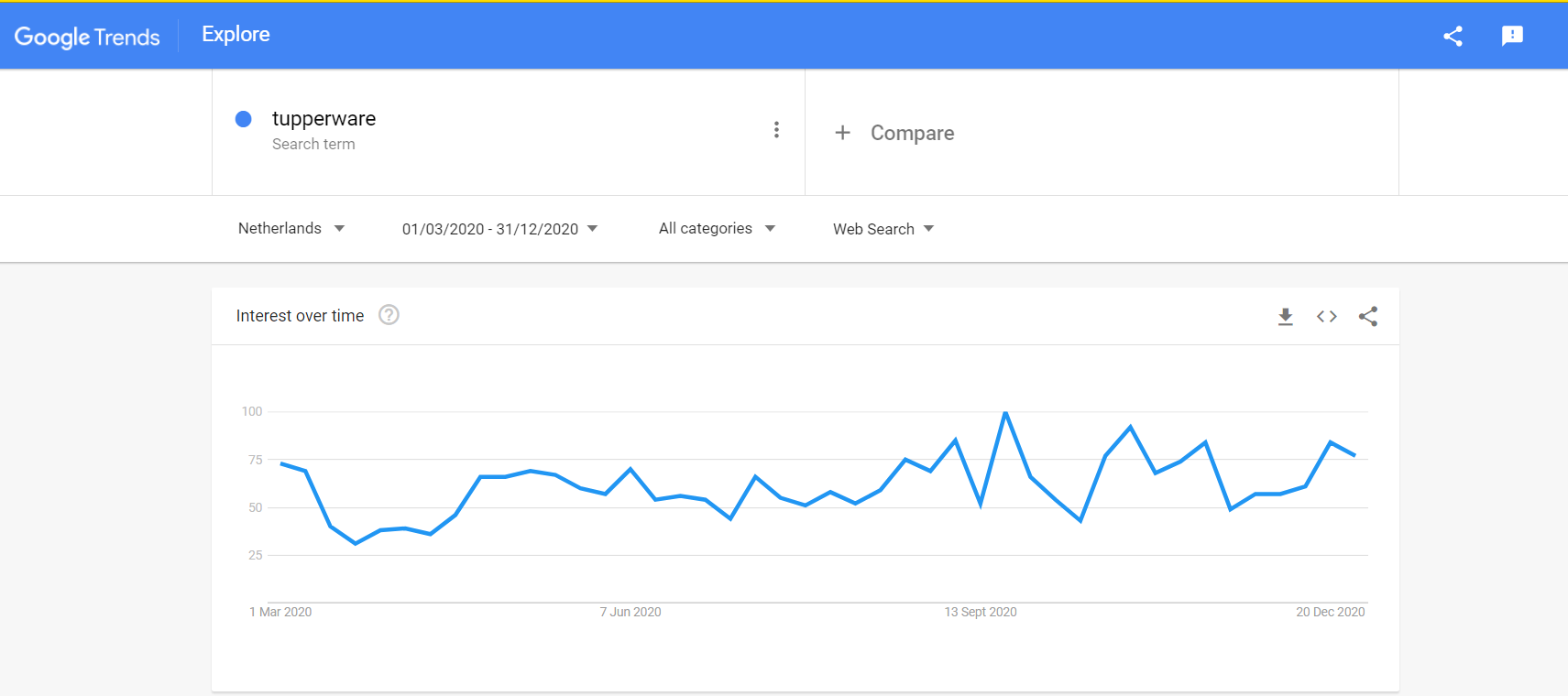
**Google Searches and Economic Performance**

Consumer confidence is one of the factors which affects demand in the economy, and we based our project idea on this theory. Given that E-commerce has boomed during the pandemic and the majority of sales are online, we think instantaneous data is more appropriate to assess the health of the economy, especially during the volatile year that was 2020. We used data about what people feel about the economy and we used this to get a correlation with economic performance. As data such as consumer confidence and economic performance are difficult to measure and obtain, we used dummy variables instead.

**Datasets**

People search for things they are interested in and this signals to the market what they are interested in buying. We use word search data in the form of Google trends as a dummy for what people feel about the economy or how much of this good they are buying. We selected five words which we thought would represent the outlook of the economy, for example, the word ‘tupperware’ corresponds to the large increase in food deliveries and thus the market for tupperware is growing now. We used the Python API Pytrends to automatically load this data into our file.   
We used data over the period of March - December 2020, as this time period captures the different waves of the virus. All word searches are in the Netherlands.



