

SP24: Management Access and Use of Big and Complex Data

Final Project Report

On

Analysis of Covid-19 Data

Alvan Dmello

2001251305

admello@iu.edu

Luddy School of Informatics, Computing and Engineering

Indiana University Bloomington

Bloomington, IN 47405, USA

Index

Pg. No.	Content
3	Introduction
3	Background
4	Methodology
13	Results
22	Discussion
24	Conclusion
24	References

Introduction:

The worldwide spread of COVID-19 has brought about substantial changes in lifestyles globally, resulting in widespread illness, disruptions to daily routines, and increased pressure on healthcare systems. Understanding the spread and impact of the virus is important for developing effective public health strategies and mitigating its effects. This project focuses on analyzing data provided by the Centers for Disease Control and Prevention (CDC) using certain tools on Google Cloud Platform to gain insights into various aspects of the COVID-19 pandemic in the United States. By examining trends in case numbers, hospitalizations and deaths, the aim is to:

Identify disparities in COVID-19 outcomes: Analyze how factors such as age, race, and ethnicity correlate with hospitalization rates, ICU admissions, and mortality. This will help uncover potential disparities in healthcare access and outcomes

Understand the factors that influenced disease severity and mortality: This involves analyzing data on comorbidities, age groups and other risk factors associated with severe COVID-19 cases and deaths.

Analyze trends over time: Understand how case numbers, hospitalizations, and deaths have changed throughout the pandemic across different demographics.

Understand the impact of healthcare utilization: By analyzing hospitalization and ICU admission data, we can assess the strain on healthcare systems and identify potential areas for improvement

This analysis provides valuable information for policymakers, healthcare professionals, and the public to better understand the dynamics of COVID-19 and make informed decisions regarding public health strategies in case another pandemic plagues us in the future. This knowledge can also inform the development of more equitable and effective public health strategies.

Background:

The emergence of the novel coronavirus, SARS-CoV-2, in late 2019 triggered a global health crisis unlike any we have witnessed. The disease, COVID-19, rapidly spread across the globe, leading the WHO declaring a pandemic around March of 2020. The virus has had a great impact on nearly every aspect of society, causing widespread illness and death, disrupting economies, and placing immense strain on healthcare systems.

In the United States, the Centers for Disease Control and Prevention (CDC) has played a critical role in monitoring and responding to the pandemic. The CDC has been responsible for tracking the spread of the virus, issuing public health guidance, and coordinating with state and local

health departments. The CDC has also been collecting vast amounts of data related to COVID-19, including case numbers, hospitalizations and deaths.

The volume of information generated from case reporting, hospital admissions and death reporting has created a vast and complex data landscape. This presents both an opportunity and a challenge. While the data holds the potential to unlock key insights into the virus and its impact, traditional methods might fall short on extracting meaningful information. Leveraging cloud computing can significantly enhance our ability to extract valuable insights from vast amounts of data. Cloud computing offers scalability, advanced analytics and is cost effective, which can prove valuable to governments and healthcare institutions to manage, analyze and take decisions on public health policies in the future in case we have to face another pandemic.

Methodology:

The project takes place in the sequence as below:

Setup and Data Collection

- Downloaded the relevant COVID-19 datasets from the CDC website
- Set up a Google Cloud Platform (GCP) account and enabled necessary APIs for Google Cloud Storage (GCS), BigQuery, and Looker.
- Utilized Google Colab later, a cloud-based Jupyter notebook environment, for data cleaning and preprocessing.

Uploading data on Cloud Storage

- The downloaded dataset is uploaded to a GCS bucket. GCS provides a scalable and durable storage solution for our data.
- Organized the data within the bucket using folders and appropriate naming conventions for easy access and management.

Data Preprocessing

- I've used Python libraries like Pandas to clean the data, handle missing values, and address any inconsistencies or errors.
- Transformed the data into a format suitable for analysis. This involved creating new variables, aggregating data and encoding categorical variables using label encoding.

BigQuery: Analysis and Insights

- The preprocessed data will be loaded into BigQuery, a serverless data warehouse on GCP.
- Utilize BigQuery's powerful querying capabilities to perform in-depth analysis of the data. This includes aggregations, joins, and models to understand specific aspects.

Looker: Visualization

- Connected Looker, a data visualization platform, to BigQuery to create interactive charts and reports.
- Designed visualizations that effectively communicate the insights acquired from the data analysis. This includes charts, graphs and other visual representations.

The project takes inspiration from the module on data lifecycle and pipelines and tries to incorporate the USGS model of data lifecycle. This module provided me with a framework which I used to build my project. I have tried to replicate the steps present in a data lifecycle using a data pipeline built with Cloud Storage, Colab, BigQuery and Looker.

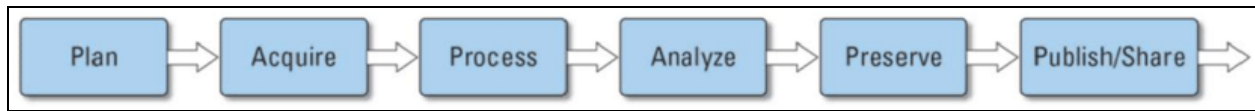


Fig1: USGS Model

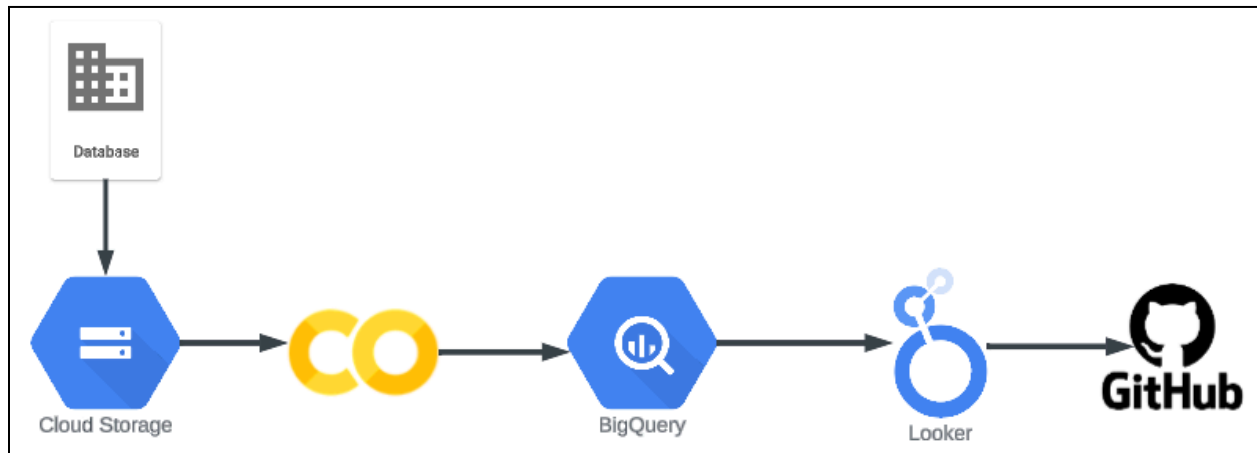


Fig2: System Architecture

Plan - Planning the architecture, the data source and other details about the project

