CISC 525 – Big Data Architectures

Final Project Report

**BHARGAV KOTIPALLI**

Contents

[Motivation 3](#_Toc57294894)

[Data Introduction 3](#_Toc57294895)

[Objectives 3](#_Toc57294896)

[System Architecture and Implementation 3](#_Toc57294897)

[Apache Hive 4](#_Toc57294898)

[Amazon QuickSight 4](#_Toc57294899)

[Experiments 5](#_Toc57294900)

[1. Use Hive to get the statistics 5](#_Toc57294901)

[2. QuickSight to analyze and visualize the data 5](#_Toc57294902)

[3. Use Jupyter notebook 5](#_Toc57294903)

[4. QuickSight to visualize Correlation 7](#_Toc57294904)

[Results 7](#_Toc57294905)

[1. From Hive 7](#_Toc57294906)

[2. From QuickSight to analyze and visualize the data 10](#_Toc57294907)

[3. From Jupyter notebook 12](#_Toc57294908)

[4. From QuickSight to visualize correlation 16](#_Toc57294909)

[Conclusion 18](#_Toc57294910)

[Reference 19](#_Toc57294911)

# Motivation

COVID-19 is in charge of the current outbreak of pneumonia worldwide. It originated in 2019 and quickly spread across multiple countries. People’s lives have been significantly impacted. Tens of thousands of people died because of COVID-19. Those who have survived this pandemic could even have trouble walking, climbing stairs or lifting objects. Thus, we would like to get the dataset of COVID-19 from 2019, explore the dataset, visualize the trend of COVID-19 growth, understand the risk factors that could be related to the deaths, and advise people to better protect themselves.

# Data Introduction

We got the dataset from Our World in Data website (<https://ourworldindata.org/coronavirus-testing>). The data is constantly updated and well maintained. Overall, this dataset covers 89% of the world’s population with 110 countries’ data [1]. There are 50 columns in the dataset, ranging from location, date, total cases, new cases, to risk factors such as GDP per capita, extreme poverty, cardiovascular death rate, diabetes prevalence, female smokers, male smokers, handwashing facilities, and etc. This dataset is so granular has all we need, and due to time constraints, we were focusing on getting to know the safest and most dangerous places based on total reported cases and deaths, understand the growth of COVID-19 and find the most relevant risk factors.

# Objectives

The project aims to assist in reducing the impact of the pandemic by identifying primary risk factors that make an individual susceptible to contracting the virus.

Objectives of the project are as follows –

1. The principal objective of the project is to study and identify primary risk factors affecting the spread of the Coronavirus using the COVID-19 testing data and employing a big data solution.
2. An auxiliary objective is to understand the extent to which countries/territories were impacted and use this to derive the correlation to achieve the main objective of identifying the risk factors.

# System Architecture and Implementation

Figure 1 is the Architecture of tools that we have been using. Once we got the original dataset, we would ingest it by Hadoop FS and save it into database Hive. Then we can run Hive queries to get data statistics and aggregated data. Afterwards, we can combine statistics with actual data and ingest into Amazon QuickSight to get good visualization. At the same time, we used Pandas library, trying to find the correlation between risk factors with the spread of COVID-19.

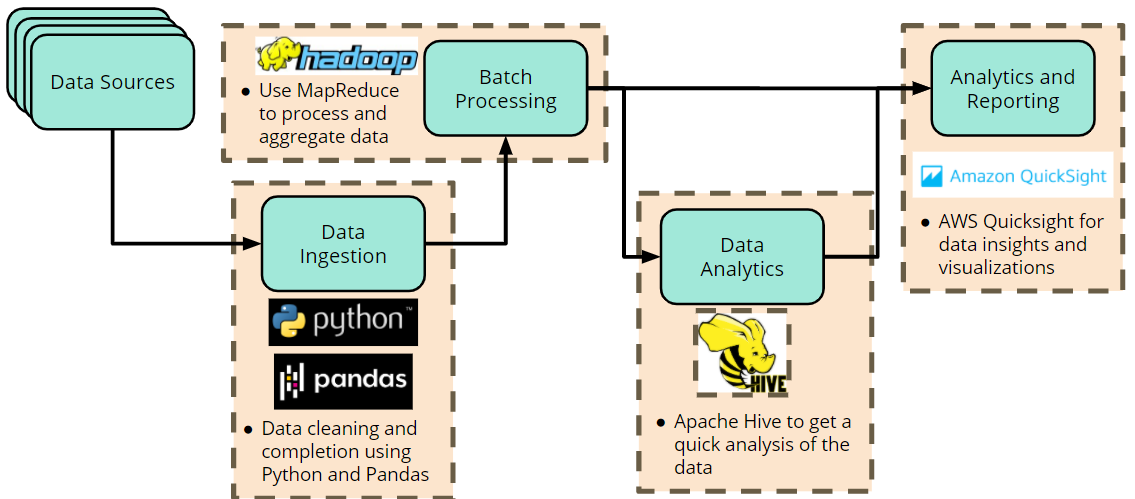


Figure 1 Architecture of our data analysis workflow

### Apache Hive

Apache Hive data warehouse software facilitates reading, writing, and managing large datasets residing in distributed storage using SQL [2]. It utilizes the advantage of HDFS and can query the dataset very easily and fast.

### Amazon QuickSight

Amazon QuickSight is a cloud- based business intelligence tool that can be used to analyze data, create visualizations, and derive insights for better decision making.

QuickSight is a fully managed service that will allow users to create and publish interactive dashboards. QuickSight can be used to access any the data set or derived tables in Hadoop Hive using a MySQL connection.

The figure below summarizes the role that Amazon QuickSight plays within a Big Data ecosystem

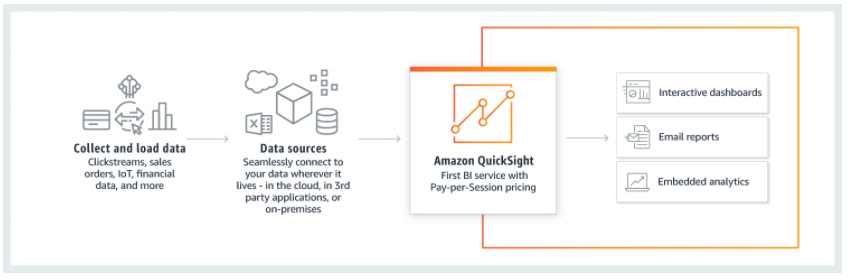


Figure 2 – How does Amazon QuickSight work? (Picture ref: Amazon QuickSight Homepage)

# Experiments

### Use Hive to get the statistics

First, we created the table called covid, defined its schema and loaded the original dataset into this covid table. Then based on the time of the experiment, which is Nov 12th 2020, we would like to find the total reported cases and deaths across the world and order them differently such that we can get the world’s safest and most dangerous countries.

