THE GEORGE WASHINGTON UNIVERSITY
WASHINGTON, DC

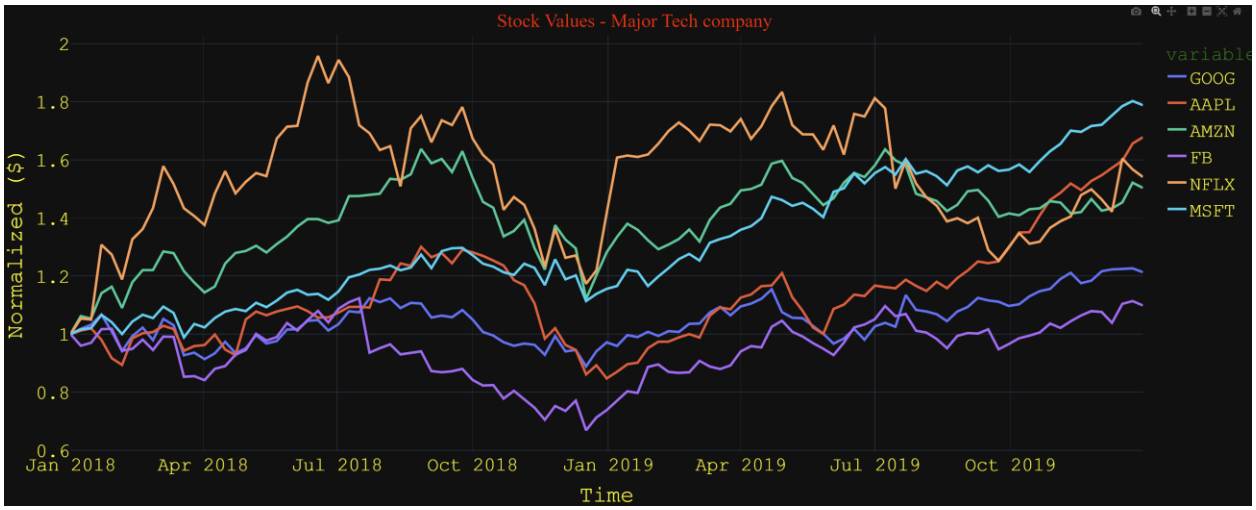
Visualization of Complex Data

DATS 6401

LAB # 4

In this assignment you will practice interactive complex data visualization using plotly express package. The PCA analysis for the feature dimension reduction will be discussed in the LAB. The dataset for this LAB is picked from the public repository inside the plotly express.

1. Load the 'stocks' data set from the plotly express and display the list of features and the last five observations on the console. Is this a time-series or non-time series dataset? What does the feature in this dataset set represent? [5 pts]
2. There are six giant tech companies in this dataset. Plot the stock values versus time in one graph. The x- axis is the date and the y axis is the stock value. Update the layout with the following settings. The graph must be displayed on the browser using the Plotly package. This plot must be open in browser. [20 pts]
    a. Title Font color 'red'
    b. All Font size thirty
    c. Title Font family 'Times New Roman'
    d. Legend title font color 'green'
    e. Rest of labels font family' Courier New'
    f. Change the x-label and y-label as shown below.
    g. x-label & y-label & legend font color yellow
    h. Linewidth four.
    i. Width two thousand height eight hundred
    j. Template plotly dark.

