CPA01 - Python Data Analysis II

due Friday 3/11 11am on mastery.cs.brandeis.edu
CS103a-Programs PIN:7297444

Motivation

Data Science is one of the fastest growing areas in Computer Science and indeed in the world workforce. The ability to find and analyze large datasets is becoming an important component of many jobs and is in general an important skill for informed citizens. The goal of this assignment is to give you the opportunity to use the pandas package to load and analyze a large interesting dataset of your choosing.

What to do

1. create a git repository called cs103a-cpa01 with a README.md file and clone it to your computer, make sure to add a .gitignore file which should have .DS_Store in it, at least.
2. find a large dataset online (say at least 10,000 data items) and copy it into your repository (if it is too large, e.g. gigabytes, then add it to the .gitignore so it won't go into your repository)
3. create a jupyter notebook called cpa01.ipynb which will hold your work
    a. create a header cell in which you
        - describe the dataset
        - give the URL of where to find the dataset and explain how to download it
        - give at least two interesting questions you have about the data
    b. load the data into a pandas dataframe
    c. use pd.describe() to get a rough overview of the data
    d. analyze your data using the following features
        - print the array of columns and the index array
        - create some simple plot of part of the data
        - create a pivot table and plot some data from that pivot table
        - use the groupby feature
    e. create a discussion cell in which you discuss what your analysis tells you about the data
4. push your changes to github
5.

\What and How to submit

1. You should make a short Zoom recording (about 1-5 minutes) stored in the cloud, showing mastery of the following six skills
    a. running the Jupyter notebook and showing the markdown and code cells you created (**Jupyter**)
    b. describe your dataset and the questions you were asking, and show the pandas code you wrote to read it in and show the columns and index and the table itself (**pandas.read_csv**)
    c. show a pivot table you created and interpret it (**pandas.pivot_table**)
    d. show the plot you made from the pivot table (**pandas,plot**) the plot should have labels on the axes and a title and a legend
    e. show your use of groupby and interpret the results (**pandas.groupby**)
    f. describe what your analysis tells you about the questions you asked (**pandas.analysis**)
    g. upload links to the github repository and a link t your Zoom movie to the CPA01 problem on the Master-Programs site (with PIN 7297444)
2. the due date is Friday 3/11 before 11am (as we start grading at 11)

Rubric

We wil grade this using a Specs grading approach looking for mastery of the following skills, so make sure that you demonstrate yourself mastering these skills in your movie!

1. Github: collaborating on github with a single branch
2. Debug: debugging using VScode
3. Jupyter: running queries and writing markdown using Jupyter lab
4. Python OOP: writing and calling methods for a Python class
5. Pylint: using pylint to write clean code
6. Python Script: modifying a console-based interactive Python script where the state is encapsulated in a Python class object

We hope and expect that everyone will demonstrate mastery of all 6 of these skills in this PA. If not, then you will have an opportunity to demonstrate mastery in later homework assignments.

For those students with more experience in Python programming, I encourage you to go beyond this assignment and do something you can add to your ePortfolio. It could help you find an internship or job. For example, you could learn to use Flask to create a web interface for the course_search app. When we get a live feed of the registrar's data later this year, it could actually be useful for Brandeis students as an alternative to using Workday to finding courses! You won't get more points, but you will build your ePortfolio!