### QuickSight to analyze and visualize the data

QuickSight was used to analyze the raw data to gain some insights into the analysis that would be needed. Any derived/queried data was also utilized to create visualizations to gain an understanding of the COVID-19 dataset. The analysis included the following –

1. Analysis of the countries affected the most by the COVID-19 Pandemic
2. Analysis of the health risk factors which could potentially affect the spread of the COVID-19 virus.
3. Analysis of the health risk factors which could potentially affect the mortality due to the COVID-19 virus.
4. Analysis of non-health factors, such as GDP per capita, poverty, handwashing facilities, etc., which could affect the spread of the COVID-19 virus or affect the mortality due to the virus.

### Use Jupyter notebook

Another tool we utilized for experiments is the Jupyter notebook based on Python. It was mainly used to find the correlation between various factors and COVID-19 cases and deaths. Below are the detailed steps.

The first step is to import matplotlib.pyplot, numpy and pandas in the notebook as well as to read the csv file we got from the dataset. A brief representation of the csv table was displayed in the output, showing the table has 56111 rows and 49 columns.

Upon a quick glance of the dataset table, we can see that within each country, the values of various factors such as “population” and “median\_age” remain the same for all the dates. Therefore, in the second step, we extracted the latest record to show only the case record on 2020-11-11 using a simple query. In this way, each country will have only one row containing all the factors we may need to analyze.

The third step is to drop the last two rows from the table which are representing “World” and “International” data using the “.head()” function. This is because we would like to analyze only the factors from different countries, not the total count from the world.

From the above three steps, we now have done a quick clean-up of the dataset and can start analyzing the result. In Step Four, what we did first in the data analysis is to sort by “total\_cases” from lowest to highest using the function of “.sort\_values(by="total\_cases", ascending=True)”. The last five rows of the output represented the top five countries who had the highest number of total cases (shown in Figure 11 in the results section).

Since we would like to see the relationship between various factors and COVID-19 cases and deaths, in the fifth step, a plot from two factors “total\_cases” and “population\_density” using scatter plot was conducted. We had originally hoped that based on the trend in the scatter chart, a possible relationship may be identified but the dots were not forming a clear trend (shown in Figure 12 in the results section).

In order to identify the possible relationships between various factors, upon discussion, we decided to run a correlation using “.corr(method ='pearson')” function on all the factors. Here is a brief explanation of the Pearson correlation coefficient according to realpython.com[3]:

The Pearson correlation coefficient measures the linear relationship between two features, as a ratio of the covariance of x and y to the product of their standard deviations. Pearson correlation coefficient can be denoted in letter r and it can be expressed mathematically in the following equation:

r = Σᵢ((xᵢ − mean(x))(yᵢ − mean(y))) (√Σᵢ(xᵢ − mean(x))² √Σᵢ(yᵢ − mean(y))²)⁻¹

The output correlation table contains 44 rows and 44 columns. The middle section from Row ”icu\_patients” to Row “stringency\_index” were all showing “NaN”. Other cells all have a value from -1 to 1. The last 15 rows represent the factors we are interested in analyzing such as “population” and “median\_age”.

With the understanding of the fact that “correlation does not indicate causation”, we can still see the strengths of the relationship between the factors in this correlation table. With “1” being the perfect positive linear relationship and “-1” being the perfect negative linear relationship, we picked up the correlation with the absolute value greater than 0.4. In other words, we identified some stronger relationships between the features (shown in Figure 15 in the results section).

### QuickSight to visualize Correlation

In this section, QuickSight was used to verify any derived correlations that provide insight into factors that affect spread of the virus or mortality due to the virus

# Results

### From Hive

As you can see in Figure 3 and Figure 4, the top safest places are mostly islands, since location advantage can better protect the places. Figure 5 and Figure 6 display the most dangerous places. As we can see clearly, the United States has way larger numbers than the rest of countries. So far, it has over 10 million total reported cases and over 240,000 reported deaths. Therefore, we would like to explore what the numbers are like for the United States per each month in 2020. The results can be seen in Table 1. We can see that in January and February, there were very few cases and no deaths. Then the number quickly increased in March and stays high ever since. Better visualization by Amazon QuickSight can help us better understand these numbers.

