Data analytics for Microsoft stock

Mini Project

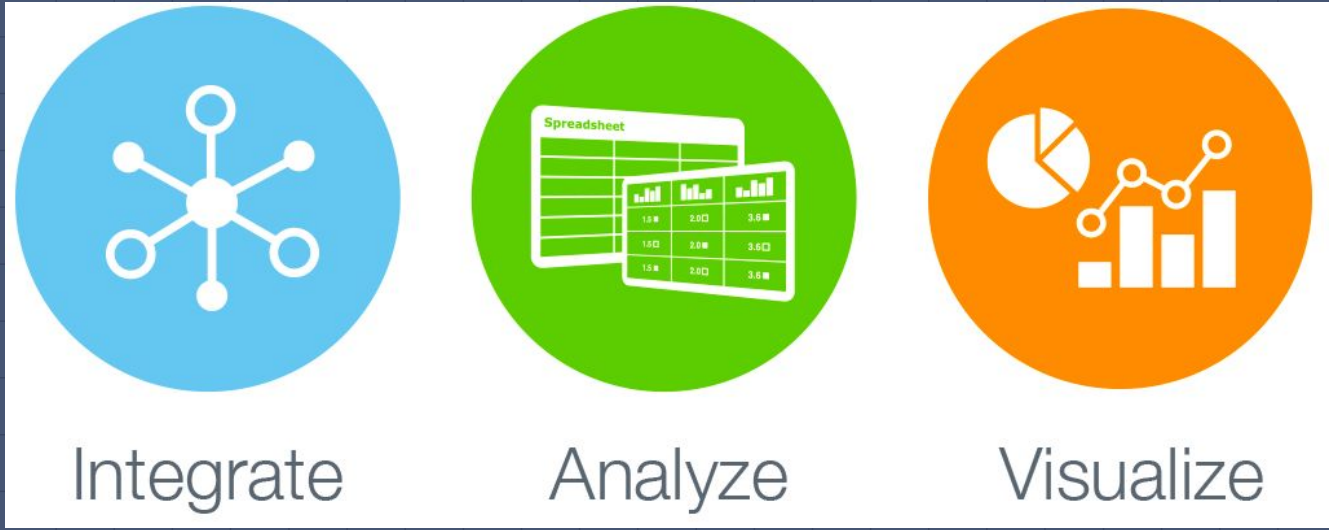
By: Surbhi Sonkiya

Project Description

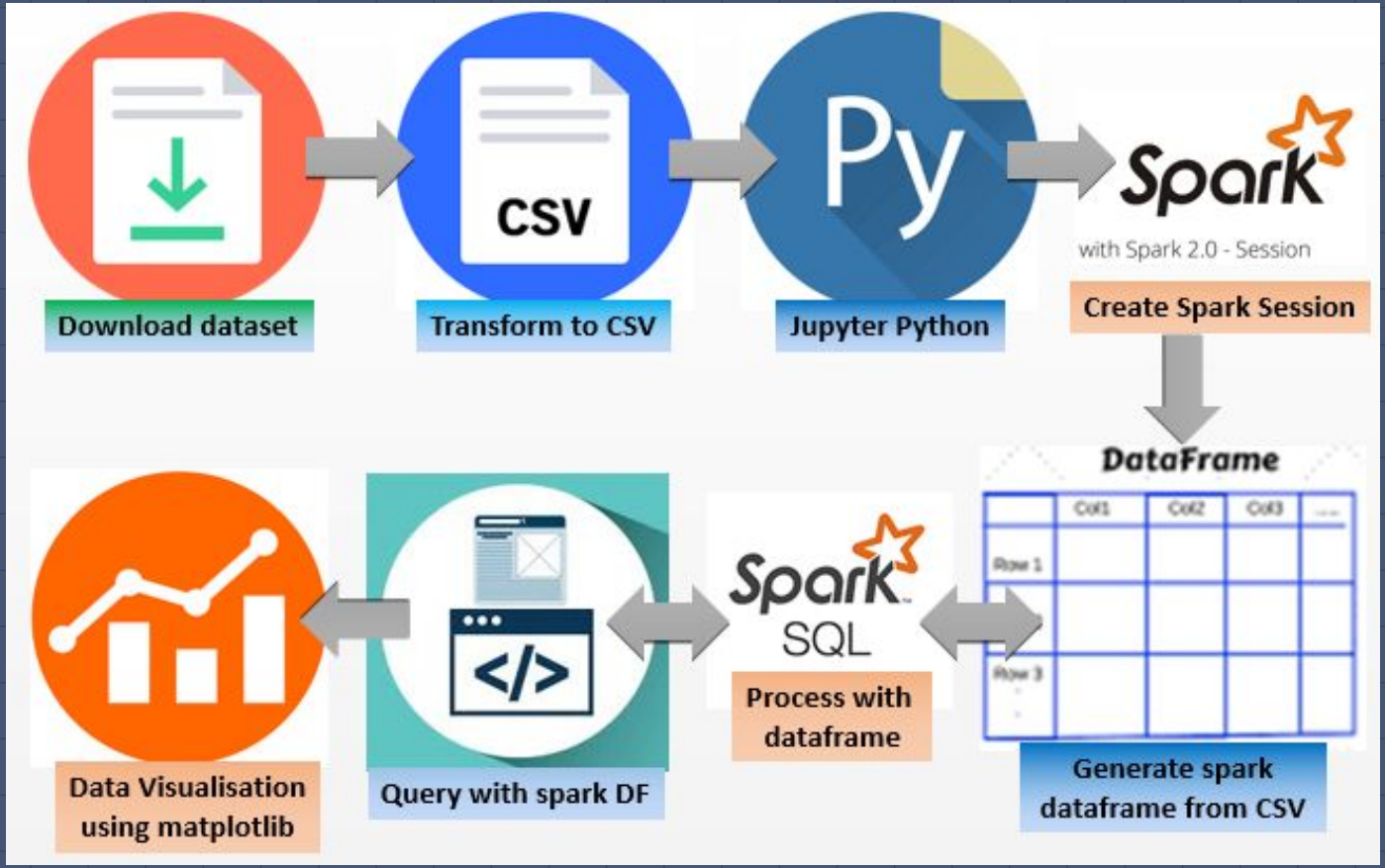
To perform data analytics on the historical stock market dataset and to visualize price variation over several years.

Project Aim

To understand basic process flow of bigdata



Process Flow



Required Dependencies

Version: Ubuntu 16.04

Version: Python 3

Version: Spark 2.2.0

Jupyter notebook

Advantages of Spark Dataframe over Spark SQL

1) Can be created from structured and semi-structured data files
2) Supports reading data from the most popular formats, including JSON files, Parquet files, Hive tables
3) It can read from local file systems, distributed file systems (HDFS), cloud storage (S3), and external relational database systems via JDBC
4) Supports any third-party data formats or sources like CSV

