

CptS 475/575: Data Science, Fall 2023

Assignment 4: Joins (Relational Data) and Visualization

Release Date: September 22, 2023 **Due Date:** September 29, 2023 (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 **PDF (or HTML) file**. The file should include the code as well as the plots and results. 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 **nycflights13** dataset (including the tables **flights**, **airlines**, **airports**, **planes** and **weather**), which we saw in class. It is available in both R and Python, however R is recommended for at least the visualization portion of the question. You can get more information about this package on github at

<https://github.com/tidyverse/nycflights13>

You can access the data tables in R by loading the nycflights13 library using the following:

```
library(nycflights13)
```

Alternatively, you can download the csv files from the Modules page on Canvas. The files are *flights.csv*, *weather.csv*, *planes.csv*, *airports.csv*, *airlines.csv*.

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

- (10 pts) Filter the dataset (using a **left join**) to display the tail number, year, month, day, hour, origin, and humidity for all flights heading to Tampa International Airport (TPA) after 12pm on November 1, 2013. How many flights happened during the given time frame that day?
- (10 pts) What is the difference between the following two joins?

```
anti_join(flights, airports, by = c("dest" = "faa"))  
anti_join(airports, flights, by = c("faa" = "dest"))
```
- (10 pts) Select the origin and destination airports and their latitude and longitude for all flights in the dataset (using one or more **inner joins**). How many flights are there in your result?
- (10 pts) Use **group_by** and **count** to get the number of flights to each unique carrier/dest combination. How many unique combination of carrier/dest are present?
- (10 pts) Produce a map that sizes each destination airport by the number of incoming flights. You may use a continuous scale for the size. Here is a code snippet to draw a map of all flight destinations, which you can use as a starting

point. You may need to install the maps packages if you have not already. Adjust the title, axis labels and aesthetics to make this visualization as clear as possible.

```
flights %>%  
  left_join(airports, c("dest" = "faa")) %>%  
  ggplot(aes(lon, lat)) +  
    borders("state") +  
    geom_point() +  
    coord_quickmap()
```

Problem 2 (30 pts). The goal of this problem to create visualizations of the US map showing the states/territories and the Gross Domestic Product (GDP) of each state/territory. For this task, you will work with the data table for GDP by States in the US provided by BEA (Bureau of Economic Analysis). The dataset can be found at

<https://apps.bea.gov/regional/histdata/>

However, we are interested only in 5 columns in this dataset: GeoName, 2019, 2020, 2021, 2022

A subset of the data (us_states_gdp.csv) containing only these 5 columns has already been made available on Canvas, in the Datasets module. The dataset contains 51 observations of 5 variables:

Name	Description
state	GeoName (State/ Territory/Federal Entity)
gdp_2019	GDP in the year of 2019
gdp_2020	GDP in the year of 2020
gdp_2021	GDP in the year of 2021
gdp_2022	GDP in the year of 2022

Create visualizations of the US map coloring the states or sizing the point/marker for the states according to the GDP for each state (one map per year). Compare the GDP of different states for all the years using the maps you generated (we recommend that you maintain a constant scale for showing the GDP in all the four maps; *hint: check min and max values in the dataset for the scale*).

You are free to choose any mapping tool you wish to produce this visualization. You can use packages available in R or Python. For example, the usmap package in R is convenient for plotting the US map. Alternatively, 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. 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.

Try to make your visualization as nice looking as possible.

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 cloud 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. An example of caption could be something like “Jenneth Joe's Essay on Life During Pandemic, written in June 2021.”