COMM 337 Group Project 1

(Due **March-31-2023 at 11:59pm**, Vancouver time)

Instructions

1. Please create your project report using Jupyter notebook. In your notebook, please indicate the group members (full name, email, student ID) in the first cell.

2. In your Jupyter notebook, use Markdown cells appropriately when discussing your analysis in texts.

3. After you complete the analysis, click on "Kernal / Restart & Run All" to ensure there is no error in your report. Use Markdown cells to clearly format your report.

4. You cannot share code (fully or partially) with other groups. Please add appropriate references in your submission.

5. Submit your **.ipynb** file on Canvas. Late submissions will not be accepted. (Note: This time, please submit the .ipynb **file** including all the code and analysis results. **Don't** submit .py file.)

Part A. Nobel Laureates [50 points]

The Nobel Prize is not a single prize, but five separate prizes that, according to Alfred Nobel's 1895 will, are awarded "to those who, during the preceding year, have conferred the greatest benefit to humankind" (Wikipedia 2022). Nobel Prizes are perhaps the worlds' most prestigious awards in their respective fields. In this project, we will use a dataset (nobel.csv) from the Nobel Foundation to know more about Nobel laureates. (Note: In this project, we need to work with date and time when analyzing the data. Pandas provides many useful methods to handle datetime data type. You may explore some of these methods and apply in this project: https://towardsdatascience.com/working-with-datetime-in-pandas-dataframe-663f7af6c587)

1. Read the data in your notebook and display the first 8 rows. How many columns and rows are there in this dataset? What are the column names?

2. All of the first 8 winners were from Europe. But that was back in 1901. Looking at all winners in this data, which gender and which country is the most commonly represented? (For country, we will use the birth_country of the winner.)

3. For the dominant country that you found in question 2, calculate the proportions of winners from that country in each decade. Make some plots to visualize your results. In which decade did the proportion reach the highest?

For the following questions, please exclude organizational laureates and focus on individual laureates. (Hint: laureate_type == 'individual')

4. Is there any gender imbalance in this data? How significant is that? Calculate the proportion of female laureates in each decade. Visualize your results and discuss.

5. For the gender imbalance that you found in question 4, is it better or worse within specific prize categories? Visualize your results for each category and discuss. Which of them has the largest gender imbalance? Which has shown some positive trend over the decades?

6. Are there any people who have won the Nobel Prize more than once? Who are they?

7. Who are the oldest and youngest people ever to have won a Nobel Prize? How old were the winners generally when they got the prize? Show the summary statistics, and plot the distribution of the age of winners.

8. For your results in question 7, does the average age of winners differ across each category? For each category, show the summary statistics and plot the distribution of the age of winners. Which categories have the largest and smallest average age of winners?

9. Make some plots to visualize the time trend of the average age of winners in each specific category per decade. What do you find?

10. Repeat the analysis in question 9, but with lifespan instead of age. What do you find?

Part B. COVID-19 [50 points]

The COVID-19 pandemic has created unprecedented disruption of our life. In this project, we will use a dataset (covid.csv) from the Government of British Columbia to generate some insights from this data.

11. Read the data (covid.csv) in your notebook and display the first 8 rows. How many columns and rows are there in this dataset? What are the column names?

12. Create a new column Month to represent the month of the Reported_Date in the data. Print the first 8 rows of updated dataframe.

13. Create a new dictionary which contains the months as keys and the number of cases for corresponding months as values. Show the dictionary. What is the largest number of cases in the dictionary?

14. Make some plots to visualize the time trend of the number of cases in every month. What do you find?

15. Is there any gender imbalance in this data? Visualize the time trend of the number of cases for each gender and discuss.

16. Create new dictionaries which contain the months as keys and the number of cases for corresponding months per each gender as values. Show the dictionaries. Find the month with the smallest number of female cases.

17. Is there any imbalance among different regions in this data? Make some plots to visualize the difference among the regions in terms of reported cases in every month, and discuss your results.

18. Calculate the cumulative reported cases in every month for each region. Print the first 8 rows. Visualize the difference among the regions in terms of cumulative reported cases in every month, and discuss your results.

19. Open question: what else can you find from this data?

20. After you complete the analysis, click on “Kernal / Restart & Run All” to ensure there is no error in your report. Use Markdown cells to clearly format your report.