IST 652

Final Project Report

Scripting for Data Analysis

Aryan Kakade & Durgesh Kulkarni
8-5-2023

IST 652 – Scripting for Data Analysis Final-Project Report

The Data and its Source- The datasets have been downloaded in structured format as a CSV file from Finance Yahoo.com. We obtained three datasets from Tesla, General Motors, and Ford that contain information about their stock prices in terms of volume, closing value, opening value, and maximum and lowest share price for the trading hour. We have also collected the revenue data for the three companies and created a data frame via web scraping.

Source of Data:- 1. https://finance.yahoo.com/

2. https://www.macrotrends.net/stocks/charts/TSLA/tesla/revenue

3. https://www.macrotrends.net/stocks/charts/F/ford-motor/revenue

4. https://www.macrotrends.net/stocks/charts/GM/general-motors/revenue

This Analysis will show which stock is better to invest money and why.

About Dataset – The three datasets each contained 1259 rows and 7 columns. For our analysis, the columns contain crucial information. Columns include:

- Date: The day of trading when all of a stock share's characteristics were recorded.
- Open = the price when the market opened in the morning.
- Close = the price when the market closed in the afternoon.
- High = the highest price during that trading day.
- Low = the lowest price during that trading day.
- Volume = number of shares of the stock traded that day.
- Adj Close (Adjusted Close) = a price adjusted to make prices comparable over time

Data Merging- To make our analysis simple, we combined the three datasets from GM, Ford, and Tesla. This allowed us to compare and integrate all the attributes into a single dataset.

Data Exploration-

- Use of the head and tail function to enhance dataset analysis.
- We checked the unique values, total count, top, frequency, standard deviation, mean, min, and maximum using the describe function.
- Before modifying the data, the data types of each column was checked and was ensured that it belong to correct data type.
- A count of all rows and columns using the shape function to reveal more specific information about the data.
- Year and month column were created so that we can perform analysis according to year and month.

- Visualization such as bar plot line plot and histogram were performed on different attributes of data to perform analysis on different attribute of data.
- Pivot and groupby function of python was used to perform categorization of data and perform analysis on data through different attributes.

Data Cleaning-

- We verified all of the null values for each dataset individually before merging the datasets. In each dataset, there were no null values.
- Dates were converted to datetime data types because this was a more useful unit for comparative analysis.
- There were no duplicate rows when the df.duplicated function was used to count them.
- For a better understanding of the data, all necessary columns were renamed and given proper labelling.

Methods of Analysis-

Questions

1. On an average which stocks are having highest closing price? Compare the stock prices with respect to each year.

Unit of Analysis:- Close

Comparison:- In this analysis, average closing price for each year of every stock was computed to study the growth of stock price over the period of years.

Year	TSLA_Close	FORD_Close	GM_Close
2018	21.069722	9.997882	36.781765
2019	18.235347	9.238651	37.562976
2020	96.665689	7.046996	30.481660
2021	259.998162	14.221587	55.883770
2022	263.093081	14.872470	40.233705
2023	175.034815	12.468765	37.011235

2. On an average which stocks are having highest volume i.e which company stocks are highly traded in the market? Compare the stock volume with respect to each year.

Unit of Analysis:- Volume

Comparison:- In this analysis, average volume for each year of every stock was computed to study the trading trend of company over the period of years.

Year	TSLA_Volume	FORD_Volume	GM_Volume
2018	1.398376e+08	4.220170e+07	1.240585e+07
2019	1.373838e+08	3.828724e+07	8.763544e+06
2020	2.259239e+08	7.670796e+07	1.538326e+07
2021	8.217489e+07	7.886394e+07	1.914958e+07
2022	8.693633e+07	7.266885e+07	1.651395e+07
2023	1.646703e+08	6.605154e+07	1.522573e+07

The Open Price Time Series Visualization gives Tesla the impression that it has always had a considerably higher corporate value than GM and Ford. However, we would need to look at the company's overall market valuation, not simply the stock price, in order to fully comprehend this. Unfortunately, the information regarding the total number of units of the stock present is not available in our current data. However, we can simply multiply the Volume column by the Open price in order to try to depict the entire amount of money moved.