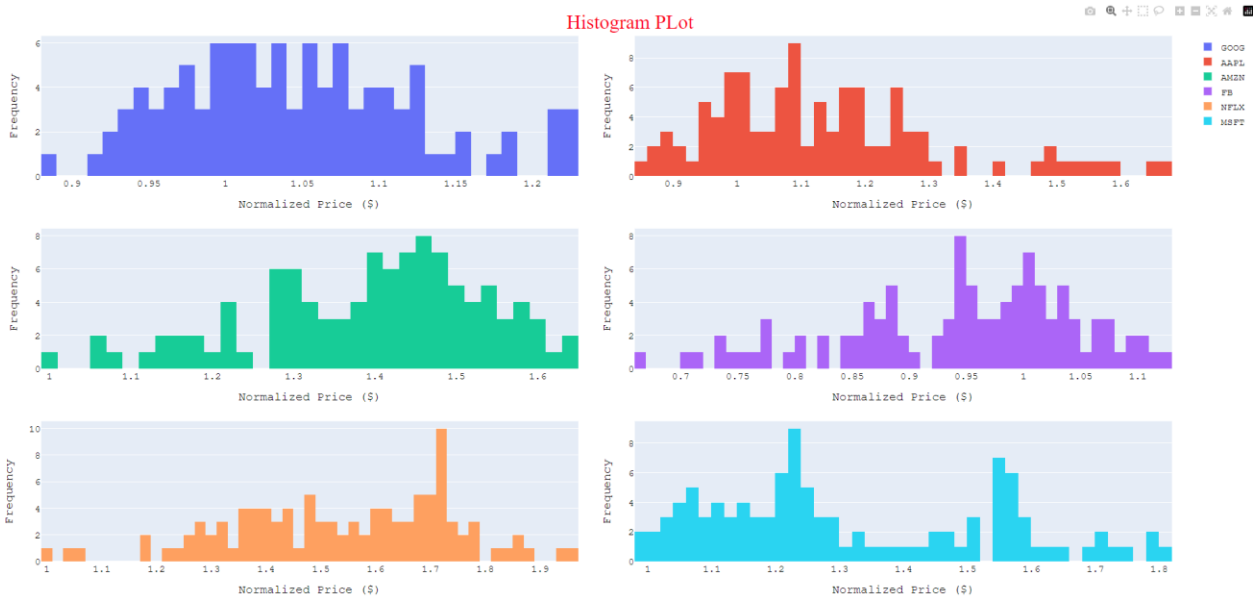


3. Graph the histogram plot of the six major tech companies in one graph (subplot not shared axis three rows and two columns). The final graph should be like the one below. Number of bins = 50.

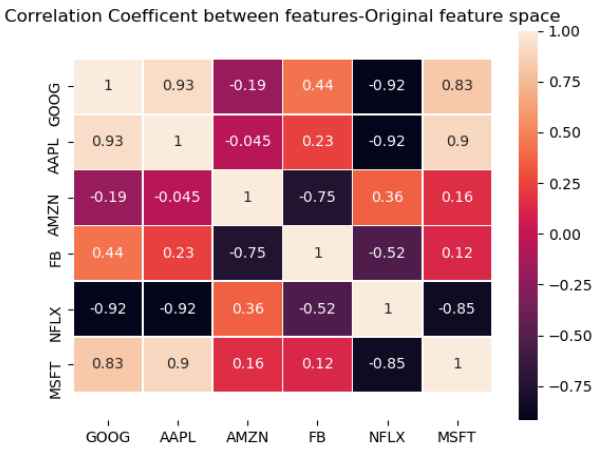
Hint: You may use the following libraries and use fig.update_layout() command. The graph must be displayed on the browser using the Plotly package. This graph must be open in browser [20 pts]

```
import plotly.express as px
from plotly.subplots import make_subplots
import plotly.graph_objects as go
fig = make_subplots(rows=3, cols = 2)
```

a. Title Font color 'red'
b. Legend & Title Font size thirty
c. Title Font family 'Times New Roman'
d. Legend title font color 'green'
e. Rest of labels font family' Courier New'
f. x-label and y-label font size fifteen.
g. x-label & y-label & legend font color black
h. Width & height: Default value



