The Impact of COVID-19 on Different Regions: A Comparative Analysis

Chenqi Zhao, Ziyue Feng, Chris Xu, Irene Yin

Abstract

In this study, we're using five different types of COVID-19 data to understand how the pandemic-related data, such as infection rate, hospitality rate, vaccination rate and Google covid-19 keyword search trend, varied from one region to another over the last three years. Moreover, the regional economic indexes were also considered in this analysis, exploring the correlation between epidemiology and economy.

By analyzing these datasets together, we aim to get a comprehensive picture of the pandemic's impact and leverage the analytic tools to help shape better public health decisions for future events.

Our analysis identifies a significant relationship between confirmed COVID-19 cases and hospitalization rates. Notably, Japan and the United States, which have high GDPs, also exhibit the highest rates of hospitalization. Furthermore, we observed a positive correlation between the frequency of keyword searches for symptoms and confirmed cases. This suggests that online search behavior could reflect and potentially predict epidemiological trends. SourceCode: https://github.com/IreneYIN7/Realtime-and-Big-Data-Analytics-FinalProject

1 Introduction

The novel coronavirus (COVID-19) pandemic has emerged as a global health crisis with profound impacts on public health, economies, and societies worldwide. The COVID-19 has indisputably been one of the most transformative global events in recent history. This paper aims to provide a comparative analysis of the impact of COVID-19 on different regions, with a focus on the relationship between confirmed cases, hospitalization rates, and economic indicators such as Gross Domestic Product (GDP).

This project leverages advanced data analytics and visualization tools, including Hadoop, Hive, and Tableau, to integrate four distinct types of daily COVID-19 datasets. These datasets encompass infection rates; vaccination data; Google search volumes for symptoms; and hospitalization and ICU admissions, along with regional economic indicators. Collectively, these datasets facilitate comprehensive analytics, offering deep insights into the national impact of COVID-19.

In light of these findings, our analysis seeks to delve deeper into the regional disparities of COVID-19's impact, examining the correlation between confirmed cases and hospitalization

rates, as well as the potential predictive value of online search behavior for epidemiological trends. By understanding these relationships, we can better prepare for and respond to current and future public health threats. If a pandemic event occurs again in the future, public health officials will be able to leverage this analysis to shape better public health decisions. In addition, this report can also be utilized by epidemiologists to gain better insights into this event. The ultimate beneficiaries are the general population, who stand to gain from improved health measures and better-prepared health systems.

2 System Design

The data pipeline designed for this study is structured to integrate and analyze diverse datasets related to the COVID-19 pandemic, as shown in Figure 1: Design Pipeline. The pipeline employs a combination of big data technologies to manage, process, and visualize data effectively, enabling a comprehensive analysis of the pandemic's impact across different regions.

