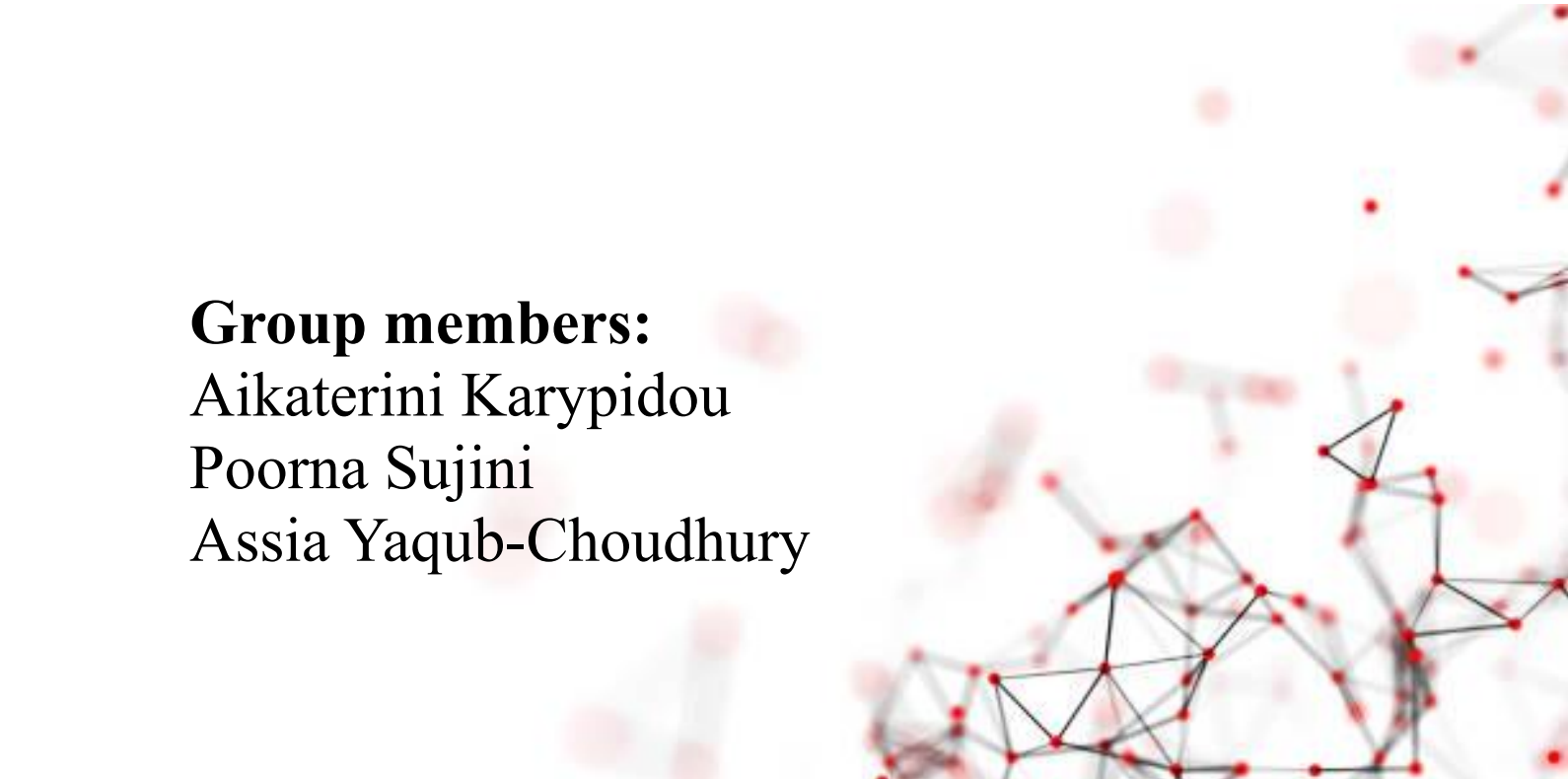
CFG Final Data Science Project

Healthcare and the Pharmaceutical Industry

Group members:
Aikaterini Karypidou
Poorna Sujini
Assia Yaqub-Choudhury



Introduction

Our project will cover the areas of global health data and the pharmaceutical industry. The project will cover 3 main questions detailed below where we will discuss the data sources, analysis of the data and a summary of each area before making the final observations.

