Part 2: Final Project
Information Management

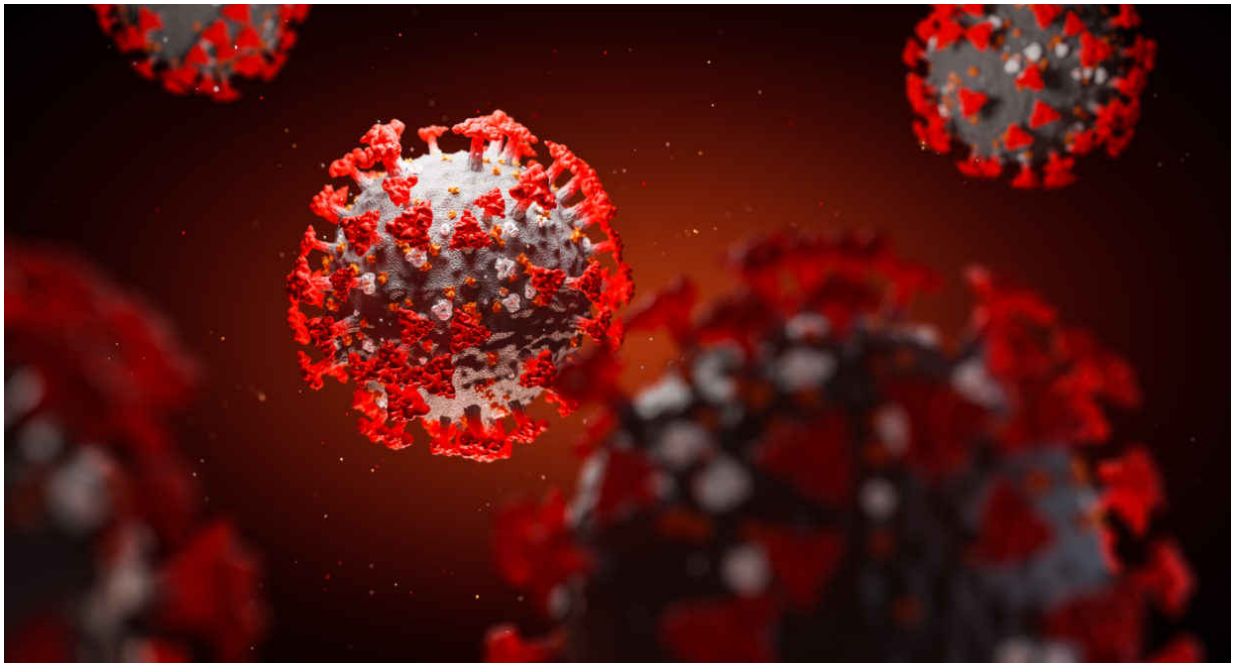
Group 2: Covid Analysis

Team Members:
David Gong
Sanyam Jain
Vaibhav Nagar
Vaishnavi Ganesh
Yash Jain

Project Narrative

Project Selection:

How did your group decide on the final project? What considerations influenced your choice?



Our group embarked on a project to visualize the global spread and treatment of COVID-19 from 2020 to 2023. We recognized the immense impact of COVID-19 worldwide and aimed to provide deep insights into the pandemic's spread while exploring various data management tools and techniques. Our focus was on visualizing trends, identifying hotspots, and analyzing the measures taken during the pandemic. The data encompassed dimensions like continents, countries, and dates, covering a range of measures such as new cases, total cases, deaths, vaccinations, and boosters.

In summary, this project was chosen for its relevance and the opportunity it presented to apply data management skills in a real-world context. Our project choice was driven by the combination of global significance, the opportunity for hands-on application of data tools, a focus on meaningful insights, comprehensive dataset dimensions, and the real-world relevance of the COVID-19 pandemic.

Data Acquisition:

Describe the process of obtaining the data for your project. Highlight any challenges faced?

Dataset

Our project obtained COVID-19 statistics from "Our World in Data." The data source comprises comprehensive COVID-19 statistics. It includes information such as the total number of cases from the beginning until the present date, the count of deaths related to COVID-19, and details on vaccinations.

However, the data acquisition process was not without challenges.

Data Challenges:

- Connection errors during data retrieval
- JSON file size limitations (16 MB) posed restrictions on data extraction
- Importing all columns into Tableau presented challenges

Licensing and Dashboard Issues:

- Licensing complications arose with MongoDB Atlas.
- Creating seamless dashboards in Tableau proved challenging.

