IE6600: Computaton and Data Visualisation

Homework 3

Prof. Mohammad Dehghani

Assignment Guidelines

1. Students need to complete the assignment **individually**.
2. All the assignments are required to be done in RStudio.
3. Provide necessary comments using '#' for better understanding of your script.
4. The code should **follow tidyverse style guide** (https://style.tidyverse.org/index.html) The tidyverse style guide has style standards for naming objects, indentation and how to write long lines of codes to name a few
5. If you take help from any external sources, please mention that in the reference. Violating academic integrity policies may include zero credit on the work.
6. The assignment report needs to include the following sections:

- **Problem statement:** A brief about your understanding on the assignment questions (maximum 3 lines)
- **Result:** What were your finding after creating the code and running it in R. This section may include:
  - Graphs / charts / plots
  - Final data frame for your result
  - Results obtained
- **Conclusion:** What were the statistical inferences and observations from the results obtained.
  - Students are not required to include codes in reports.

Deliverables:

1. **Please submit a **.rmd* file which includes your code and can be knit into a PDF(recommended)** or submit a **.zip* file including the following items
   i. **R script** (just 1 file including all your codes)
   ii. **HW Report**: Report with a maximum length of 10 pages including all appendices, tables, and graphs if any.
2. All of the above mentioned files have to be labeled as: 'HW # - IE 6600 – Sec # - <Student Name>'
3. Submit your HW deliverables via CANVAS

Task 1

From the " wine_data.csv " answer the following questions using data wrangling functions from relevant packages.

- Write a code to calculate the frequency count of " variety " variable from the dataset. Display top 10 variety by count
- Write a code to calculate the average points by country
- Which province has the highest average price?
- Which province in the US has the highest average price?
- From the " designation " variable calculate the number of 20 year old wine

Task 2

Write a code to compute the number of farmers market by states based on the Month in Season1Date. Present the total number of active farmer's markets by month from the above result.

Sample output:

The below table should only be considered as a reference as how the output should look like. Students need to generate the entire long form table. Leave rows which have 'NA', '-'.

States	January	February
Vermont	23	2162
Ohio	3523	4366
Maine	26574	2605
California	854	3086
New York	976	3397

Task 3

What are the monthly active farmer's markets in different cities in the state of California based on the month in updateTime. Ignore rows if they do not have month details.

Cities	Month	Number of Farmer's Markets
City_Name_1	3	210
City_Name_2	5	561
City_Name_3	7	782
City_Name_4	6	153
City_Name_5	12	198

Task 4

The attached dataset "airlines_delay" has information related to aircraft delays of airlines operating within the United States. The variable description is as follows:

- **arr_flights** : Number of aircraft arriving
- **arr_del15** : Number of aircraft that were delayed beyond 15 minutes. A delay incident is recorded if an aircraft gets delayed beyond 15 mins.
- **carrier_delay (in mins)**: The cause of the cancellation or delay was due to circumstances within the airline's control (e.g. maintenance or crew problems, aircraft cleaning, baggage loading, fueling, etc.
- **weather_delay (in mins)**: Significant meteorological conditions (actual or forecasted) that, in the judgment of the carrier, delays or prevents the operation of a flight such as tornado, blizzard or hurricane
- **nas_delay (in mins)**: Delays and cancellations attributable to the national aviation system that refer to a broad set of conditions, such as non-extreme weather conditions, airport operations, heavy traffic volume, and air traffic control
- **late_aircraft_delay** : A previous flight with same aircraft arrived late, causing the present flight to depart late
- **security_delay (mins)**: Delays or cancellations caused by evacuation of a terminal or concourse, re-boarding of aircraft because of security breach, inoperative screening equipment and/or long lines in excess of 29 minutes at screening areas
- **arr_delay (in mins)**: Sum of all the aforementioned delays

Compute the total number of late aircraft delays (in mins) for each carrier. Use a bar plot to show the top 10 carriers by late aircraft delays and create a stacked bar plot by adding additional variable " year " to the bar plot. Output should look like: