Kwasi Mensah

IST 652

Final Project Proposal

I am currently working independently under the temporary team name "Team Explore". My focus is on analyzing COVID-19 trends in the US and comparing them to trends in other nations. I plan to explore the vaccination rates, both in the US to look into their impact on the decrease in COVID-19 cases. Additionally. Lastly, I intend to develop a model to see if there are relations to the inflation and unemployment rate during the time and the Covid-19 pandemic.

Below, are the business questions I intend to answer using Python for data analysis, based on the data I have collected:

1. What was the overall trajectory of COVID-19 cases from 2019 to date?
   1. Investigate nationwide trends in the United States and dissect state-level variations.
   2. Compare COVID-19 trends across 14 different countries in comparison to the United States.
2. What percentage of the population received COVID-19 vaccinations by the conclusion of 2022?
3. Did vaccination rates contribute to the decline in observed COVID-19 cases?
4. Compare the types of vaccinations administered and their distribution rates or quantities to the public.
5. Explore inflation and unemployment rates and trace their progression during the COVID-19 pandemic.
   1. Which features most influence inflation and unemployment rates during the COVID-19 pandemic

For the analysis to be conducted the pandas, warnings, scikit-learn, requests, BeautifulSoup, and github libraries were imported. The primary dataset, sourced from the Our World In Data website, contains 332,463 rows and 67 columns in its uncleaned state. I have also collected a dataset providing state-level statistics on the COVID-19 pandemic, focusing on vaccination rates. This dataset consists of 54,628 rows and 16 columns in its uncleaned state. These data sets were collected using the github library. An API key was created to access public GitHub repositories available online. This library was used to look at what files were available in the OWID repository, and from here I used the “get\_raw\_url” function created to retrieve the URL for the file to be downloaded in the python environment.

In addition, the US vaccination rate on a state level was retrieved as well from GitHub. I then attempted to get census data from a government website by web scraping, but received an error so this data as well as the unemployment data from the BLS website were both imported manually. Lastly, I was able to scrape the data on the US inflation rate using the BeautifulSoup and request python packages. I created functions to modularize my script for this Muli-Dimensional COVID-19 analysis. The functions are as followed:

* Data Handling Functions
  + safe\_float\_conversion:
    - Transforms a value into a numerical format, ensuring that it becomes a floating-point number. If any issues arise during the conversion, it will gracefully handle the error and returns “None” if there is an error
  + get\_raw\_url:
    - Gets the direct web link (URL) to a file stored in a GitHub repository. It enables seamless access to the file for further processing.
  + write\_file:
    - Retrieves a CSV file from a provided URL. It then translates the contents into a structured format for data analysis.
  + upload\_file
    - Uploads local files that are either CSV or Excel files into the python environment, and then converts the file into a usable data frame.
* Data Exploration Functions
  + explore\_df:
    - Provides an overview of a data frame, offering key insights such as the initial rows, column count, data types, and summary statistics.
  + subset\_columns:
    - Subsets specific columns.
  + drop\_column:
    - Removes a particular column from a data frame.
  + rename\_columns:
    - Changes the name of a column within a data frame.
  + scrape\_table\_from\_website:
    - Extracts tabular data from an HTML table on a website, enabling integration of external data.
  + map\_month:
    - Changes the numeric representation of a month into its corresponding three-letter abbreviation.
* Data Visualization Functions
  + visualization\_line:
    - Creates a line plot visualization using Plotly Express.
  + visualization\_bar:
    - Creates a bar chart visualization using Plotly Express.
  + visualization\_scatter:
    - Creates a scatter plot using Plotly Express.

Once the data was obtained, I performed data scrubbing and exploration on COVID-19-related datasets. The first step involves calling the function explore\_df on a data frame ‘owid\_covid\_data’, which was retrieved from the OWID GitHub repository. This function provides a detailed overview of the dataset, including the number of columns, the number of rows, and the data types of each column. In addition, it presents a descriptive statistical summary, showcasing key statistics for various columns like total cases, new cases, total deaths, and others. This summary helps in understanding the distribution and characteristics of this and the other datasets that are being analyzed.

Next, a subset of columns is created, containing columns such as continent, location, date, total cases, total deaths, and various demographic and health-related variables. This new data frame was named covid\_data. The countries of interest, including United States, Canada, India, Italy, Greece, China, United Kingdom, Brazil, Russia, Japan, France, Germany, Mexico, Peru, and Iran. This data frame is then filtered to include only the data corresponding to these selected countries. This new filtered dataset is assigned to the variable covid. The code then renames the column 'location' to 'country' in the covid dataset. This step helps to provide a clearer representation of the data. I filled any missing values in the covid dataset with zeros to ensure that the dataset is complete and ready for further analysis.