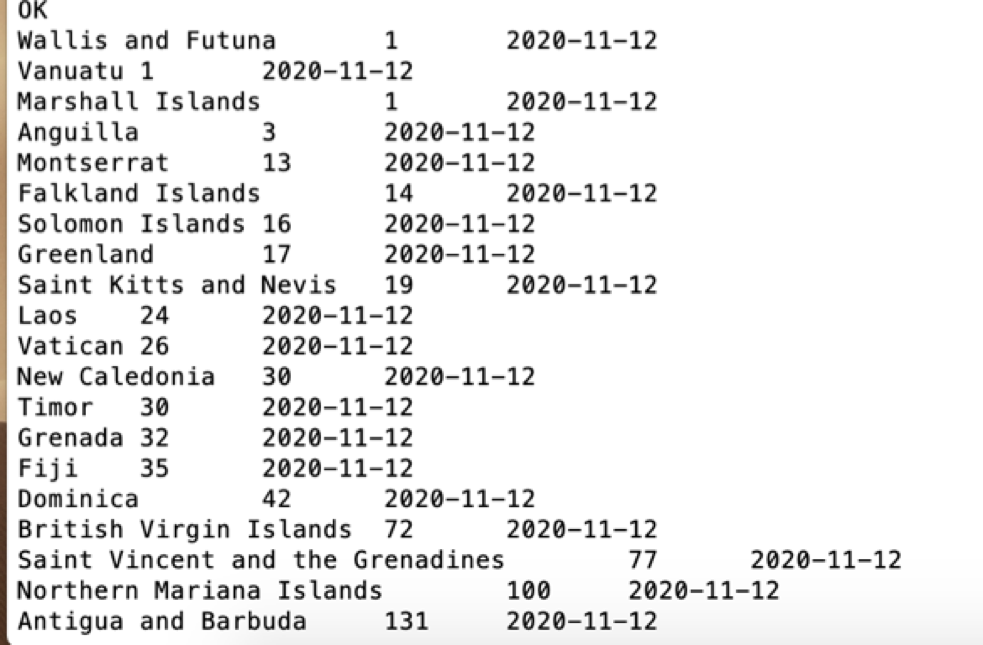


Figure 3 - The top "safest" place based on the total reported cases

**

Figure 4 The top safest place based on total reported deaths

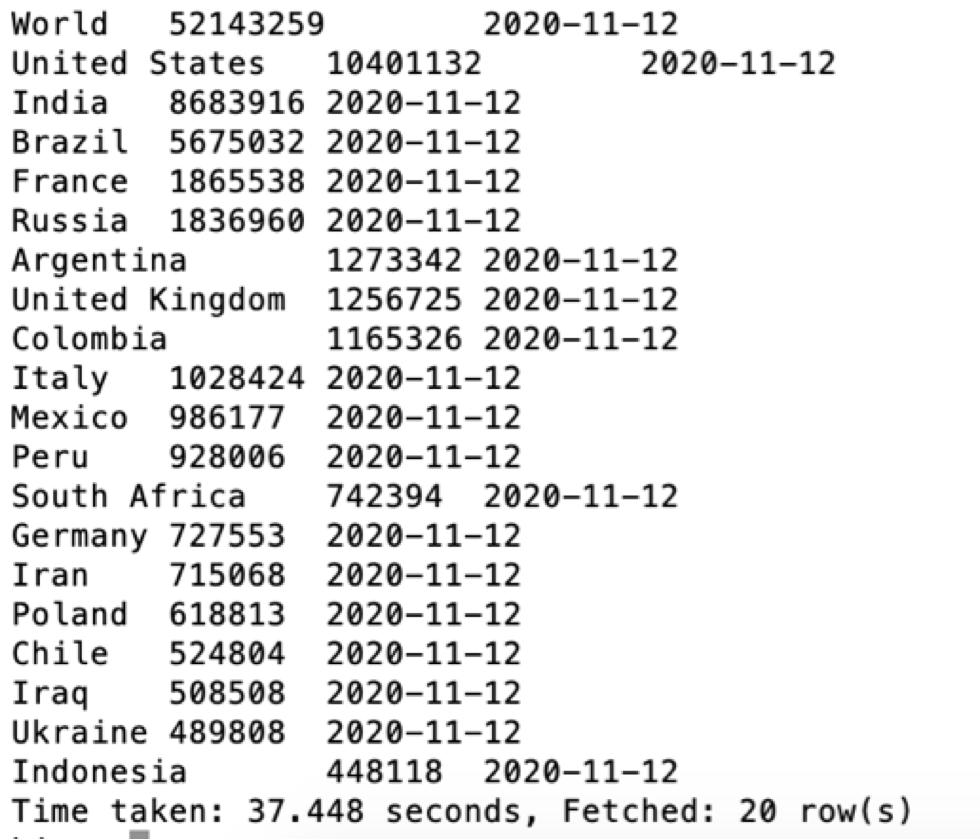
**

Figure 5 The top dangerous places based on total reported cases

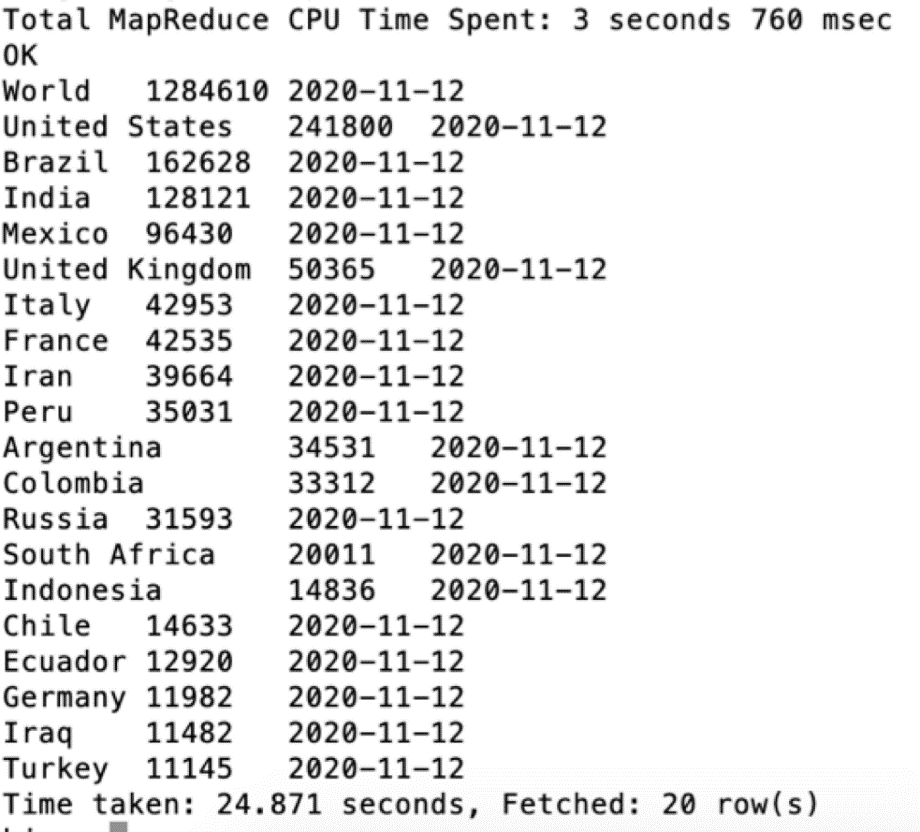


Figure 6 The top dangerous places based on total reported deaths

|  |  |  |
| --- | --- | --- |
| ***Month*** | ***Total cases*** | ***Total Death*** |
| ***Jan*** | *6* | *0* |
| ***Feb*** | *60* | *0* |
| ***Mar*** | *164554* | *3170* |
| ***Apr*** | *875289* | *57796* |
| ***May*** | *730475* | *42815* |
| ***Jun*** | *820168* | *22359* |
| ***Jul*** | *1904462* | *25930* |
| ***Aug*** | *1502149* | *30999* |
| ***Sep*** | *1193898* | *22929* |
| ***Oct*** | *1856366* | *23710* |
| ***Nov*** | *1353705* | *12092* |

Table 1 Monthly reported cases and deaths for United States in 2020

### From QuickSight to analyze and visualize the data

QuickSight was used to get context on the potential analysis that can be performed on the data. One of the initial analysis shows the 20 countries with the highest recorded cases of COVID-19 pandemic over the last 12 months. This is illustrated in figure 7.

