# **Big Data Final Project**

## **Yalamanchili Krishna Suhruth**

## -**1st Year 2nd Semester.**

**Introduction:**

After the global pandemic the stress about averting another pandemic and keeping track of any possible situation is of high priority. After the novel Coronavirus we have been awakened to the sad reality of administration shortcomings and reporting difficulties.

Analyzing the information in real-time and producing insights for it might be a difficult adaptation to the heroes in the line of public health that is where machine learning tools come handy.

With data being consumed at the maximum in today’s time its only viable that we use the same data analysis techniques in the sector of public health to understand their present outcome if there is a chance for an outbreak or eradication of any future distresses. This report endeavors to leverage Big Data tools to analyze trends in disease-associated mortality and to provide a clearer picture of the pandemic's landscape.

**Background:**

In the rapidly evolving landscape of global health, the stakes are incredibly high. Epidemics such as COVID-19, recurrent threats like influenza, and emerging concerns such as Respiratory Syncytial Virus (RSV) demand a robust response reported by reliable data. As stated, before the critical nature of timely and accurate health data cannot be overlooked, as it can help make public health decisions affecting millions. However, the data due to abundance comes with its own challenges: extracting insights from vast and complex datasets requires advanced computational tools and fool-proof analysis techniques.

This project is rooted in the belief that through intelligent data analysis, we can discern the diseases impact across various demographics. It is an exploration of how the intersection of healthcare data and machine learning can lead to a deeper understanding and an efficient approach to managing public health crises. Taking advantage of the might of GCP's infrastructure, including the seamless data handling capabilities of BigQuery and the computational environment provided by VM instances, the project sets out to create an analytical model that sifts through the vast data to spotlight the crucial signals in health data.

The ultimate objective is to find trends and visualize the data in hand to extract evident information and build on it. Since this is a huge dataset, and this is something that is at the essence of time it is very beneficial that we get to do it in the cloud. Since working on the cloud will help us for scalability, with several people getting access to it innovation will be of key essence and since it is on the cloud modularity is also an added bonus. The people dealing with a particular problem can lead with only the disease they are worried about since we are reporting about different diseases. Through this initiative, we anticipate unveiling insights that can drive effective strategies and interventions, ultimately enhancing the resilience of health systems against the onslaught of infectious diseases.

**Proposed Approach:**

This project leverages the powerful computational capabilities of Google Cloud Platform (GCP) to conduct a detailed analysis of public health data, particularly focusing on the mortality rates associated with COVID-19, influenza, and RSV across various demographics. The approach combines numerical data analysis and visualization techniques within a Jupyter Notebook environment hosted on a GCP Virtual Machine (VM). This hybrid methodology ensures faster data processing, insightful analytics, and dynamic reporting capabilities.

**Dataset Description:**

The dataset contains the provisional percent of total deaths by week for COVID-19, Influenza, and Respiratory Syncytial Virus for deaths occurring among residents in the United States, by sex, age group, and race and Hispanic origin. Provisional data are based on non-final counts of deaths based on the flow of mortality data in the National Vital Statistics System.

The dataset is provided by the National Center for Health Statistics and this dataset is available to the whole public without any restrictions.

**A screenshot of a computer

Description automatically generated**

***Fig 1: Data Description-columns and column information type***

Researchers and policymakers can utilize this dataset to track the impact of respiratory diseases over time and across demographics, informing public health responses and resource allocation. It is also an essential tool for understanding trends, potential health disparities, and the efficacy of interventions.

There are 14.4K rows and 17 columns.

The description of the dataset that is columns describe what the column is and what is the dataset of the column with its description.

**A screenshot of a phone

Description automatically generated**

***Fig 2: Data Description-columns and column information type***

**Methodology:**

Google Cloud Platform (GCP) is a comprehensive set of cloud computing services provided by Google. These services encompass various functions such as computation, data storage, data analysis, and machine learning. As part of a course project, I became acquainted with GCP's user interface and decided to leverage its capabilities for my project.

Since wanted to establish the whole project in the cloud, I used Google Cloud Platform, where first I created a project named- mgmt\_final. Then later I created a virtual machine instance to keep all the information intact and to establish jupyter notebook work with it.

To get my work done in the GCP platform I enabled few APIs such as

* BigQuery API
* Dataflow API
* Cloud Storage API
* Compute Engine API

Then later created a bucket in GCP, since Once data is in a bucket, it can be easily accessed by various services within GCP, such as BigQuery for analysis, Cloud Machine Learning Engine for model training, or Dataflow for data processing.