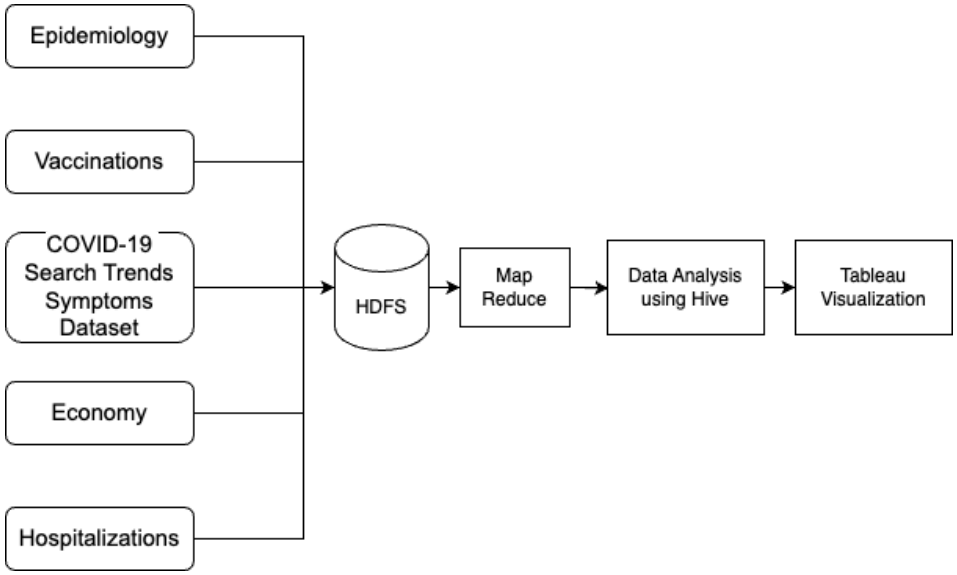

Figure 1: Design Pipeline

The pipeline integrates multiple datasets, including epidemiology data, vaccination records, Google search trends for COVID-19 symptoms, economic indicators, and hospitalization figures. These datasets are collected daily and vary widely in format and granularity. To manage this data efficiently, all datasets are first ingested into the Hadoop Distributed File System (HDFS). HDFS offers a reliable and scalable storage solution, suitable for handling the large volume of data generated by the pandemic.

Once the data is stored in HDFS, the next stage in the pipeline is data processing. This is achieved through a MapReduce framework, which allows for processing large data sets with a distributed algorithm on a Hadoop cluster. The MapReduce job performs two major functions: it parses the raw data to filter out unnecessary information, and it aggregates data points based on various parameters like region, date, and type of data (e.g., infection rates, economic impact).

Following processing, the data is analyzed using Hive. Hive provides a SQL-like interface to query data stored in a Hadoop environment, enabling complex analytical queries that facilitate the extraction of meaningful patterns and trends. This step is crucial as it transforms the processed data into actionable insights, highlighting correlations such as between the rates of infection and economic impacts, or vaccination rates and reductions in case numbers.

The final component of the pipeline is data visualization, executed using Tableau. Tableau interfaces with Hive to fetch the analytical results and presents them in an intuitive graphical format. This visualization supports the comparative analysis by illustrating trends and differences across regions, making it easier to communicate findings and insights effectively to both technical and non-technical stakeholders.

3 Methodology

3.1 DataSource Profiling

3.1.1 COVID-19 Epidemiology

This dataset records information about the infections of Covid-19 for different countries on a daily-based. It contains 12,525,825 instances and has 10 variables: date, location key, new confirmed cases, new recovered cases, new tested cases, cumulative confirmed cases, cumulative decreased cases, cumulative recovered cases, and cumulative tested cases. The size of this dataset is 520.0 MB.

Below is the table schema of the datasets:

Name	Type	Description	Example
date	String	ISO 8601 date (YYYY-MM-DD) of the datapoint	2020-03-30
key	String	Unique string identifying the region	CN_HB
new_confirmed1	Integer	Count of new cases confirmed after positive test on this date	34
new_deceased1	integer	Count of new deaths from a positive COVID-19 case on this date	2
new_recovered1	Integer	Count of new recoveries from a positive COVID-19 case on this date	13
new_tested2	Integer	Count of new COVID-19 tests performed on this date	13

cumulative_confirmed3	integer	Cumulative sum of cases confirmed after positive test to date	6447
cumulative_deceased3	integer	Cumulative sum of deaths from a positive COVID-19 case to date	133
cumulative_recovered3	integer	Cumulative sum of recoveries from a positive COVID-19 case to date	133
cumulative_tested2	integer	Cumulative sum of COVID-19 tests performed to date	133

Table 1: Covid-19 Epidemiology Dataset Schema

	date	location_key	new_confirmed	new_deceased	new_recovered	new_tested	cumulative_confirmed	cumulative_deceased	cumulative_recovered	cumulative_tested
0	2020-01-01	AD	0	0	0	0	0	0	0	0
1	2020-01-02	AD	0	0	0	0	0	0	0	0
2	2020-01-03	AD	0	0	0	0	0	0	0	0
3	2020-01-04	AD	0	0	0	0	0	0	0	0
4	2020-01-05	AD	0	0	0	0	0	0	0	0

Figure 2: Snippet of dataset – COVID-19 epidemiology

Figure 1 shows the snippet of the dataset COVID-19 epidemiology, if the value is NaN, represents the empty value, then replace that to value 0.

