Data Engineering NETW908

Final Project

1 Abstract:

In this project you will learn how to extract,explore and understand financial data regarding the stock market.The means by which you will extract this data is through the IEX cloud API.

2 Data extraction:

Please refer to the excel sheet to know which stocks you will be working on.
https://docs.google.com/spreadsheets/d/1quEFwaeUGjRSoM0UZ8l5mmoDwUDd6k6G5a9lL3u8WRU/edit?usp=sharing

Symbols	Duration	Group
MMM,AXP,AIG,ADB,ACN,CHRW,CSCO,DVN,XOM,GM	2nd quarter of 2020	1
HAL,AAPL,TSLA,HPQ,ICE,ISRG,IVZ,MTD,MRNA,NVDA	1st quarter of 2021	2
OKE,OXY,MS,MCHP,MKTX,MRO,ES,EBAY,DHR,CVS	4th quarter of 2020	3
NFLX,SCHW,CERN,CNP,CDAY,FB,ADI,BKR,CVX,DVN	4th quarter 2019	4
EOG,F5,FRC,BEN,APH,ANSS,ABT,ABBV,UPS,FDX	3rd quarter of 2020	5

The features required to be extracted are the following (symbol,open,close,high,low,volume,change percent,date,company name and sector).All data extracted should be stored in a single dataframe.Show the first and last 5 rows of the dataframe as well general information about the dataframe.
Please refer to the data extraction notebook in week 9.

3 Description:

1. Data cleaning

All prices should be rounded to 2 decimal places and the change percent feature should be presented as a percentage and rounded to 2 decimal places as well.

2. Understand relationships of your data's features:

The first step in this project would be to understand the relationships between the features of your data. For each stock,Calculate the correlation coefficient between each numeric feature and the close price.

3. Variable Transformation

Variables in the dataset have several types, some of which can be easily handled by linear models while others not, like for example categorical variables or date/time variables. Thus, while exploring your dataset you need to identify variables of these types and transform them into numerical variables. Please note that no more than 20 additional columns should be created.

4. Date Discretization

Discretize the dates into weeks whereby you would have a new column denoting the week number of the quarter and another column denoting the range of each week.(i.e week number : 1 range: 23/04/2020-30/04/2020).

5. Detecting Outliers

A common way of identifying whether outliers are present in the dataset is using boxplot, which shows any outliers as black dots available below the 25th quartile or above the 75th quartile. Use this method to detect the presence of any outliers of each stock in any variable in the dataset.

6. Variable Normalization

Normalizing variables' values enhances the performance of ML algorithms. Examine the values of the dataset's variables and standardize them.

Bonus - Create a dashboard to visualise some data. For the first three tasks you should create a function and call this function on multiple stocks showing your output.For the rest, a function should be created and called only one time to achieve the task.

1- Visualize the average closing price per week for a stock.

2- Visualize the average change per week for a stock.

3- Visualize the average volume per week for a stock.

4- In one graph, visualize the Average price per week of each stock.

5- In one graph, Visualise Average closing price of each sector.

4 Deliverables:

You are required to submit a Jupyter notebook showing how you performed each step, showing its results and commenting on these results.

5 References:

- https://iexcloud.io/docs/api/#api-reference

Deadline: 9th January 2022, submit your Jupyter notebook by mail to badr.tarek@guc.edu.eg