Despite these obstacles, we effectively addressed them. Python was employed for data cleaning, and the refined data was successfully loaded into MongoDB for database creation. This database was then connected to Tableau for visualization, overcoming the challenges encountered during the process.

Data Sampling:

Provide a brief sampling of the raw data you worked with?

The raw data comprised over 350,000 rows and 67 columns, providing a comprehensive view of the pandemic. It included various parameters such as infection rates, location, mortality, and a time frame from 2020 to 2023. This rich dataset offered a wide array of information for analysis, from new and total cases to vaccination details.

```
import numpy as np
import pandas as pd
```

```
df= pd.read_csv('owid-covid-data.csv')
df.head(3)
```

```
C:\Users\vaish\AppData\Local\Temp\ipykernel_13540\929516247.py:1: DtypeWarning: Columns (33) have mixed types. Specify dtype option on import or set low_mem
  df= pd.read_csv('owid-covid-data.csv')
```

	iso_code	continent	location	date	total_cases	new_cases	new_cases_smoothed	total_deaths	new_deaths	new_deaths_smoothed	...	male_smokers	handwa
0	EST	Europe	Estonia	6/19/2022	564276.0	822.0	117.429	2462.0	2.0	0.286	...	39.3	
1	OWID_ASI	NaN	Asia	11/20/2023	NaN	NaN	NaN	NaN	NaN	NaN	...	NaN	
2	BGR	Europe	Bulgaria	11/20/2023	NaN	NaN	NaN	NaN	NaN	NaN	...	44.4	

3 rows × 67 columns

```
[ ] df.info()
```

```
 11  new_cases_per_million                  347550 non-null  float64
 12  new_cases_smoothed_per_million         346291 non-null  float64
 13  total_deaths_per_million               297655 non-null  float64
 14  new_deaths_per_million                 347624 non-null  float64
 15  new_deaths_smoothed_per_million        346394 non-null  float64
 16  reproduction_rate                      184817 non-null  float64
 17  icu_patients                           37793 non-null   float64
 18  icu_patients_per_million               37793 non-null   float64
 19  hosp_patients                          39165 non-null   float64
 20  hosp_patients_per_million              39165 non-null   float64
 21  weekly_icu_admissions                  10293 non-null   float64
 22  weekly_icu_admissions_per_million      10293 non-null   float64
 23  weekly_hosp_admissions                 23444 non-null   float64
```

Covid Dataset

Data Cleaning (ETL):

Share insights into how you cleaned the data. If ETL tools were used, specify and briefly explain?

We utilized Python for data cleaning, focusing on removing null values, fixing data granularity, and improving overall data quality. This process was critical in preparing the data for effective analysis and visualization. The cleaned data was then stored in MongoDB, a NoSQL database, ensuring quick and seamless extraction for further processing and visualization.

```
data_missing_value = df_clean.isnull().sum().reset_index()
data_missing_value.columns = ['feature','missing_value']
data_missing_value['percentage'] = round((data_missing_value['missing_value']/len(df_clean))*100,3)
data_missing_value = data_missing_value.sort_values('percentage', ascending=False).reset_index(drop=True)
data_missing_value = data_missing_value[data_missing_value['percentage']>0]
data_missing_value
```

	feature	missing_value	percentage
0	weekly_icu_admissions	346940	97.119
1	weekly_icu_admissions_per_million	346940	97.119
2	excess_mortality_cumulative_per_million	345022	96.582
3	excess_mortality	345022	96.582
4	excess_mortality_cumulative	345022	96.582
...	...	...	...
58	new_deaths_smoothed	10839	3.034
59	new_cases	9683	2.711
60	new_cases_per_million	9683	2.711
61	new_deaths	9609	2.690
62	new_deaths_per_million	9609	2.690

