TEXAS TECH UNIVERSITY®

Business Intelligence
ISQS 6339

Final Project

Group members:

Nicholas Small
Gatimbirizo Daniel Mucyo
Olivio Ogbebor

October 9, 2020

Contents

Introduction

The infamous COVID-19 pandemic has been of great disrupt to various economies worldwide, especially the US economy. Essential businesses that helped enrich the US economy were shut down for weeks at a time which played a major role in the degradation of the US economy and other economies worldwide. Because of this predicament, millions of jobs have been lost and most unfortunately millions have died worldwide since the inception of this pandemic.

This project covers the various elements surrounding the COVID-19 pandemic in our world today. The project entails two datasets related to COVID-19 cases worldwide and COVID-19 in the U.S . In this report, effective data analysis will be performed on these datasets in order to generate more insights on the COVID-19 pandemic cases in various countries worldwide as well as a close look to the COVID-19 pandemic cases in the U.S.

Time Series Datasets

Data analysis

Three datasets relating to the COVID-19 pandemic globally were collected. The first datasets(df1) answer the question of how many confirmed COVID-19 cases are there in each country in the world. The second dataset (df2) answers the question of how many lives have been lost as a result of the COVID-19 pandemic and the third dataset (df3) answers the question how many people have recovered from the COVID 19 pandemic in each country in the world.

The valuable data items that are present in these respective datasets are items pertaining to Province/State, Country/Region, date, number of cases confirmed, number of deaths and number of recovered patients depending on what dataset is in question. The Latitude and Longitude of the countries will not be useful in our analysis; therefore, we have dropped those columns.

The remaining data sets can now be effectively merged and manipulated to directly give us more insight on how many cases are currently active, new cases that are arising, new deaths5 and newly recovered patients. We can also then utilize the data to make inference on the direction the COVID 19 pandemic is moving; if things are looking to get better or worse in the near future. The variables; “confirmed”, “recovered” and “deaths” in the merged dataset can be used to determine the number of active cases and can also be useful in determining the number of new cases, new deaths and newly recovered patients.

Data Cleaning

After pulling the data from the web using the io library, the next step was to clean out the data.

Initially, the quality of the data was quite mediocre because there were some missing data and some useless columns.

The first task performed in cleaning the data was to pivot the table by changing the rows into columns and columns into rows using the “melt” function. This allowed passing values for confirmed cases, death and recovered cases as attributes and dates as entities. It also helped in restructuring the data for better analysis and visualization.

In all datasets, Province/State contained missing data probably because only few countries in the world have provinces or states. Province/State variables are data keys that can be used in joining the datasets. For that reason, the missing data will be left as is as the Province/State column was used in merging the data. After merging the datasets, the “recovered” column had 3640 values missing. The missing values were filled with 0 using the “fillna” function because the missing values were from different countries, so using the average would add bias to the data analysis.

The columns that were not useful in the data analysis were also dropped. The “Lat” and ‘Long” columns were dropped using the “drop function” as they had no role in the data analysis

Data Merging

The common elements between the three datasets were Province/State, Country/Region and Date, so all three datasets were merged using a left join based on these attributes. The merged dataset is now comprised of confirmed cases, deaths and recovered cases. Merging the datasets made the data more valuable because new insights about the COVID-19 such as active cases i.e. (confirmed – recovered – deaths) could be generated due to the combined datasets. With further data manipulation, these variables can be utilized in determining the number of new cases, new deaths, and newly recovered patients.

Demographic Datasets

Data analysis

The second three datasets provide a better feel of the location (state and county) in the U.S that are affected by COVID 19. It also gives us a broader perspective of the total deaths in each county, how many of those deaths were as a result of COVID 19, and those that were as a result of other causes. This dataset also provides more insight on how different races have been affected by the COVID 19 pandemic. The potential valuable data in all three datasets are those pertaining to location (FIPS Code, FIPS County Code) which answers the question of where the COVID 19 patients are located. Also, the “Non-Hispanic White”, “Non-Hispanic Black” and “Hispanic” data in the second data frame provides information about the number of affected patients in each race. Then the “total deaths”, “Deaths involving COVID-19” and “Deaths from All Causes” can help determine how many deaths are actually caused by COVID-19 as opposed to other death causes.

