Andrew Plum

Professor Ma

CS 479

2/23/2024

Data Cleansing and Exploratory Data Analysis

1. Loading, Transforming and Saving the Dataset

The logical organization of the data changed when I added two extra columns and changed the data formats of two columns. The physical organization of the data changed from being contained in CSV and TXT files to being stored in a data frame whilst working with the data to only being contained in CSV files. I loaded the initial dataset stored in a CSV file using read.csv(). From there, I dropped all rows with null values. I reformatted the DATE column into the date format as well as converting the SP500 column from char format to numeric format. I then added two columns. I added a column called TEXT\_OF\_SPEECH whose rows contained all of the text of its respective speech listed in CHAIR\_POWELL\_FED\_SPEECH\_NAMES and a column called SP500\_DAILY\_GAIN which calculated the daily percentage change over each trading day. Note for this calculation, the first row couldn’t be calculated because there was no initial value to subtract the final value from so I calculated this before I dropped all of the rows with NULL values (I dropped all rows with NULL values in the SP500 column prior to this for the calculation). After this, I performed the exploratory data analysis where I made times series, histograms, a word cloud, and an initial and final summary of the data frame. I chose to do a time series plot over a scatter plot or box plot primarily because my dataset has a time series feature. I created histograms because they seemed like they could be insightful. And I made a word cloud of the text in the speeches since that is a good way to visualize the most prevalent text in them. After the exploratory data analysis, I saved the data frame as a CSV file using write.csv().

1. Features of Dataset Found from Exploratory Data Analysis

A screenshot of a computer

Description automatically generated

A screenshot of a computer

Description automatically generated

A graph showing the price of a stock market

Description automatically generated

A graph showing a line graph

Description automatically generated

A graph of a sound wave

Description automatically generated

A graph with lines on it

Description automatically generated

A graph with a bar graph

Description automatically generated with medium confidence

A graph of a bar graph

Description automatically generated

A graph of a graph

Description automatically generated

A graph of a number

Description automatically generated

A close up of words

Description automatically generated

There were some things I already knew about the dataset which could be seen in the exploratory data analysis I did. For example, I knew that the S&P 500 was relatively volatile during the onset of 2020 due to the uncertainty surrounding the Covid 19 pandemic and this could be seen in the “S&P 500 Daily % Gain Time Series All Days” time series. Looking at this time series, although I knew the time around 2020 had some sharp volatility, it surprised me to see that the time around 2022 was more consistently volatile for a more sustained amount of time. Another thing that I noticed comparing this time series and the “S&P 500 Daily % Gain Time Series Fed Speech Days” time series was that volatility of the year 2020 was less extreme when looking at only the days Jerome Powell, the Federal Reserve Chair, spoke (this was the latter time series). In this latter time series, there was more volatility in 2018, 2019, and 2022. From personal memory, I remember that there was more discourse in financial markets about interest rates changing, and I believe this helps explain the discrepancy between the two time series. On the days that Jerome Powell gave a speech, it seems the stock market was primarily concerned with any change in monetary policy signaled by him instead of being concerned about Covid. This makes sense to me because the Federal Reserve mainly exerts control over the economy through monetary policy with abilities like indirectly adjusting interest rates which in turn affects the stock market. Another big take away, which explains the consistent volatility in the S&P 500 in 2022 was that inflation was a relatively big concern. I knew previously that inflation was the highest it had been in roughly 40 years in 2022 and this was primarily why it was the largest word by quite a bit in the word cloud I made.

1. Difficulties Encountered and Solutions Developed in Conversion Process

Although I was able to figure everything out, there were some problems along the way that I encountered. I initially had a problem I think with whitespace when generating the word cloud. It did not seem to be considering all of the words from the speeches and I believe this was because the program would stop retrieving the text from each speech once it encountered whitespace (although regular spaces seemed to be ignored) with the initial method I used. I changed my approach to where I stored each word in a single character vector and then used that vector to generate the word cloud. The other issue I ran into was the data formats of the vectors when generating the time series. This took me longer than I wanted to troubleshoot because it was not giving me any errors since technically the time series were generating; they just did not look the way they should. I fixed this by converting the DATE column from char format to date format and changing SP500 from char format to numeric format. Overall, both problems were not too difficult to address as I was able to solve them after looking at the relevant R programming documentation.