Ultimately, we are aiming to draw some conclusions about whether or not the pharmaceutical industry is actually meeting global healthcare needs, or only the healthcare needs of the wealthiest nations.

Background

We will be asking three main questions in our project, with the aim of trying to tell a story about the relationship between global health and the pharmaceutical industry. Our questions will be:

- How does the wealth of a country impact the leading causes of death?
- How were the profits of the pharmaceutical companies affected since 2019 and up to 2021 by the Covid Vaccination and the lockdown.
- Best-selling pharmaceutical drug by each top 10 pharma company and what it treats, compared to top causes of death for 2021.

We believe that this information will be useful for public knowledge as well as government decision makers to ensure Pharmaceutical companies are held accountable for the drugs they are producing and also, that governments can allocate research funding based on the needs of the population.

Steps Specifications

In our first discussion our group highlighted a few topics that we were interested in and it transpired that all of us had an interest in healthcare. We then brainstormed areas within healthcare that we could base our project on. We narrowed these ideas to the top 3 and then did some research to find out which area would have the most reliable data. We finally decided we wanted to find out if Pharmaceutical companies were meeting the global need for medication or just the needs of the most wealthy countries and so we framed this into 3 key steps which became our three project questions above. We broke down our data needs to the below:

- Data to find the top causes of death for as many countries in the world
- Research country wealth indicators and get this data for the same countries as above
- The top 10 Pharma companies in the world and their stock prices for the last 5 years
- Best selling drugs from top 10 Pharma companies and the condition it treats.

Each of us spent a couple of days trying to gather as many data sets as we could, above using online resources and then we reviewed the quality of the data and how reliable the sources were. Each of us then took one of our key project questions and cleaned and analysed the data to best answer each of the project questions. The final question relied on data that was also gathered for the first and second questions so where this was the case, we worked together on these data sets. Throughout the process we discussed all aspects of the project

as a team to ensure we were all on the same page and reviewed each other's code to make sure the data was being handled and analysed correctly and efficiently.

Below is a summary of our data collection and types of analysis for the respective questions.

Question 1) We found the top causes of death data set in the World Health Organisation (WHO) through Kaggle and country wealth indicators in the World bank data library, we decided to go with GDP per capita for our wealth indicator as it was the best indicator for standard of living for the majority of the population. We had great difficulty with formatting this data as we wanted to filter the top 5 causes of death by country. The way the raw table was formed made it very difficult to write queries for fetching the data so after many attempts we decided the best way to deal with it was to transpose the table according to our needs. After reviewing each country's data, there were some null values for covid deaths i.e Greenland, North Korea and Turkmenistan. We decided to drop them as Covid was a cause with special importance. With the GDP data from the world bank we decided to drop the countries with null values for 2020 as the data started from 1960 and most of the null values came from countries that no longer exist and the 2021 data was very incomplete. As the data was so similar for both years we decided to stick with 2020 for this. We used the wealthiest country data and pulled it together with the top causes of death. We counted the number of times each disease appears and we managed to get our first statistics. We then found the count for the occurrence of diseases within the poor countries and used these data frames to produce 2 bar charts one for top causes of death for the richest countries and one for the top causes of death for the poorest countries which are described below.

Question 2) We retrieved stock market data for the last 3 years using the Yahoo data API of the top 10 global pharmaceutical companies. The choice of these dates was based on the idea that we want to see how the pandemic has affected the stock market. We chose to retrieve data from 2019 that was the year before the pandemic to have a view of the "normal" market, 2020 that was the year the lockdown began and 2021 was when the first vaccines went on the market. For this we made 10 different calls to this API to retrieve the data for all 10 companies, and then save this data to 10 separate CSV files. We created a list of dictionaries (within a dictionary), with the company names and tickers and a for loop that would feed this information to a function that calls the API for the dates we need and saves it into a CSV. We also added a print function to the loop to detect any issue with the API call. We could see which of the calls is failing so we can target the problem more effectively. We then displayed this data in a line plot to effectively visualise the data. To get accurate data for company revenue and EPS growth (which is the earnings per share) we had to manually sift through the annual reports of each company for 2021 and save this data into a CSV file which we could then use for analysis. With this new data file we were able to plot a pie chart for company revenues and EPS growth for closer analysis.