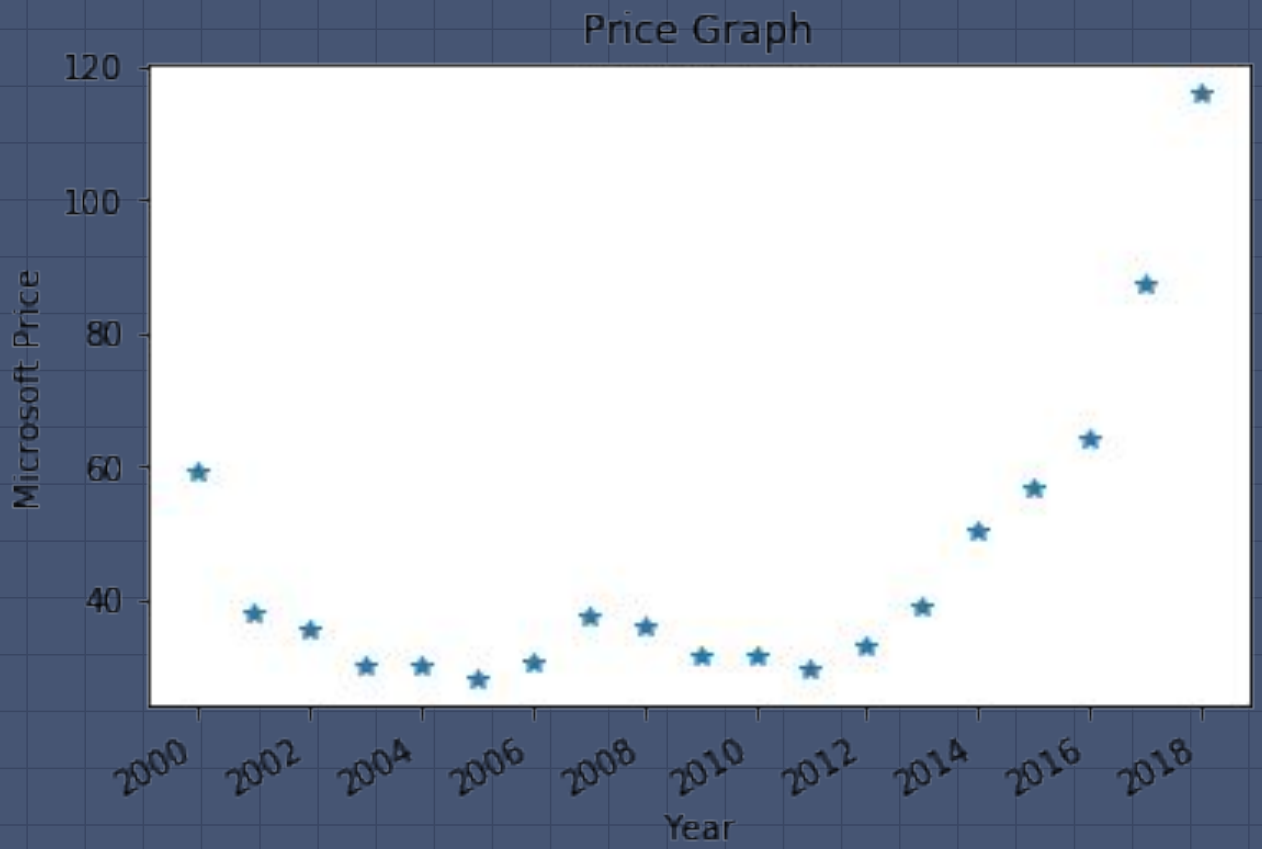
Code - Data Analytics

- ★ What are the column names?
- ★ What does the Schema look like?
- ★ Extract the first 5 columns.
- ★ Use describe function to learn about the DataFrame.
- ★ Format the column values to limit the values to show up to two decimal places.
- ★ What was the peak ‘High’ and which day was the peak ‘High’ in price?
- ★ What is the mean of the ‘Close’ column?
- ★ What is the max and min of the ‘Volume’ column?
- ★ How many days was the closing lower than avg. value of ‘Close’?
- ★ What percentage of the time was the ‘High’ greater than threshold?
- ★ What is the max ‘High’ per year?
- ★ What is the average ‘Close’ for each Calendar Month across all the years?