It also comes with a bonus of scalability since we can add more information or update the storage details according to our needs. Later on we can modify accordingly. Google Cloud Storage integrates seamlessly with other GCP services. This infrastructure setup is a precursor to the intensive data processing and analysis tasks that follow in your project's lifecycle.A screenshot of a computer

Description automatically generated

*Fig 3: Confirmation of the enabled APIs*

Later on, added my information that is the cdc\_table about provisional deaths onto my bucket.

A screenshot of a computer

Description automatically generated

*Fig 5: Storing the csv file in the bucket.*

As shown the table was added to the bucket as cdc-table.

After uploading the table onto a bucket, to continue with the motive of completing the task in a cloud platform a virtual machine was established. A virtual machine (VM) is a digital version of a physical computer. Virtual machine software can run programs and operating systems, store data, connect to networks, and do other computing functions, and requires maintenance such as updates and system monitoring.

But the creation of the VM comes with its own challenges where initiating a network is of key essence, a VPC network helps for enhanced connection to our VM and also offers native internal passthrough for Network Load Balancers and proxy systems . This helps us connect our VM to multiple subnets and keep track of the firewall rules as to what IP addresses are allowed to be used and what are restricted. The main rule during establishing a VPC network is first allowing ingress and then establishing 0.0.0/0 as IPV4 ranges.

After creating the VPC and connecting it to the network initiation in the VM we can move ahead with establishing the Jupyter Notebook. So far we have created a bucket stored the table in it and then created VM for establishing cloud compatibility.

To establish local connectivity for the VM we further use the external IP with the command: scp -i "path:" " mgmtfinal-420614-5015c736d243.json " kryala@external-ip:/home/kryala/ after doing this the external IP with the SSH private key will be established for the username. After establishing the external IP address into the username with the SSH key which is *mgmtfinal-420614-5015c736d243.json*, we can proceed onto working in the Jupyter notebook.A screenshot of a computer

Description automatically generated

*Fig 6: Initializing Jupyter notebook in the VM via the SSH browser.*

Firstly, we download jupyter notebook onto the VM and then we begin by working on the external address. Since there is a new port that is available, we then add the new port into the TCP port address in the firewall rule, in our case that will be 8888, and have a secure token which was ‘MyS3cureT0ken!’, this is the token which will act as a password to further login to our jupyter notebook established on the VM.

Firstly in the jupyter notebook we have to establish several instructions which can only be possible via imports, for the working of the visualization we had to import:

*import pandas as pd:*

This is useful to read our table/csv file as a dataframe to further work with our implementation.

*import seaborn as sns:*

High level interface built on top of matplotlib for data visualization.

*import matplotlib.pyplot as plt:*

The classic matplotlib is used for implementing data visualization such as histogram, scatter plot and other basic visualizations.

*from google.cloud import bigquery:*

This is the main import for our project as this what helps with the connection to the cloud platform- GCP in our case. We can run bigquery as well as since we already have google.cloud we can use those functionality to make changes in our cloud module from the jupyter notebook all happening on the cloud.

When made these imports since running on GCP on student account and not admin account, there was an external restriction to bypass which created another virtual environment env and activated it and doing that was able to import and use the modules I desired to.

To establish connections the command I used was:

*import os*

*from google.cloud import bigquery*

*os.environ['GOOGLE\_APPLICATION\_CREDENTIALS'] = '/home/kryala/mgmtfinal-420614*

*client = bigquery.Client()*

*datasets = list(client.list\_datasets())*

*if datasets:*

*print("Datasets in project {}:".format(client.project))*

*for dataset in datasets:*

*print("\t{}".format(dataset.dataset\_id))*

*else:*

*print("{} project does not contain any datasets.".format(client.project)*

Which gave the result cdc-analysis which is the bucket we had created confirming that the connection had been established.

Now after connecting the VM’s jupyter notebook to the bucket and also the final project which was created, we can then move to the visualizations implementation.

After completing the visual implementation I realized a real time dashboard which accurately represents the trends would be a very good way to show findings as well and Google Cloud Platform is very well equipped to visualize findings with the help of Looker studio, I first used Bigquery where data was added in a table and ran SQL queries on it and further used visualizations in Looker by exporting results option.

A screenshot of a computer

Description automatically generated

*Fig 7: Initializing Dataset and creating table in BigQuery*

**Findings:**

After establishing a jupyter notebook connection for the VM instance we can proceed with the data visualization task that was intended to the table that was added to the bucket cdc-analysis with the name cdc-table.

The number of columns present in the table are:

Index(['data\_as\_of', 'start\_date', 'end\_date', 'group', 'year', 'month',

'mmwr\_week', 'weekending\_date', 'state', 'demographic\_type',

'demographic\_values', 'pathogen', 'deaths', 'total\_deaths',

'percent\_deaths', 'provisional', 'suppressed'],

dtype='object')

The main objective of the project was to do better reporting and to analyze any trends that can be concluded based on the visualization that we obtain for that the best visualization options were histogram, line chart, area chart and pie chart to view distribution.

