Week 6 Tutorial: Data Wrangling in R

POP77001 Computer Programming for Social Scientists

Loading Dataset

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

```
library("readr")
library("dplyr")
```

```
PATH <- "../data/kaggle_survey_2022_responses.csv"

# As the header of this dataset is composite (consisting of 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 = 1)
```

```
kaggle2022 <- readr::read_csv(PATH, col_names = names(questions), skip = 2)
```

```
head(kaggle2022, 1)
```

```
# A tibble: 1 × 296
  `Duration (in seconds)` Q2    Q3    Q4    Q5    Q6_1  Q6_2  Q6_3  Q6_4  Q6_5
                    <dbl> <chr> <chr> <chr> <chr> <chr> <chr> <chr> <chr> <chr>
1                     121 30-34 Man   India No    <NA>  <NA>  <NA>  <NA>  <NA>
# ℹ 286 more variables: Q6_6 <chr>, Q6_7 <chr>, Q6_8 <chr>, Q6_9 <chr>,
#   Q6_10 <chr>, Q6_11 <chr>, Q6_12 <chr>, Q7_1 <chr>, Q7_2 <chr>, Q7_3 <chr>,
#   Q7_4 <chr>, Q7_5 <chr>, Q7_6 <chr>, Q7_7 <chr>, Q8 <chr>, Q9 <chr>,
#   Q10_1 <chr>, Q10_2 <chr>, Q10_3 <chr>, Q11 <chr>, Q12_1 <chr>, Q12_2 <chr>,
#   Q12_3 <chr>, Q12_4 <chr>, Q12_5 <chr>, Q12_6 <chr>, Q12_7 <chr>,
#   Q12_8 <chr>, Q12_9 <chr>, Q12_10 <chr>, Q12_11 <chr>, Q12_12 <chr>,
#   Q12_13 <chr>, Q12_14 <chr>, Q12_15 <chr>, Q13_1 <chr>, Q13_2 <chr>, …
```

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

```
# A tibble: 1 × 10
  `Duration (in seconds)` Q2    Q3    Q4    Q5    Q6_1  Q6_2  Q6_3  Q6_4  Q6_5
```

Exercise 1: Summarise Categorical Variable

- Load the dataset (as a local file).
- Consider country of residence reported by respondents (question Q4).
- 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 Variable

- When analysing categorical data (particularly using it as independent variables in regression) it is common to construct 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:

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

- This process of replacing actual labels (e.g. 'female' and 'male' in the example above) with binary values is called creating a dummy variable in statistics and one-hot encoding in computer science.

Dummy Variables

- 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:

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

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 row above), we can be sure that this observation is from a respondent who is in age group 18-24.

Exercise 2: Pivoting Tables

- Now let's construct such design matrix with dummy variables for respondents' age group in Kaggle survey.
- First, check what levels does the variable age group take (question Q2).
- Since we are making use of only a small portion of the data in this exercise, make the survey dataset more manageable by subsetting the columns Q2 to Q5.
- Check the function `model.matrix()` from base R and apply it to the dataset to get a design matrix (you need to specify formula as the first argument).
- This might be not the most usual example of pivoting data frame (as while the number of columns increases, the number of rows remains the same), but it gives you a sense of what it can entail.
- To simplify working with the dataset, let's also create a unique id for each respondent (you can use `seq_along()` function in combination with any other variable to do so).
- Finally, use `pivot_wider` function from `tidyr` package to create a separate column for each age group.
- If the original pivoting produced columns that are populated by values of the categorical variable and `NA`'s, use `mutate` function to replace them with 0's and 1's.
- Finally, use `pivot_longer` function to convert this representation of the dataset back into its original form.
- You might also need to use `dplyr::filter()` function to remove redundant rows.

Week 6: Assignment 2

- Functions and data wrangling in R
- Due by 12:00 on Monday, 27th October