Homework 4

Instructions

For this homework you will create a python notebook (`.ipynb` file) using Google Colab and submit a link to your notebook. This can be done by clicking on the 'share' icon in the top right. Please make sure to change the sharing settings so that anyone with the link can see the notebook. No edits should be made after the due date!

The purpose of this homework is to practice summarizing data using `pandas` and `matplotlib` and do some more advanced function writing. Most homework assignments will have a part that pushes you beyond what was in the lectures! Learning to search for the right questions and browsing stackoverflow are really life skills that we should hone :)

Summarizing Student Data Graphically

For this part, we'll revisit the StudentData.txt data that comes from the UCI machine learning repository. Information about the variables in the dataset can be found here. I want you to look at the math scores data set.

You should read up on the variables. The dataset is generally about math scores (G1, G2, G3) for students from two different schools. They also measure a bunch of things about the students' home life. Hopefully you can make some interesting connections!

Bring in Homework 3 Code

- Copy your code and markdown cells from homework 3 that read in the data and summarized it numerically.
- We were essentially starting an EDA there.
- **Our goal is to now add to this basic EDA by including graphs that describe the variables.**

Task 1

This data has many categorical variables and a few numeric. You should add the following:

Categorical variables

- Create a stacked bar graph and a side-by-side bar graph. Give relevant `x` and `y` labels, and a title for the plots.

Numeric variables (and across groups)

The numeric variables are age, absences, and the three test grades variables (G1, G2, and G3).

- Create a histogram, kernel density plot, and boxplot for two of the numeric variables across one of the categorical variables (that is, create graphs that can compare the distributions across the groups). For at least one of the kernel density plots across groups, make sure that the graphs are overlayed on the same plot. Add appropriate labels and titles.
- Create two scatterplots relating G3 to other numeric variables (G3 on the y-axis). Color the points by a categorical variable in each. Add appropriate labels and titles.

After each summary or graph, you should discuss what is interesting about it or what it tells you!

Plotting the NFL Data

For this part we'll read in the NFL Box Score data that we read in class videos. The data is available here.

You may not be familiar with (American) football, but each row of this dataset represents information about one particular game. The most important thing is the score for the home and away teams (`AQ1`, ..., `AFinal`, `HQ`, ..., `HFinal`). Other variables like yards gained (passing or rushing or combined) can be good indicators of score. Things like turnovers, penalties, etc. can also be indicators of how the came went.

Task 1: Read in the data

- You can either read this dataset from the URL or download it and read it in locally. Check out the first few observations of the data.
- Convert the `homeTeam`, `awayTeam`, `day`, `stadium`, `startTime`, `toss`, `roof`, and `surface` to `category` type variables.

Task 2: Summarize the data

- I want you to look at some trends during the regular season. This means you should remove any data where the week is not 1 through 17.
- You should summarize some of the variables grouped by season and week, season alone, and week alone (three different scenarios).
  - Produce some common numeric summaries of variables across these different groups.
  - Similarly, produce some common plots over time (mostly line plots I'd think!)
- Write at least one function that can be used to easily create a plot for this data.
  - For instance, you might have a function that takes in a numeric variable and a statistic (or a categorical variable, etc.). Then the function plots the average (or some other statistic) of the numeric variable across the seasons.
  - This is just an example! You get to decide what might be useful to do here.

After each graph, you should discuss what is interesting about it or what it tells you!

Submission

Now you can grab the 'share' link and submit that! Please make sure to change the sharing settings so that anyone with the link can see the notebook. Good luck and let us know if you run into issues.