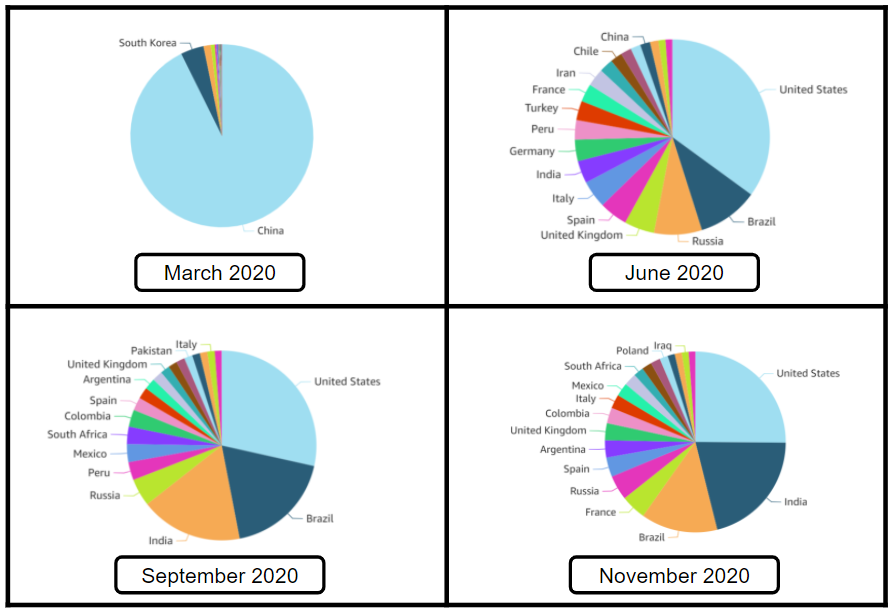


Figure 7 – Pie Chart showing countries (top 20) with highest cases over time

The data queries from Hive helped derive the results in QuickSight. The next steps included the analysis of the risk factors in the countries that have the highest recorded number of cases. The data for prevalence of male and female smokers is presented below in figure 8 –

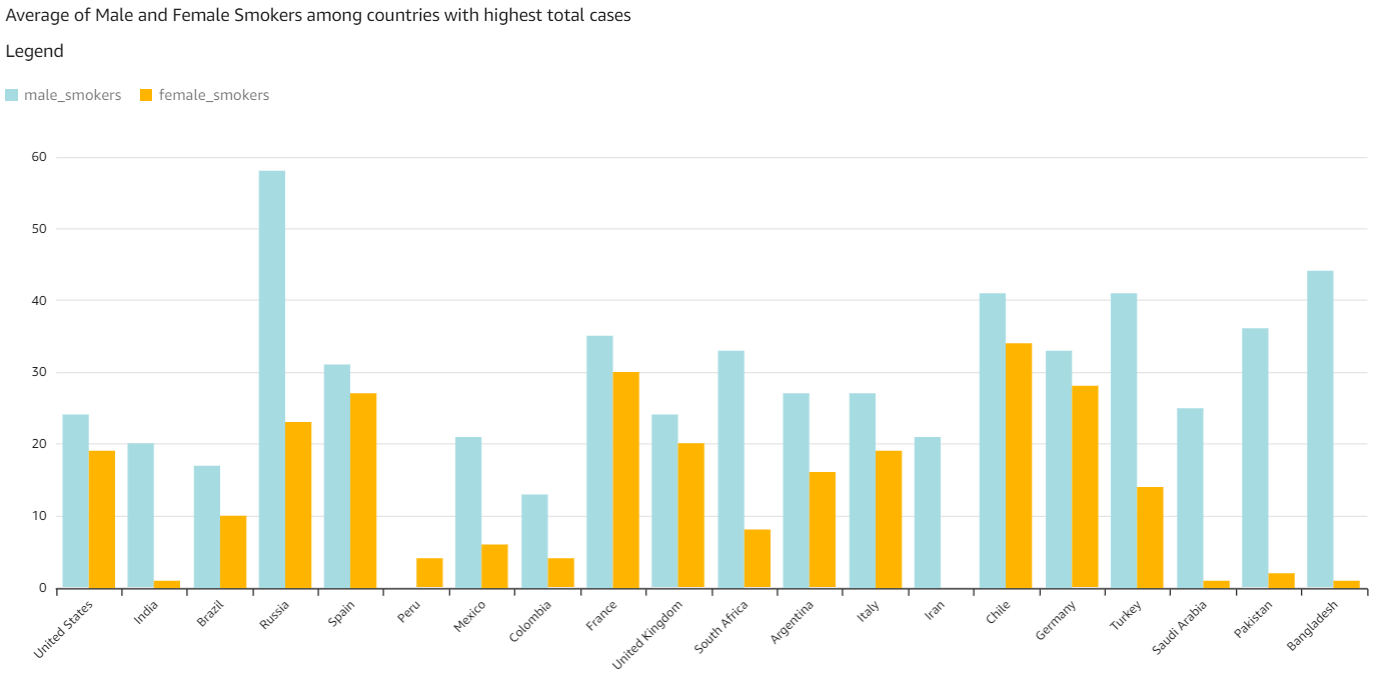


Figure 8 – Male and Female smokers for the countries with the highest cases

The consequent analysis was of the diabetes prevalence and the mortality rate among the countries with the highest recorded cases. The analysis revealed that the mortality and diabetes prevalence follow a similar trend in the countries. The visualization is shown in figure 9.

