ASSIGNMENT 2 FRONT SHEET

Qualification	**BTEC Level 5 HND Diploma in Computing**		
Unit number and title	**Unit 14: Business Intelligence**		
Submission date	August 22, 2023	**Date Received 1st submission**	
Re-submission Date		**Date Received 2nd submission**	
Student Name	Nguyen Tuan Kiet	**Student ID**	GDD210002
Class	GCD1002	**Assessor name**	Nguyen Si Thin

Student declaration

I certify that the assignment submission is entirely my own work and I fully understand the consequences of plagiarism. I understand that making a false declaration is a form of malpractice.

Student's signature	

Grading grid

P3	P4	P5	P6	M3	M4	D3	D4

❒ **Summative Feedback:**	❒ **Resubmission Feedback:**	
Grade:	**Assessor Signature:**	**Date:**
Internal Verifier's Comments:		
Signature & Date:		

UNIVERSITY of GREENWICH

Alliance with FPT Education

BTEC

Assignment 2:

Business Intelligence.

First and foremost, the first report would not have been accomplished without the assistance of friends and others. I'd want to convey my heartfelt gratitude to **Dr. Nguyen Si Thin**, who made helpful feedback to help me complete the assignment more efficiently.

Furthermore, I would like to express my sincere thanks to the authors for providing me with some important material that helped me create a better assignment report.

Table of Contents

Table of Figures

LEARNING OUTCOME 3: DEMONSTRATE THE USE OF BUSINESS INTELLIGENCE TOOLS AND TECHNOLOGIES

1. Understanding Business Intelligence and its Tools (P3)

1.1. Introduction to Business Intelligence (BI)

1.1.1. Definition of BI and its importance in modern business operations

Business Intelligence (BI) refers to a technology-driven process that involves collecting, analyzing, and presenting data to provide actionable insights for informed decision-making by executives, managers, and workers within an organization. It encompasses the entire journey from data collection and preparation to analysis and visualization, culminating in the delivery of reports, dashboards, and other informative outputs. (Stedman, 2023)

Business Intelligence (BI) is the use of data, technologies, and skills to make informed decisions that improve business performance1. BI can help businesses to:

- **Analyze customer behavior:**

Business Intelligence tools allow organizations to gather and analyze data related to customer behavior. This includes information about purchasing patterns, preferences, demographics, and interactions with the company. By understanding customer behavior, businesses can tailor their products, services, and marketing strategies to better meet customer needs. For instance, analyzing which products customers frequently buy together can lead to more effective cross-selling and upselling strategies. (Sharma, 2023)

- **Track and optimize operations:**

BI enables businesses to track and monitor various operational aspects, such as production processes, supply chain management, inventory levels, and employee performance. By closely monitoring these operational metrics, organizations can identify bottlenecks, inefficiencies, and areas for improvement. This information empowers them to streamline processes, reduce costs, and enhance overall operational efficiency. (Sharma, 2023)

- **Spot market trends:**

BI tools provide the capability to analyze historical and real-time data to identify emerging market trends and patterns. By monitoring shifts in consumer preferences, industry trends, and competitive landscape, businesses can proactively adjust their strategies. This allows them to stay ahead of market changes, launch products aligned with trends, and position themselves strategically to capture new opportunities. (Sharma, 2023)

- **Predict success:**

Business Intelligence leverages data analysis techniques, including predictive analytics, to forecast future outcomes based on historical data patterns. By analyzing factors that contributed to past successes, businesses can develop models to predict potential success for new products, campaigns, or initiatives. This enables them to make more informed decisions about resource allocation and investments. (Sharma, 2023)

- **Increase profit:**

BI's impact on increasing profits stems from its ability to optimize various aspects of business operations. By analyzing data, businesses can identify cost-saving opportunities, optimize pricing strategies, and allocate resources more efficiently. Additionally, understanding customer behavior and preferences helps tailor marketing efforts, resulting in improved customer retention, higher sales, and ultimately, increased profitability. (Sharma, 2023)

1.1.2. Overview of how BI supports decision-making and problem-solving

Business Intelligence (BI) is a comprehensive approach that empowers organizations to make informed decisions and tackle challenges through the effective utilization of data, technology, and analytical skills. It serves as a vital tool for decision-makers and problem solvers by offering a range of functionalities aimed at enhancing understanding, optimizing performance, and driving success across various business aspects:

- **Generating Reports with Valuable Insights:**

BI tools enable the creation of detailed reports that consolidate relevant data from various sources. These reports provide key performance indicators (KPIs), metrics, and trends in a structured format. Decision-makers can analyze these insights to understand the current state of the business, identify strengths and weaknesses, and make informed decisions to address challenges and opportunities. (Chopra, 2021)

- **Visualizing Data to Understand Patterns and Trends:**

BI emphasizes data visualization techniques to present complex information in an easy-to-understand visual format. Through charts, graphs, heatmaps, and interactive dashboards, decision-makers can quickly identify patterns, trends, and anomalies within the data. Visualizations make it simpler to grasp the significance of data points and relationships, leading to quicker and more accurate decisions. (Chopra, 2021)

- **Increasing Benchmarking Standards to Compare Performance:**

BI facilitates benchmarking, which involves comparing an organization's performance against industry standards, competitors, or its own historical data. This comparison provides insights into areas where the business is excelling or lagging. Decision-makers can then take actions to close performance gaps, set higher standards, and drive continuous improvement. (Chopra, 2021)

- **Unlocking the Power of Unstructured Data for New Perspectives:**

BI tools can process and analyze unstructured data, such as social media posts, customer reviews, and text documents. By extracting meaningful insights from this unstructured data, businesses gain new perspectives and insights that might not be apparent through traditional analysis. This expanded view can reveal customer sentiments, emerging trends, and potential risks. (Chopra, 2021)

- **Tracking Progress on a Daily Basis for Better Performance Management:**

BI solutions enable the monitoring of key metrics and performance indicators on a daily basis. This real-time tracking offers a clear view of progress towards goals and objectives. Decision-makers can promptly identify deviations from expected outcomes and take corrective actions to ensure that projects, initiatives, and operations stay on track. (Chopra, 2021)

- **Accessing Real-Time Sales Intelligence to Optimize Sales Strategies:**

BI tools can integrate with sales data systems to provide real-time sales intelligence. Sales teams and managers can access up-to-date information about sales performance, customer behavior, and market trends. This enables agile decision-making, adjustment of sales strategies, and quick responses to market changes, resulting in optimized sales performance. (Chopra, 2021)

1.2.Real Examples of BI Application in Businesses

This section showcases real-world case studies of businesses that have harnessed the power of Business Intelligence (BI) to achieve remarkable improvements and successes. Through the strategic application of BI tools and methodologies, these organizations have transformed their operations, enhanced customer experiences, and achieved significant growth.

Amazon:

Figure 1: 1.2_Illustration_Amazon

- **Explanation**: Amazon, being a global e-commerce leader, handles immense data to provide personalized user recommendations and optimize pricing. BI tools enable Amazon to process vast amounts of data quickly, facilitating real-time personalized recommendations and targeted marketing. Additionally, BI assists in logistics and supply chain management through predictive analysis, optimizing shipping schedules, routes, and product groupings. This leads to improved customer experience, efficient operations, and Amazon's position as a leading e-retailer. (Rudenko, 2022)

Coca-Cola Bottling Company (CCBC):

Figure 2: 1.2_Illustration_Coca-Cola

- **Explanation**: Coca-Cola Bottling Company faced challenges with manual report preparation and data analysis from various systems. The adoption of BI technologies streamlined data analysis and reporting processes. BI tools allowed data analysts to create analytical reports more efficiently, reducing manual work and optimizing the workflow. This enhancement led to quicker access to analytical insights for leadership, enabling them to make informed decisions and improve overall team performance. (Rudenko, 2022)

American Express:

Figure 3: 1.2_Illustration_American Express

- **Explanation**: American Express utilizes BI to prevent fraudulent transactions, adhere to compliance regulations, and enhance customer retention. By analyzing vast amounts of transactional data, BI tools identify suspicious activities and ensure compliance with anti-money laundering (AML) regulations. This contributes to a secure payment environment and builds trust with customers. Efficient data analysis also helps American Express optimize customer acquisition strategies, leading to improved overall online customer experiences. (Rudenko, 2022)

Starbucks:

Figure 4: 1.2_Illustration_Starbucks

- **Explanation**: Starbucks employs BI through its Loyalty Card program to collect and analyze individualized purchase data. BI tools process terabytes of data to predict customer preferences and offer personalized promotions. This enhances customer retention by providing targeted offers that align with customers' interests. Starbucks' use of BI results in consistent revenue growth, increased customer trust, and a strong community of loyal customers. (Rudenko, 2022)

In each of these cases, BI tools contribute to success by:

- **Processing Big Data**: Handling massive amounts of data efficiently to derive meaningful insights.
- **Personalization**: Enabling personalized customer experiences through tailored recommendations and offers.

- **Operational Optimization**: Streamlining logistics, supply chain management, and fraud detection processes.
- **Efficient Reporting**: Streamlining reporting processes, allowing leadership to access timely insights.
- **Data-Driven Decision-Making**: Empowering organizations to make informed decisions based on data insights.
- **Enhanced Customer Experience**: Providing better services and experiences to customers through data-driven strategies.

1.3.Business Intelligence Techniques and Tools

1.3.1. Collection Techniques:

Data cleansing methods:

Definition:

Data cleansing, also known as data cleaning or data scrubbing, is a critical process in data management that involves identifying and rectifying inaccuracies, inconsistencies, and errors within datasets. The goal of data cleansing is to ensure that the data used for analysis and decision-making is accurate, reliable, and consistent. (Tableau, n.d.)

Basic steps:

- **Step 1: Remove Duplicate or Irrelevant Observations**

In the first step of data cleansing, the focus is on eliminating redundancy and maintaining relevance. Duplicate observations, often arising from data collection from various sources, are identified and removed to ensure data accuracy and prevent skewing analysis outcomes. Similarly, observations that don't align with the specific problem under scrutiny—irrelevant data—are excluded, streamlining analysis efficiency and honing in on the intended insights. (Tableau, n.d.)

- **Step 2: Fix Structural Errors**

Addressing structural errors is pivotal in ensuring consistent and reliable data. This step involves rectifying issues like inconsistent naming conventions, typographical errors, or improper capitalization. By standardizing these inconsistencies, the data is organized coherently, preventing mislabeling or misinterpretation caused by variations in data representation. (Tableau, n.d.)

- **Step 3: Filter Unwanted Outliers**

Identifying and handling outliers is essential for maintaining data accuracy and preventing undue influence on analysis results. Outliers, data points that deviate significantly from the norm, are evaluated for their validity and relevance. While they may reveal interesting patterns or provide valuable insights, outliers stemming from data-entry errors or outliers not aligned with the analysis objective can be removed to ensure robust and accurate outcomes. (Tableau, n.d.)

- **Step 4: Handle Missing Data**

Addressing missing data is a crucial aspect of data cleansing. Given that many analysis algorithms cannot handle missing values, strategies are devised to manage this situation. Either observations with missing values can be dropped, albeit with potential information loss, or missing values can be estimated based on other available data points. An alternative approach involves adapting data utilization to accommodate null values effectively. (Tableau, n.d.)

- **Step 5: Validate and Quality Assurance (QA)**

In the final step, the cleaned data is subjected to validation and quality assurance processes. This involves assessing the data for logical coherence and adherence to the appropriate rules and standards of its field. The data's alignment with initial hypotheses or theories is checked, and trends are sought to guide further analysis. Any discrepancies or anomalies identified are examined to determine whether they stem from data quality issues, ensuring that the data's integrity and reliability are upheld. (Tableau, n.d.)

Data labeling methods:

Definition:

Data labeling methods refer to the techniques and processes employed to assign meaningful annotations, categorizations, or classifications to raw data instances. The objective of data labeling is to transform unstructured or raw data into structured and labeled datasets that can be utilized for training machine learning algorithms. This process involves assigning context, meaning, and relevance to data points, enabling machine learning models to recognize patterns, make predictions, and perform various tasks accurately. Data labeling methods play a crucial role in supervised learning, where models learn from labeled examples to generalize and make predictions on new, unlabeled data. These methods involve human annotators or automated processes to assign appropriate labels, annotations, or categories to data, making it suitable for building and training predictive models. (IBM, n.d.)

Basic steps:

- **Step 1: Define Labeling Guidelines:**

Before starting the data labeling process, establish clear and comprehensive guidelines for annotators. These guidelines should outline the categories, classes, or labels that need to be assigned to the data. Precise instructions on how to interpret and apply labels ensure consistency across annotations. (Kudan, 2023)

- **Step 2: Select and Train Annotators:**

Choose annotators who possess domain knowledge and expertise relevant to the dataset. Provide comprehensive training on the labeling guidelines, emphasizing examples and edge cases. Consistent and accurate annotations rely on well-trained annotators who understand the nuances of the task. (Kudan, 2023)

- **Step 3: Sample Data Selection:**

Select a representative subset of the data to train and validate the annotators' understanding of the labeling guidelines. This sample serves as a benchmark for consistency and quality assurance, allowing us to assess annotators' performance before labeling the entire dataset. (Kudan, 2023)

- **Step 4: Annotation Process:**

Annotators review individual data instances and apply the appropriate labels based on the established guidelines. This step requires meticulous attention to detail to ensure accurate labeling and prevent errors. Collaborative tools and platforms can aid in efficient annotation. (Kudan, 2023)

- **Step 5: Quality Control and Review:**

Implement a review process to maintain labeling accuracy and consistency. Reviewers, often more experienced annotators or domain experts, assess the labeled data against the guidelines. Any discrepancies or errors are identified and corrected, ensuring high-quality annotations. (Kudan, 2023)

- **Step 6: Iterative Refinement:**

Continuously engage with annotators and reviewers to address questions, clarify guidelines, and refine the labeling process. This iterative approach improves labeling accuracy over time and ensures a deep understanding of the data and task. (Kudan, 2023)

- **Step 7: Benchmark and Agreement Measurement:**

Calculate inter-annotator agreement to gauge the consistency of labeling among different annotators. This benchmark helps assess labeling quality and identify areas that may need further clarification or training. (Kudan, 2023)

- **Step 8: Scale Up and Label Full Dataset:**

Once the labeling process is refined and quality is assured, proceed to label the entire dataset using the guidelines and insights gained from the training and validation phases. (Kudan, 2023)

- **Step 9: Monitor and Maintain Quality:**

Maintain ongoing monitoring of labeling quality even after the full dataset is labeled. Regularly review labeled data for accuracy and address any emerging challenges promptly. (Kudan, 2023)

- **Step 10: Feedback Loop and Continuous Learning:**

Establish a feedback loop between annotators, reviewers, and data analysts. Encourage a culture of continuous learning and improvement by incorporating feedback to enhance the labeling process over time. (Kudan, 2023)

1.3.2. Analysis Techniques:

Creating Dashboards

In the realm of Business Intelligence (BI), analysis techniques play a crucial role in transforming raw data into actionable insights. These techniques encompass a range of methods and tools that facilitate data exploration, interpretation, and decision-making. Three fundamental analysis techniques are generating reports, executing queries, and creating dashboards. (DashThis, n.d.)

Generating Reports:

Reports are essential tools for presenting structured information extracted from the underlying data. BI systems allow users to generate customized reports that display key performance indicators, metrics, and trends. Reports provide a clear snapshot of specific aspects of the business, aiding in informed decision-making. Through drag-and-drop interfaces and pre-designed templates, users can select data fields, apply filters, and aggregate data to create comprehensive reports that highlight performance, identify areas for improvement, and support strategic planning. (DashThis, n.d.)

Executing Queries:

Querying data involves extracting specific information from databases by formulating queries based on user-defined criteria. BI platforms offer query tools that enable users to interactively retrieve relevant data segments. SQL (Structured Query Language) is commonly used to construct queries that retrieve data from relational databases. BI query tools provide an intuitive interface to build, modify, and execute queries, empowering users to explore data and extract insights in real-time. The ability to query data flexibly enhances the analysis process by facilitating ad-hoc exploration and uncovering patterns. (DashThis, n.d.)

Creating Dashboards:

Dashboards are dynamic visual representations of data insights that provide a comprehensive overview of business performance. BI dashboards combine various visual components such as charts, graphs, gauges, and maps to convey information intuitively. Users can customize dashboards to display relevant metrics, KPIs, and trends in real-time, enabling quick decision-making. These interactive dashboards facilitate monitoring and analysis across multiple dimensions, departments, and timeframes. With features like drill-down capabilities and interactivity, dashboards empower users to explore data at different levels of granularity and gain deeper insights. (DashThis, n.d.)

1.3.3. Analytic Techniques:

In the realm of Business Intelligence (BI), analytic techniques serve as powerful tools for extracting meaningful insights and predicting future trends from data. Two key analytic techniques that contribute to informed decision-making are regression analysis and machine learning for predictive insights.

Regression analysis for predictive insights:

Regression analysis is a statistical method used to explore the relationship between variables and predict the outcome of a dependent variable based on one or more independent variables. Specifically, linear regression seeks to find the best-fitting line that represents the relationship between variables. This line allows us to make predictions about the dependent variable's values based on the values of the independent variables. By calculating coefficients for each independent variable, regression analysis quantifies how much each variable contributes to the variation in the dependent variable. For instance, in a sales context, linear regression can help understand how changes in advertising expenditures influence sales. This technique provides insights into causality, allowing businesses to identify key drivers and make informed decisions. (Ray, 2015)

Machine learning for predictive insights

Machine learning is a data-driven approach where algorithms learn patterns from historical data to make predictions or decisions without explicit programming. In the realm of Business Intelligence, machine learning is harnessed for predictive insights. Supervised machine learning involves training a model on labeled data, where the relationship between inputs (features) and outputs (labels) is known. This model can then predict labels for new, unseen data. For example, a retail company can predict customer churn based on historical behavior data. Unsupervised machine learning, on the other hand, identifies hidden patterns within unlabeled data, revealing clusters or associations that might not be apparent through manual analysis. These patterns can offer valuable insights into customer segmentation, product recommendations, or anomaly detection. (Castillo, 2021)

Both regression analysis and machine learning for predictive insights elevate the capabilities of Business Intelligence. Regression analysis provides a structured approach to understanding relationships and making predictions based on established variables. Machine learning, on the other hand, offers a dynamic and adaptable way to predict outcomes by learning patterns from data. The insights gained from these techniques empower organizations to anticipate trends, optimize strategies, and make proactive decisions, thereby enhancing operational efficiency and competitiveness. By incorporating these advanced analytic techniques into BI workflows, businesses can leverage data-driven insights to drive innovation and success.

1.3.4. BI Tools:

Programming Tools:

Programming tools are essential for performing various data-related tasks in Business Intelligence. Python, R, and SQL are prominent examples.

- **Python**: Python is a versatile programming language with a wide array of libraries for data manipulation, analysis, and visualization. Its simplicity and robust ecosystem make it an excellent choice for data cleaning, transformation, statistical analysis, and even machine learning.
- **R**: R is a programming language and software environment specifically designed for statistical computing and graphics. It provides a vast array of packages for statistical analysis, data visualization, and modeling. Researchers and analysts often use R for data exploration, hypothesis testing, linear and nonlinear modeling, and creating visual representations of data through various plotting libraries.
- **SQL**: SQL (Structured Query Language) is a domain-specific language used for managing and manipulating relational databases. It enables users to perform tasks like retrieving, inserting, updating, and deleting data from databases. SQL is integral to Business Intelligence as it allows users to interact with databases and perform complex queries to extract valuable insights from structured data.

Database Management Systems:

Database Management Systems (DBMS) are software applications that facilitate the storage, retrieval, and management of data in a structured manner. Examples include MySQL, MongoDB, and PostgreSQL.

- **MySQL**: MySQL is a popular open-source relational database management system. It offers a reliable way to store structured data and supports SQL queries for data retrieval and manipulation. It's often used for traditional structured data storage and management.
- **MongoDB**: MongoDB is a NoSQL database that excels in handling unstructured or semi-structured data. It uses a flexible document-oriented model, making it suitable for scenarios where data types vary widely. MongoDB is well-suited for applications requiring scalability and handling diverse data formats.

Data Visualization Platforms:

Data visualization platforms are crucial for transforming complex data into easily understandable visual representations.

- **Microsoft Power BI**: Microsoft Power BI is a powerful data visualization and business intelligence tool. It allows users to connect to various data sources, create interactive dashboards, reports, and visualizations. Power BI is user-friendly and suitable for creating compelling visuals to communicate insights.
- **Tableau**: Tableau is another popular data visualization platform that empowers users to create interactive and shareable dashboards and reports. It offers a wide range of visualization options and helps users explore data visually, uncover patterns, and make data-driven decisions effectively.

2. Designing a Business Intelligence Tool for Decision Support (P4)

2.1. Team and Dataset Introduction

2.1.1. Introduce about the team: K-BI

In the year 2020, the world experienced an unprecedented global pandemic caused by the outbreak of COVID-19. The virus quickly spread across continents, affecting millions of lives and posing significant challenges to healthcare systems, economies, and societies worldwide.

K-BI, which stands for "Knowledge-Business Intelligence," is a specialized sub-organization established under the umbrella of the World Health Organization (WHO) to address the specific needs arising from the COVID-19 pandemic. As a division of the WHO, K-BI focuses on harnessing data-driven insights to provide comprehensive statistics and business intelligence related to COVID-19 cases worldwide. The primary objective of K-BI is to facilitate evidence-based decision-making, support public health responses, and coordinate international efforts to combat the virus effectively.

K-BI is tasked with compiling comprehensive statistics related to COVID-19 cases, including the number of confirmed infections and fatalities over time. Their dataset contains valuable information that sheds light on the progression of the pandemic and its effects on different regions and communities globally.

2.1.2. Description of the dataset used for the project.

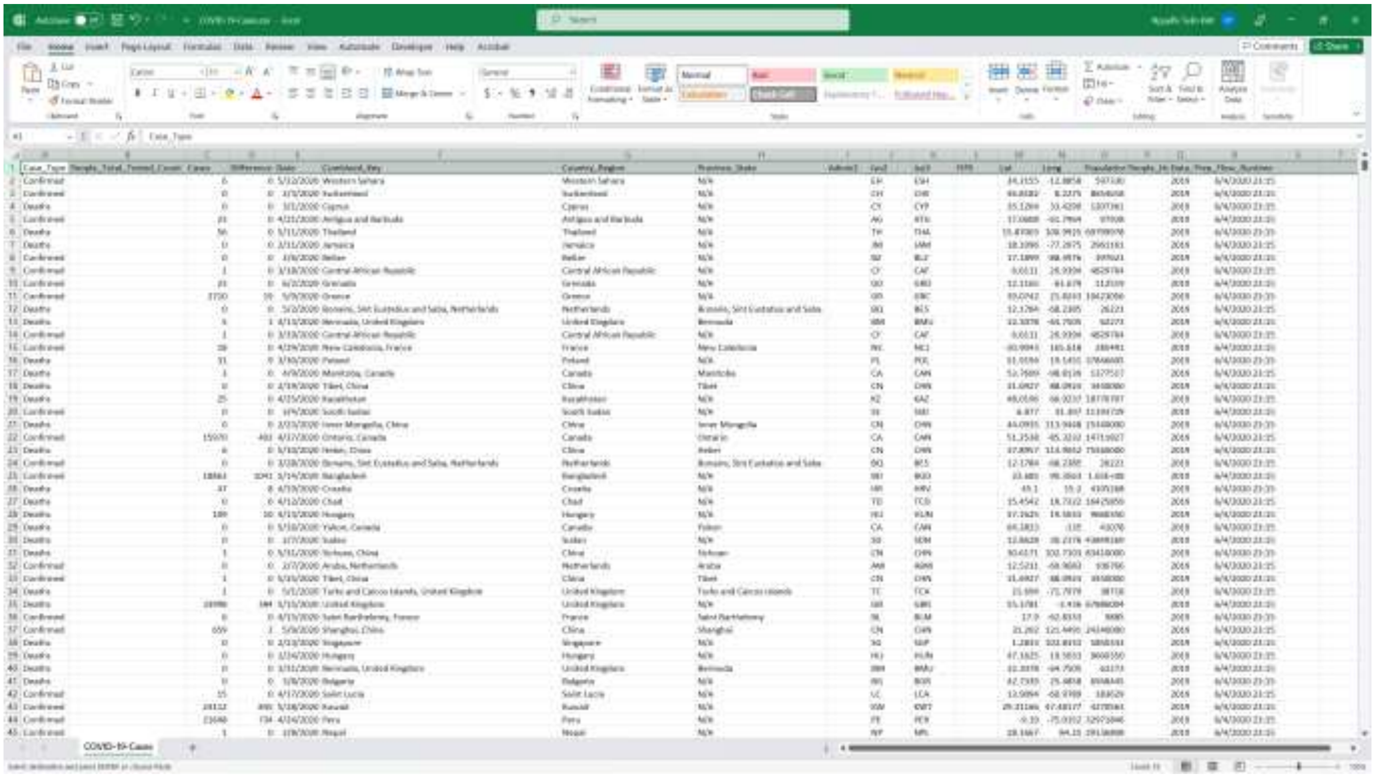


Figure 5: 2.1.1_K-BI_Business Dataset

Data Source:
https://data.world/covid-19-data-resource-hub/covid-19-case-counts/workspace/file?filename=COVID-19+Cases.csv

These columns include:

- **Case_Type**: This column specifies the type of COVID-19 case, distinguishing between confirmed infections, deaths, and possibly other classifications.
- **People_Total_Tested_Count**: The number of individuals who have undergone COVID-19 testing, providing insight into testing efforts and identifying potential trends in infection rates.
- **Cases Difference**: The difference in the number of cases compared to the previous recorded date, facilitating the observation of daily changes and trends.
- **Date**: The date on which the COVID-19 data was recorded, enabling temporal analysis and trend identification.
- **Combined_Key**: A key that uniquely identifies a specific location or region, facilitating data aggregation and regional comparisons.

- **Country_Region**: The name of the country or territory to which the data pertains, helping to categorize and compare data on a global scale.
- **Province_State**: If applicable, this column specifies the name of the administrative division within a country or region, offering a more granular analysis.
- **Admin2**: Further administrative division information, if relevant, to pinpoint specific localities within a province or state.
- **iso2 and iso3**: The ISO country codes, which are standardized country codes, aiding in data integration with external datasets.
- **FIPS**: The Federal Information Processing Standards code, used for data processing and linking with administrative boundaries in the United States.
- **Lat and Long**: Latitude and longitude coordinates of the region, essential for geographical mapping and spatial analysis.
- **Population_Count**: The total population count of the region, providing context for understanding the pandemic's impact on different population sizes.
- **People_Hospitalized_Cumulative_Count**: The cumulative number of individuals hospitalized due to COVID-19, offering insights into the severity of cases and healthcare system stress.
- **Data_Source**: The source from which the data is derived, ensuring transparency and facilitating data credibility.
- **Prep_Flow_Runtime**: Information related to data preparation and processing times, important for data quality assessment

2.2.Pre-processing and Data Preparation

2.2.1. General Data Description:

Before diving into the specific data processing steps, it's important to get a general overview of the dataset. Understand the structure, the meaning of each column, and how the data is organized. This step helps us become familiar with the data we're working with and prepares our for subsequent processing steps.

Loading the Dataset:

```
og_data = pd.read_csv(
    'C:/Users/PC/Desktop/Learning/Tableau/Dashboard_Covid/COVID-19-Cases.csv',
low_memory=False)
```

The code reads a CSV file named 'COVID-19-Cases.csv' located at the specified path using the pd.read_csv() function from the pandas library. The low_memory=False parameter is used to disable the memory optimization mode. This can be useful for large datasets to prevent potential dtype inference issues.

Displaying the First 10 Rows:

```
print(og_data.head(10))
```

The head() method is used to display the first 10 rows of the dataset.

```
Output exceeds the size limit. Open the full output data in a text editor
   Case_Type  People_Total_Tested_Count  Cases  Difference       Date  \
0  Confirmed                        NaN      6           0  5/22/2020
1  Confirmed                        NaN      0           0   2/3/2020
2     Deaths                        NaN      0           0   3/1/2020
3  Confirmed                        NaN     23           0  4/21/2020
4     Deaths                        NaN     56           0  5/11/2020
5     Deaths                        NaN      0           0  2/11/2020
6  Confirmed                        NaN      0           0   2/6/2020
7  Confirmed                        NaN      1           0  3/18/2020
8  Confirmed                        NaN     23           0   6/2/2020
9  Confirmed                        NaN   2710          19   5/9/2020

               Combined_Key            Country_Region Province_State Admin2  \
0            Western Sahara            Western Sahara            NaN    NaN
1               Switzerland               Switzerland            NaN    NaN
2                    Cyprus                    Cyprus            NaN    NaN
3       Antigua and Barbuda       Antigua and Barbuda            NaN    NaN
4                  Thailand                  Thailand            NaN    NaN
5                   Jamaica                   Jamaica            NaN    NaN
6                    Belize                    Belize            NaN    NaN
7  Central African Republic  Central African Republic            NaN    NaN
8                   Grenada                   Grenada            NaN    NaN
9                    Greece                    Greece            NaN    NaN

  iso2 iso3  FIPS        Lat        Long  Population_Count  \
...
6  2019 Novel Coronavirus COVID-19 (2019-nCoV) Da...   6/4/2020 11:15:39 PM
7  2019 Novel Coronavirus COVID-19 (2019-nCoV) Da...   6/4/2020 11:15:39 PM
8  2019 Novel Coronavirus COVID-19 (2019-nCoV) Da...   6/4/2020 11:15:39 PM
9  2019 Novel Coronavirus COVID-19 (2019-nCoV) Da...   6/4/2020 11:15:39 PM
```

Displaying the Last 10 Rows:

```
print(og_data.tail(10))
```

The tail() method is used to display the last 10 rows of the dataset.

```
Output exceeds the size limit. Open the full output data in a text editor
        Case_Type  People_Total_Tested_Count  Cases  Difference       Date  \
950660  Confirmed                        NaN      0           0  1/22/2020
950661     Deaths                        NaN      0           0  1/22/2020
950662  Confirmed                        NaN      0           0  1/22/2020
950663  Confirmed                        NaN      0           0  1/22/2020
950664  Confirmed                        NaN      0           0  1/22/2020
950665     Deaths                        NaN      0           0  1/22/2020
950666  Confirmed                        NaN      0           0  1/22/2020
950667     Deaths                        NaN      0           0  1/22/2020
950668  Confirmed                        NaN      0           0  1/22/2020
950669  Confirmed                        NaN      0           0  1/22/2020

                     Combined_Key Country_Region Province_State     Admin2  \
950660        Grady, Georgia, US             US        Georgia      Grady
950661       Noble, Oklahoma, US             US       Oklahoma      Noble
950662       Payne, Oklahoma, US             US       Oklahoma      Payne
950663    Reynolds, Missouri, US             US       Missouri   Reynolds
950664     Maury, Tennessee, US             US      Tennessee      Maury
950665     Missoula, Montana, US             US        Montana   Missoula
950666   Ontonagon, Michigan, US             US       Michigan  Ontonagon
950667  Sargent, North Dakota, US            US   North Dakota    Sargent
950668        Ohio, Kentucky, US             US       Kentucky       Ohio
950669    Marion, Mississippi, US            US    Mississippi     Marion

       iso2 iso3     FIPS        Lat        Long  Population_Count  \
...
950666  6/4/2020 11:15:39 PM
950667  6/4/2020 11:15:39 PM
950668  6/4/2020 11:15:39 PM
950669  6/4/2020 11:15:39 PM
```

2.2.2. Statistical Data Description:

Compute basic statistical measures for numerical columns in the dataset. This includes calculating measures like mean, median, standard deviation, and quartiles. These statistics provide insights into the central tendency and distribution of the data, helping to identify outliers or anomalies.

Output of og_data.info():

```
print(og_data.info())
```

The info() method gives an overview of the dataset's structure, including the data types and the number of non-null values for each column.

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 950670 entries, 0 to 950669
Data columns (total 18 columns):
 #   Column                               Non-Null Count   Dtype
---  ------                               --------------   -----
 0   Case_Type                            950670 non-null  object
 1   People_Total_Tested_Count            6048 non-null    float64
 2   Cases                                950670 non-null  int64
 3   Difference                           950670 non-null  int64
 4   Date                                 950670 non-null  object
 5   Combined_Key                         950670 non-null  object
 6   Country_Region                       950670 non-null  object
 7   Province_State                       901260 non-null  object
 8   Admin2                               878580 non-null  object
 9   iso2                                 949590 non-null  object
 10  iso3                                 949860 non-null  object
 11  FIPS                                 849690 non-null  float64
 12  Lat                                  921780 non-null  float64
 13  Long                                 921780 non-null  float64
 14  Population_Count                     921780 non-null  float64
 15  People_Hospitalized_Cumulative_Count 6048 non-null    float64
 16  Data_Source                          950670 non-null  object
 17  Prep_Flow_Runtime                    950670 non-null  object
dtypes: float64(6), int64(2), object(10)
memory usage: 130.6+ MB
None
```

The dataset has a total of 950,670 entries and 18 columns. The columns have varying counts of non-null values, indicating the presence of missing data. There are columns with data types: float64 (6 columns), int64 (2 columns), and object (10 columns). The memory usage of the dataset is approximately 130.6+ MB.

Output of og_data.describe():

```
print(og_data.describe())
```

The describe() method provides basic statistics for numerical columns (data type: int or float).

```
Output exceeds the size limit. Open the full output data in a text editor
       People_Total_Tested_Count          Cases      Difference           FIPS  \
count               6.048000e+03  950670.000000  950670.000000  849690.000000
mean                1.672996e+05     273.518654       7.388438   30335.484906
std                 2.643341e+05    5187.497792     166.491030   15196.097280
min                 3.000000e+00       0.000000  -10034.000000      60.000000
25%                 2.980700e+04       0.000000       0.000000   18169.000000
50%                 7.798700e+04       0.000000       0.000000   29171.000000
75%                 1.897678e+05       4.000000       0.000000   45079.000000
max                 2.293032e+06  614941.000000   33274.000000   56045.000000

                 Lat           Long  Population_Count  \
count  921780.000000  921780.000000      9.217800e+05
mean       37.115970     -83.442661      2.260264e+06
std         9.763286      38.621432      2.661885e+07
min       -51.796300    -174.159600      8.600000e+01
25%        34.068596     -97.698949      1.202300e+04
50%        38.154335     -89.205569      2.969950e+04
75%        41.772338     -81.926357      9.792800e+04
max        71.706900     178.065000      1.380004e+09

       People_Hospitalized_Cumulative_Count
count                           6048.000000
mean                            2625.715278
std                             9465.118234
min                                0.000000
...
25%                                0.000000
50%                              325.500000
75%                             1697.500000
max                            89995.000000
```

- **Cases**: The mean number of confirmed cases is approximately 273, with a standard deviation of 5187. The minimum is 0, and the maximum is 614941.
- **Difference**: The mean difference (likely daily case changes) is around 7.4, with a standard deviation of 166.49. The minimum is -10034, indicating potential errors or negative values.
- **FIPS**: The Federal Information Processing Standards code statistics show the range of values and the standard deviation.

Output of og_data.describe(include='object'):

```
print(og_data.describe(include='object'))
```

This part of the output provides statistics for categorical columns (data type: object).

```
       Case_Type       Date    Combined_Key Country_Region Province_State  \
count     950670     950670          950670         950670         901260
unique         2        135            3521            187            135
top    Confirmed  5/22/2020  Western Sahara             US          Texas
freq      475335       7042             270         879930          69120

            Admin2    iso2    iso3  \
count       878580  949590  949860
unique        1901     215     216
top     Unassigned      US     USA
freq         13770  878580  878580

                                              Data_Source  \
count                                              950670
unique                                                  1
top     2019 Novel Coronavirus COVID-19 (2019-nCoV) Da ...
freq                                               950670

           Prep_Flow_Runtime
count                 950670
unique                     1
top     6/4/2020 11:15:39 PM
freq                  950670
```

- **Case_Type**: There are two unique case types (Confirmed and Deaths) with a frequency of occurrences.
- **Country_Region**: The most common country region is the US.

2.2.3. Missing Data Handling:

Identify columns with missing data and decide how to handle them. Options include removing rows with missing values, imputing missing values with the mean or median, or using more advanced imputation techniques. It's important to choose an approach that preserves the integrity of the data.

Calculating Percentage of Missing Values:

```
print(og_data.isnull().sum()*100/len(og_data))
```

The COVID-19 dataset exhibits varying degrees of missing data across its columns. Notably, the "People_Total_Tested_Count" and "People_Hospitalized_Cumulative_Count" columns show extremely high percentages (99.36%) of missing values, potentially suggesting incomplete or inconsistent reporting. Columns related to geographic details like "Province_State," "Admin2," "FIPS," "Lat," "Long," and "Population_Count" have relatively moderate proportions of missing values, ranging from around 3% to 10.62%. In contrast, essential attributes such as case counts, type, date, country, and data source exhibit no missing values. Addressing these missing values will depend on the intended analysis and the significance of the affected columns, with careful consideration given to the potential biases that imputation or removal of these values might introduce into the analysis.

```
Case_Type                               0.000000
People_Total_Tested_Count              99.363817
Cases                                   0.000000
Difference                              0.000000
Date                                    0.000000
Combined_Key                            0.000000
Country_Region                          0.000000
Province_State                          5.197387
Admin2                                  7.583073
iso2                                    0.113604
iso3                                    0.085203
FIPS                                   10.621982
Lat                                     3.038909
Long                                    3.038909
Population_Count                        3.038909
People_Hospitalized_Cumulative_Count   99.363817
Data_Source                             0.000000
Prep_Flow_Runtime                       0.000000
dtype: float64
```


Nguyen Tuan Kiet
GDD210002

Identifying Columns with Missing Values:

```
# Find columns with missing values
columns_with_missing_values = og_data.columns[og_data.isnull().sum() > 0]

# Display the list of columns with missing values
print("Columns with missing values:")
print(columns_with_missing_values)
```

```
Columns with missing values:
Index(['People_Total_Tested_Count', 'Province_State', 'Admin2', 'iso2', 'iso3',
       'FIPS', 'Lat', 'Long', 'Population_Count',
       'People_Hospitalized_Cumulative_Count'],
      dtype='object')
```

The condition og_data.isnull().sum() > 0 returns a boolean Series indicating columns with missing values, and the .columns attribute retrieves the column names.

Dropping Columns with Missing Values:

```
# Drop columns with missing values: People_Total_Tested_Count,
People_Hospitalized_Cumulative_Count
og_data.drop(columns=['People_Total_Tested_Count',
                      'People_Hospitalized_Cumulative_Count'], inplace=True)
```

This code drops columns 'People_Total_Tested_Count' and 'People_Hospitalized_Cumulative_Count' from the dataset. Columns with a very high percentage of missing values, such as 99.36% in this case, might indicate that the data in these columns is not consistently or accurately reported. This can severely compromise the integrity and quality of the dataset, potentially leading to incorrect or misleading analyses.

Filling Missing Values in 'Population_Count' Column:

```
og_data['Population_Count'].fillna(og_data['Population_Count'].mean(),
inplace=True)
```

This code fills missing values in the 'Population_Count' column with the mean of the non-missing values in the same column.

Filling missing values in the 'Population_Count' column is a reasonable approach due to the relatively low percentage of missing values (3.04%). This column contains information about the population count of each region, which can be a crucial factor in understanding the impact of the COVID-19 pandemic on different areas. Since the missing values represent only a small portion of the dataset, imputing them with a suitable method, such as using the mean population count or employing a more advanced imputation technique, can help provide a completer and more accurate picture of the pandemic's effects.

Displaying Missing Values After Handling:

```
print(og_data.isnull().sum()*100/len(og_data))
```

```
Case_Type              0.000000
Cases                  0.000000
Difference             0.000000
Date                   0.000000
Combined_Key           0.000000
Country_Region         0.000000
Province_State         5.197387
Admin2                 7.583073
iso2                   0.113604
iso3                   0.085203
FIPS                  10.621982
Lat                    3.038909
Long                   3.038909
Population_Count       0.000000
Data_Source            0.000000
Prep_Flow_Runtime      0.000000
dtype: float64
```

2.2.4. Data Consistency Check:

Check for inconsistencies in the data, such as incorrect values or formats. This step helps maintain data accuracy and reliability.

```
print(og_data['Case_Type'].unique())
```

```
['Confirmed' 'Deaths']
```

The 'Case_Type' column in the COVID-19 dataset contains two unique values: 'Confirmed' and 'Deaths'. These values indicate the type of COVID-19 cases reported in the dataset. The presence of only these two distinct values suggests that the dataset includes records for confirmed infections and reported deaths due to COVID-19. This information will be essential for further analysis and visualization, allowing us to distinguish between different case types and understand the progression and impact of the pandemic more effectively.

2.2.5. Visualize Data:

Create visualizations to explore the data further. This could involve plotting time series data to observe trends in COVID-19 cases, mapping geographical data using latitude and longitude coordinates, and generating line charts or bar charts to understand the distribution of numerical variables. Visualization helps uncover patterns and insights that might not be evident from the raw data.

Line Chart

The Confirmed Cases in the World Chart

```
# Subplot between 'Case_Type' and Date

## 'Confirmed' cases and 'Date' are plotted
### Filter the data to only include 'Confirmed' cases
confirmed_data = backup_df[backup_df['Case_Type'] == 'Confirmed']
# Convert the 'Date' column to datetime type
confirmed_data['Date'] = pd.to_datetime(confirmed_data['Date'])

# Sort the data by date in ascending order
confirmed_data = confirmed_data.sort_values(by='Date')

# Create the subplot
fig, ax = plt.subplots(figsize=(10, 6))

# Plot the data using seaborn's lineplot
sns.lineplot(x='Date', y='Cases', data=confirmed_data, ax=ax, color='orange')

# Set labels and title
ax.set_xlabel('Date')
ax.set_ylabel('Confirmed Cases')
ax.set_title('Confirmed COVID-19 Cases Over Time')

# Rotate x-axis labels for better visibility
plt.xticks(rotation=45)

# Show the plot
plt.tight_layout()
plt.show()
```

The code snippet creates a subplot to visualize the trend of 'Confirmed' COVID-19 cases over time. The data is filtered to include only 'Confirmed' cases, and the 'Date' column is converted to the datetime data type to facilitate time-based plotting. The filtered data is then sorted chronologically based on the 'Date' column.

Using the Seaborn library's lineplot function, the code generates a line plot with the 'Date' on the x-axis and the 'Cases' (number of confirmed cases) on the y-axis. The line plot appears in orange color to distinguish it from other visualizations. Labels for the x-axis, y-axis, and the plot title are set appropriately, with the x-axis labels rotated for better readability.

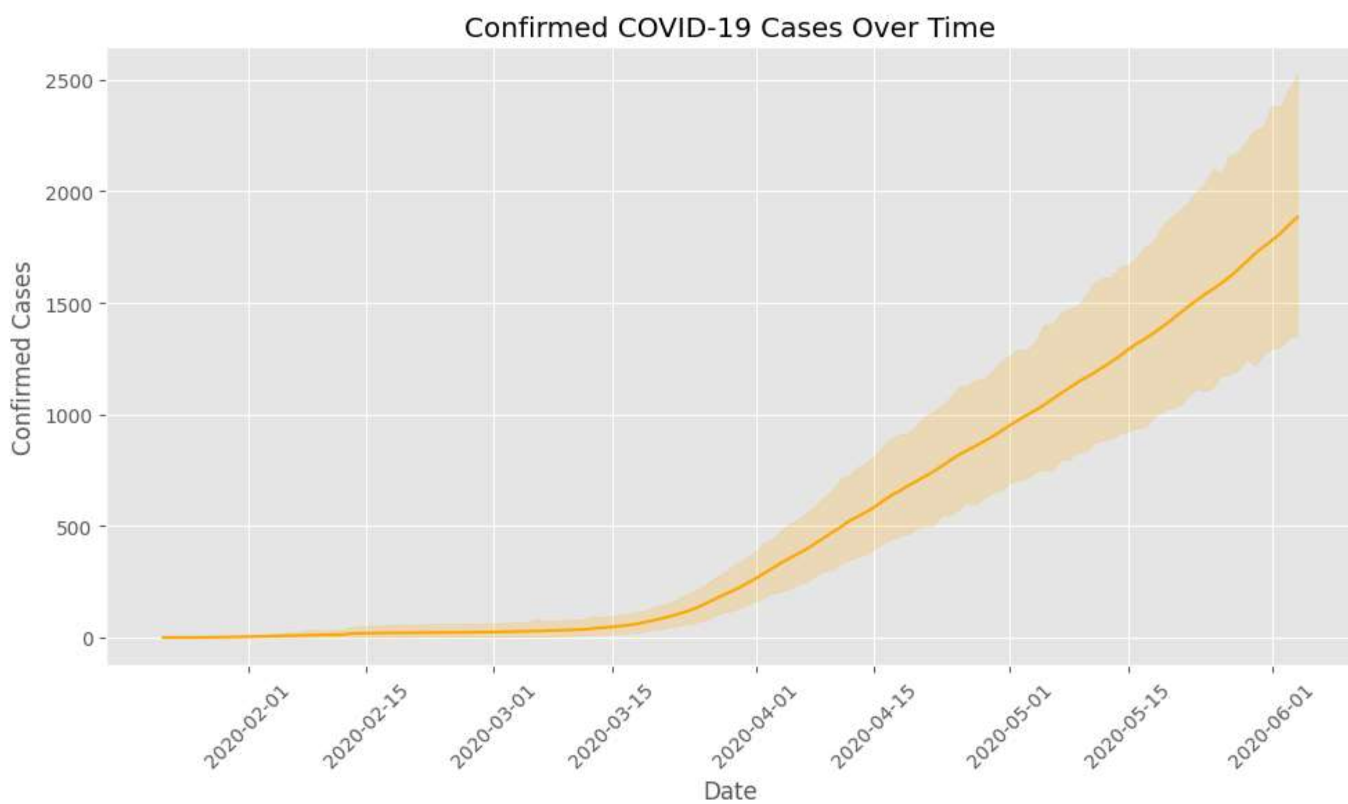


Figure 6: 2.2.5_Confirmed COVID-19 Cases Over Time

The line graph illustrates the number of confirmed COVID-19 cases over a period of six months from January to June 2020. It is clear that the number of cases increased steadily throughout the time frame. In January, there were only a few confirmed cases of COVID-19, less than 100. However, by February, the number of cases rose sharply to over 500. The trend continued in March and April, reaching over 1000 and 1500 cases respectively. In May, the number of cases slowed down slightly, but still increased to over 2000. Finally, in June, the number of cases reached its peak at around 2500.

The Deaths Cases in the World Chart

```
## 'Confirmed' cases and 'Date' are plotted
### Filter the data to only include 'Confirmed' cases
deaths_data = backup_df[backup_df['Case_Type'] == 'Deaths']
### Convert the 'Date' column to datetime type
deaths_data['Date'] = pd.to_datetime(deaths_data['Date'])

### Sort the data by date in ascending order
deaths_data = deaths_data.sort_values(by='Date')

### Create the subplot
fig, ax = plt.subplots(figsize=(10, 6))

### Plot the data using seaborn's lineplot
sns.lineplot(x='Date', y='Cases', data=deaths_data, color='red', ax=ax)

### Set labels and title
ax.set_xlabel('Date')
ax.set_ylabel('Deaths Cases')
ax.set_title('Deaths COVID-19 Cases Over Time')

### Rotate x-axis labels for better visibility
plt.xticks(rotation=45)

### Show the plot
plt.tight_layout()
plt.show()
```

The code snippet serves the purpose of visualizing the trend of 'Deaths' COVID-19 cases over time. Similar to the previous code snippet, the data is filtered to include only 'Deaths' cases, and the 'Date' column is converted to the datetime data type. Subsequently, the filtered data is sorted chronologically based on the 'Date' column.

Using the Seaborn library's lineplot function, the code generates a line plot with the 'Date' on the x-axis and the 'Cases' (number of deaths) on the y-axis. The line plot is colored in red to distinguish it from previous visualizations. Labels for the x-axis, y-axis, and the plot title are appropriately set, with x-axis labels rotated for improved readability.

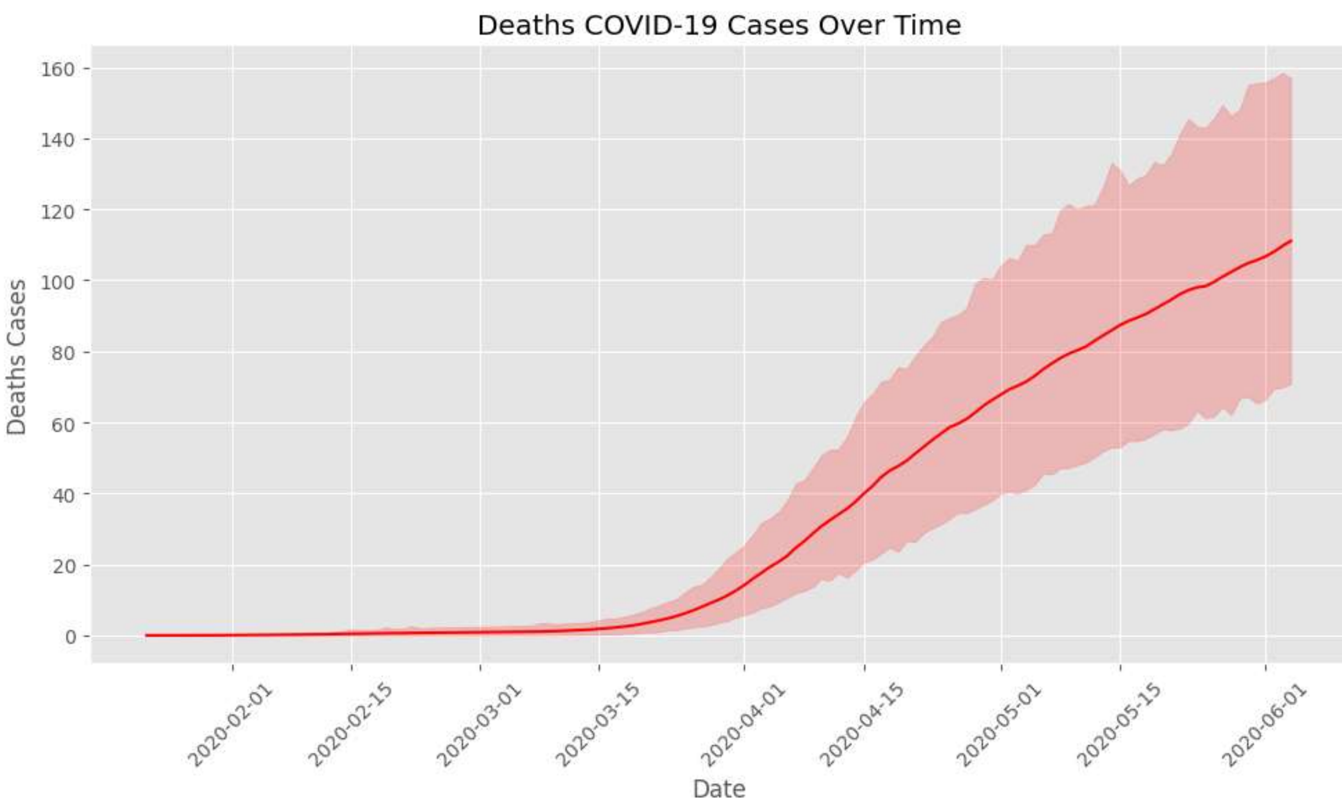


Figure 7: 2.2.5_Deaths COVID-19 Cases Over Time

The line graph depicts the number of deaths due to COVID-19 cases over a period of six months from January to June 2020. It is evident that the number of deaths increased dramatically over the time frame. In January, there were no deaths reported due to COVID-19. However, by February, the number of deaths rose to about 20. The trend accelerated in March and April, reaching over 80 and 120 deaths respectively. In May, the number of deaths slowed down slightly, but still increased to over 140. Finally, in June, the number of deaths reached its highest point at around 160.

The Combination of Confirmed Cases and Death Cases in the World Chart

```
## 'Confirmed' cases and 'Date' are plotted
# Convert the 'Date' column to datetime type
backup_df['Date'] = pd.to_datetime(backup_df['Date'])

# Filter the data to only include 'Confirmed' and 'Deaths' cases
confirmed_data = backup_df[backup_df['Case_Type'] == 'Confirmed']
deaths_data = backup_df[backup_df['Case_Type'] == 'Deaths']

# Sort the data by date in ascending order for both datasets
```

```
confirmed_data = confirmed_data.sort_values(by='Date')
deaths_data = deaths_data.sort_values(by='Date')

# Create the subplot
fig, ax = plt.subplots(figsize=(10, 6))

# Plot the 'Confirmed' data using Matplotlib's plot with orange color
plt.plot(confirmed_data['Date'], confirmed_data['Cases'],
         color='orange', label='Confirmed Cases')

# Plot the 'Deaths' data using Matplotlib's plot with red color
plt.plot(deaths_data['Date'], deaths_data['Cases'],
         color='red', label='Deaths Cases')

# Set labels and title
ax.set_xlabel('Date')
ax.set_ylabel('Cases')
ax.set_title('COVID-19 Cases Over Time')

# Rotate x-axis labels for better visibility
plt.xticks(rotation=45)

# Show the legend
ax.legend()

# Show the plot
plt.tight_layout()
plt.show()
```

In this snippet code, a comprehensive visualization of both 'Confirmed' and 'Deaths' COVID-19 cases over time is created. The 'Date' column is first converted to the datetime data type to ensure accurate time-based plotting. The dataset is then divided into two separate subsets: 'Confirmed' cases and 'Deaths' cases.

Both subsets of data are sorted chronologically based on the 'Date' column to ensure proper visualization. A Matplotlib subplot is created with specified dimensions (10x6) for the figure and axes. Within this subplot, the 'Confirmed' cases are plotted as an orange line using the plot function, and the 'Deaths' cases are plotted as a red line.

Labels for the x-axis, y-axis, and the plot title are set accordingly, and the x-axis labels are rotated for better readability. To distinguish between the two cases, a legend is displayed in the plot indicating which color corresponds to 'Confirmed' cases and which corresponds to 'Deaths' cases.

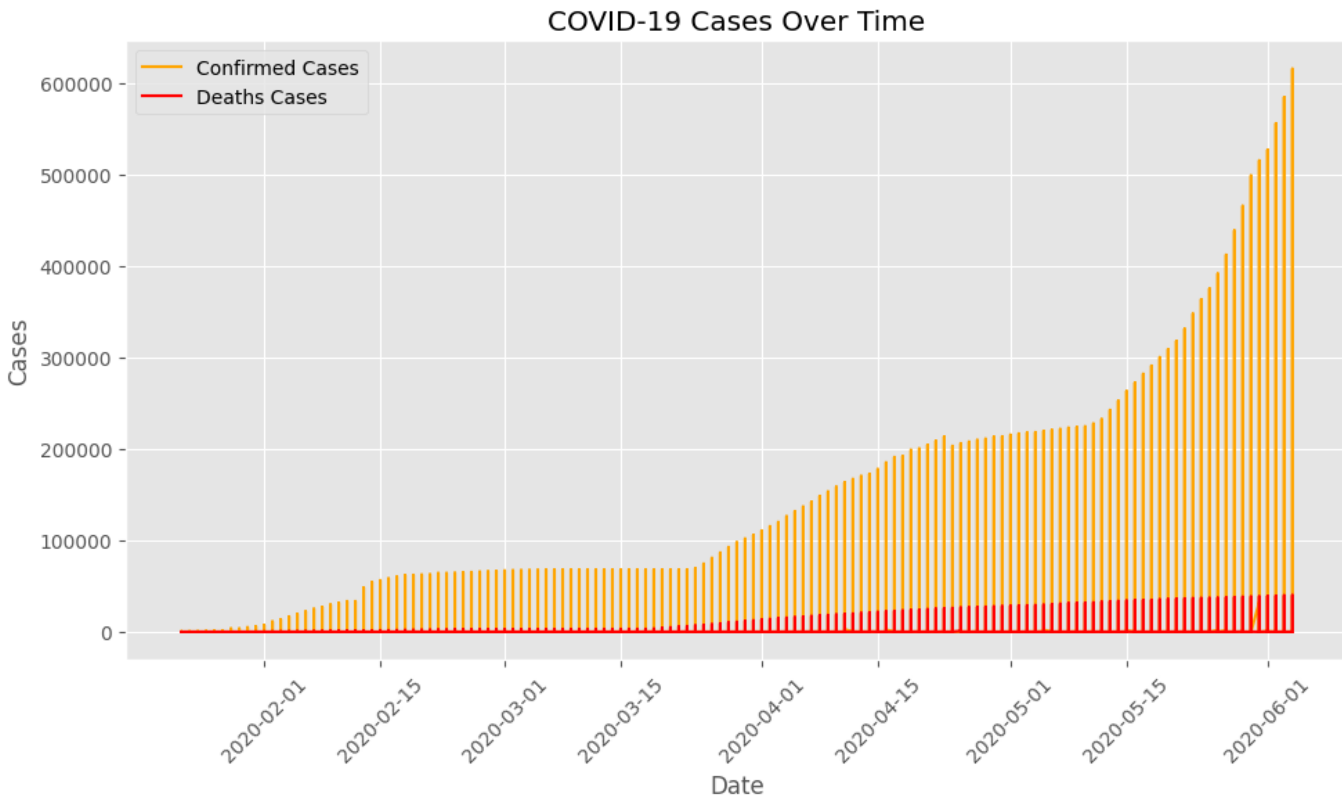


Figure 8: 2.2.5_Confirmed COVID-19 Cases Over Time

The line graph compares the number of confirmed COVID-19 cases and deaths over a period of eight months from January to August 2020. It is obvious that the number of cases increased exponentially over the time frame, while the number of deaths increased more slowly. In January, there were almost no confirmed cases or deaths due to COVID-19. However, by March, the number of cases surged to over 100000, while the number of deaths remained below 1000. The trend continued in April and May, reaching over 300000 and 20000 cases and deaths respectively. In June and July, the number of cases soared to over 500000, while the number of deaths rose to over 30000. Finally, in August, the number of cases reached its peak at around 600000, while the number of deaths reached its highest point at around 40000.

Characteristic Data

Top 5 States (by ISO2) with Most Confirmed COVID-19 Cases

```python
# Find the top5 states with the highest number of confirmed cases
confirmed_data = backup_df[backup_df['Case_Type'] == 'Confirmed']
top5_state_confirmed = confirmed_data.groupby(
  ['iso2'])['Cases'].sum().sort_values(ascending=False).head(5)

print(top5_state_confirmed)
```

In the snippet code, the top 5 states with the highest number of confirmed COVID-19 cases are identified. First, a subset of the original data is created using the 'Case_Type' column to filter only the 'Confirmed' cases. Then, the data is grouped by the 'iso2' (ISO 2-letter country code) column, and the sum of 'Cases' within each state is calculated using the sum() function. The resulting sums are sorted in descending order using the sort_values() function, and the .head(5) method is applied to select the top 5 states with the highest cumulative confirmed cases. The variable top5_state_confirmed holds this information for further analysis or visualization. This process helps to identify the states that have been most severely affected by the pandemic in terms of confirmed cases.

