Advanced Data Analysis and Machine Learning (2023-2024)
Computational Engineering
LUT University

Google Stock Dataset: Week 2

Alejandro López Vargas
Sina Saeedi
Pau Ventura Rodríguez

November 13, 2023

1. Introduction

Visit our GitHub to gain access to the script.
In this project, we aim to analyze the historical data of Google's (GOOGL) stock price and create a predictive model for it. The dataset includes the following variables: Date, Open, High, Low, Close, Volume, and Name. The 'Name' variable is constant and represents the company's name, which is Google in this case, making it unnecessary for our analysis.
Here are brief explanations about each of the variables:

1. **Date:** This variable represents the specific dates corresponding to each data point in the dataset. It serves as the timeline for the stock's historical performance.

2. **Open:** The opening price of the Google stock on a particular date. This is the price at which the stock begins trading at the beginning of the trading day.

3. **High:** The highest price at which Google stock traded during a given day. It provides insights into the highest level the stock reached within a trading session.

4. **Low:** The lowest price at which Google stock traded during a given day. It indicates the lowest level the stock reached within a trading session.

5. **Close:** The closing price of Google stock on a specific date. It is the final price at which the stock is traded for the day, representing the market sentiment at the close of the trading session.

6. **Volume:** The total number of shares of Google stock traded on a given day. Volume is a crucial indicator of market activity, providing insights into the level of interest and participation in the stock.

Project's Goal:
The main objective of this project is to develop a predictive model that can anticipate future closing prices(Close) of Google stock based on historical patterns and trends. By drawing insights from the information contained in the Open, High, Low, Close, and Volume variables, we aim to create a reliable model that can assist investors and traders in making reasonable decisions about Google stock.

2. Strategies

1. As the interval is closed within a day, a good way to start would be to predict the high and the low values, to know which one is the stock value that is the best for buying (high) and selling (low) in the day.

2. It could be that the high or low value is never actually reached, and the model would freeze during that day as it would never buy nor sell. In that case, a possible improvement would be to predict at what time of the day will the high/low happen.

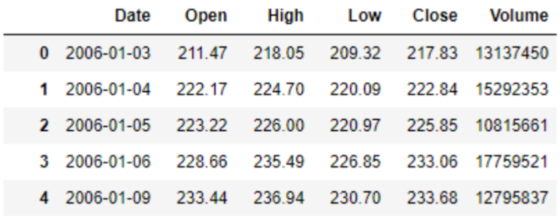
This implementation is not possible as it does not have pieces of information smaller than a day.

3. However, we have the value of the stock at certain points of the day (at the beginning and at the close), which means that we have two points in time where we know exactly the value of the stock. Another implementation would then be to find within a week, the best possible time to buy or sell, buying on the lowest day and selling on the highest day. This would ensure that the money is constantly moving and the freezing point mentioned in (1) is not reached.

4. A more basic model would be to, given the open price, try to predict if the closing price will be higher than the open, and thus it would be worth it to buy on the open and then to sell in the close.

3. Preprocessing

Before performing any visualization, we have to convert the Date column on the dataframe to format Datetime in order to give it an ordinal value. The column "Name" only has one unique value, which does not give us any information on the target and will then be dropped.
We also have to check if there are any NaN values, as they could disturb our analysis. As we will see afterwards, the stock market is only open on business days, which means we do not have data for all days of the year. Furthermore, some holidays did also not register data like Christmas, which we will create in order to have an equally spaced dataframe. To fill up those values, we can use Backfilling (setting up the values with the next datapoint's values), Frontfilling (setting up the values with the previous datapoint's values), Meanfilling (assigning the mean value of that feature, which in this case will not yield good results due to the increasing tendency). In this case, we will use Frontfilling.
Note that the price on the open is not equal to the price of the close from the day before, which might require a bit more understanding of the dataset.

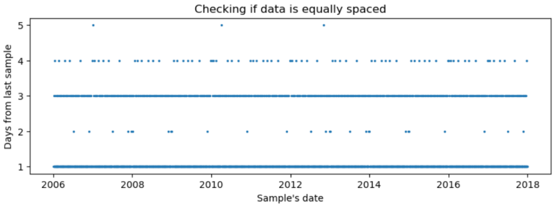


	Date	Open	High	Low	Close	Volume
0	2006-01-03	211.47	218.05	209.32	217.83	13137450
1	2006-01-04	222.17	224.70	220.09	222.84	15292353
2	2006-01-05	223.22	226.00	220.97	225.85	10815661
3	2006-01-06	228.66	235.49	226.85	233.06	17759521
4	2006-01-09	233.44	236.94	230.70	233.68	12795837

4. Data Visualization

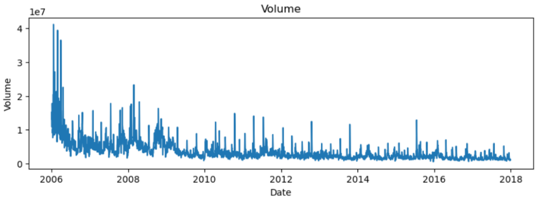
4.1. Date Distribution

Another important step is to check if all the data points are equally distanced between them.