Next, I retrieved another related to COVID-19 dataset from the OWID repository on GitHub, referring to vaccinations, and named it state\_vacc. A subset of columns containing information about total vaccinations, people vaccinated, people fully vaccinated, and total boosters administered was then created. An exploration for the data frame is performed on this subset using the explore\_df function. The 'date' column in the vaccination dataset is then converted to datetime format, and any rows containing missing values in the vaccination dataset are subsequently removed to ensure data integrity and accuracy. Lastly, a summary statistic is generated for the vaccination dataset, providing key statistics for variables like total vaccinations, people vaccinated, people fully vaccinated, and total boosters administered. This summary serves as a reference for understanding the distribution of vaccination data. This dataset was then aggregated by the maximum total vaccinations, and number of people vaccinated, when grouped by state and year.

The census data was saved to a data frame named ‘census\_data’. I drop specific rows and columns, renamed specific columns for improved readability, and remove unwanted characters from the 'State' column, so that there are no issues when merging the data sets. The data types of the columns are examined, and unique values in the 'State' column are shown to ensure that there are 50 states being analyzed only. The inflation rate undergoes a similar process, as it is cleaned, and a dictionary is used to replace column names with more descriptive ones that indicate the month instead of a numeric value. In addition, the 'Avail.Sept.13' string is removed. Lastly for the unemployment data set uploads, column names are renamed, and any unnecessary rows and columns are dropped, and a new column is added to calculate the average unemployment rate for each year.

For analysis first I wanted to answer, What was the overall trajectory of COVID-19 cases from 2019 to date? Surprisingly we see that the number of cases has increased as the years progress, which is shocking because it is not in the news as often. When I compare COVID-19 trends across 14 different countries in comparison to the United States, I noticed that the US had the highest number of recorded cases. This could be for many of reasons, one may be that the datasets have more consistent data for the US when compared to other nations. Additionally, other nations could have altered their initial numbers during the beginning of the pandemic before quarantine were enforced in respective nations. We see the top 5 nations with the most COVID-19 cases in order was the US with around 103 Million, China with around 99 million, India with around 45 million, France with around 40 million, and Germany with around 38 million cases.

When looking at what percentage of the population received COVID-19 vaccinations by the conclusion of 2022 in the US, we see that about 65% of the population were vaccinated, which could have attributed to the unemployment rate decreasing. When looking at our data we see that the vaccination rates contribute to the decline in observed COVID-19 cases and overall deaths. We looked at the number of vaccinations distributed by manufacturers to compare the types of vaccinations administered and their distribution rates or quantities to the public. We see the top 7 vaccines distributed worldwide were Pfizer/ BioNTech with 1,121,398,127,590 vaccines distributed, Moderna with 306,646,392,939 vaccines distributed, Oxford/AstraZeneca with 93,766,996,577 vaccines distributed, Sinopharm/Beijing with 39,412,498,036 vaccines distributed, Johnson & Johnson with 29,416,335,861vaccines distributed, Sinovac with 21,412,504,419 vaccines distributed, and Sputnik V with 12,879,314,452 vaccines distributed.

When we drilldown on the US vaccine distribution specifically, we see that Pfizer/BioNTech distributed the most vaccines within the states with around 120 billion vaccines distributed, followed by Moderna with 80 billion vaccines distributed, and Johnson & Johnson with around 6.5 billion vaccines distributed. Additionally, we also examined the nationwide trends in the United States and dissect state-level variations in vaccination rates for each State’s residents. We see that the top 10 states with the highest vaccination rates in 2021 were Vermont with around 75% vaccinated, Maine with around 74% vaccinated, Rhode Island with around 73.8% vaccinated, Massachusetts with around 73.5% vaccinated, Connecticut with around 73%. vaccinated, New York with around 70% vaccinated, Maryland with around 69% vaccinated, New Jersey with around 67.5% vaccinated, Virginia with around 67% vaccinated, Washington State with around 66% vaccinated.

When we look at the year 2022 were Rhode Island with around 85% vaccinated, Massachusetts with around 83% vaccinated, Vermont with around 82% vaccinated, Connecticut with around 81% vaccinated, Maine with around 80.5%. vaccinated, Hawaii with around 80% vaccinated, New York with around 79.5% vaccinated, Maryland with around 78% vaccinated, New Jersey with around 76% vaccinated, California with around 75% vaccinated.

