  Median :4.000   Median :3.000   Mode  :character   Median :4.00
##  Mean   :3.955   Mean   :3.545                      Mean   :3.37
##  3rd Qu.:4.000   3rd Qu.:4.000                      3rd Qu.:4.00
##  Max.   :5.000   Max.   :5.000                      Max.   :4.50
##
```

```
##  Post_diploma       Patente_nonne        Matematica     Donne_ICT
##  Length:22          Length:22          Min.   :0.300   Length:22
##  Class :character   Class :character   1st Qu.:1.000   Class :character
##  Mode  :character   Mode  :character   Median :1.500   Mode  :character
##                                        Mean   :1.427
##                                        3rd Qu.:2.000
##                                        Max.   :2.000
##
##  Ambiente_importanza  PM25_oggi           PM10_oggi         Liberta
##  Min.   : 6.000      Length:22          Min.   : 20.00   Min.   :1.00
##  1st Qu.: 8.000      Class :character   1st Qu.: 23.75   1st Qu.:4.25
##  Median : 9.000      Mode  :character   Median : 30.00   Median :6.50
##  Mean   : 8.682                         Mean   : 45.85   Mean   :6.00
##  3rd Qu.:10.000                         3rd Qu.: 52.50   3rd Qu.:7.75
##  Max.   :10.000                         Max.   :230.00   Max.   :8.00
##                                         NA's   :2
##     Paese              Lavoro             Reddito      Temperatura_estate
##  Length:22          Length:22          Min.   :1.000   Length:22
##  Class :character   Class :character   1st Qu.:3.500   Class :character
##  Mode  :character   Mode  :character   Median :4.000   Mode  :character
##                                        Mean   :4.579
##                                        3rd Qu.:6.000
##                                        Max.   :8.000
##                                        NA's   :3
##  Temparatura_innalzamento Fonte_informazione Uso_tecnologie
##  Min.   : 1.50            Length:22          Min.   :0.0000
##  1st Qu.: 2.00            Class :character   1st Qu.:0.0000
##  Median : 6.25            Mode  :character   Median :1.0000
##  Mean   : 7.45                               Mean   :0.5455
##  3rd Qu.:10.00                               3rd Qu.:1.0000
##  Max.   :40.00                               Max.   :1.0000
##  NA's   :2
##  Orientamento_post_diploma Emancipazione_Femminile Ambiente_Inquinamento
##  Min.   :0.0000            Min.   :0.0000          Min.   :0.0000
##  1st Qu.:0.0000            1st Qu.:0.0000          1st Qu.:0.0000
##  Median :1.0000            Median :0.0000          Median :0.0000
##  Mean   :0.6818            Mean   :0.3182          Mean   :0.3636
##  3rd Qu.:1.0000            3rd Qu.:1.0000          3rd Qu.:1.0000
##  Max.   :1.0000            Max.   :1.0000          Max.   :1.0000
##
##  Reddito_Lavoro   Liberta_civili   Surriscaldamento_globale
##  Min.   :0.0000   Min.   :0.0000   Min.   :0.0000
##  1st Qu.:0.0000   1st Qu.:0.0000   1st Qu.:0.0000
##  Median :0.0000   Median :0.0000   Median :0.0000
##  Mean   :0.4091   Mean   :0.2273   Mean   :0.1818
##  3rd Qu.:1.0000   3rd Qu.:0.0000   3rd Qu.:0.0000
##  Max.   :1.0000   Max.   :1.0000   Max.   :1.0000
##
```

Grazie a questi risultati possiamo farci un'idea di cosa abbiamo tra le mani e possiamo iniziare a porci delle domande: le variabili sono state importate come ci aspettavamo? Ci sono delle variabili di un tipo diverso da quello che ci aspettavamo? Dobbiamo trasformare le variabili categoriche in factor? Ci sono valori mancanti (NA)? Ci sono valori non conformi al dominio di appartenenza?
A volte, se non si è abbastanza attenti in questa fase, si ottengono dei risultati anomali e, nel migliore dei

casi, ce ne accorgiamo e lo sistemiamo; in altri casi, l'anomalia potrebbe essere meno percettibile e potremmo non accorgerci dell'erroneità del risultato.
Per avere risposta a tutte queste domande, potremmo aver bisogno di descrizioni più dettagliate, ma per ora andiamo avanti con l'analisi. Avremo tempo per andare più in profondità con il nostro progetto.

Filtrare e correggere il dataset

In alcuni casi, ci sono molte informazioni nei dataset non necessarie alla nostra analisi. Oppure ci sono delle variabili che bisogna trasformare (ad esempio, numeri salvati come stringhe). Ora vediamo come filtare righe, colonne e trasformare le variabili. Per fare ciò, useremo il package dplyr (se non l'hai installato, esegui il comando "install.packages('dplyr')").