Data Cleaning

The overall quality of the second datasets was quite low. Nevertheless, adequate data munging, mining and cleaning were utilized to produce a data frame of very good quality and that became eminent in the data analysis of the project. Again, the necessary columns that would not aid our cause in the data analysis were dropped. Columns such as “incident rate”, “Combined Key”, “Admin2”, “UID”, “People_Tested”, “People_Hospital”, “Date as of”, “start week”, “first week” and more were dropped from all three datasets depending on what dataset they were present in. Missing race values for blacks and Hispanics were also infilled with the mean of the respective columns. Some columns such as ‘Total deaths’ and ‘COVID-19 Deaths’ had to be converted to numeric data types using the “to_numeric” function. To foster data analysis, a column showing the death per race was created by multiplying the number of COVID-19 deaths by each race value. A left join was then utilized to merge the datasets together based on FIPS which was present in all three data sets. Now the combined dataset is more valuable together and can help determine how different races are impacted by COVID-19, where these cases are present and what areas are experiencing more cases. Better insights can now be generated on the active COVID cases and a clearer picture of how many death cases are actually caused by COVID-19 can be derived.

Data Merging

The common element between the three datasets was the FIPS county code which is the Federal Information Processing Standard. All three datasets were merged using a left join based on the FIPS attribute. Merging the datasets made the data more valuable because new insights about COVID-19 in the U.S was generated, such as confirmed cases, deaths and active cases by race in each state/county.

Visualizations analysis

The main goal for the visualization of the datasets was to adhere as much as possible to good visualization characteristics. Visual Perception approach and analogical reasoning are visualization characteristics that dominates the most in our analysis through association, differentiation, comparison, exploration and validation. To achieve our visualization, Tableau was utilized, which enabled the creation of several graphs and an interactive dashboard that followed good visualization principles.

Nevertheless, not all our visualization graphs adhered to the concept of natural processing. Plots in Fig 1 adhere to the natural processing concept because the graph and charts can be understood easily. Fig 2 does not adhere to this concept because the user will spend more time trying to understand that each race has its own row. This blur can be optimized by comparing race side by side. Fig 3 and Table 1 also adhere to the natural processing concept and some knowledge will be required to understand the matrix.