63 rows × 3 columns

Cleaning the data by taking care of nulls and missing values

```
[ ] nan_rows = df[df['continent'].isna()]
    if not nan_rows.empty:
        nan_locations = nan_rows['location'].unique()
        print("Values in 'location' column for NaN rows in 'continent':")
        print(nan_locations)
    else:
        print("No NaN values in 'continent' column.")
```

```
Values in 'location' column for NaN rows in 'continent':
['Lower middle income' 'Upper middle income' 'World' 'High income'
 'Low income']
```

```
[ ] show_list = ['Lower middle income', 'Upper middle income', 'World', 'High income', 'Low income']
    selected_rows = category_counts_df.loc[category_counts_df['index'].isin(show_list)]

    # Print the result
    print(selected_rows)
```

```
                  index  location
7    Lower middle income      1418
9    Upper middle income      1418
10                 World      1418
13           High income      1417
134           Low income      1414
```

```
df.loc[df['location'] == 'Lower middle income', 'continent'] = 'Unknown'
```

```
[ ] df.loc[df['location'] == 'Upper middle income', 'continent'] = 'Unknown'
```

```
[ ] df.loc[df['location'] == 'World', 'continent'] = 'Unknown'
```

Cleaning the data for specific columns like Location, Income, continents

Cleaned Data Sample:

Include a sampling of the cleaned data to showcase the quality and structure achieved?

The cleaned dataset represented a significant improvement in quality and structure over the raw data. Our efforts in cleaning the data using Python resulted in a dataset with enhanced granularity and categorization, although specific samples of the cleaned data are not provided in the document.

```
[ ] # Print the result
    print("Columns with null values:")
    print(columns_with_nulls)

    Columns with null values:
    ['population_density', 'median_age', 'aged_65_older', 'aged_70_older', 'gdp_per_capita', 'extreme_poverty', 'cardiovasc_death_rate', 'diabetes_prevalence',
```

```
[ ] columns_to_replace = ['population_density', 'median_age', 'aged_65_older', 'aged_70_older', 'gdp_per_capita', 'extreme_poverty', 'cardiovasc_death_rate', '
    df[columns_to_replace] = df[columns_to_replace].fillna(0)
```

```
[ ] columns_to_replace = ['handwashing_facilities']  # Replace with your column names
    df[columns_to_replace] = df[columns_to_replace].fillna(200)
```

```
[ ] # Identify columns with null values
    columns_with_nulls = df.columns[df.isnull().any()].tolist()

    # Print the result
    print("Columns with null values:")
    print(columns_with_nulls)

    Columns with null values:
    []
```

```
    # Delete those rows from the DataFrame
    df = df[df['continent'] != 'Unknown']
```

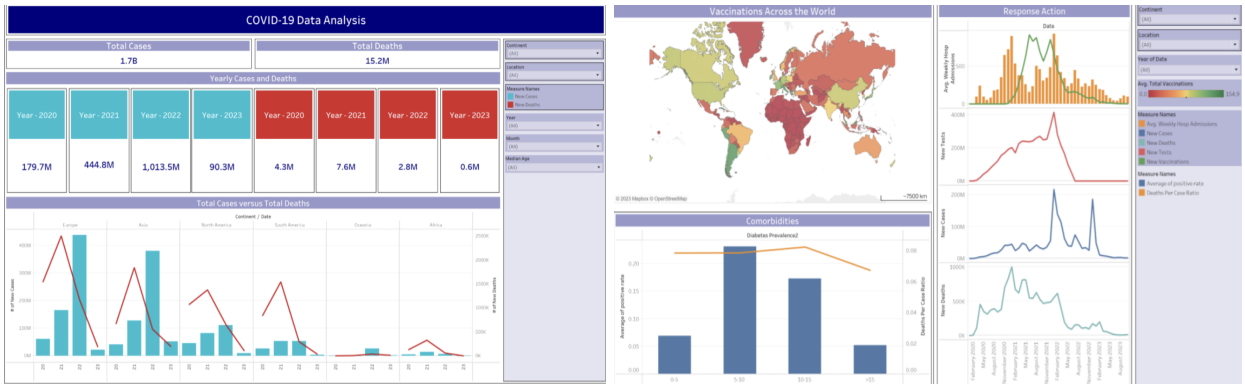
```
[ ] # Specify the file path where you want to save the CSV file
    csv_file_path = 'C:/Vaishnavi files/01 MSBA Files/04 Subjects/02 Fall/06 Information Management/Project/IM_data_cleaned1.csv'

    # Save the DataFrame as a CSV file
    df.to_csv(csv_file_path, index=False)
```

Research Questions:

Outline the types of questions your group aimed to answer through the project?