Acquire - Gather data from the CDC website and store it in Cloud storage

Process - Cleaning and manipulating data in Colab

Analyze - Analyze the cleaned data in BigQuery and Looker

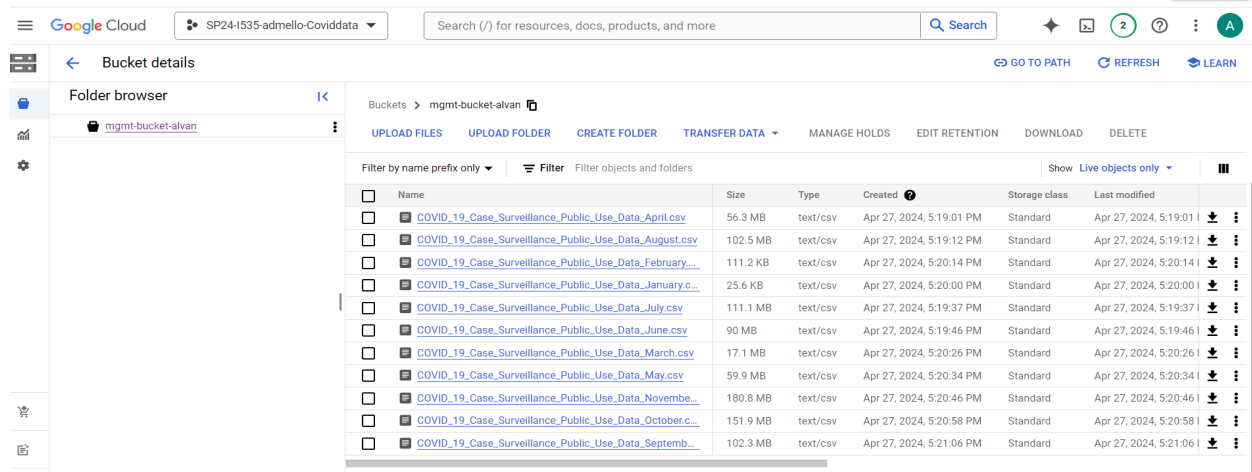
Preserve - The data is stored in BigQuery where it can be used again for analysis

Publish - The report along with the code and data is published on [Github](#)

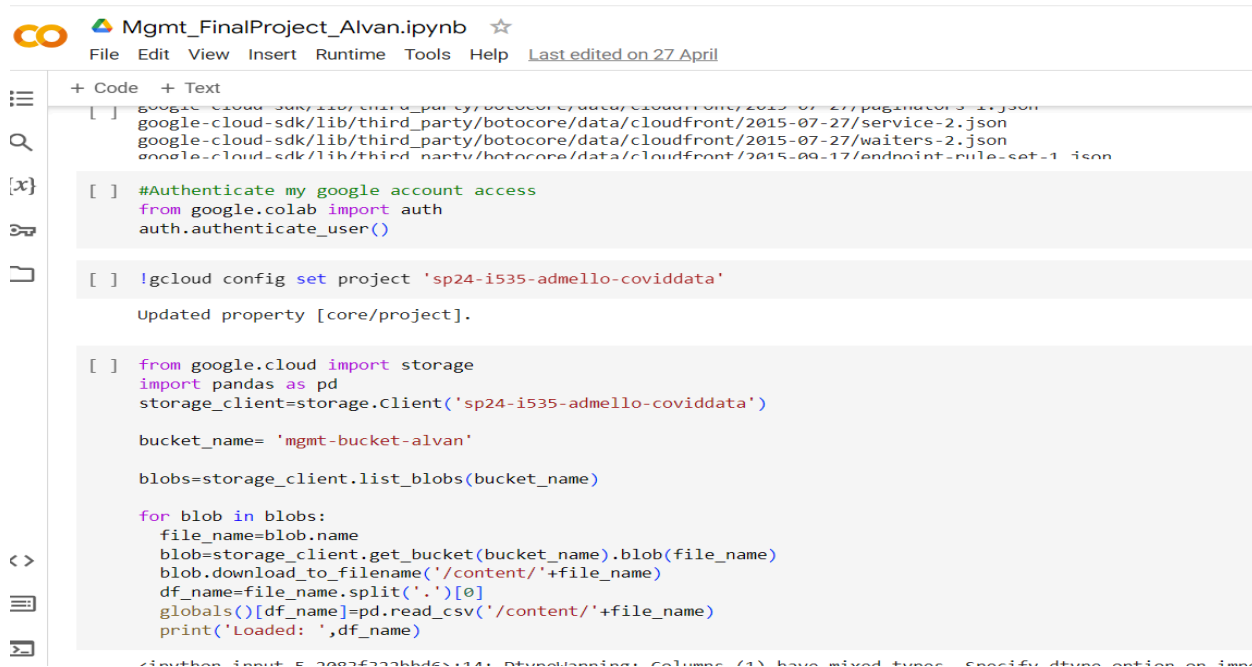
Steps Performed:

1. I downloaded the dataset from the [Center for Disease Control and Prevention \(CDC\)](#) website. This dataset encapsulates the year of 2020 when the United States encountered the inception of the Covid-19 outbreak. Spanning a size of around **914MB**, it encompasses **8 million rows** and **11 columns** in its raw form

2. Subsequently, after collecting the data, I uploaded it onto a Cloud Storage Bucket within GCP.



3. Using a series of commands, I established authentication within the notebook to establish connectivity with GCP, facilitating the ingestion of data into the Colab environment for preprocessing.



4. I combined all the individual files into a singular master dataset. Following this, a thorough examination of the dataset's composition was conducted through a review of its metadata. This helped me understand the data types and other details about various features present in the data.

```
[ ] Covid = pd.concat([COVID_19_Case_Surveillance_Public_Use_Data_January, COVID_19_Case_Surveillance_Public_Use_Data_February,
                      COVID_19_Case_Surveillance_Public_Use_Data_March, COVID_19_Case_Surveillance_Public_Use_Data_April,
                      COVID_19_Case_Surveillance_Public_Use_Data_May, COVID_19_Case_Surveillance_Public_Use_Data_June,
                      COVID_19_Case_Surveillance_Public_Use_Data_July, COVID_19_Case_Surveillance_Public_Use_Data_August,
                      COVID_19_Case_Surveillance_Public_Use_Data_September, COVID_19_Case_Surveillance_Public_Use_Data_October,
                      COVID_19_Case_Surveillance_Public_Use_Data_November], ignore_index=True)
```

```
[ ] Covid.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 8405079 entries, 0 to 8405078
Data columns (total 11 columns):
#   Column                                Dtype
---  -
0   cdc_report_dt                         object
1   pos_spec_dt                          object
2   onset_dt                             object
3   current_status                       object
4   sex                                  object
5   age_group                            object
6   Race and ethnicity (combined)        object
7   hosp_yn                              object
8   icu_yn                               object
9   death_yn                             object
10  medcond_yn                           object
dtypes: object(11)
memory usage: 705.4+ MB
```