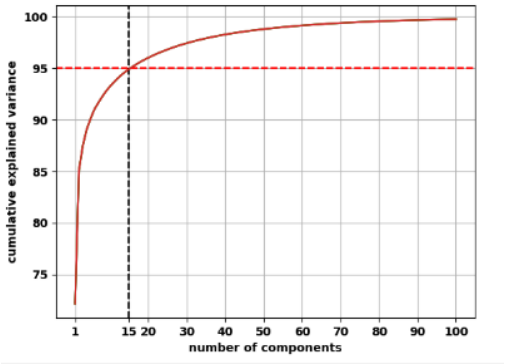
4. Consider each company stock as a feature that needs to be fed to a ML model. The target is not given in this problem. You need to perform a complete PCA analysis of the 'stocks' dataset and answer the following questions and tasks: [55 pts] [No need to use plotly]
    a. Using the following library standard (standardize) the feature space.
    `from sklearn.preprocessing import StandardScaler`
    b. Find the singular values and condition number for the original feature space and display it on the console.
    c. Find the correlation coefficient matrix between all features of the original feature space and use the seaborn heatmap to display the result. The numbers in the heatmap below may not be accurate.



d. Perform a PCA analysis using the following library in python and assume the threshold for explained variance to be 95%. How many features should be removed per the PCA analysis and assumed threshold? Explain your answer. Then display the 'explained variance ratio' of the original feature space versus reduced feature space on the console. [No need to use plotly]

```
from sklearn.decomposition import PCA
```

e. Graph the 'cumulative explained variance' in percentage versus the 'number of components. Draw a vertical and horizontal dash lines [black and red] that shows 95% explained variance and the optimum number of features to be displayed on the x axis as shown below [numbers in the graph are dummy] [No need to use plotly]



f. Find the singular values and condition number for the reduced feature space and display it on the console. Compare the singular values and condition numbers of the original data and the reduced feature data. Explain your answer.

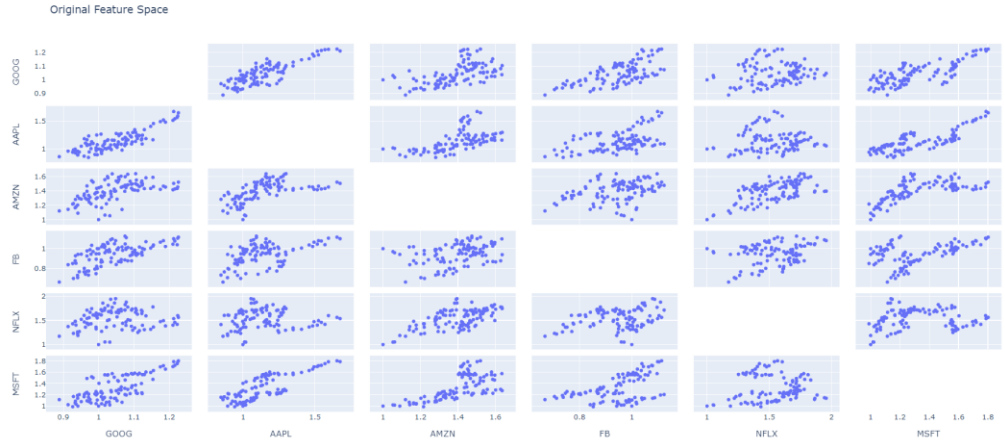
g. Find the correlation coefficient matrix between all features of the reduced feature space and use the seaborn heatmap to display the result. Explain the results from the table.

h. Create a new Dataframe with the columns that are transformed. Name the columns as 'Principal col 1', 'Principal col 2' ,… Your code should be generalized and should work for any number of components in the PCA analysis. Display the first five rows of the newly created Dataframe.

i. Plot the line plot versus date (like question 2) for the transformed feature (reduced dimension feature space). [Use plotly]

j. Graph the line histogram plot (like question 3) for the transformed feature (reduced dimension feature space). Graph the histogram into one column and multiple rows. [Use plotly]

k. Using the Plotly express and scatter_matrix command, plot the interactive scatter matrix for the 'original feature space' and the 'reduced feature space.' The original feature space scatter plot should look like the figure below. Does the plot for reduced features male sense. Explain why? [Use plotly]



Upload a formal **report (as a single pdf**) plus **the .py file** through canvas by the due date.