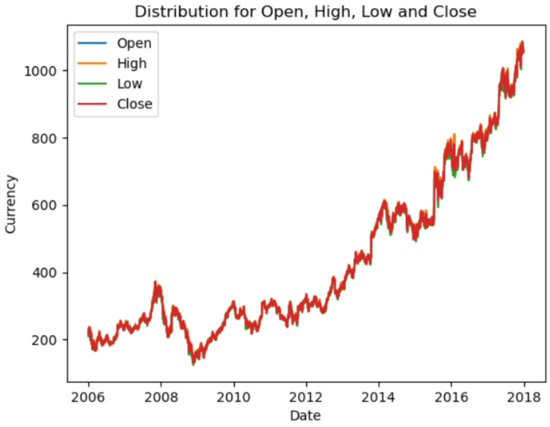
The general case is that two days consecutively have samples assigned. Observing our samples we can see that in general we have samples for 5 days of a week, and then there are two days missing (days from the last sample= 3). This makes sense as the stock market is only open on weekdays, so we do not have data on weekends. However, this is not the only case, as there have been times when the stock market has been closed for 1 day, 3 days, or even 4 days at maximum. This might occur on holidays or dates like Christmas. These anomalies can alter our analysis and we will focus on them lately.

4.2. Volume Distribution

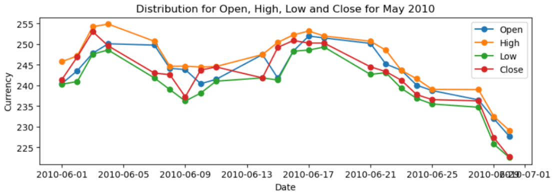


It seems like Google's stock market has lowered its popularity over the years, with some peaks throughout time, but the general tendency seems to have decreased.

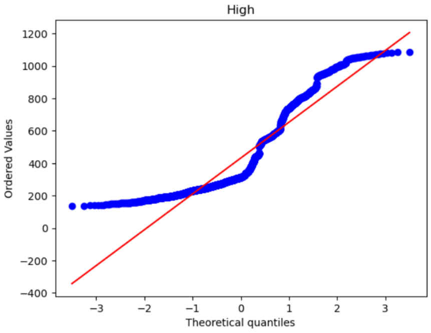
4.3. Prices Distribution



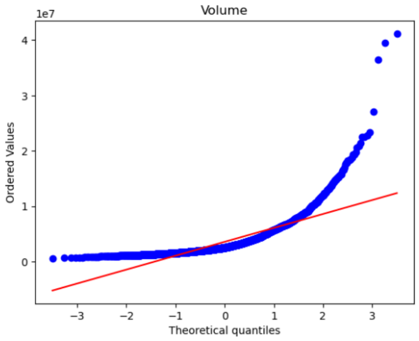
Although the measures are overlapping, we can see the shape and behavior of the prices during the years. It seems like until 2010 it remained on average constant, but after that it started increasing drastically, going from 200 in 2010 to surpassing 1000 in 2018.
If we take a deeper look into a short period of time, we can further understand the variables.



As we can see, the High variable works as an upper bound for all other variables, whereas the Low variable works as a lower bound for them.
For further analysis, we will check if our features follow any normal distribution, which is not usually the case for time series. For doing so, we will use a probplot.



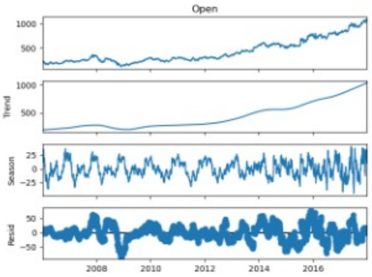
All prices variables follow this same probplot, which indicates that for values higher than 600 the distribution is more close to a normal than for values under 600. Anyways, these features do definitely not follow a normal distribution.

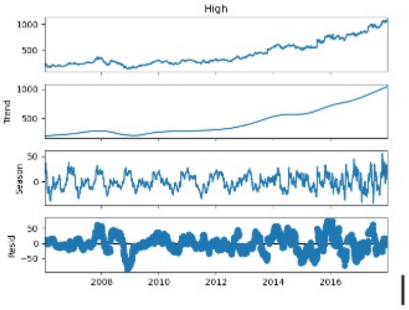


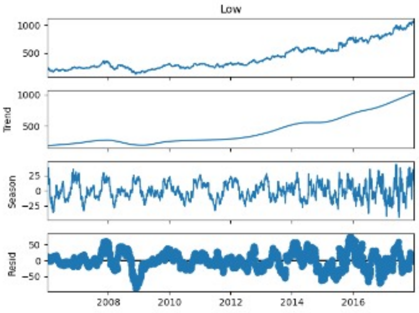
The Volume variable does not as well follow a normal distribution, as most values are very far from the red line.