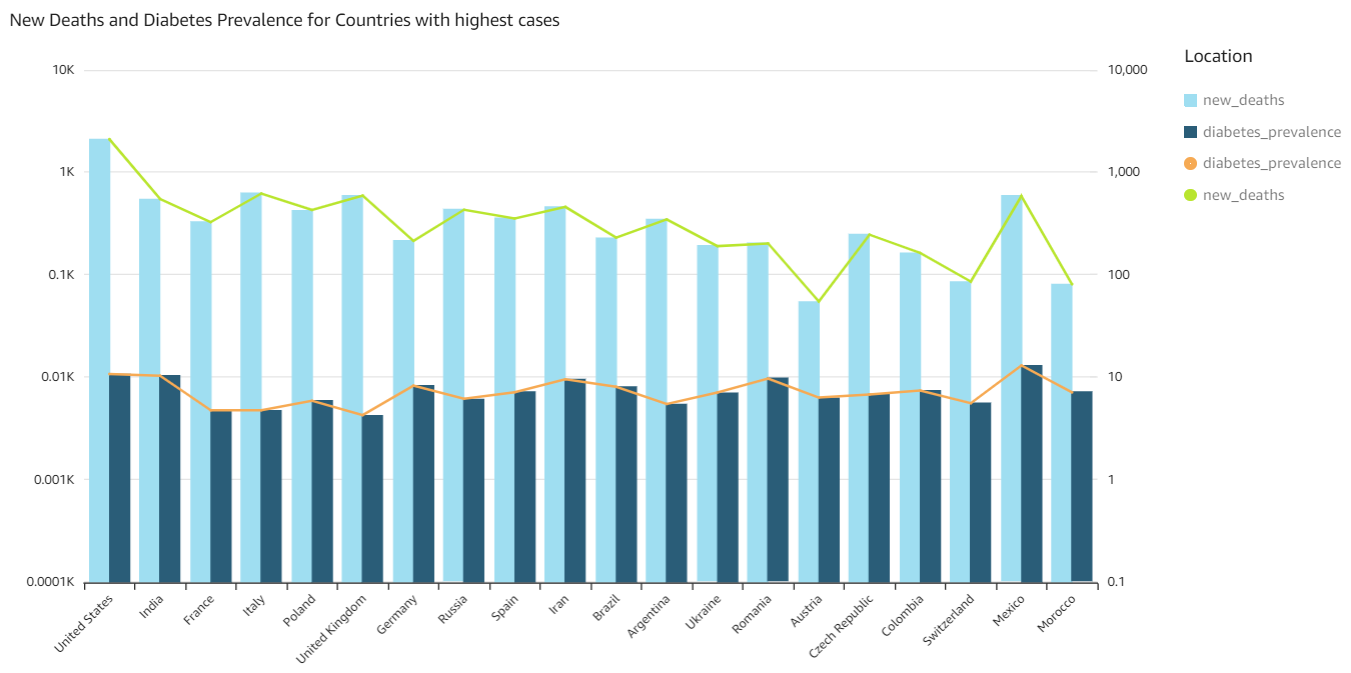


Figure 9 - Correlation between diabetes prevalence and mortality due to COVID-19 for the countries with the highest cases

Another notable analysis was conducted between the mortality rate and the cardiovascular mortality rate. This revealed that there is no apparent correlation between the two factors among the countries with the highest recorded cases. This is displayed in figure 10.

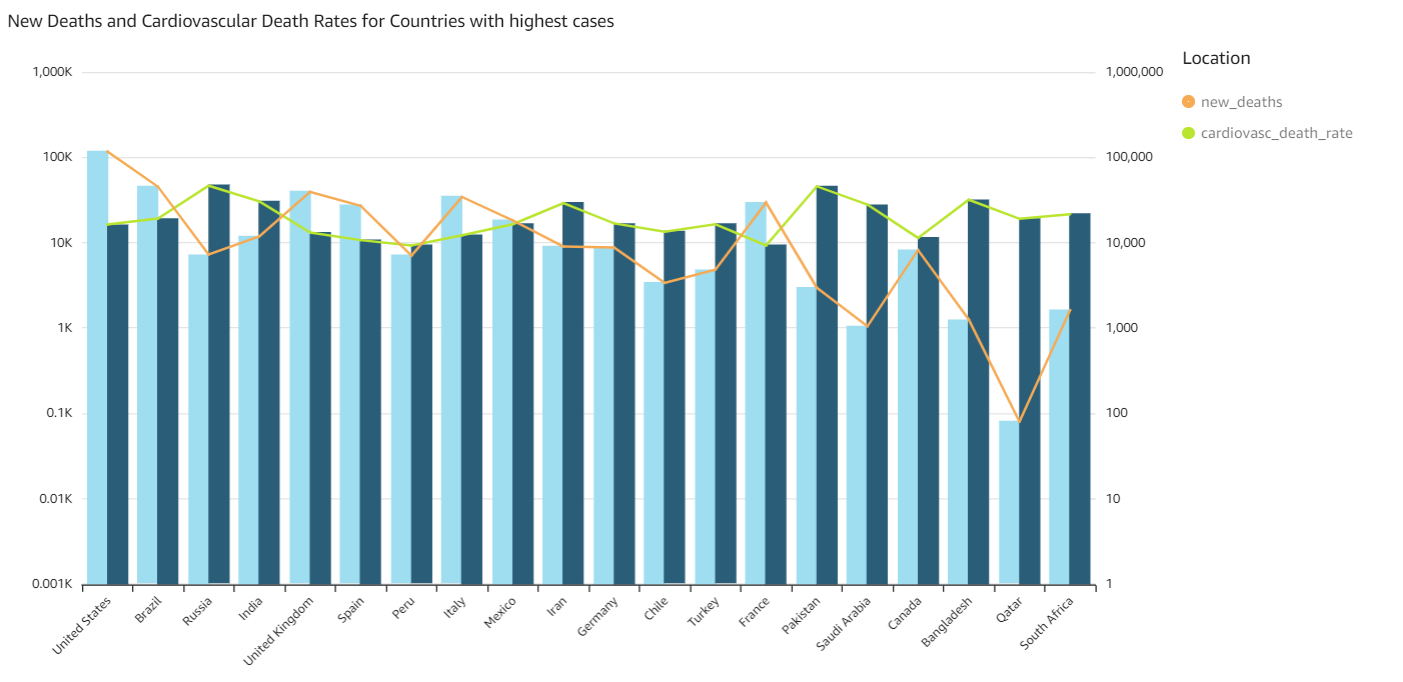


Figure 10 - Correlation between cardiovascular mortality and mortality due to COVID-19 for the countries with the highest cases

### From Jupyter notebook

The Step Four generated the table of countries in the latest record with total cases from lowest to highest. As you can see from the last two rows in the figure 11, the US, India, Brazil, France and Russia are the top 5 countries with the highest total cases. This result matches the result we got from Hive.