5. Recognizing the presence of substantial null values within two specific columns of `post_spec_dt` and `onset_dt`, I dropped these columns since they would significantly skew the analysis.

```
[ ] round(100*(Covid.isnull().sum()/len(Covid.index)), 10)
```

```
cdc_report_dt          0.000000
pos_spec_dt            65.844592
onset_dt               47.698802
current_status         0.000000
sex                    0.000214
age_group              0.001059
Race and ethnicity (combined) 0.000083
hosp_yn                0.000000
icu_yn                 0.000000
death_yn               0.000000
medcond_yn             0.000000
dtype: float64
```



```
▶ Covid = Covid.drop(['pos_spec_dt','onset_dt'],axis=1)
Covid.info()
```

```
↳ <class 'pandas.core.frame.DataFrame'>
RangeIndex: 8405079 entries, 0 to 8405078
Data columns (total 9 columns):
#   Column                                Dtype
---  -
0   cdc_report_dt                        object
1   current_status                      object
2   sex                                 object
3   age_group                           object
4   Race and ethnicity (combined)      object
5   hosp_yn                             object
6   icu_yn                              object
7   death_yn                           object
```

6. Subsequent to this, any residual rows containing null values across the remaining columns were removed. Given the dataset's voluminous nature, this selective removal of rows would not have a significant impact on the overall analysis.

```
[ ] Covid.dropna(subset=['age_group', 'sex','Race and ethnicity (combined)'], inplace=True)
```

```
▶ Covid.isnull().sum()
```

```
↳ cdc_report_dt                0
   current_status              0
   sex                         0
   age_group                   0
   Race and ethnicity (combined) 0
   hosp_yn                     0
   icu_yn                      0
   death_yn                    0
   medcond_yn                  0
   dtype: int64
```

7. For further refinement, certain string-formatted rows were converted into integer format, thereby enhancing their utility for subsequent analysis. Below is an example of this transformation on one of the columns.

```
[ ] def convert_to_numeric(value):
    if value == 'Yes':
        return 1
    elif value == 'No':
        return 0
    else:
        return 0 # For 'unknown', return 0

# Apply the function to the column
Covid['hosp_yn'] = Covid['hosp_yn'].apply(convert_to_numeric)
```

8. After completion of all the preprocessing steps, the refined dataset was ingested into BigQuery. I created a dedicated dataset within BigQuery where this data could be stored. This serves as a data warehouse where data can be stored and used for analysis as and when required.

```
▶ from google.cloud import bigquery
  from google.oauth2 import service_account

project_id = "sp24-i535-admello-coviddata"
dataset_id = "mgmt_dataset_alvan"
table = "covid_final_data"
table_id = "{}.{}.{}".format(project_id, dataset_id, table)

def load_table_dataframe(project_id, table_id):
    # Automatically uses default credentials
    client = bigquery.Client(project=project_id)

    job_config = bigquery.LoadJobConfig(write_disposition="WRITE_TRUNCATE")

    job = client.load_table_from_dataframe(
        Covid, table_id, job_config=job_config
    )
    job.result()

    data = client.get_table(table_id)
    return data

data = load_table_dataframe(project_id, table_id)
```

Mgmt Access and Use of Big Data

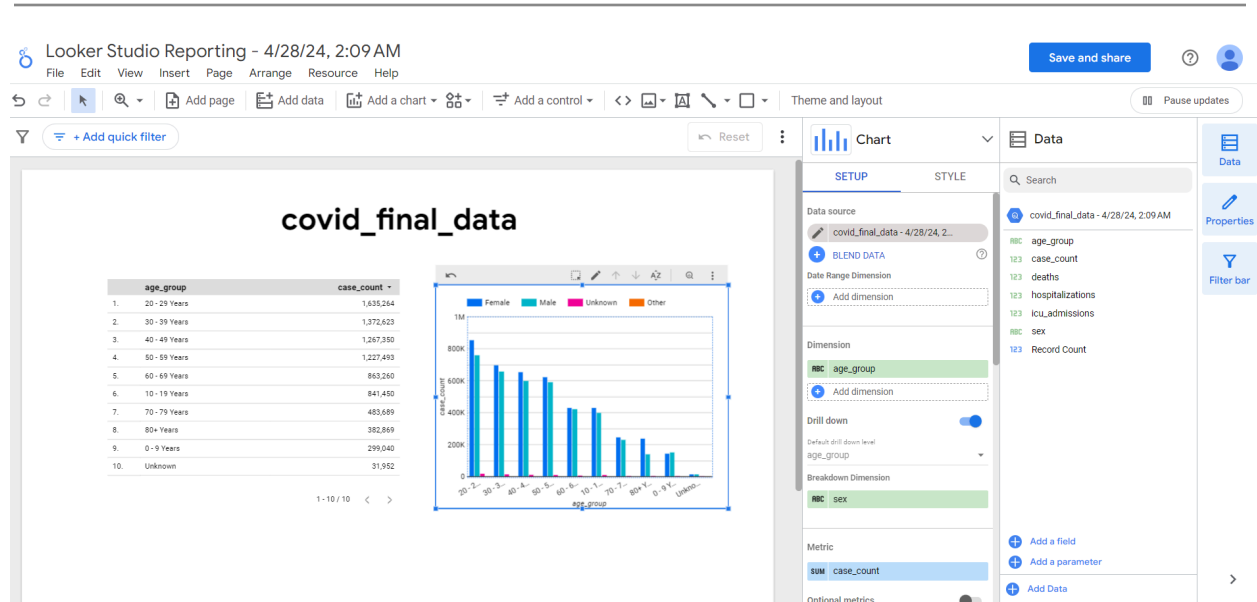
Row	cdc_report_dt	current_status	sex	age_group	Race_and_ethnicity_combined	hosp_yn
1	2020-01-22	Laboratory-confirmed case	Female	10 - 19 Years	Hispanic/Latino	0
2	2020-01-21	Laboratory-confirmed case	Female	20 - 29 Years	White, Non-Hispanic	0
3	2020-01-14	Laboratory-confirmed case	Female	20 - 29 Years	White, Non-Hispanic	0
4	2020-01-21	Laboratory-confirmed case	Female	20 - 29 Years	White, Non-Hispanic	0
5	2020-01-23	Laboratory-confirmed case	Female	20 - 29 Years	Asian, Non-Hispanic	0
6	2020-01-14	Laboratory-confirmed case	Female	20 - 29 Years	American Indian/Alaska Native...	0
7	2020-01-21	Laboratory-confirmed case	Female	20 - 29 Years	American Indian/Alaska Native...	0
8	2020-01-21	Laboratory-confirmed case	Male	20 - 29 Years	American Indian/Alaska Native...	0
9	2020-01-01	Probable Case	Female	20 - 29 Years	Multiple/Other, Non-Hispanic	0
10	2020-01-20	Laboratory-confirmed case	Female	20 - 29 Years	Multiple/Other, Non-Hispanic	0
11	2020-01-14	Laboratory-confirmed case	Female	20 - 29 Years	Unknown	0
12	2020-01-14	Laboratory-confirmed case	Female	20 - 29 Years	Unknown	0
13	2020-01-24	Laboratory-confirmed case	Female	20 - 29 Years	Unknown	1
14	2020-01-24	Laboratory-confirmed case	Female	20 - 29 Years	Unknown	0

9. Once the dataset was safely stored in BigQuery, I carefully ran a series of queries to dig out key insights. These queries helped me understand how the pandemic affected various demographics in various ways.

