An Exploration of Drug Development Pipelines and the Financials of the Top 20 Largest Publicly Traded Pharmaceutical Companies

Benjamin T. McCaffrey

Northwest Missouri State University, Maryville MO 64468, USA
s538239@nwmissouri.edu

Abstract. There are many large pharmaceutical companies in the publically traded space. This paper explores the financials and drug development pipelines for each of the top 20 largest companies.

Keywords: Pharmaceutical, Drug, Treatment, Market Cap., RD

1 Introduction and Goals

The top twenty largest pharmaceutical companies by market cap (as of August 1, 2022) are as follows: Johnson & Johnson, Eli Lilly, Pfizer, Novo Nordisk, Roche, AbbVie, Merck, Astrazeneca, Novartis, Bristol-Myers Squibb, Amgen, Sanofi, CSL, GlaxoSmithKline, Gilead Sciences, Vertex Pharmaceuticals, Moderna, Regeneron Pharmaceuticals, Bayer, and Daiichi Sankyo (Figure 1)[1]. These 20 pharmaceutical companies collectively have over a 1000 drugs/treatments for sale on the U.S. market (Figure 2) and have a combined market capitalization of over $3.3 trillion.

Some of these companies are familiar household names, while others are, presumably, more obscure. The goal of this research is to dive deeper into the publicly available information for each of these mammoth pharmaceutical companies and highlight their financials, drug/treatment developments, and forecast future R&D expenses. Most of us are on a medication, or might know someone one medication, that is researched, manufactured, and sold by one of these publicly traded companies. Therefore, this information could be of short-term use for the average retail investor that is interesting in knowing more about pharmaceutical companies.

The way this paper is structured is relatively straight forward. First, the data was collected/extracted from the online sources, then parsed through and cleaned. Next, the data was examined in a process known as exploratory data analysis, where parts of it were examined with visualizations and statistics [2]. And finally, predictive analysis was performed on a select set of the company financials and the overall results were interpreted.