**LINE PLOT:**

**A graph with a line

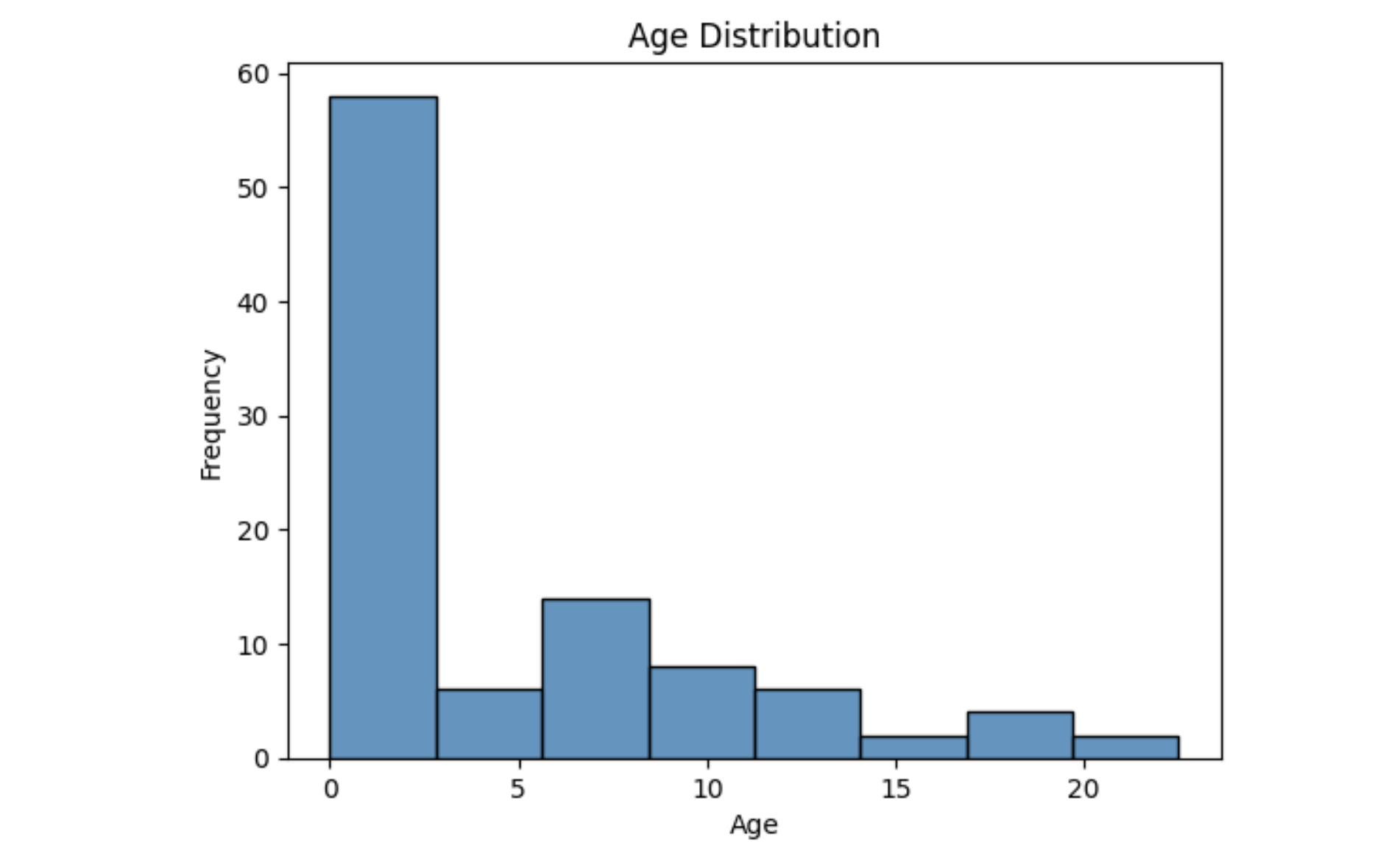
Description automatically generated**

*Fig 8: Line Plot indicating the monthly trend of deaths due to the pathogens.*

This line plot illustrates the monthly trend of percent deaths due to diseases such as COVID-19, influenza, and RSV. The plot shows a time series analysis from the combined years' data, highlighting significant fluctuations in mortality rates, which could correspond to seasonal patterns or specific public health responses. The visualization clearly indicates there have been multiple deaths from 2021-2022, this might be due to the third wave with the Delta variant and also an outbreak of influenza.

