DCU School of Computing
CA682 Assignment Report

Declaration on Plagiarism

This form must be filled in and completed by the student submitting an assignment

Name:	SOUMI MITRA
Student Number:	20210300
Programme:	MSc in Computing – Data Analytics (Full Time)
Module Code:	CA682
Assignment Title:	Data Visualisation
Submission Date:	18 Dec 2020
Module Coordinator:	Dr Suzanne Little

I declare that this material, which I now submit for assessment, is entirely my own work and has not been taken from the work of others, save and to the extent that such work has been cited and acknowledged within the text of my work. I understand that plagiarism, collusion, and copying are grave and serious offences in the university and accept the penalties that would be imposed should I engage in plagiarism, collusion or copying. I have read and understood the Assignment Regulations. I have identified and included the source of all facts, ideas, opinions, and viewpoints of others in the assignment references. Direct quotations from books, journal articles, internet sources, module text, or any other source whatsoever are acknowledged and the sources cited are identified in the assignment references. This assignment, or any part of it, has not been previously submitted by me or any other person for assessment on this or any other course of study.

I have read and understood the referencing guidelines found at http://www.dcu.ie/info/regulations/plagiarism.shtml, https://www4.dcu.ie/students/az/plagiarism and/or recommended in the assignment guidelines

Name: Soumi Mitra

Date: 17 Dec 2020

STOCK MARKET DATA VISUALIZATION

Abstract

Stock market investment decision can be made based on the current market situation and the historical data analysis on specific stocks. And price changing of different stocks for a long-term period indicates the potential connection between the listed companies. This Stock Market Data Visualization project aims to analyse stocks of various companies to specify what companies can be more beneficial to invest upon. Most of the people have very limited knowledge on stock market. To make a decision for investment will require time, knowledge and awareness on stock market historical data, and how the data varies over time. Visualization of this data can make people understand the data and they will get a clear idea about stock market operations and take proper decision for investing on various stocks. Visual portrayal is one of the most efficient approaches to help speculators to have a reasonable outline of developments of the securities exchange, just as giving a more profound comprehension of every individual stock. The utilization of chart drawing strategy can offer pictured information with explicit characteristics, for example, weight data, accompanies graphical associations between every information component. The major question here isn't just to furnish clients with an extensive showcase of enormous diagrams on the screen, yet additionally an easy-to-understand traversable visual structure for clients perusing through the structure to locate a specific detail of applicable information.

Dataset(s)

I've collected the datasets in .csv format from the following URLs.
https://www.kaggle.com/aceofit/stockmarketdatafrom1996to2020
https://www.kaggle.com/hk7797/stock-market-india
https://www.kaggle.com/camnugent/sandp500
Also, I've scrapped some data from https://finance.yahoo.com/ using Excel.

From Kaggle, I've taken 2GB of data and from Yahoo Finance I've taken some smaller datasets.
Here all the aspects of Big data are present in the datasets.

1. **Velocity**: It is the measure of how fast the data is coming. In Yahoo Finance, every minute data is getting loaded and backup is also stored.
2. **Volume**: It is the quantity of the data. As the dataset size is 2GB, the volume is huge and in Yahoo Finance also, data is getting uploaded every minute, then the volume is also increasing.
3. **Variety:** In these sites, data is available in multiple formats. But I've chosen the csv formatted data for the analysis.

I've considered the data from multiple tables (more than 100) having 619041 rows and 7 columns. I'll explain those in detail in the next section.

Data Exploration, Processing, Cleaning and/or Integration

After collecting the data, I did some filtering to get only the required data for the visualization. I've considered the following three indexes of stock market:

- S&P 500
- Dow Jones Industrial Average
- Nasdaq Composite

The datasets have the following columns:

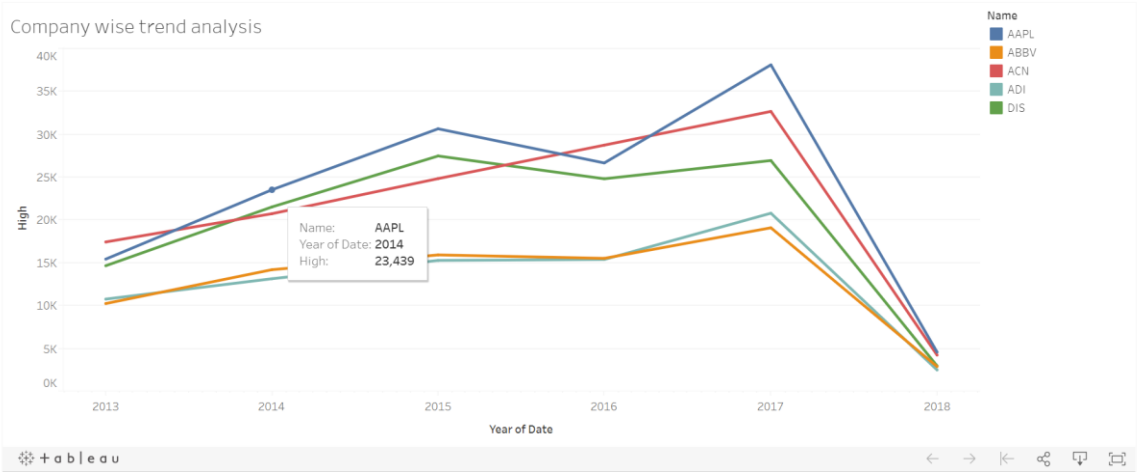
Date - in format: yy-mm-dd
Open - price of the stock at market open (this is NYSE data so all in USD)
High - Highest price reached in the day
Low Close - Lowest price reached in the day
Volume - Number of shares traded
Name/Company Name - the stock's ticker name
Last Price – Last price reached in the day

I've used the Date, High, Volume and Name fields in most of the visualizations.

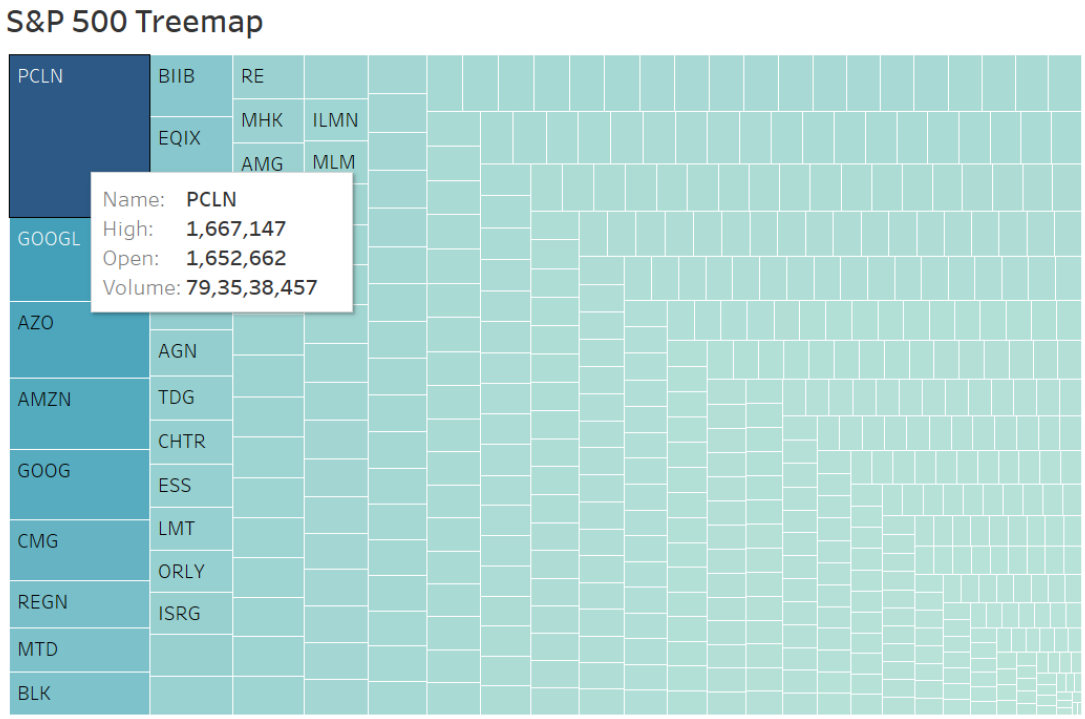
Visualisation

I've used line graph and tree maps for visualization. Here different colour combinations are used to distinguish the shares of different companies.