Our project sought to answer several key questions:

- How has the COVID-19 scenario evolved historically?
- What insights can be derived by filtering data by continent, country, date, and age group?
- How can these insights contribute to informed decision-making, especially in addressing the dynamic nature of the COVID-19 pandemic?
- These questions guided our analysis, allowing us to explore the pandemic from various perspectives.



Above dashboards that helped us answer these questions

Tools and Technology:

Detail the tools and technologies used during the project. Mention any specific requirements?

We employed a combination of tools for this project:

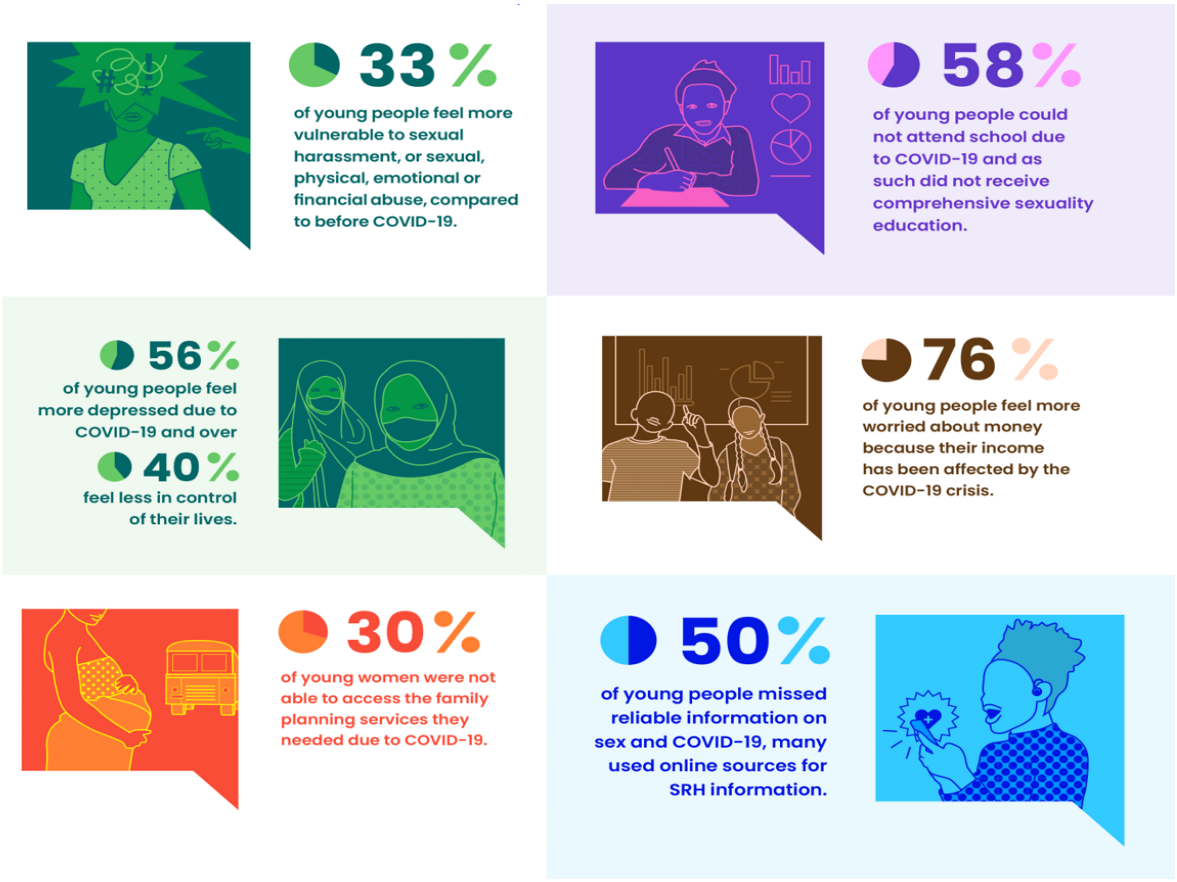
- Python: Used for data exploration and cleaning.
- MongoDB: Served as the database for storing the cleaned data.
- Tableau: Used for data visualization, helping us extract meaningful insights from the data.

Real-Life Impact:

Explain how the outcomes of your work could be applied to address real-life problems or challenges

The real-life impact of our project was multifaceted:

- Historical Perspective: We gained an understanding of the historical evolution of the COVID-19 pandemic.
- Granular Insights: By focusing on specific data points like continent, country, date, and age group, we could derive detailed insights.
- Informed Decision-Making: The project facilitated key insights to address the dynamic nature of COVID-19, contributing to more informed public health strategies and emergency responses.