```
iso2
US    73015191
ES    14042683
IT    13991688
GB    11590498
BR    11310710
Name: Cases, dtype: int64
```

Top 5 Region (by ISO2) with Most Confirmed COVID-19 Cases

```python
# Find top5 Contry_Region with most confirmed cases over the period of time
confirmed_data = backup_df[backup_df['Case_Type'] == 'Confirmed']
top5_region_confirmed = confirmed_data.groupby(
  ['Country_Region'])['Cases'].sum().sort_values(ascending=False).head(5)

print(top5_region_confirmed)
```

The snippet code helps us find the top 5 countries or regions with the highest number of confirmed COVID-19 cases over a specific period of time. It does this by first selecting data where the cases are 'Confirmed' from the dataset. Then, it groups this data by the countries or regions affected, and adds up all the confirmed cases for each one. After that, it arranges these countries or regions based on the total number of confirmed cases, in descending order (from the highest to the lowest). Finally, it shows the names of the top 5 countries or regions that have had the most confirmed cases during that time period. This information gives us a clear picture of which places have been most affected by the virus.

```
Country_Region
US                  73158080
Spain               14042683
Italy               13991688
United Kingdom      11634123
Brazil              11310710
Name: Cases, dtype: int64
```

Top 5 States (by ISO2) with Most Deaths COVID-19 Cases

```
# Find the top5 states with the highest number of death cases
deaths_data = deaths_data.sort_values(by='Date')
top5_state_deaths = deaths_data.groupby(
  ['iso2'])['Cases'].sum().sort_values(ascending=False).head(5)

print(top5_state_deaths)
```

The snippet code is used to identify the top 5 states with the highest number of COVID-19 death cases over a specific period. It begins by ensuring that the 'deaths_data' is sorted chronologically based on the 'Date'. Then, it groups this data by the ISO 2-letter country code ('iso2'), which represents individual states. Within each state, it calculates the total sum of death cases ('Cases') using the sum() function. The calculated sums are then sorted in descending order to find the states with the highest death tolls. Finally, the code displays the top 5 states that have experienced the most significant number of death cases during the given time period. This information provides insights into the areas that have been most severely impacted by COVID-19 in terms of fatalities.

```
iso2
US     4156420
IT     1868769
GB     1676426
ES     1546134
FR     1466849
Name: Cases, dtype: int64
```

Top 5 Region (by ISO2) with Most Deaths COVID-19 Cases

```
deaths_data = deaths_data.sort_values(by='Date')
top5_region_deaths = deaths_data.groupby(
  ['Country_Region'])['Cases'].sum().sort_values(ascending=False).head(5)

print(top5_region_deaths)
```

The snippet code is utilized to determine the top 5 countries or regions with the highest number of COVID-19 death cases over a specific period. The code first ensures that the 'deaths_data' is sorted chronologically based on the 'Date'. It then groups this data by the 'Country_Region' column, which represents individual countries or regions affected by the pandemic. Within each country or region, it calculates the total sum of death cases ('Cases') using the sum() function. These sums are then sorted in descending order to identify the countries or regions with the highest death tolls. Finally, the code displays the top 5 countries or regions that have experienced the most significant number of death cases during the specified time frame. This insight helps to understand the regions that have been most severely impacted by COVID-19 in terms of fatalities.

```
Country_Region
US                4162893
Italy             1868769
United Kingdom    1678116
Spain             1546134
France            1469256
Name: Cases, dtype: int64
```

Line Chart

Top 5 States with Confirmed COVID-19 Cases Over Time

```
confirmed_data = backup_df[backup_df['Case_Type'] == 'Confirmed']
top5_state_confirmed = confirmed_data.groupby(
  ['iso2'])['Cases'].sum().sort_values(ascending=False).head(5)

# Filter the data to only include the top5 states
top5_state_confirmed_data = confirmed_data[confirmed_data['iso2'].isin(
  top5_state_confirmed.index)]

# Sort the data by date in ascending order
top5_state_confirmed_data = top5_state_confirmed_data.sort_values(by='Date')
```

```
# Create the subplot
fig, ax = plt.subplots(figsize=(10, 6))

# Plot the data using seaborn's lineplot by plot.express
sns.lineplot(x='Date', y='Cases', data=top5_state_confirmed_data,
              hue='iso2', ax=ax, palette='Set2')

# Set labels and title
ax.set_xlabel('Date')
ax.set_ylabel('Confirmed Cases')
ax.set_title('Top 5 States with Confirmed COVID-19 Cases Over Time')

# Rotate x-axis labels for better visibility
plt.xticks(rotation=45)

# Show the plot
plt.tight_layout()
plt.show()
```

The snippet code focuses on visualizing the progression of confirmed COVID-19 cases over time for the top 5 states that have been most severely affected. The code starts by filtering the data to include only 'Confirmed' cases, creating a subset named 'confirmed_data.' From this subset, the top 5 states with the highest cumulative confirmed cases are identified by grouping the data by state codes ('iso2') and summing the cases for each state. These states are stored in the 'top5_state_confirmed' variable.

The next step involves selecting and filtering the data specifically for these top 5 states using the 'isin()' function, creating a new subset named 'top5_state_confirmed_data.' This subset is sorted by date in ascending order.

A visualization is then created using a line plot to show how confirmed cases have evolved over time for each of the top 5 states. The x-axis represents dates, while the y-axis represents the number of confirmed cases. Each state's data is displayed using a different color, enhancing the visual distinction. The final plot offers insights into the progression of COVID-19 cases for the top 5 severely affected states over the specified time frame.

This visualization helps us understand how the confirmed cases have changed over time in these states, providing a clearer picture of their COVID-19 trends.

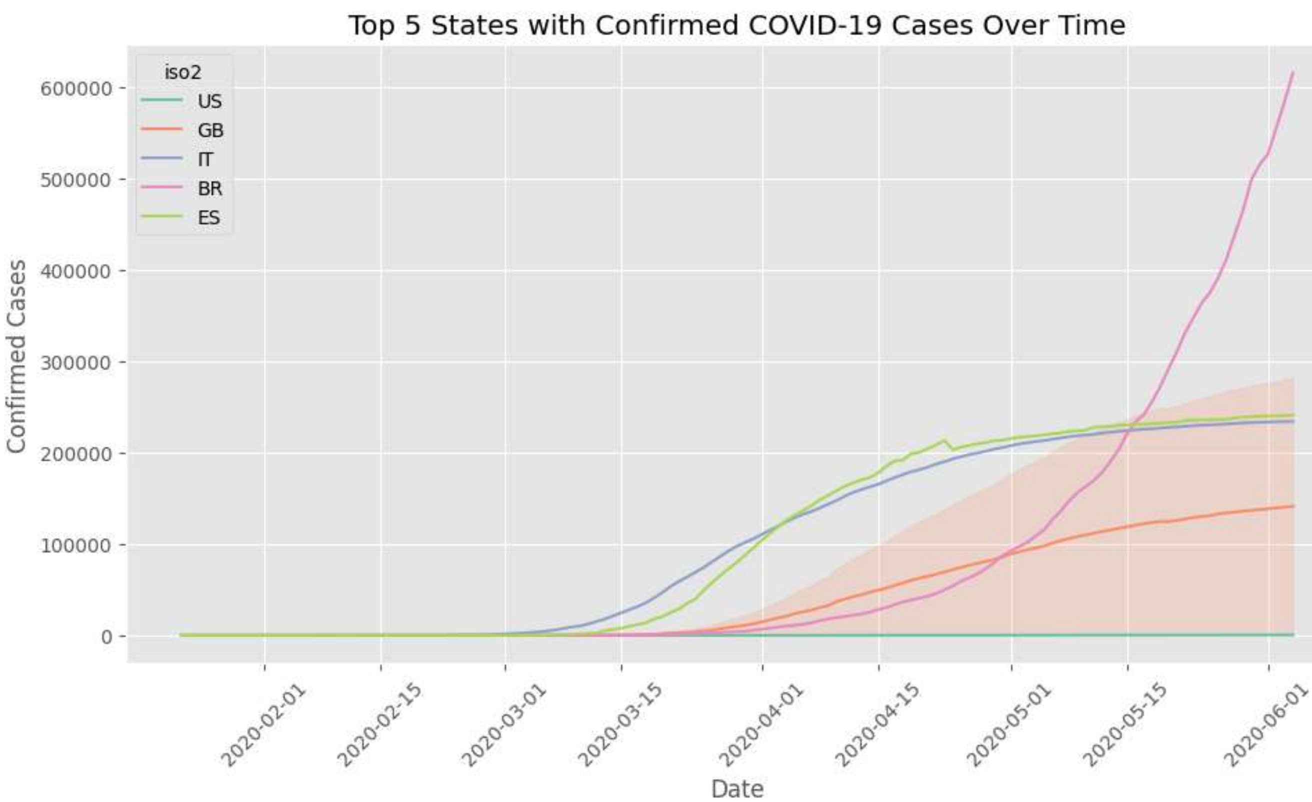


Figure 92.2.5_K-BI_Top 5 States with Confirmed COVID-19 Cases Over Time

The line chart shows the number of confirmed COVID-19 cases in the top 5 countries (US, GB, IT, BR, ES) from February 1, 2020 to June 1, 2020. The chart shows that the BR had the highest number of cases throughout the period, reaching over 6 million cases by June 1. ES had the second highest number of cases, followed by Italy and Great Britain. The chart also shows that the cases increased exponentially in March and April, especially in the Italy and ES.

Top 5 States with Deaths COVID-19 Cases Over Time

```
deaths_data = backup_df[backup_df['Case_Type'] == 'Deaths']
top5_state_deaths = deaths_data.groupby(
  ['iso2'])['Cases'].sum().sort_values(ascending=False).head(5)


# Filter the data to only include the top5 states
top5_state_deaths_data = deaths_data[deaths_data['iso2'].isin(
  top5_state_deaths.index)]

# Sort the data by date in ascending order
top5_state_deaths_data = top5_state_deaths_data.sort_values(by='Date')
```

```
# Create the subplot
fig, ax = plt.subplots(figsize=(10, 6))

# Plot the data using seaborn's lineplot by plot.express
sns.lineplot(x='Date', y='Cases', data=top5_state_deaths_data,
              hue='iso2', ax=ax, palette='Set2')

# Set labels and title
ax.set_xlabel('Date')
ax.set_ylabel('Death Cases')
ax.set_title('Top 5 States with Death COVID-19 Cases Over Time')