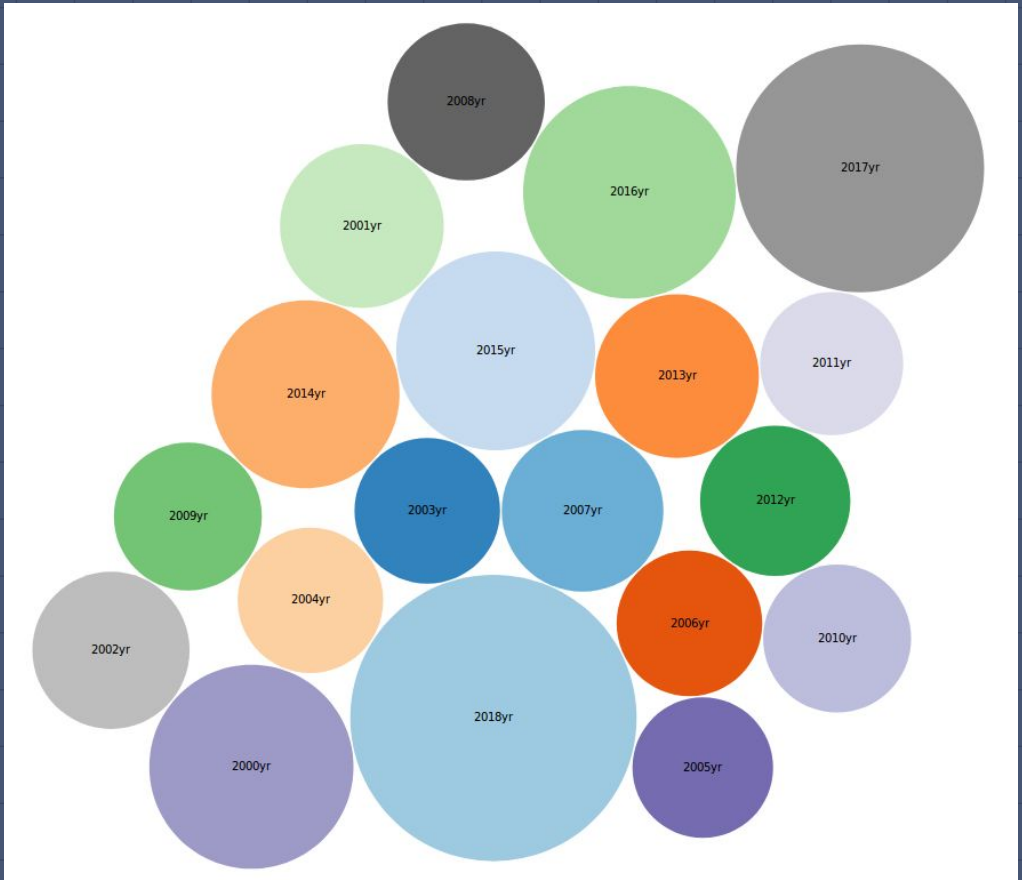
Code - Feature Engineering

- ★ Create a new feature called HV Ratio
- ★ It is the ratio of the High Price versus volume of stock traded for a day.