Question 3) The only place we were able to reliably find the best-selling drugs was also in the annual reports of each pharma company, as explained above, this information had to be manually sifted and added to the table of key annual report data we created to help us to answer questions 2 and 3. We also used the new table created from question 1 with top causes of death according to GDP per capita to filter the top 10 wealthiest countries and the bottom 10 on this list. We then found the top selling drugs for each company which was included in the annual report table we created for question 2 and tabulated it with the

illnesses that they treat and then compared this information with the top causes of death for the wealthiest and least wealthy nations.

Implementation and Execution

As every member of the group wanted the opportunity to be able to analyse data for the project we each took one question (Sujini-Q1, Katerina-Q2, Assia-Q3) to lead. Our allocated question was our main focus but we assisted each other at each stage where needed and reviewed each other's code and analysis along the way. As there was very little coding required for Question 3 and a great deal of coding work needed for question 1 and how interlinked these questions were, Assia and Sujini worked together on both of these questions.

We used Python Libraries NumPy, Matplotlib and Pandas to clean and analyse our data. The biggest admin challenge was to sift through the annual reports for data to help with answering questions 2 and 3 and to curate this information into a new dataset which we could then use for both questions. We also had to filter through 30 causes of death to display the top 5 causes for each country. This data then had to be extracted into a new table to make it easier to read and analyse alongside the GDP per capita for each of those countries. We decided to use GDP per capita as the measure for country wealth as this information was readily available from reliable sources and is a good indicator of population living standards.

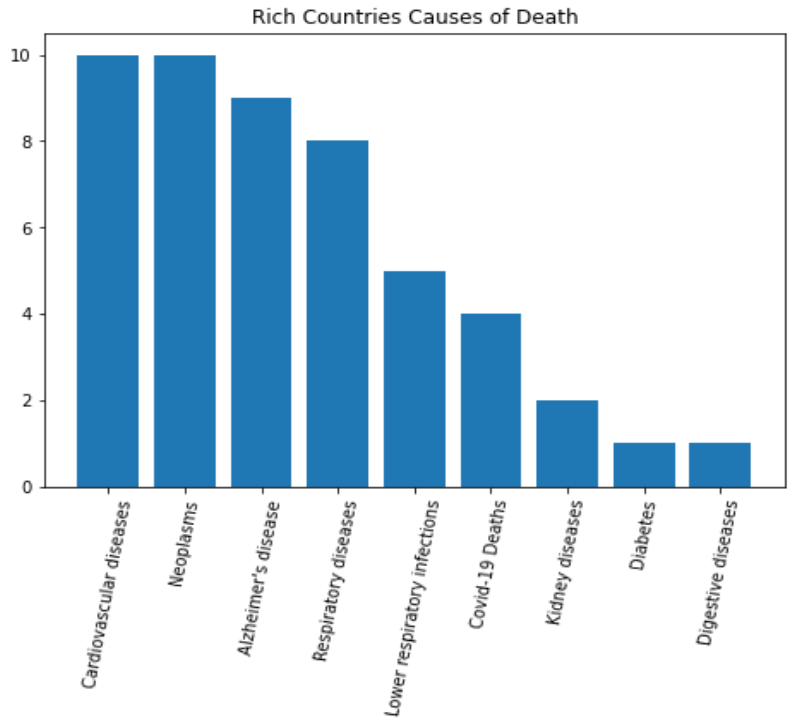
The biggest issue we had was making the data work for question 1 as we could not get the top 5 causes of death for each country because they were rows and not columns. We had to change the data set and transpose the data to get the top 5 for each and this involved a great deal of trial and error and data manipulation so that we could use it. In question 2 we also had to make 10 calls to an API to get the stock market data. For this we had to make a list of dictionaries and create a nested loop to make all 10 calls and save them to CSV files. This ended up working really well for our data.