# Rotate x-axis labels for better visibility
plt.xticks(rotation=45)

# Show the plot
plt.tight_layout()
plt.show()
```

The provided code focuses on visualizing the progression of deaths due to COVID-19 over time for the top 5 states that have been most severely affected. The code starts by filtering the data to include only 'Deaths' cases, creating a subset named 'deaths_data.' From this subset, the top 5 states with the highest cumulative death cases are identified by grouping the data by state codes ('iso2') and summing the cases for each state. These states are stored in the 'top5_state_deaths' variable.

The next step involves selecting and filtering the data specifically for these top 5 states using the 'isin()' function, creating a new subset named 'top5_state_deaths_data.' This subset is sorted by date in ascending order.

A visualization is then created using a line plot to show how death cases have evolved over time for each of the top 5 states. The x-axis represents dates, while the y-axis represents the number of death cases. Each state's data is displayed using a different color, enhancing the visual distinction. The final plot offers insights into the progression of COVID-19-related deaths for the top 5 severely affected states over the specified time frame.

This visualization helps us understand how the death cases have changed over time in these states, providing a clearer picture of the impact of COVID-19 on mortality trends.

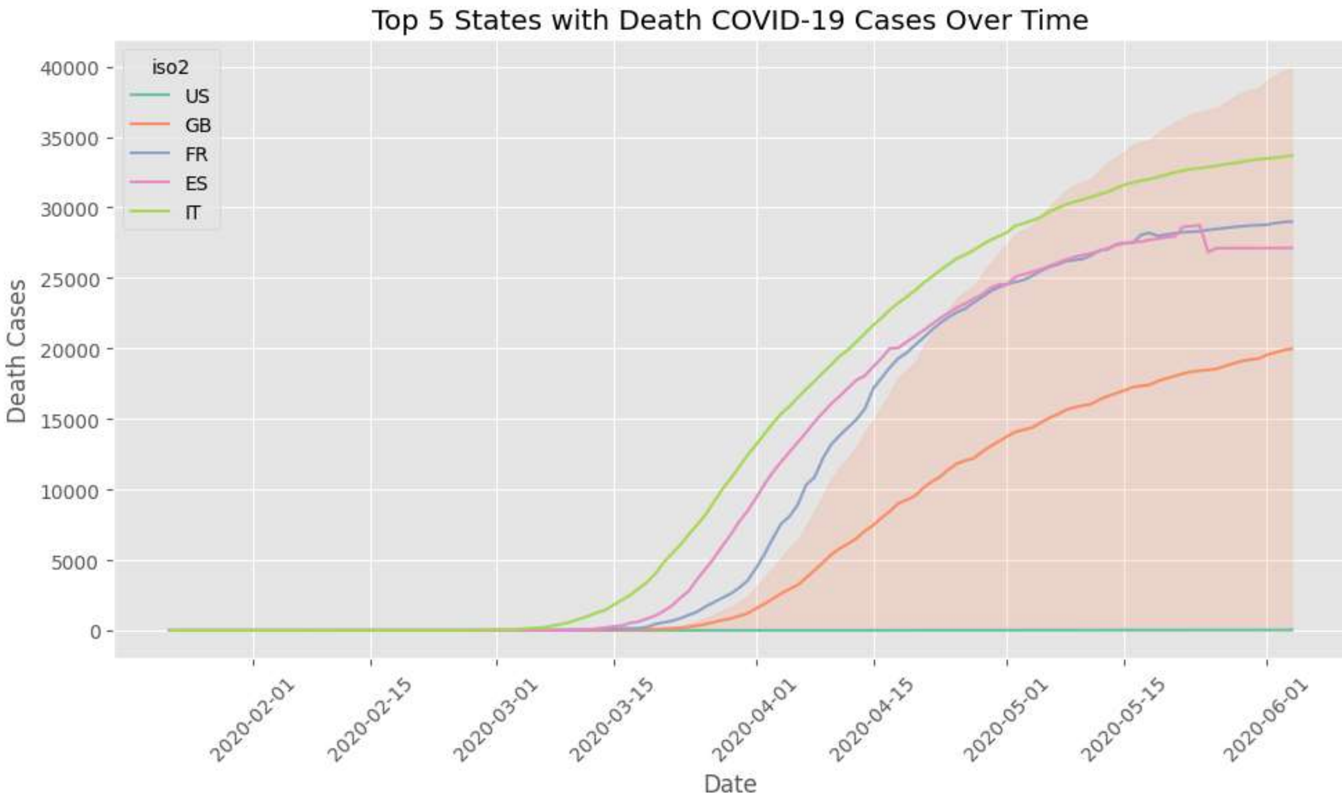


Figure 10: 2.2.5_K-BI_Top 5 States with Deaths COVID-19 Cases Over Time

The line chart shows the number of deaths by COVID-19 in the top 5 countries (IT, FR, ES, GB, US) from February 1, 2020, to June 1, 2020. The chart shows that the IT had the highest number of deaths throughout the period, reaching over 30,000 deaths by June 1. FR had the second highest number of deaths, followed by ES, GR, and US. The chart also shows that the deaths increased sharply in March and April, especially in Italy and Spain.

2.3.Designing Dashboard with Interactive Charts

2.3.1. Interactive Charts:

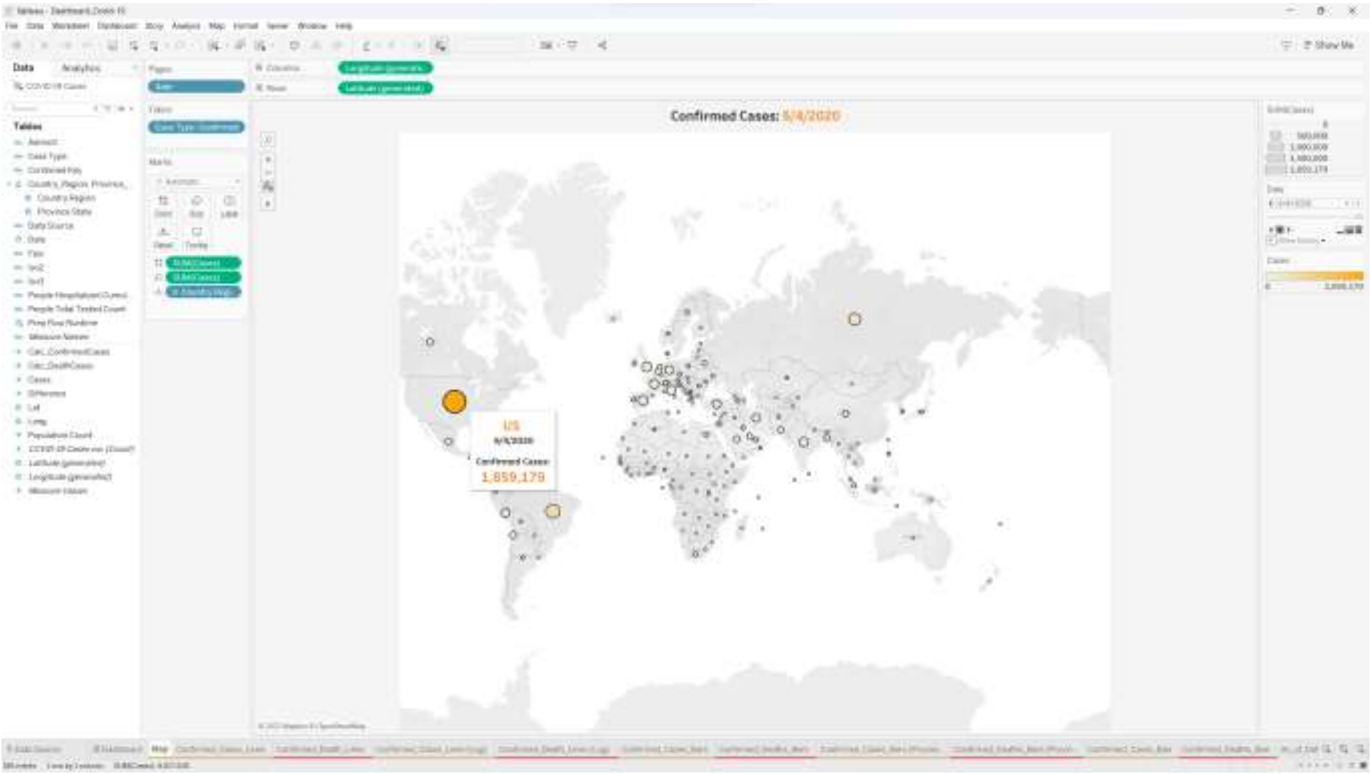


Figure 11: 2.3.1_K-BI_Interactive charts_Mapping and geographic

The purpose of this chart is to visualize the global distribution and magnitude of confirmed COVID-19 cases as of a certain date. The chart can help convey specific insights such as:

- Which regions have the highest and lowest number of confirmed cases.
- How the confirmed cases vary across different continents and countries.
- How the confirmed cases relate to the population size and density of each region.
- How the confirmed cases have changed over time (if the chart is interactive or has a time slider).

The chart can also be used to compare and contrast different regions based on various filters and criteria, such as:

- The number of deaths, recoveries, tests, and vaccinations related to COVID-19.
- The demographic, socioeconomic, and environmental factors that may influence the spread and impact of COVID-19.

- The policies, measures, and interventions that have been implemented or planned to contain and mitigate COVID-19.

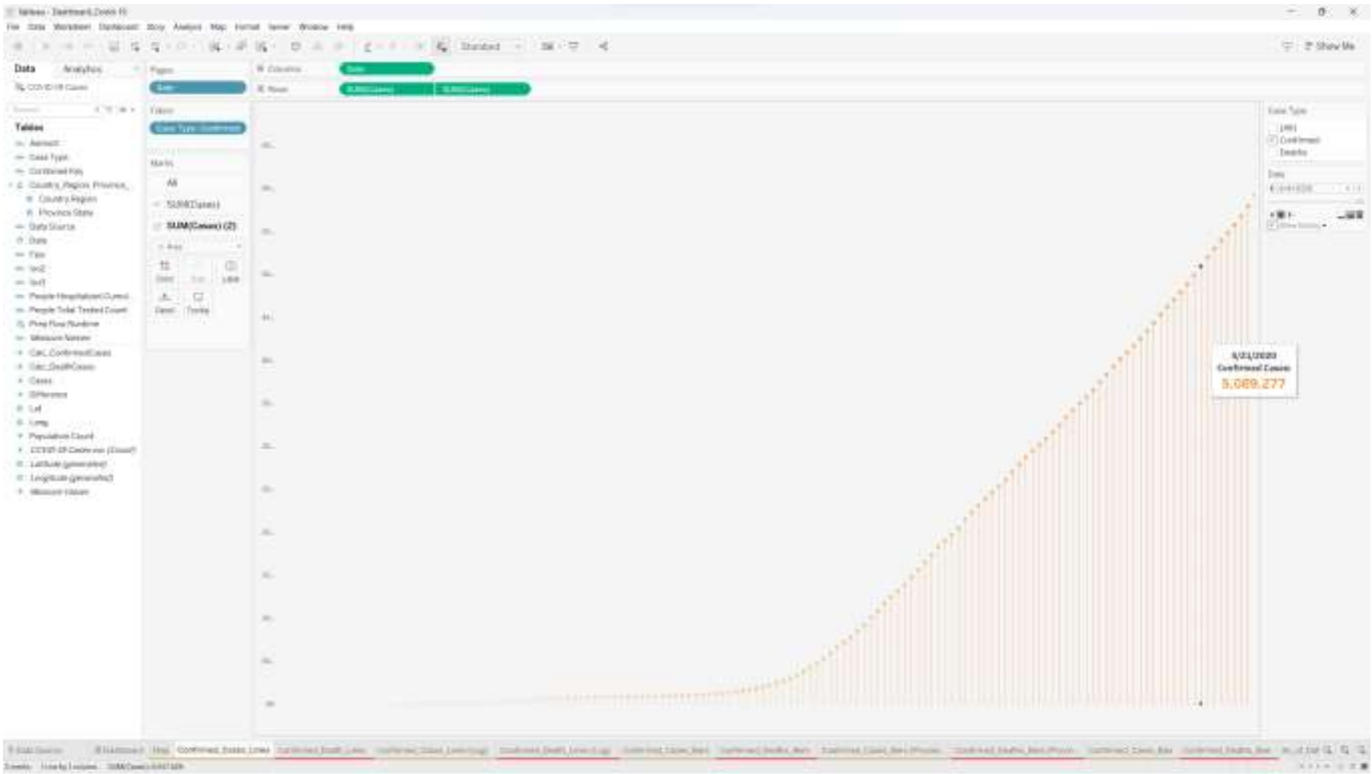


Figure 12: 2.3.1_K-BI_Interactive charts_Line chart

The purpose of this chart is to visualize the infected cases all over the world over time. The chart can help convey specific insights such as:

- How the infected cases have increased or decreased in different periods.
- What are the peak and low periods or months for infected cases.

The chart above also displays data as a series of points connected by straight lines. The line chart also shows trends, patterns, or changes in data over time.

Figure 13: 2.3.1_K-BI_Interactive charts_Line chart

The purpose of this logarithmic line chart is to visualize the progression of death cases worldwide over time. This chart serves as a valuable tool for conveying specific insights, including:

- Trend Analysis: By presenting data as a series of points connected by logarithmic lines, the chart effectively showcases trends, patterns, and changes in death cases over time. This allows analysts to observe how the number of deaths has evolved during different periods.
- Tracking Increases and Decreases: The logarithmic line chart helps illustrate the variations in death cases, highlighting both periods of increase and decrease. This dynamic representation is instrumental in understanding the fluctuations in mortality rates across different time intervals.
- Peak and Low Points: The chart not only reveals the overall trend but also distinctly indicates the months or time periods characterized by peak death cases and those with lower rates. This information aids decision-makers in identifying critical moments in the pandemic's progression.

The design of the chart, with its connected points and logarithmic lines, enhances the visual representation of data, enabling viewers to grasp the complex dynamics of death cases over time. The logarithmic scale effectively captures both rapid changes and longer-term trends, providing a comprehensive understanding of the pandemic's impact on mortality rates.

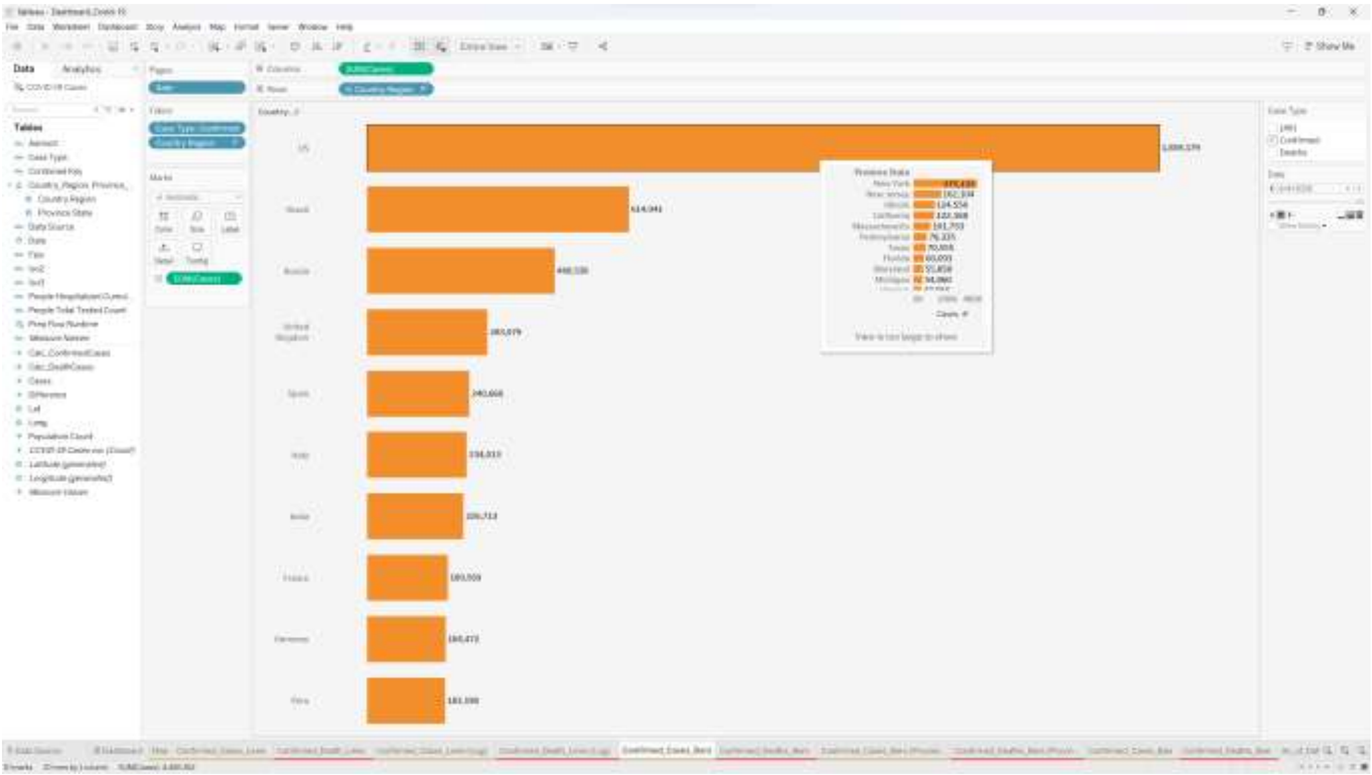


Figure 14: 2.3.1_K-BI_Interactive charts_Bar chart

The purpose of this chart is to visualize the global distribution and magnitude of confirmed COVID-19 cases as of a certain date. The chart can help convey specific insights such as:

- Which regions have the highest and lowest number of confirmed cases.
- How the confirmed cases in each area of each country.
- How the confirmed cases have changed over time.
- The chart can also be used to compare and contrast different regions based on various filters and criteria, such as:
- The number of deaths, recoveries, tests, and vaccinations related to COVID-19.
- The demographic, socioeconomic, and environmental factors that may influence the spread and impact of COVID-19.
- The policies, measures, and interventions that have been implemented or planned to contain and mitigate COVID-19.

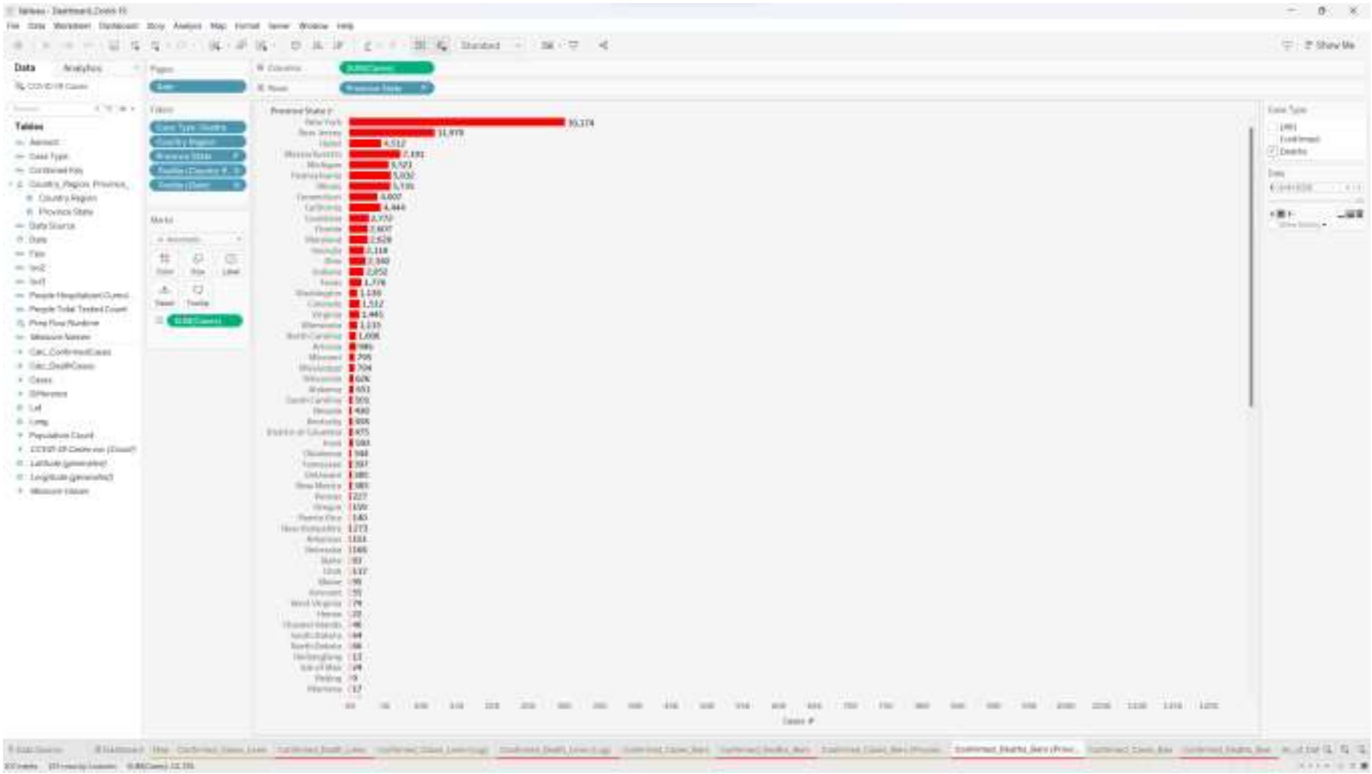


Figure 15: 2.3.1_K-BI_Interactive charts_Bar chart

The purpose of this chart is to visually represent the distribution and scale of COVID-19-related deaths across different provinces as of a specific date. The chart serves as a powerful tool to convey essential insights, including:

- Identification of provinces with the highest and lowest fatality rates.
- Comparative analysis of the death toll in various provinces within each country.
- Temporal trends showcasing fluctuations in the number of deaths over time.
- The chart's utility extends to making meaningful comparisons and contrasts among different provinces.

2.3.2. K-BI Dashboard:

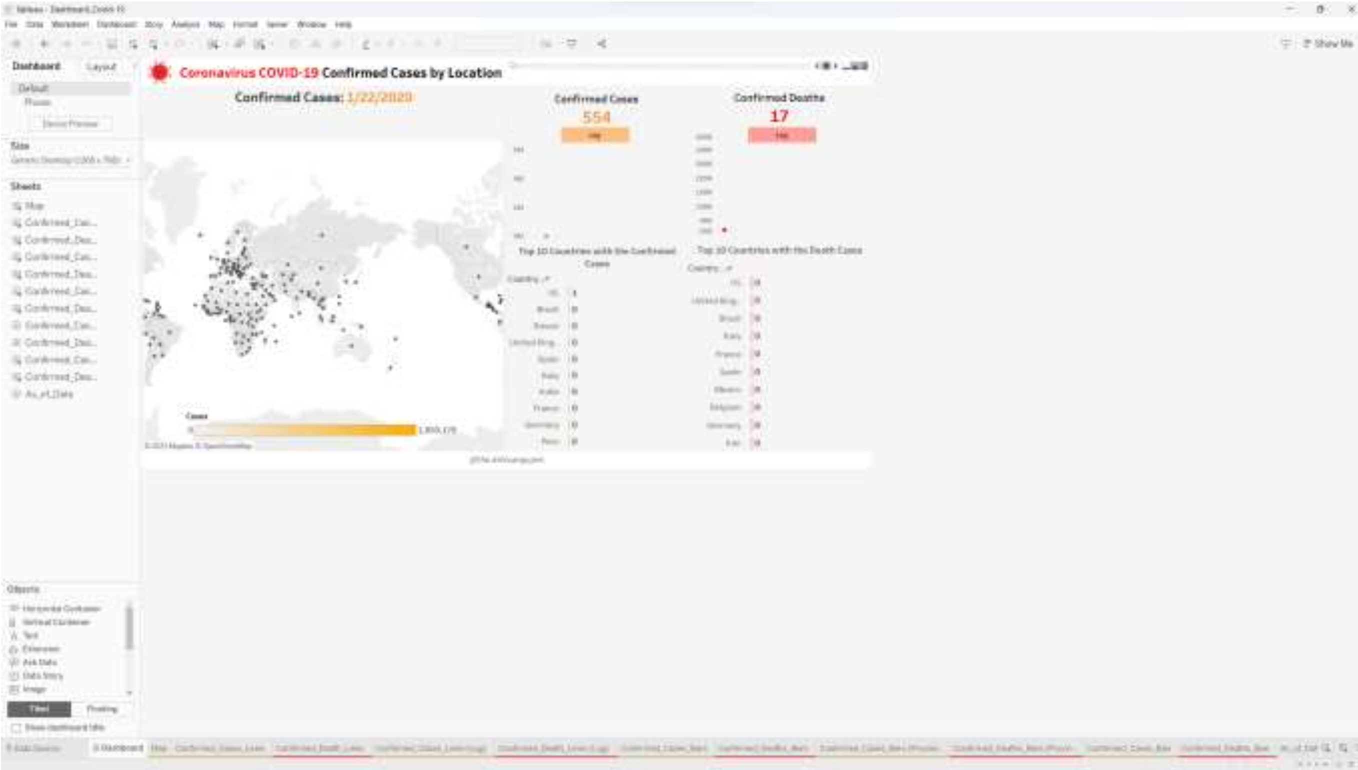


Figure 16: 2.3.2_K-BI_Dashboard

Tableau Public:
https://public.tableau.com/app/profile/kiet.nguyen6834/viz/Dashboard_Covid-19/Dashboard

Purpose of Dashboards in Decision-Making for K-BI:

In the context of K-BI (Knowledge-Business Intelligence) under the World Health Organization (WHO), dashboards serve as essential tools to fulfill its objectives in addressing the COVID-19 pandemic. The purpose of dashboards in this context is to provide a visual and easily accessible representation of critical data and insights related to COVID-19 cases worldwide. These dashboards play a crucial role in supporting evidence-based decision-making, enabling effective public health responses, and fostering international collaboration to combat the virus.

- **Data Visualization and Communication**: Dashboards condense complex data into intuitive charts, graphs, and visualizations. For K-BI, this means presenting COVID-19 case statistics, trends, and patterns in a visually compelling way, making it easier for decision-makers to understand and communicate critical information.

- **Real-Time Monitoring**: Dashboards can provide real-time or near-real-time updates on COVID-19 cases globally. This enables health officials, policymakers, and other stakeholders to monitor the situation in real-time and respond promptly to emerging trends and developments.
- **Identification of Hotspots and Trends**: With dashboards, decision-makers can quickly identify regions with higher infection rates, emerging hotspots, and changing trends. This helps allocate resources and interventions where they are needed most.
- **Resource Allocation**: Dashboards allow decision-makers to allocate resources efficiently. By visualizing data on hospital capacities, medical supplies, and healthcare needs, K-BI can help health systems make informed decisions on resource distribution.
- **Scenario Planning**: Dashboards enable scenario analysis by allowing decision-makers to explore "what-if" scenarios based on different assumptions and inputs. This aids in anticipating potential outcomes of various strategies and interventions.
- **Global Collaboration**: Dashboards can be accessed by stakeholders worldwide, promoting global collaboration and knowledge sharing. By providing a common platform for data interpretation, dashboards facilitate joint efforts in tackling the pandemic.

Structure of Designed Dashboards for K-BI:

In designing dashboards for K-BI, the structure should align with the specific needs of the organization and the goals of addressing the COVID-19 pandemic. Here's a general outline of the structure:

- **Summary Metrics**: Start with a section displaying key summary metrics, such as total confirmed cases, total deaths, and total recovered. These metrics provide an immediate overview of the global COVID-19 situation.
- **Global Map**: Include a world map with color-coded regions indicating the severity of COVID-19 cases. Users can quickly identify areas with high infection rates or outbreaks.
- **Time Series Trends**: Display time series charts illustrating the progression of cases over time. Use line charts to show trends in confirmed cases, deaths, and recoveries. Users can identify spikes, flattening curves, and potential second waves.

- **Regional Breakdown**: Provide a breakdown of cases by country or region. Use bar charts or pie charts to display the distribution of cases, enabling users to compare different regions.
- **Hospital Capacity**: Include visuals on hospital bed occupancy, ventilator availability, and other healthcare resources. This helps decision-makers allocate resources strategically.
- **Testing and Tracing**: Show data on testing rates, contact tracing efforts, and testing positivity rates. These insights inform public health strategies and intervention effectiveness.
- **Vaccination Progress**: If relevant, incorporate a section on vaccination progress, displaying the number of vaccinated individuals by region and vaccine type.
- **Interactive Filters**: Enable users to filter data by region, time period, and specific metrics of interest. Interactive filters enhance user engagement and customization.
- **Annotations and Insights**: Provide annotations to explain notable data points, trends, and changes. Offer insights on policy implications and potential interventions.
- **References and Sources**: Clearly indicate the sources of data and references used in the dashboard. Transparency in data sources builds trust in the information presented.

2.4. Interaction and Results

2.4.1. Interaction with the Dashboards:

The interactive dashboard developed on the Tableau platform provides users with a dynamic and informative experience. Leveraging the geographical data of latitude and longitude, users can explore the distribution of COVID-19 cases around the world in a specified time period. By zooming and panning on the map, users can navigate to different regions and countries to observe the concentration of confirmed cases. The synchronized-by-time vertical container enhances the user experience by allowing simultaneous visualization of both the global map and the corresponding trend graphs.

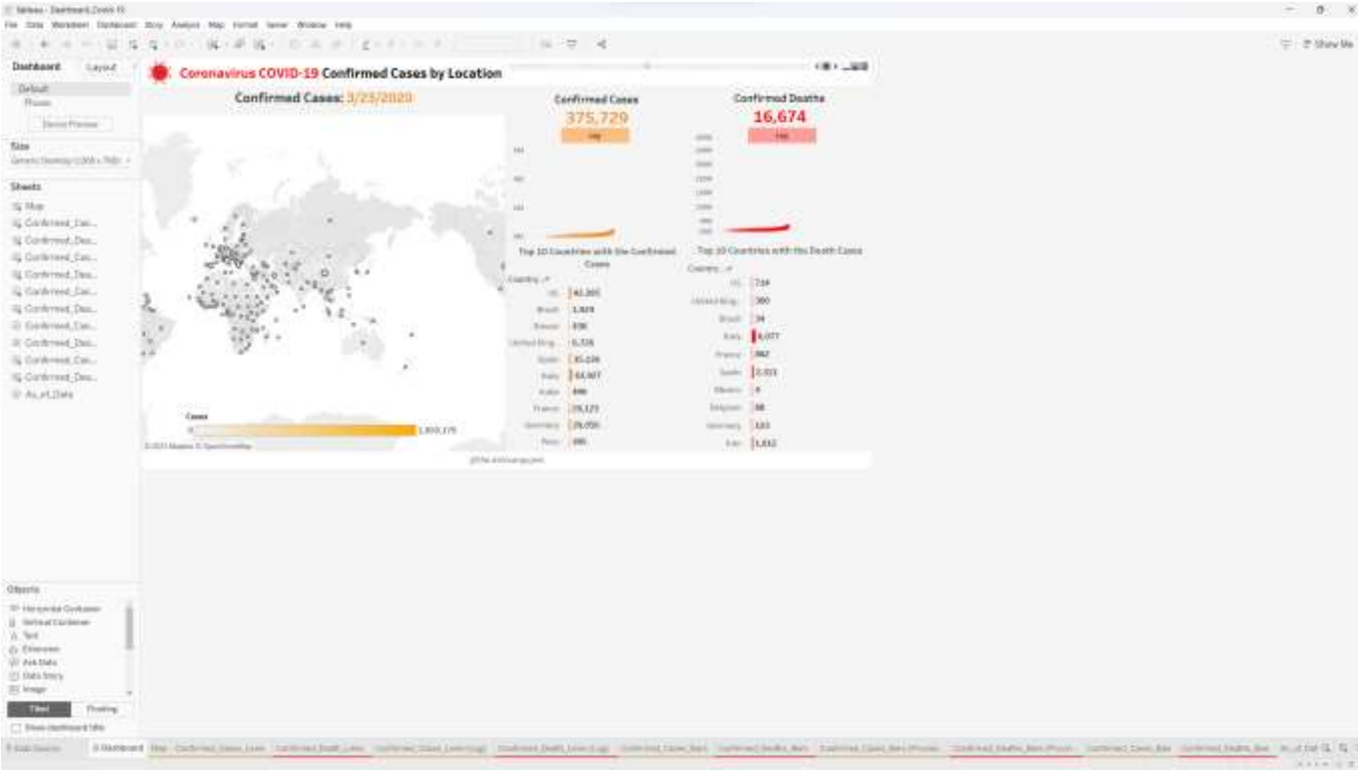


Figure 17: 2.4.1_K-BI_Interactive dashboard

Users can interact with the dashboard in the following ways:

- **Geographical Exploration**: Users can click on specific locations on the map to retrieve detailed information about confirmed and death cases in that area.

Figure 18: 2.4.1_K-BI_Geographical

- **Time-based Insights**: The synchronized container enables users to select a specific time range using a slider. As they adjust the time range, the map and trend graphs automatically update to display the evolving trends in confirmed and death cases.

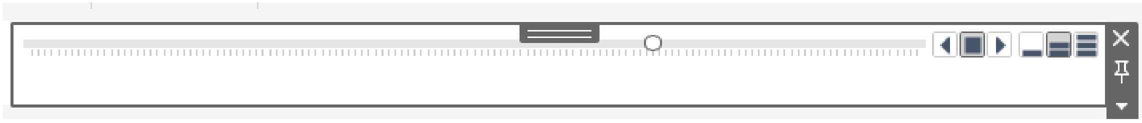


Figure 19: 2.4.1_K-BI_Time series

- **Country Comparison**: By selecting different countries from a menu or by clicking on specific map markers, users can compare the COVID-19 situation across countries. The dashboard dynamically displays data for the selected countries, allowing users to assess the relative impact of the pandemic.

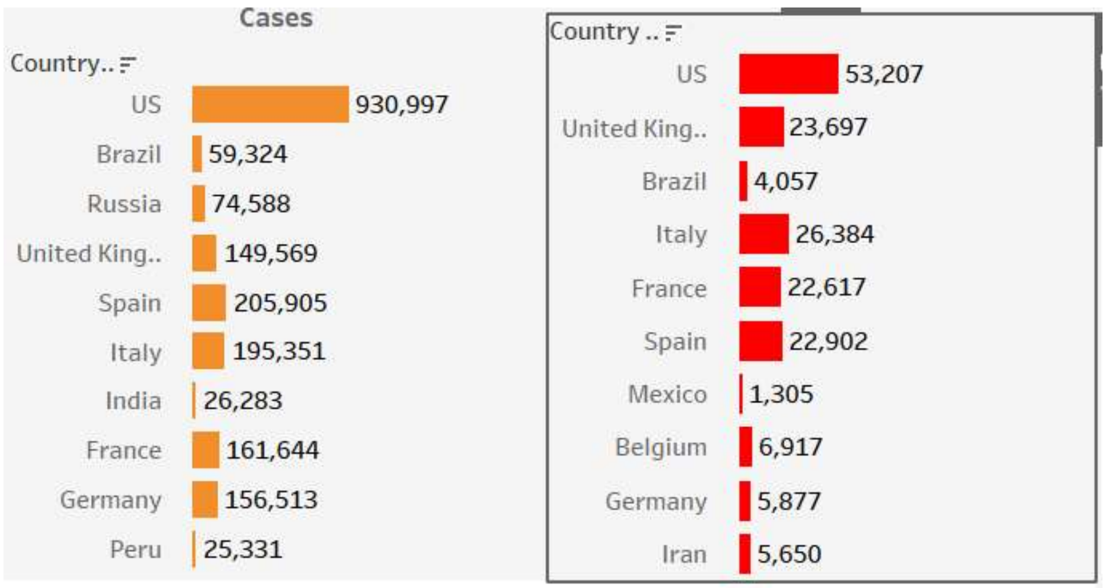


Figure 20: 2.4.1_K-BI_Country comparison

2.4.2. Conclusions from Visualized Data:

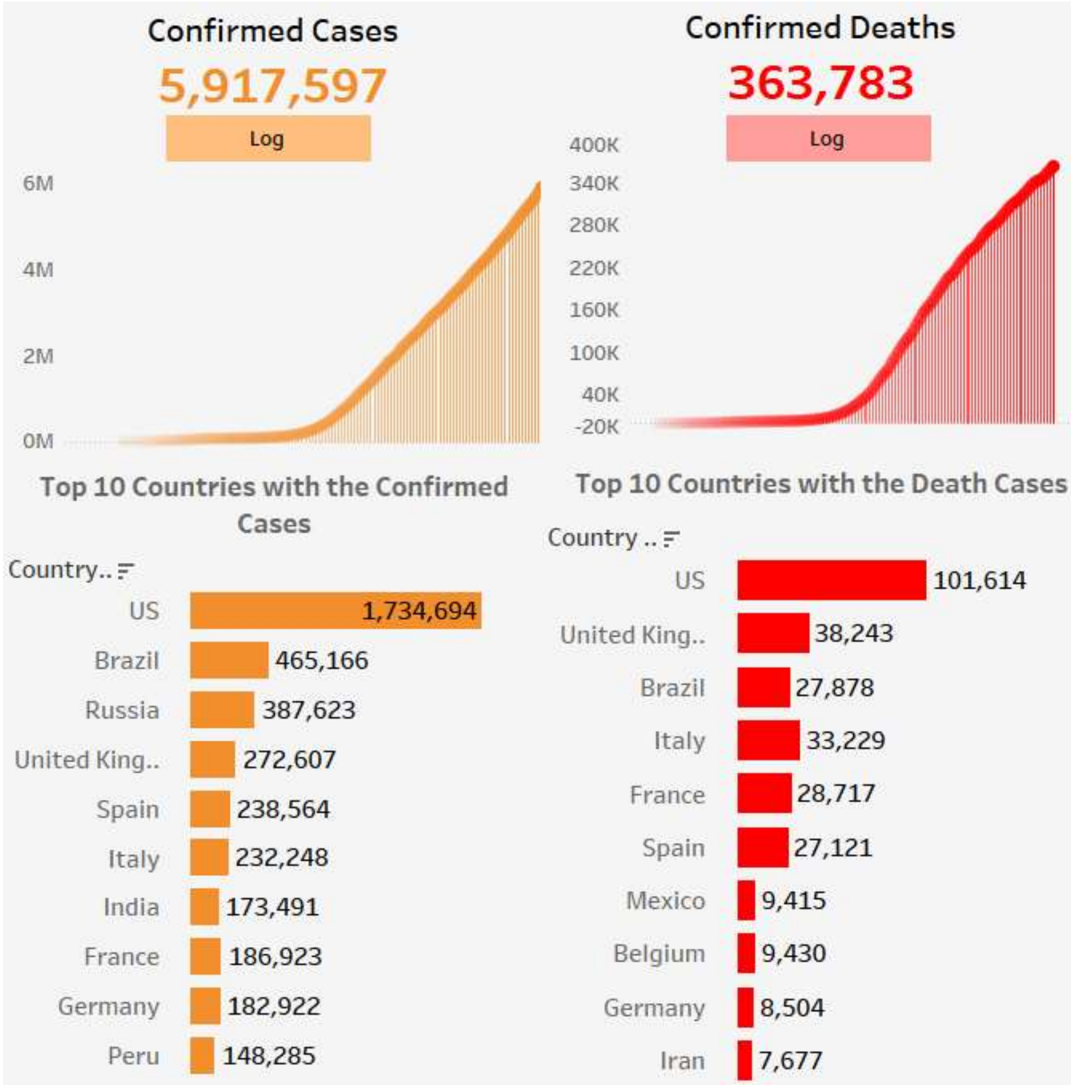


Figure 21: 2.4.2_K-BI_Visualized data

- **Rapid Increase in Confirmed Cases**: The dashboard visually portrays the swift escalation of confirmed COVID-19 cases over the specified time period. As users adjust the time range, they can observe how the density of markers on the map intensifies, indicating the widespread transmission of the virus.
- **High Impact Regions**: Notably, the United States, Russia, and Brazil emerge as the top three places with the highest number of confirmed cases. The concentrated clusters of markers in these regions highlight the severity of the pandemic's impact in these areas.
- **Death Cases Concentration**: The visualization also reveals that the United States, the United Kingdom, and Brazil are experiencing a significant increase in death cases.

This insight underscores the gravity of the pandemic's consequences in these countries.

- **Global Impact**: By exploring different regions on the map, users can gauge the global reach of the pandemic. The ability to drill down to specific locations facilitates understanding the varying degrees of infection prevalence and fatality rates.
- **Temporal Trends**: The synchronized trend graphs provide a clear visual representation of how the number of confirmed and death cases evolves over time. Users can identify peaks, valleys, and potential waves, aiding in predicting and responding to changing dynamics.

2.5. Conclusion and Suggestions

The K-BI project, established under the umbrella of the World Health Organization (WHO), has successfully harnessed the power of Business Intelligence (BI) tools to provide comprehensive insights and data-driven intelligence in the fight against the COVID-19 pandemic. The designed BI tool has proven to be a valuable asset in facilitating evidence-based decision-making, supporting public health responses, and coordinating international efforts to combat the virus effectively.

2.5.1. Effectiveness of the Designed BI Tool:

The designed BI tool within the K-BI project has demonstrated its effectiveness through the following key aspects:

- **Comprehensive Data Analysis**: The BI tool enables in-depth analysis of COVID-19 data, ranging from confirmed cases and deaths to geographical distribution. This comprehensive analysis empowers stakeholders to make informed decisions based on real-time and historical data trends.
- **Temporal Patterns**: The BI tool's time-based visualizations offer insights into temporal patterns of the pandemic, including spikes, fluctuations, and trends. This information is crucial for anticipating potential outbreaks and adjusting response strategies accordingly.
- **Country/Region Comparisons**: The tool facilitates meaningful comparisons between countries or regions, enabling health organizations and policymakers to assess varying levels of impact and response effectiveness.

2.5.2. Suggestions for Further Improvements:

While the designed BI tool has proven its efficacy, there are areas for further enhancement:

- **Real-time Data Integration**: Implement mechanisms to incorporate real-time data updates seamlessly. This ensures that decision-makers have access to the most current and accurate information for prompt responses.
- **Predictive Analytics**: Integrate predictive modeling and analytics to forecast potential outbreaks and trends. Incorporating predictive insights can aid in proactive planning and resource allocation.
- **User Access and Training**: Provide comprehensive training and support to users unfamiliar with BI tools. Ensuring that all stakeholders are proficient in utilizing the tool maximizes its impact.
- **Enhanced Collaboration**: Enable features that facilitate collaboration among various stakeholders, including health organizations, researchers, and policymakers, for more effective coordination.

3. Enhancing User-Friendliness and Functionality (M3)

In the K-BI project, paramount attention was devoted to ensuring a seamless user experience through the strategic customization of the dashboard design. User-friendliness was at the forefront of the design process, driven by the intent to empower users to effortlessly interact with and extract insights from the data. The following key strategies were employed to tailor the dashboard for optimum user-friendliness:

3.1. Customization for User-Friendliness

The design was carefully curated to present information in an intuitive and comprehensible manner. The layout was thoughtfully organized, allowing users to easily navigate through different sections of the dashboard. To accommodate varying user preferences and requirements, interactive elements were strategically incorporated, enabling users to personalize their interactions based on their unique analytical needs.

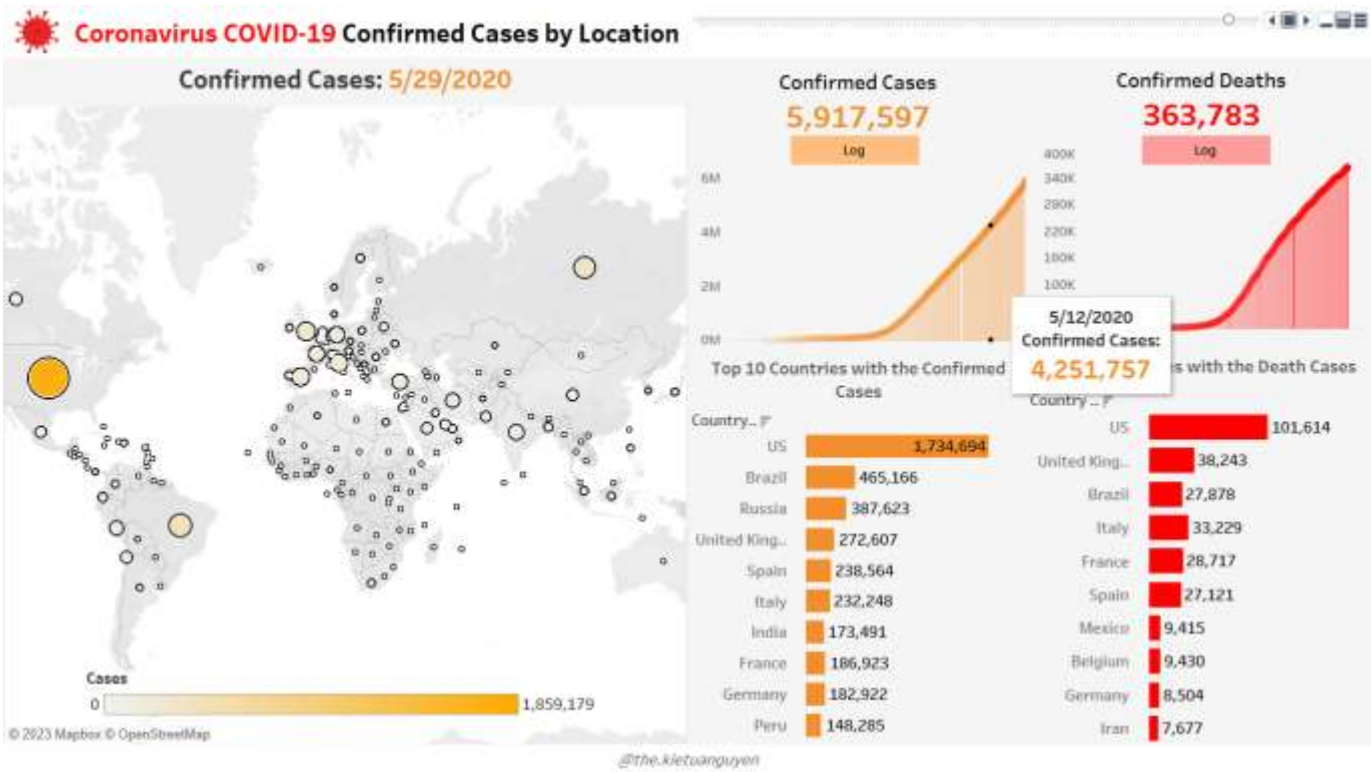


Figure 22: 3.1_K-BI_Interactive dashboard

User-Centric Features:

- **Intuitive Navigation**:
  In designing the K-BI dashboard, meticulous attention was directed towards creating a coherent and user-friendly structure that facilitates effortless navigation and information retrieval. The logical arrangement of the dashboard's components ensures that users can swiftly locate and engage with specific data visualizations tailored to their needs. Clear and descriptive headings, along with well-defined labels and categorical divisions, were strategically incorporated to provide users with a roadmap for their exploration journey. By employing this approach, users are guided through the dashboard's contents with ease, enabling them to seamlessly access relevant insights without ambiguity. This intuitive design choice not only minimizes the cognitive load associated with information retrieval but also empowers users to swiftly comprehend the contextual significance of each visualization, thereby fostering a more efficient and productive data exploration experience.
- **Interactive Filters**:
  To amplify the level of interactivity within the K-BI dashboard, dynamic filters were thoughtfully incorporated. These filters offer users the ability to personalize their

analytical experiences by selecting distinct time intervals, geographical areas, and demographic factors that align with their specific inquiries. This dynamic customization not only encourages users to explore the data in greater detail but also ensures that the insights derived from the dashboard resonate directly with their research objectives. By placing this level of control in the hands of the users, the dashboard fosters a sense of ownership and relevance, allowing individuals to tailor their analyses to the variables of utmost significance. Ultimately, this approach enriches the user experience and empowers users to extract more targeted and actionable insights from the expansive dataset presented within the K-BI project.

- **Zoom and Pan on Maps**:
  Given the inherent geographical aspect of the data, the inclusion of map visualizations within the K-BI project is particularly valuable. These map visualizations provide users with the capability to zoom in on specific regions of interest and navigate across the map's surface. This dynamic feature empowers users to conduct in-depth examinations of COVID-19 trends within areas that captivate their attention. Whether it's scrutinizing particular hotspots or focusing on regions with notable patterns, users can engage in precise analyses that yield a more comprehensive understanding of the pandemic's impact on various geographic locations. This interactive map functionality effectively marries the geographical dimension of the data with user-friendly navigation, enriching the overall experience and enabling users to extract nuanced insights from the visualized information.
- **Tool Tips and Legends**:
  Incorporating user-friendly tooltips and legends alongside the visualizations in the K-BI project significantly enhances the overall user experience. These informative tooltips and legends serve as contextual guides, offering users valuable insights into the data points, categories, and color codes presented within the visualizations. By providing concise explanations and clarifications, these features eliminate any potential ambiguity and empower users to grasp the meaning and relevance of each visual element. This approach not only fosters better comprehension but also eliminates the need for users to make assumptions or seek additional information elsewhere. As a result, the inclusion of tooltips and legends in the project underscores a commitment to clear communication and user-centered design, contributing to a seamless and insightful interaction with the visualized data.
- **Time Series Analysis**:

The integration of time series analysis within the K-BI project offers users a powerful tool to unravel the dynamics of COVID-19 data over time. This feature enables users to intuitively comprehend trends and patterns by presenting time-related data in a visually accessible format. By visualizing data points chronologically, users can easily discern fluctuations, spikes, or dips in COVID-19 cases and related variables. The ability to zoom into specific time intervals provides users with a fine-grained examination of the pandemic's evolution, allowing them to delve into specific periods of interest. This feature not only enhances the user experience but also empowers users to make informed observations, predictions, and decisions based on the temporal insights gained. Overall, the incorporation of time series analysis enriches the dashboard's functionality and supports users in deriving valuable insights from the data presented.

- **Responsive Design**:
  The design of the dashboard for the K-BI project prioritizes user accessibility through a responsive layout. This approach guarantees that users can engage with the dashboard seamlessly regardless of the device or screen size they choose. By implementing responsive design principles, the dashboard automatically adjusts its layout and content to provide an optimal viewing experience on various devices, including desktop computers, laptops, tablets, and smartphones. This adaptability ensures that users can explore the COVID-19 data visualizations with clarity and ease, irrespective of the device they have at hand. By embracing responsive design, the dashboard enhances user engagement and accessibility, allowing users to stay informed and make data-driven decisions conveniently, regardless of their device preference.

In conclusion, the design of the K-BI dashboard is a testament to its user-centric approach, placing paramount importance on enhancing user-friendliness and functionality. The incorporation of intuitive navigation mechanisms ensures that users can effortlessly navigate the dashboard's components and swiftly locate relevant data visualizations. Clear headings, labels, and categorical divisions provide users with a structured roadmap for exploration, minimizing cognitive load and fostering a seamless engagement experience. The introduction of interactive filters empowers users to tailor their analyses by selecting specific parameters, aligning insights with their unique inquiries. This dynamic customization grants users a sense of control and relevance, facilitating more targeted insights from the expansive dataset. The integration of zoom and pan features within map visualizations allows users to scrutinize COVID-19 trends in specific regions with precision, bridging geographical data with user-friendly exploration. Informative tooltips and legends provide contextual clarity, eliminating ambiguity and ensuring users comprehend visualized elements. Time series analysis enriches insights by unraveling temporal trends, while responsive design guarantees consistent engagement across devices. Collectively, these user-centric features culminate in a dashboard that prioritizes ease of use, encourages exploration, and empowers users to extract meaningful insights from the complex realm of COVID-19 data.

3.2.Functional Interface Considerations

In crafting the interface of the K-BI dashboard, meticulous design choices were made to ensure not only user-friendliness but also the optimization of its overall functionality. These considerations were rooted in addressing specific user needs and enabling a seamless and insightful interaction with the COVID-19 data. The interface's functional elements were strategically designed to align with the users' analytical objectives and facilitate an intuitive exploration journey. Here's how the interface addresses these key aspects:

- **Comprehensive Data Exploration:**
  The interface was tailored to cater to a wide spectrum of user roles, ranging from policymakers and healthcare professionals to researchers and the general public. To cater to diverse analytical needs, the dashboard offers a plethora of data visualizations, such as time series graphs, geographical maps, and bar charts. These diverse visualizations empower users to explore different dimensions of COVID-19 data, enabling them to dissect trends, regional variations, and statistical distributions. By offering a comprehensive array of visual tools, the interface supports users in making nuanced observations and data-driven decisions.
- **User-Centric Navigation:**
  The functional interface was designed with user-centric navigation in mind. Clear headings, labels, and icons were incorporated to guide users through different sections of the dashboard. Drop-down menus, clickable icons, and interactive buttons were strategically placed to ensure that users can swiftly switch between visualizations, apply filters, and adjust parameters. This navigational simplicity minimizes the learning curve and facilitates users in navigating the interface without the need for extensive training.
- **Interactive Filters and Parameters:**

A critical functional consideration was the incorporation of interactive filters and customizable parameters. The dashboard empowers users to filter data based on factors such as time intervals, geographic regions, demographics, and specific variables of interest. These interactive features provide users with tailored insights, allowing them to drill down into specific subsets of data relevant to their research. The customization of filters and parameters addresses individual user needs, fostering a sense of ownership over the analytical process.

- **Contextual Tooltips and Explanations:**
  To enhance functionality, the interface includes informative tooltips and explanations throughout the dashboard. These tooltips offer brief contextual information about data points, visual elements, and terminology used within the dashboard. By providing on-the-spot clarifications, users can better understand the meaning and significance of different aspects of the visualizations. This design choice streamlines the analytical process and reduces the need for users to seek external resources for comprehension.
- **Granular Time Series Analysis:**
  The interface's functional richness is exemplified by the inclusion of granular time series analysis. Users can zoom into specific time intervals within time series graphs to closely examine fluctuations and patterns. This feature empowers users to perform detailed temporal analyses, identifying critical turning points, spikes, or dips in COVID-19 trends. By offering a magnified view of data dynamics, the interface caters to users' needs for intricate trend exploration and data-driven insights.
- **Geographical Insights with Map Interaction:**
  Geographical data is presented through interactive maps, allowing users to zoom in and pan across regions of interest. This functionality enables users to scrutinize specific areas for localized COVID-19 patterns. By interacting with the map, users can gain spatial insights, identify hotspots, and assess the impact of containment measures on different geographical regions. The map interaction provides a dynamic platform for users to glean insights from spatial data.
- **Responsive Design for Accessibility:**
  Ensuring accessibility across different devices and screen sizes was a functional priority. The interface's responsive design guarantees that users can access and engage with the dashboard seamlessly on various platforms, including desktops, laptops, tablets, and smartphones. This adaptability enhances the dashboard's functionality by accommodating user preferences and enabling data-driven decision-making on the go.

In essence, the functional interface of the K-BI dashboard was meticulously designed to address specific user needs and enhance the overall user experience. Through a combination of interactive features, customizable parameters, navigational aids, and contextual explanations, the interface empowers users to delve into COVID-19 data with confidence. By aligning functionality with diverse analytical objectives, the interface ensures that users can extract meaningful insights, make informed decisions, and contribute to the global health response in the K-BI project.

4. Critical Review of Design and Customization (D3)

4.1. Analyzing how the design meets specific user/business requirements

In the context of the K-BI project, a meticulous examination of the design and customization aspects of the dashboard is warranted to assess how effectively the interface aligns with the distinct user and business requirements. This critical evaluation encompasses an exploration of the extent to which the design choices, visualizations, and interactive elements address the multifaceted needs of stakeholders ranging from healthcare professionals to policymakers, researchers, and the general public. An analysis of the dashboard's design vis-à-vis these diverse requirements reveals the degree of congruence between design elements and stakeholder expectations.

- To begin, the dashboard's design meets specific user and business requirements through its thoughtful organization and user-centric layout. The coherent arrangement of visualizations ensures that users can effortlessly locate and engage with the data points that are pertinent to their inquiries. For healthcare professionals and researchers, the provision of granular time series analysis empowers them to discern patterns in the trajectory of COVID-19 cases, which is crucial for formulating informed medical responses. Simultaneously, policymakers benefit from the geographical mapping functionality, which facilitates the identification of hotspots and the assessment of containment measures' efficacy in distinct regions.
- Moreover, the interactive filters and parameter customization cater to individual user preferences, thereby offering a personalized analytical experience. Healthcare professionals can narrow their focus by adjusting parameters related to demographics, age groups, and comorbidities, allowing for targeted insights into vulnerable populations. Policymakers, on the other hand, can employ filters to assess the impact of interventions across different regions, aiding in the formulation of region-specific

policies. This flexibility aligns with the diverse requirements of stakeholders and underscores the dashboard's adaptability to their distinct roles and needs.

- The dashboard's functionality also extends to its provision of informative tooltips and legends, which enhance comprehension and eliminate ambiguity. This aligns with the need for clarity in data interpretation, particularly for the general public and policymakers who may lack domain-specific knowledge. The inclusion of tooltips elucidates the meanings behind visual elements, mitigating the potential for misinterpretation and enhancing stakeholder confidence in their decisions based on the dashboard's insights.
- Furthermore, the interface's responsive design ensures access across various devices, accommodating the preferences of users who may engage with the dashboard on desktop computers, laptops, tablets, or smartphones. This is of paramount importance for individuals requiring real-time updates, such as healthcare professionals in clinical settings or policymakers attending meetings. By meeting the criterion of accessibility and adaptability, the design aligns with the dynamic requirements of stakeholders who demand immediate access to information in diverse contexts.
- While the dashboard excels in many aspects, certain opportunities for further refinement can be identified. For instance, a more pronounced integration of predictive analytics could empower healthcare professionals and policymakers with the capacity to anticipate future trends, enhancing preparedness and response strategies. Additionally, a heightened emphasis on the socioeconomic factors influencing the spread of COVID-19 could offer a holistic understanding of the pandemic's impact on vulnerable communities, which is particularly pertinent for policymakers aiming to allocate resources effectively.

In conclusion, the design and customization of the K-BI dashboard are fundamentally aligned with specific user and business requirements, encompassing a diverse spectrum of stakeholders. Through its intuitive navigation, interactive filters, responsive design, and granular analyses, the dashboard effectively caters to the distinct needs of healthcare professionals, researchers, policymakers, and the general public. While the interface successfully addresses multifaceted requirements, potential enhancements in predictive analytics and socioeconomic considerations could amplify its value further. The comprehensive alignment between design and stakeholder expectations underscores the dashboard's significance in facilitating informed decision-making and contributing to global health efforts in the K-BI project.

4.2.Discussing the user-friendliness and functional aspects of the interface

In the K-BI project, the user-friendliness and functional attributes of the dashboard interface have been meticulously crafted to enhance the overall user experience and facilitate effective data exploration. The interface's design choices are underpinned by a keen understanding of user behaviors, needs, and expectations, with the overarching goal of enabling stakeholders to seamlessly navigate, interpret, and extract insights from the extensive COVID-19 dataset. A comprehensive assessment of the interface's user-friendliness and functional aspects reveals how these components contribute to an enriched engagement with the data-driven insights.