GitHub Repository: https://github.com/BMccaf/Capstone-Project

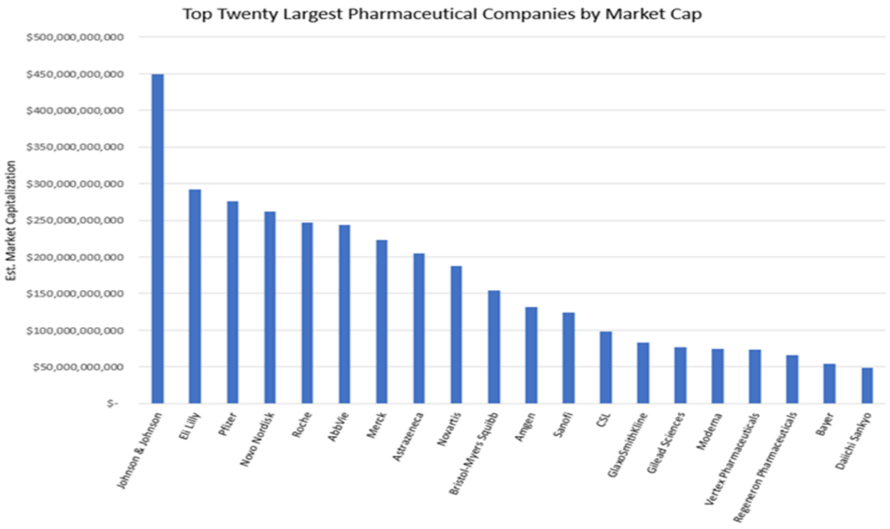


Fig. 1. Current Market Capitalizations

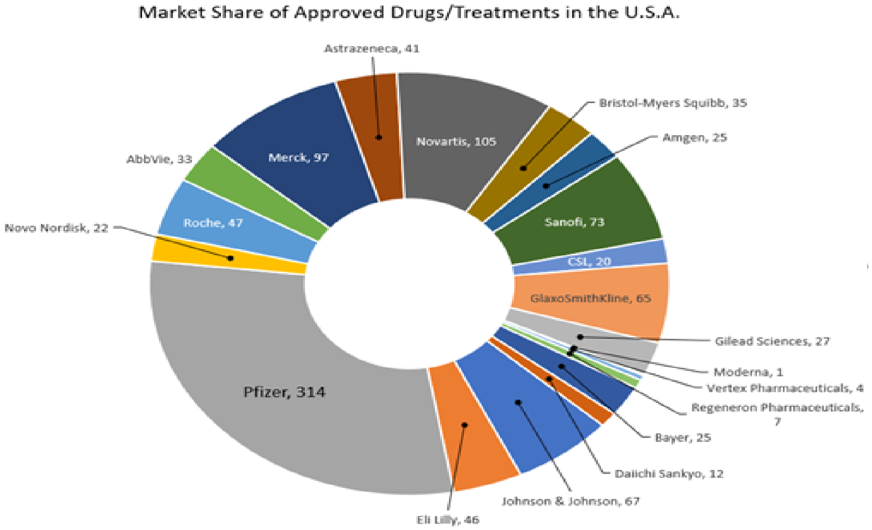


Fig. 2. Total Market Share

2 Data Extraction/Collection and Cleaning

Obviously, the first step of the research process was to identify the top pharmaceutical companies by current market cap. A Python HTML web scraping tool was utilized to extract information from an up-to-date market cap website to create a list of publicly traded pharmaceutical companies [5]. The resulting Excel file had 100 rows and included the following attributes: *Rank, Company Name, Ticker Symbol, Country of Origin, and Est. Market Capitalization.* The list was then manually narrowed down to a top 20 because these companies were slightly more recognizable and had more of an effect on the overall drug market in the U.S.A.

The following three companies were excluded from the list of the top 20: CVS Health (CVS), Merck KGaA (MRK.DE), and Zoetis (ZTS). CVS Health, which should have been in the 11th spot, was excluded because they deal most often with healthcare management, and information on their development pipeline was confusing and incomplete in spots. Merck KGaA, which should have been in the 15th spot, was excluded because they are the parent holding company of Merck and have mostly the same drugs in development. And finally, Zoetis which should have been in the 16th spot, was excluded because they only make treatments for animals, and information on their development pipeline was lacking. Where these three were excluded, the company with the next highest estimated market cap moved up the list.

Once the top twenty list was established, the next step in the process was to find information regarding each company's development pipeline and financials. As a publicly traded pharmaceutical company, the U.S. Securities and Exchange Commission (SEC) requires each to make their quarterly and annual financial statements available to the public, as well as disclose information about their research and development pipelines [3]. Therefore, the public facing website for each company was used to gather the required information (Example: [4]).

The values for the following attributes were gathered from the SEC quarterly reports filed by each company: *Total Liabilities, R&D Expense 3rd Quarter 2021, R&D Expense 3rd Quarter 2020, R&D Expense 3rd Quarter 2019, and R&D Expense 3rd Quarter 2018.*

The values for the following attributes were gathered from the clinical trial and or drug development pipeline information available on the website for each company: *Number of Research Areas Covered by R&D Pipeline, No. of Ongoing Phase 1 Clinical Trails, No. of Ongoing Phase 2 Clinical Trails, No. of Ongoing Phase 3 Clinical Trails, and Filed.*

The values for the attribute *D/E Ratio* were calculated with the following formula:

D/E Ratio = Total Liabilities / Est. Market Capitalization

Example - Johnson and Johnson (JNJ) D/E Ratio:

0.23 = $101,367,000,000 / $449,870,000000

The values for the attribute *Total in Pipeline* were calculated with the following formula:

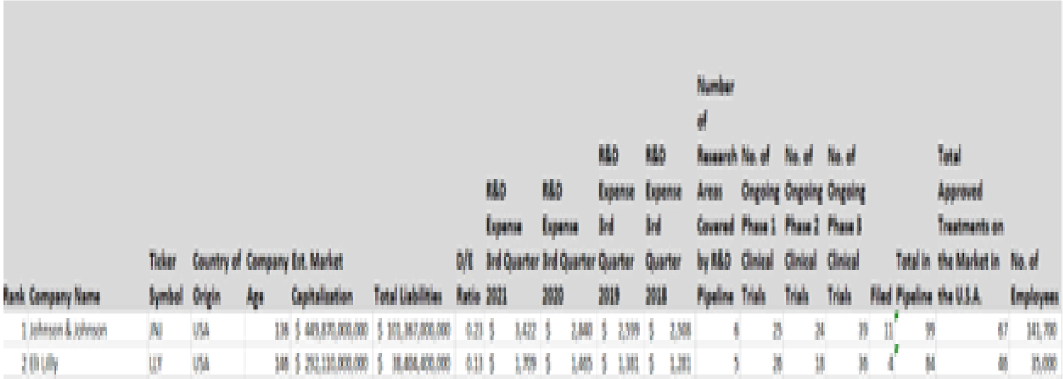
Total in Pipeline = *No. of Ongoing Phase 1 Clinical Trails + No. of Ongoing Phase 2 Clinical Trails + No. of Ongoing Phase 3 Clinical Trails + Filed.*