**HISTOGRAM:**

The histogram showcases the distribution of percent deaths, providing insights into the commonality of different mortality rates. This visualization helps identify the most frequent percent death rates and outliers in the data. Such a distribution can inform public health officials about the typical severity of outbreaks and guide preparedness measures. The histogram shows that multiple deaths happened during infant age might be due to do immature immunity.



*Fig 9: Histogram representing frequency of deaths vs Age.*

**LINE PLOT FROM PIVOT TABLE:**

For this particular visualization, limited the querying and the results to 1000 rows to get more particular results. The command for that is:

*from google.cloud import bigquery*

*client = bigquery.Client()*

*query = """*

*SELECT \**

*FROM `mgmtfinal-420614.cdc\_analysis.cdc\_table`*

*LIMIT 1000*

*"""*

*query\_job = client.query(query)*

*df = query\_job.to\_dataframe()*

*df.head()*

A graph with colored lines and dots

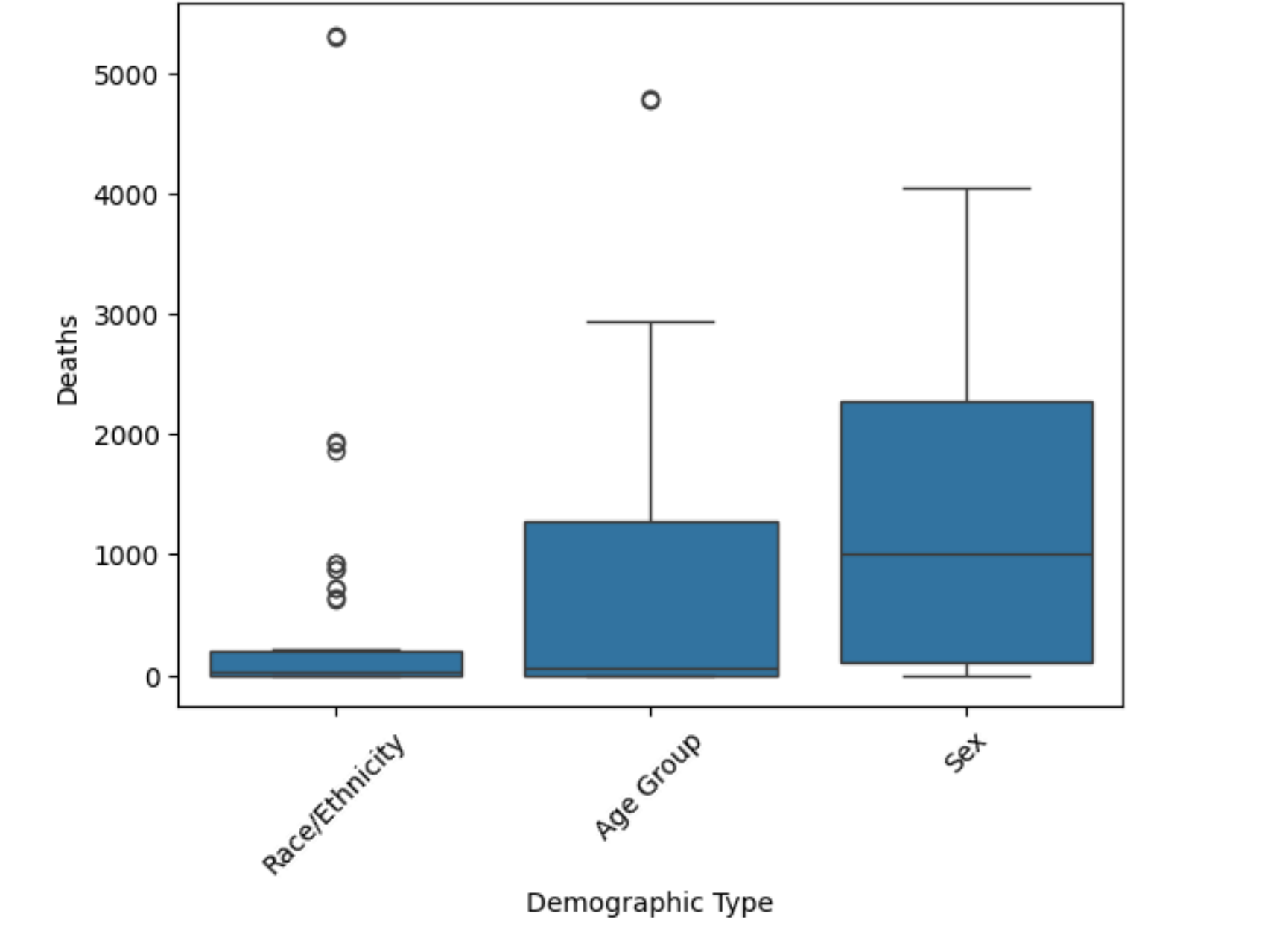
Description automatically generated

*Fig 10: Line plot for the pivot table showing seasonal trends based on months.*

This line plot illustrates the variation in total deaths by month and year. It reveals potential seasonal patterns or spikes in deaths that could correlate with flu seasons or pandemic waves. The visualization effectively demonstrates year-over-year trends and variations, offering insights into how during what month of the year the total deaths spiked and where there were less, like in peak winter December and January the deaths spiked indicating seasonal impact.