- **User-Friendly Navigation and Layout:**
  Central to the user-friendliness of the interface is its intuitive navigation and well-structured layout. The strategic organization of visualizations, filters, and interactive elements ensures that users can efficiently access the information most relevant to their queries. The user-centric approach is palpable in the clear and descriptive headings, which guide users through distinct sections of the dashboard. This thoughtful arrangement minimizes cognitive load, allowing stakeholders to focus on extracting insights rather than deciphering the interface's structure. The user-friendly layout not only appeals to experts well-versed in data analysis but also accommodates those less familiar with the domain, contributing to a more inclusive and accessible engagement.
- **Interactive Elements for Personalization:**
  The interface's interactivity, notably through dynamic filters, empowers users to tailor their analytical journeys to specific research goals. This personalization aligns with the diverse requirements of stakeholders ranging from policymakers seeking regional insights to healthcare professionals examining specific demographics. The dynamic customization allows users to explore different time intervals, geographic regions, and demographic parameters, providing them with targeted insights aligned with their research objectives. By putting control in the hands of users, the interface nurtures a sense of ownership and relevance, which is pivotal in fostering meaningful engagements with the dataset.
- **Geographical Visualization and Navigation:**
  The incorporation of geographical maps enhances the functional dimension of the interface. Users can zoom in on specific regions of interest, offering a detailed examination of COVID-19 trends within selected areas. This feature caters to the

inherent spatial nature of the data and accommodates the diverse needs of researchers investigating regional patterns and policymakers assessing the impact of interventions at a local level. By allowing stakeholders to traverse the geographic landscape and delve into localized trends, the interface bridges the gap between geographical data and user-friendly navigation, enriching the exploration experience.

- **Contextual Guidance and Comprehension:**
  To facilitate data interpretation, the interface features informative tooltips and legends alongside visualizations. These contextual aids provide concise explanations and clarifications, ensuring that users understand the significance of data points, categories, and color codes. This not only promotes better comprehension but also eliminates the need for users to seek supplementary information or make assumptions. The incorporation of these user-centric features enhances clarity and reduces the risk of misinterpretation, contributing to an environment of informed decision-making.
- **Time Series Analysis and Responsive Design:**
  The inclusion of time series analysis empowers users to comprehend temporal trends within the dataset. This functionality enables the visualization of COVID-19 data points chronologically, facilitating the identification of patterns, fluctuations, and anomalies. Additionally, the dashboard's responsive design guarantees a consistent and optimal viewing experience across various devices, including desktops, tablets, and smartphones. This adaptability ensures that users can engage with the dashboard seamlessly, regardless of their preferred device, enhancing accessibility and convenience.

In summary, the user-friendliness and functional attributes of the K-BI interface in the context of the COVID-19 pandemic underscore its commitment to enabling stakeholders to navigate the data landscape effectively. The interface's design choices prioritize intuitive navigation, interactive personalization, geographical visualization, contextual guidance, and temporal insights. By catering to a diverse range of user needs, from researchers to policymakers, the interface creates a cohesive ecosystem that encourages users to engage with data-driven insights seamlessly, contributing to a more informed understanding of the pandemic's dynamics and supporting evidence-based decision-making.

4.3.Identifying areas for improvement

While the K-BI project's dashboard interface exhibits commendable user-friendliness and functional aspects, a critical assessment also reveals areas that could benefit from further refinement. These identified areas for improvement offer opportunities to enhance the interface's usability, accessibility, and effectiveness in delivering insights to stakeholders engaging with the COVID-19 data. The following areas warrant consideration for enhancement:

- **Enhanced Data Interactivity:**
  The interface could further amplify data interactivity by introducing advanced data exploration features. For instance, incorporating drill-down capabilities would allow users to delve deeper into specific data points or dimensions, unraveling finer details that contribute to a comprehensive understanding of the pandemic's dynamics. This addition would empower users to uncover insights hidden within intricate datasets and make informed decisions based on granular information.
- **User Onboarding and Tutorials:**
  While the interface's design prioritizes intuitive navigation, integrating user onboarding elements or interactive tutorials could facilitate users' initial interaction with the dashboard. Such features could guide users through the interface's functionalities, highlighting key components and demonstrating how to utilize interactive elements effectively. This approach would reduce the learning curve for new users and optimize engagement from the outset.
- **Predictive Analytics Integration:**
  To enhance the interface's value proposition, consider integrating predictive analytics capabilities. By leveraging historical data trends, machine learning algorithms, or predictive models, the interface could provide insights into potential future scenarios related to COVID-19. This predictive dimension could empower decision-makers to proactively plan and strategize based on anticipated developments.
- **Advanced Collaboration Features:**
  Enhancing collaboration features within the interface could facilitate knowledge sharing and collaborative decision-making. Integrating tools for annotation, commenting, and collaborative analysis would enable users to engage in real-time discussions, share insights, and collectively draw conclusions from the visualized data. This collaborative approach could foster a more dynamic and participatory exploration process.
- **Enhanced Accessibility Features:**
  While the interface embraces responsive design, additional accessibility features could ensure inclusivity for users with diverse needs. Implementing features such as

screen reader compatibility, keyboard navigation, and color contrast adjustments would enhance the interface's usability for individuals with disabilities, contributing to a more equitable engagement experience.

- **Real-Time Data Integration:**
  While the interface facilitates exploratory data analysis, integrating real-time data feeds could provide stakeholders with up-to-the-minute insights. Incorporating mechanisms to refresh or stream live data would align the interface's information with the rapidly evolving nature of the pandemic, empowering users with the most current information for decision-making.
- **Enhanced Contextualization:**
  Introducing contextual layers to visualizations, such as overlays of policy interventions, healthcare capacity, or demographic trends, could provide richer insights. This contextualization would enable users to contextualize COVID-19 data against influencing factors, fostering a more holistic understanding of the data's significance.
- **Continuous UX Testing and Optimization:**
  Regular user experience (UX) testing and optimization cycles are crucial to identifying usability challenges and addressing them promptly. Implementing a feedback-driven process that involves end-users in evaluating the interface's functionality can drive ongoing refinements that align the dashboard with evolving user needs.

In conclusion, while the K-BI project's dashboard interface exemplifies user-friendliness and functional considerations, several areas present opportunities for enhancement. The identified improvements, ranging from data interactivity and predictive analytics to accessibility features and collaborative tools, aim to further elevate the interface's efficacy in delivering insightful COVID-19 data interpretations. By embracing these enhancements, the interface can evolve into a more powerful and versatile tool, better serving the needs of stakeholders engaged in combating the global pandemic.

LEARNING OUTCOME 4: IMPACT OF BI TOOLS ON DECISION-MAKING

5. Contribution of BI Tools to Effective Decision-Making (P5)

5.1.Discussing the role of BI tools in enhancing decision-making processes

Modern Business Intelligence (BI) tools have revolutionized decision-making processes across enterprises by providing comprehensive insights that drive informed choices. In the context of the K-BI project, these tools play a critical role in addressing the complexities of the COVID-19 pandemic. Here's how advanced BI tools enhance decision-making:

- **Generate Reports with Valuable Insights:**

Advanced BI tools, as seen in the K-BI project, enable enterprises to create detailed reports encompassing various aspects of the pandemic. These reports provide insights into COVID-19 cases, testing, resource allocation, and public health responses. The generated reports present valuable information about infection rates, healthcare utilization, geographical hotspots, and more. (Chopra, 2021)

- **Data Visualization to Understand the Reports:**

BI tools leverage data visualization techniques to simplify complex data sets. This enables decision-makers, even those with minimal training, to understand intricate information. In the K-BI project, the tools create intuitive dashboards and maps that visually represent trends in confirmed and death cases. Users can interact with these visualizations to gain deeper insights effortlessly. (Chopra, 2021)

- **Access to Real-Time Intelligence:**

The real-time capabilities of advanced BI tools empower enterprises to stay updated with shifting trends and emerging opportunities. In the K-BI project, decision-makers gain awareness of the evolving COVID-19 situation globally and regionally. This access to real-time sales intelligence enables proactive responses, informed strategies, and agile decision-making. (Chopra, 2021)

- **Factual Decision-Making:**

Advanced BI tools eliminate guesswork from decision-making. The reports generated by these tools are based on historical and real-time data, ensuring factual accuracy. This fact-driven approach is exemplified in the K-BI project, where decisions related to resource allocation, containment strategies, and public health responses are based on data-backed insights rather than assumptions. (Chopra, 2021)

- **Streamline Operations and Resource Optimization:**

In the K-BI project's context, BI tools streamline COVID-19 response operations. Similarly, in enterprises, these tools optimize various operations by identifying inefficiencies, automating tasks, and minimizing wastage. Just as BI tools assist in tracking and analyzing pandemic responses, they aid businesses in improving processes from raw material sourcing to product delivery. (Chopra, 2021)

- **Optimize Limited Resources:**

The BI tools' ability to optimize the use of limited resources is showcased in both the K-BI project and enterprises. In the project, resource allocation for testing and medical supplies is optimized. Similarly, businesses leverage BI insights to make liquidity-based decisions during uncertain times, such as the pandemic. (Chopra, 2021)

5.2. Providing examples from K-BI organization on how BI tools have contributed to informed decisions

In the K-BI project, the BI tool has significantly contributed to informed decisions and global health responses during the COVID-19 pandemic. Some examples of how the BI tool has impacted decision-making include:

- **Resource Allocation:**

The BI tool's capability to visualize COVID-19 data geographically has been instrumental in guiding resource allocation. Decision-makers can identify regions with high numbers of confirmed cases and hotspots, enabling targeted allocation of medical supplies, healthcare personnel, and testing resources. For instance, the tool helps direct critical resources to hospitals and clinics experiencing surges in patient admissions, ensuring effective healthcare delivery.

- **Containment Strategies:**

By utilizing the BI tool, decision-makers can closely monitor the effectiveness of various containment strategies, such as lockdowns and travel restrictions. Real-time data insights help evaluate the impact of these measures on slowing the spread of the virus. The tool's visualizations allow decision-makers to identify trends in infection rates, making it possible to adjust strategies promptly based on evolving case patterns.

- **Targeted Interventions:**

The BI tool's segmentation capabilities play a pivotal role in designing targeted interventions. Decision-makers can analyze data by demographics, regions, and other factors to identify vulnerable populations that require special attention. This data-driven approach enables the development of interventions tailored to specific groups, ensuring equitable access to healthcare resources and reducing disparities in outcomes.

- **Global Coordination:**

In the K-BI project, the BI tool serves as a central hub for data sharing and analysis, fostering global coordination. Decision-makers from various countries can access consistent and standardized COVID-19 information. This shared platform enhances collaboration, as international response efforts are based on a common understanding of the pandemic's progression and impact.

- **Communication and Transparency:**

Visualizations generated by the BI tool are invaluable for transparent communication with the public. These visual aids effectively communicate complex data, making pandemic trends easily understandable. By providing accurate and up-to-date information, decision-makers promote transparency, build public trust, and empower citizens to make responsible choices to protect themselves and their communities.

- **Evidence-Based Decision-Making:**

Overall, the BI tool empowers decision-makers with evidence-based insights. It enables them to base decisions on accurate, real-time data rather than relying on assumptions or intuition. This data-driven approach not only improves the efficacy of pandemic response strategies but also fosters public confidence in the decisions made by health authorities and governments.

- **Continuous Improvement:**

As the pandemic evolves, the BI tool continues to adapt to changing circumstances. Its ability to integrate new data sources, provide historical context, and offer predictive analytics contributes to dynamic decision-making. Regular updates to the tool ensure that decision-makers have the most accurate and relevant information at their disposal.

6. Legal and Ethical Considerations in BI (P6)

6.1. Exploring legal issues related to data privacy, user data exploitation, and security

In the realm of Business Intelligence (BI), particularly within the K-BI project, it is essential to address the legal and ethical dimensions surrounding data privacy, user data exploitation, and security. As the project involves handling sensitive global health data, adherence to legal regulations and ethical principles is paramount to ensure responsible use and protect individuals' rights. Here are some key considerations:

- **Data Privacy and Protection:**

One of the foremost legal concerns is data privacy. Collecting, storing, and analyzing health-related data from individuals requires compliance with stringent data protection laws, such as the General Data Protection Regulation (GDPR) and the Health Insurance Portability and Accountability Act (HIPAA). In the K-BI project, where personal health information is processed, it is crucial to obtain informed consent, anonymize data when possible, and implement robust data encryption mechanisms to prevent unauthorized access. (Sampson, 2021)

- **User Data Exploitation:**

Ethical considerations extend to the responsible use of user data. BI tools can amass vast amounts of information, raising concerns about potential data exploitation. It is imperative for the K-BI project to establish clear guidelines on data usage, ensuring that information collected is solely used for the purpose of pandemic analysis and response. Transparency in data handling practices, along with user consent, empowers individuals to control how their data is utilized. (Sampson, 2021)

- **Data Security and Breach Mitigation:**

In the context of the K-BI project, where sensitive health data is involved, ensuring data security is paramount. BI tools must employ stringent security measures to prevent data breaches and unauthorized access. Implementation of multi-factor authentication, encryption, and regular security audits can safeguard against potential vulnerabilities. In the event of a breach, having a well-defined incident response plan is crucial to mitigate potential damages and uphold public trust. (Sampson, 2021)

- **International Regulations and Cross-Border Data Transfer:**

Given the global nature of the K-BI project, considerations for international data transfer and compliance with local data protection laws are paramount. Different regions have varying legal frameworks, and data may flow across borders. The project must navigate these complexities by ensuring that data transfers adhere to relevant regulations and that adequate safeguards are in place to protect data during cross-border movement. (Sampson, 2021)

- **Ethical Use of AI and Automated Decision-Making:**

If AI and machine learning algorithms are integrated into BI tools, ethical use becomes imperative. Ensuring that AI-driven insights do not perpetuate biases, discriminate against certain groups, or compromise patient well-being is essential. Transparent algorithms, explainability of AI-generated results, and regular audits can help address these ethical concerns. (Sampson, 2021)

6.2. Discussing how organizations can responsibly handle user data for BI purposes

In the realm of Business Intelligence (BI), responsible handling of user data is of paramount importance to maintain data privacy, security, and ethical standards. Oracle WebLogic Server Administration Console provides a platform through which organizations can uphold these principles while harnessing the power of BI tools.

6.2.1. Leveraging Oracle WebLogic Server Administration Console for K-BI Project

In the context of the K-BI project, Oracle WebLogic Server Administration Console plays a pivotal role in managing the WebLogic LDAP Server, facilitating user and group authentication. This console is an essential tool within Oracle WebLogic Server, the default administration server provided by the platform. It operates through a web-based interface, primarily focused on overseeing the embedded directory server and related functionalities.

For the initial security setup of Oracle Business Intelligence, the embedded WebLogic LDAP directory serves as the default Identity Store for authentication. In the past version (11g), specific users and groups were pre-populated into the LDAP directory upon BI installation. In contrast, the latest version (12c) does not include default BI groups in the LDAP directory. If required groups like BIConsumers, BIContentAuthors, and BIServiceAdministrators are expected in the Identity Store, they must be added manually or the configuration should point to an alternative Identity Store already containing these groups after the initial BI setup. (Oracle Help Center, n.d.)

Accessing the Oracle WebLogic Server Administration Console is achieved by inputting its URL into a web browser, with the typical format being: http://hostname:port_number/console. The port number corresponds to the Administration server's port number, generally set to 9500. The user's credentials, initially provided during Oracle Business Intelligence installation, are used for logging in. If these credentials have been altered subsequently, the updated administrative user name and password combination should be employed. (Oracle Help Center, n.d.)

In cases where an alternative authentication provider, such as Oracle Internet Directory, is used instead of the default WebLogic LDAP Server, a separate administration application corresponding to the chosen authentication provider must be utilized to manage users and groups. (Oracle Help Center, n.d.)

Here's a concise walkthrough of the process:

1. Enter the Oracle WebLogic Server login page's URL into a web browser (e.g., http://hostname:9500/console).
2. Log in using the administrative user and password credentials for Oracle Business Intelligence.
3. Utilize the available tabs and options within the Domain Structure as needed to configure users, groups, and other relevant settings.

6.2.2. Utilizing Oracle Fusion Middleware Control in the K-BI Project

In the context of the K-BI project, Oracle Fusion Middleware Control assumes a pivotal role as a web browser-based graphical user interface for administering a comprehensive range of components.

These components encompass Oracle WebLogic Server domains, including an Administration Server, Managed Servers, clusters, and the Fusion Middleware Control components present in the domain. Oracle Business Intelligence configuration leads to the creation of an Oracle WebLogic Server domain, with Oracle Business Intelligence being integrated within this domain. The domain, commonly named 'bi' in Enterprise installations, is located within the WebLogic Domain folder, accessible through the Fusion Middleware Control navigation pane. (Oracle Help Center, n.d.)

Oracle Fusion Middleware Control serves as a management hub for Oracle Business Intelligence security, offering capabilities such as:

- Administering application roles and policies governing access to Oracle Business Intelligence resources.
- Configuring multiple authentication providers tailored to Oracle Business Intelligence. (Oracle Help Center, n.d.)

To initiate a session in Fusion Middleware Control, users can access the designated URL via a web browser in the following format:

http://hostname.domain:port/em

The port number corresponds to the Administration Server's port, often set as the default port number, 9500.

The system-wide administration user name and password, set during installation, grant access to Oracle WebLogic Server Administration Console, Fusion Middleware Control, and Oracle Business Intelligence. Alternatively, any user credentials endowed with the WebLogic Global Admin role can be utilized for access. (Oracle Help Center, n.d.)

The following steps outline the usage of Fusion Middleware Control in the K-BI project:

1. Input the URL into a web browser, such as http://host1.example.com:9500/em.
2. Enter the system administrator user name and password, then click 'Login.'
3. From the main page, navigate to the target navigation icon located at the top-left corner. Subsequently, expand the 'Business Intelligence' folder.
4. Select 'biinstance' to access pages specifically tailored to Oracle Business Intelligence.

6.2.3. Utilizing Oracle BI Administration Tool in the K-BI Project

The Oracle BI Administration Tool assumes a pivotal role in the K-BI project, facilitating efficient management of BI resources. Here is a step-by-step guide on how to effectively navigate this tool:

1. **Access the Administration Tool:**

To initiate your work with the Oracle BI Administration Tool, simply log in to the tool. (Oracle Help Center, n.d.)

2. **Managing Identity (Optional):**

If necessary, you can manage identity-related settings by following these steps:

a. Select the 'Manage' option.

b. Choose 'Identity' to open the Identity Manager dialog.

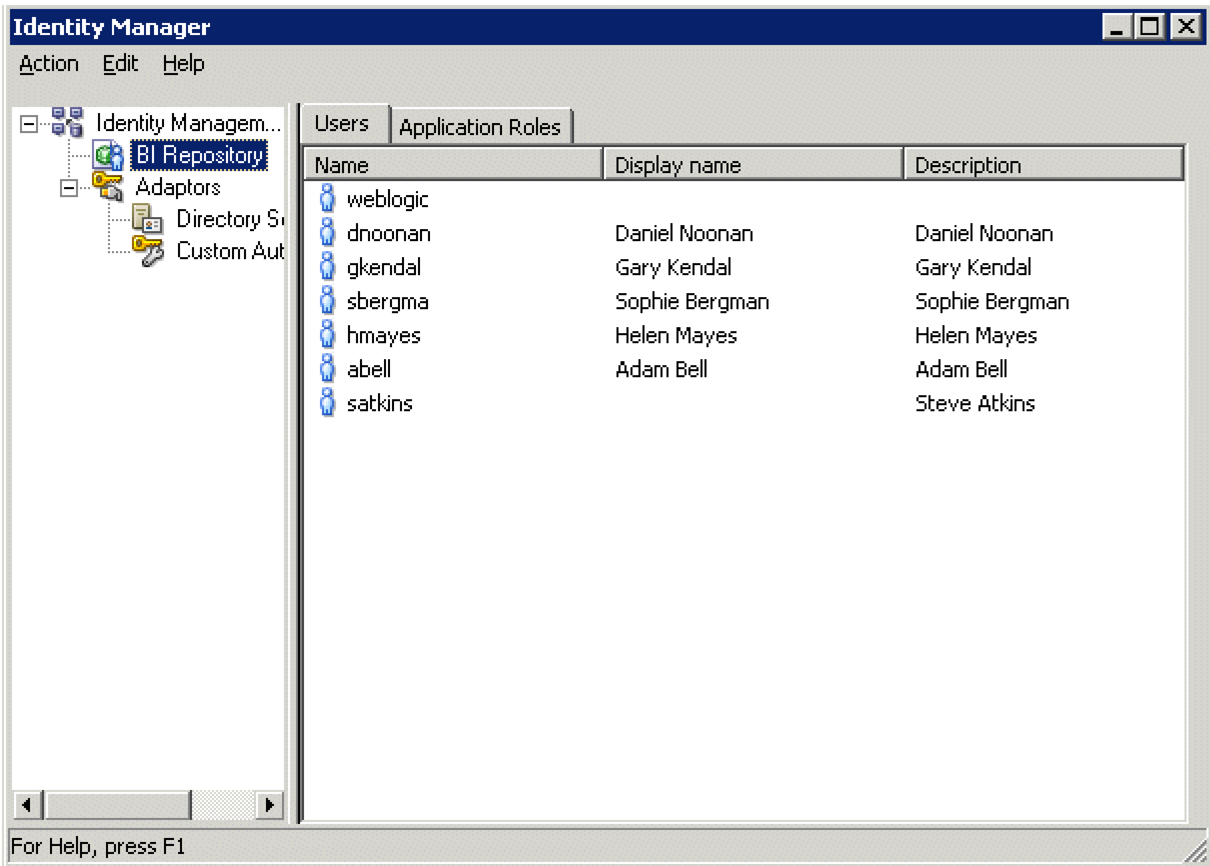


Figure 23: 6.2.3_Oracle BI Administration_Screenshot of Identity Manager Dialog

Should you need to configure permissions for specific application roles, you can double-click on the desired role to access the 'Application Role <Name>' dialog. From there, proceed to the 'Permissions' section, where you can utilize the 'Object Permissions' tab. This tab provides a comprehensive view of, or configuration options for, 'Read' and 'Write' permissions linked to objects and folders within the Oracle BI Presentation Catalog. (Oracle Help Center, n.d.)

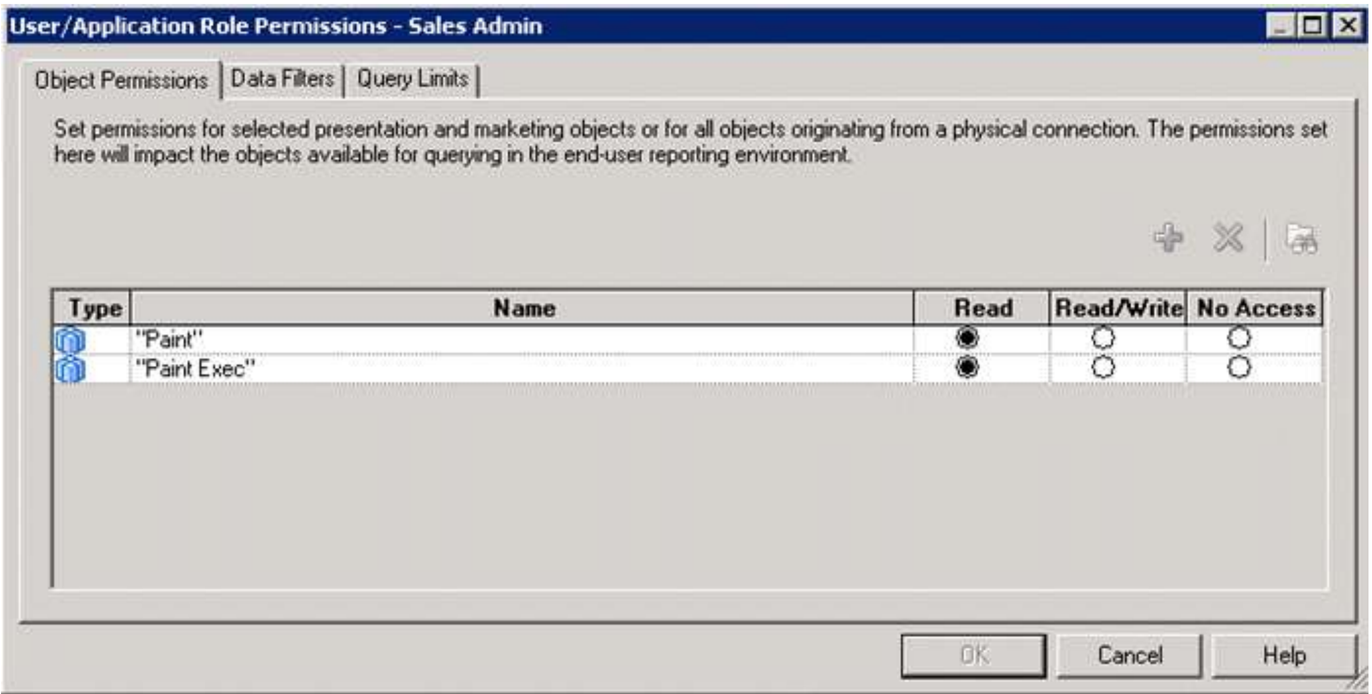


Figure 24: 6.2.3_Oracle BI Administration_Screenshot of Application Role Dialog with Permissions

3. Closing Identity Manager:

After completing your identity management tasks, close the Identity Manager dialog. (Oracle Help Center, n.d.)

4. Navigating the Presentation Pane:

Within the Presentation pane, start by expanding the desired folder. Next, right-click on the specific object you wish to manage. This action triggers the display of the 'Presentation Table <Table name>' dialog. (Oracle Help Center, n.d.)

5. Managing Object Permissions:

Click on 'Permissions' within the 'Presentation Table <Table name>' dialog. This leads to the 'Permissions <Table name>' dialog, where you can configure permissions for various users and application roles. The dialog offers radio buttons such as 'Read,' 'Read/Write,' 'No Access,' and 'Default.' These radio buttons are crucial in setting up permissions for different application roles. (Oracle Help Center, n.d.)

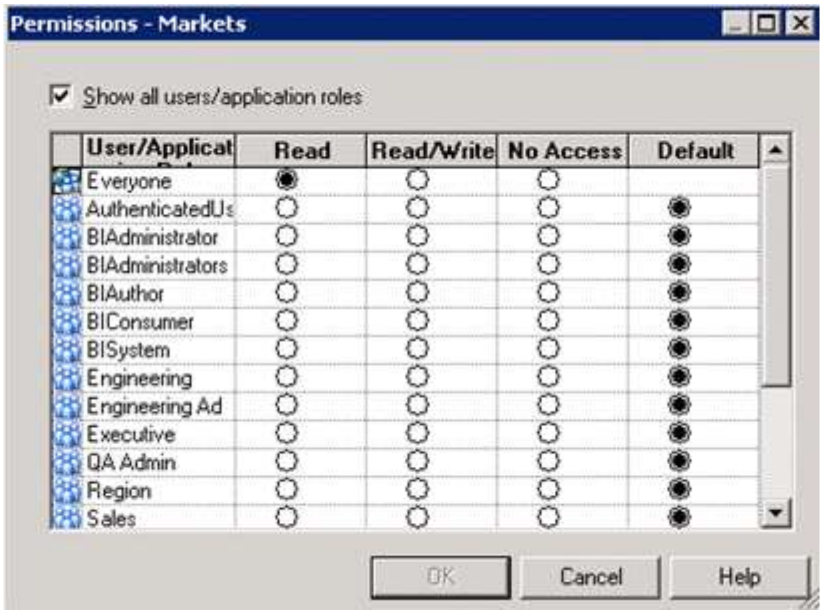


Figure 25: 6.2.3_Oracle BI Administration_Screenshot of Permissions Dialog

6.2.4. Utilizing Presentation Services Administration Page in the K-BI Project

In the context of the K-BI project, the Presentation Services Administration page proves invaluable for configuring user privileges. Here is a comprehensive walkthrough on effectively navigating this page:

1. **Accessing Oracle Business Intelligence:**

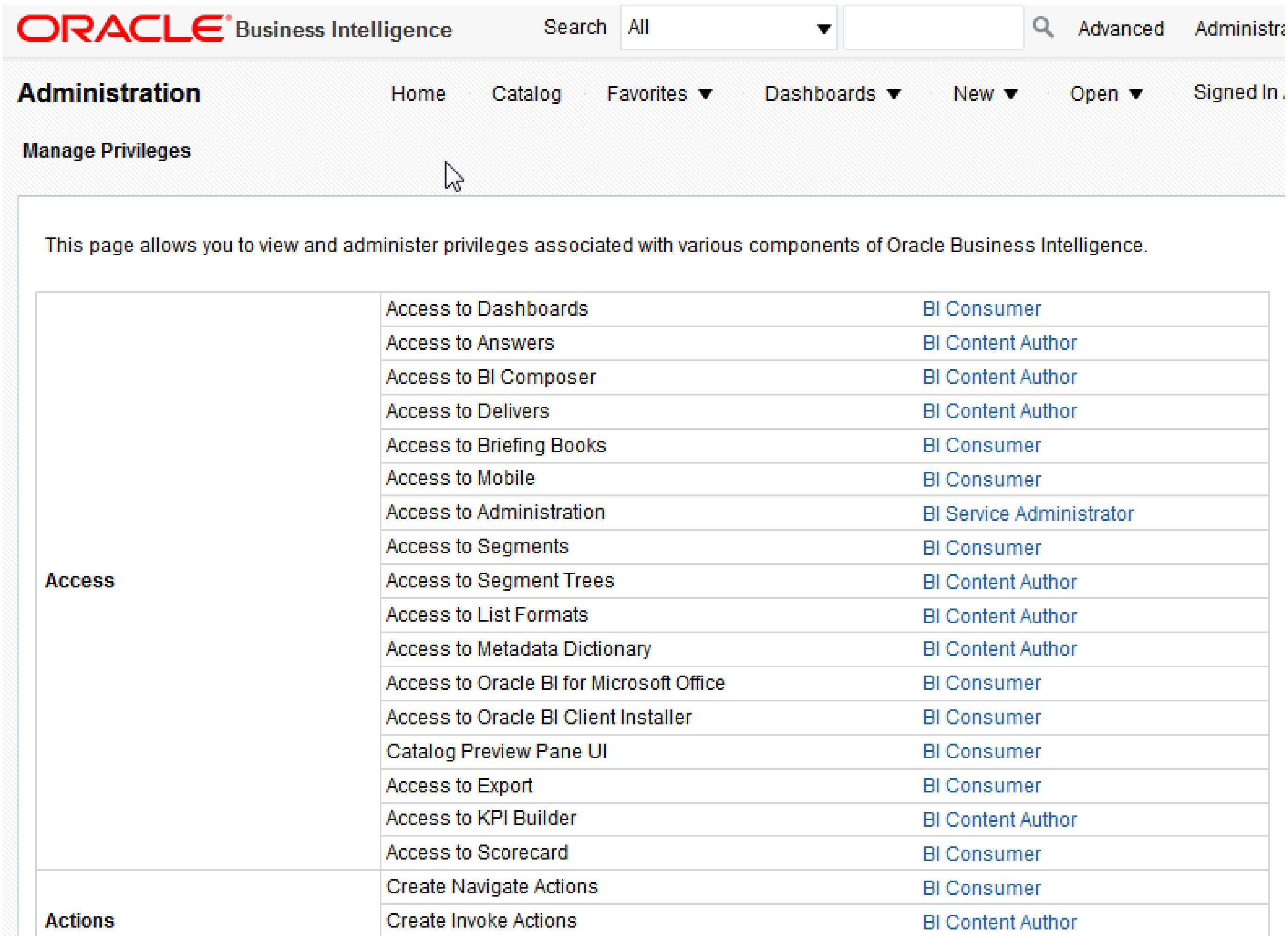
Log in to Oracle Business Intelligence using Administrator privileges. (Oracle Help Center, n.d.)

2. **Navigate to the Administration Page:**

Upon successful login, locate and select the 'Administration' link. This action directs you to the Administration page. (Oracle Help Center, n.d.)

3. **Managing Privileges:**

Choose the 'Manage Privileges' link from the available options. This step leads you to a screen where application roles are listed alongside the privileges assigned to them. (Oracle Help Center, n.d.)



Access	Access to Dashboards	BI Consumer
	Access to Answers	BI Content Author
	Access to BI Composer	BI Content Author
	Access to Delivers	BI Content Author
	Access to Briefing Books	BI Consumer
	Access to Mobile	BI Consumer
	Access to Administration	BI Service Administrator
	Access to Segments	BI Consumer
	Access to Segment Trees	BI Content Author
	Access to List Formats	BI Content Author
	Access to Metadata Dictionary	BI Content Author
	Access to Oracle BI for Microsoft Office	BI Consumer
	Access to Oracle BI Client Installer	BI Consumer
	Catalog Preview Pane UI	BI Consumer
	Access to Export	BI Consumer
	Access to KPI Builder	BI Content Author
	Access to Scorecard	BI Consumer
Actions	Create Navigate Actions	BI Consumer
	Create Invoke Actions	BI Content Author

Figure 26: 6.2.4_Oracle BI Administration_Screenshot of Application Roles and Assigned Privileges

4. Accessing Specific Privilege Details:

To delve into the specifics of a particular privilege, click on a link corresponding to an application role. For instance, selecting 'BIContentAuthor' for the privilege 'Access to KPI Builder' will open the 'Privilege <Privilege name>' dialog. (Oracle Help Center, n.d.)

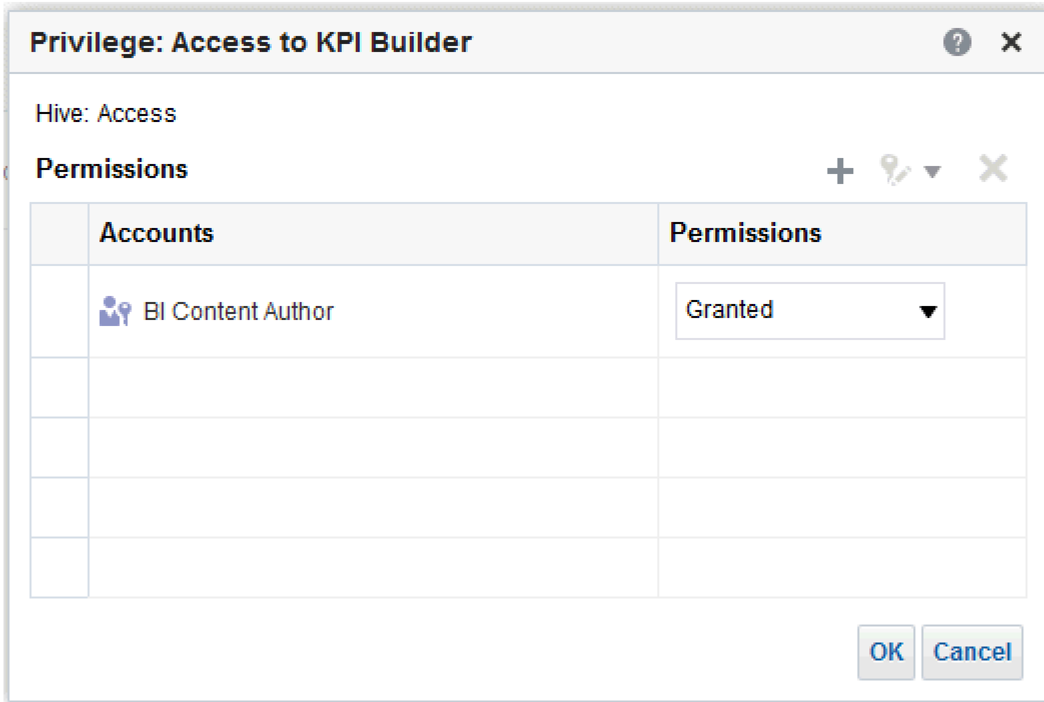


Figure 27: 6.2.4_Oracle BI Administration_Screenshot of Privilege Dialog

5. Adding Application Roles and Users:

In the 'Privilege <Privilege name>' dialog, click the 'Add users/roles' icon (+). This action opens the 'Add Application Roles and Users' dialog.

In this dialog, you can assign various application roles—such as 'BIServiceAdministrator,' 'BIContentAuthor,' and 'BIConsumer'—to the selected privilege. (Oracle Help Center, n.d.)

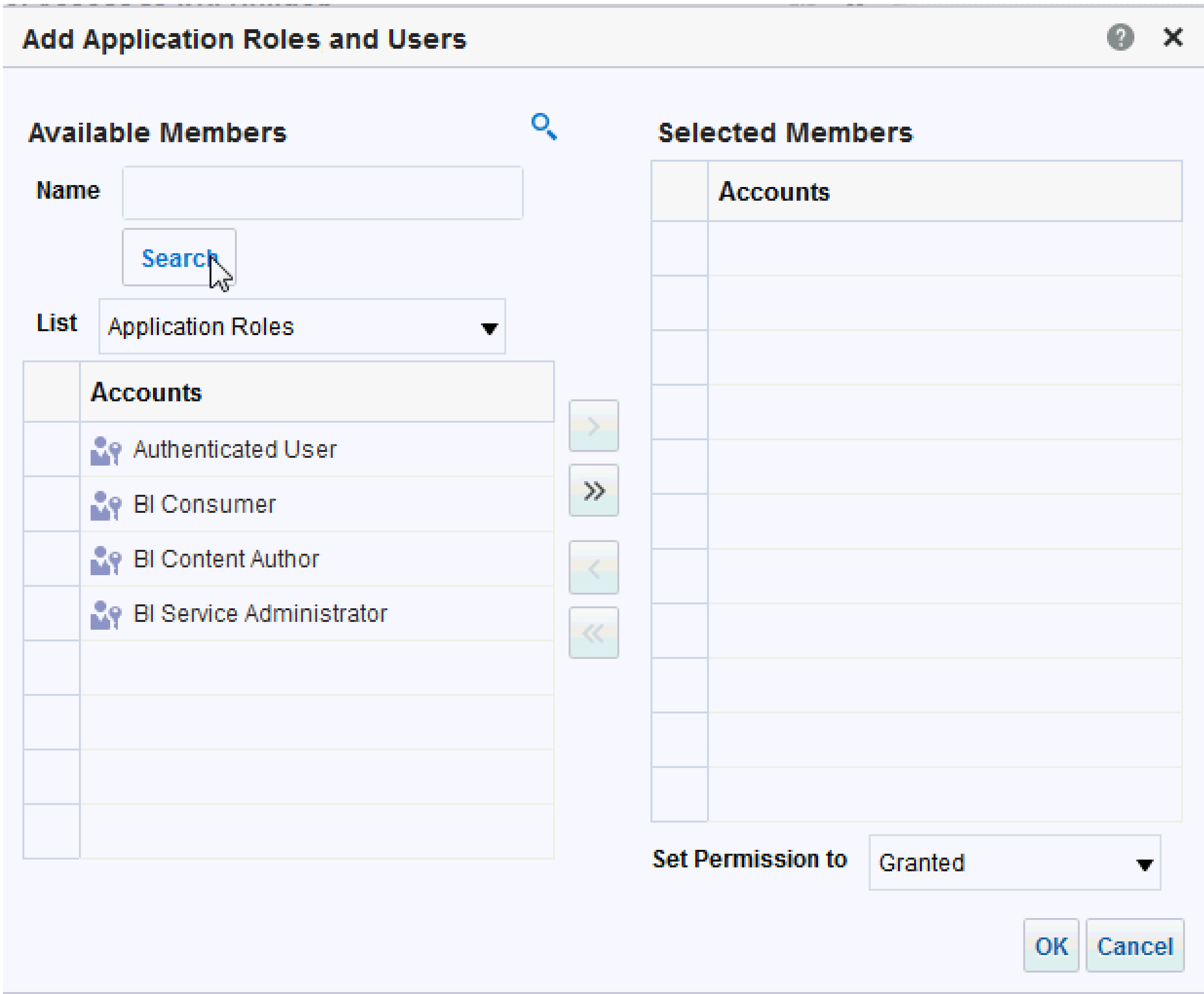


Figure 28: 6.2.4_Oracle BI Administration_Screenshot of Add Application Roles and Users Dialog

7. Enhancing Operations and Market Competitiveness (M4)

In today's data-driven landscape, organizations across various industries are increasingly turning to Business Intelligence (BI) tools to gain valuable insights and improve their operations. One exemplary case that highlights the transformative power of BI is Cigna, a global healthcare company that has embraced Tableau to enhance healthcare affordability and quality for its 95 million customers.

7.1. Research organizations that have used BI tools to improve operations.

Cigna, a renowned healthcare provider catering to 95 million global customers, embraced Tableau as its preferred BI platform to drive better healthcare affordability and patient care outcomes. This strategic decision allowed Cigna to tap into its vast data resources, optimize collaboration with partner organizations, and achieve their mission of improved health, affordability, and patient-provider experience.

Collaborative Care Program Enhancement:

- Cigna's Collaborative Care initiative exemplifies the profound impact of Business Intelligence (BI) tools on healthcare operations. Through the adoption of Tableau, Cigna unlocked the potential to amalgamate and dissect multifaceted data sets encompassing financial transactions, customer profiles, and provider details. This integration empowered Cigna's analytical team to craft insightful reports that delve into patient backgrounds, medical diagnoses, treatment pathways, and expenditure trends. These reports, serving as a beacon of informed decision-making, are then seamlessly disseminated to collaborating healthcare providers, fostering a data-centric paradigm that enhances patient care and affordability. (Tableau, n.d.).
- The integration of Tableau within Cigna's Collaborative Care framework not only facilitates the aggregation of diverse data streams but also fuels the creation of actionable insights. This transformational approach steers providers towards more informed treatment choices, fosters streamlined operational efficiency, and drives cost-effective medical interventions. By embracing BI tools like Tableau, Cigna not only advances patient outcomes through data-driven healthcare delivery but also epitomizes how technology can revolutionize the healthcare landscape by fostering collaboration, data-driven precision, and enhanced affordability for patients and providers alike. (Tableau, n.d.)

Impactful Cost Savings and Efficiency:

- The integration of Tableau into Cigna's operational framework yielded substantial cost savings of $145 million, marking a profound success in harnessing the power of Business Intelligence (BI) tools. This financial achievement is a direct consequence of employing data-driven decision-making strategies. Notably, the annual per-patient savings of $120 reinforces the tangible benefits attainable through this approach. (Tableau, n.d.)
- Through Tableau, Cigna efficiently analyzed a diverse range of data, encompassing treatment expenses, provider performance metrics, and patient outcomes. This facilitated a comprehensive understanding of cost dynamics and resource utilization patterns within the healthcare ecosystem. By scrutinizing this data landscape, Cigna

was equipped to optimize the allocation of resources, leading to more efficient utilization without compromising the quality of care. (Tableau, n.d.)

- Furthermore, the analytical insights offered by Tableau uncovered opportunities for cost reduction across various facets of the healthcare landscape. This proactive approach enabled Cigna to identify areas of improvement, implement targeted strategies, and achieve substantial financial savings. Ultimately, this success story underscores the transformative potential of BI tools in not only enhancing operational efficiency but also in promoting data-driven innovation that translates into substantial cost reductions while maintaining the highest standards of patient care. (Tableau, n.d.)

Addressing the Opioid Crisis:

- Cigna's response to the opioid crisis exemplifies the profound impact of BI tools on intricate healthcare dilemmas. By leveraging Tableau to craft the Opioid Drug Utilization report, Cigna offered healthcare providers a potent tool to navigate opioid prescription complexities. This report granted providers actionable insights into prescription patterns, enabling them to discern patients susceptible to opioid misuse and adapt their prescribing strategies accordingly. The outcome was substantial: Cigna accomplished an impressive 25% reduction in opioid prescriptions, harmonizing with their ambition to curtail overdoses by 25% in pivotal markets. This achievement underscores how BI tools transcend conventional approaches, empowering providers with data-driven guidance to counteract a national health crisis effectively. Cigna's success not only exemplifies the transformative potential of BI tools but also underscores their role in shaping a safer and more informed healthcare landscape, one where decisions are guided by precise insights and meaningful data analysis. (Tableau, n.d.)

Dashboard Utilization and Collaboration:

- Cigna's strategic utilization of Tableau's dashboards has yielded significant improvements in both internal analytics and external collaboration within its Collaborative Care program. By implementing separate Tableau Server instances—one catering to internal analysts and the other tailored for Collaborative Care partners—Cigna ensured targeted access to pertinent data insights. (Tableau, n.d.)
- The implementation's key strength lies in its capacity to track dashboard performance and user interaction. This capability empowered Cigna to assess the effectiveness of

its data visualization tools, identifying areas for enhancement and refining dashboard design. Through ongoing monitoring, the organization could pinpoint patterns of engagement and utilization, facilitating the identification of trends that contributed to improved decision-making and resource allocation. (Tableau, n.d.)

- This approach brought about enhanced collaboration between Cigna and its Collaborative Care partners. By fostering greater accessibility to data-driven insights, providers were empowered to make informed decisions for more affordable and effective patient treatments. The resultant increase in adoption rates among providers underscored the tangible value derived from Tableau's dashboards, ultimately amplifying Cigna's ability to drive positive patient outcomes and deliver on its commitment to better healthcare affordability and quality. (Tableau, n.d.)