COVID-19 measures have huge impact on young people's lives

Presentation Slides

TEXAS
The University of Texas at Austin

WHAT STARTS HERE CHANGES THE WORLD

FALL 2023

TEXAS
The University of Texas at Austin

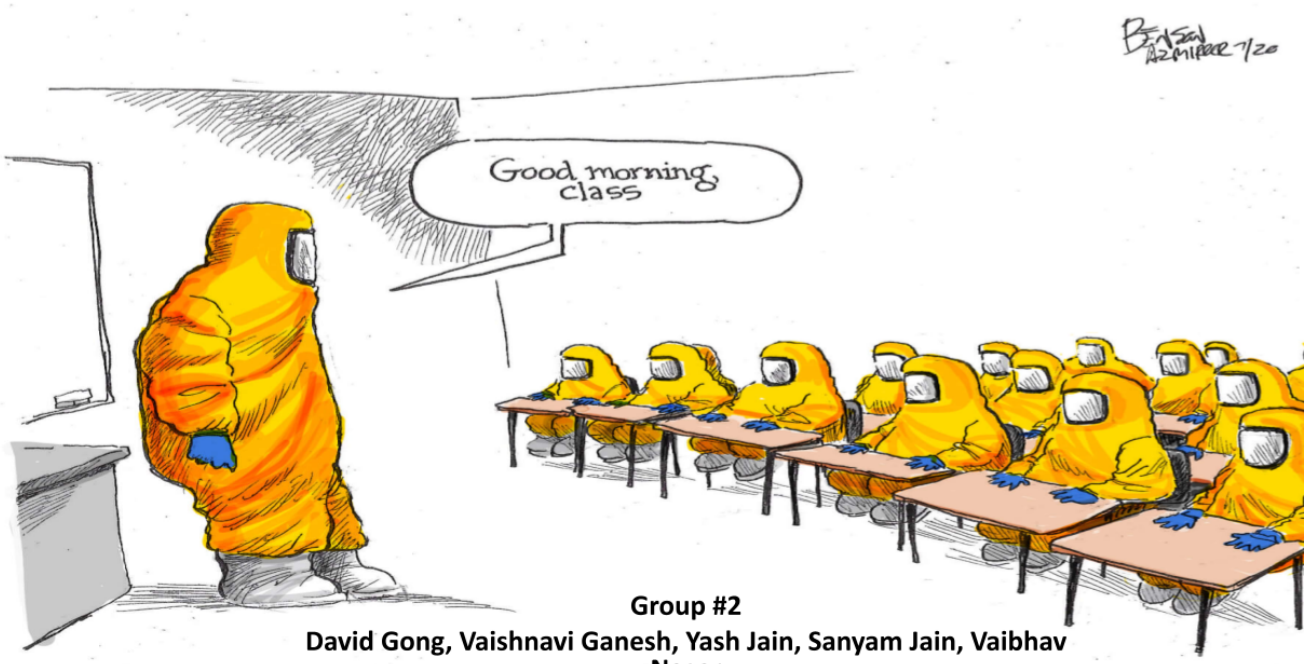
MIS 381N
INFORMATION MANAGEMENT

Final Group Project Presentation

GROUP #: 2
UNIQUE #: 04945
UT McCombs School of Business (MSBA)
The University of Texas at Austin

TEXAS
The University of Texas at Austin

WHAT STARTS HERE CHANGES THE WORLD



Group #2
David Gong, Vaishnavi Ganesh, Yash Jain, Sanyam Jain, Vaibhav Nagar
Date: 11/27

TEXAS
The University of Texas at Austin
WHAT STARTS HERE CHANGES THE WORLD

Agenda

- Introduction
- Project Overview
- Information Resources
- Tools and Technology in use
- Data Management Strategy
- Challenges
- Results/Demo
- Ongoing and future steps
- Conclusion
- Acknowledgement

Introduction

About

- Visualizing COVID-19 spread and treatment from 2020-2023
- Relevance : Covid was a huge event impacting the entire world

Goal

- Provide insights into the spread of the pandemic
- Explore various data management tools and techniques