In dataframe, add a new column named "Total Traded" that is equal to the Open Price times the Volume Traded.

3. On an average which stocks are having highest total traded volume i.e which company stocks are highly traded in the market in terms of prices? Compare the total traded column with respect to each year.

Unit of Analysis:- Total traded volume

Comparison:- In this analysis, average total traded volume for each year of every stock was computed to study the total trading trend of company over the period of years.

Year	TSLA_Total_Traded	FORD_Total_Traded	GM_Total_Traded
2018	2.941456e+09	4.164173e+08	4.565476e+08
2019	2.490886e+09	3.527928e+08	3.285970e+08
2020	1.932218e+10	5.229914e+08	4.601738e+08
2021	2.135260e+10	1.140682e+09	1.065080e+09
2022	2.211110e+10	1.145318e+09	6.811781e+08
2023	2.847383e+10	8.211408e+08	5.633885e+08

Daily Change in Percent

To start, we'll figure out the daily percentage change. The formula below defines daily percentage change:

rt=(pt/(pt-1))-1

As a result, the price at time t is defined as being equal to the price at time t-1 (the previous day) minus 1. In essence, this only tells you your percent gain (or loss) if you bought the stock one day and sold it the next. While this is incredibly helpful in assessing the stock's volatility, it isn't necessarily useful when trying to anticipate the stock's future values. The stock is more erratic from one day to the next if daily returns have a wide distribution.

Let’s Determine which stock is the most stable by computing the percent returns and plotting them with a histogram.Each dataframe should have a new column named returns. From the Close price column, this column will be determined. There are two ways to do this: either a straightforward calculation using the.shift() method, which uses the formula above, or the built-in 'pct_change' method in pandas.

4. On an average which stocks are having highest total percent returns compare the returns column with respect to each year.

Unit of Analysis:- Returns

Comparison:- In this analysis, average returns for each year of every stock was computed to study the which stock is best to invest in.

Year	TSLA_Returns	FORD_Returns	GM_Returns
2018	0.001508	-0.002131	-0.000340
2019	0.001391	0.000924	0.000476
2020	0.010047	0.000395	0.001252
2021	0.002188	0.003758	0.001652
2022	-0.003293	-0.001828	-0.001825
2023	0.004432	0.000570	0.000068

Cumulative Return**:** Daily returns are helpful, but they don't immediately show the investor how much money they have made so far, especially if the stock is extremely volatile. The day the investment is made is used to calculate the cumulative return. You are making profits if the cumulative return is greater than 1, otherwise you are losing money.

The question we are attempting to answer using daily cumulative returns is the following: If I had invested $1 in the company at the start of the time series, how much would it be worth today? Due to the fact that it will consider daily returns, this differs from simply looking at the stock price at the time being.

5. On an average which stocks are having highest cumulative percent returns compare the returns column with respect to each year.

Unit of Analysis:- Cumulative Returns

Comparison:- In this analysis, average cumulative returns for each year of every stock was computed to study the which stock provide best returns.

Year	TSLA_Cummulative_Return	FORD_Cummulative_Return	GM_Cummulative_Return
2018	1.075797	0.888837	1.001143
2019	0.930691	0.821944	1.022400
2020	4.933601	0.626957	0.829659
2021	13.269727	1.265266	1.521061
2022	13.427685	1.323173	1.095093
2023	8.933387	1.109321	1.007382

6. On an average which companies generated higher revenue throughout the decade

Unit of Analysis:- Revenue

Comparison:- In this analysis, average revenue for each year of every stock was computed to study the which stock provide best returns.

