Javier Lozano

Dr. Steve Levkoff

ECON 494-01 – Intro to Business Analytics

October 25, 2020

Project I: Descriptive Analytics Data Mining Project – Summary Report

1. **Collecting and Cleaning the Data in Excel**
2. I looked up the S&P 500 Index and three individual stocks, Apple (AAPL), Microsoft (MSFT), and Facebook (FB) on Yahoo! Finance, that represent companies in the technology industry and are within the S&P 500 Index. I downloaded the .csv files to being my cleaning process in Excel.
3. I cleaned the data from each .csv file using Excel and deleted four columns in each file. Those columns were the “Open”, “High”, Low”, and “Close”. The idea behind that being that the only important variables to consider in this analysis are the “Date”, “Adj. Close”, and “Volume”. I manually added two more columns that calculate the percentage change of the adjusted closing stock price of the companies and index from one year to the next and another column highlighting if COVID-19 was present or not at the time with “0” being there was no coronavirus yet and “1” being coronavirus was present. This column would act as the analysis’ categorical variable. According to my online research, COVID-19 was first officially recorded in Wuhan City, China in December 2019.
4. The dates that are being analyzed are from the past 36 months, starting on October 2017 and ending in September 2020. The adjusted close amends a stock’s closing price to reflect that stock’s value after accounting for any corporate actions, this allows us to have a detailed analysis of the companies’ and index’s performance over the last year. The volume is the physical number of shares traded of the stock for the given period, in this case starting from October 2017 and ending in September 2020.
5. The main purpose of the period selected (10/1/17 – 9/1/20) is to identify the impact of COVID-19 on the companies’ and the S&P 500’s share prices accounting for dividend yields, stock splits, and new offerings. With having a longer period to reference the adjusted closing prices and volumes when COVID-19 was not present, it will be easier to identify the impact of the pandemic in the graphs created using the R code. The key things to consider are if the pandemic affected the companies and index significantly or not, and if so if they were able to recover or if they currently are on the road to recovery. In other words, if the goal is to identify and illustrate the companies’ returns distribution before and after COVID-19.
6. **Importing the Data & Preliminary Exploratory Analysis**
7. After identifying what the data is for and its purpose, the data of the companies and index are imported to R Studio. There are two options presented in the R code, the option to import the data sets from the internet or locally. The purpose of this being that not everyone who would look at the project will have the exact data locally, so it is pivotal to make the data sets available to everyone so that there are no problems when it comes to analyzing the data with the specific columns and variables.
8. Once imported, it is then viewed and used for generating summary statistics for each data set. The functions utilized were View() and summary().
9. AAPL, MSFT, and FB are all competitors in the technology sector. In this analysis, we are looking into their performance, along with the S&P 500's, to see if we can identify significant impacts due to COVID-19. To get a solid idea of the adjusted close trends, a summary for each data set’s adjusted close was ran.
10. The same process of viewing the summary after isolating the adjusted close variable was applied to the volume variable for each data set to identify the shares sold during the period.
11. To clearly have a visual reference of the overall performance of the companies’ and the index’s adjusted closes and volumes over the period selected, histograms with normal calibrated density curves best represent the visual activity of the data sets to any person running or viewing the code. Making it even easier to comprehend the data, Apple’s histograms will have a red density curve, Microsoft’s is green, Facebook’s is dark blue, and the S&P 500’s is yellow.
12. By looking at the histograms with the density curves it can be clear that there is a skew to the right of each data set’s graph. This skewness represents the negative impact that COVID-19 had on the companies and index on a broad scale by showing the increase in the adjusted close prices. To clearly identify the periods in which COVID-19 truly started to impact the company it is crucial to dive deeper into a more specific analysis. For that I installed the ggplot2 data visualization package for statistical programming to optimize my graphs and data visualizations.
13. **ggplot2 Exploratory Analysis**
14. Once ggplot2 is part of your machine, the idea is to begin connecting the pieces and optimizing your data information. In my case, I grabbed each company’s and index’s adjusted close and volume information and set up the adjusted close as my x variable and the volume as my y variable. This allows anyone to clearly see the volume of how many shares were sold over their current closing price.
15. The first action taken was creating a scatter plot for each data set with the dots being the same color as the density curve lines back when we crated histograms. The idea of the color-correlation is to allow a visual connection across the entire work process. As for the scatter plots, each data set has its peak shares sold with their corresponding value. For example, for AAPL it’s 6e+09 shares sold at around $63, for MSFT it’s 1.6e+09 shares sold at around $158, for FB it’s 1e+09 shares sold at around %158, and the S&P 500 Index fund’s shares sold at a certain value are going to be much greater given that it’s an index fund. Its peak shares sold was at 1.6e+11 at around $2,610. This data illustrates that the other points highlight the number of shares sold and their corresponding adjusted closing price, but it is crucial to keep the peaks in mind. In all the scatter plots created in this case, if examined from a general perspective, their peaks are at around the same coordinates in their corresponding scatter plots, except for MSFT’s peak which is a little bit more to the center than the rests which are to the left.
16. In order to obtain a more specific view of COVID-19 affecting the companies and index and where their peaks fall, I added the facet\_wrap() function to each data’s scatter plot in order to identify the shares sold and their adjusted closing price for when coronavirus wasn’t present and when it was present (keeping the color-correlation). The facet\_wrap() function generates to separate scatter plots using the same information from the selected data set selected. The reason for two scatter plots is because there is a categorical variable, in this case “COVID.19”, that allocates the data into categories. The categories are “0” for when there was no pandemic and “1” when there was pandemic. After creating the two scatter plots for each data set, I noticed that there is constant trend between all of them. There are way more points plotted in the pre-COVID plot than in the post-COVID plot, but in the post-COVID plot the peak of shares sold of every data set is presented in the post-COVID plots, except for Facebook’s plots. Facebook’s peak is in the pre-COVID plot. The illustration that shows there are more points plotted in the pre-COVID scatter plot indicates that there were more shares sold at the time because the pre-COVID term is longer than the post-COVID time frame, and there is also a much more dense concentration in low adjusted closing prices in the pre-COVID scatter plot as opposed to the post-COVID scatter plot where the adjusted closing price is not as concentrated in some particular adjusted closing prices but spread out evenly across the whole plot. This indicates that as the market got riskier to invest in, its unpredictability led traders to keep on trading or to completely stop their bets, despite the extreme downward and upward movements in the stock market during this past year since the pandemic began.