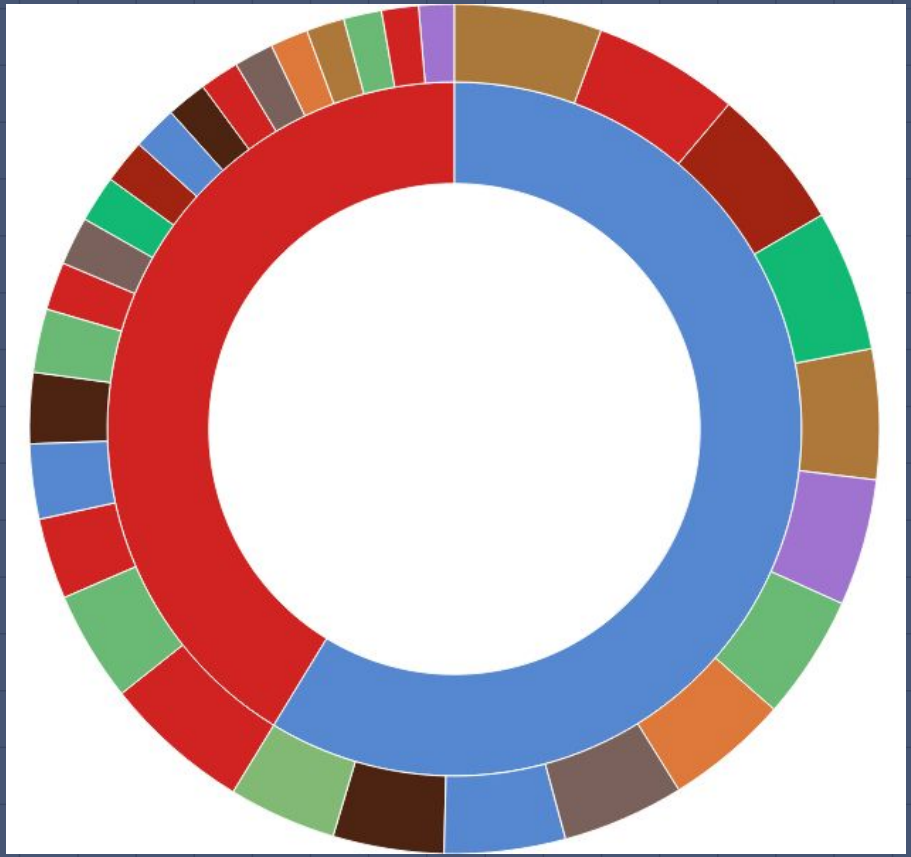
Visualization - matplotlib



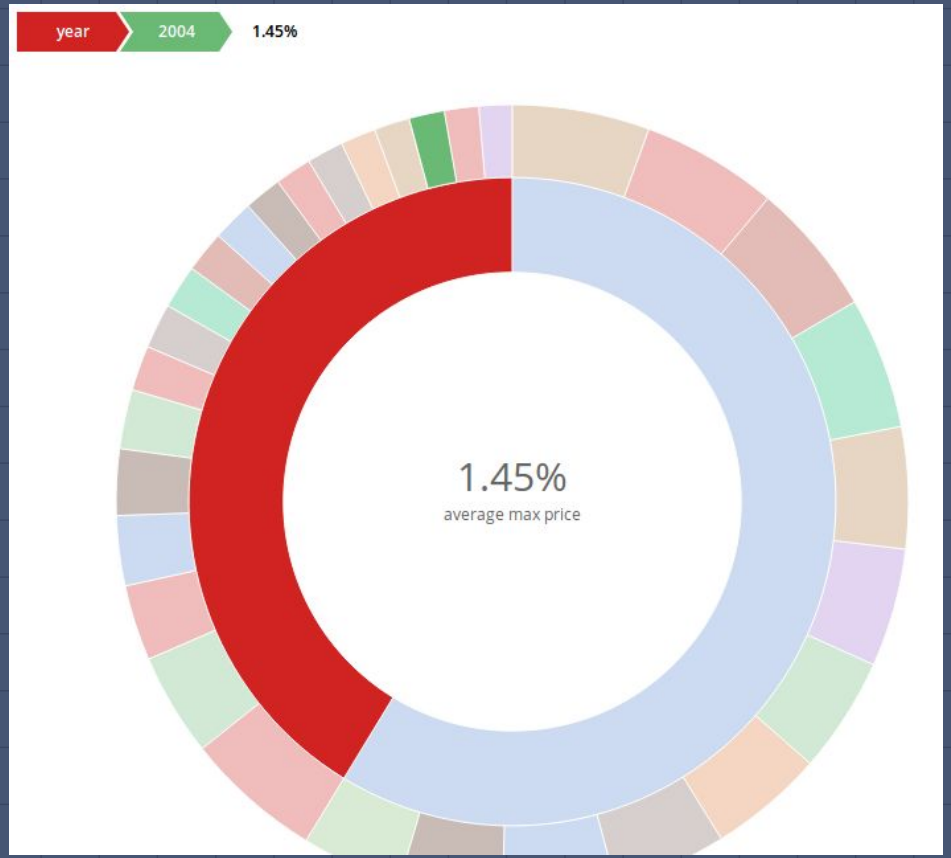
Visualization - D3 (bubble)