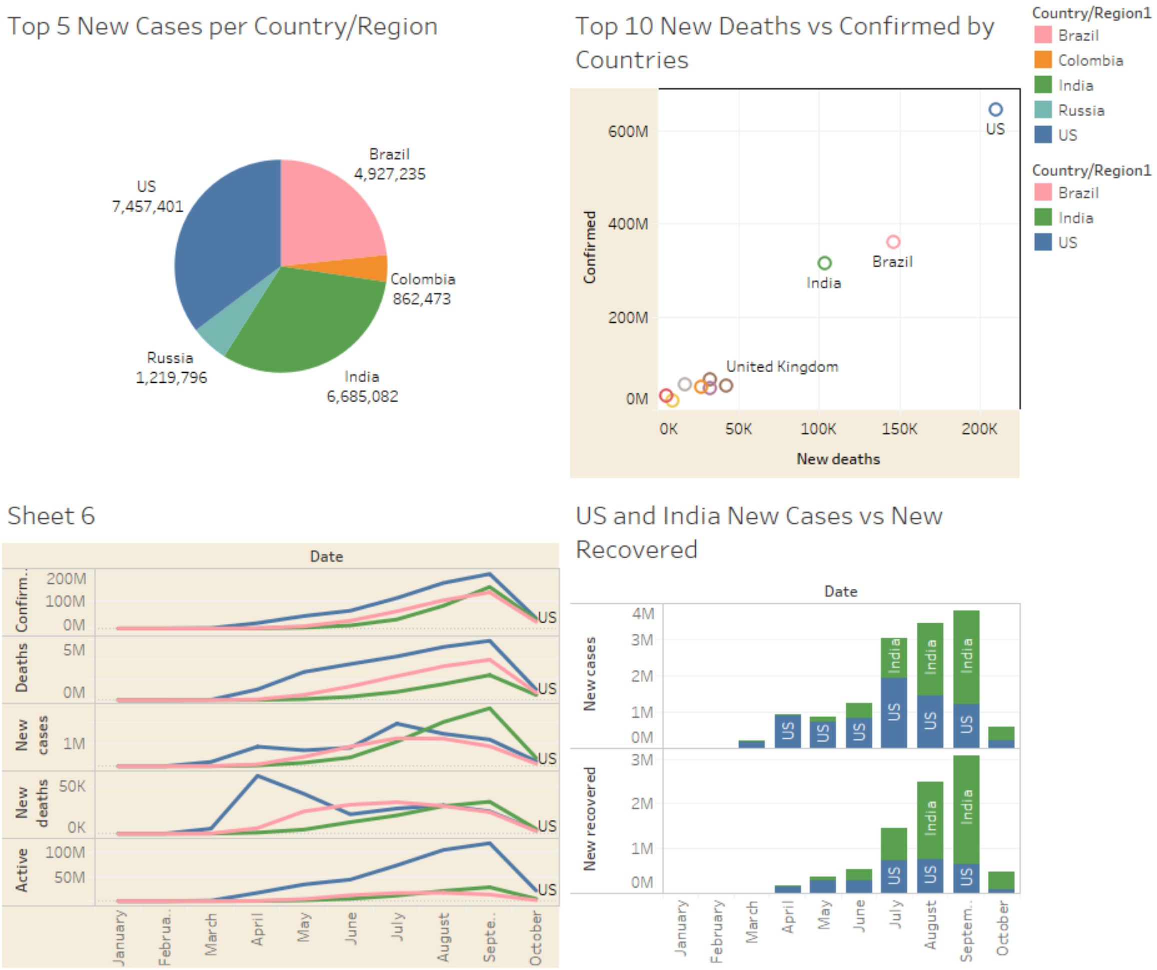


Fig 1: Dashboard of Covid-19 worldwide comparison

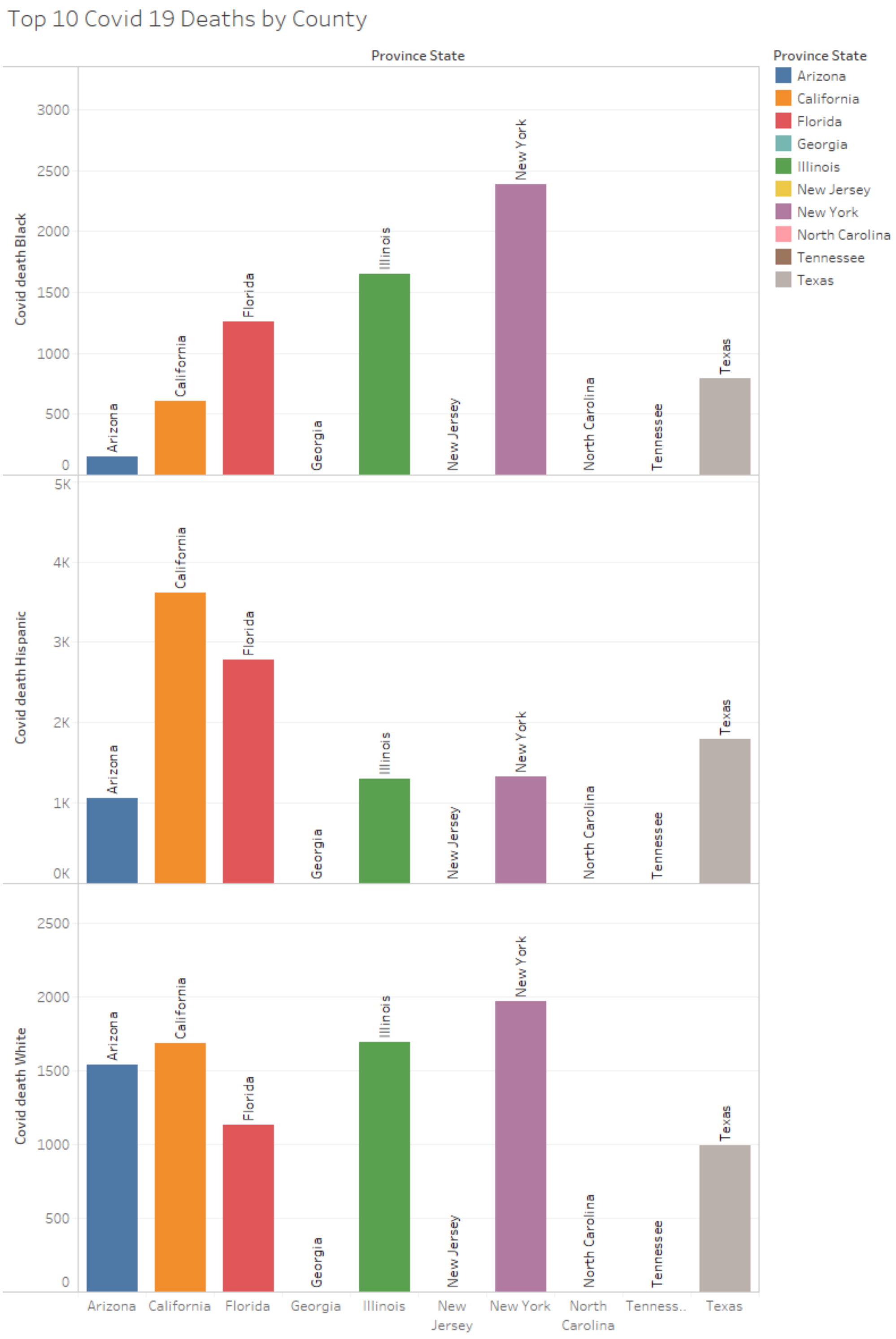
From the first datasets, the above interactive dashboard was created and includes a pie chart which shows the top 5 newest cases per country worldwide.