Project Overview

Focus:
Visualizing the global spread of COVID-19

Parameters:
Infection rates, Location, and Mortality

Visualisation:
Trends, hotspots, and Measures

Information Sources

Our World in Data

- **Data dimensions:** Continents, countries, dates (2020-2023)
- **Measures:** New cases, total cases, new deaths, total deaths, vaccinations, boosters, etc
- **Huge dataset** - over 350k rows and 67 columns
- **EDA(Python)**: Missing Values, Fill null Values (Continent), Zeroes, Categorical data, Remove unknown locations

TEXAS
The University of Texas at Austin
WHAT STARTS HERE CHANGES THE WORLD

Tools and Technologies

MongoDB.		+ableau
Data storage and NoSQL	Python for data exploration	Tableau for data visualization

Data Management Strategy

Python
(Cleaning)

- Cleaned the covid dataset using python
- Removed Nulls, fixed granularity and data quality

MongoDB
(DataBase)

- Fed the data to MongoDB to create our database
- Created a connection to Tableau

Tableau
(Visualization)

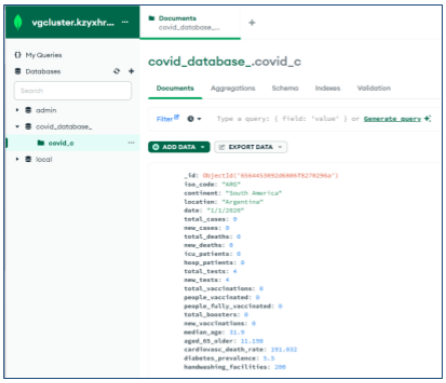
- Fetched the data from the MongoDB cluster to create dashboards
- Derived Insights

