MDS5102 Python Programming

Assignment 3

Due Date: 23 December, 2021

Assignment Description:
This assignment will worth **10%** of the final grade.

You should write your code for each question in a .py file (please name it using the question name, e.g. q1.py). Please pack all your python codes into a single zip file, name it using your student ID (e.g. if your student ID is 123456789 and to submit Assignment 3, then the file should be named as 123456789_Assignment3.zip or 123456789_A3.zip), and then **submit the *.zip** file via BlackBoard.

Please also **write a report**, which provide the details of your codes. (Note that the report should be submitted as PDF). The report should also be included in the .zip file as well.
Please note that, the teaching assistant may ask you to explain the meaning of your program, to ensure that the codes are indeed written by yourself. Plagiarism will NOT be tolerated. We may check your code using Blackboard.

Only packages mention in Lecture are allowed to use, e.g. Numpy, Pandas, Matplotlib, Sklearn.

This assignment is due on **23:59PM, 23 Dec (Wednesday)**. For each day of late submission, you will lose 10% of your mark in corresponding assignment. If you submit more than three days later than the deadline, you will receive zero in this assignment.

A well-documented Report is necessary for this assignment.

Question 1 Data Visualization (4%×5=20% of this assignment)
SP500_stocks_5yr.csv attached provides stock prices from 2013/2/8 to 2018/2/7 for all companies on the S&P 500 index. The file has the following columns.

date:	In format of yyyy-mm-dd
open:	The price of the stock when the market opens in the morning
close:	The price of the stock when the market closed in the evening
high:	Highest price the stock reached during that day
low:	Lowest price the stock reached on that day
volume:	The total amount of stocks traded on that day
Name:	Stock's ticker name

All prices are in the unit of USD. Write programs to conduct following actions.

(1) Ask users to **input** a stock code. **Print warning massage** and wait for new input if input code not found in the given dataset. Filter the data of input stock code out if input name matched.
Print the shape of result data and **show** the start, end date and the number of trade days covered, then **save** the filtered data as a new csv file named "{stock code}.csv" in the same directory.

(2) Ask users to **input** 2 dates, both in format of "xxxx (Year)-xx (Month)-xx (Day)", and **Print warning massage** and wait for new input if input date is not covered by result in part (1).
Take the earlier one as the start day and the later one as end day and **keep only** the data in part (1)

within this period. **Plot** all the filtered data except "Volume" in terms of date in different subplots in a big figure. You should add title to each subplot to show the item plotted and other necessary labels.

(3) As the stock prices are highly volatile and change quickly with time, moving average can be used to observe a trend. Ask users to **input** a positive integer as window length to computing moving average. **Add Plots** of moving average of each item to the corresponding plots in (2).

(4) Focusing on the "Open" & "Close", **plot** these two items of filtered data in part 2 in one scatter plot in terms of date. **Mark** open price as "o" and close price as "x". **Colorize** data as red if close price is larger than open price, green if not.

(5) **Plot** bar chart to show volume data of filtered results in part 2 with respect to date.

Question 2 Data Aggregation (5%×4=20% of this assignment)

Score.csv attached contains 1000 samples of scores from three exams and a variety of personal, social and economic factors that have interaction effects upon them.

(1) **Read** the given data as a data frame and **present** data type of each column.

(2) **Present** the statistic (min, max, mean, standard deviation) of all scores for different parental levels of education.

(3) For each combination of ethnicity and lunch, **present** the percentage of male and female.

(4) **Evaluate** the effectiveness of test preparation course and lunch status using statistics of score. Identify which factor has more effect.

Question 3 KNN Classifier (20% of this assignment)

Split the first 200 sample in ***Score.csv*** into training data, the last 50 samples into test data. Given the scores of reading and writing, use KNN classifier to predict whether the sample has completed test preparation course. **Scikit-learn** package can be used for model selection and training. Things you are required to do:

(1) **Visualize** samples in training set and use different colors to **identify** samples with different course status.

(2) Take K=10, **train** a KNN classifier using training samples, and **visualize** predictions of trained model in the plots of part (1), mark predictions differently from training samples.

(3) **Evaluate** performance, e.g. accuracy, of model in part (2). **Change** K to smaller or larger value, **re-train** KNN model and **compare** performance of all models at least 3 times. Briefly **describe** the phenomenon you observed.

Question 4 Linear Classification (20% of this assignment)

Continued from question 3. Use a linear classifier to predict status of preparation course of samples in test data.

(1) Conduct **prediction** on test set and **Evaluate** performance of the trained linear classifier. **Visualize** training samples, predictions and separation line represented by trained model. Use annotations properly in your plots.

(2) **Varying** the number of training samples N, up to 950, and **re-train** the linear classifier. **Visualize** separation lines of different models and evaluate their performance.

(3) **Visualize** the trend of model accuracy with respect to N. Briefly **conclude** your experiments.

Question 5 Regression (20% of this assignment)

Split the first 900 samples in ***Score.csv*** into training set and the last 100 samples into test set. Use a linear regression model to study the trend of **reading score** with respect to **match score**.

(1) **Train** a linear regression model using train set and **conduct** prediction on test set.

(2) **Plot** all original samples in blue, your prediction in red. And **plot** line represented by your regression model.

(3) **Evaluate** the performance of your model.

End of Assignment