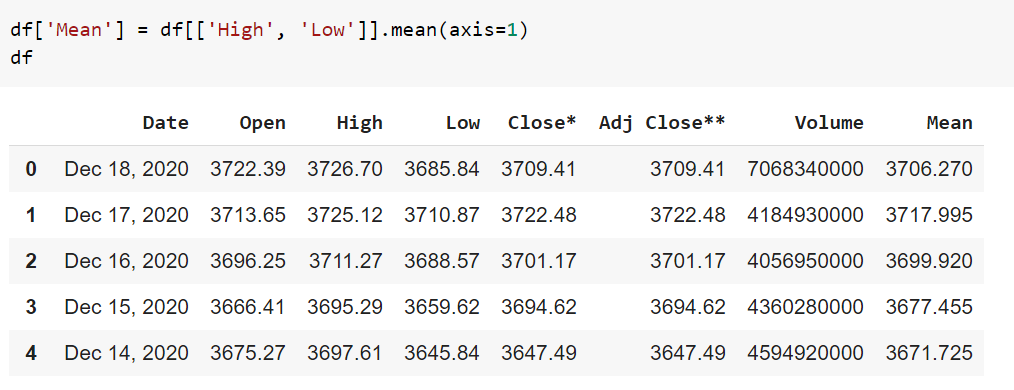
Our project essentially consists of working with 2 datasets: word search frequencies, and stock market data. We used S&P 500 data as a dummy for economic performance. We did some web scraping to load this data frame into our file.

**Hypothesis**

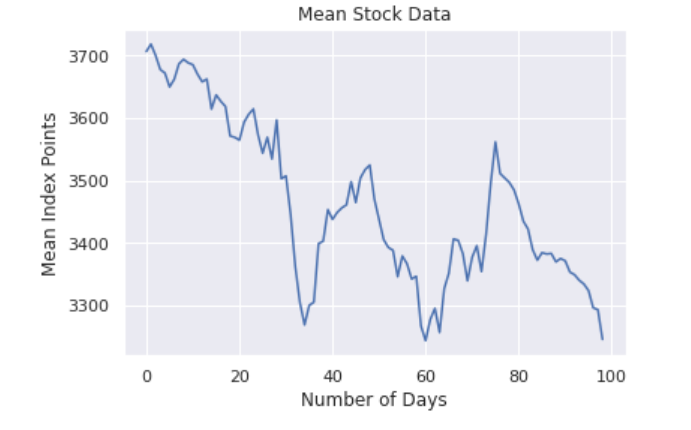
We hypothesize that increased frequency of words that are positive indicators of consumer preference will correspond to a positive trend in S&P 500, and an analogous argument for negative indicators. E.g increased searches of the word ‘buy’ will correspond to a positive S&P trend.

**Methodology**

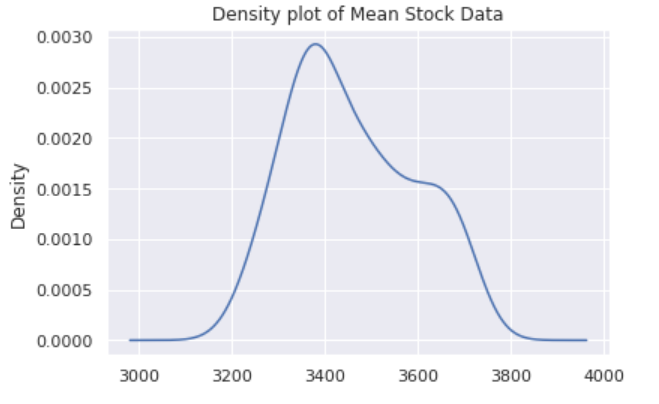
The stock market data had several columns, so we did some cleaning and data manipulation to obtain the average index points on each date, which was not available in the original data frame. All the data had to be converted into a numeric form first.



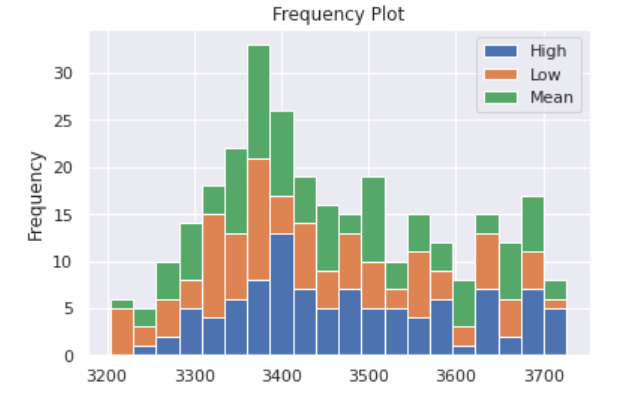
We then observed what the stock data looked like over time using several visualisations:

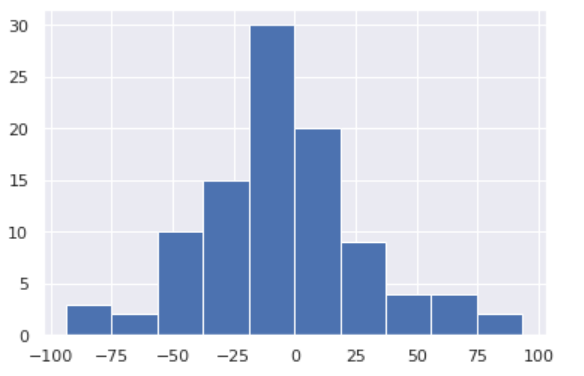


Here we see that the general trend is declining, which we know is true as markets generally performed poorly during the pandemic.