TEXAS
The University of Texas at Austin
WHAT STARTS HERE CHANGES THE WORLD

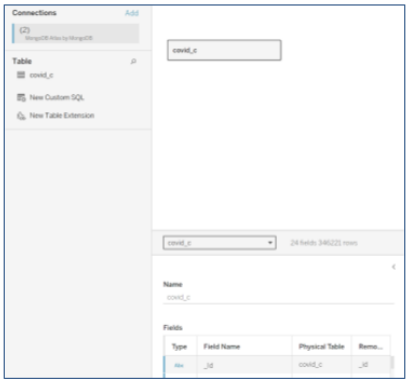
Data Management Strategy

Python

MongoDB



Tableau

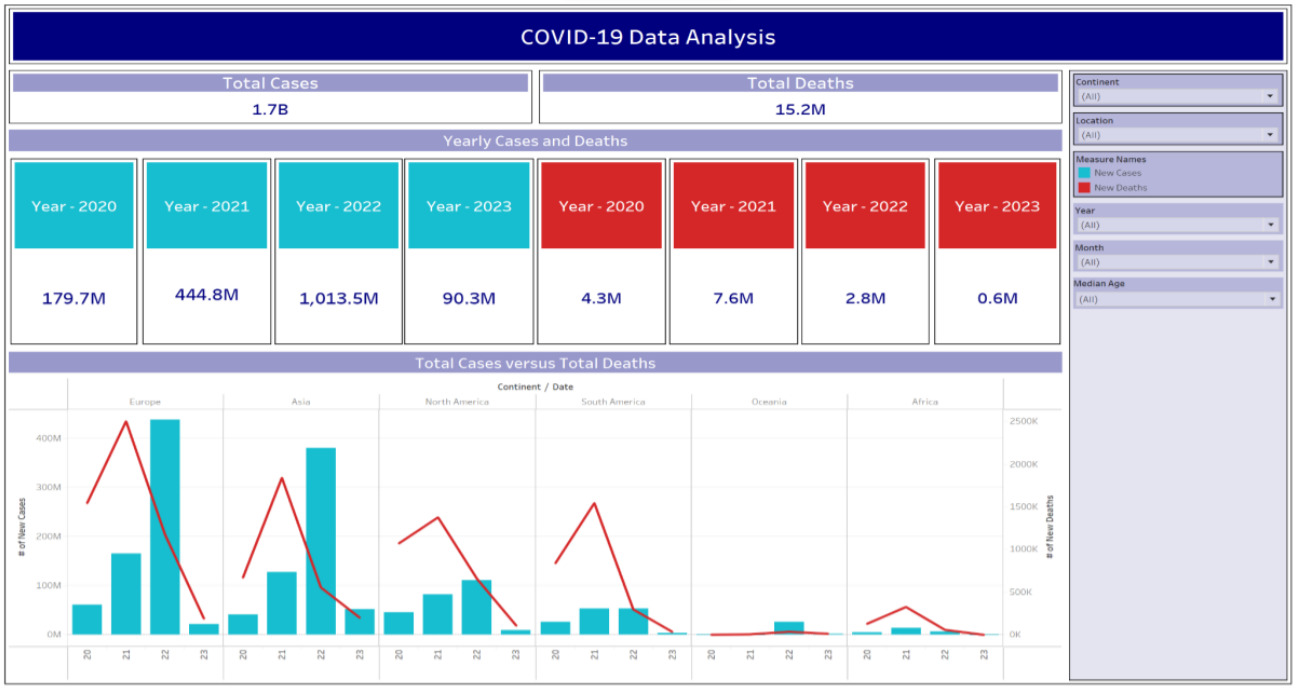


Challenges

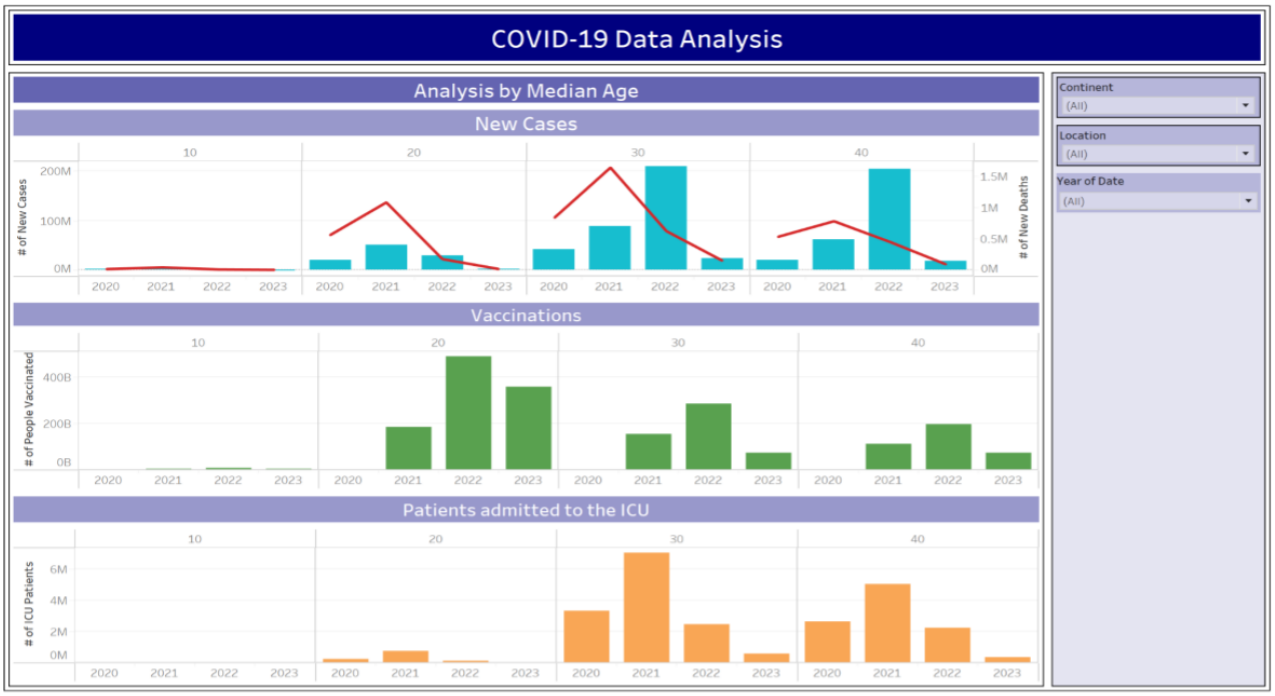
MongoDB	– Connection error – JSON File size restriction (16 mb) – Not all columns imported to Tableau
Atlas	– Licensing issues (Single cluster free) – Dashboarding not very seamless
Data	– Missing values and Null – Continent and location info