The scatter chart shows the new deaths vs confirmed cases by the top 10 countries worldwide.

The plot on the left compares active cases, new deaths, new cases, deaths and confirmed cases by the top three countries worldwide over time. It can be observed that in all three visualizations the US has the highest number of cases.

The bar chart on the right compares newly recovered and new cases in US vs India.

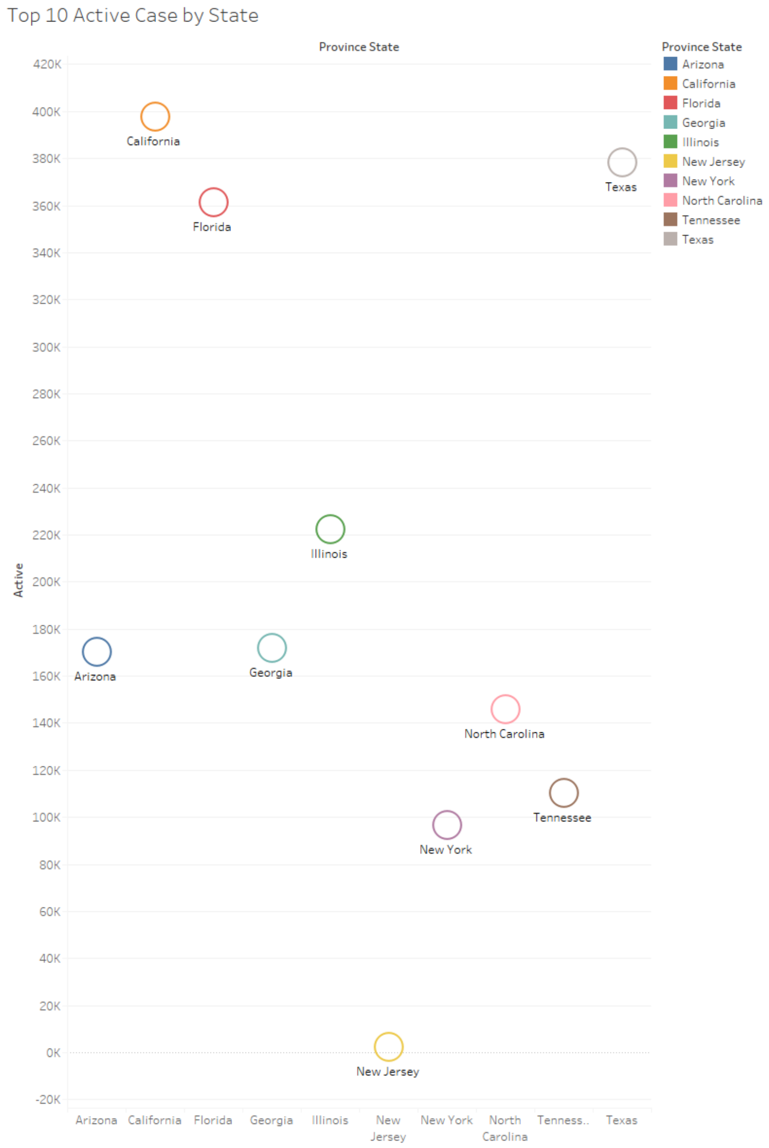
With the second dataset, which comprises of data about COVID-19 in the U.S, visualizations were created in Tableau using bar charts as shown below, comprising of data on COVID-19 deaths in each state by race. It can be easily observed that black and white people died the most in almost all states.



Sum of Covid death Black, sum of Covid death Hispanic and sum of Covid death White for each Province State. Color shows details about Province State. The marks are labeled by Province State. The data is filtered on County Name, which has multiple members selected. The view is filtered on Province State, which keeps 10 of 52 members.