```
#install.packages('dplyr')
library(dplyr)
```

Selezionare solo alcune colonne

```
dataset %>% select(columns)
```

Esempio

```
data %>% select(Patente_nonne, Reddito)
```

```
## # A tibble: 22 x 2
##    Patente_nonne Reddito
##    <chr>            <dbl>
##  1 Una                  4
##  2 Nessuna             NA
##  3 Entrambe             3
##  4 Nessuna              4
##  5 Nessuna              5
##  6 Una                  6
##  7 Una                  8
##  8 Entrambe             7
##  9 Una                  1
## 10 Nessuna              6
## # ... with 12 more rows
```

Selezionare solo alcune righe

```
dataset %>% filter(condition)
```

Esempio

```
data %>% filter(Patente_nonne == "Una")
```

```
## # A tibble: 6 x 29
##   Indirizzo       Genere Statistica stat_software Stat_importante Stat_difficile
##   <chr>           <chr>  <chr>      <chr>                   <dbl>          <dbl>
## 1 Tradizionale    Femmi~ No         No                          4              3
## 2 Scienze Applic~ Uomo   No         Si                          4              3
## 3 Scienze Applic~ Masch~ No         Si                          4              4
## 4 Quadriennale    Donna  No         No                          5              3
```

```
## 5 Scienze Applic~ Femmi~ No         No                          4              4
## 6 Scienze Applic~ Femmi~ Si         No                          5              4
## # ... with 23 more variables: Motivazione <chr>, Smartphone <dbl>,
## #   Post_diploma <chr>, Patente_nonne <chr>, Matematica <dbl>, Donne_ICT <chr>,
## #   Ambiente_importanza <dbl>, PM25_oggi <chr>, PM10_oggi <dbl>, Liberta <dbl>,
## #   Paese <chr>, Lavoro <chr>, Reddito <dbl>, Temperatura_estate <chr>,
## #   Temparatura_innalzamento <dbl>, Fonte_informazione <chr>,
## #   Uso_tecnologie <dbl>, Orientamento_post_diploma <dbl>,
## #   Emancipazione_Femminile <dbl>, Ambiente_Inquinamento <dbl>,
## #   Reddito_Lavoro <dbl>, Liberta_civili <dbl>, Surriscaldamento_globale <dbl>
```

Esempio

```
#possiamo anche concatenare le operazioni
data %>%
  select(Patente_nonne, Reddito) %>%
  filter(Patente_nonne == "Una")
```

```
## # A tibble: 6 x 2
##   Patente_nonne Reddito
##   <chr>           <dbl>
## 1 Una                 4
## 2 Una                 6
## 3 Una                 8
## 4 Una                 1
## 5 Una                 4
## 6 Una                 3
```

Esempio

```
#seleziono le righe che non hanno valori mancanti in nessuna colonna
data %>% filter(if_all(everything(), ~ !is.na(.)))
```

```
## # A tibble: 17 x 29
##    Indirizzo     Genere  Statistica stat_software Stat_importante Stat_difficile
##    <chr>         <chr>   <chr>      <chr>                   <dbl>          <dbl>
##  1 Tradizionale  Femmina No         No                          4              3
##  2 Scienze Appl~ Maschio No         No                          5              3
##  3 Scienze Appl~ Maschio No         No                          3              3
##  4 Scienze Appl~ Uomo    No         Si                          4              3
##  5 Scienze Appl~ Maschio No         Si                          4              4
##  6 Scienze Appl~ Maschio Si         No                          5              3
##  7 Quadriennale  Donna   No         No                          5              3
##  8 Scienze Appl~ Maschio Si         No                          4              3
##  9 Tradizionale  Femmina Si         No                          4              5
## 10 Tradizionale  Femmina No         No                          3              5
## 11 Scienze Appl~ Maschi~ No         No                          5              4
## 12 Scienze Appl~ Femmina No         No                          3              4
## 13 Scienze Appl~ Maschi~ No         No                          4              4
## 14 Scienze Appl~ Maschio No         No                          4              3
## 15 Scienze Appl~ Femmina No         No                          4              4
## 16 Tradizionale  Maschio No         No                          4              3
## 17 Scienze Appl~ Femmina Si         No                          5              4
## # ... with 23 more variables: Motivazione <chr>, Smartphone <dbl>,
```

