COLLEGE OF ENGINEERING
COMPUTER SCIENCE
VIRGINIA TECH.

Machine Learning, I

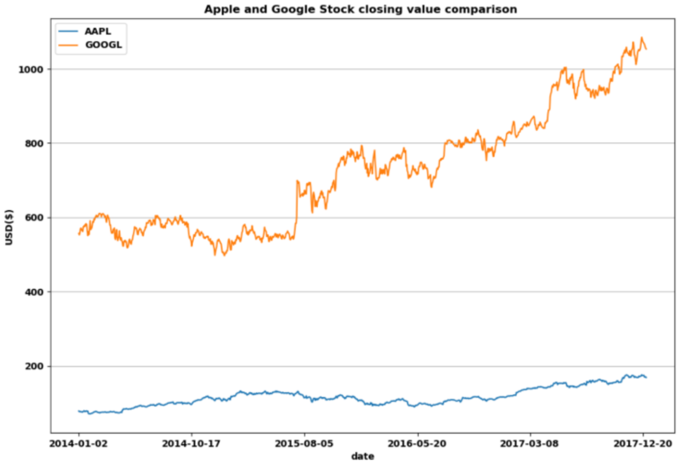
CS 5805

Homework # 2

In this assignment, you will learn how to create Dataframe satisfying specific criteria, graph time series data, apply rolling window, aggregation, and discretization of dataset. All numbers must be displayed with 3-digit decimal precision.

1. Develop a python program that loads the 'stock prices.csv' dataset from the course GitHub using the Pandas package. [6pts]
    a. Does the dataset contain missing entries? Display the missing features and the number of missing entries on the console.
    b. Replace the missing values with 'mean'.
    c. Run a test to check that all missing values are filled and cleaned. Show the result of the test on the console and explain how the dataset is cleaned. The rest of the questions must be answered using the cleaned dataset. Display on the console that missing values are fixed.

2. Using python, develop a program that answers the following questions: [9pts]
    a. How many unique companies are listed in the dataset. List the unique name of companies on the console.
    b. Which of the predictors are quantitative and which are qualitative?
    c. Create a new Dataframe that only includes 'GOOGLE' and 'APPLE' stock with all the original predictors in place. Plot the closing stock value for google and apple in one graph versus time as shown below. Note: Figure size = (12,8). Make sure the time x-axis (x-tick) is not crowded and displays the date as shown below.

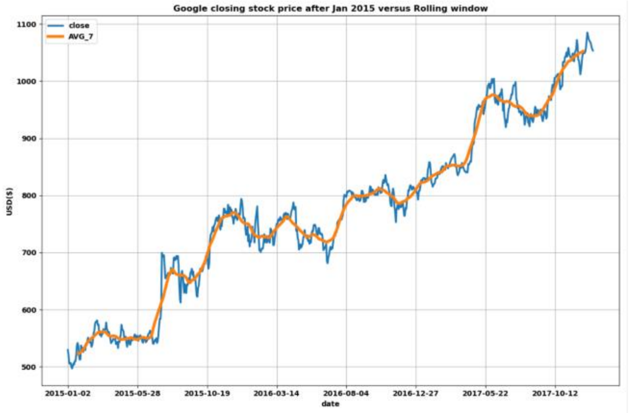


3. Create a new Dataframe that is aggregated the ‘symbol’ with the summation operation. Compare the number of objects in the cleaned data set versus the aggregated data set. Display the first 5 rows of the aggregated dataset on the console. [9pts]

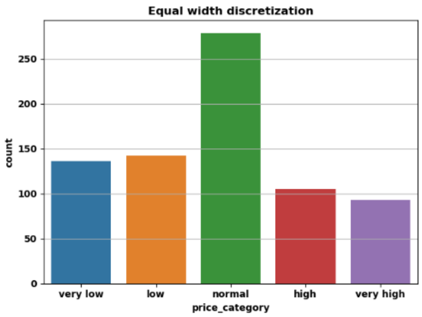
4. Create a new Dataframe that is sliced from the cleaned dataset with three features: ‘symbol’, ‘close’ and ‘volume’. Then aggregate the sliced dataset over ‘symbol’ feature with the mean and variance operation. Find the company that has the maximum variance in the closing cost. Display a message on the console. Hint: Use the np.argmax and np.max. [9pts]