Row	age_group	sex	case_count	hospitalizations	icu_admissions	deaths
1	0 - 9 Years	Female	143494	2015	154	32
2	0 - 9 Years	Male	151635	2411	218	39
3	0 - 9 Years	Other	7	0	0	0
4	0 - 9 Years	Unknown	3904	22	0	0

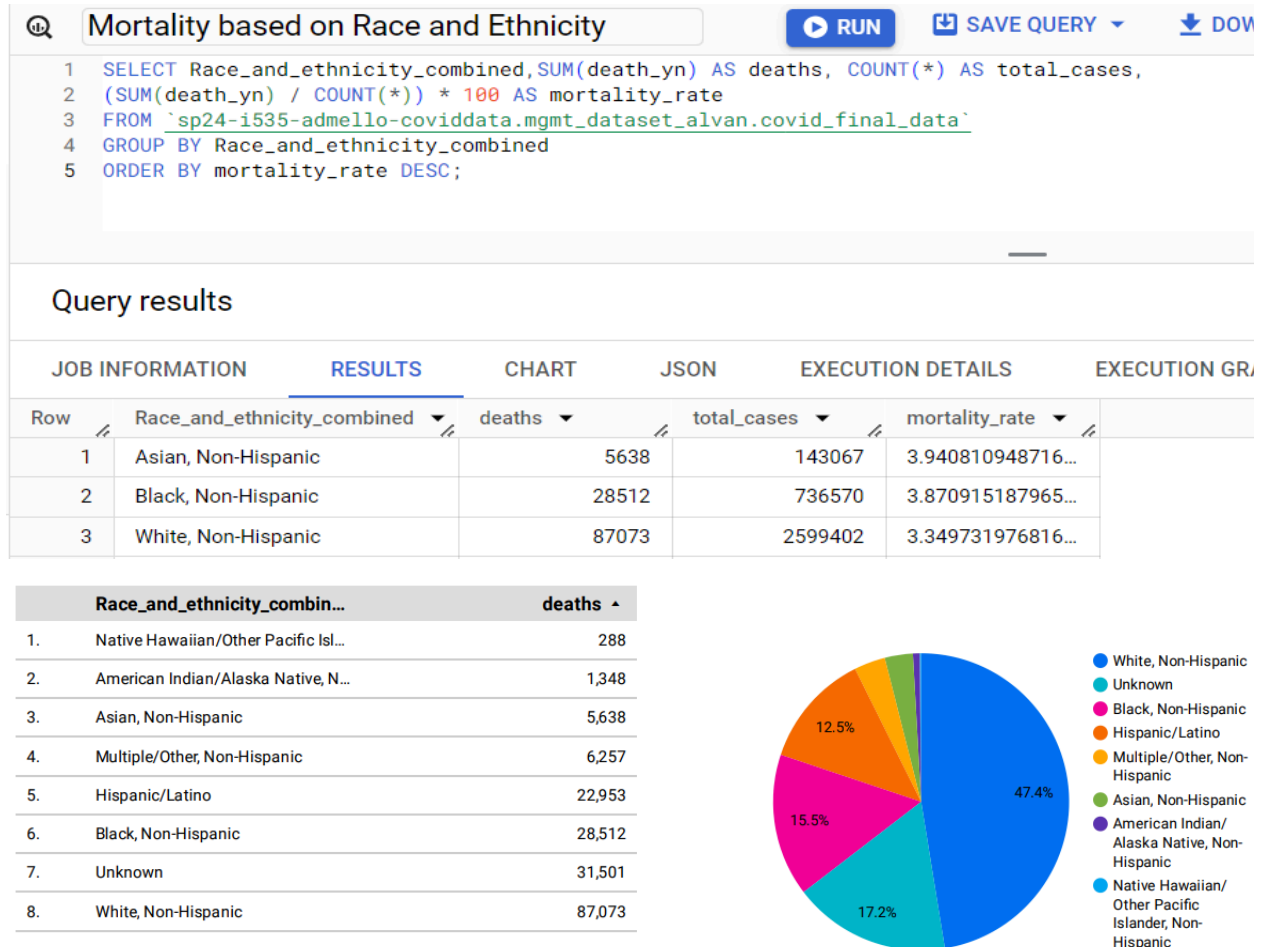
10. Finally, the extracted insights were visually presented using Looker. This visual depiction helps establish a deeper comprehension of underlying trends within the dataset, thereby fortifying subsequent analytical interpretations.

Mgmt Access and Use of Big Data



Results:

Query 1: This query examines potential disparities in mortality rates among different racial and ethnic groups.



From the plot we can observe that, whites have the highest number of covid cases followed by unknown and then blacks. Although the white population comprises 60% of the population, the total portion of deaths they have is 47%. Whereas the blacks comprise only about 12% of the nation's population. This might be due to the fact that the african-american community disproportionately faces poverty and its associated challenges, such as crowded housing, food insecurity, and limited access to healthcare due to historic racism. Albeit further investigation is crucial to understand the complex interplay of these factors and determine the specific causes. Similarly, the presence of a significant "Unknown" category for deaths raises issues about data collection and reporting accuracy. This might be possible since the pandemic happened unexpectedly and governments across the world were not prepared to handle a pandemic.

Query 2: This query shows the distribution of cases, hospitalizations, ICU admissions, and deaths across different age groups and sexes, highlighting potential disparities.

Case Distribution RUN SAVE QUERY DOWNLOAD SHARE SCHEDULE MORE This

```

1 SELECT age_group,sex,COUNT(*) AS case_count,SUM(hosp_yn) AS hospitalizations,SUM(icu_yn) AS icu_admissions,
2 SUM(death_yn) AS deaths
3 FROM `sp24-1535-admello-coviddata.mgmt_dataset_alvan.covid_final_data`
4 GROUP BY age_group,sex
5 ORDER BY age_group,sex;

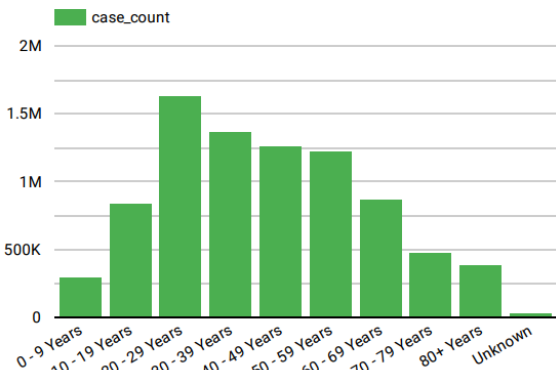
```