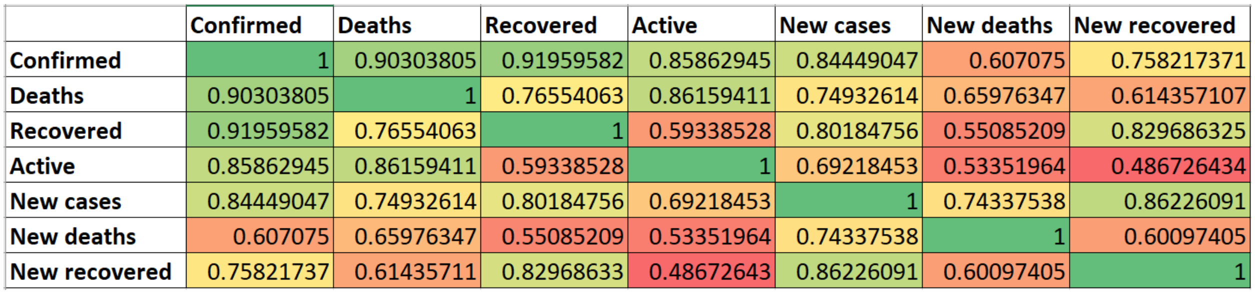
Fig 2: Covid-19 deaths by race in U.S

A bubble plot for the top 10 active cases by state was also created as shown below. From the bubble plot, it can be observed that California and Texas have the most active cases.



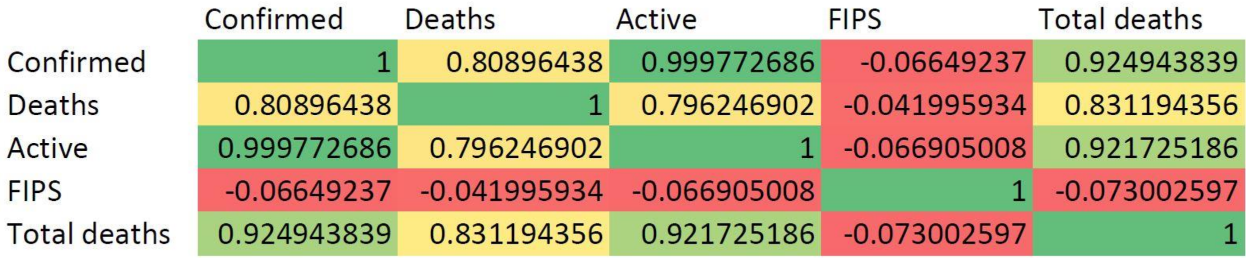
Sum of Active for each Province State. Color shows details about Province State. The marks are labeled by Province State. The data is filtered on County Name, which has multiple members selected. The view is filtered on Province State, which keeps 10 of 52 members.

Fig 3: Active case by states in U.S



	Confirmed	Deaths	Recovered	Active	New cases	New deaths	New recovered
Confirmed	1	0.90303805	0.91959582	0.85862945	0.84449047	0.607075	0.758217371
Deaths	0.90303805	1	0.76554063	0.86159411	0.74932614	0.65976347	0.614357107
Recovered	0.91959582	0.76554063	1	0.59338528	0.80184756	0.55085209	0.829686325
Active	0.85862945	0.86159411	0.59338528	1	0.69218453	0.53351964	0.486726434
New cases	0.84449047	0.74932614	0.80184756	0.69218453	1	0.74337538	0.86226091
New deaths	0.607075	0.65976347	0.55085209	0.53351964	0.74337538	1	0.60097405
New recovered	0.75821737	0.61435711	0.82968633	0.48672643	0.86226091	0.60097405	1