We used agile development where possible to streamline and improve our process. We broke the project down into weekly sprints to move the project along at a steady speed. The sprints consisted of one 30-minute sprint planning meeting every Tuesday evening via Zoom, where we plan the work for the next sprint and one 30 min backlog refinement meeting on a Friday evening via Zoom to clean up any unfinished work, sort issues ready for the next sprint. Due to time constraints, we did not conduct daily scrum meetings, but we created our own Slack channel for any pertinent ad hoc updates. We used a shared Github repository to create a branch for each question. We then uploaded our Jupyter notebooks with our Python code onto the corresponding branch so that we were able to review each other’s code and suggest improvements etc to inform the next sprint. We used the version control function to track any updates and note all feedback. We also set up a shared google drive to store links and documents for admin purposes, so we were all able to access and contribute to this.

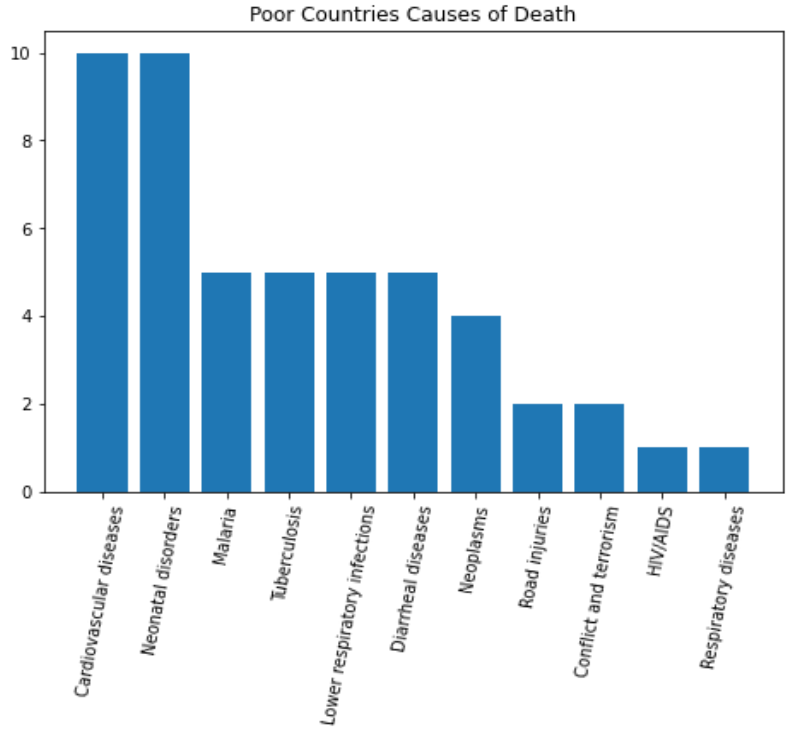
We wrote up the analysis and tests for our code as we went and brought all of this together in the final sprint to finalise the report.

Result Reporting

How does the wealth of a country impact the leading causes of death?