Query results

[SAVE RESULTS](#)

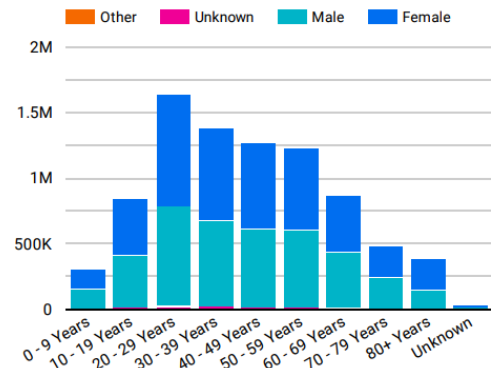
JOB INFORMATION		RESULTS	CHART	JSON	EXECUTION DETAILS		EXECUTION GRAPH	
Row	age_group	sex	case_count	hospitalizations	icu_admissions	deaths		
1	0 - 9 Years	Female	143494	2015	154	32		
2	0 - 9 Years	Male	151635	2411	218	39		
3	0 - 9 Years	Other	7	0	0	0		
4	0 - 9 Years	Unknown	3904	22	0	0		
5	10 - 19 Years	Female	432858	3993	259	58		
6	10 - 19 Years	Male	398220	3235	317	76		
7	10 - 19 Years	Other	18	0	0	0		
8	10 - 19 Years	Unknown	10354	15	2	1		
9	20 - 29 Years	Female	854734	16540	871	295		

	age_group	case_count
1.	0 - 9 Years	299,040
2.	10 - 19 Years	841,450
3.	20 - 29 Years	1,635,264
4.	30 - 39 Years	1,372,623
5.	40 - 49 Years	1,267,350
6.	50 - 59 Years	1,227,493
7.	60 - 69 Years	863,260
8.	70 - 79 Years	483,689
9.	80+ Years	382,869
10.	Unknown	31,952



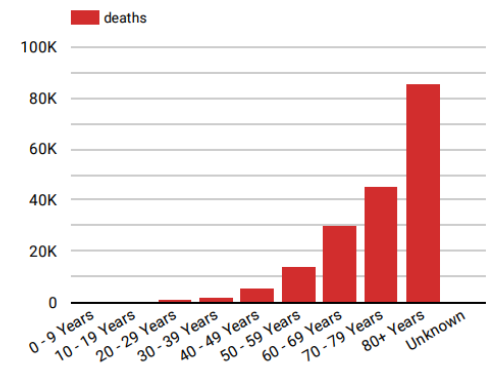
From the above plot, we can observe that the age group of 20-29 years has the highest number of COVID cases. This might be due to the fact that individuals in this age group tend to be more socially active, engaging in activities that involve close contact with others, such as attending social gatherings, going to bars and restaurants, or participating in group activities. Younger individuals are more likely to experience asymptomatic or mild COVID-19 infections, which can lead to unknowingly spreading the virus to others. Also, young adults may be more likely to get tested due to employment requirements, travel protocols, or participation in social activities that necessitate testing.

	age_group ^	case_count
1.	0 - 9 Years	299,040
2.	10 - 19 Years	841,450
3.	20 - 29 Years	1,635,264
4.	30 - 39 Years	1,372,623
5.	40 - 49 Years	1,267,350
6.	50 - 59 Years	1,227,493
7.	60 - 69 Years	863,260
8.	70 - 79 Years	483,689
9.	80+ Years	382,869
10.	Unknown	31,952



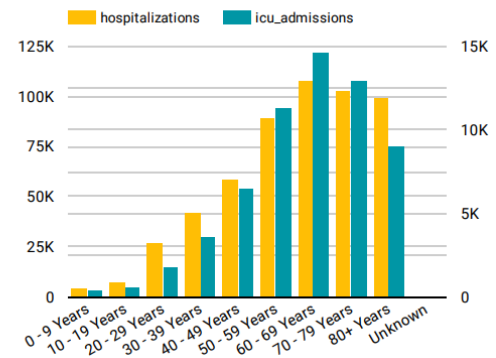
From the above plot, we can observe that the disease has had an almost similar impact on males and females across all age groups. This might be due to the fact that our society has witnessed a convergence of gender roles, with both men and women engaging in a wider range of activities and occupations, leading to similar exposure risks in many contexts.

	age_group ^	case_count
1.	0 - 9 Years	299,040
2.	10 - 19 Years	841,450
3.	20 - 29 Years	1,635,264
4.	30 - 39 Years	1,372,623
5.	40 - 49 Years	1,267,350
6.	50 - 59 Years	1,227,493
7.	60 - 69 Years	863,260
8.	70 - 79 Years	483,689
9.	80+ Years	382,869
10.	Unknown	31,952



From the above plot, we can observe that the age group of 80+ years has the highest number of COVID deaths. This might be due to the fact that their immune systems naturally weaken due to their age, making them more susceptible to infections and less able to fight off viruses effectively. Older adults are more likely to have pre-existing health conditions such as heart disease, diabetes, respiratory illnesses, and weakened lung function, which can significantly increase the risk of severe complications and death from COVID-19. Another rare possibility might be that a significant proportion of individuals in the 80+ age group reside in nursing homes or long-term care facilities. These places can unfortunately facilitate rapid transmission of the virus due to close living quarters and shared care among residents.

	age_group ^	case_count
1.	0 - 9 Years	299,040
2.	10 - 19 Years	841,450
3.	20 - 29 Years	1,635,264
4.	30 - 39 Years	1,372,623
5.	40 - 49 Years	1,267,350
6.	50 - 59 Years	1,227,493
7.	60 - 69 Years	863,260
8.	70 - 79 Years	483,689
9.	80+ Years	382,869
10.	Unknown	31,952



From the plot above, we can observe that individuals in the age group of 60-69 years have the highest number of hospitalizations and ICU admissions. While generally healthier than older age groups, individuals in their 60s may still experience age-related decline in immune function and increased prevalence of health conditions. These conditions can significantly increase the risk of severe COVID-19 illness requiring hospitalization. People in their 60s may also be more proactive in seeking medical attention due to their age, leading to higher rates of hospitalization and ICU admissions.

Query 3: This query provides a monthly breakdown of hospitalizations, ICU admissions, and deaths, allowing us to observe trends over time.

