CSE 519 -- Data Science (Fall 2024)
Prof. Steven Skiena
Homework 1: Exploratory Data Analysis in iPython
Due: Thursday, September 26, 2024 (11:59 PM)

This homework will investigate doing exploratory data analysis in iPython. The goal is to get you fluent in working with the standard tools and techniques of exploratory data analysis, by working with a data set in a domain where you have some basic sense of familiarity.

This homework is based on the Kaggle New York Stock Exchange data set, which provides daily stock prices on major stocks with interesting metadata. Specifically, this data set contains:

- **prices.csv**: raw, as-is daily prices. Most of the data spans from 2010 to the end 2016, but for companies new on stock market date range is shorter. There have been approximately 140 stock splits in that time, this set doesn't account for that.
- **prices-split-adjusted.csv**: same as prices, but there have been added adjustments for splits.
- **securities.csv**: general description of each company with division on sectors
- **fundamentals.csv**: metrics extracted from annual SEC 10K fillings (2012-2016), should be enough to derive most of popular fundamental indicators.

You will need to submit your code files in two different formats (.ipynb and .pdf). Make sure to have your code documented with proper comments and the exact sequence of operations you used to produce the resulting tables and figures. The submission steps are discussed below.

Data downloading

First of all, you need to join Kaggle and download the data here. The description of the data can also be found at this page.

Python Installation

Instead of installing python and other tools manually, we suggest installing **Anaconda**, which is a Python distribution with a package and environment manager. It simplifies a lot of common problems when installing tools for data science. More introduction can be found here. Installation instructions can be found here.

Another option can be using **Google Colaboratory.** This is another option for those who want to run their Jupyter notebook remotely instead of installing the required packages locally. Colab allows you to write and execute Python in your browser, with

- Zero configuration required

- Free access to GPUs
- Easy sharing

If you are an expert of Python and data science, what you need to do is install some packages relevant to data science. Packages that I believe you will definitely use for this homework include:

- pandas
- scikit-learn
- numpy
- Matplotlib
- seaborn (maybe)

The Google colab notebook (must access using @stonybrook.edu, not @cs.stonybrook.edu) contains boilerplate code to download the data to your google drive and a dictionary containing the features along with its data type. **Make a copy of the notebook before you start your HW.**

Tasks (100 pts)

1. Identify which companies in the dataset entered after data collection started (presumably new companies) and which left the exchange before the end date for data collection (perhaps going bankrupt). (10 points)
2. Order the companies by how big/interesting/important they are, so we can restrict analysis to the top *n* companies to work with to keep future computations reasonable. Describe what ranking feature you think best here and why, and what *n* works for you for future analysis **given your limited compute environment.** (10 points)
3. There are two datafiles, one with raw prices, the other adjusted by stock splits. Compare the pre-and-post split files to see if you can identify the dates/companies of the roughly 140 stock splits and what the ratio of each of the splits were. (15 points)
4. Pairs trading is an investment strategy which relies on identifying pairs of stocks which move in the same direction each day – if stock A goes up (down) on a given day, then stock B likely goes up (down) the same day. So:
    a. Construct an appropriate daily time series for each stock reflecting how much it goes up or down each day.
    b. Construct a pairwise correlation matrix measuring how in sync these movements are among all pairs of your *n* top stocks. **Present this pairwise correlation matrix in a way to make its lesson as clear as possible to the viewer**.
    c. Identify which pairs are most and least strongly correlated in their movements, and propose some reasonable explanations **why** this is the case.
    d. The securities.csv file contains the economic sector which each company participates in. Do companies within the same economic sector have stronger or weaker price correlations than those in different sectors?
    (20 points)

5. Plot the distributions of frequency of daily price movements according to your statistic. What type of classical distribution does this look like, and are there any surprising deviations from the theoretical distribution? (10 points)
6. Create **three** plots of your own using the dataset that you think reveal something very interesting. Explain what it is, and anything else you learned from your exploration. (15 points)
7. The fundamentals.csv file contains four years worth of profitability data on each of the companies. Perhaps the most important number to reflect how profitable the company is the *earnings per share*. There are two predictive tasks here:
    i. Use linear regression to predict the earnings per share for company X in year Y using the other variables from fundamentals.csv for year Y
    ii. Use linear regressions to predict the earnings per share for company X in year Y using the variables from fundamentals.csv for year Y-1.

    How good are these models and how can you tell? (10 points)
8. Now repeat this exercise to try to build better models for parts i and ii. I bet that (hint) improved data preparation/normalization/feature engineering will help. Maybe an algorithm other than linear regression (e.g. Random Forest, Nearest Neighbor, etc) will prove better training. [Note: *scikit-learn* is a user-friendly library which is used to perform data loading, pre-processing, transformations, algorithms and metrics needed for Data Science and Machine learning] Compare their performance and explain your reasoning for the differences in their performances. (10 points)

Be honest. This is your first modeling experience, and I am hoping to see you learned something, not that you got a great performance on the model.

Rules of the Game

1. This assignment must be done **individually by each student**. It is not a group activity.
2. If you do not have much experience with Python and the associated tools, this homework will be a substantial amount of work. Maybe check out the online course https://learn.deeplearning.ai/courses/ai-python-for-beginners. Get started on the assignment as early as possible!
3. All of your written responses should be put in the appropriate place in your notebook template. Get the template notebook form from here. *You must access the notebook via your @stonybrook.edu email address, not @cs.stonybrook.edu!*
   You are allowed to add more cells, but definitely fill out the cells we give.
4. I will give a brief introduction to topics like linear regression in detail before the HW is due, with detailed treatment to come. Muddle along for now, and we will understand the issues better when we discuss them in the course. Feel free to read ahead in the book.
5. KISS is an important philosophy in data science: *keep it simple, stupid*.. For this homework that means that you should start by making a pass through the assignment

doing simple things, instead of over-optimizing each part. *Then* go back and improve things where it counts once you know how simple works.

6. You may use ChatGPT if you want, provided that you cite it through the class policy in the syllabus. But you will be doing yourself a terrible disservice if you use this to fake Python programming rather than learn it – now is your chance!
7. You will submit your code so we can run it through MOSS to detect copying and plagiarism. Do your own work!!
8. Our class Piazza account is an excellent place to discuss the assignment. Check it out at piazza.com/stonybrook/fall2024/cse519.
9. It is good to learn about the stock market, but the goal of investing is long term returns, not short term speculation. I personally invest in low-cost mutual funds or ETFs that diversify by tracking broad market indices like the S&P 500. Perhaps now (when you are poor) is a good time to create a Roth IRA retirement account and make whatever investments you can while your tax rate is low.

Submission

Submit everything through Google classroom. As mentioned above, you will need to upload:

1. The Jupyter notebook all your work is in (.ipynb file), derived from the provided template
2. PDF (export the notebook as a pdf file)

These files should be named with the following format, where the italicized parts should be replaced with the corresponding values:

1. cse519_hw1_*lastname_firstname_sbuid*.ipynb
2. cse519_hw1_*lastname_firstname_sbuid*.pdf