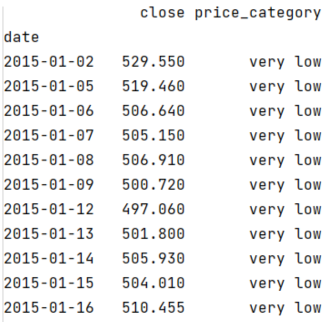
5. Create a new Dataframe that shows only the Google stock closing cost after 2015-01-01 only. Display the first 5 rows of the newly constructed dataset on the console. [9pts]

6. Plot the closing cost in the previous question versus the time versus the rolling window and mean method [window size = 30 days]. How many observations will be missed when rolling window applies? What are the advantages and disadvantages of rolling windows? Note: The final plot should look like below. Hint: Use the df.rolling() with center=True. [10pts]



7. Discretize the ‘close’ feature in the created dataset in question 5 into 5 <u>equal-width bins</u>: ‘very low’, ‘low’, ‘normal’, ‘high’, ‘very high’. Name the new categorical feature as ‘price_category’. Plot the count plot of the new added categorical feature. Display the created dataset on the console using the df.to_string() function. [6pts]

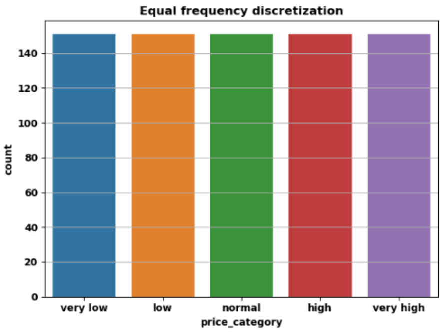


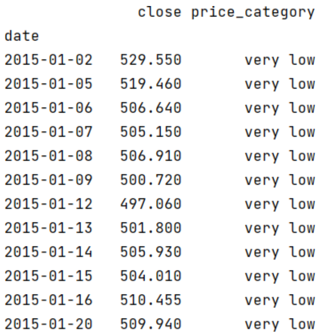


```
                close price_category
date
2015-01-02   529.550       very low
2015-01-05   519.460       very low
2015-01-06   506.640       very low
2015-01-07   505.150       very low
2015-01-08   506.910       very low
2015-01-09   500.720       very low
2015-01-12   497.060       very low
2015-01-13   501.800       very low
2015-01-14   505.930       very low
2015-01-15   504.010       very low
2015-01-16   510.455       very low
```

8. Plot the histogram plot of the ‘close’ feature in the previous question (number of bins = 5) . Does the histogram plot make sense compared to the count plot in the previous question. Explain your answer. [7pts]

9. Discretize the ‘close’ feature in the created dataset in question 5 into <u>equal frequencies</u> (q = 5). Bins: ‘very low’, ‘low’, ‘normal’, ‘high’, ‘very high’. Name the new categorical feature as ‘price_category’. Plot the count plot of the new added categorical feature. Display the created dataset on the console using the df.to_string() function. [6pts]





```
                close price_category
date
2015-01-02   529.550       very low
2015-01-05   519.460       very low
2015-01-06   506.640       very low
2015-01-07   505.150       very low
2015-01-08   506.910       very low
2015-01-09   500.720       very low
2015-01-12   497.060       very low
2015-01-13   501.800       very low
2015-01-14   505.930       very low
2015-01-15   504.010       very low
2015-01-16   510.455       very low
2015-01-20   509.940       very low
```

10. Using the definition of covariance matrix estimate the covariance matrix <u>without</u> using built-in python function (for the dataset in question 5 including all features (open, high, low, close and volume). For the covariance matrix estimation, you cannot use the built-in function like .cov().The estimated covariance matrix needs to be constructed using the definition of covariance matrix. Display the covariance matrix contents on the console. [10pts]

11. Estimate the covariance matrix for the previous question with the built-in python function. cov(). Display the result on the console. The answer to this question must be identical to the answer to the previous question. Write down your observations about the values on diagonal and off diagonal of the covariance matrix. What is the linear relationship between features in this dataset? [10pts]

Upload the solution as a **<u>report (as a single pdf</u>**) plus **<u>the .py file</u>** through Canvas by the due date.