Trend of Cases Over Time RUN SAVE QUERY DOWNLOAD SHARE SCHE

```

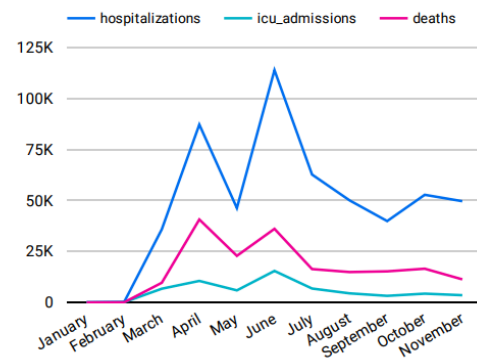
1 SELECT Year,Month,COUNT(*) AS total_cases,SUM(hosp_yn) AS hospitalizations,SUM(icu_yn) AS icu_admissions,
2     SUM(death_yn) AS deaths
3 FROM `sp24-i535-admello-coviddata.mgmt_dataset_alvan.covid_final_data`
4 GROUP BY Year, Month
5 ORDER BY Year, Month;
6

```

Query results

JOB INFORMATION		RESULTS	CHART	JSON	EXECUTION DETAILS		EXECUTION GRAPH	
Row	Year	Month	total_cases	hospitalizations	icu_admissions	deaths		
1	2020	1	232	18	6	1		
2	2020	2	1037	238	82	51		
3	2020	3	161476	35830	6700	9630		
4	2020	4	538947	87257	10535	40679		
5	2020	5	582192	46296	5895	22839		
6	2020	6	880120	113963	15434	36095		
7	2020	7	1079629	62739	6797	16343		
8	2020	8	991092	50060	4430	14884		

total_cases		Month
1.	232	January
2.	1037	February
3.	161476	March
4.	538947	April
5.	582192	May
6.	880120	June
7.	1079629	July
8.	991092	August
9.	989488	September
10.	1460003	October
11.	1720774	November



From the graph above we can observe that the number of cases, hospitalizations and ICU admissions spike in April, decline in May, and experience a massive surge around July before steadily declining. The initial spike in April and the subsequent surge around July could be attributed to the introduction and spread of a new variant with increased virulence or ability to cause severe illness, leading to higher rates of hospitalization, ICU admission, and death. Also, the surge in cases and severe outcomes in the first wave could have strained healthcare system resources, potentially impacting the quality of care and leading to increased hospitalization and ICU admissions in the second wave.

Query 4: This chart compares the monthly trends of laboratory-confirmed, hospitalizations and total COVID-19 cases

Monthly Fluctuations in cases and Lab tests

```

1 SELECT Month AS month,
2       SUM(CASE WHEN current_status = 'Laboratory-confirmed case' THEN 1 ELSE 0 END) AS lab_confirmed_cases,
3       SUM(hosp_yn) AS hospitalizations, COUNT(*) AS total_cases
4 FROM `sp24-1535-admello-coviddata.mgmt_dataset_alvan.covid_final_data`
5 GROUP BY month
6 ORDER BY month;

```

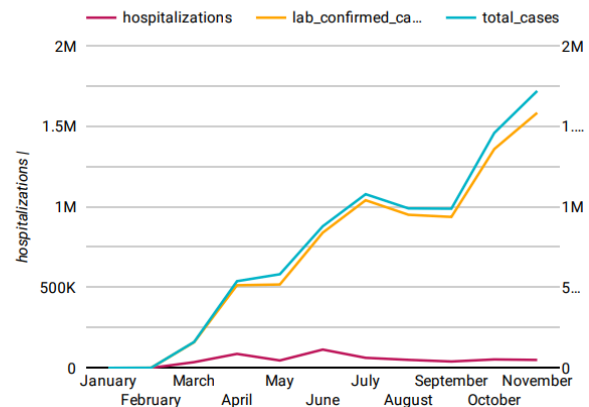
Query results

[SAVE RE](#)

JOB INFORMATION	RESULTS	CHART	JSON	EXECUTION DETAILS	EXECUTION GRAPH
Row	month	lab_confirmed_cases	hospitalizations	total_cases	
1	1	206	18	232	
2	2	966	238	1037	
3	3	159563	35830	161476	
4	4	513722	87257	538947	
5	5	518406	46296	582192	
6	6	840457	113963	880120	
7	7	1042076	62739	1079629	
8	8	951547	50060	991092	

Results per page: 50 ▼

lab_confirmed_cases	month
1. 1585030	November
2. 1359831	October
3. 938233	September
4. 951547	August
5. 1042076	July
6. 840457	June
7. 518406	May
8. 513722	April
9. 159563	March
10. 966	February
11. 206	January



From the graph above we can observe that the number of cases steadily keeps on increasing over the months. It also has slightly flat lines around the months of May and August. This might be indicative of easing of public health measures such as mask mandates or social distancing requirements following the two waves in April and July observed in the previous graph. This could facilitate transmission and contribute to rising case numbers. Another possibility might be the improvements in testing capacity and availability over time which could lead to more cases being detected and reported, contributing to the increase.

Query 5: This chart compares the rate of hospitalization based on race and ethnicity

Hospitalization by Race and Ethnicity
▶ RUN
📄 SAVE QUERY ▾

```

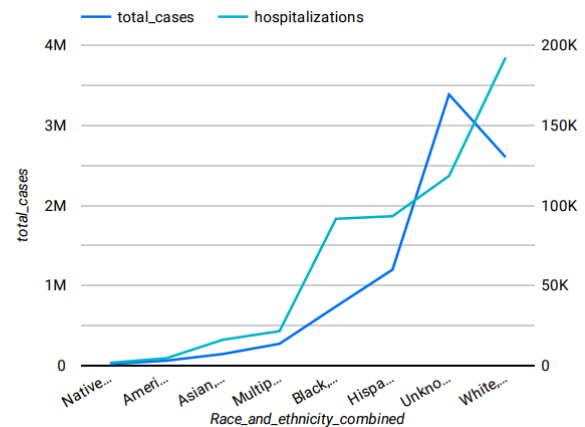
1 SELECT
2   Race_and_ethnicity_combined,
3   COUNT(*) AS total_cases,
4   SUM(hosp_yn) AS hospitalizations
5 FROM `sp24-i535-admello-coviddata.mgmt_dataset_alvan.covid_final_data`
6 WHERE Race_and_ethnicity_combined IS NOT NULL
7 GROUP BY Race_and_ethnicity_combined
8 ORDER BY total_cases DESC;

```

Query results

JOB INFORMATION	RESULTS	CHART	JSON	EXECUTION DETAILS	EXEC
Row	Race_and_ethnicity_combined ▾	total_cases ▾	hospitalizations ▾		
1	Unknown	3382684	118300		
2	White, Non-Hispanic	2599402	192167		
3	Hispanic/Latino	1195726	93148		
4	Black, Non-Hispanic	736570	91534		
5	Multiple/Other, Non-Hispanic	270503	21388		
6	Asian, Non-Hispanic	143067	16066		

	Race_and_ethnicity_combin...	total_cases ▾
1.	Unknown	3,382,684
2.	White, Non-Hispanic	2,599,402
3.	Hispanic/Latino	1,195,726
4.	Black, Non-Hispanic	736,570
5.	Multiple/Other, Non-Hispanic	270,503
6.	Asian, Non-Hispanic	143,067
7.	American Indian/Alaska Native, N...	59,842
8.	Native Hawaiian/Other Pacific Isl...	17,196



From the graph above, we can observe that a higher proportion of individuals in the unknown category are not hospitalized as compared to other ethnicities. There is a possibility that individuals experiencing homelessness could be included within the "unknown" category when collecting data on race/ethnicity. Data collection often relies on self-reporting of race and ethnicity and they might choose not to disclose their ethnicities. Homeless individuals often lack health insurance and face financial barriers to accessing healthcare services, including hospitalization. Negative past experiences or fear of discrimination might lead to distrust of the healthcare system among homeless individuals, discouraging them from seeking hospital care.