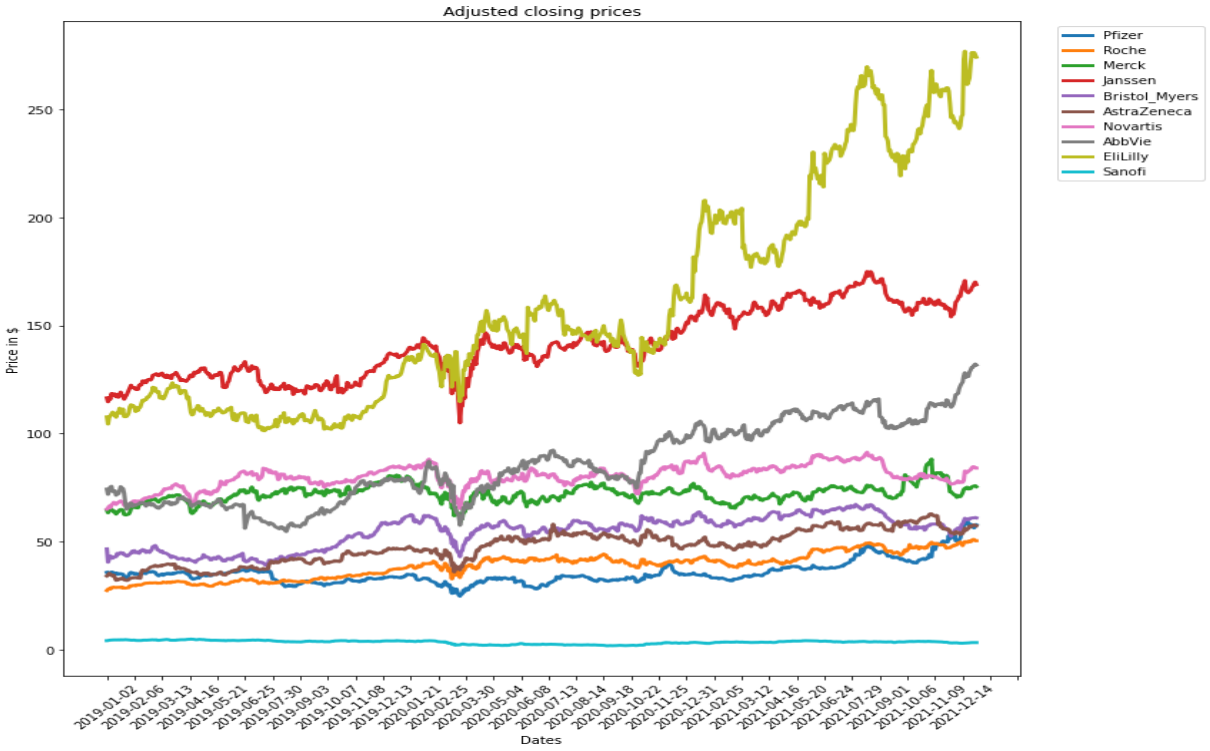
This diagram shows on the X-axis the top causes of death for the top 10 wealthiest nations in the world according to the GDP per capita, the Y-axis counts the amount of instances each disease appears in the list of top 10. As you can see from the data Cardiovascular Diseases (heart), Neoplasms (Cancer), Alzheimer's and Respiratory Diseases (lung) have a high frequency of appearances. These diseases, especially lungs and heart disease are usually associated with poor diet and exercise. Covid-19 makes an appearance here but only makes it to the top 5 in 4 of our richest countries



This diagram shows the top causes of death for the 10 poorest nations in the world according to the GDP per capita, the Y-axis counts the amount of instances of disease that appear in the list of top 10. Again, heart disease appears in the top 5 of every country in this data set, but so does Neonatal disorders. This contains data on infant mortality as well as mothers who die in childbearing/birthing related instances, which doesn't even appear on the list for rich countries. Malaria, Tuberculosis, and Diarrheal Diseases also make a strong appearance here and do not appear in the first diagram. Successful vaccine programs can irradiate Malaria and TB which is why richer nations are not as badly affected. Diarrheal diseases can be a result of unsafe drinking water and unsafe

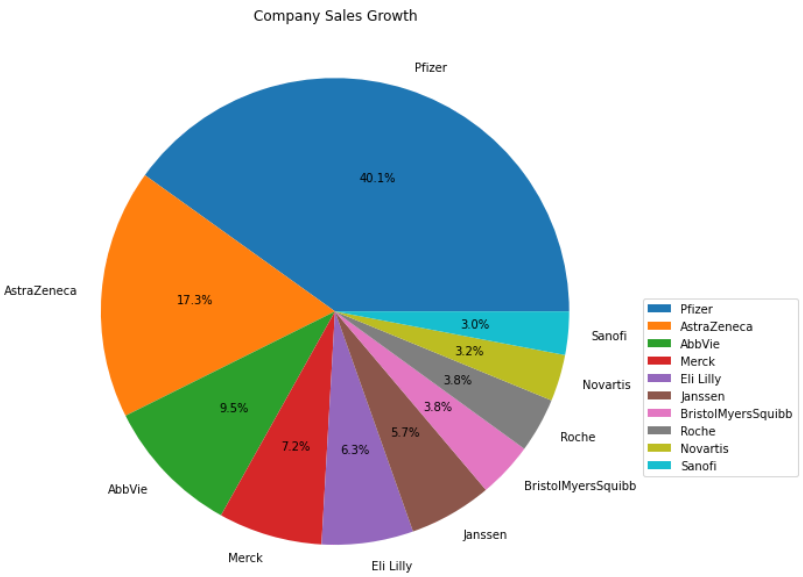
food storage and processing. It is interesting to note that covid-19 does not appear once in this list