Example - Eli Lilly (LLY) Total in Pipeline:

84 = 26 + 18 + 36 + 4

The values for the attribute *Total Approved Treatments on the Market in the U.S.A.* were obtained by manually counting the drugs listed under each company's profile on the website www.drugs.com, which is a reputable source for current and discontinued drug information.

And finally, the values for the attributes *Age of Company* and *No. of Employees* were obtained from the profile information listed for each company on the website of the stock trading information company Yahoo Finance.



Rank	Company Name	Ticker Symbol	Country of Origin	Company Age	Est. Market Capitalization	Total Liabilities	D/E Ratio	R&D Expense 3rd Quarter 2021	R&D Expense 3rd Quarter 2020	R&D Expense 3rd Quarter 2019	R&D Expense 3rd Quarter 2018	Number of Research Areas Covered by R&D Pipeline	No. of Ongoing Phase 1 Clinical Trials	No. of Ongoing Phase 2 Clinical Trials	No. of Ongoing Phase 3 Clinical Trials	Filed	Total in Pipeline	Total Approved Treatments on the Market in the U.S.A.	No. of Employees
1	Johnson & Johnson	JNJ	USA	136	$ 445,875,000,000	$ 101,367,000,000	0.23	$ 3,422	$ 2,840	$ 2,599	$ 2,508	6	25	24	39	11	99	67	141,700
2	Eli Lilly	LLY	USA	146	$ 252,110,000,000	$ 38,606,400,000	0.13	$ 1,799	$ 1,465	$ 1,381	$ 1,381	5	26	18	36	4	84	46	35,000

Fig. 3. Sample of Final Data Set (20 Attributes & 20 Records)

3 Exploratory Data Analysis

During the exploratory data analysis phase of this project, some peculiar outliers and strong correlations were identified. To help in the analysis of the data, a variety of different tools were used, from Python code using the Pandas and Matplotlib libraries, to Excel Bar charts and Pivot tables. Collectively, these helped led to some intriguing insights about the way these companies are valued.

Below is a pictogram highlighting the correlation between the different qualitative variables from the data set (Figure 4).

Looking at these variables more closely, it can be inferred that the *No. of Ongoing Phase 3 Clinical Trails* and the *No. of Employees* that a company has appear highly correlated (R squared = 0.5216; Figure 5).

Interestingly, it can be inferred that the attributes *Age of Company* and *Total Approved Treatments on the Market in the U.S.A.* are not strongly correlated (R squared = 0.1947; Figure 6). This was surprising, because common sense would have you believe that the longer a company has been around, the more likely

