CICS 397A – Fall 2020

Homework 2

Due Monday, October 19th at 11:59pm ET

You are encouraged to discuss the assignment in general with your classmates, and may optionally collaborate with one other student. If you choose to do so, you must indicate with whom you worked. Multiple teams (or non-partnered students) submitting the same code or output will be considered plagiarism.

Code must be written in a reasonably current version of Python (>3.0). You are free to use Python's standard modules for data structures and utilities, as well as the modules mentioned in class (pandas, seaborn, etc.).

Data Exploration

The goal of this assignment is to give you experience doing exploratory data analysis and creating plots in Python. The precise details of what you produce will vary with the type of data you're working with, so there's a lot of choice involved. Our intention is for you to work further with the data you collected in Homework 1, but that is not required – you are welcome to modify the data set (for instance, adding more rows or additional columns), or use a different data set altogether.

In addition to your writeup, you should submit all code used to calculate your summary and produce your plots.

1. Data Summary

Your first task is to write up a short (4-6 sentence) summary of your data set. You should include relevant details such as:

- The types of the variables
- Details about the prevalence of missing values
- Summary statistics for each column. For example, numeric attributes you might include means, medians, ranges, etc.; for categorical attributes, modes, uniques, etc.
- Any other interesting findings

2. Plottin' Fools

Your next job is to make some plots that highlight interesting aspects of your data. Choose at least **four** from the list below, or come up with your own ideas (this website has a ton of examples for inspiration).

Plotting ideas:

- Histogram and/or density curve showing the distribution of a column or columns
- Same as above, but showing different curves separated by a second attribute
- Scatter plot based on the values of two columns
- Same as above, but color the dots based on a third attribute
- Line plots showing values over time
- Bar plot of showing value counts for a categorical attribute
- Map showing locations for location data
- Heat maps showing associations between columns
- Sankey diagrams for categorical data
- Venn diagrams for categorical data
- Any of the above with numeric data binned into a new categorical attribute
- Word clouds generated by text data
- Bar plot showing word or bigram frequencies for text

You should produce all of your plots using `seaborn`, `matplotlib`, or another Python plotting library. If you have questions about whether a particular type of plot is appropriate for your data, come talk to us in office hours or post on CampusWire.

Some helpful links:
https://www.data-to-viz.com
https://matplotlib.org/
https://seaborn.pydata.org/

Requirements and Grading

You must create a Python script called `exploration.py` which contains the code for reading in your data from csv, calculating your summary statistics, and producing your plots. You may also store your code in a Jupyter notebook if you'd like.

Think of this as a mini-project. Maximum points will be awarded for plots that are well-produced, revealing, and creative.

What to Submit

You should submit:

- Your code, contained in a file called `exploration.py` (or `exploration.ipynb` if you use a notebook).
- The data file that you used called `data.csv`
- A `summary.txt` (or `summary.pdf`) containing your data summary from part 1, along with descriptions of your plots from part 2
- A collection of at least four plots, saved as pdf files