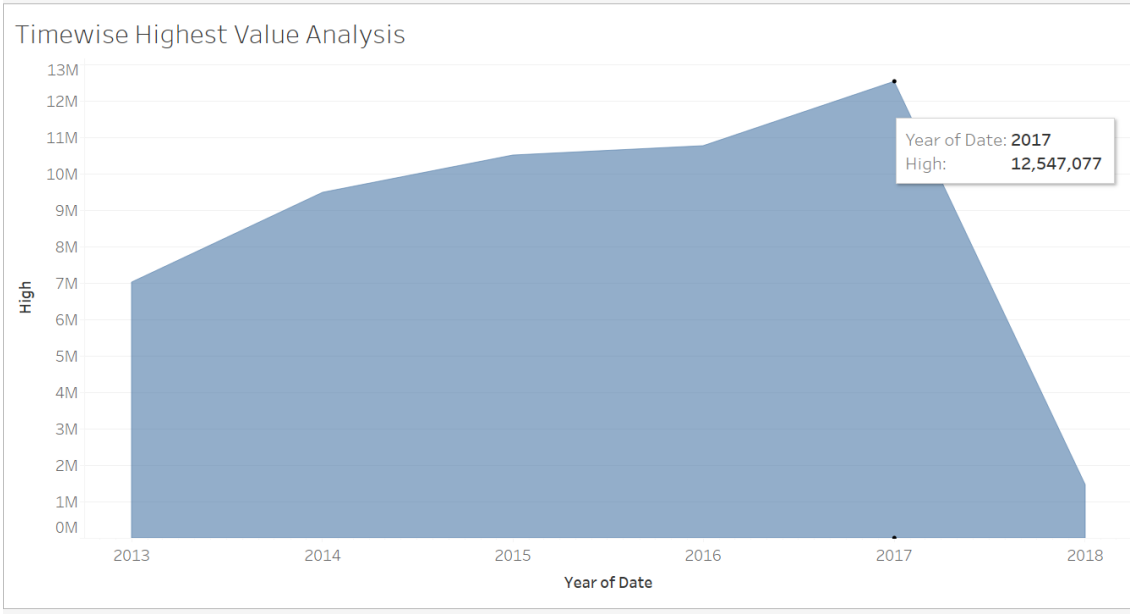
Tool used: I've used Tableau Public 2020.3 for visualizing the data.



In this graph, I've shown the trend of different company shares over time. Here, I've added a filter in the Company Name column so that the trend can be visualized more clearly. Here, I've considered all the S&P 500 shares. I've taken 5 years timeframe to draw the chart/graph. Here different colours are used to distinguish different companies.