By using the MapReduce and pandas python library for the data profiling, we have the following table below:

Variable	Maximum	Minimum
Cumulative Confirmed Cases	92,440,495	0
Cumulative Recovered Cases	234,584,140	-3
Cumulative Tested Cases	9,214,000,000	0

Table 2: Data profiling of COVID-19 Epidemiology

3.1.2 COVID-19 Vaccinations

The dataset utilized for analysis comprises vaccination records, containing comprehensive data pertaining to the distribution and administration of COVID-19 vaccines. It includes columns detailing the number of new individuals vaccinated, the count of fully vaccinated individuals, the administered doses, and cumulative figures for these categories. While the dataset also encompasses specific vaccination information for individual vaccines such as Pfizer, we have chosen not to utilize this segment due to variations in data availability across different regions worldwide. Our analysis primarily focuses on cumulative data, necessitating thorough data

profiling and cleaning to extract the relevant columns. Furthermore, we aim to assess the impact of COVID-19 both within the United States, on a state level, and globally by country. To achieve this, we apply filters and mappings to key columns, delineating locations to their respective states within the US or countries worldwide. The size of the data file is 157MB and below is the data schema:

Name	Type	Description	Example
date	string	ISO 8601 date (YYYY-MM-DD) of the datapoint	2021-02-07
key	string	Unique string identifying the region	ID
new_persons_vaccinated	integer	Count of new persons which have received one or more doses	7222
cumulative_persons_vaccinated	integer	Cumulative sum of persons which have received one or more doses	782391
new_persons_fully_vaccinated	integer	Count of new persons which have received all doses required for maximum immunity	1923
cumulative_persons_fully_vaccinated	integer	Cumulative sum of persons which have received all doses required for maximum immunity	1239472
new_vaccine_doses_administered	integer	Count of new vaccine doses administered to persons	9146
cumulative_vaccine_doses_administered	integer	Cumulative sum of vaccine doses administered to persons	923449

${statistic}_${vaccine}	integer	Statistic value corresponding to a specific vaccine such as new_persons_vaccinated_moderna	1035

Table 3: Covid-19 Vaccinations Dataset Schema

To accomplish the filtering and mapping of location keys, the program parses each line, splitting the string using specified delimiters. The resulting elements consist of a date and a location key, which requires filtering. If the location key lacks underscores, it denotes a country code, which is then converted to the corresponding full country name. For example, "GB" becomes "United Kingdom" in the output file. Conversely, if the location key contains underscores, it signifies a state within the United States, and the output reflects the full state name. For instance, "AR_B_357" is transformed to "Arkansas", with "AR" representing the state code. Addressing the challenge of filtering cumulative data involves identifying null fields within the CSV file. Utilizing a regular expression pattern such as "^-?\\d+$" facilitates matching numerical strings like "12341" or "-74".

For enhanced file formatting and seamless integration with external tables in Hive, the reducer class adopts a strategy where all information is written into the key field in Text format, with the key itself being a null string. Extracted fields are then delineated by commas to ensure clarity and accessibility.

In addition to data extraction, we employ MapReduce for data profiling, enabling us to gain fundamental insights into our data files. Using the MinMaxCount mapper and reducer for each data file, we extract the minimum and maximum values of columns, along with the total number of rows associated with specific date and location pairs. Our findings reveal instances where certain columns exhibit unexpected negative values, despite an expected logical minimum of 0. Consequently, this necessitates filtering out these invalid data points during both data ingestion and subsequent analysis.

3.1.3 COVID-19 Search Trends symptoms dataset

This dataset reflects trends in symptom search patterns, but the source dataset contains search trends for hundreds of symptoms. When the symptom data in the region does not meet the quality or privacy thresholds for weekly resolution, the data is empty, so the dataset has lots of empty data. In order to better compare with other COVID-19 data sets, we specifically extracted the symptoms most closely related to COVID-19: ageusia, anosmia, common cold, cough, fatigue, fever, headache and vomiting. Even so, some symptom data are still empty, and we unify the empty data to 0. Because everyone has different symptoms of covid-19 due to their physical fitness, we summed the corresponding data of relevant symptoms to obtain a score representing the search volume, and obtained the changing trend of the score each day.

In addition, this data set covers data from different countries and regions and even different zip codes in different regions. In the project, we focus on analyzing countries around the world or different regions in the United States, so we only extracted data from relevant countries and regions in the United States. , and map the location abbreviation to the full name of the region.