Query 6: This chart compares the monthly trend of hospitalization among different age groups

Monthly Hospitalization trend
▶ RUN
📄 SAVE QUERY ▾
⬇️ DOWNLO

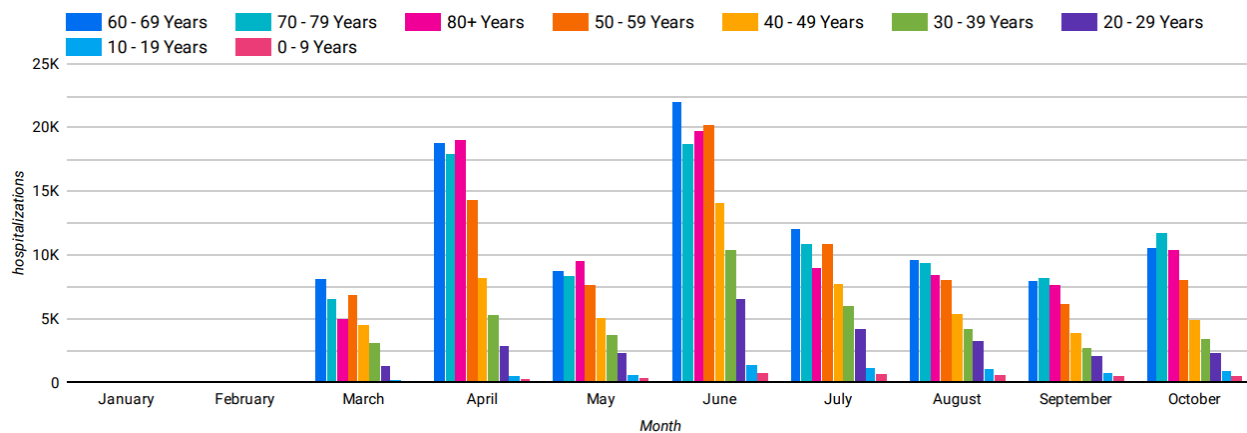
```

1 SELECT Month,age_group,COUNT(*) AS case_count,SUM(hosp_yn) AS hospitalizations,
2 FROM `sp24-i535-admello-coviddata.mgmt_dataset_alvan.covid_final_data`
3 WHERE age_group IS NOT NULL
4 GROUP BY Month, age_group
5 ORDER BY Month, age_group;

```

Query results

JOB INFORMATION		RESULTS	CHART	JSON	EXECUTION DETAILS	EXECUT
Row	Month ▾ ↑	age_group ▾	case_count ▾	hospitalizations ▾		
1	1	0 - 9 Years	6	0		
2	1	10 - 19 Years	16	0		
3	1	20 - 29 Years	41	1		
4	1	30 - 39 Years	39	1		
5	1	40 - 49 Years	34	3		



From the graph above, we can observe that the age group 60-69 has the highest hospitalization overall apart from certain pockets where age groups 70-79 and 80+ have higher hospitalization. This is possible due to the vaccine policy. The initial vaccine rollout likely prioritized older age groups (80+ and 70-79) due to their higher vulnerability. As vaccination coverage increased in these groups, their hospitalization rates might have declined compared to the 60-69 group, who might have been vaccinated later. The temporary shifts observed in May and August where the older age groups were more affected might be due to the waves being at their peak.

Model Evaluation

In addition, I've developed a classification model using BigQuery, aiding in predicting hospitalization likelihood based on age, gender, and comorbidities. After training the model on our dataset, we can see that it has an accuracy of around 80%. We can tweak the model parameters to get the precision and recall we desire. This model gives us the likelihood of how prone a person recently diagnosed with covid is, to be hospitalized based on whether he or she has comorbidities.

Model_Evaluation							
<pre> 1 SELECT 2 * 3 FROM 4 ML.EVALUATE(MODEL `mgmt_dataset_alvan.hospitalization_model`, 5 (6 SELECT 7 age_group, 8 sex, 9 medcond_yn, 10 hosp_yn AS label 11 FROM 12 `sp24-i535-admello-coviddata.mgmt_dataset_alvan.covid_final_data` </pre>							
Query results							
<div> JOB INFORMATION RESULTS CHART JSON EXECUTION DETAILS EXECUTION GRAPH </div>							
Row	precision	recall	accuracy	f1_score	log_loss	roc_auc	
1	0.833333333333...	0.046728971962...	0.794	0.088495575221...	0.519831594196...	0.716886113886...	

Prediction:

Here we can see that a male in the age group of 45-50 having no comorbidities would most likely not require hospitalization based on the probability prediction of the model. This can serve as a tool for healthcare professionals to decide the course of treatment and the next steps with regards to the patient as soon as they are diagnosed with covid and thereby save time.

Prediction							
<pre> 1 SELECT 2 * 3 FROM 4 ML.PREDICT(MODEL `mgmt_dataset_alvan.hospitalization_model`, 5 (6 SELECT 7 '45-54' AS age_group, 8 'male' AS sex, 9 CAST(0.0 AS INT64) AS medcond_yn -- Cast FLOAT64 to INT64 10) 11) 12); </pre>							
Query results							
<div> JOB INFORMATION RESULTS CHART JSON EXECUTION DETAILS EXECUTION GRAPH </div>							
Row	predicted_label	predict... label	predict... prob	age_group	sex	medcond_yn	
1	0	1	0.235886747491...	45-54	male	0	
		0	0.764113252508...				

Discussion:

Interpretation:

The COVID-19 pandemic presents a complex and evolving landscape with various factors influencing its impact across different demographics and time periods. While Whites initially showed the highest case numbers, potentially due to increased testing and reporting, the African-American community faced a disproportionate burden in terms of severe outcomes and mortality due to pre-existing healthcare disparities and socioeconomic challenges. Age played a significant role, with younger adults (20-29) experiencing higher case numbers due to increased social activity and potential asymptomatic spread, while older adults (80+) faced the highest risk of death due to weakened immune systems and underlying health conditions. Hospitalization rates were highest among the 60-69 age group, likely due to a combination of age-related health decline, active lifestyles, and greater access to healthcare.

Fluctuations in cases, hospitalizations, and deaths over time suggest the influence of factors such as seasonal changes, the emergence of new variants, and the implementation of public health interventions. The presence of a significant "unknown" category in data reporting highlights the challenges in capturing accurate information, particularly for marginalized groups like homeless individuals who may face barriers to healthcare access and data collection efforts. These observations underscore the need for targeted interventions for certain ethnic groups and establishing community medical centers where their populations are higher in numbers. It is also essential to develop culturally competent educational material and other outreach programs to increase awareness on what to do during a pandemic and how to keep yourself prepared. There also seems a need to implement standardized data collection protocols across healthcare systems and public health agencies to ensure consistency and accuracy in reporting so that there is no skew in analysis. Also it is necessary to push for policies that address various social determinants of health such as poverty and lack of access to facilities.