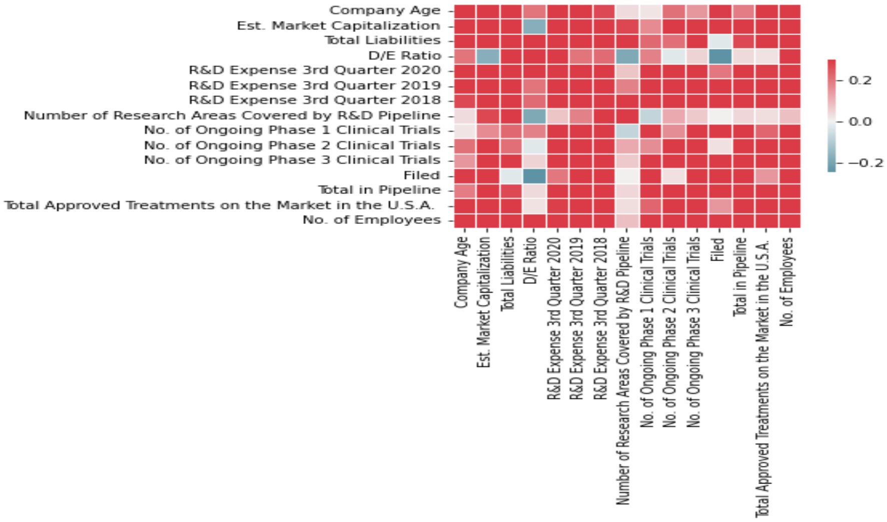


Fig. 4. Correlation Between Quantitative Variables

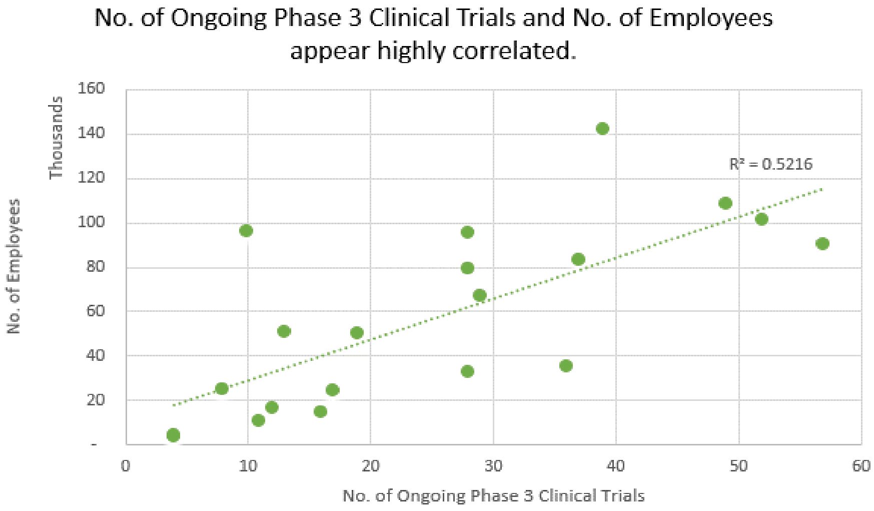


Fig. 5. Correlation Between Phase 3 Trials and No. of Employees

they are to have a lot of treatments approved for sale. If the outlying value for Pfizer was removed, the R squared for the remaining values would be 0.0424.

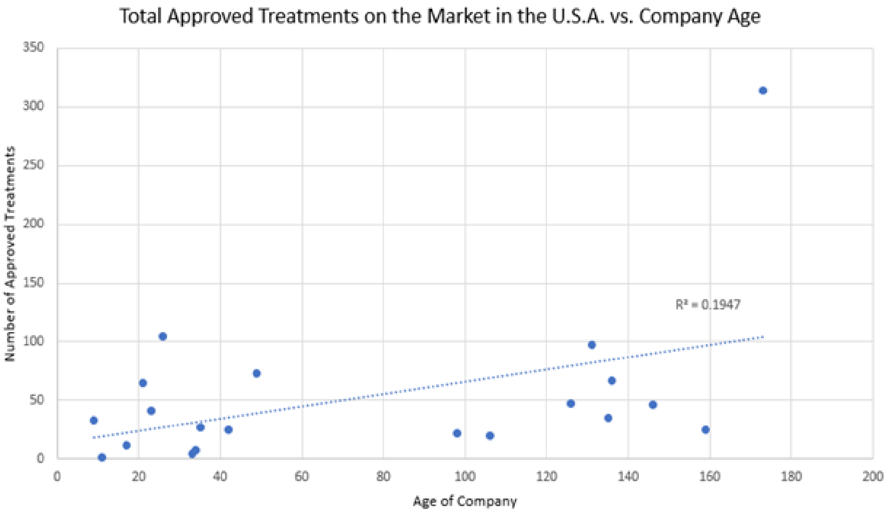


Fig. 6. Correlation Between Approved Treatments and Company Age

As mentioned above, the exploratory analysis yielded some interesting outliers. One outlier is that Pfizer had a noticeably higher number of approved treatments for sale on the U.S. market (314), when it had the third highest market cap (Figure 7). Another interesting outlier is that Moderna only has one approved treatment on the market (SARS-CoV-19 vaccine) yet they have the 16th highest pharmaceutical market cap ($74.3 Billion).

4 Predictive Analysis

After the exploratory data analysis was finished, it was decided that it would be interesting to forecast/predict the Research and Development (R&D) expenses for the end of the third quarter 2022 (September 30th) based on the previous four years of records. Because of the small size of the final data set, splitting it into separate training, validation, and testing data sets was unnecessary.