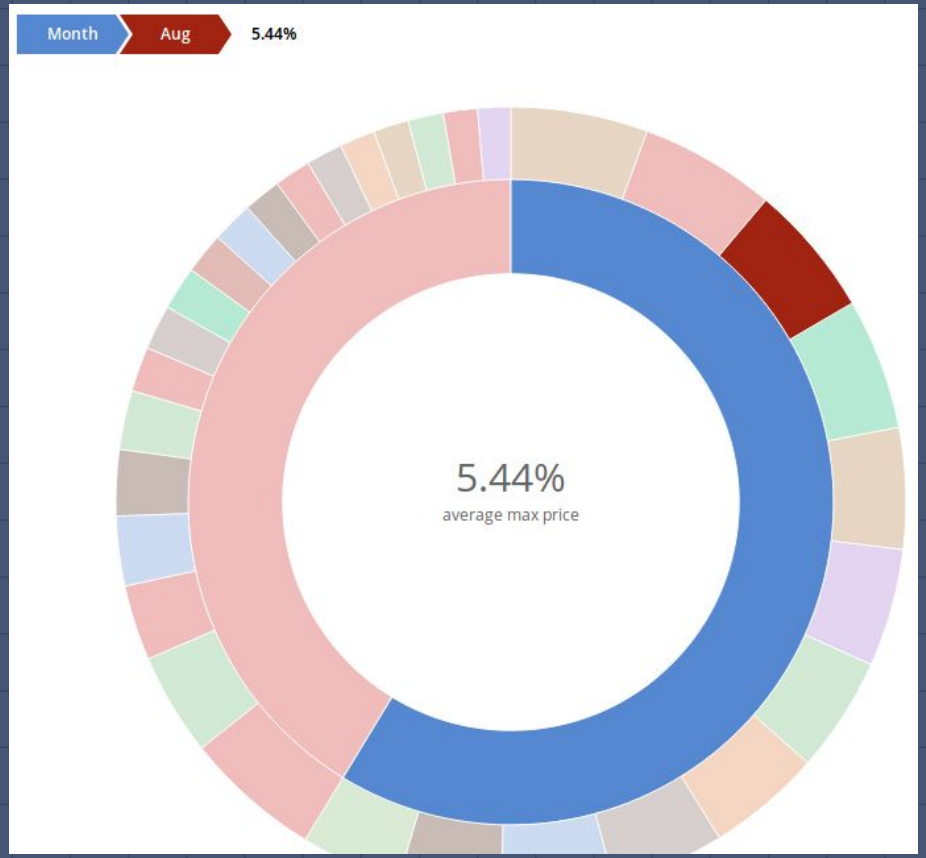


Visualization - D3 (Sunburst)



Visualization - D3 (Sunburst)





Future Scope

- ★ A machine learning model can be trained using the past dataset and supervised machine learning algorithms can be applied to predict the future price of the stock.
- ★ It will help investors know the predicted price of the stock based on the data analytics performed on the past dataset.
- ★ This will also give some insight to the investors whether to invest in a particular stock or not. How profitable the stock can be in future and how much can they expect to benefit from it.

References

1) https://spark.apache.org/docs/2.2.0/streaming-programming-guide.html#dataframe-and-sql-operations
2) https://spark.apache.org/docs/2.2.0/sql-programming-guide.html
3) http://moonshinenews.com/tips-leverage-cloud-big-data/
4) https://www.bloomberg.com/professional/blog/3-ways-big-data-changing-financial-trading/
5) https://www.analyticsvidhya.com/blog/2014/08/big-data-profit-stock-market/
6) https://databricks.com/blog/2015/02/17/introducing-dataframes-in-spark-for-large-scale-data-science.html
7) http://www.signalsolver.com/download-historical-stock-price-data-excel/
8) https://d3js.org/

Thank you!