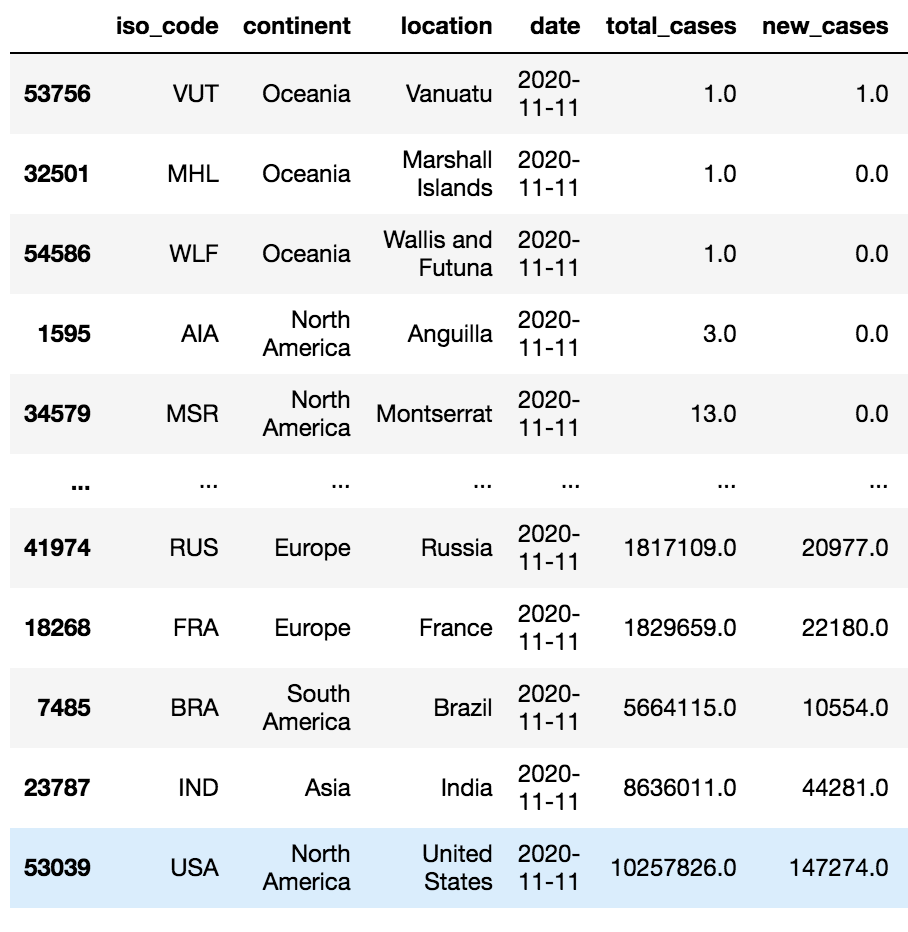


Figure 11 The countries with highest total cases

In Step Five, it is possible to conclude based on this scatter chart of Figure 12 that there is no clear link between total cases and population density.

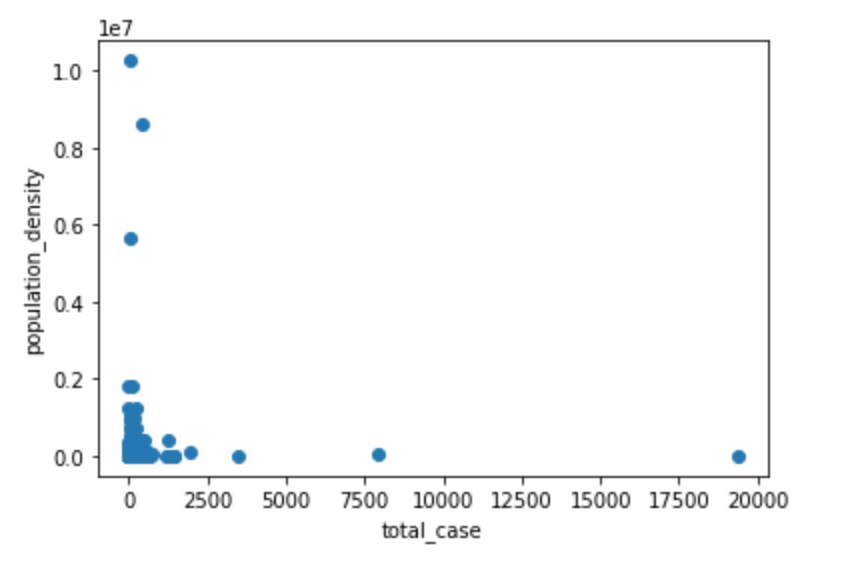
**

Figure 12 plot scatter of total\_cases and population\_density

Figure 13 shows the first column from the correlation table with “total\_cases” VS various factors such as “population” and “median\_age”. Based on these numbers, there is no clear indication of any strong correlation among these factors.

**

Figure 13 correlation between total\_cases and various factors

However, picking up correlations with absolute value greater than 0.4 such as Figure 14 shows below, we can see the strengths of the relationship between “total\_cases” and “total\_deaths” are relatively strong. This matches our intuition that the higher the population, the higher total cases and total deaths.

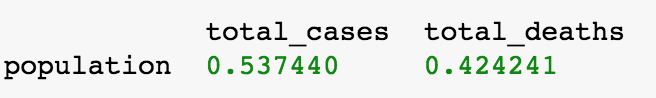


Figure 14 correlation between total\_cases and total\_deaths

Based on Figure 15 below, four possible related factors of COVID-19 were identified:

1. Age (based on median age VS total cases, median age VS new cases smoothed, age 65 older VS new cases smoothed, age 70 older VS new cases smoothed)
2. Finance (based on gdp per capita VS total cases, extreme poverty VS total cases, extreme poverty VS total deaths)
3. Female smokers (based on female smokers VS total cases, female smokers VS new cases, female smokers VS new cases smoothed, female smokers VS new deaths, female smokers VS new deaths smoothed)
4. Handwashing facilities (based on handwashing facilities VS total cases, handwashing facilities VS new cases, handwashing facilities VS new cases smoothed, handwashing facilities VS total deaths, handwashing facilities VS new deaths, handwashing facilities VS new deaths smoothed)

Therefore, some possible lessons we got from the results are:

1. Older generations and female smokers may be more susceptible to COVID-19. Groups that match these profiles should pay special attention.
2. Handwashing helps. Actions like increasing the handwashing facilities as well as making handwashing a habit should be encouraged.

