Assignment 2: Design and Build an Interactive Data Science Application: Writeup

Team:

1) Sundar Anand (Andrew id: sundara)
2) Annie Johnson (Andrew id: anniej)

Your repository should also include a write-up with the following components:

- **A clear description of the goals of your project.** Describe the question that you are enabling a user to answer. The question should be compelling and the solution should be focused on helping users achieve their goals.

In this project we took up 3 datasets related to the COVID-19 pandemic. One dataset has time-series data regarding COVID-19 active cases and the vaccination count for 179 countries. The second dataset has population by country data. The third dataset has global stock data (in United States Dollar) of 4 companies (Facebook, Zoom, Uber Eats and Moderna) from 2012 for Facebook all the way to 2021. Through this analysis we plan to address multiple questions related to the pandemic and the economic conditions during that time.

The questions that we address include:

Question 1)

a) Which countries were most impacted by the COVID-19 pandemic?

b) Which countries faced the least impact from the COVID-19 pandemic?

Question 2)

a) Which countries had the best vaccination rate for the pandemic period?

b) Which countries had the worst vaccination rate for the pandemic period?

Question 3)

a) Which countries had the highest mortality rate during the pandemic period?

b) Which countries had the lowest mortality rate during the pandemic period?

Question 4)

Which companies might be worth investing in during the pandemic period?

Question 5)

What is the maximum profit percentage for each company over the pandemic period?

Addressing these questions would help users understand the result of the COVID-19 pandemic in terms of lives impacted and global economic conditions. One would assume that the pandemic would negatively impact both of these variables but the results of our analysis show otherwise.

- **A rationale for your design decisions.** How did you choose your particular visual encodings and interaction techniques? What alternatives did you consider and how did you arrive at your ultimate choices?

To answer the above questions we used the following metrics

Question 1)

a) Which countries were most impacted by the pandemic?

b) Which countries faced the least impact from the COVID-19 pandemic?

Metric used: A percentage metric based on the number of active cases in a country is used to select the top/bottom 10 countries. This percentage metric is calculated by taking the total number of active cases in a country and dividing it by the country's population and multiplying it by 100.

Visual encoding used: A scatter plot is used here where the size of each bubble is set based on the active cases percentage metric. A bar chart was considered as an alternative but we found that in a scatter plot, we could modify the bubble size based on the percentage metric and this seemed like a better way to visualize the data.

Conclusion: British Virgin Islands has the most active covid percentage of 5.5% and Peru and China have the least active covid percentage of 0% and 0.00013% respectively.

Question 2)

a) Which countries had the best vaccination rate for the pandemic period?

b) Which countries had the least vaccination rate for the pandemic period?

Metric used: A percentage metric based on the number of vaccinated people in a country is used to select the top/bottom 10 countries. This percentage metric is calculated by taking the total number of vaccinated people in a country and dividing it by the country's population and multiplying it by 100.

Visual encoding used: A scatter plot is used here where the size of each bubble is set based on the vaccinated people percentage metric. A bar chart was considered as an alternative but we found that in a scatter plot, we could modify the bubble size based on the percentage metric and this seemed like a better way to visualize the data.

Conclusion: Tanzania and Haiti have the least vaccination covid percentage of 0% and 0.057% respectively and Malta and Cayman Islands have the most vaccination covid percentage of 74.92% and 74.88% respectively.

Question 3)

a) Which countries had the highest mortality rate during the pandemic period?

b) Which countries had the lowest mortality rate during the pandemic period?

Metric used: A percentage metric based on the number lives lost in a country is used to select the top/bottom 10 countries. This percentage metric is calculated by taking the total number of deaths in a country and dividing it by the country's population and multiplying it by 100.

Visual encoding used: A scatter plot is used here where the size of each bubble is set based on the mortality percentage metric. A bar chart was considered as an alternative but we found that in a scatter plot, we could modify the bubble size based on the percentage metric and this seemed like a better way to visualize the data.

Conclusion: United States and Brazil have the most death covid rate of 108.7 and 96.04 respectively per day. There are a few countries with a mortality of 0.

Question 4)

Which companies might be worth investing in during the pandemic period?

Metric used: We observed the difference in stock values for each of the 4 companies between the dates: 10-10-2019 and 28-09-2021. These 2 dates are chosen as we have stock data for all 4 companies in this range and it marked the before-pandemic stock value and the during-pandemic stock value for each company.

Visual encoding used: A line graph with y axis as the stock value and x axis as the company was used to show the difference in stock value before and after the pandemic for each company. A box plot was considered as an alternative but box plots had other details like quartiles that were depicted in the plot. This did not make sense in this setting, therefore we chose to use a line graph itself to visualize the data.

Conclusion: The company that had the largest stock increment was Moderna.

Question 5)

What is the maximum profit percentage for each company over the pandemic period?