How were the profits of the pharmaceutical companies affected since 2019 and up to 2021 by the Covid Vaccination and the lockdown:



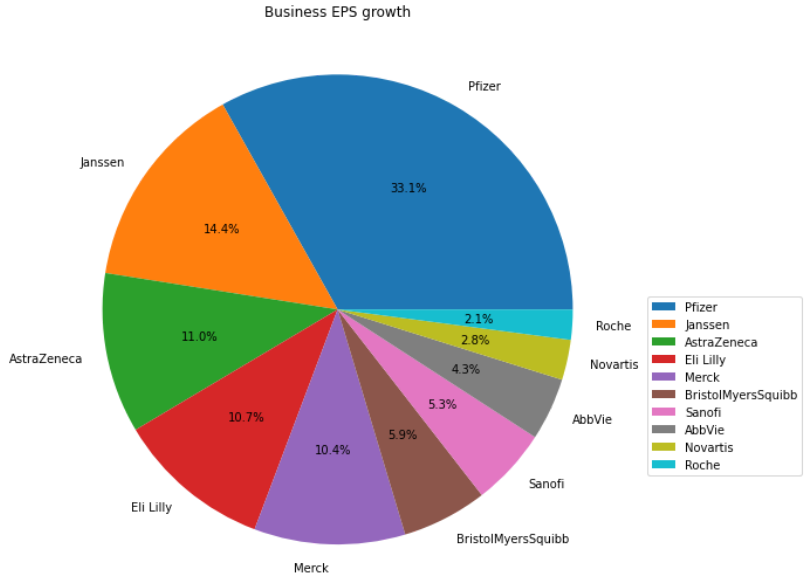
As we can see from the diagram, all the companies apart from Sanofi are gaining value for the whole time period (2019 - 2021). Eli Lilly has skyrocketed between September 2020 and July 2021. As we read on Forbes this happened because Eli Lilly was developing a drug for Alzheimer's that was expected to be approved (3rd most common cause of death in wealthiest countries), In March 2021, Eli Lilly published details of the 2nd trial of the same drug and the findings were mixed. The stock value dropped a lot to recover again by the end of April.

All the companies had a drop in February and March of 2020. This was the beginning of the lockdown in the Western World. The whole market dropped significantly those two months. Merk and AstraZeneca have dropped in value for the last two months of 2021.



Here is an illustration of the companies' sales growth for 2021. Pfizer and Astra Zeneca are the top two companies that raised their sales, Pfizer growing their sales by 40.1% from the previous year and Astra Zeneca by 17.1%. Together they have made massive sales growth in just a short period of time. These two

companies also happen to have released the first 2 Covid-19 vaccines which were being produced and sold en masse during this period. Janssen also released a Covid-19 vaccine but much later than the other two and was not as widely adopted, so they saw growth but not as much. We will look into the top selling drug for each company later in the report.



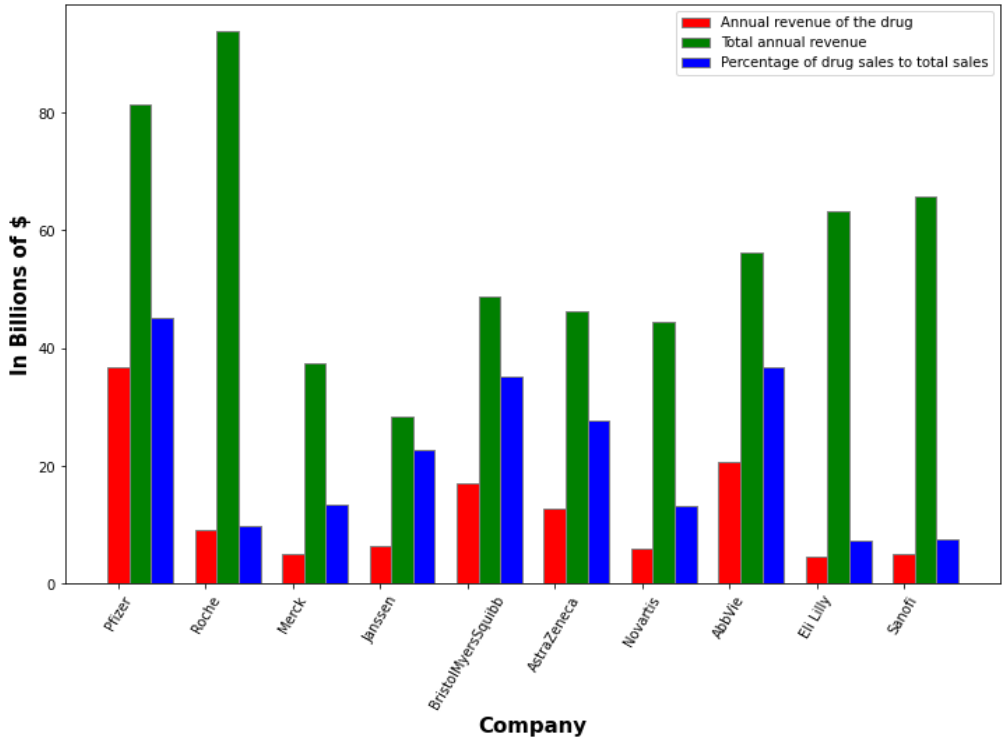
This diagram shows the earnings per share growth for each company. Again, you can see massive growth for Pfizer and significant growth for AstraZeneca and Janssen. We chose EPS as an indicator because it is a value that is independent from the amount of shares a company might have chosen to sell.

