CptS 475/575: Data Science, Fall 2024

Assignment 4: Joins (Relational Data) and Visualization

Release Date: September 23, 2024 **Due Date:** October 2, 2024 (11:59 pm)

General instruction: This assignment has **three problems.** The first problem is on Joins (relational data from the data wrangling series of topics), and the second and third problems are on visualization.

Your solution will be submitted as a single **PDF (or HTML) file**. You are encouraged to use R Markdown or a similar tool (like Jupyter) to prepare your file.

Problem 1 (50 pts). This problem will involve the Lahman dataset (including the tables Batting, Teams, Salaries, and Managers). It is available in R by loading the **Lahman** library using the following command:

```
library(Lahman)
```

Alternatively, you can download the csv files from the Modules page on Canvas. The files are *Batting.csv, Teams.csv, Salaries.csv,* and *Managers.csv*. You can use *Lahman_Desc.txt* (also from Modules) to check the column descriptions for each dataset.

We will first use joins to search and manipulate the dataset, then we will produce a flight count visualization.

a) (10 pts) Filter the dataset (using a **left join**) to display the `playerID`, `yearID`, `teamID`, `stint`, `G` (games played), `HR` (home runs), and `salary` for all players who hit more than 30 home runs in a single season and played for a team in New York (teamID "NYA" or "NYN") between 2010 and 2020. How many players match these criteria?

b) (10 pts) What is the difference between the following two joins? Do not show the result of these anti_joins in your submission.

```
anti_join(Salaries, Batting, by = c("playerID" = "playerID"))
anti_join(Batting, Salaries, by = c("playerID" = "playerID"))
```

What is the difference between semi_join and anti_join? Provide an example using the Salaries and Batting tables.

c) (10 pts) Select the `teamID`, `yearID`, and the total number of runs batted in (`RBI`) for each team in the American League (AL) for the year 2015 (using one or more inner joins with the Teams and Batting tables). How many total home runs were hit by American League teams in 2015?

d) (10 pts) Using the Managers and Teams tables, determine the number of seasons each manager managed a team. Use group_by and count to get the number of unique `managerID` and `teamID` combinations. How many unique combinations of `managerID` and `teamID` are present? Are there any players with unusually high number of years as a manager?

e) (10 pts) Using the provided template as a start, produce a horizontal bar plot that shows the number of wins for the top 10 teams in 2019. Adjust the axis labels to clearly represent the teams and the number of wins. Add a meaningful title to the plot, and include the number of wins as text on each bar for clarity.

```
Teams %>%
  filter(yearID == 2019) %>%
  select(teamID, W) %>%
  ggplot(aes(x = reorder(teamID, W), y = W)) +
    geom_bar(stat = "identity", fill = "steelblue") +
    coord_flip()
```

Problem 2 (30 pts). The goal of this problem to create a visualization of the US map showing the states/territories and the number of presidential votes received during an election year. For this task, you will work with the us-presidents.csv dataset. The dataset can be found on the Modules page on Canvas.

There dataset consists of 612 observations of 4 variables: year, state, state_po, office, totalvotes.

For this question, you will create **two** visualizations of the US map for **two** presidential years of your choice coloring the states or sizing the point/marker for the states according to the number of total votes received from that state for the presidential election.

Compare both maps and comment on any observations.

You are free to choose any mapping tool you wish to produce this visualization. Try to make your visualization as nice looking as possible. You can use the state column directly to visualize the observations or you could get the coordinates for each state (depending on the tool and your visualization). Research how this can be done and use what you find. The dataplusscience.com website has some blogs about mapping that you may find useful. After you have coordinates you can use different methods for mapping. You can use packages available in R or Python. Another simple method is probably through https://batchgeo.com/features/map-coordinates/ . However, you can also use d3 to map the locations, if you want to learn something that you could use for other projects later.

Problem 3 (20 pts). Create a word cloud for an interesting (relatively short, say a couple of pages) document of your own choice. Examples of suitable documents include: summary of a recent project you are working or have worked on; your own recent Statement of Purpose or Research Statement or some other similar document.

You can create the word clouds in R using the package called *wordcloud* or you can use another tool outside of R such as *Wordle.* If you do this in R, you will first need to install *wordcloud* (using install.packages("wordcloud")) and then load it (using library(wordcloud)). Then look up the documentation for the function called *wordcloud* in the package with the same name to create your cloud. Note that this function takes many arguments, but you would be mostly fine with the default settings. Only providing the text of your words may suffice for a minimalist purpose.

You are welcome (and encouraged) to take the generated word cloud and manipulate it using another software to enhance its aesthetic. If you have used Wordle instead of R, Wordle gives you functionalities to play with the look of the word cloud you get. Experiment till you get something you like most.

Your submission for this would include the figure (cloud) and a brief caption that describes the text for the cloud. For example, it could be something like ``Jenneth Joe's Essay on Life During Pandemic, written in June 2021."