**BOX PLOT:**

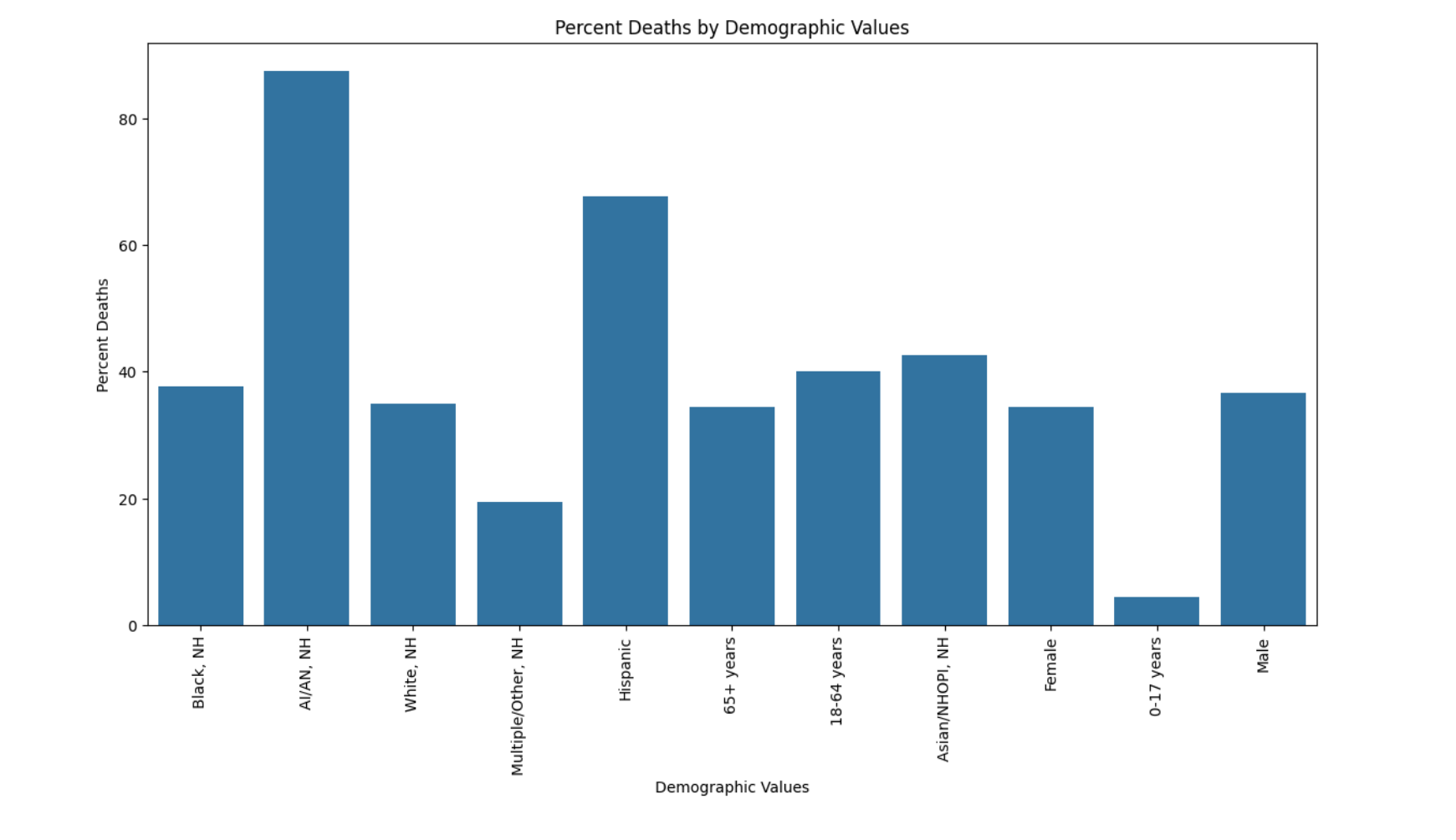
This boxplot categorizes deaths by different demographic types, revealing the median, quartiles, and outliers within each category. This is a very good indication to categorize a target demographic to know what sector got hit the most and what can be first targeted to control any outbreak. In this case sex acted as a major distinguisher.