3.1.4 Economy by region & COVID-19 Hospitalizations

The economy dataset has 4 columns: location key, GDP, GDP per capita, HCI. The hospitalization dataset has 8 columns: location key, date, new patient, cumulative patient, current patient, new/cumulative/current ICU patient and patients used ventilator. At the first glance, there are a few columns in both dataset that need to be removed, HCI and all columns related to patients with ventilators, because of the percentage of missing values. In addition, all columns related to the current patient needed to be removed because of the inaccuracy of the data during collection(noted in the datasource README file).

Two jobs were performed. First, in mapper, the hospitalization dataset was cleaned up and ranked by date and country. The reducer simply outputs the values. In the second job, one mapper was used to map both the economy and hospitalization dataset. In mapper, the column length is used to differentiate the two dataset, and concat a string “CASE” or “ECON” to output to the reducer. The reducer sums all new hospitalized patients and new icu patients within one region. At last, concat with the economy data as output.

3.3 Data Integration

We use Hive to combine different sets of data. First, we store all the data files in the Hadoop file system. Then, we create external tables in Hive for each dataset. This way, even if we accidentally delete a table, the data is still safe in the file system. After setting up these tables, we import the combined data into them. We do all this using NYU Dataproc, making sure our data is organized and accessible for analysis.

For storing joined data, we impose the same approach. For each joined table, we first create corresponding external tables in Hive, then we import joined data into the newly created table.

3.4 Data visualization

In the data visualization section, we employ Tableau for the task. We extract the joined table from Hive and import it into Tableau. Initially, we create line charts and bar charts to illustrate the data trends. Once we have identified the primary trends, we proceed to generate line charts depicting the relationship between two variables, revealing any shared trends. Additionally, we calculate the correlation coefficient and p-value between these variables to ascertain their statistical relationship, providing us with a deeper understanding of the results.

4 Result

Our initial observation reveals a notable similarity in the trends between the number of confirmed cases and the count of hospitalized individuals. However, upon closer examination

via the correlation coefficient, we identify a very weak positive linear relationship between the two variables. The statistical significance is evident from an extremely small p-value of < 0.05, indicating that this correlation is highly unlikely to occur by chance alone. Although the relationship is not strongly linear, the confidence in its existence is bolstered by this statistical significance (Figure 3,4).