17. To make the information presented in the facet wrapped information much clearer to the eye, I decided to add a trend line connecting the plotted information. It demonstrates the clear trend that prices before the pandemic were way less than what they are worth currently with the pandemic present and a shows the breathing room that the post-COVID dots have as opposed to the pre-COVID dots. The main difference across all the data sets is that Facebook’s peak, as mentioned before, is in the pre-COVID plot yet it still demonstrates way more dots given that the pre-COVID term that is being analyzed in the data sets was longer than the post-COVID term.
18. The main idea I wanted to illustrate was the company and index’s consistency in adjusted closing prices and their corresponding volume. In simpler terms identify visually the number of shares sold overall over the past 3 years with their corresponding adjusted closing price. For that I created violin charts for each data set by adding the geom\_violin() function. I’d say every data set shared the similar aspect of having a wide base and a pyramid-like shape with the peak being extended into a small upside down triangle at the top, representing the dense amount of shares sold at the prices in the x-axis and having the upside down triangle at the top represent the peak of shares sold at their corresponding adjusted closing price.
19. The violin plots with the facet\_wrap function are to be considered as well because those give a more specific sense of the number of shares sold over the past 3 years indicating their periods in time (which are pre-COVID and post-COVID). The clear trend between all of them is that the majority of their shares sold at a high adjusted closing price were in the post-COVID category and including their peaks (except for Facebook once again, which has its mostly sold shares still in the post-COVID category but its peak is in the pre-COVID category). This shows the amount of diversity in the adjusted closing prices post-COVID. And this one more demonstrates that as the market gets riskier, its unpredictability leads traders to keep on trading or to completely stop in both an optimistic and a pessimistic fashion and mainly without as much knowledge of the market, despite the sudden movements in the stock market during this past year since the coming of COVID-19.
20. The next thing added to keep analyzing the relationship between the adjusted closing price and the volume of shares sold pre and post COVID-19 is the box and whiskers plot for each data set. In this case I only selected the adjusted closing price to be our data that would be depicted into groups of numerical data through its quartiles. The lines outside the box indicate variability outside the upper and lower quartiles. These box and whiskers plots, like the facet wrapped violin plots, are yet another visualization of how the adjusted closing prices varied and increased more during the pandemic as opposed to when there was no pandemic. Highlighting how uncertainty leads to more conservative actions by each company. What the box and whiskers plot does not illustrate is how many observations are in each graph, so to understand the observations and their gravity a scatter plot would be more useful. But when it comes to analyzing only the diversification of the adjusted close price, then box and whiskers plot is direct and concise.
21. The frequency polygon is like a histogram, compares data and displays a cumulative frequency distribution. It uses a line graph to represent the quantitative data. When these graphs are created from a general perspective, it reflects how there have been more months throughout the past 3 years that have had low adjusted closing prices for each data set. By not taking account the periods of when COVID-19 was present and when it was not, the data can still be able to illustrate the impact of how COVID-19 made the adjusted closing prices rise.
22. Lastly, to demonstrate the TIDY data, I created a histogram with the geom\_vline() and geom\_density() functions to illustrate my main point that I had made way in the beginning when I used regular histograms to illustrate the data. The adjusted closing prices increased drastically when the pandemic began. There is a skewness to the right given the greater number of observations that are pre-COVID rather than post-COVID. However, it is important to remember that pre-COVID adjusted closing prices were much lower that what they are now. This tells us that there the companies that we analyzed have been decreasing dividends, stock splits, or new stock offerings ultimately affecting the S&P 500 Index’s adjusted closing price as well.
23. **Conclusion**

In conclusion, the companies’ and index’s returns distribution are much higher for when there no pandemic, but the adjusted closing prices are higher for when there was pandemic. The illustrated graphs in the R code clearly demonstrated the impact of COVID-19 by showing how the adjusted closing prices rose and how the adjusted closing prices before were not as large as they are today. So, once I understood this, I did further research upon these findings that the code provided, and I found an article about one of the companies that was examined. The article said that Apple’s dividend percentage has indeed been falling recently and that it is better to invest in companies with higher dividend yields. Since the data sets of the companies had high adjusted closing prices recently, simultaneously it makes sense that the S&P 500 has also high adjusted closing prices, especially since these tech companies are a part of the index fund and it is safe to assume that they all have been decreasing their dividend yields, stock splits, and new offerings. As for their roads to recovery, you can look at the data sets using the View() function, take a look at the “Adj.Close.Change” column, and see that this month’s adjusted closing price has actually decreased, so hopefully this is a good signal foreshadowing the companies’ higher share prices increasing to what they were before the pandemic. Although I would still argue that the future remains unclear and companies will keep on working on a more conservative environment until the normality begins to reach us faster or as competitors get more competitive and leave companies no other option but to follow their rhythms.