*Fig 11: Box Plot revealing which demographic of people were effected most by deaths*

**BAR PLOT:**

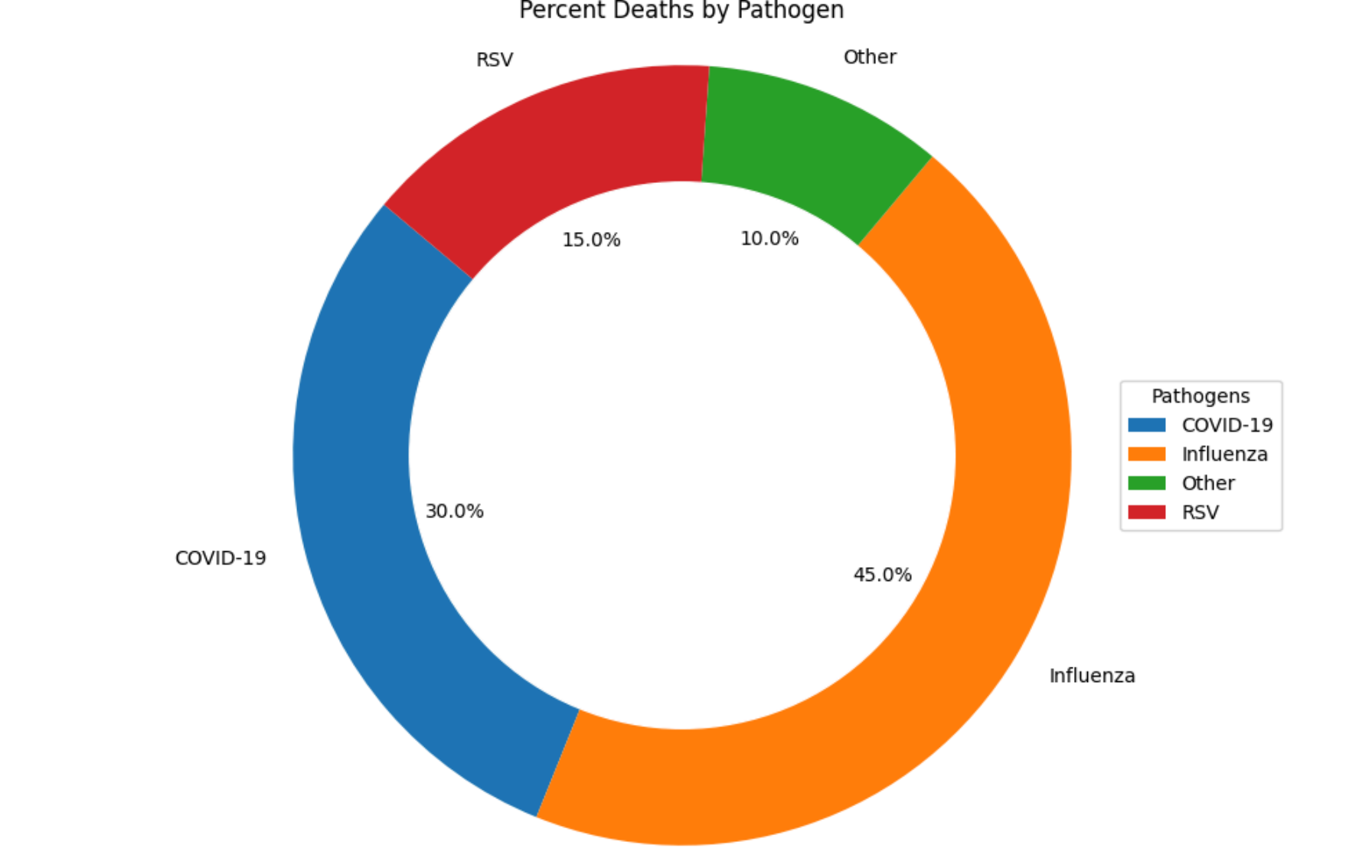
A bar plot to illustrate percentage deaths across different demographic values offers insights into which groups within the demographics are most susceptible to the diseases studied. This can be crucial for targeted public health responses and for understanding the impact of socio-economic factors on disease mortality. This bar plot shows AI/AN and Hispanic were severely hit might be due to genetics or the food consumption or the living habits, this can help understand what is the triggering point for the pathogen.



*Fig 12: Bar plot to understand the distribution of graphs among different demographics.*

**PIE PLOT:**

This pie chart provides a straightforward visual comparison of the contribution of each pathogen to total percent deaths. It highlights the relative impact of each disease, allowing for easy comparison and potentially guiding resource allocation and public health prioritization. With this pie plot we know COVID-19 accounted for almost 50 percent of the deaths, this chart can be majorly used for reporting purposes rather finding trends.