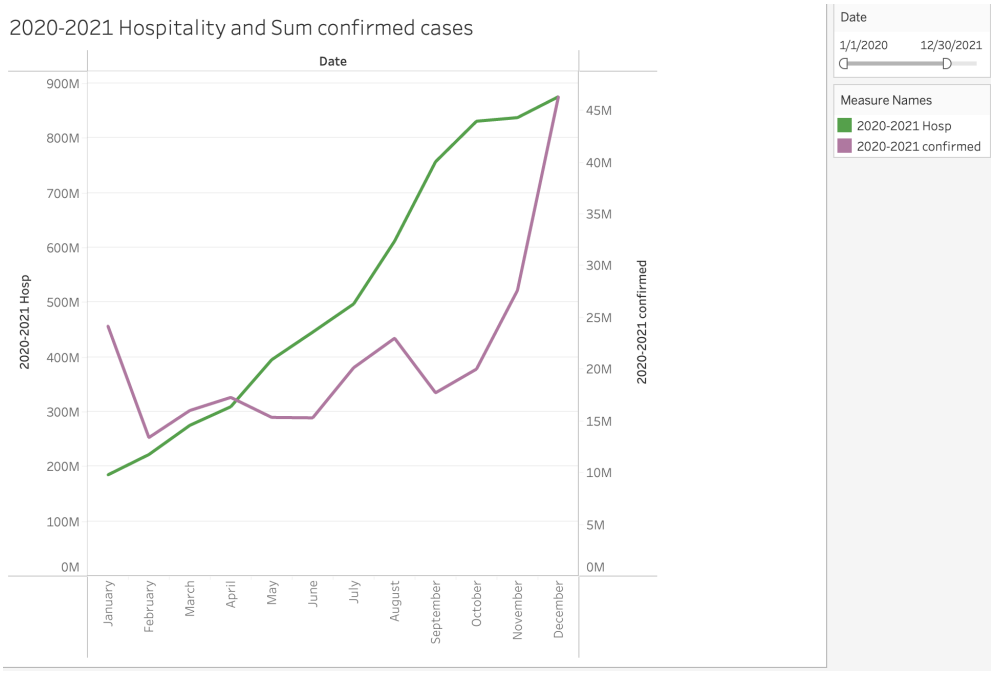

Figure 3: 2020-2021 Hospitality and Sum confirmed cases.

```
[15] # Calculate Pearson correlation and p-value
     correlation, p_value = pearsonr(filtered_df['sumConfirmed'], filtered_df['cu_hosp'])

     print(f"Correlation Coefficient: {correlation}")
     print(f"P-Value: {p_value}")

     Correlation Coefficient: 0.055297751910917875
     P-Value: 8.206520307219955e-10
```

Figure 4:Pearson correlation and p-value of sum confirmed cases and cumulative hospitality.

Another noteworthy finding pertains to the correlation between the hospitality rate and economic indicators (Figure 4). Interestingly, both Japan and the US exhibit high hospitality rates alongside the highest GDP figures (Figure 5).

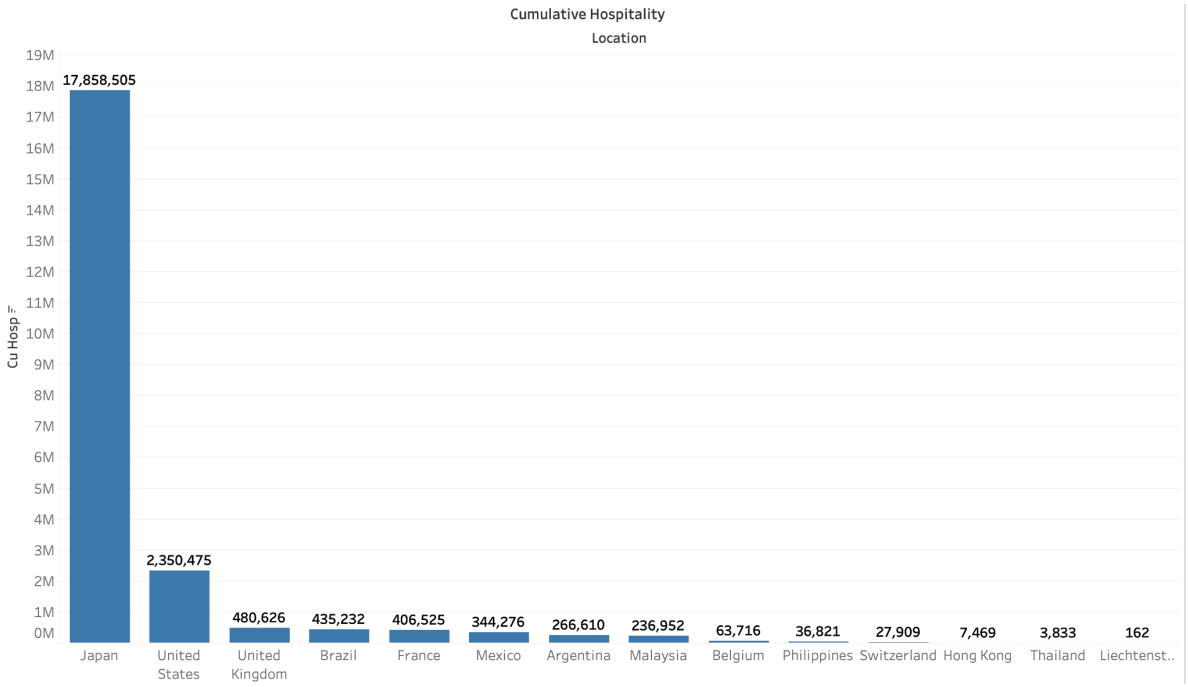

Figure 4: Cumulative Hospitality by different countries.