Technologies and Skills utilized:

Tools utilized are:

Google Cloud Storage - I utilized this for storing data. I also connected the Colab notebook to GCS to pull data from it.

Google Colaboratory - I used colab for cleaning and preprocessing data obtained from cloud storage. Colab provided me with the interface to bridge the gap in my pipeline and helped me clean the data and push it into BigQuery.

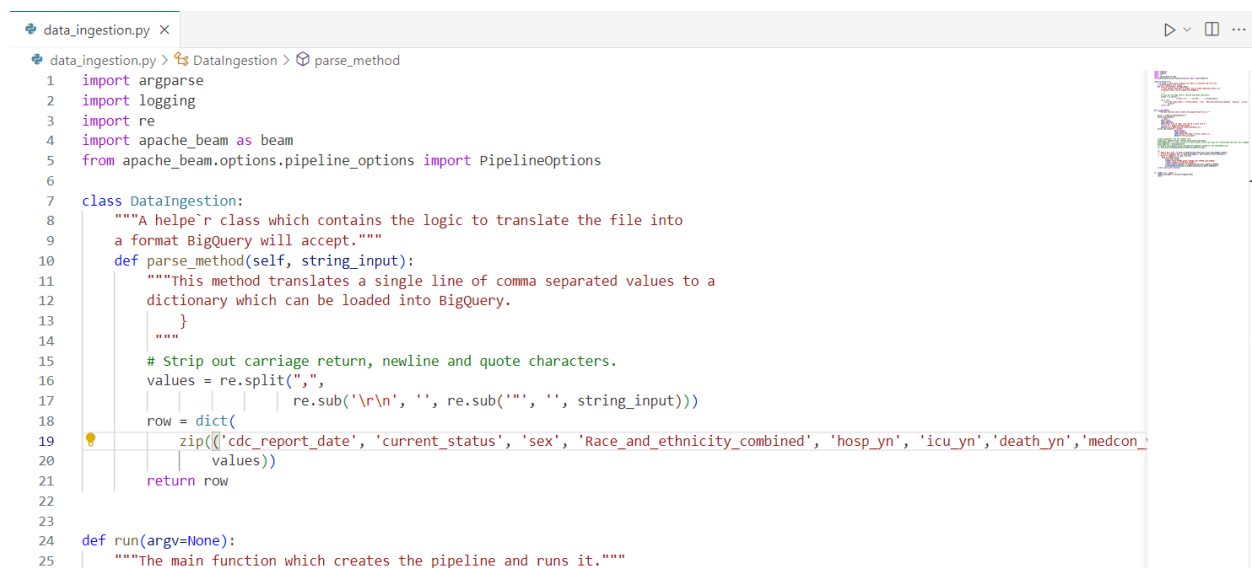
BigQuery - This was used to run queries which would retrieve data that can be used for analysis. I also built a classification model on BigQuery which

Looker - This was used for visualization of the query outputs. I could create charts which provided clear trends and helped with the analysis

The concepts and tools introduced in this course were extremely informative. Modules like Cloud computing introduced me to the fundamentals of GCP which helped me understand the various services offered by GCP and their features. Listening to different perspectives on the same topics through weekly discussions contributed to ideation and conception of the project and also helped improve my communication and writing skills. I could also improve on skills like data analysis and visualization through assignments in modules like Processing and Analytics.

Challenges and Failures:

I tried implementing the data preprocessing portion of the project in Dataflow using Apache Beam but was not successful in doing so. My aim was to implement a streaming data pipeline so that data could be ingested periodically and queries could be scheduled from time to time to give updated analysis on a fixed time basis. However, the code was complex and despite referring to multiple videos and tutorials, I could not implement it. I've attached a sample snippet below where I try to implement the data ingestion pipeline. That is why I switched my focus on using Google Colab which could very easily integrate with various Google Cloud services. Colab provided a reliable platform to clean and preprocess the data and effectively bridged the gap in the data pipeline. Apart from that, I did not face any hurdles during the project. Any problem that I encountered could be troubleshooted with a little bit of research.



```
data_ingestion.py X
data_ingestion.py > DataIngestion > parse_method
1 import argparse
2 import logging
3 import re
4 import apache_beam as beam
5 from apache_beam.options.pipeline_options import PipelineOptions
6
7 class DataIngestion:
8     """A helper class which contains the logic to translate the file into
9     a format BigQuery will accept."""
10    def parse_method(self, string_input):
11        """This method translates a single line of comma separated values to a
12        dictionary which can be loaded into BigQuery.
13        """
14        # Strip out carriage return, newline and quote characters.
15        values = re.split(",",
16                           re.sub('\r\n', '', re.sub("'", '', string_input)))
17        row = dict(
18            zip(['cdc_report_date', 'current_status', 'sex', 'Race_and_ethnicity_combined', 'hosp_yn', 'icu_yn', 'death_yn', 'medcon',
19                values))
20        return row
21
22
23
24 def run(argv=None):
25     """The main function which creates the pipeline and runs it."""
```

Conclusion:

This project underscores the vital role of data analysis in understanding the complexities of the COVID-19 pandemic and its impact on diverse populations. Google Cloud Platform (GCP) provides the essential infrastructure for this analysis, offering powerful tools for managing, processing, and visualizing large-scale datasets. GCP services like BigQuery and Cloud Storage enable efficient data handling, while also providing machine learning tools to support the development of predictive models for forecasting. Additionally, data visualization tools like Looker Data Studio facilitate clear communication of public health insights to relevant stakeholders.

Further scope can include applications such as outbreak tracking and vaccination management, promoting data interoperability between healthcare systems, and expanding access to cloud resources for public health organizations, particularly in under-resourced settings. By prioritizing ethical considerations and data privacy, GCP can continue to empower data-driven decision-making and strengthen public health initiatives, ultimately contributing to improved health outcomes for communities worldwide.

References:

https://cloud.google.com/bigquery/docs/create-machine-learning-model#create_your_dataset

<https://github.com/mponce/google-cloud-dataflow-pipeline>

<https://github.com/vigneshSs-07/Cloud-AI-Analytics/blob/main/etl-dataflow-GCP/commands.md>

https://github.com/GoogleCloudPlatform/professional-services/blob/main/examples/dataflow-pyth-non-examples/batch-examples/cookbook-examples/pipelines/data_enrichment.py

<https://www.cloudskillsboost.google/focuses/3460?parent=catalog>

<https://chat.openai.com/>

<https://console.cloud.google.com/vertex-ai/generative/multimodal/>

<https://www.kaggle.com/code/shahendayoussef/covid-19-case-surveillance#Classification-Models>

<https://www.kaggle.com/datasets/arashnic/covid19-case-surveillance-public-use-dataset/data>

https://data.cdc.gov/Case-Surveillance/COVID-19-Case-Surveillance-Public-Use-Data/vbim-akqf/about_data