```
## #   Post_diploma <chr>, Patente_nonne <chr>, Matematica <dbl>, Donne_ICT <chr>,
## #   Ambiente_importanza <dbl>, PM25_oggi <chr>, PM10_oggi <dbl>, Liberta <dbl>,
## #   Paese <chr>, Lavoro <chr>, Reddito <dbl>, Temperatura_estate <chr>,
## #   Temparatura_innalzamento <dbl>, Fonte_informazione <chr>,
## #   Uso_tecnologie <dbl>, Orientamento_post_diploma <dbl>,
## #   Emancipazione_Femminile <dbl>, Ambiente_Inquinamento <dbl>,
## #   Reddito_Lavoro <dbl>, Liberta_civili <dbl>, Surriscaldamento_globale <dbl>
```

Esempio

```
#seleziono le colonne che non hanno valori mancanti
data %>% select_if(~ !any(is.na(.)))
```

```
## # A tibble: 22 x 24
##    Indirizzo      Genere Statistica stat_software Stat_importante Stat_difficile
##    <chr>          <chr>  <chr>      <chr>                  <dbl>          <dbl>
##  1 Tradizionale   Femmi~ No         No                         4              3
##  2 Scienze Appli~ Masch~ No         No                         4              3
##  3 Scienze Appli~ Femmi~ Si         No                         4              4
##  4 Scienze Appli~ Masch~ No         No                         5              3
##  5 Scienze Appli~ Masch~ No         No                         3              3
##  6 Scienze Appli~ Uomo   No         Si                         4              3
##  7 Scienze Appli~ Masch~ No         Si                         4              4
##  8 Scienze Appli~ Masch~ Si         No                         5              3
##  9 Quadriennale   Donna  No         No                         5              3
## 10 Scienze Appli~ Masch~ Si         No                         4              3
## # ... with 12 more rows, and 18 more variables: Motivazione <chr>,
## #   Smartphone <dbl>, Post_diploma <chr>, Patente_nonne <chr>,
## #   Matematica <dbl>, Donne_ICT <chr>, Ambiente_importanza <dbl>,
## #   PM25_oggi <chr>, Liberta <dbl>, Paese <chr>, Temperatura_estate <chr>,
## #   Uso_tecnologie <dbl>, Orientamento_post_diploma <dbl>,
## #   Emancipazione_Femminile <dbl>, Ambiente_Inquinamento <dbl>,
## #   Reddito_Lavoro <dbl>, Liberta_civili <dbl>, Surriscaldamento_globale <dbl>
```

Trasformare i tipi delle variabili

```
#trasformare una variabile esistente
dataset %>% mutate(column = func(column))

#aggiungere una variabile nuova
dataset %>% mutate(column = func(some_other_column))
```

Esempio

```
#trasformiamo Reddito in una stringa
data %>% mutate(Reddito = as.character(Reddito))
```

```
## # A tibble: 22 x 29
##    Indirizzo      Genere Statistica stat_software Stat_importante Stat_difficile
##    <chr>          <chr>  <chr>      <chr>                  <dbl>          <dbl>
##  1 Tradizionale   Femmi~ No         No                         4              3
##  2 Scienze Appli~ Masch~ No         No                         4              3
##  3 Scienze Appli~ Femmi~ Si         No                         4              4
##  4 Scienze Appli~ Masch~ No         No                         5              3
```

```
##  5 Scienze Appli~ Masch~ No         No                          3             3
##  6 Scienze Appli~ Uomo   No         Si                          4             3
##  7 Scienze Appli~ Masch~ No         Si                          4             4
##  8 Scienze Appli~ Masch~ Si         No                          5             3
##  9 Quadriennale   Donna  No         No                          5             3
## 10 Scienze Appli~ Masch~ Si         No                          4             3
## # ... with 12 more rows, and 23 more variables: Motivazione <chr>,
## #   Smartphone <dbl>, Post_diploma <chr>, Patente_nonne <chr>,
## #   Matematica <dbl>, Donne_ICT <chr>, Ambiente_importanza <dbl>,
## #   PM25_oggi <chr>, PM10_oggi <dbl>, Liberta <dbl>, Paese <chr>, Lavoro <chr>,
## #   Reddito <chr>, Temperatura_estate <chr>, Temparatura_innalzamento <dbl>,
## #   Fonte_informazione <chr>, Uso_tecnologie <dbl>,
## #   Orientamento_post_diploma <dbl>, Emancipazione_Femminile <dbl>,
## #   Ambiente_Inquinamento <dbl>, Reddito_Lavoro <dbl>, Liberta_civili <dbl>,
## #   Surriscaldamento_globale <dbl>
```

Nota che ora la variabile Reddito è un char (chr).