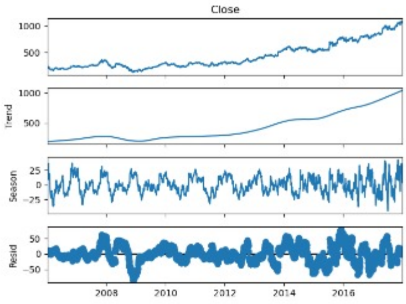
4.4. Trends

The features open, close, high and low have very similar distributions, so they will also have similar trends and seasonalities as we can see below:

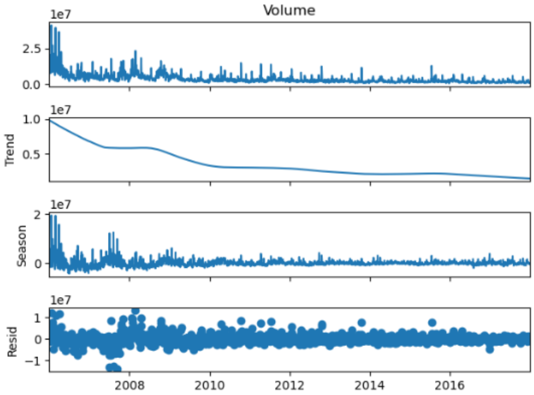








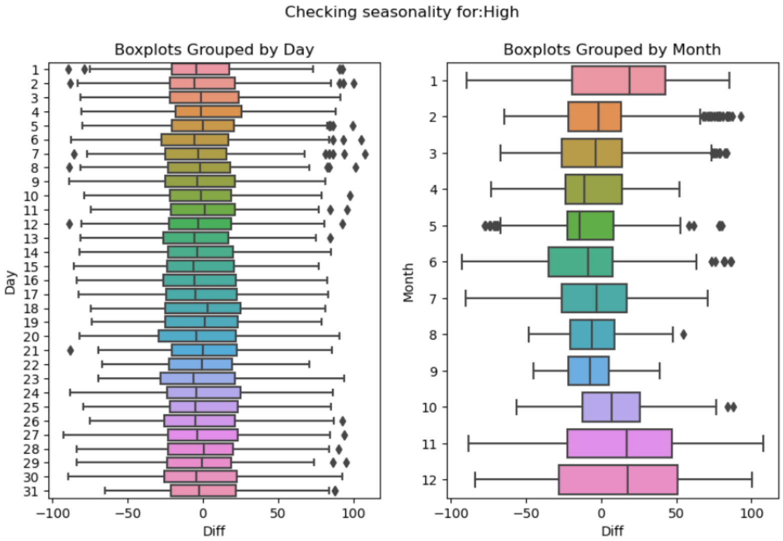
The trend is clearly increasing throughout the years. For the seasonalities, at least until 2014, there seems to be a small peaks followed by a large peak and a big decrease for approximately each year.



The volume however, looks very noisy, specially on the first years (probably with the boom) the season captures a lot of variance, which we were not able to detect any pattern.

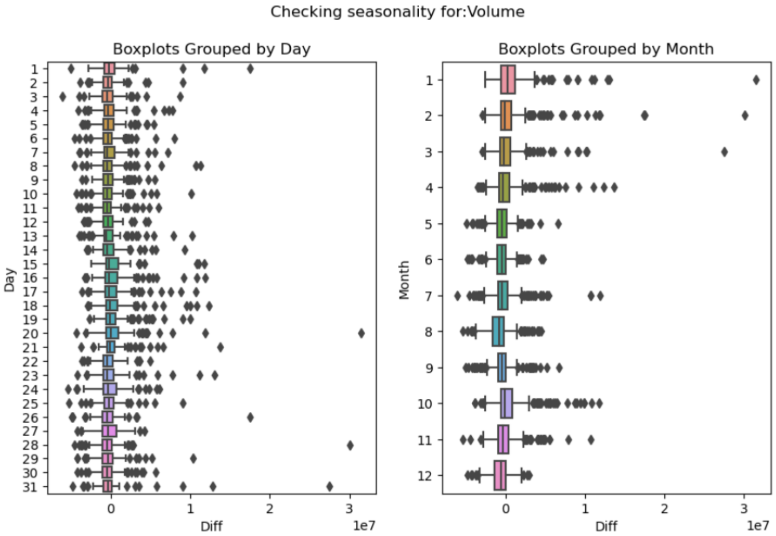
4.5. Looking for seasonality

To look for seasonality, we take out the trends for each variable and then group by month and days. The following boxplots then represent the samples distribution for each group.



The days seem to have not a lot of variance, whereas although the months are not completely different between each other, the distributions seem to be different. In winter months (November, December and January) the mean is a bit higher, which indicates that the price tends to be higher on those months overall (then probably a good strategy would be to buy in summer and sell in winter). However, winter months tend to have a higher variance as well.

Other price variables (low, open, close) have a very similar distribution, so we won't comment on them (you can check the code for that).



For the volume, we know that this variable is very noisy, and this can be seen on the amount of outliers. Overall, we cannot take that many insights from their boxplots, as the distributions are very similar among months and days.