Metric used: Between 10-10-2019 and 28-09-2021 the highest and lowest stock value for each company was found and the difference in

Visual encoding used: A line graph with y axis as the stock value and x axis as date was used to show the maximum profit that could be made for each company and the time taken (in days) to achieve that profit. A bar chart consisting of 2 bars for each company, where 1 bar indicated maximum profit percentage and the other indicted time taken to achieve that profit percentage was considered as an alternative. But we felt that this would be too many bars to keep track of and that we had to simplify the visualization. Therefore, we decided to use a line graph itself to visualize both the profit percentage and the time taken to achieve this profit using a single line for each company.

Conclusion:
It was seen that in terms of maximum profit percentage that could be obtained, Moderna was the wisest choice to invest in. Moderna had a profit percentage of 3305% over 22 months during the

pandemic period. But in terms of maximum profit percentage in the least amount of time, Zoom was the best choice having a net profit percentage of 823% in one year.

- **An overview of your development process.** Describe how the work was split among the team members. Include a commentary on the development process, including answers to the following questions: Roughly how much time did you spend developing your application (in people-hours)? What aspects took the most time?

Development process

1) Choosing the domain for collecting the datasets:
   We wanted to truly estimate the devastation caused by the COVID-19 pandemic both in terms of human lives and economic resources. Therefore decided to find datasets related to the COVID cases in each country and the economic growth of companies belonging to various sectors. We chose 1 company each from 4 different sectors- Facebook from the Entertainment/Social media sector, Zoom from the Video communication platform companies, UberEats from the online food ordering services, and Moderna from the Medical sector.

2) Choosing the questions:
   We split the questions into 2 parts, where one set of questions were used to address the impact of the virus in terms of the number of human lives that were impacted and another set of questions to address the economic conditions over the same period of time. Although, the exact questions were framed after some analysis on the dataset was conducted.

3) Data Collection:
   The datasets that were collected from Kaggle include:
   1) COVID-19 World Vaccination Progress (https://www.kaggle.com/gpreda/covid-world-vaccination-progress)
   2) World population data (https://www.kaggle.com/tanuprabhu/population-by-country-2020)
   3) Global Stocks:
   Facebook (https://www.kaggle.com/varpit94/facebook-stock-data)
   Zoom (https://www.kaggle.com/kannan1314/zoom-stock-price-all-time)
   UberEats (https://www.kaggle.com/varpit94/uber-stock-data)
   Moderna (https://www.kaggle.com/akpmpr/covid-vaccine-companies-stock-data-from-2019)

4) Data Cleaning:
   The datasets were cleaned in a jupyter notebook. The data was treated for missing values, availability of data only for intermittent dates and inconsistent naming of same countries. A few unnecessary columns were dropped and new columns of data such as percentage of vaccinated people per country were added. The details of the data cleaning steps that were performed on each of the datasets have been clearly explained in the data_cleaning.ipynb notebook.

5) Data Merging:
   Finally, once all the 3 datasets were cleaned, they were all merged on the date column to get a single dataset having all the information regarding the number of active cases in the world each day, the number of vaccinated people in the world each day, the number of deaths in the world every day and the price of the stock for each company every day.

6) Data Pre-processing:
   The data pre-processing steps to obtain the required metrics to plot each visualization was performed on the merged dataset.

7) Data Analysis and Visualization:
   Based on the first few plots that were made on the data, the metrics and charts to answer our questions were finalized. Then we performed visualization and made inferences based on the results that were observed.

8) Write-up:
   A comprehensive write-up to summarize our project was created.

   Sundar Anand: Worked on developing the right kind of visualizations and analysis to address the questions related to the impact of the pandemic in terms of human lives (Questions: 1, 2, 3). (People hours: 20)
   Annie Johnson: Worked on developing the right kind of visualizations and analysis to address the questions related to the impact of the pandemic in terms of global economic conditions (Questions: 4, 5). (People hours: 20)

   Coming up with a plan to clean the data and then figuring out which visual encoding would best depict the answer to each question were the most time taking aspects of the assignment.

   <u>Conclusion:</u>
   From the analysis it is clear that there isn't much of an overlap between the countries that fall in the category of the highest percentage of active cases and the countries that fall in the category of highest vaccination percentage. This indicates that countries with more number of active cases are probably having such a high number as their vaccination percentage is low. There are a few countries like the UK and the USA which fall in the top 10 countries having the highest active cases and highest mortality. This also makes sense as the pandemic being a deadly one would result in more death in countries having more active cases.

   When we analyse the visualizations on the stock data, we can come to the conclusion that even though the most economically progressive countries like the USA and UK had probably the highest impact in terms of human lives lost, they were still able to generate so much profit in various sectors. The pandemic did not slow down the global economic growth (at least for Facebook, Zoom, UberEats and Moderna). In fact, these companies used the pandemic as an opportunity to flourish despite social distancing and working online rather than in-person.