```
#ri-trasformiamo Reddito in un numero
data %>% mutate(Reddito = as.numeric(Reddito))
```

```
## # A tibble: 22 x 29
##    Indirizzo      Genere Statistica stat_software Stat_importante Stat_difficile
##    <chr>          <chr>  <chr>      <chr>                  <dbl>          <dbl>
##  1 Tradizionale   Femmi~ No         No                         4              3
##  2 Scienze Appli~ Masch~ No         No                         4              3
##  3 Scienze Appli~ Femmi~ Si         No                         4              4
##  4 Scienze Appli~ Masch~ No         No                         5              3
##  5 Scienze Appli~ Masch~ No         No                         3              3
##  6 Scienze Appli~ Uomo   No         Si                         4              3
##  7 Scienze Appli~ Masch~ No         Si                         4              4
##  8 Scienze Appli~ Masch~ Si         No                         5              3
##  9 Quadriennale   Donna  No         No                         5              3
## 10 Scienze Appli~ Masch~ Si         No                         4              3
## # ... with 12 more rows, and 23 more variables: Motivazione <chr>,
## #   Smartphone <dbl>, Post_diploma <chr>, Patente_nonne <chr>,
## #   Matematica <dbl>, Donne_ICT <chr>, Ambiente_importanza <dbl>,
## #   PM25_oggi <chr>, PM10_oggi <dbl>, Liberta <dbl>, Paese <chr>, Lavoro <chr>,
## #   Reddito <dbl>, Temperatura_estate <chr>, Temparatura_innalzamento <dbl>,
## #   Fonte_informazione <chr>, Uso_tecnologie <dbl>,
## #   Orientamento_post_diploma <dbl>, Emancipazione_Femminile <dbl>,
## #   Ambiente_Inquinamento <dbl>, Reddito_Lavoro <dbl>, Liberta_civili <dbl>,
## #   Surriscaldamento_globale <dbl>
```

Nota che ora la variabile Reddito è un double (dbl).

```
#aggiungiamo una variabile
data %>% mutate(Reddito_al_quadrato = Reddito^2)
```

```
## # A tibble: 22 x 30
##    Indirizzo      Genere Statistica stat_software Stat_importante Stat_difficile
##    <chr>          <chr>  <chr>      <chr>                  <dbl>          <dbl>
##  1 Tradizionale   Femmi~ No         No                         4              3
##  2 Scienze Appli~ Masch~ No         No                         4              3
##  3 Scienze Appli~ Femmi~ Si         No                         4              4
##  4 Scienze Appli~ Masch~ No         No                         5              3
##  5 Scienze Appli~ Masch~ No         No                         3              3
```

```
##  6 Scienze Appli~ Uomo   No         Si                         4             3
##  7 Scienze Appli~ Masch~ No         Si                         4             4
##  8 Scienze Appli~ Masch~ Si         No                         5             3
##  9 Quadriennale   Donna  No         No                         5             3
## 10 Scienze Appli~ Masch~ Si         No                         4             3
## # ... with 12 more rows, and 24 more variables: Motivazione <chr>,
## #   Smartphone <dbl>, Post_diploma <chr>, Patente_nonne <chr>,
## #   Matematica <dbl>, Donne_ICT <chr>, Ambiente_importanza <dbl>,
## #   PM25_oggi <chr>, PM10_oggi <dbl>, Liberta <dbl>, Paese <chr>, Lavoro <chr>,
## #   Reddito <dbl>, Temperatura_estate <chr>, Temparatura_innalzamento <dbl>,
## #   Fonte_informazione <chr>, Uso_tecnologie <dbl>,
## #   Orientamento_post_diploma <dbl>, Emancipazione_Femminile <dbl>,
## #   Ambiente_Inquinamento <dbl>, Reddito_Lavoro <dbl>, Liberta_civili <dbl>,
## #   Surriscaldamento_globale <dbl>, Reddito_al_quadrato <dbl>
```