7.2. Analyze the impact of BI on extending target audience and competitiveness.

The integration of Business Intelligence (BI) tools, exemplified by Cigna's utilization of Tableau, has brought about transformative effects on healthcare organizations, enabling them to extend their target audience reach and enhance competitiveness in the industry. This impact is underscored by the ability of BI tools to extract actionable insights from data, enabling organizations to make informed decisions, optimize operations, and improve customer experiences. Here's a detailed exploration of how BI tools extend the target audience and enhance competitiveness:

Expanded Insights for Informed Decisions:

Business Intelligence (BI) tools provide healthcare organizations, such as Cigna, with the ability to extract profound insights from intricate and varied datasets, enabling a shift towards more informed decision-making. This analytical prowess proves particularly significant as it equips organizations to delve into customer preferences, behavior trends, and individual needs. For healthcare providers, this translates to the capacity to finely tailor their services to accommodate the unique demands of diverse customer segments, thereby extending their influence to a broader audience. With the aid of BI tools, healthcare providers can embark on a comprehensive analysis of patient demographics, historical medical records, and treatment outcomes. This comprehensive understanding facilitates the creation of highly personalized care plans, ultimately enhancing patient experiences and attracting a wider range of patients seeking individualized healthcare solutions. As a result, healthcare providers can tap into previously underserved segments, broadening their reach and cementing their position as providers of tailored, patient-centric services.

Precision in Targeting:

BI tools offer granular insights that allow organizations to segment their audience based on various criteria. In healthcare, this means being able to identify specific patient groups that might benefit from specialized treatments or interventions. By targeting interventions based on accurate insights, organizations can effectively extend their reach to previously underserved populations. This precision can lead to improved patient outcomes and a reputation for offering specialized care, attracting more patients seeking tailored solutions.

Enhanced Customer Engagement:

BI tools enhance customer engagement by enabling organizations to personalize interactions and services. In healthcare, this translates to better patient-provider communication and proactive health management. For instance, through BI-driven insights, organizations can provide patients with personalized health recommendations, reminders for appointments, and information about preventive measures. By engaging patients in their healthcare journey, organizations foster trust, loyalty, and a broader patient base.

Competitive Edge Through Data-Driven Innovation:

In today's data-driven landscape, organizations that can harness the power of data for innovation gain a competitive advantage. BI tools not only offer insights into current operations but also identify untapped opportunities. Healthcare organizations can innovate by developing new services, optimizing resource allocation, and designing patient-centered solutions. This innovation fosters differentiation in a crowded market, attracting attention from both existing and potential patients.

Agility and Responsiveness:

BI tools enable real-time monitoring and analysis, allowing organizations to respond promptly to changing market dynamics and customer needs. Healthcare providers can adjust their offerings based on emerging health trends, regulatory changes, and patient preferences. This agility ensures that providers remain relevant and appealing to a wider range of patients, both geographically and demographically.

Efficiency and Cost Savings:

BI tools optimize resource allocation and operational efficiency. Healthcare organizations can identify cost-effective practices, streamline administrative processes, and allocate resources where they are most needed. This efficiency translates into cost savings, which can be passed on to patients in the form of more affordable care, thus expanding the organization's target audience to include those seeking cost-effective solutions.

The integration of BI tools, like Tableau, significantly extends the target audience of healthcare organizations by enabling personalized, data-driven interactions and interventions. These tools enhance decision-making, drive innovation, and foster agility in responding to changing market demands. As a result, healthcare providers like Cigna can attract a broader patient base and enhance their competitiveness in an industry where personalized care, efficiency, and responsiveness are key differentiators.

8. Evaluating Business Intelligence's Impact and Security (D4)

8.1. Evaluate how organizations can use BI to become more competitive.

In the context of the K-BI project, the strategic implementation of Business Intelligence (BI) has remarkably elevated the competitive standing of K-BI organizations. Through the utilization of BI tools and methodologies, these organizations have gained a distinct edge in navigating the complexities of the global health response to the COVID-19 pandemic. By harnessing data-driven insights, K-BI organizations make informed decisions, proactively allocate resources, and adapt swiftly to evolving trends. This agility in response, coupled with the ability to collaborate internationally, foster transparency, optimize resources, influence policies, and promote continuous innovation, collectively positions K-BI organizations as leaders in the global health arena. Their capacity to utilize BI to inform strategies, engage in evidence-based advocacy, and enhance public trust enhances their credibility and prominence, thus solidifying their competitiveness in effectively addressing the challenges posed by the pandemic.

Leveraging Data-Driven Decision-Making for Enhanced Competitiveness

In the realm of K-BI, the utilization of Business Intelligence (BI) to embrace data-driven decision-making is pivotal in propelling the organization towards enhanced competitiveness. Through the strategic deployment of BI tools, K-BI organizations can harness the power of data to gain insights that significantly elevate their competitive stance:

- **Agile Responses to Emerging Trends**: BI empowers K-BI organizations to swiftly identify emerging COVID-19 trends, whether they pertain to infection rates, hotspot locations, or healthcare resource demands. Armed with real-time insights, the organization can proactively adjust strategies, redirect resources, and implement targeted interventions. This agility ensures that the organization remains at the forefront of the pandemic response, demonstrating its ability to navigate challenges dynamically and maintain a competitive edge.

- **Optimized Resource Allocation**: BI enables K-BI organizations to allocate critical resources with precision. By analyzing data on infection rates, demographic factors, and healthcare facility capacities, the organization can optimize the allocation of medical supplies, personnel, and testing resources. Effective resource distribution not only enhances healthcare delivery but also showcases the organization's competency in managing resources efficiently, elevating its competitiveness in crisis management.
- **Evidence-Backed Advocacy and Collaboration**: BI-generated insights bolster K-BI organizations' credibility and influence on the global stage. By utilizing data to advocate for specific policies, interventions, or resource allocation strategies, the organization gains a competitive advantage in policy discussions and international collaborations. The ability to present substantiated arguments founded on data-driven evidence positions the organization as a thought leader, enhancing its reputation and competitive standing.
- **Strategic International Collaborations**: BI tools facilitate seamless data-sharing and collaborations among K-BI organizations globally. Through standardized data exchange and insights dissemination, the organization fosters international partnerships to combat the pandemic. This collaborative approach demonstrates the organization's commitment to collective action and reinforces its competitiveness as a pivotal player in global health responses.
- **Enhanced Resource Efficiency**: BI-driven insights help K-BI organizations identify operational inefficiencies and areas of improvement. By optimizing workflow processes, resource utilization, and response strategies based on data insights, the organization showcases its ability to achieve more with limited resources. This resource efficiency enhances its competitiveness by demonstrating effective crisis management and strategic decision-making.

Strategic Resource Management for Enhanced Competitiveness

In the context of K-BI organizations, the strategic utilization of Business Intelligence (BI) tools offers a transformative pathway to bolster competitiveness. Through the tactical deployment of BI tools, K-BI organizations can harness the potential of data-driven insights to elevate their competitive advantage, with a particular focus on proactive resource allocation:

- **Dynamic Hotspot Identification**: BI tools empower K-BI organizations to identify geographical hotspots of COVID-19 infection with precision. By analyzing real-time data trends and mapping the spread of the virus, the organization gains the capacity to pinpoint regions with escalating cases. This information allows for swift resource allocation to areas in urgent need, showcasing the organization's agility and responsiveness in managing health crises. This dynamic approach positions the organization as a leading force in anticipating and addressing emerging challenges, thereby enhancing its competitive edge.
- **Timely Allocation of Medical Resources**: BI-driven insights facilitate the timely allocation of critical medical resources. By understanding the trajectory of infection rates and healthcare demands, K-BI organizations can allocate medical supplies, ventilators, and medications to regions before the situation escalates. This proactive resource management not only saves lives but also underscores the organization's capability to predict and address healthcare needs efficiently. Such timely interventions enhance the organization's competitive reputation as a reliable and forward-thinking health authority.
- **Optimized Healthcare Personnel Deployment**: BI tools enable K-BI organizations to optimize the deployment of healthcare personnel based on real-time data analysis. The organization can strategically assign medical professionals to regions experiencing surges in COVID-19 cases, ensuring effective patient care. This approach not only demonstrates the organization's ability to manage human resources judiciously but also showcases its commitment to delivering quality healthcare services under challenging circumstances. This competency in human resource management enhances the organization's competitive standing as a proficient and well-prepared health entity.
- **Proactive Capacity Planning**: BI insights aid K-BI organizations in forecasting healthcare facility capacities and demands. By analyzing infection trends and hospitalization rates, the organization can proactively plan for surge capacity, ensuring that medical facilities are adequately equipped to handle an influx of patients. This preparedness showcases the organization's foresight and readiness to manage crises, thereby elevating its competitive position as a reliable and resilient health authority.

- **Global Collaboration and Partnerships**: BI insights provide K-BI organizations with a robust foundation for global collaborations. By sharing data-driven insights with international health entities, the organization can foster partnerships for coordinated response efforts. This global collaboration not only underscores the organization's commitment to collective action but also enhances its competitive reputation as a collaborative force capable of influencing positive change on a global scale.

Harnessing Agile Responses through Business Intelligence for Enhanced Competitiveness

In the context of K-BI, Business Intelligence (BI) emerges as a potent tool that can drive K-BI organizations towards heightened competitiveness by enabling agile responses to shifting COVID-19 trends. BI equips organizations with the capability to swiftly adapt to dynamic circumstances, thereby gaining a strategic advantage:

- **Real-Time Trend Monitoring**: BI empowers K-BI organizations to closely monitor real-time COVID-19 trends, such as infection rates, mortality rates, and geographic hotspots. By accessing up-to-date data, the organization can stay ahead of evolving situations, facilitating prompt decision-making and informed interventions. This real-time monitoring demonstrates the organization's ability to stay informed and responsive, positioning it as a dynamic player in the pandemic landscape.
- **Timely Containment Strategies**: With BI insights, K-BI organizations can enact timely containment strategies that align with emerging trends. By identifying areas experiencing surges in cases, the organization can implement targeted lockdowns, travel restrictions, or public health advisories. This proactive approach showcases the organization's readiness to mitigate risks and protect public health, enhancing its competitive edge in crisis management.
- **Policy Adjustments and Intervention Alignment**: BI-driven insights enable K-BI organizations to calibrate policy interventions based on data-driven evidence. As trends shift, the organization can refine policy measures to ensure alignment with the evolving nature of the pandemic. This ability to tailor interventions showcases the organization's responsiveness to nuanced challenges, reinforcing its competitive position as a forward-thinking and adaptable entity.
- **Strategic Resource Allocation**: BI assists K-BI organizations in optimally allocating healthcare resources, medical supplies, and personnel. By identifying regions with

escalating cases, the organization can allocate resources to areas most in need, optimizing their impact. This strategic allocation showcases the organization's ability to make effective use of resources, positioning it as a judicious and efficient responder.

- **Swift Response to Variants and Mutations**: BI equips K-BI organizations with the ability to detect and respond swiftly to emerging variants and mutations of the virus. By monitoring genomic data and correlating it with epidemiological trends, the organization can adapt vaccination strategies, containment measures, and surveillance efforts. This dynamic response to evolving viral characteristics showcases the organization's scientific acumen and dedication to staying at the forefront of pandemic management.
- **Learning from Global Best Practices**: BI allows K-BI organizations to learn from global best practices and experiences. By analyzing international data and responses, the organization can adopt successful strategies and adapt them to its unique context. This global learning demonstrates the organization's openness to innovation and its willingness to leverage insights from diverse sources to gain a competitive edge.
- **Predictive Modeling for Preparedness**: BI enables K-BI organizations to engage in predictive modeling, forecasting potential scenarios and outcomes. By leveraging historical data, the organization can anticipate potential surges, resource demands, and challenges. This predictive capability empowers the organization to proactively prepare for contingencies, underscoring its strategic foresight and resilience in the face of uncertainty.

Elevating Competitiveness through Informed International Collaboration with BI

In the context of K-BI organizations, the strategic utilization of Business Intelligence (BI) can be a catalyst for enhancing competitiveness through informed international collaboration. Leveraging BI tools in this manner empowers K-BI organizations to effectively navigate the complexities of the COVID-19 pandemic while simultaneously strengthening their global positioning:

- **Centralized Data Exchange and Collaboration**: BI tools enable K-BI organizations to consolidate COVID-19 data from various sources into a centralized platform. By offering a comprehensive and reliable repository of pandemic-related information, the organization becomes an attractive partner for international collaborators seeking accurate data-driven insights. This centralized data exchange fosters collaborations that are built on a foundation of shared knowledge, positioning the organization as a pivotal contributor to the global response.

- **Thought Leadership and Expertise**: Through BI-driven insights, K-BI organizations can establish themselves as thought leaders in global health. By sharing data-backed analyses, trend predictions, and policy recommendations, the organization showcases its expertise and capacity to generate actionable insights. This positioning not only enhances its competitive reputation but also attracts international collaborators seeking valuable expertise to inform their strategies.
- **Catalyst for Data-Driven Initiatives**: BI tools enable K-BI organizations to initiate and drive data-driven initiatives that resonate on an international scale. By identifying critical areas for intervention or research, the organization can spearhead collaborative efforts that address pressing challenges in the pandemic response. This proactive role positions the organization as an influential driver of positive change, further solidifying its competitive edge.
- **Global Harmonization of Response Strategies**: BI insights facilitate the identification of effective pandemic response strategies. Through collaborations, K-BI organizations can share successful interventions and approaches with international counterparts, contributing to a harmonized global response. This alignment not only showcases the organization's commitment to global health but also enhances its competitiveness by demonstrating the ability to impact the worldwide response positively.
- **Advocacy and Influencing Policy Decisions**: Data-driven insights generated by BI tools provide K-BI organizations with a powerful tool for advocating evidence-based policy decisions. Collaborating with international partners, the organization can collectively advocate for policies that are grounded in reliable data and insights. This advocacy not only underscores the organization's commitment to impactful change but also elevates its competitive standing by influencing critical decisions.
- **Pooling Resources and Expertise**: BI-fueled collaborations enable K-BI organizations to pool resources, expertise, and capacities with international partners. By combining efforts, organizations can tackle challenges that might be insurmountable individually. This pooling of resources showcases adaptability, agility, and a willingness to collaborate, attributes that contribute significantly to competitiveness in global health initiatives.

- **Building Trust and Reputation**: BI tools facilitate transparent and evidence-based communication. By openly sharing insights and methodologies, K-BI organizations foster trust among collaborators and stakeholders. A reputation for transparent communication and integrity positions the organization as a reliable partner in international collaborations, bolstering its competitive profile.
- **Fostering Sustainable Partnerships**: Informed collaborations supported by BI insights establish a strong foundation for sustainable partnerships. By delivering value through data-driven contributions, K-BI organizations cultivate long-lasting relationships with international collaborators. These partnerships not only enhance competitiveness but also create avenues for continuous knowledge exchange and shared problem-solving.

8.2.Discuss potential market advantages and security concerns.

In the realm of the K-BI project, the strategic implementation of Business Intelligence (BI) not only enhances the organization's competitive standing but also provides a range of potential market advantages. Simultaneously, this integration raises critical security concerns that must be addressed to ensure the confidentiality, integrity, and availability of sensitive data.

8.2.1. Potential Market Advantages:

Informed Decision-Making: Leveraging BI tools equips K-BI organizations with data-driven insights that underpin informed decision-making. This advantage resonates in the marketplace, as organizations that can demonstrate their ability to make strategic choices grounded in accurate and up-to-date information gain a competitive edge.

Agility and Responsiveness: BI enables organizations to swiftly adapt to shifting circumstances, whether related to emerging trends or evolving regulations. This agility positions K-BI organizations as nimble and capable of navigating the challenges posed by the pandemic, fostering a reputation for responsive crisis management.

Thought Leadership: Organizations that effectively utilize BI to generate meaningful insights become thought leaders in their respective fields. Sharing well-informed analyses, predictions, and policy recommendations elevates the organization's reputation, making it a sought-after partner for collaborations and international initiatives.

Enhanced Collaboration: BI tools facilitate data-driven collaborations that transcend geographical boundaries. K-BI organizations can form partnerships with international entities, leveraging shared data for coordinated responses. This collaborative prowess enhances market visibility and underscores the organization's commitment to global health.

Influence on Policies: The ability to advocate for evidence-based policies positions K-BI organizations as influential players in shaping public health strategies. This influence resonates in the market, as the organization's recommendations carry the weight of data-driven insights.

Improved Public Trust: Transparent communication of insights through BI engenders public trust. Organizations that are seen as transparent and responsible stewards of data are more likely to attract public confidence, enhancing their reputation and market credibility.

8.2.2. Security Concerns:

Data Privacy: BI involves the collection, processing, and sharing of sensitive data. Ensuring data privacy is paramount, especially in the healthcare sector. Unauthorized access to personal health information can lead to breaches, legal consequences, and loss of public trust.

Cybersecurity Threats: As BI tools store and transmit critical data, they become targets for cyberattacks. Malicious actors may attempt to compromise systems, leading to data breaches or disruptions in critical operations. Robust cybersecurity measures are essential to mitigate these risks.

Data Integrity: Maintaining the accuracy and integrity of data is crucial. Unauthorized modifications or tampering can lead to incorrect analyses and decisions. Ensuring data integrity through encryption, access controls, and validation processes is essential.

Availability: BI tools heavily rely on technology infrastructure. Downtime or disruptions in availability can impede critical decision-making and responses. Robust redundancy and disaster recovery plans are necessary to ensure continuous access to BI resources.

Compliance and Regulations: Healthcare data is subject to stringent regulations and compliance requirements, such as HIPAA (Health Insurance Portability and Accountability Act). Failure to adhere to these regulations can result in legal consequences and damage to the organization's reputation.

User Training and Awareness: BI tools are effective only when users understand their functionalities and best practices. Inadequate training can lead to unintentional data leaks or misuse, emphasizing the need for ongoing user education.

Vendor and Third-Party Risk: If BI tools are sourced from third-party vendors, their security practices and data handling must be thoroughly evaluated. Weaknesses in vendor security can pose risks to the organization's data.

Data Governance: Proper data governance frameworks must be in place to ensure that data is accurate, consistent, and used appropriately. Inaccurate or inconsistent data can lead to flawed analyses and misguided decisions .

References

Castillo, D., 2021. *https://www.seldon.io/machine-learning-regression-explained.* [Online]
Available at: https://www.seldon.io/machine-learning-regression-explained
[Accessed 20 August 2023].

Chopra, Y., 2021. *10 Ways the Advanced BI Tools Help in Decision Making.* [Online]
Available at: https://www.datatobiz.com/blog/advanced-bi-tools-help-in-decision-making/
[Accessed 20 August 2023].

DashThis, n.d. *How to build a stunning data analytics dashboard in minutes: A step-by-step guide.* [Online]
Available at: https://dashthis.com/blog/data-analytics-dashboard/
[Accessed 20 August 2023].

IBM, n.d. *What is data labeling?*. [Online]
Available at: https://www.ibm.com/topics/data-labeling
[Accessed 20 August 2023].

Kudan, N., 2023. *Your guide to data labeling methods.* [Online]
Available at: https://toloka.ai/blog/data-labeling-methods/
[Accessed 20 August 2023].

Oracle Help Center, n.d. *Using Tools to Configure Security in Oracle Business Intelligence.* [Online]
Available at: https://docs.oracle.com/middleware/12212/biee/BIESC/GUID-318425C8-720F-44C3-9C37-579A1AC167CB.htm#BIESC5965
[Accessed 20 August 2023].

Ray, S., 2015. *7 Regression Techniques You Should Know!*. [Online]
Available at: https://www.analyticsvidhya.com/blog/2015/08/comprehensive-guide-regression/
[Accessed 20 August 2023].

Rudenko, E., 2022. *Successful Business Intelligence Examples.* [Online]
Available at: https://nix-united.com/blog/examples-of-business-intelligence-systems-for-improving-decision-making/
[Accessed 20 August 2023].

Sharma, R., 2023. *What is Business Intelligence? Operations, Types of Users & Advantages.* [Online]
Available at: https://www.upgrad.com/blog/what-is-business-intelligence/
[Accessed 20 August 2023].

Stedman, C., 2023. ***What is business intelligence?.*** **[Online]**
Available at: https://www.techtarget.com/searchbusinessanalytics/definition/business-intelligence-BI
[Accessed 20 August 2023].

Tableau, n.d. ***Cigna embraces Tableau, improves healthcare affordability and care for 95M customers.*** **[Online]**
Available at: https://www.tableau.com/solutions/customer/cigna-embraces-tableau-improves-healthcare-affordability-and-care-for-95m-customers
[Accessed 20 August 2023].

Tableau, n.d. ***Guide To Data Cleaning: Definition, Benefits, Components, And How To Clean Your Data.*** **[Online]**
Available at: https://www.tableau.com/learn/articles/what-is-data-cleaning
[Accessed 20 August 2023].

Sampson, F. (2021). **Data Privacy and Security: Some Legal and Ethical Challenges.** In: Jahankhani, H., Kendzierskyj, S., Akhgar, B. (eds) Information Security Technologies for Controlling Pandemics. Advanced Sciences and Technologies for Security Applications. Springer, Cham. https://doi.org/10.1007/978-3-030-72120-6_4

Appendices

Data Source:
https://data.world/covid-19-data-resource-hub/covid-19-case-counts/workspace/file?filename=COVID-19+Cases.csv

Tableau Public:
https://public.tableau.com/app/profile/kiet.nguyen6834/viz/Dashboard_Covid-19/Dashboard

BUSINESS INTELLIGENCE

Nguyen Tuan Kiet – GDD210002

ABOUT US

K-BI, which stands for "Knowledge-Business Intelligence," is a specialized sub-organization established under the umbrella of the World Health Organization (WHO) to address the specific needs arising from the COVID19 pandemic.

DESCRIPTION FOR THE DATASET

Case_Type	People_Total_Tested_Count	Cases	Difference	Date	Combined_Key	Country_Region	Province_State	Admin2	iso2	iso3	FIPS	Lat	Long	Populatior	People_Hc	Data_	Prep_Flow_Runtime
Confirmed		6	0	5/22/2020	Western Sahara	Western Sahara	N/A		EH	ESH		24.2155	-12.8858	597330		2019	6/4/2020 23:15
Confirmed		0	0	2/3/2020	Switzerland	Switzerland	N/A		CH	CHE		46.8182	8.2275	8654618		2019	6/4/2020 23:15
Deaths		0	0	3/1/2020	Cyprus	Cyprus	N/A		CY	CYP		35.1264	33.4299	1207361		2019	6/4/2020 23:15
Confirmed		23	0	4/21/2020	Antigua and Barbuda	Antigua and Barbuda	N/A		AG	ATG		17.0608	-61.7964	97928		2019	6/4/2020 23:15
Deaths		56	0	5/11/2020	Thailand	Thailand	N/A		TH	THA		15.87003	100.9925	69799978		2019	6/4/2020 23:15
Deaths		0	0	2/11/2020	Jamaica	Jamaica	N/A		JM	JAM		18.1096	-77.2975	2961161		2019	6/4/2020 23:15
Confirmed		0	0	2/6/2020	Belize	Belize	N/A		BZ	BLZ		17.1899	-88.4976	397621		2019	6/4/2020 23:15
Confirmed		1	0	3/18/2020	Central African Republic	Central African Republic	N/A		CF	CAF		6.6111	20.9394	4829764		2019	6/4/2020 23:15
Confirmed		23	0	6/2/2020	Grenada	Grenada	N/A		GD	GRD		12.1165	-61.679	112519		2019	6/4/2020 23:15
Confirmed		2710	19	5/9/2020	Greece	Greece	N/A		GR	GRC		39.0742	21.8243	10423056		2019	6/4/2020 23:15
Deaths		0	0	5/2/2020	Bonaire, Sint Eustatius and Saba, Netherlands	Netherlands	Bonaire, Sint Eustatius and Saba		BQ	BES		12.1784	-68.2385	26221		2019	6/4/2020 23:15
Deaths		5	1	4/13/2020	Bermuda, United Kingdom	United Kingdom	Bermuda		BM	BMU		32.3078	-64.7505	62273		2019	6/4/2020 23:15
Confirmed		1	0	3/19/2020	Central African Republic	Central African Republic	N/A		CF	CAF		6.6111	20.9394	4829764		2019	6/4/2020 23:15
Confirmed		18	0	4/29/2020	New Caledonia, France	France	New Caledonia		NC	NCL		-20.9043	165.618	285491		2019	6/4/2020 23:15
Deaths		31	9	3/30/2020	Poland	Poland	N/A		PL	POL		51.9194	19.1451	37846605		2019	6/4/2020 23:15
Deaths		3	0	4/9/2020	Manitoba, Canada	Canada	Manitoba		CA	CAN		53.7609	-98.8139	1377517		2019	6/4/2020 23:15
Deaths		0	0	2/19/2020	Tibet, China	China	Tibet		CN	CHN		31.6927	88.0924	3440000		2019	6/4/2020 23:15
Deaths		25	0	4/25/2020	Kazakhstan	Kazakhstan	N/A		KZ	KAZ		48.0196	66.9237	18776707		2019	6/4/2020 23:15
Confirmed		0	0	3/4/2020	South Sudan	South Sudan	N/A		SS	SSD		6.877	31.307	11193729		2019	6/4/2020 23:15
Deaths		0	0	2/23/2020	Inner Mongolia, China	China	Inner Mongolia		CN	CHN		44.0935	113.9448	25340000		2019	6/4/2020 23:15
Confirmed		15970	402	4/27/2020	Ontario, Canada	Canada	Ontario		CA	CAN		51.2538	-85.3232	14711827		2019	6/4/2020 23:15
Deaths		6	0	5/10/2020	Hebei, China	China	Hebei		CN	CHN		37.8957	114.9042	75560000		2019	6/4/2020 23:15
Confirmed		0	0	3/28/2020	Bonaire, Sint Eustatius and Saba, Netherlands	Netherlands	Bonaire, Sint Eustatius and Saba		BQ	BES		12.1784	-68.2385	26221		2019	6/4/2020 23:15
Confirmed		18863	1041	5/14/2020	Bangladesh	Bangladesh	N/A		BD	BGD		23.685	90.3563	1.65E+08		2019	6/4/2020 23:15
Deaths		47	8	4/19/2020	Croatia	Croatia	N/A		HR	HRV		45.1	15.2	4105268		2019	6/4/2020 23:15
Deaths		0	0	4/12/2020	Chad	Chad	N/A		TD	TCD		15.4542	18.7322	16425859		2019	6/4/2020 23:15
Deaths		109	10	4/13/2020	Hungary	Hungary	N/A		HU	HUN		47.1625	19.5033	9660350		2019	6/4/2020 23:15
Deaths		0	0	5/16/2020	Yukon, Canada	Canada	Yukon		CA	CAN		64.2823	-135	41078		2019	6/4/2020 23:15
Deaths		0	0	2/7/2020	Sudan	Sudan	N/A		SD	SDN		12.8628	30.2176	43849269		2019	6/4/2020 23:15
Deaths		3	0	5/31/2020	Sichuan, China	China	Sichuan		CN	CHN		30.6171	102.7103	83410000		2019	6/4/2020 23:15
Confirmed		0	0	2/7/2020	Aruba, Netherlands	Netherlands	Aruba		AW	ABW		12.5211	-69.9683	106766		2019	6/4/2020 23:15
Confirmed		1	0	5/25/2020	Tibet, China	China	Tibet		CN	CHN		31.6927	88.0924	3440000		2019	6/4/2020 23:15
Deaths		1	0	5/1/2020	Turks and Caicos Islands, United Kingdom	United Kingdom	Turks and Caicos Islands		TC	TCA		21.694	-71.7979	38718		2019	6/4/2020 23:15
Deaths		33998	384	5/15/2020	United Kingdom	United Kingdom	N/A		GB	GBR		55.3781	-3.436	67886004		2019	6/4/2020 23:15
Confirmed		6	0	4/15/2020	Saint Barthelemy, France	France	Saint Barthelemy		BL	BLM		17.9	-62.8333	9885		2019	6/4/2020 23:15
Confirmed		659	2	5/9/2020	Shanghai, China	China	Shanghai		CN	CHN		31.202	121.4491	24240000		2019	6/4/2020 23:15
Deaths		0	0	2/23/2020	Singapore	Singapore	N/A		SG	SGP		1.2833	103.8333	5850343		2019	6/4/2020 23:15
Deaths		0	0	1/24/2020	Hungary	Hungary	N/A		HU	HUN		47.1625	19.5033	9660350		2019	6/4/2020 23:15
Deaths		0	0	1/31/2020	Bermuda, United Kingdom	United Kingdom	Bermuda		BM	BMU		32.3078	-64.7505	62273		2019	6/4/2020 23:15
Deaths		0	0	3/8/2020	Bulgaria	Bulgaria	N/A		BG	BGR		42.7339	25.4858	6948445		2019	6/4/2020 23:15
Confirmed		15	0	4/17/2020	Saint Lucia	Saint Lucia	N/A		LC	LCA		13.9094	-60.9789	183629		2019	6/4/2020 23:15
Confirmed		24112	845	5/28/2020	Kuwait	Kuwait	N/A		KW	KWT		29.31166	47.48177	4270563		2019	6/4/2020 23:15
Confirmed		21648	734	4/24/2020	Peru	Peru	N/A		PE	PER		-9.19	-75.0152	32971846		2019	6/4/2020 23:15
Confirmed		1	0	2/8/2020	Nepal	Nepal	N/A		NP	NPL		28.1667	84.25	29136808		2019	6/4/2020 23:15

- **Case_Type**: This column specifies the type of COVID-19 case, distinguishing between confirmed infections, deaths, and possibly other classifications.
- **People_Total_Tested_Count**: The number of individuals who have undergone COVID-19 testing, providing insight into testing efforts and identifying potential trends in infection rates.
- **Cases Difference**: The difference in the number of cases compared to the previous recorded date, facilitating the observation of daily changes and trends.
- **Date**: The date on which the COVID-19 data was recorded, enabling temporal analysis and trend identification.
- **Combined_Key**: A key that uniquely identifies a specific location or region, facilitating data aggregation and regional comparisons.
- **Country_Region**: The name of the country or territory to which the data pertains, helping to categorize and compare data on a global scale.
- **Province_State**: If applicable, this column specifies the name of the administrative division within a country or region, offering a more granular analysis.
- **Admin2**: Further administrative division information, if relevant, to pinpoint specific localities within a province or state.
- **iso2 and iso3**: The ISO country codes, which are standardized country codes, aiding in data integration with external datasets.
- **FIPS**: The Federal Information Processing Standards code, used for data processing and linking with administrative boundaries in the United States.
- **Lat and Long**: Latitude and longitude coordinates of the region, essential for geographical mapping and spatial analysis.
- **Population_Count**: The total population count of the region, providing context for understanding the pandemic's impact on different population sizes.
- **People_Hospitalized_Cumulative_Count**: The cumulative number of individuals hospitalized due to COVID-19, offering insights into the severity of cases and healthcare system stress.
- **Data_Source**: The source from which the data is derived, ensuring transparency and facilitating data credibility.
- **Prep_Flow_Runtime**: Information related to data preparation and processing times, important for data quality assessment

PREPROCESSING AND DATA PREPARATION

Loading the Dataset:

```
og_data = pd.read_csv(
    'C:/Users/PC/Desktop/Learning/Tableau/Dashboard_Covid/COVID-19-Cases.csv',
low_memory=False)
```

Displaying the First 10 Rows:

```
Output exceeds the size limit. Open the full output data in a text editor
   Case_Type  People_Total_Tested_Count  Cases  Difference       Date  \
0  Confirmed                        NaN      6           0  5/22/2020
1  Confirmed                        NaN      0           0   2/3/2020
2     Deaths                        NaN      0           0   3/1/2020
3  Confirmed                        NaN     23           0  4/21/2020
4     Deaths                        NaN     56           0  5/11/2020
5     Deaths                        NaN      0           0  2/11/2020
6  Confirmed                        NaN      0           0   2/6/2020
7  Confirmed                        NaN      1           0  3/18/2020
8  Confirmed                        NaN     23           0   6/2/2020
9  Confirmed                        NaN   2710          19   5/9/2020

               Combined_Key            Country_Region Province_State Admin2  \
0            Western Sahara            Western Sahara            NaN    NaN
1               Switzerland               Switzerland            NaN    NaN
2                    Cyprus                    Cyprus            NaN    NaN
3       Antigua and Barbuda       Antigua and Barbuda            NaN    NaN
4                  Thailand                  Thailand            NaN    NaN
5                   Jamaica                   Jamaica            NaN    NaN
6                    Belize                    Belize            NaN    NaN
7  Central African Republic  Central African Republic            NaN    NaN
8                   Grenada                   Grenada            NaN    NaN
9                    Greece                    Greece            NaN    NaN

  iso2 iso3  FIPS        Lat        Long  Population_Count  \
...
6  2019 Novel Coronavirus COVID-19 (2019-nCoV) Da...  6/4/2020 11:15:39 PM
7  2019 Novel Coronavirus COVID-19 (2019-nCoV) Da...  6/4/2020 11:15:39 PM
8  2019 Novel Coronavirus COVID-19 (2019-nCoV) Da...  6/4/2020 11:15:39 PM
9  2019 Novel Coronavirus COVID-19 (2019-nCoV) Da...  6/4/2020 11:15:39 PM
```

STATISTICAL DATA DESCRIPTION

Overview of the dataset's structure

These statistics provide insights into the central tendency and distribution of the data, helping to identify outliers or anomalies.

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 950670 entries, 0 to 950669
Data columns (total 18 columns):
 #   Column                              Non-Null Count   Dtype
---  ------                              --------------   -----
 0   Case_Type                           950670 non-null  object
 1   People_Total_Tested_Count           6048 non-null    float64
 2   Cases                               950670 non-null  int64
 3   Difference                          950670 non-null  int64
 4   Date                                950670 non-null  object
 5   Combined_Key                        950670 non-null  object
 6   Country_Region                      950670 non-null  object
 7   Province_State                      901260 non-null  object
 8   Admin2                              878580 non-null  object
 9   iso2                                949590 non-null  object
 10  iso3                                949860 non-null  object
 11  FIPS                                849690 non-null  float64
 12  Lat                                 921780 non-null  float64
 13  Long                                921780 non-null  float64
 14  Population_Count                    921780 non-null  float64
 15  People_Hospitalized_Cumulative_Count  6048 non-null    float64
 16  Data_Source                         950670 non-null  object
 17  Prep_Flow_Runtime                   950670 non-null  object
dtypes: float64(6), int64(2), object(10)
memory usage: 130.6+ MB
None
```

STATISTICAL DATA DESCRIPTION

Basic statistics for numerical columns

- **Cases**: The mean number of confirmed cases is approximately 273, with a standard deviation of 5187. The minimum is 0, and the maximum is 614941.
- **Difference**: The mean difference (likely daily case changes) is around 7.4, with a standard deviation of 166.49. The minimum is -10034, indicating potential errors or negative values.
- **FIPS**: The Federal Information Processing Standards code statistics show the range of values and the standard deviation.

```
Output exceeds the size limit. Open the full output data in a text editor
       People_Total_Tested_Count          Cases     Difference           FIPS  \
count               6.048000e+03  950670.000000  950670.000000  849690.000000
mean                1.672996e+05     273.518654       7.388438   30335.484906
std                 2.643341e+05    5187.497792     166.491030   15196.097280
min                 3.000000e+00       0.000000  -10034.000000      60.000000
25%                 2.980700e+04       0.000000       0.000000   18169.000000
50%                 7.798700e+04       0.000000       0.000000   29171.000000
75%                 1.897678e+05       4.000000       0.000000   45079.000000
max                 2.293032e+06  614941.000000   33274.000000   56045.000000

                 Lat           Long  Population_Count  \
count  921780.000000  921780.000000      9.217800e+05
mean       37.115970     -83.442661      2.260264e+06
std         9.763286      38.621432      2.661885e+07
min       -51.796300    -174.159600      8.600000e+01
25%        34.068596     -97.698949      1.202300e+04
50%        38.154335     -89.205569      2.969950e+04
75%        41.772338     -81.926357      9.792800e+04
max        71.706900     178.065000      1.380004e+09

       People_Hospitalized_Cumulative_Count
count                           6048.000000
mean                            2625.715278
std                             9465.118234
min                                0.000000
...
25%                                0.000000
50%                              325.500000
75%                             1697.500000
max                            89995.000000
```

STATISTICAL DATA DESCRIPTION

Statistics for categorical columns

- **Case_Type**: There are two unique case types (Confirmed and Deaths) with a frequency of occurrences.
- **Country_Region**: The most common country region is the US.

