Week 6 Tutorial: Data Wrangling in R

POP77001 Computer Programming for Social Scientists

Module website: tinyurl.com/POP77001

Loading the dataset

- Replace filepath with the location of the file on your computer

Loading the dataset

- Replace filepath with the location of the file on your computer

In [2]:
```
library("readr")
library("dplyr")
```

```
Attaching package: ‘dplyr’


The following objects are masked from ‘package:stats’:

    filter, lag


The following objects are masked from ‘package:base’:

    intersect, setdiff, setequal, union
```

In [3]:

```r
PATH <- "../data/kaggle_survey_2021_responses.csv"

# As the header of this dataset is composite (consisting ot 2 rows)
# we start by reading in the first 2 rows and then using the header
# of that 'header' dataset for the actual full dataset
questions <- readr::read_csv(PATH, n_max = 2)
```

```
Rows: 2 Columns: 369

── Column specification ────────────────────────────────────────────────
────────────────────────────────────────────────────────────────────────
────────────────────────────

Delimiter: ","
chr (369): Time from Start to Finish (seconds), Q1, Q2, Q3, Q4,
Q5, Q6, Q7_P...


i Use `spec()` to retrieve the full column specification for thi
s data.
i Specify the column types or set `show_col_types = FALSE` to qu
iet this message.
```

In [4]:
```
kaggle2021 <- readr::read_csv(PATH, col_names = names(questions), skip
```

```
Rows: 25973 Columns: 369

── Column specification ────────────────────────────────────────────────
────────────────────────────────────────────────────────────────────────
──────────────────────

Delimiter: ","
chr (360): Q1, Q2, Q3, Q4, Q5, Q6, Q7_Part_1, Q7_Part_2, Q7_Par
t_3, Q7_Part_...
dbl   (1): Time from Start to Finish (seconds)
lgl   (8): Q30_B_Part_1, Q30_B_Part_2, Q30_B_Part_3, Q30_B_Part
_4, Q30_B_Par...


i Use `spec()` to retrieve the full column specification for thi
s data.
i Specify the column types or set `show_col_types = FALSE` to qu
iet this message.
```

In [4]:
```
kaggle2021 <- readr::read_csv(PATH, col_names = names(questions), skip
```

```
Rows: 25973 Columns: 369

── Column specification ────────────────────────────────────────────────
────────────────────────────────────────────────────────────────────────
────────────────────────

Delimiter: ","
chr (360): Q1, Q2, Q3, Q4, Q5, Q6, Q7_Part_1, Q7_Part_2, Q7_Par
t_3, Q7_Part_...
dbl   (1): Time from Start to Finish (seconds)
lgl   (8): Q30_B_Part_1, Q30_B_Part_2, Q30_B_Part_3, Q30_B_Part
_4, Q30_B_Par...


i Use `spec()` to retrieve the full column specification for thi
s data.
i Specify the column types or set `show_col_types = FALSE` to qu
iet this message.
```

In [5]:
```
head(kaggle2021, 1)
```

```
  Time from Start to Finish (seconds) Q1    Q2  Q3    Q4
Q5
1 910                                 50-54 Man India Bachelo
r’s degree Other
  Q6         Q7_Part_1 Q7_Part_2 Q7_Part_3 … Q38_B_Part_3 Q38_B
_Part_4
```

```
1 5-10 years Python    R          NA         … NA            NA
  Q38_B_Part_5 Q38_B_Part_6 Q38_B_Part_7 Q38_B_Part_8 Q38_B_Par
t_9
1 NA           NA           NA           NA           NA
  Q38_B_Part_10 Q38_B_Part_11 Q38_B_OTHER
1 NA             NA             NA
```

In [6]:
```
questions[,1:10]
```

```
  Time from Start to Finish (seconds) Q1
1 Duration (in seconds)               What is your age (# year
s)?
2 910                                 50-54
  Q2
1 What is your gender? - Selected Choice
2 Man
  Q3
1 In which country do you currently reside?
2 India
  Q4
1 What is the highest level of formal education that you have a
ttained or plan to attain within the next 2 years?
2 Bachelor’s degree
  Q5
1 Select the title most similar to your current role (or most r
ecent title if retired): - Selected Choice
2 Other
  Q6
1 For how many years have you been writing code and/or programm
ing?
2 5-10 years
  Q7_Part_1
1 What programming languages do you use on a regular basis? (Se
lect all that apply) - Selected Choice - Python
2 Python
  Q7_Part_2
1 What programming languages do you use on a regular basis? (Se
```

```
lect all that apply) - Selected Choice - R
2 R
  Q7_Part_3
1 What programming languages do you use on a regular basis? (Se
lect all that apply) - Selected Choice - SQL
2 NA
```

Exercise 1: Summarise categorical variable

- Load the dataset (as local file)
- Consider country of residence reported by respondents (question Q3).
- Make sure you can select the column both using both it name and index
- Calculate the percentages of top 3 countries of residence in the sample

Dummy variables

- When analysing categorical data (particularly using it as indepedent variables in regression) it is common to contruct design matrices, where categorical variables are represented by 1's and 0's depending on whether it is true or not for a given observation.
- For example, gender of respondents in survey can be represented by this matrix below, where 1's indicate whether a given respondent is female and 0's if they are male:

$$\begin{matrix} female \\ \begin{bmatrix} 1 \\ 0 \\ 1 \\ \vdots \\ 1 \end{bmatrix} \end{matrix}$$

Dummy variables continued

- A more complex example would be when instead of having just two levels of a categorical (i.e. factor in R) variable, we have multiple different values that a variable might take.
- For instance, a variable like age group might be represented as follows:

$$\begin{matrix} 25-34 & 35-44 & 45-64 & 65+ \end{matrix}$$
$$\begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 1 \\ \vdots & \vdots & \vdots & \vdots \\ 0 & 0 & 0 & 0 \end{bmatrix}$$

Where the first row corresponds to a respondent who is between 25 and 34 years old, the second to someone between 35 and 44 and the third one to a participant who is older than 65. Note that the number of columns in this matrix is one lower than the number of levels of our imaginary categorical variable age. We are omitting the baseline (reference) category. You can see that we can establish belonging to this category from the information provided in the matrix. If the values in all columns are 0 (such as in the last

Exercise 2: Pivoting tables

Week 6: Assignment 2

- Functions and data wrangling in R
- Due by 12:00 on Monday, 24th October