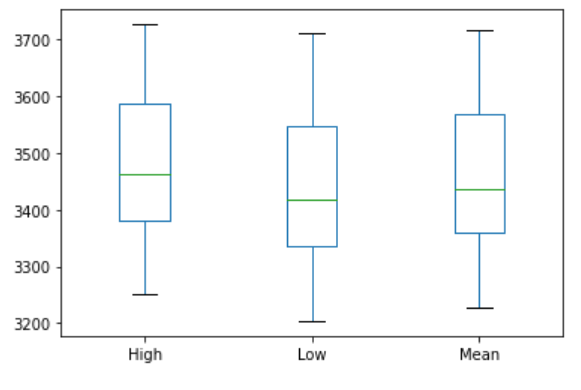


The following histogram and frequency plots show the average difference between the mean stock data and between the High and Low values. We visualise them to get an idea of the deviation of the Mean data we are using.

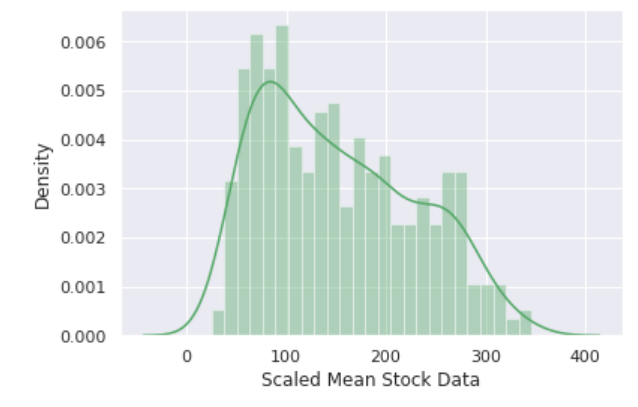




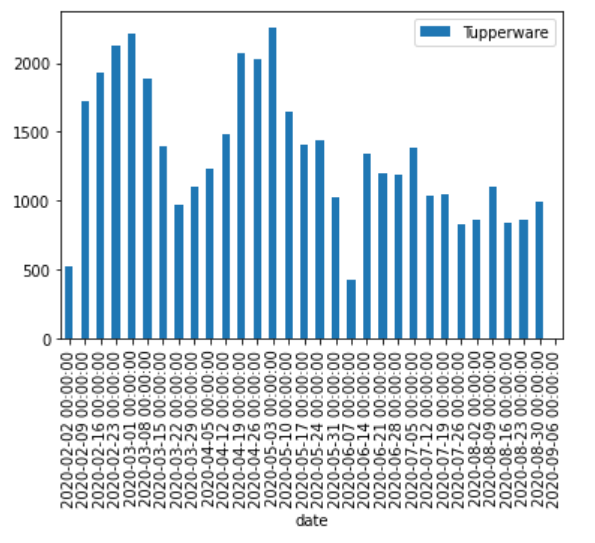
This boxplot affirms that it is a good decision to use the mean stock data.