Table 1: Matrix Table for Covid-19



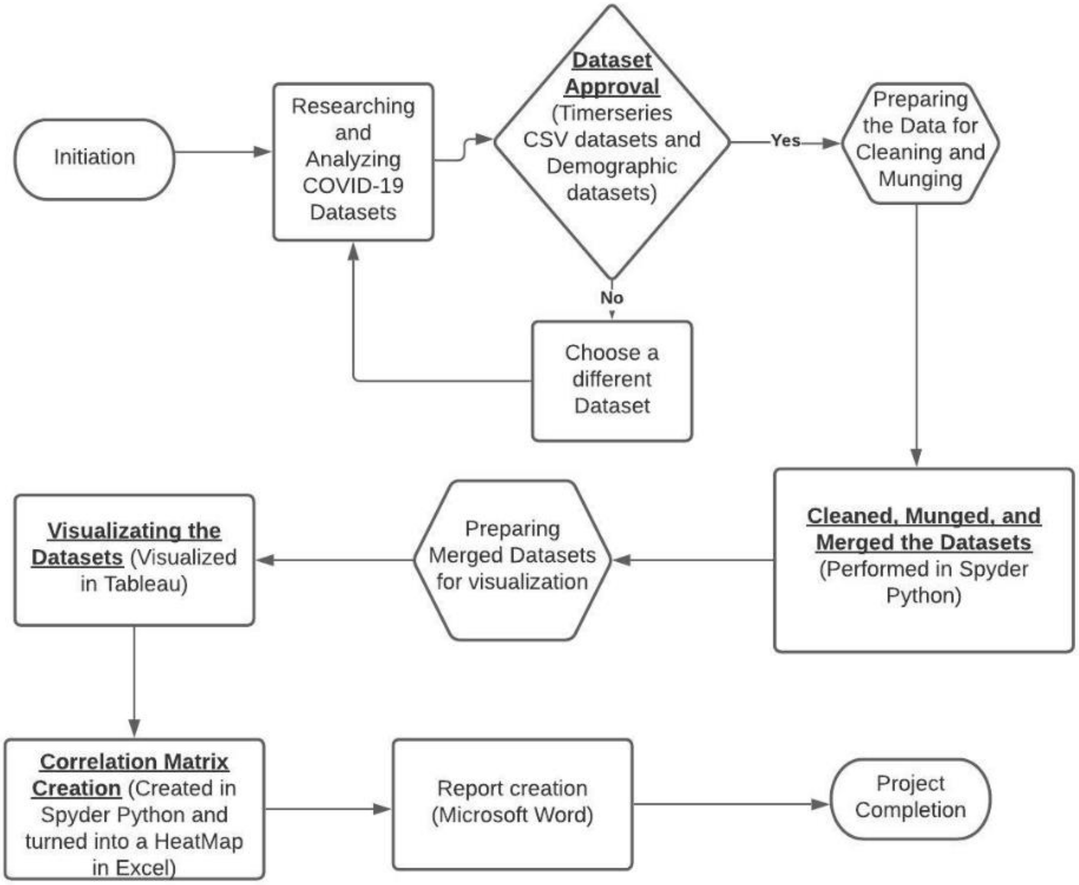
	Confirmed	Deaths	Active	FIPS	Total deaths
Confirmed	1	0.80896438	0.999772686	-0.06649237	0.924943839
Deaths	0.80896438	1	0.796246902	-0.041995934	0.831194356
Active	0.999772686	0.796246902	1	-0.066905008	0.921725186
FIPS	-0.06649237	-0.041995934	-0.066905008	1	-0.073002597
Total deaths	0.924943839	0.831194356	0.921725186	-0.073002597	1

Table 2: Matrix Table for Covid-19

Displayed above are the correlation matrix of the time series and the demographic datasets. The analysis of the correlation matrix for the time series dataset shows all positive correlation. Confirmed cases and Deaths, Recovered and Deaths, and Recovered and Confirmed all have a high positive correlation while newly recovered and active cases are shown to have a low correlation.

In the demographic dataset, there is a negative correlation between FIPS and all other variables, while active and confirmed cases have the highest correlation. Using these correlations, it can be determined that the datasets demonstrate both positive and negative relationships for COVID, that can provide useful information to healthcare professionals and the government of various countries.

Flow diagram


Initiation
Researching and Analyzing COVID-19 Datasets
Dataset Approval (Timerseries CSV datasets and Demographic datasets)
Yes
No
Preparing the Data for Cleaning and Munging
Choose a different Dataset
Cleaned, Munged, and Merged the Datasets (Performed in Spyder Python)
Preparing Merged Datasets for visualization
Visualizating the Datasets (Visualized in Tableau)
Correlation Matrix Creation (Created in Spyder Python and turned into a HeatMap in Excel)
Report creation (Microsoft Word)
Project Completion

Code instruction

A zip file containing two .py files for the codes utilized in data munging, cleaning and analysis has been submitted. Both codes are very well commented step by step and should be easy to understand. All datasets are read directly from the web. The only manipulation needed will be to access the csv files, which can be easily done by renaming the file path to match the operating system used to access the files.