Week 7 Tutorial Worksheet

AY23/24 Semester 2

No submission required

Question 1. International vistor arrivals (revisited)

In this question, we will revisit the tourist data in Week 5. The file `tourist.xlsx` contains monthly international visitor arrivals in Singapore from July to December in 2022, originally retrieved from the Singapore Department of Statistics.

1. Read the data into `R` as `qn1_1` and transform it into a tidy format. Transform the data so that it follows the structure below. The first variable `k` is an unique identifier for each duration category, $k = 1, 2, .., 13$.

```
glimpse(qn1_1)
```

```
## Rows: 78
## Columns: 5
## $ k        <int> 1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 1, 2, 3, 4, 5, 6, ~
## $ duration <chr> "Under 1 Day (Number)", "1 Day (Number)", "2 Days (Number)", ~
## $ year     <int> 2022, 2022, 2022, 2022, 2022, 2022, 2022, 2022, 2022, 2022, 2~
## $ month    <ord> Jul, Jul, Jul, Jul, Jul, Jul, Jul, Jul, Jul, Jul, Jul, Jul, J~
## $ arrivals <dbl> 108581, 88056, 91996, 109801, 75171, 44932, 29829, 23829, 305~
```

2. Compute the **month-on-month** change for tourist arrivals for the 13 arrival duration categories from August to December in 2022. For instance, in August, we can compute for each duration category `k`:

$$mom\_change_k = \frac{\text{Aug}_k - \text{Jul}_k}{\text{Jul}_k} \qquad k = 1, 2, ..., 13$$

   Store the result as a new column named `mom_change` in `qn1_2`. The columns of `qn1_2` should follow the structure of the following:

```
head(qn1_2, 2)
```

```
## # A tibble: 2 x 6
##       k duration              year month arrivals mom_change
##   <int> <chr>                <int> <ord>    <dbl>      <dbl>
## 1     1 Under 1 Day (Number)  2022 Aug     125981       16.0
## 2     2 1 Day (Number)        2022 Aug      96875       10.0
```

Question 2. NBER research papers

In this question, we will use data from a TidyTuesday project about the National Bureau of Economic Research (NEBR) research papers from 1973 to 2021.

The data description is available in the link below.

> https://github.com/rfordatascience/tidytuesday/blob/master/data/2021/2021-09-28/readme.md.

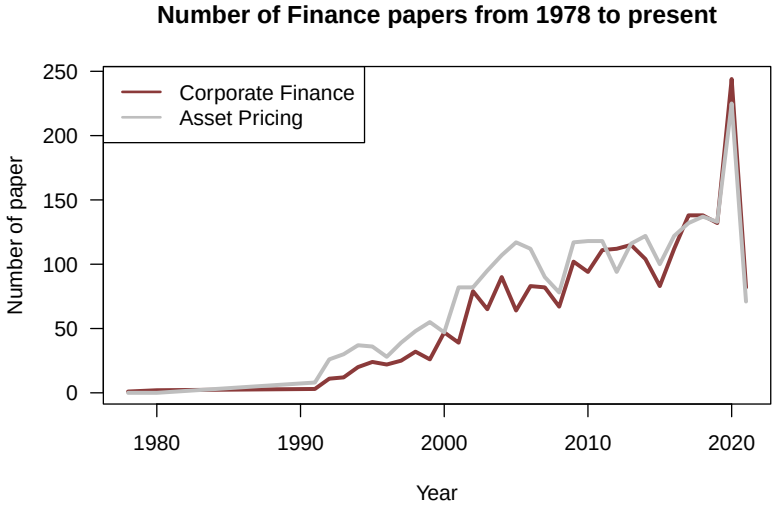
There's also a blog post here that discusses the data. The analyses might be advanced for us at the moment, but it could be helpful to see how others have explored the data.

1. There are five data sets in total. Follow the instructions on the web site and read in all of them.
2. Use the tables to create a data frame that contains all **Finance** papers and their program descriptions. After that, create a table that counts the number of papers in each sub-categories in each year. Save the resulting data frame as a new object named `qn2_2`.

```
glimpse(qn2_2)
```

```
## Rows: 64
## Columns: 4
## $ year             <dbl> 1978, 1980, 1991, 1991, 1992, 1992, 1993, 1993, 1994,~
## $ program_desc     <chr> "Corporate Finance", "Corporate Finance", "Asset Pric~
## $ program_category <chr> "Finance", "Finance", "Finance", "Finance", "Finance"~
## $ n                <int> 1, 2, 8, 3, 26, 11, 30, 12, 37, 20, 36, 24, 28, 22, 3~
```

3. From `qn2_2`, we observe that the first Finance research paper started in 1978 and started to grow in the early 1990s. We can visualize the growth across time in a line chart.
   - Prepare the data and recreate, as much as you can, the graph below.
   - Additionally, explore the data to understand the reason for the apparent decrease in the number of Finance papers in 2021. Briefly discuss your findings in an `Rmd` text section titled "Question 2.3".



4. Use the tables to create a data frame that contains all papers and their authors. Save the data frame as a new object called `qn2_4`. In this data frame, each row should represent a paper-author pair.

5. Inspect a few rows in `qn2_4`. We observe that some papers have multiple authors (e.g., the paper `w0008` has two authors: *Merle Yahr Weiss* and *Robert E Lipsey*), and some authors have written more than one paper (e.g., `w0004` and `w0009` are written by the same author: *Lee A Lillard*).

   Conduct the following tasks, and save the resulting data frame as `qn2_5`.

   - Create a new variable called `decade` to represent the decade in which each paper was written.
   - For each paper, calculate the total number of authors. After that, compute the average number of authors per paper by each decade and save the result in a new column named `avg_n_authors`.

   The final data frame should follow the structure of the following:

```
head(qn2_5, 3)
```

```
## # A tibble: 3 x 2
##   decade avg_n_authors
##   <fct>          <dbl>
## 1 1970s           1.43
## 2 1980s           1.58
## 3 1990s           1.87
```

Requirements

- After answering 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 (tibbles): `qn1_1`, `qn1_2`, `qn2_2`, `qn2_4`, `qn2_5`.
  - A graph for Question 2.3.
  - An `Rmd` text section titled "Question 2.3".
- **You do not need to submit your worksheet this week.** If you are unsure about any of the requirements above, please reach out to your TA as soon as possible.