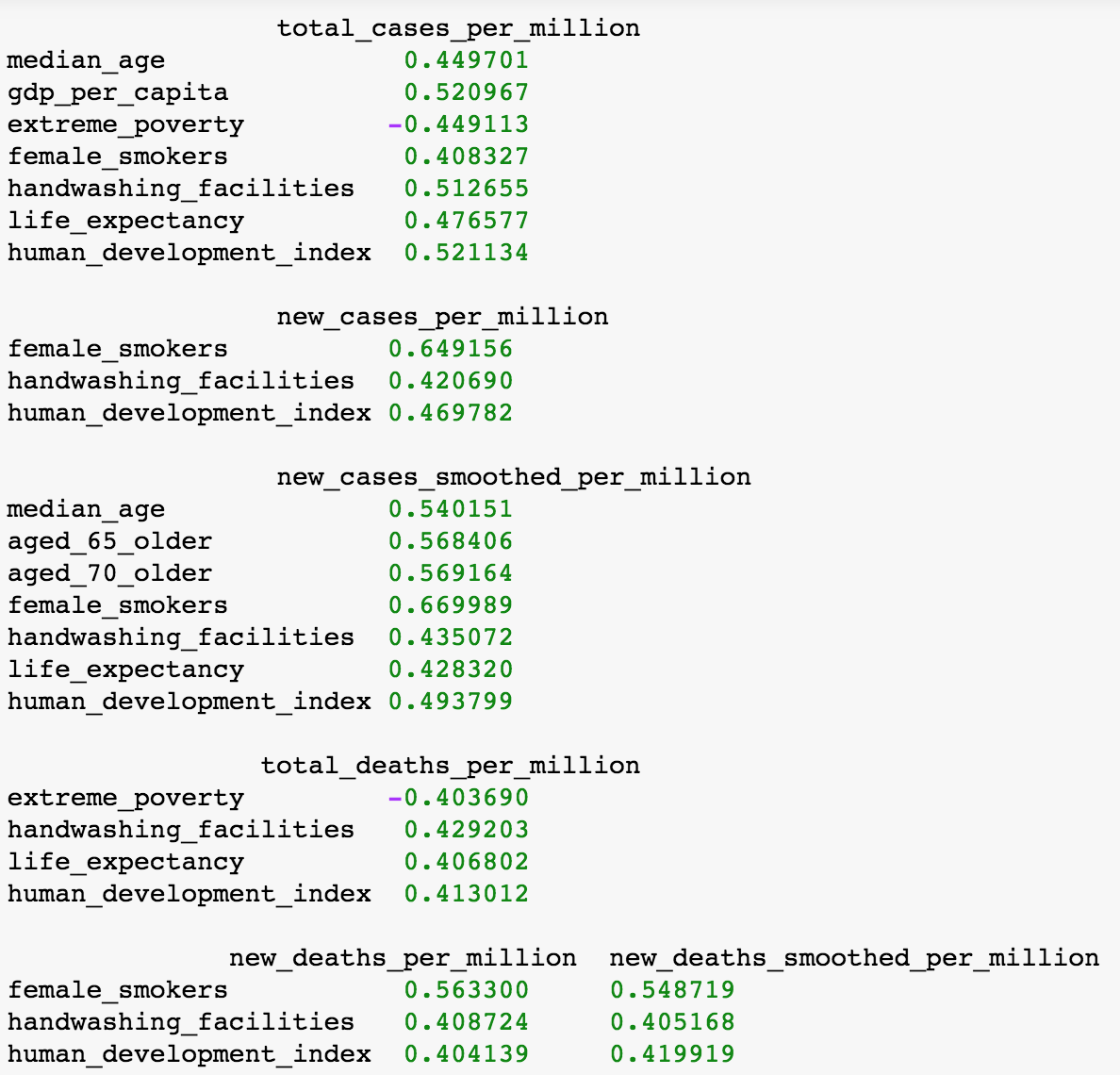


Figure 15 correlation among factors with absolute value greater than 0.4

### From QuickSight to visualize correlation

The correlation metric derived from the data analysis helps understand that there is evidence of dependency between the mortality rate and the habit of smoking in the female population. This relation is visually depicted in the graph in the figure 16. As expected, the trend lines for the ‘new deaths’ recorded follow the trend line for the population of female smokers in the top 20 countries with the highest recorded cases.