*Fig 13: Percent of deaths by each pathogen*

**Looker Studio Visualizations:**

* In Bigquery ran SQL function to obtain state wise breakdown of deaths and pathogen, which is very useful in obtaining geographical analysis of the impact.

After getting the results exported the SQL query to Looker studio where implemented Heatmap and geographic distribution.

A screenshot of a computer

Description automatically generated

*Fig 14: SQL Query to get state wise comparison for deaths*

The link of the lookerstuido dashboard is provided below:

<https://lookerstudio.google.com/reporting/c6023057-b41e-4073-8941-f39a7df3a9d1>

The dashboard looks like: A screenshot of a computer

Description automatically generated

*Fig 15: Looker studio dashboard for state wise death query*

This is a very impactful way to understand which areas have been hit the most and where the need of the hour for change. By hovering the region, we can get further information as well.

* For the second visualization in Looker Studio, I found was to compare the deaths over time to establish a time-series trend to understand where were right and wrong in a particular time period and what we need to replicate/avoid.

The SQL query for which, ran in Bigquery is :

A screenshot of a computer

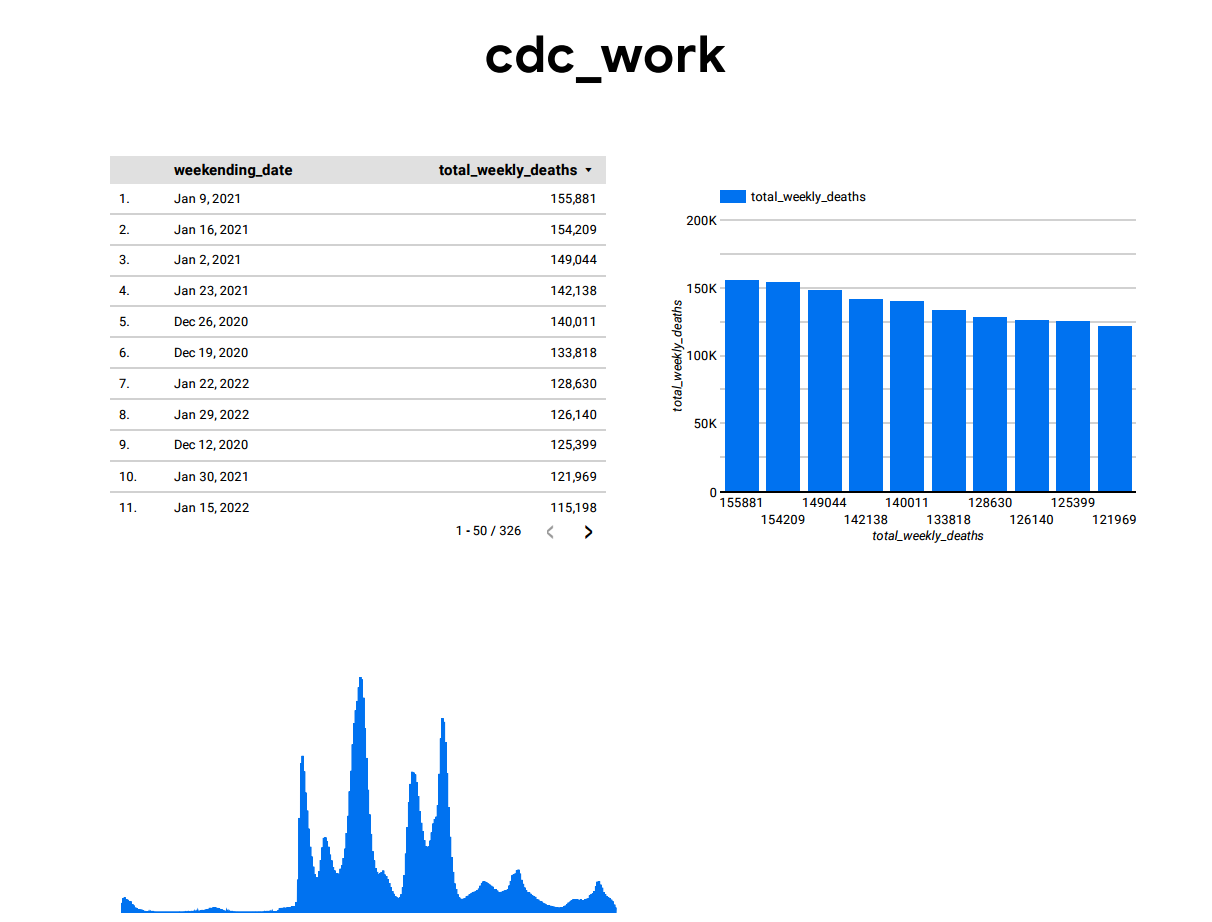
Description automatically generated

*Fig 16: SQL query to compare deaths frequency over time*

In this used the explore data option in which Looker Studio is provided.