Year	TSLA_Revenue	Ford_Revenue	GM_Revenue
2010	9.078205e+06	8.107075e+18	9.225085e+18
2011	9.896462e+06	8.479333e+18	9.497592e+18
2012	7.662568e+07	8.932830e+18	9.826844e+18
2013	1.538579e+11	9.392589e+18	1.012135e+19
2014	2.394632e+11	8.967587e+18	9.904348e+18
2015	3.037345e+11	1.006285e+19	5.747597e+18
2016	5.713075e+14	9.663590e+18	8.911847e+18
2017	8.220746e+14	1.033159e+19	9.428834e+18
2018	1.806671e+15	1.044834e+19	9.599839e+18
2019	1.846158e+15	9.928842e+18	7.706589e+18
2020	2.686219e+15	8.988094e+18	9.379589e+18
2021	4.429784e+18	9.419589e+18	8.396067e+18
2022	6.079554e+18	1.099985e+19	1.077710e+19
2023	2.332900e+04	4.147400e+04	3.998500e+04

Program Summary:

- Imported necessary libraries.
- Connected Google colab with the Google drive.
- Reading the csv file (dataset) using pd.read_csv function.
- Using head and tail function on dataset to get insights about the data.
- Converted the 'Date' column in the Tesla GM, and Ford dataframe from a string format to a datetime format using the pandas to_datetime() function.
- Checked for the null values using isnull function to drop them from the dataset.
- Renamed all the necessary columns in all the three dataframes.
- Merged all the three datasets using inner join on date column.
- Extracted year from date column as an important unit for comparative analysis.
- Used groupby function to group the data by year and then calculated the mean values of the volumes for Tesla, GM, and Ford columns for each year.
- Created a line plot showing the adjusted close prices of Tesla, Ford, and GM stock over time.
- Created a line plot of the daily trading volume for the stocks of Tesla, Ford, and GM for the entire time period.
- Compared and plotted the total volume of trades for each stock of Tesla, Ford, and GM over the time period.
- Plotted the daily returns of Tesla, Ford, and GM stocks over the entire time period.
- Created a histogram plot of each stock's daily returns. Three histograms were plotted, one for each stock of GM, Ford, and Tesla.
- Grouping the stock returns of Tesla, Ford, and GM by year and then calculating the mean returns for each year.
- The cumulative returns for each stock in the dataframe were calculated. Using the cumprod() function we determine the cumulative product of 1 plus daily return and created a new column to represent each stock's cumulative return.
- Plotted the cumulative returns for Tesla, Ford, and GM over period of time.
- Using the groupby function, we calculated the mean of the cumulative returns for the stocks of Tesla, GM, and Ford by year.
- Created a bar plot of the yearly mean cumulative return for the stocks of Tesla, Ford, and GM.
- We created a new dataframe using the groupby function by month and calculated the average of Tesla, Ford, and GM's cumulative returns.
- Gathered the revenue data by web scraping.
- Combined all the revenue data in single data frame.
- Created barplot by different years to study revenue data.

Description of result of the analysis:-

The data analysis program processed a dataset of stock market from a csv file for a different type of companies for different dates. Also, we had got revenue data for companies through web scraping. After performing various steps such as data cleaning, exploration and comparison following conclusion were drawn.

a. Overall Tesla has good attributes to invest in as it has highest growth in stock prices, volume and total traded.

b. Tesla has highest cumulative and daily returns compared to GM and Ford therefore there is high chance that one will get highest profit by investing in Tesla but also tesla has highest fluctuations in returns.

From our analysis, we can see that Tesla's stock outperformed Ford and GM in terms of price and cumulative returns over the given period. Ford and GM stocks have remained relatively stable in comparison, and their performance has been impacted by various factors like economic conditions, production issues, and the COVID-19 pandemic. We also observed that Tesla's trading volume increased significantly from 2019 onwards, while Ford and GM's trading volume remained relatively stable.

Our analysis suggests that investing in Tesla's stock might provide higher returns in the long run, but it is always recommended to do thorough research and analysis before making any investment decisions. The stock market is volatile, and several factors impact the performance of stocks, making it important to have a diversified investment portfolio.

Overall, our analysis provides insights into the performance of three significant players in the automobile industry and can be helpful in making informed investment decisions.

Project Contribution:-

1. Aryan Kakade :- Data collection, pre-processing and munging. Also contributed to analysis for close and volume.

2. Durgesh Kulkarni:- Data Visualization and web scraping. Also created analysis for all remaining unit of analysis