Week 5 Tutorial Worksheet

AY23/24 Semester 2

Submission: End of tutorial day

Question 1. International visitor arrivals in Singapore

The file `tourist.xlsx` was downloaded from the Singapore Department of Statistics.

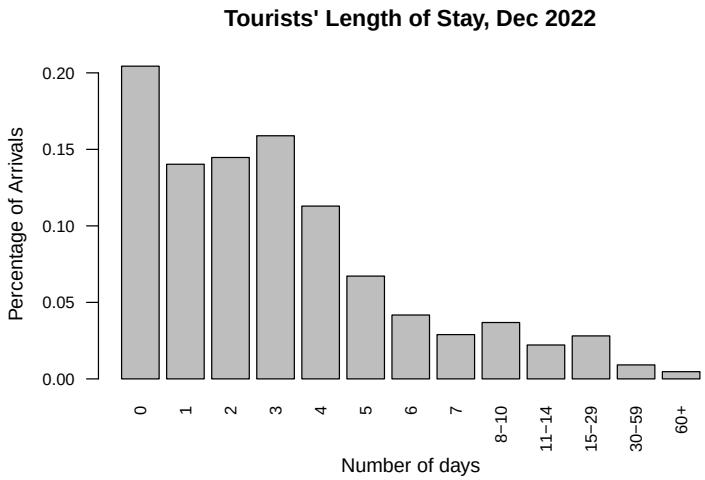
1. Read the data in as `qn1_1`. Since the data flows in the middle of the spreadsheet, use the `range` or `skip` argument to specify the appropriate data range. This will simplify the data cleaning process in the subsequent steps. *Hint: After reading in the right range, the first few rows of* ***`qn1_1`*** *would read as:*

```
head(qn1_1)
```

```
## # A tibble: 6 x 7
##   `Data Series`           `2022 Dec` `2022 Nov` `2022 Oct` `2022 Sep` `2022 Aug`
##   <chr>                        <dbl>      <dbl>      <dbl>      <dbl>      <dbl>
## 1 Total International Vi~     815654     775804     805979     691440     685839
## 2 Under 1 Day (Number)        166698     146629     136573     134931     125981
## 3 1 Day (Number)              114412     108266     102074      97676      96875
## 4 2 Days (Number)             118019     124149     117744     104369     101746
## 5 3 Days (Number)             129601     132535     140122     114944     112391
## 6 4 Days (Number)              92109      86753      98830      76710      74885
## # i 1 more variable: `2022 Jul` <dbl>
```

2. Examine the data frame and rectify any data issues you identify using base `R` syntax. Save your resulting data frame as an object named `qn1_2`.

3. Recreate the bar plot to show the distribution of tourist arrival length in December 2022.



Question 2. YRBSS questionnaires

The file `yrbss.csv` contains a subset of data retrieved from the Youth Risk Behavior Surveillance System (YRBSS). You can read the data documentation here.

In the following questions, we use the data to practice our data manipulation skills **using functions in base `R`**.

1. Read the data into `R` as an object named `yrbss`. Conduct the following tasks and overwrite `yrbss` with the resulting data frame.

   - Remove rows with missing values (if any).
   - Remove duplicated rows (if any).
   - Rename the columns `record` as `id`, and `stweight` as `weight_kg`.
   - Convert the `grade` variable into numeric.

2. Continue working on the cleaned data in `yrbss`. Subset female youth with BMI lower than 15 and then extract the following columns: `id`, `age`, `race4`, `weight_kg`, and `bmi`. Store your result in an object named `qn2_2`.

3. Create a column `height_m` in `qn2_2`, computed based on the BMI formula:

$$BMI = \frac{weight(kg)}{height^2(m)}$$

   Hint: After this, the first few rows of ***`qn2_2`*** *would read as the following:*

```
head(qn2_2)
```

```
##           id          age                     race4 weight_kg     bmi height_m
## 1123 1313049 16 years old                     White     40.82 14.9936 1.649998
## 2480  635432 13 years old          Hispanic/Latino     27.67 13.1605 1.450001
## 3856  771312 16 years old          All other races     41.73 14.7853 1.679999
## 4179  932782 15 years old                     White     44.45 14.8518 1.730001
## 6001 1314901 17 years old Black or African American     45.36 14.8114 1.750002
## 6065  637858 16 years old          Hispanic/Latino     52.16 13.5777 1.959998
```

Requirements

- After you answer all questions in the `Rmd`, file, hit the **Knit** button. Make sure your `Rmd` can knit to HTML.
- The code in your `Rmd` file should create the following data frames: `qn1_1`, `qn1_2`, `yrbss`, and `qn2_2`
- The knitted HTML file should contain a bar plot for Question 1.
- **Submit your `Rmd` file to Canvas after your tutorial session.**
- Reach out to your tutor as soon as possible if you are unsure about our submission requirements.