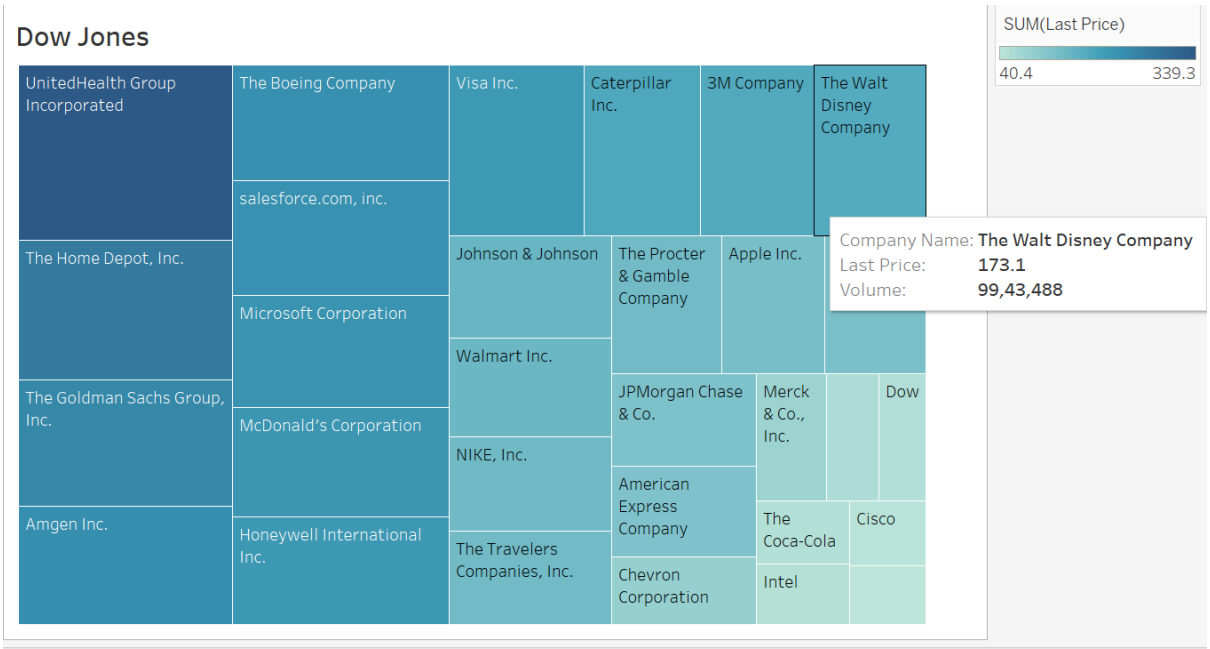


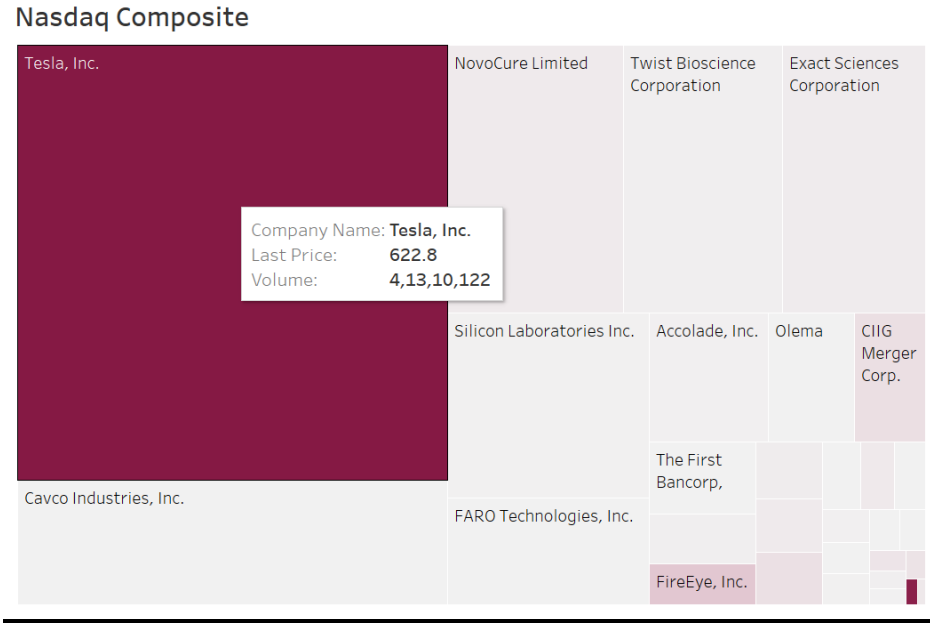
In this tree map, I've shown all the company shares as per High, Open and Volume. It will help the investors to get a clear idea about which stock has been traded more in the market and which will be more beneficial to invest on.
Here dark to light shaded colours are used to distinguish different companies.



In this line graph/area graph, I've plotted the data as per highest value over time so that investors can get an idea about whether S&P 500 stocks are beneficial or not.

In the following two tree maps, I've shown all the Dow Jones and Nasdaq Composite shares as per last price and volume so that investors can get a clear idea about the current price of the stocks and take decision for short term investment.





Conclusion

The biggest problem I faced was to collect the data in correct format and then filtering it as in most of the datasets are available company wise so getting data for all the stocks in one frame was an issue.
But then I've done some data scraping using excel from Yahoo Finance and got the historical data and components sheet to analyse the indexes well.
This visualization will work for both long term and short-term investments.

For S&P 500 data, I've used the visualization for long term investments whereas for Dow Jones and Nasdaq, the visualization was done for short term investments as I've considered the data on the basis of last updated price only.

References

I've collected the datasets from the following URLs.

- https://www.kaggle.com/aceofit/stockmarketdatafrom1996to2020
- https://www.kaggle.com/hk7797/stock-market-india
- https://www.kaggle.com/camnugent/sandp500
- https://finance.yahoo.com/

To do the visualizations, I've taken reference from the following sites and done the visualizations as per my choice.

- https://medium.com/analytics-vidhya/visualizing-the-growth-and-composition-of-s-p-500-5873d2cfac15
- https://towardsdatascience.com/visualizing-the-stock-market-with-tableau-c0a7288e7b4d
- https://www.youtube.com/watch?v=byiXWnkcLB8