The link for the visualization is:

<https://lookerstudio.google.com/reporting/0fa025f3-c173-47cb-b560-b0014d2d8cc3>

The visualization looked like

*Fig 17: Looker studio dashboard for SQL query to get time trends for death count*

This visualization can help us understand a time trend, by hovering over the graph that is on the peaks and the troughs the count is also indicated with the dates, which can very useful in the reporting process.

**Discussion:**

The findings of this project provide great scope for further discussion and development. The achievement of this project is to provide visual trends and information in the sector of public health for forecasting future epidemics based on the diseases which are a growing problem. The visualization acts as a key guiding metric to understand what is going on and where further work needs to be done. This project was commenced using the Google Cloud Platform which in our case was used as PaaS(Platform as a Service), since we used GCP for complete stretch that is implemented a VM, called a bucket for storage, customized VPC and then further used Jupyter Notebook from the SSH browser of the VM with the use of the SSH private key. The course provided great exposure to google cloud with several qwiklabs assignment where we must create our VM instance, create our own bucket for storage, and in the week where we had to data pipeline helped me come up with the idea of visualizing our results. The VM instance assignment helped me understand where virtual machine is used, and how it can be authenticated with the bucket where I even used the SSH browser for the first time. The pipeline assignment ends with visualizing the information we added in the pipeline, so I took elements from the VM, bucket storage, and the pipeline assignment, integrated them to form an overall project to forecast public health results in a cloud platform. Anyone after logging into the GCP account can access the VM and then after accessing the VM can open the SSH browser, authenticate the external IP address they are prompted with and then further turn on the Jupyter Notebook to find the work done. They can either check the notebook and do developments but also can create DataProc clusters or perform any form of updating or scalability for their requirements which was the mission of the project.

The barriers that I had to jump through were majorly with the integration of the VM to my local system, as in the assignments they were pre-built and I have never worked with a customized VPC before for further access, so coming up with that idea and that work through took longer than I anticipated. The other issue was with the firewall rules for the VPC as I had to particularly mention the IP address ranges and the ports and the ingress rule to properly use my cloud platform for integration with the local system for accessing jupyter on the external IP. Finding the private key which is usually in the .json file was tricky and the scp command was also difficult. The final barrier was faced in imports since using this as a student account I was not allowed to make imports of my choice for which I had to choose the bypassing mechanism of activating a virtual environment. So that I would consider a failure as a virtual environment has to be invoked every time the jupyter notebook has to do its job.

Scope for further improvement majorly consists of real-time interactive dashboards and implementation of machine learning predictive models.

**Conclusion:**

This comprehensive analysis of public health data, particularly focusing on the mortality rates due to COVID-19, influenza, and RSV, reflects the huge impact these pathogens have on various demographics across different time frames. Using the advanced computational capabilities of the Google Cloud Platform, the integration of numerical and textual data analysis has significantly enhanced our understanding of disease dynamics, boosting our ability to forecast and respond to public health threats more effectively. The scalability of cloud solutions ensures that as data volumes grow and the granularity of data improves, our analytical capabilities can expand, enhancing our responsiveness to public health crises.

In conclusion, this project highlights the critical role of data-driven insights in shaping public health policy and practice, setting a benchmark for future epidemiological research. The methodologies and technologies used provide a blueprint for similar health data analyses, potentially leading to more informed and effective public health strategies. As we continue to grow on these analytical models and expand our data sources, we anticipate a future where public health responses are not only more proactive and precise but also more impactful.

**References:**

1) Olawade, D. B., Wada, O. J., David-Olawade, A. C., Kunonga, E., Abaire, O., & Ling, J. (2023). Using artificial intelligence to improve public health: A narrative review. Frontiers in Public Health, 11. https://doi.org/10.3389/fpubh.2023.1196397

2)Google Cloud Platform Quests- [Creating a Virtual Machine](https://www.cloudskillsboost.google/focuses/3563?parent=catalog), [Ingesting Data to BigQuery](https://www.cloudskillsboost.google/focuses/3692?parent=catalog)

[Rent-a-VM to process Earthquake Data](https://www.cloudskillsboost.google/focuses/1846?parent=catalog)

3)Google Cloud Platform documentation <https://www.cdc.gov/flu/weekly/index.htm>

4) Weekly U.S. Influenza Surveillance Report- <https://www.cdc.gov/flu/weekly/index.htm>