Figure 8 shows the forecasted third quarter R&D expenses for each of the 20 companies. Surprisingly, only one company is forecast to spend less on R&D in the 3rd quarter 2022 and that is AbbVie ($1.771 Billion to $1.453 Billion). The company that is projected to spend the most on R&D is Bayer ($1.844 Billion to $2.696 Billion).

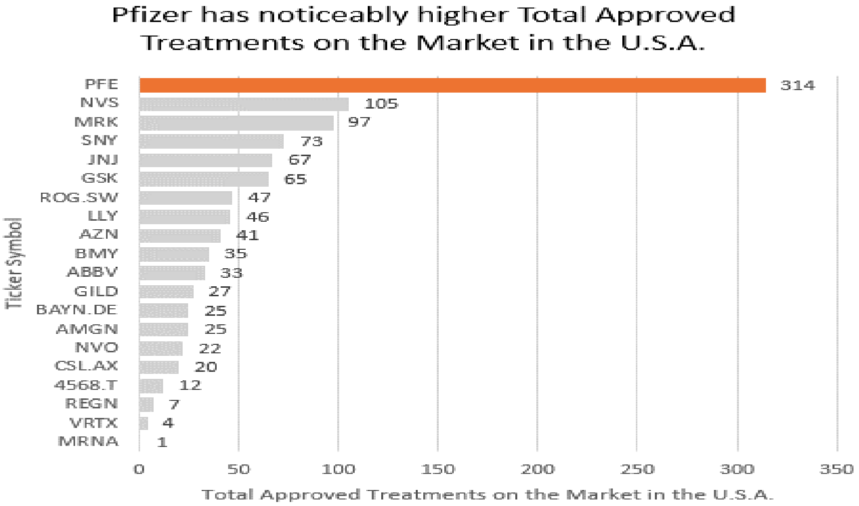


Fig. 7. Total Approved Treatments on the Market

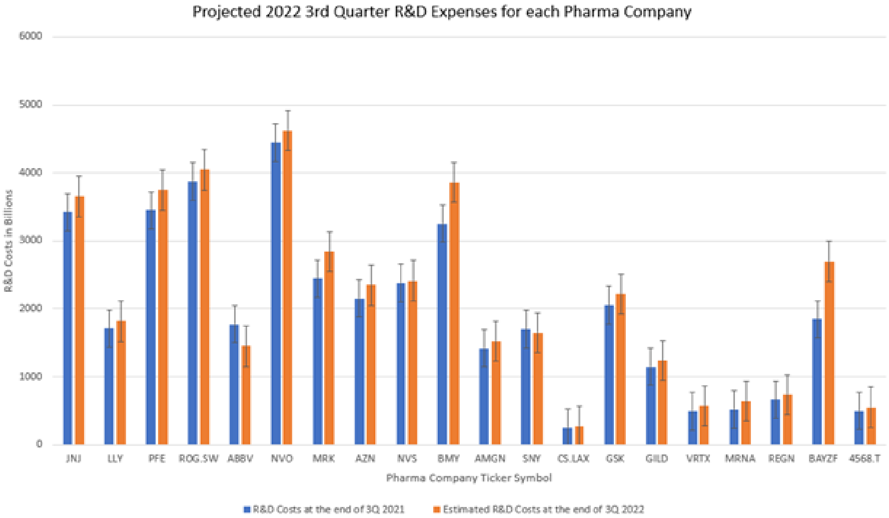


Fig. 8. Projected Expenses

Figure 9 shows that the top 20 largest pharmaceutical companies combined are projected to spend $42.662 Billion on R&D expenses at the end of the third quarter 2022. This would represent an increase of $3.173 Billion year-over-year.