Lastly, I decided to Explore inflation and unemployment rates and trace their progression during the COVID-19 pandemic. In doing so I decided to create 2 models to compare to see if features from the COVID-19 data set acquired were influential to the inflation and unemployment rates in the US during that time. First, I combined the average inflation and unemployment rates by year to our subset covid dataset, and I used this data frame to create a linear regression model using the sklearn package in python. I use the unemployment and the inflations rate averages as our dependent variable, and the 'life\_expectancy', 'total\_vaccinations', 'total\_deaths', 'people\_vaccinated', 'people\_fully\_vaccinated', 'total\_boosters', 'month', 'Year' columns as the independent features from the covid subset of data. The data was split into x and y training and testing sets, with 30% of the data being the test size, and the random state ‘11’ specified for reproducibility. The training data is then fitted, and the predicted values of the x testing set are stored in the variable y\_pred.

To evaluate the model, the Mean Square Error is calculated using the function provided to us in the sklearn library in python. We get a MSE of around 4.6311, which is a moderate score telling us the level of accurate prediction this model may make. Additionally, we investigated the r-squared of the linear regression model using another function also provide in the library, and when doing so we received the r-squared of about 47% meaning that 47% of the change in the unemployment rate and inflation rates were contributed by factors during the pandemic. For comparison, I also created a Random Forest Model to compare models, but for this model the MSE received was well below zero indicating that the data may be subjected to overfitting.

In conclusion, the trajectory of COVID-19 cases showed a consistent increase over time, particularly in the US. The highest recorded cases were in the United States, followed by China, India, France, and Germany. By the end of 2022, approximately 65% of the US population received COVID-19 vaccinations, potentially impacting the decline in cases and deaths. Vaccination distribution was led by Pfizer/BioNTech, followed by Moderna, Johnson & Johnson in the United States. Among individual states, Vermont, Maine, and Rhode Island consistently had the highest vaccination rates. When looking into the inflation and unemployment rates during the pandemic our linear regression model using features from the COVID-19 dataset achieved a moderate level of accuracy. The R-squared value of approximately 47% indicated that these factors contributed to nearly half of the changes in unemployment and inflation rates during the pandemic.

**References**

1. Our World in Data. (n.d.). Coronavirus (COVID-19) Deaths. Retrieved from <https://ourworldindata.org/covid-deaths>
2. Worldometer. (n.d.). Countries where Coronavirus has spread. Retrieved from <https://www.worldometers.info/coronavirus/countries-where-coronavirus-has-spread/>
3. U.S. Census Bureau. (n.d.). State Population Totals: 2020-2022. Retrieved from <https://www.census.gov/data/tables/time-series/demo/popest/2020s-state-total.html>
4. US Inflation Calculator. (n.d.). Current US Inflation Rates: 2000-2023. Retrieved from <https://www.usinflationcalculator.com/inflation/current-inflation-rates/>
5. U.S. Bureau of Labor Statistics. (n.d.). Databases, Tables & Calculators by Subject. Retrieved from <https://www.bls.gov/data/>
6. Our World in Data. (n.d.). Coronavirus Pandemic (COVID-19). Retrieved from <https://ourworldindata.org/coronavirus>
7. scikit-learn. (n.d.). **sklearn.metrics.r2\_score**. Retrieved from <https://scikit-learn.org/stable/modules/generated/sklearn.metrics.r2_score.html>
8. scikit-learn. (n.d.). **sklearn.metrics.mean\_squared\_error**. Retrieved from <https://scikit-learn.org/stable/modules/generated/sklearn.metrics.mean_squared_error.html>
9. scikit-learn. (n.d.). **sklearn.ensemble.RandomForestClassifier**. Retrieved from <https://scikit-learn.org/stable/modules/generated/sklearn.ensemble.RandomForestClassifier.html>
10. Stack Overflow. (2020). Is there a way to import CSV file from GitHub automatically to my Jupyter Notebook? Retrieved from <https://stackoverflow.com/questions/60609760/is-there-a-way-to-import-csv-file-from-github-automatically-to-my-jupyter-notebo>
11. Stack Overflow. (2018). How to drop rows from pandas data frame that contains a particular string in a particular column? [duplicate] Retrieved from <https://stackoverflow.com/questions/28679930/how-to-drop-rows-from-pandas-data-frame-that-contains-a-particular-string-in-a-p>