Best-selling pharmaceutical drug by each top 10 pharma company and what it treats, compared to top causes of death for 2021.

Companies_2021	Top Selling Drug	What it treats
Pfizer	Comirnaty	Covid 19
Roche	Ocrevus	Multiple Sclerosis
Merck	Keytruda	Cancer
Janssen	Stelara	Crohn's disease
BristolMyersSquibb	Revlimid	Cancer
AstraZeneca	Tagrisso	Cancer
Novartis	Cosentyx	Psoriasis/arthritis
AbbVie	Humira	Arthritis
Eli Lilly	Trulicity	Diabetes
Sanofi	Dupixent	Eczema

The table above shows the top 10 Pharmaceutical companies in the world along with their best selling drug for 2021. The third column shows what disease the top selling medication treats. From the above you can see that Cancer, Covid 19, Crohn's(digestive disease), and diabetes cover the majority of the top selling drugs, and these diseases all feature highly in the list of diseases for the top 10 rich countries we looked at in question 1. The only disease mentioned in the list from the top 10 poor countries is Cancer (Neoplasm) and even then, it doesn't feature highly in that top 10. Conditions such as TB, Malaria, HIV/AIDS and neonatal

disorders can all be addressed with suitable vaccine programmes and medications, but this requires the money to buy it.



Here, we can see that Pfizer, Bristol Myers Squibb and AbbVie made about 40% of their revenue from a single product.

Conclusion

Our aim was to look into whether or not the global Pharma industry is meeting the medicinal needs of the poorest nations as well as the richest nations in the world. We believe we made a first step with this study but a lot more in depth analysis would need to be done to find out where it is that the industry can improve the outcomes of populations in poorer nations. We can see from the data that where there is money the industry will follow. For example, with Pfizer and the covid-19 vaccine mentioned above. Although it was not a huge killer in relation to other diseases, the amount of money thrown to the Pharma industry to tackle this problem made them react quickly and produce an effective vaccine within months. Whereas very treatable illnesses like Malaria and TB are still prevalent in poorer nations when medical treatment exists. This is a part of a bigger problem however, many of the residents of the poorest countries do not have universally available appropriate medical infrastructure which means they suffer a great deal from very treatable illnesses.

References:

Yahoo stock API
Sanofi
AbbVie
AstraZeneca
Bristol Myers Squibb
Janssen
Elli Lilly
Merck & co
Novartis
Pfizer
Roche Pharmaceuticals
Kaggle db
World Bank

Financial press articles:
https://www.forbes.com/sites/greatspeculations/2021/11/05/whats-happening-with-eli-lilly-stock/?sh=6575edd921e1
https://www.marketwatch.com/story/lillys-stock-is-down-after-sharing-additional-clinical-data-about-its-alzheimers-disease-drug-candidate-2021-03-15
https://www.forbes.com/sites/greatspeculations/2021/10/04/will-eli-lilly-stock-rebound-after-an-11-fall-last-month/?sh=35eb8d2f4462
cover image