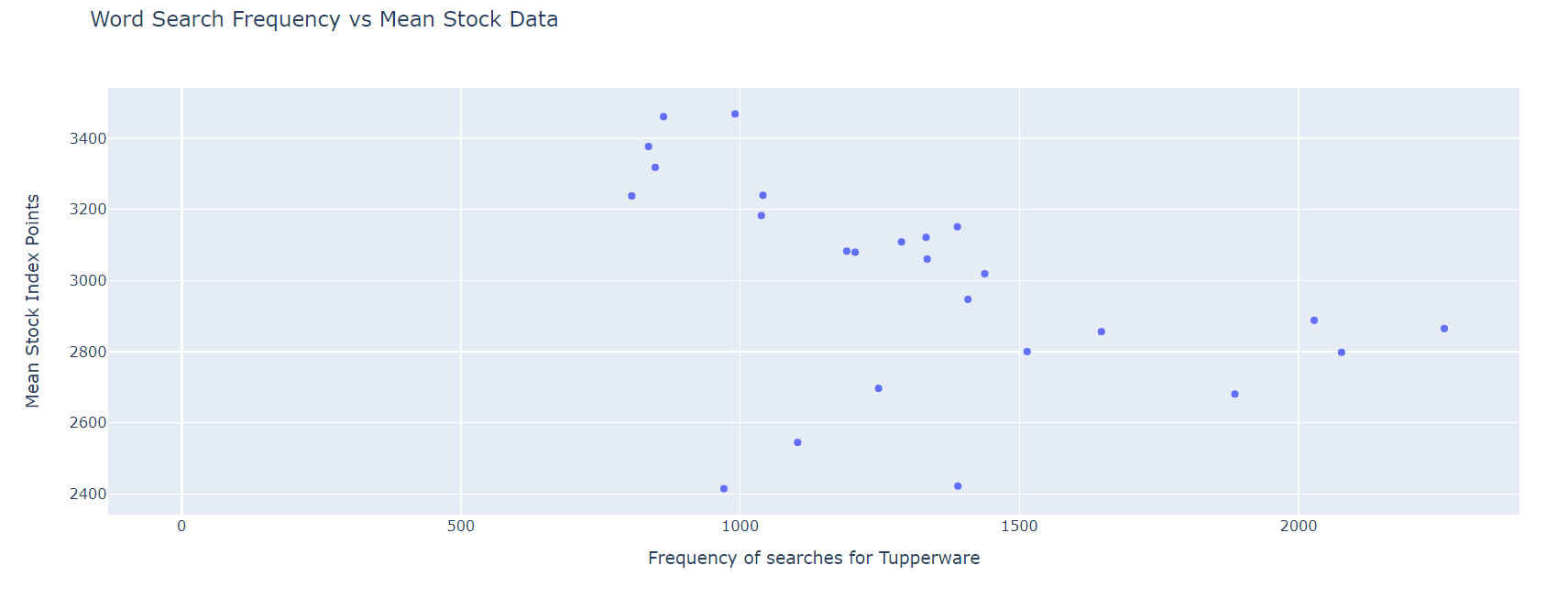
To study the statistical properties of the stock data further, I created and sorted the data into bins and scaled it to display the distribution curve:



We can also see the declining trend here.

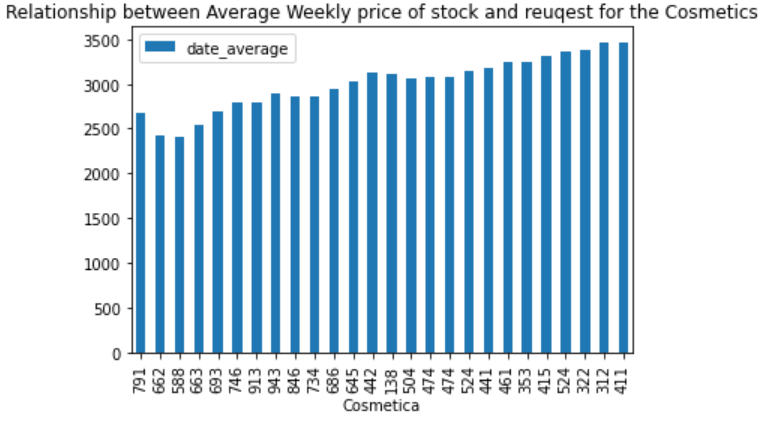
Now, on to our Pytrends data. We plotted the word search frequencies for each word. This is the frequency graph for the word Tupperware on a selected subset of dates:  


We then merged our two datasets with the date as an index to finally observe the relation between word search frequency and mean stock data. Below is the scatterplot for the word tupperware:



We do not see a clear relation here.

With a different plot type, we can also observe the relationship for the word ‘cosmetics’:



We see that there is generally a positive trend between word search frequency and stock price here, which is in agreement with our hypothesis.

Initially, I also wanted to make word clouds and use the stock price data as a reference for the size of the words, but unfortunately did not get around to it. This could be a possible extension for the project.

**Potential Issues and how we (did not) fix them**

Since the relationship we are trying to assess has multiple underlying factors, we might be omitting important variables and not accounting for error terms.

The stock data is not available for weekends and public holidays, so we instead took the weekly stock price average.

What is good about our project is that we can validate our model to an extent by comparing against official statistics about quarterly economic performance. Qualitatively speaking, our model works to an extent.

**Conclusion**

We employed a simple methodology which gives us sufficient results in the scope of this project and allows us to simply visualise some complex aspects about the economy. The graphical relationships suggest the correlation that we expected for cosmetics, but not tupperware, and the correlation factor for cosmetics is not too high. This may be due to missing variables as mentioned above, and also because we used S&P data which is based on US markets. It was difficult to get exact information about each industry, such as, the market for cosmetics. So we used general economic performance data instead. Also, cosmetics is a broad term for an industry whereas Tupperware is just one good, so perhaps applying our methodology to whole industries rather than specific goods is better.

As an extension, we can perform linear regression on the data to get a more accurate mathematical relationship between our two variables.