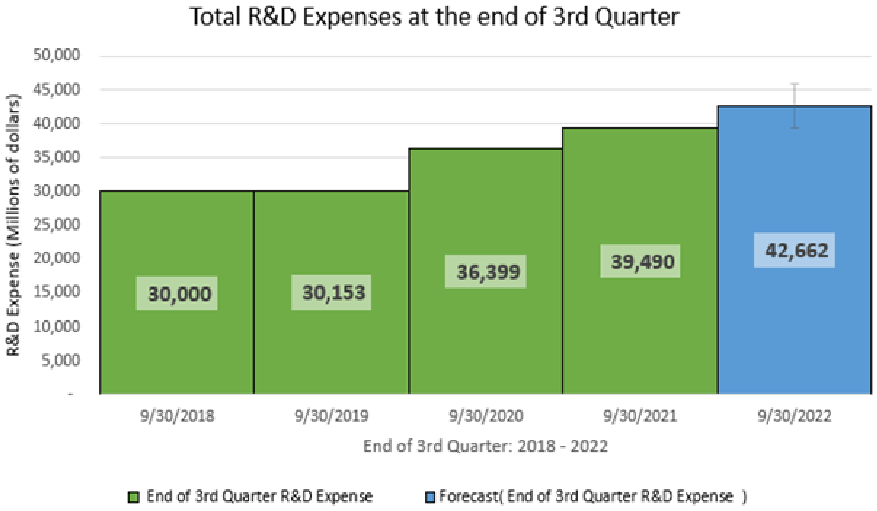


Fig. 9. Total Projected Expenses

5 Interpretation of Results

This research has sorted through drug development pipelines and select financial aspects of the top 20 largest publicly traded pharmaceutical companies. In addition to exploring data for each company, forecasts were made for future R&D expenses both individually and combined. The key takeaways from this research are as follows:

1. Excluding some outlying examples such as Moderna, in general, the longer a pharmaceutical company has existed the more likely they are to have a high number of treatments approved for the U.S. market and, in turn have a high market cap. 2. In general, the more employees that a pharmaceutical company has, the more likely they are to have drugs/treatments farther along in the clinical trial pipeline. 3. Excluding forecasts for AbbVie, all included pharmaceutical companies are projected to spend more money on R&D by the end of the third quarter 2022.

The research and final conclusions for this project were not without their limitations. Overall, the largest limitation for this research is the size of the data set that was used. Typically, a data science project would involve a data set with hundreds to thousands of records. In this case, it was decided to narrow down the data set to just twenty records because a lot of the needed financial and pipeline information turned out to be difficult to hunt down for each company. Even though the companies are required to make this information available to the public, often they burry it under mountains of irrelevant financial information, presentations, and other company publications. This point sort of leads into the next large limitation. That is, the fact that every company uses different medical jargon to describe many of the same things. For example, one company might have their drug development pipeline divided up into very separate and distinct therapeutic areas (i.e. metabolism, cardiovascular, virology, and vaccines, etc.) and another was grouped together in broader terms (i.e. Heart related, pathogenic diseases, etc.) The best effort was made to divide these groupings into collective groups that matched well with each other.

Another limitation associated with this project is the fact that many of these companies are performing clinical trials on drugs that have already been approved for different uses. This can lead to their total number of drugs in development being further divided into "new molecular entities" and "Lifetime Management Projects (LMP)" or in other words, new uses for previously made drugs.

And finally, another known limitation associated with this research is in relation to the forecasted R&D expenses for the end of the third quarter. These forecasts were made with only four years' worth of data. To get a more precise measurement for projected expenses, it would have been good to extend the number of years included, such as the previous ten years' worth of data. As alluded to earlier, a reason for only including four years was the difficultly and time commitment it took to hunt down the financial documents of each company.

If given more time, it would be interesting to dive deeper into the data and possibly increase the number of records that could be included. It is a hope that this work can serve as a reference for future research into the fascinating world of large pharmaceutical entities and their developments.

6 References

So as not to create an overly complicated reference list, the individual websites for each company values were excluded.

[1] "Largest pharma companies by market cap," .com. https://companiesmarketcap.com/pharmaceuticals/largest-pharmaceutical-companies-by-market-cap/ .

[2] K. Lee, "Data Science Project from Scratch - Part 4 (Exploratory Data Analysis)," *Youtube*. Apr. 10, 2020. Accessed: Aug. 03, 2022. [Lecture]. Available: https://www.youtube.com/watch?v=QWgg4w1SpJ8.

[3] "SEC Requirements for Public Companies, SEC Requirements to Go Public, Going Public Attorneys, OTC Attorneys, — Anthony L.G., PLLC," *Legalandcompliance.com*, 2019. http:

//www.legalandcompliance.com/securities-resources/sec-requirements-for-public-companies/ .

[4]“Investor Information,” Investors. https://www.investor.jnj.com/

[5]A. The Analyst, “Amazon Web Scraping Using Python — Data Analyst Portfolio Project,” Youtube. Aug. 24, 2021. Accessed: Aug. 03, 2022. [Lecture]. Available: https://www.youtube.com/watch?v=HiOtQMcI5wg.