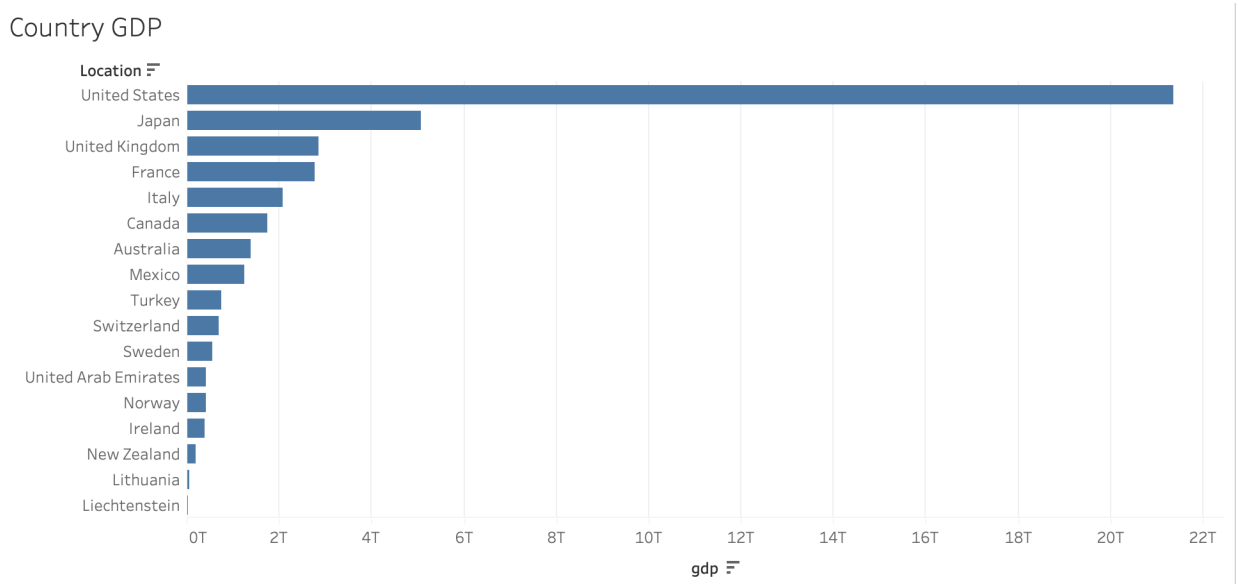

Figure 5: GDP of each country.

Additionally, our analysis uncovers an intriguing relationship between Google search trends and the number of confirmed cases (Figure 6). Specifically, we observe a positive correlation between Google keyword searches related to symptoms and the count of confirmed cases. The statistical significance of this relationship, supported by a p-value < 0.05, suggests a noteworthy association between symptom-related searches and confirmed case numbers.

