THE GEORGE WASHINGTON UNIVERSITY
WASHINGTON, DC

Visualization of Complex Data

DATS 6401

LAB # 2

In this LAB, you will connect to Yahoo stock database using Python Pandas package & visualize the stock values for major giant companies through pandas package. Display numbers with 2-digit decimal precision.

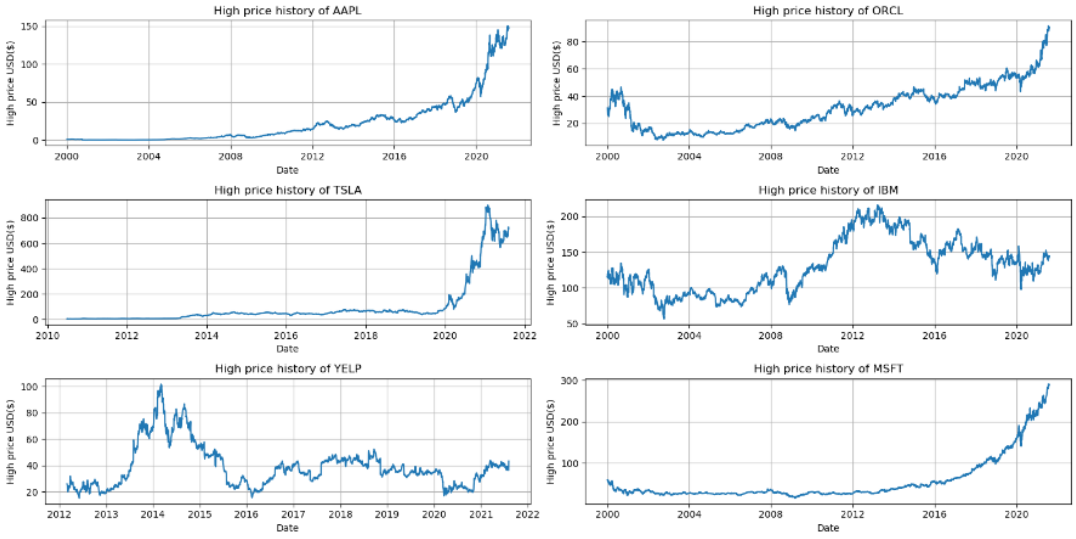
1. Using the pandas_datareader package connect to yahoo database and load the stock value for the following giant companies. Pick the start date as '2000-01-01' and the end date "June 6th, 2023".

   ```
   stocks = ['AAPL','ORCL', 'TSLA', 'IBM','YELP', 'MSFT']
   ```

   You will need the following package to be able to connect to yahoo API. Make sure to use the updated version of the pandas and pandas data_reader (You can use the "pip install --upgrade pandas" and "pip install --upgrade pandas-datareader")

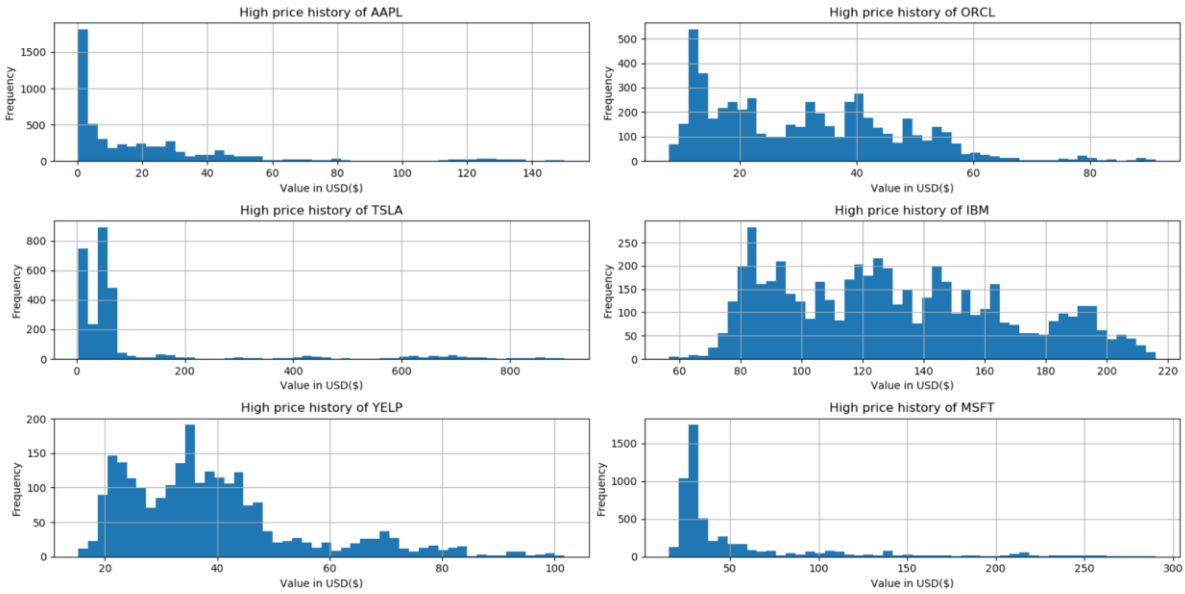
   ```
   import pandas_datareader as web
   ```

2. The database contains six features: "High," "Low," "Open", "Close", "Volume", "Adj Close" in USD ($). Using the panda package and subplot from matplotlib package, plot the "High" columns for all companies in one figure with three rows and two columns graph. Make sure to add title, legend, x-label. y-label and grid to your plot. The plot below is only for demonstration and may not be the solution. Fig size = (16,8)



3. Repeat question 2 for, "Low", "Open", "Close", "Volume", "Adj Close".

4. Using the panda package and hist command, plot the histogram plot of the "High" columns for all companies in a three-by-two graph. Make sure to add title, legend, x-label. y-label and grid to your plot. The plot below is only for demonstration and may not be the solution. # of bins = 50



5. Repeat question 4 for, "Low", "Open", "Close", "Volume", "Adj Close".

6. Using the technique covered in class, estimate the covariance matrix for the features in the "AAPL" company [You are NOT allowed to use. cov() built-in function]. Display the estimated covariance matrix using the pretty table package with an appropriate title. Which two features have the highest correlation, and which two features have the lowest correlation?

7. Repeat question 6 for, "ORCL", "TSLA", "IBM", "YELP" and "MSFT".

8. Use pandas package to plot the scatter matrix. Using pandas package plot the scatter matrix plot of the "AAPL" company with the following parameters: hist_kwds= {'bins': 50} , alpha = 0.5, s = 10, diagonal = 'kde'. Hint: you can use the following command: pd.plotting.scatter_matrix(). Write down your observations about the correlation between features.

9. Repeat question 8 for, "ORCL", "TSLA", "IBM", "YELP" and "MSFT".

Submission guidelines:

1. The softcopy of the developed Python code .py must also be submitted separately. Please make sure the developed python code runs without any error by testing it through PyCharm software. The developed python code with any error will subject to 50% points penalty.
2. Add an appropriate x-label, y-label, legend, and title to each graph.
3. Write a report and answer all the above questions. Include the required graphs in your report.
4. Submission: report (pdf format) + .py file. The python file is a supporting file and will not replace the solution. A report that includes the solution to all questions is required and will be graded only.
5. The python file must regenerate the provided results inside the report.