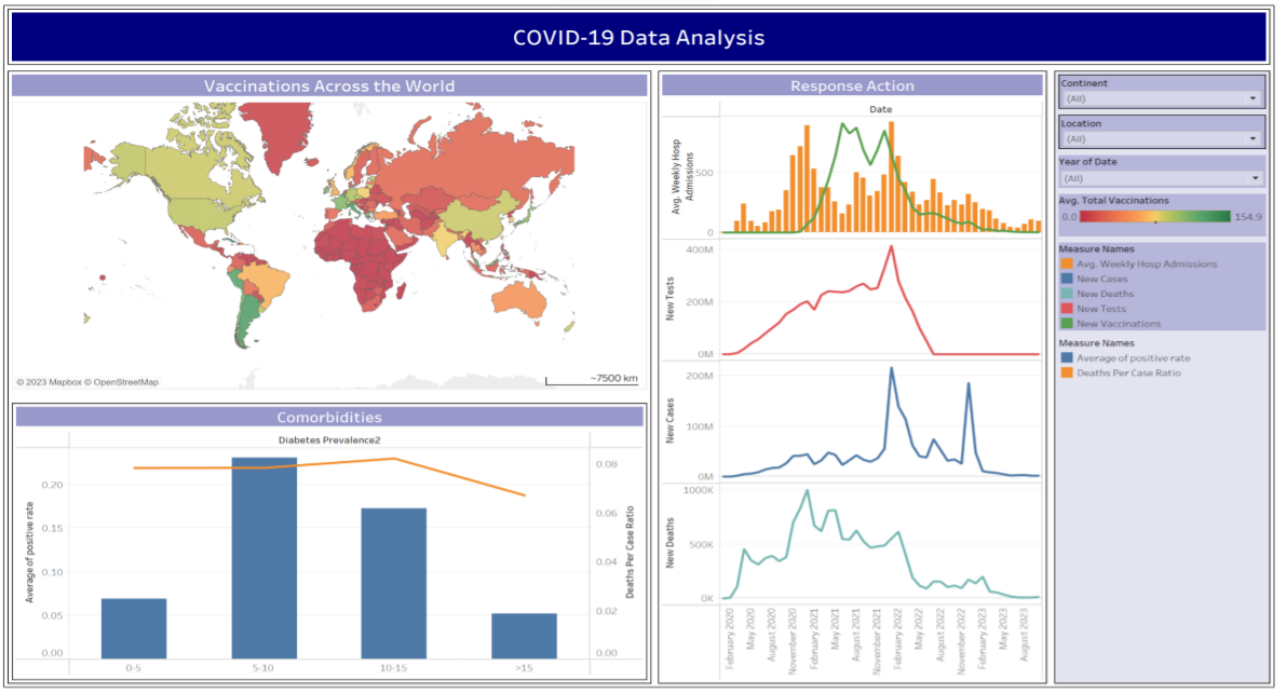
TEXAS
The University of Texas at Austin
WHAT STARTS HERE CHANGES THE WORLD

Results/Demo



TEXAS
The University of Texas at Austin
WHAT STARTS HERE CHANGES THE WORLD





TEXAS
The University of Texas at Austin
WHAT STARTS HERE CHANGES THE WORLD

Future Steps

- **Real-time Data**: Check on data in future

- **Granularity**: Go more specific for more insights (i.e. cities, demographics)

- **Predictions**: Forecast trends

- **Data Quality**: Get more comprehensive updated data

Conclusion

Data Analysis:

- **Historical Perspective:** Comprehend the historical evolution of COVID scenario
- **Dataset:** Tried using JSON Dataset (GridFS issue)
- **Granular Insights:** Filtering data by continent, country, date, and age group for detailed understanding
- **Informed Decision-Making:** Derived key insights to effectively address the dynamic nature of the COVID

Key Takeaways:

- **Project Scope:** Navigated the database ecosystem, starting with data retrieval from the COVID website
- **Data Processing:** Pre-processed the acquired dataset to enhance its quality and relevance
- **Storage Optimization:** Efficiently stored in a NoSQL database, ensuring quick and seamless extraction
- **Visualization and Insights:** Leveraging Tableau, we extracted meaningful insights from the data
- **Seamless Execution:** Entire project executed end-to-end, learning database management and analysis

TEXAS
The University of Texas at Austin
WHAT STARTS HERE CHANGES THE WORLD

Acknowledgements

- **Dataset:** https://ourworldindata.org/covid-deaths
- **Dashboard:** Referenced the Austin Covid Dashboard
- **Snowflake Assignment** : Exposure to Semi Structured Data Warehousing
- Our team for being so cooperative

Professor Joshi for being such a great professor :)

End of Report