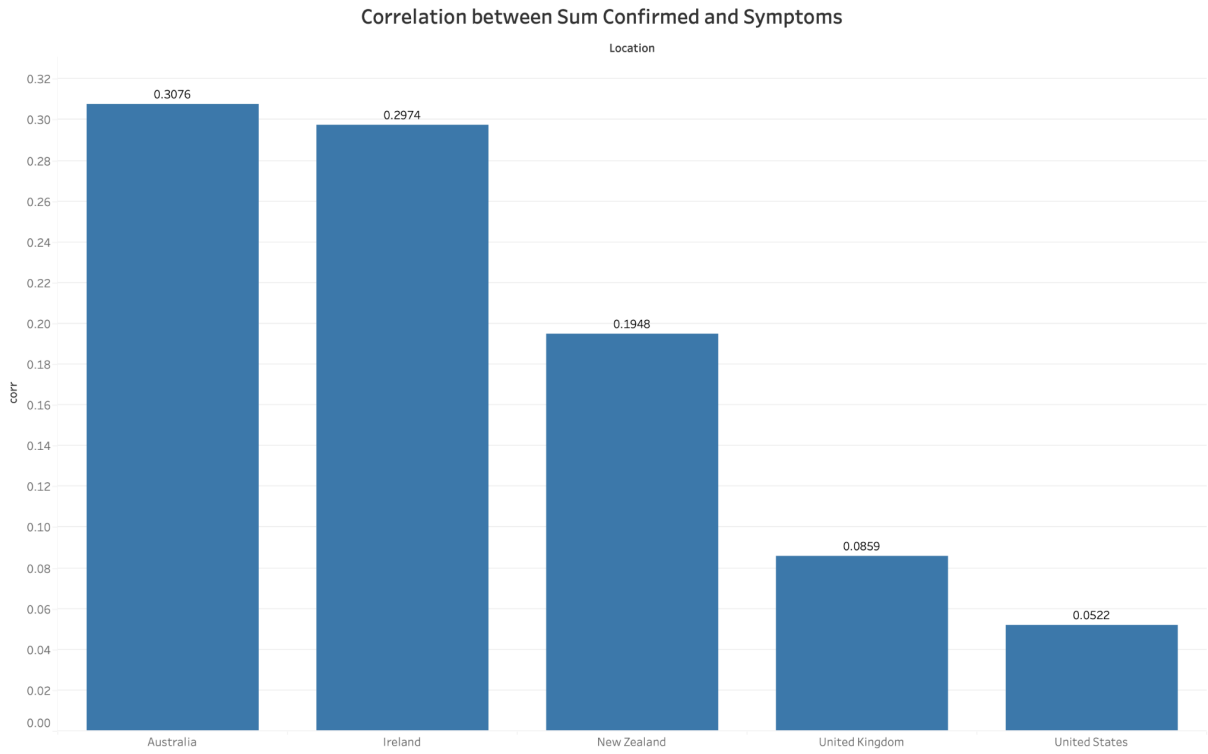

Figure 6: Correlation between Sum Confirmed and Symptoms

```
correlation, p_value = pearsonr(dftrend['sumConfirmed'], dftrend['symptoms'])

print(f"P-Value: {p_value}")
```

```
P-Value: 2.7685238505616954e-12
```

Figure 7: Pearson correlation and p-value of sumConfirmed and symptoms.

5 Conclusion

We performed data cleaning, profiling and analyzing to 5 covid-19 datasets, epidemiology, vaccination, Hospitalization, Google keyword search trend and economy by regions. The dataset was cleaned and profiled individually with Hadoop. Then all datasets were integrated using Hive. In the end, the Hive tables were exported to Tableau for visualization.
After evaluation, three results were concluded:

1.Sum of the confirmed cases and the number of people who have been hospitalized have similar trends. As the number of confirmed cases goes up, the number of hospitality also goes up.
2.Japan and the US have the most hospitality rate while having the highest GDP.
3.The google keyword search trend for symptoms has positive correlations with the number of confirmed cases, suggesting that online search behaviors can reflect and potentially predict epidemiological trends.

6 Reference

1. covid-19-open-data: https://github.com/GoogleCloudPlatform/covid-19-open-data
2. Guimarães, R., Policena, G., Paula, H., Pedroso, C., Pinheiro, R., Itria, A., Neto, O., Teixeira, A., Silva, I., Oliveira, G., & Batista, K. (2022). Analysis of the impact of coronavirus disease 19 on hospitalization rates for chronic non-communicable diseases in Brazil. PLoS ONE, 17. https://doi.org/10.1371/journal.pone.0265458.

1. COVID-19 Epidemiology
Data Size: 520.9 MB
Site:
https://github.com/GoogleCloudPlatform/covid-19-open-data/blob/main/docs/table-epidemiology.md
Profile & clean: Irene Yin

2. COVID-19 Vaccinations
Data Size: 157 MB
Site:
https://github.com/GoogleCloudPlatform/covid-19-open-data/blob/main/docs/table-vaccinations.md
Profile & Clean: Ziyue Feng

3. COVID-19 Search Trends symptoms dataset
Data Size: 1.9 GB
Site:
https://github.com/GoogleCloudPlatform/covid-19-open-data/blob/main/docs/table-search-trends.md
Profile & clean: Chenqi Zhao

4. Economy by region
Data Size: 10 KB
Site:
https://github.com/GoogleCloudPlatform/covid-19-open-data/blob/main/docs/table-economy.md
Profile & clean: Chris Xu

5. COVID-19 Hospitalizations
Data Size: 63.4 MB
Site:
https://github.com/GoogleCloudPlatform/covid-19-open-data/blob/main/docs/table-hospitalizations.md
Profile & clean: Chris Xu