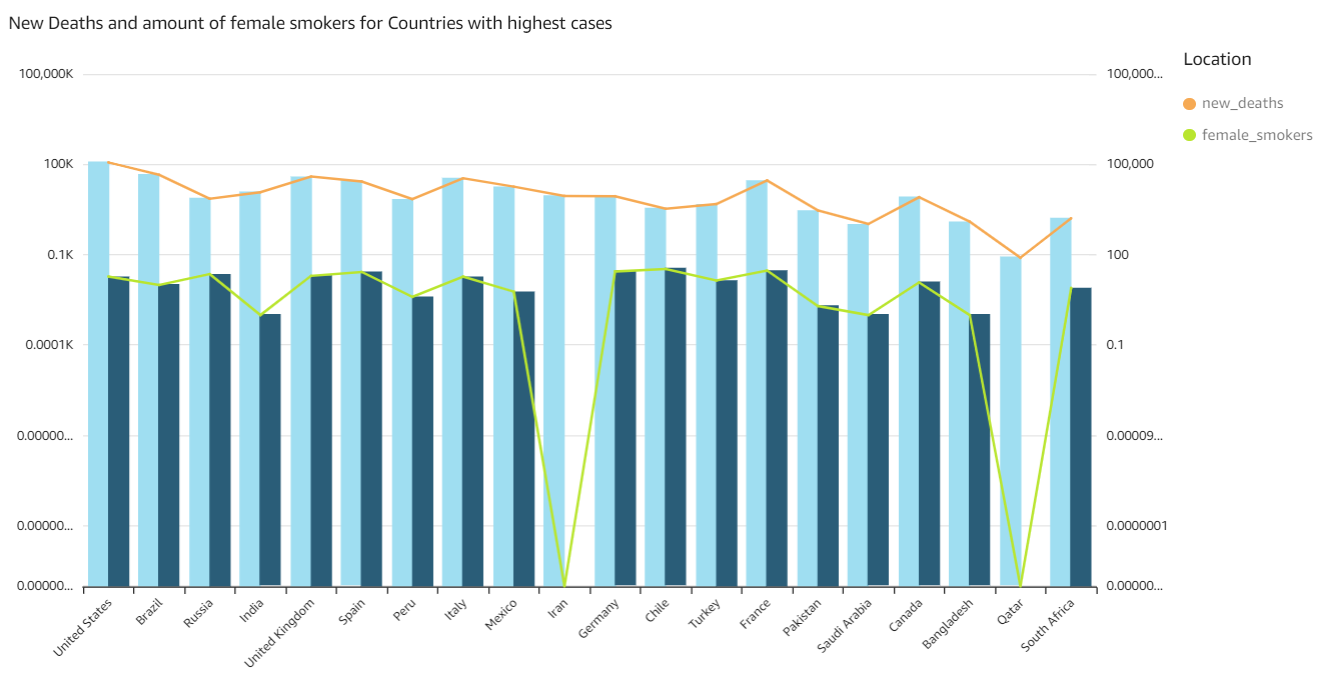


Figure 16 - Correlation between female smokers and mortality due to COVID-19 for the countries with the highest cases

Similarly, the analysis revealed a high correlation between the age and the total cases recorded in any country. This result is shown to coincide with the data visualization in figure 17 displaying the graphs of median age and total recorded cases for the top 20 countries affected by COVID-19.

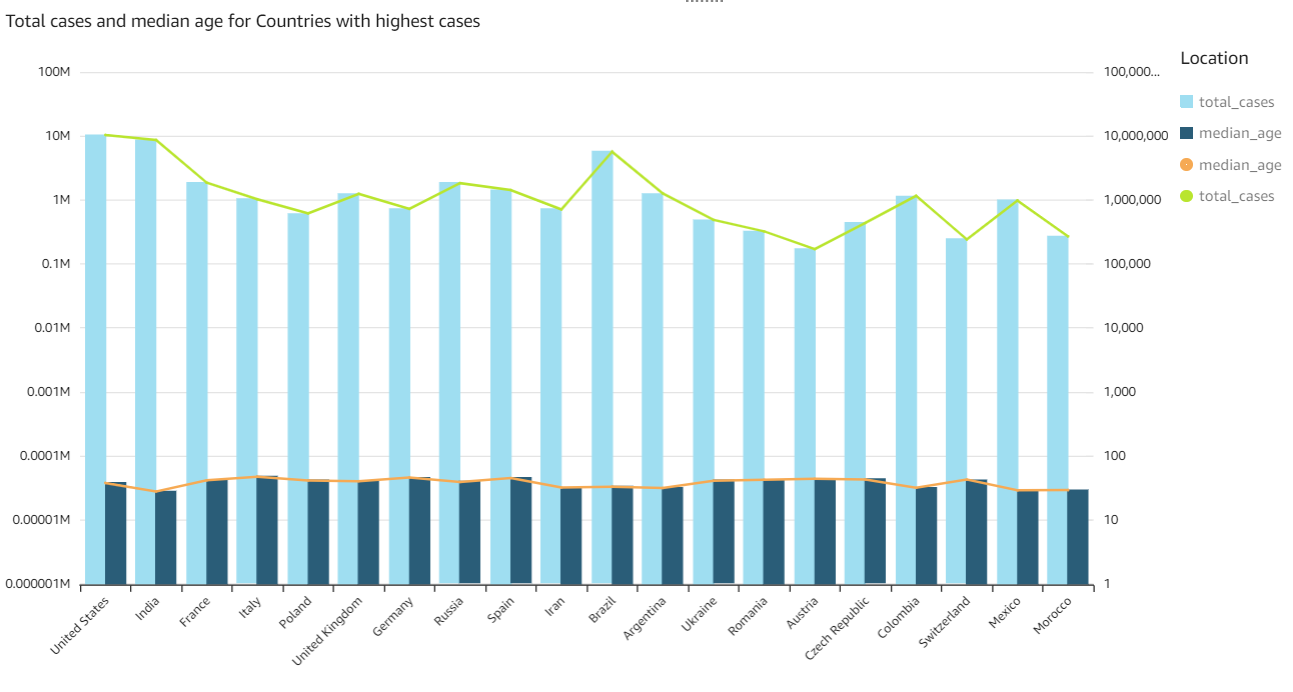


Figure 17 - Correlation among total cases and median age for countries with highest cases.

# Conclusion

In this project, the team was able to implement the following –

1. Ingest data using the HDFS and perform analysis using Hive
2. Use Hive to derive secondary data for analysis and visualizations.
3. Analyze and visualize data to get insights and inform the next steps for in depth data analysis.
4. Verify data insights and correlations using data engineering techniques.

The analysis of the data revealed that there is evidence of the following –

1. Mortality rate due to COVID-19 is affected by smoking in the female population. This suggests that female smokers are at a greater risk from the COVID-19 virus.
2. Median age is directly related to the mortality due to COVID-19. It can be concluded that older population is more susceptible to the COVID-19 virus.
3. Increasing the handwashing facilities as well as making handwashing a habit should be encouraged to reduce the risk of contracting COVID-19.

# Reference

[1] Hasell, J., Mathieu, E., Beltekian, D. *et al.* A cross-country database of COVID-19 testing. *Sci Data* **7**, 345 (2020). <https://doi.org/10.1038/s41597-020-00688-8>

[2] Apache Hive Homepage. Website - <https://hive.apache.org/>

[3] NumPy, SciPy, and Pandas: Correlation With Python. Website - <https://realpython.com/numpy-scipy-pandas-correlation-python/>

[4] Amazon QuickSight: Overview and Review. Website -<https://www.xplenty.com/blog/amazon-quicksight-overview-and-review/>

[5] Tutorial Points – Hadoop MapReduce. Website - <https://www.tutorialspoint.com/hadoop/hadoop_mapreduce.htm#:~:text=MapReduce%20is%20a%20processing%20technique,(key%2Fvalue%20pairs).>