	Case_Type	Date	Combined_Key	Country_Region	Province_State
count	950670	950670	950670	950670	901260
unique	2	135	3521	187	135
top	Confirmed	5/22/2020	Western Sahara	US	Texas
freq	475335	7042	270	879930	69120

	Admin2	iso2	iso3
count	878580	949590	949860
unique	1901	215	216
top	Unassigned	US	USA
freq	13770	878580	878580

	Data_Source
count	950670
unique	1
top	2019 Novel Coronavirus COVID-19 (2019-nCoV) Da ...
freq	950670

	Prep_Flow_Runtime
count	950670
unique	1
top	6/4/2020 11:15:39 PM
freq	950670

MISSING DATA HANDLING

Identifying Columns with Missing Values

```
# Find columns with missing values
columns_with_missing_values = og_data.columns[og_data.isnull().sum() > 0]

# Display the list of columns with missing values
print("Columns with missing values:")
print(columns_with_missing_values)
```

```
Columns with missing values:
Index(['People_Total_Tested_Count', 'Province_State', 'Admin2', 'iso2', 'iso3',
       'FIPS', 'Lat', 'Long', 'Population_Count',
       'People_Hospitalized_Cumulative_Count'],
      dtype='object')
```

Dropping Columns with Missing Values:

```
# Drop columns with missing values: People_Total_Tested_Count,
People_Hospitalized_Cumulative_Count
og_data.drop(columns=['People_Total_Tested_Count',
                      'People_Hospitalized_Cumulative_Count'], inplace=True)
```

MISSING DATA HANDLING

Filling Missing Values in 'Population_Count' Column

```
og_data['Population_Count'].fillna(og_data['Population_Count'].mean(),
inplace=True)
```

Displaying Missing Values After Handling:

```
print(og_data.isnull().sum()*100/len(og_data))
```

```
Case_Type             0.000000
Cases                 0.000000
Difference            0.000000
Date                  0.000000
Combined_Key          0.000000
Country_Region        0.000000
Province_State        5.197387
Admin2                7.583073
iso2                  0.113604
iso3                  0.085203
FIPS                 10.621982
Lat                   3.038909
Long                  3.038909
Population_Count      0.000000
Data_Source           0.000000
Prep_Flow_Runtime     0.000000
dtype: float64
```

DATA CONSISTENCY CHECK

Check for inconsistencies in the data, such as incorrect values or formats. This step helps maintain data accuracy and reliability

```
print(og_data['Case_Type'].unique())
```

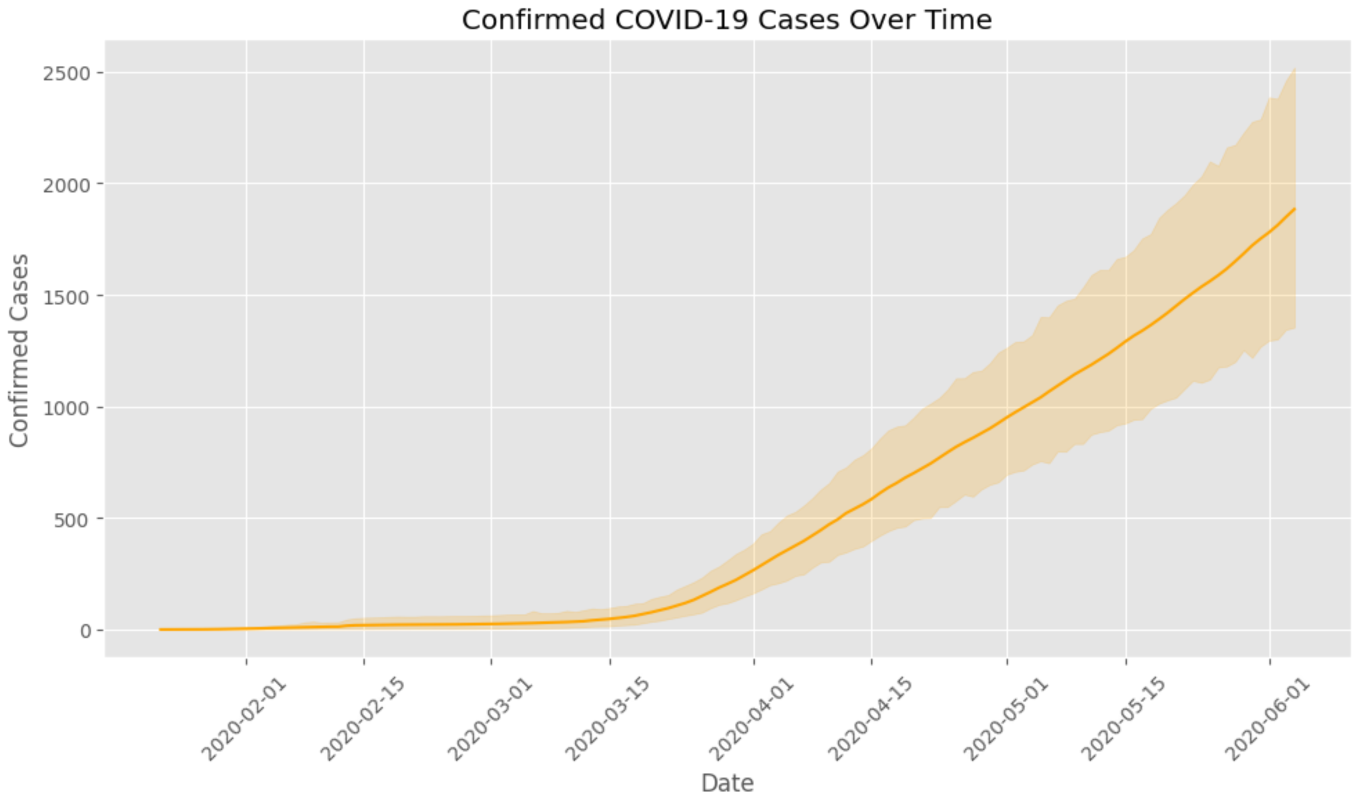
```
['Confirmed' 'Deaths']
```

The 'Case_Type' column in the COVID-19 dataset contains two unique values: 'Confirmed' and 'Deaths'. These values indicate the type of COVID-19 cases reported in the dataset. The presence of only these two distinct values suggests that the dataset includes records for confirmed infections and reported deaths due to COVID-19.

VISUALIZE DATA

The Confirmed Cases in the World Chart

```
# Subplot between 'Case_Type' and Date

## 'Confirmed' cases and 'Date' are plotted
### Filter the data to only include 'Confirmed' cases
confirmed_data = backup_df[backup_df['Case_Type'] == 'Confirmed']
# Convert the 'Date' column to datetime type
confirmed_data['Date'] = pd.to_datetime(confirmed_data['Date'])

# Sort the data by date in ascending order
confirmed_data = confirmed_data.sort_values(by='Date')

# Create the subplot
fig, ax = plt.subplots(figsize=(10, 6))

# Plot the data using seaborn's lineplot
sns.lineplot(x='Date', y='Cases', data=confirmed_data, ax=ax, color='orange')

# Set labels and title
ax.set_xlabel('Date')
ax.set_ylabel('Confirmed Cases')
ax.set_title('Confirmed COVID-19 Cases Over Time')

# Rotate x-axis labels for better visibility
plt.xticks(rotation=45)

# Show the plot
plt.tight_layout()
plt.show()
```

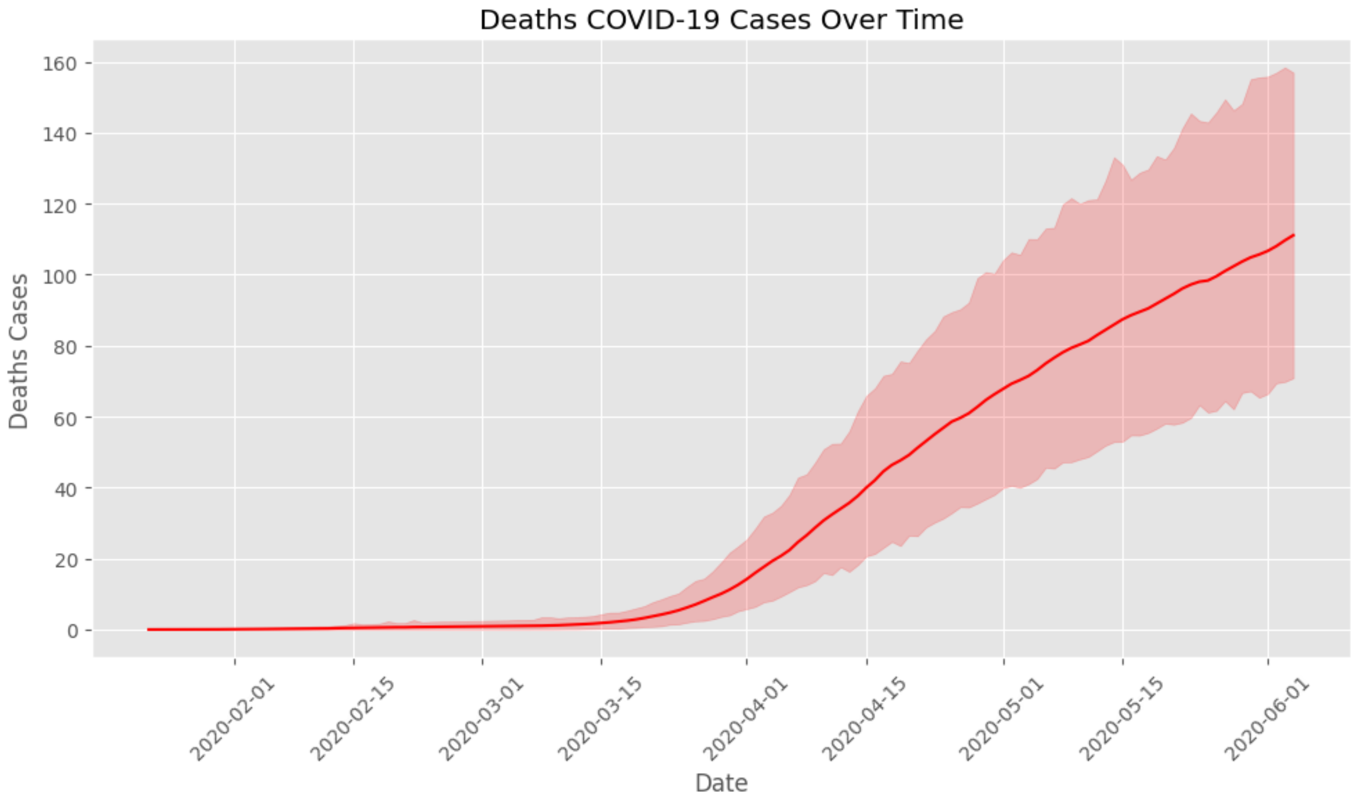


The line graph illustrates the number of confirmed COVID-19 cases over a period of six months from January to June 2020. It is clear that the number of cases increased steadily throughout the time frame. In January, there were only a few confirmed cases of COVID-19, less than 100. However, by February, the number of cases rose sharply to over 500. The trend continued in March and April, reaching over 1000 and 1500 cases respectively. In May, the number of cases slowed down slightly, but still increased to over 2000. Finally, in June, the number of cases reached its peak at around 2500.

VISUALIZE DATA

The Deaths Cases in the World Chart

```
## 'Confirmed' cases and 'Date' are plotted
### Filter the data to only include 'Confirmed' cases
deaths_data = backup_df[backup_df['Case_Type'] == 'Deaths']
### Convert the 'Date' column to datetime type
deaths_data['Date'] = pd.to_datetime(deaths_data['Date'])

### Sort the data by date in ascending order
deaths_data = deaths_data.sort_values(by='Date')

### Create the subplot
fig, ax = plt.subplots(figsize=(10, 6))

### Plot the data using seaborn's lineplot
sns.lineplot(x='Date', y='Cases', data=deaths_data, color='red', ax=ax)

### Set labels and title
ax.set_xlabel('Date')
ax.set_ylabel('Deaths Cases')
ax.set_title('Deaths COVID-19 Cases Over Time')

### Rotate x-axis labels for better visibility
plt.xticks(rotation=45)

### Show the plot
plt.tight_layout()
plt.show()
```

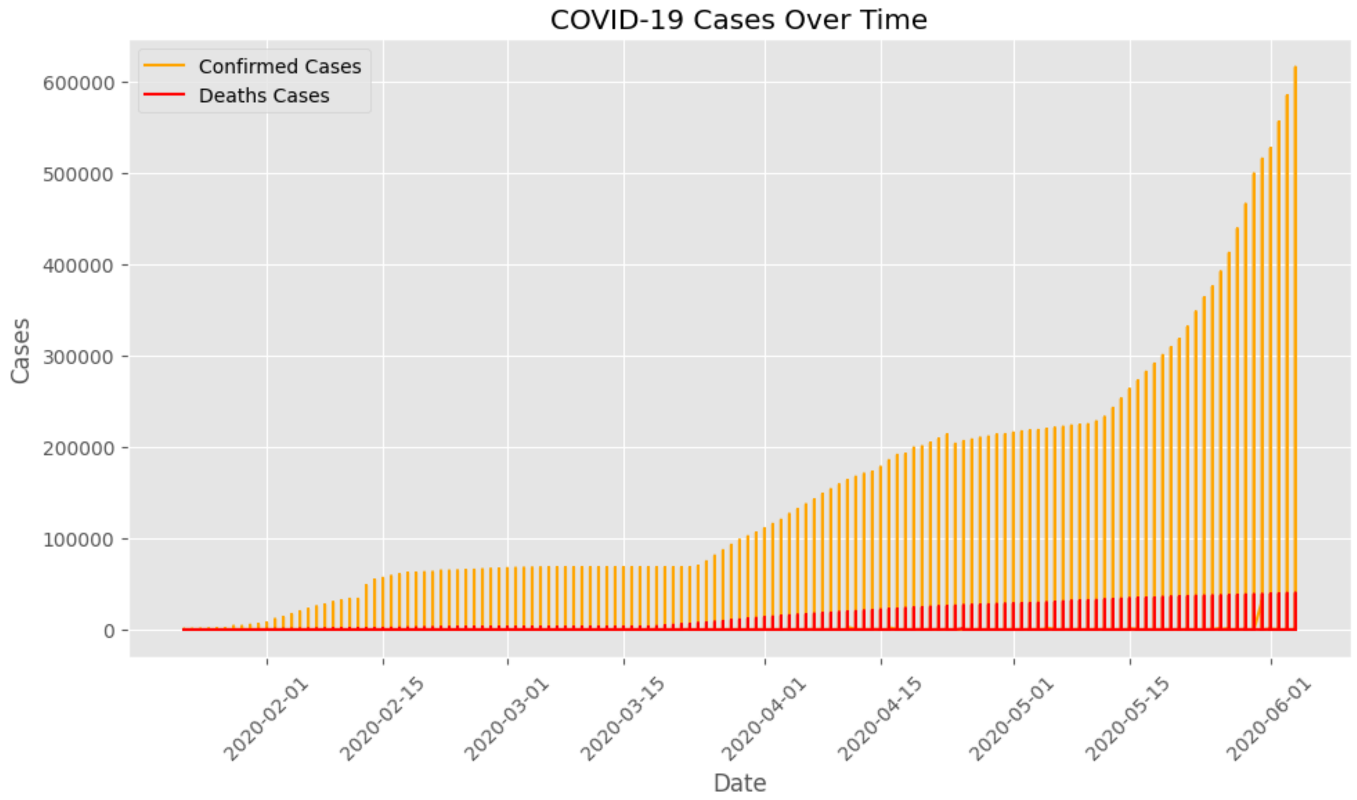


The line graph depicts the number of deaths due to COVID-19 cases over a period of six months from January to June 2020. It is evident that the number of deaths increased dramatically over the time frame. In January, there were no deaths reported due to COVID-19. However, by February, the number of deaths rose to about 20. The trend accelerated in March and April, reaching over 80 and 120 deaths respectively. In May, the number of deaths slowed down slightly, but still increased to over 140. Finally, in June, the number of deaths reached its highest point at around 160.

VISUALIZE DATA

The Combination of Confirmed Cases and Death Cases in the World Chart

```python
## 'Confirmed' cases and 'Date' are plotted
# Convert the 'Date' column to datetime type
backup_df['Date'] = pd.to_datetime(backup_df['Date'])

# Filter the data to only include 'Confirmed' and 'Deaths' cases
confirmed_data = backup_df[backup_df['Case_Type'] == 'Confirmed']
deaths_data = backup_df[backup_df['Case_Type'] == 'Deaths']

# Sort the data by date in ascending order for both datasets
confirmed_data = confirmed_data.sort_values(by='Date')
deaths_data = deaths_data.sort_values(by='Date')

# Create the subplot
fig, ax = plt.subplots(figsize=(10, 6))

# Plot the 'Confirmed' data using Matplotlib's plot with orange color
plt.plot(confirmed_data['Date'], confirmed_data['Cases'],
         color='orange', label='Confirmed Cases')

# Plot the 'Deaths' data using Matplotlib's plot with red color
plt.plot(deaths_data['Date'], deaths_data['Cases'],
         color='red', label='Deaths Cases')

# Set labels and title
ax.set_xlabel('Date')
ax.set_ylabel('Cases')
ax.set_title('COVID-19 Cases Over Time')

# Rotate x-axis labels for better visibility
plt.xticks(rotation=45)

# Show the legend
ax.legend()

# Show the plot
plt.tight_layout()
plt.show()
```



The line graph compares the number of confirmed COVID-19 cases and deaths over a period of eight months from January to August 2020. It is obvious that the number of cases increased exponentially over the time frame, while the number of deaths increased more slowly. In January, there were almost no confirmed cases or deaths due to COVID-19. However, by March, the number of cases surged to over 100000, while the number of deaths remained below 1000. The trend continued in April and May, reaching over 300000 and 20000 cases and deaths respectively. In June and July, the number of cases soared to over 500000, while the number of deaths rose to over 30000. Finally, in August, the number of cases reached its peak at around 600000, while the number of deaths reached its highest point at around 40000.

VISUALIZE DATA

Top 5 States (by ISO2) with Most Confirmed COVID-19 Cases

```
# Find the top5 states with the highest number of confirmed cases
confirmed_data = backup_df[backup_df['Case_Type'] == 'Confirmed']
top5_state_confirmed = confirmed_data.groupby(
  ['iso2'])['Cases'].sum().sort_values(ascending=False).head(5)

print(top5_state_confirmed)
```

```
iso2
US    73015191
ES    14042683
IT    13991688
GB    11590498
BR    11310710
Name: Cases, dtype: int64
```

VISUALIZE DATA

Top 5 Region (by ISO2) with Most Confirmed COVID-19 Cases

```python
# Find top5 Contry_Region with most confirmed cases over the period of time
confirmed_data = backup_df[backup_df['Case_Type'] == 'Confirmed']
top5_region_confirmed = confirmed_data.groupby(
  ['Country_Region'])['Cases'].sum().sort_values(ascending=False).head(5)

print(top5_region_confirmed)
```

```
Country_Region
US                  73158080
Spain               14042683
Italy               13991688
United Kingdom      11634123
Brazil              11310710
Name: Cases, dtype: int64
```

VISUALIZE DATA

Top 5 States (by ISO2) with Most Deaths COVID-19 Cases

```python
# Find the top5 states with the highest number of death cases
deaths_data = deaths_data.sort_values(by='Date')
top5_state_deaths = deaths_data.groupby(
  ['iso2'])['Cases'].sum().sort_values(ascending=False).head(5)

print(top5_state_deaths)
```

```
iso2
US    4156420
IT    1868769
GB    1676426
ES    1546134
FR    1466849
Name: Cases, dtype: int64
```

VISUALIZE DATA

Top 5 Region (by ISO2) with Most Deaths COVID-19 Cases

```
# Find top5 Contry_Region with most death cases over the period of time
deaths_data = deaths_data.sort_values(by='Date')
top5_region_deaths = deaths_data.groupby(
  ['Country_Region'])['Cases'].sum().sort_values(ascending=False).head(5)

print(top5_region_deaths)
```

```
Country_Region
US                 4162893
Italy              1868769
United Kingdom     1678116
Spain              1546134
France             1469256
Name: Cases, dtype: int64
```

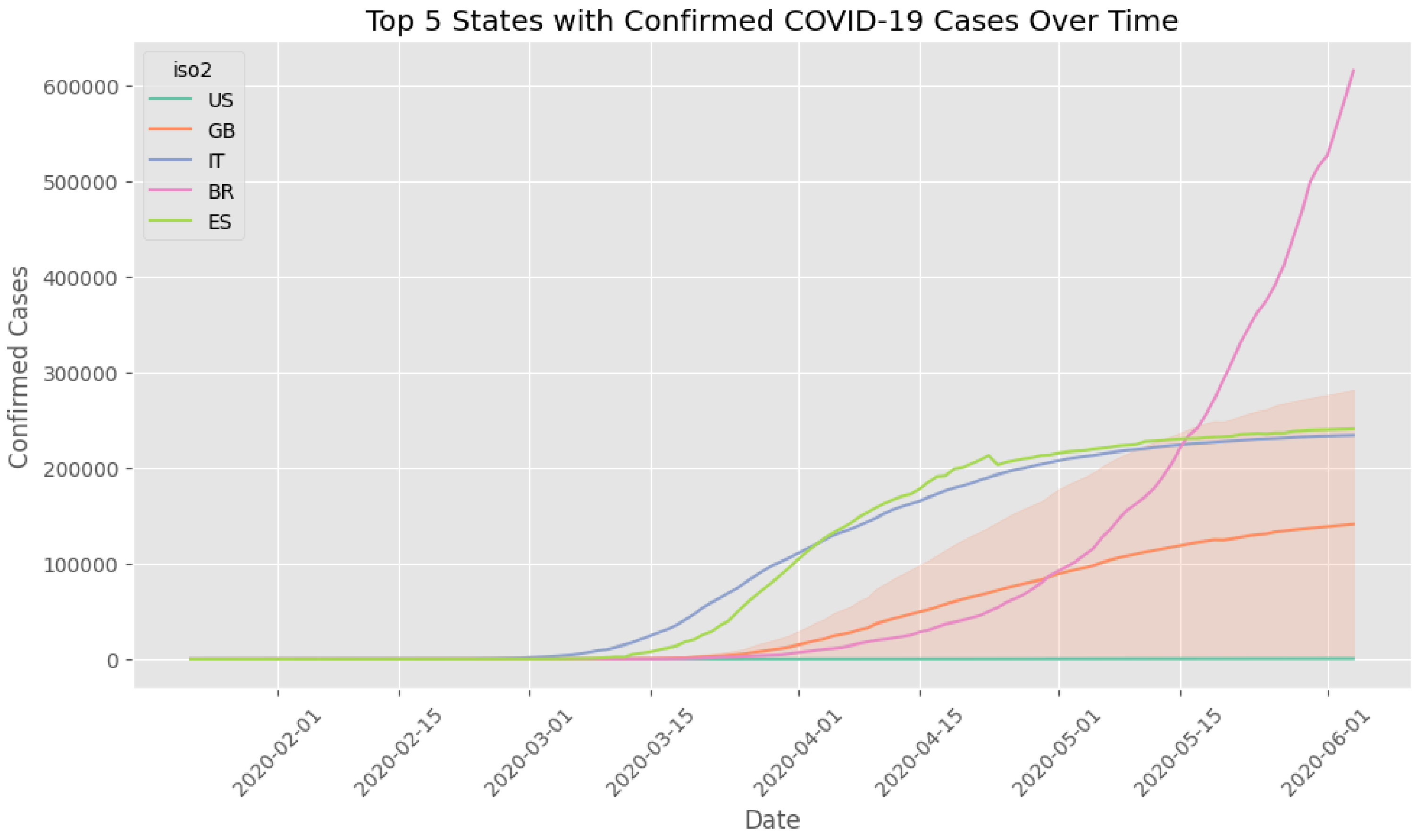
VISUALIZE DATA

Top 5 States with Confirmed COVID-19 Cases Over Time

```
confirmed_data = backup_df[backup_df['Case_Type'] == 'Confirmed']
top5_state_confirmed = confirmed_data.groupby(
  ['iso2'])['Cases'].sum().sort_values(ascending=False).head(5)

# Filter the data to only include the top5 states
top5_state_confirmed_data = confirmed_data[confirmed_data['iso2'].isin(
  top5_state_confirmed.index)]

# Sort the data by date in ascending order
top5_state_confirmed_data = top5_state_confirmed_data.sort_values(by='Date')

# Create the subplot
fig, ax = plt.subplots(figsize=(10, 6))

# Plot the data using seaborn's lineplot by plot.express
sns.lineplot(x='Date', y='Cases', data=top5_state_confirmed_data,
              hue='iso2', ax=ax, palette='Set2')

# Set labels and title
ax.set_xlabel('Date')
ax.set_ylabel('Confirmed Cases')
ax.set_title('Top 5 States with Confirmed COVID-19 Cases Over Time')

# Rotate x-axis labels for better visibility
plt.xticks(rotation=45)

# Show the plot
plt.tight_layout()
plt.show()
```

VISUALIZE DATA

The line chart shows the number of confirmed COVID-19 cases in the top 5 countries (US, GB, IT, BR, ES) from February 1, 2020 to June 1, 2020. The chart shows that the BR had the highest number of cases throughout the period, reaching over 6 million cases by June 1. ES had the second highest number of cases, followed by Italy and Great Britain. The chart also shows that the cases increased exponentially in March and April, especially in the Italy and ES.



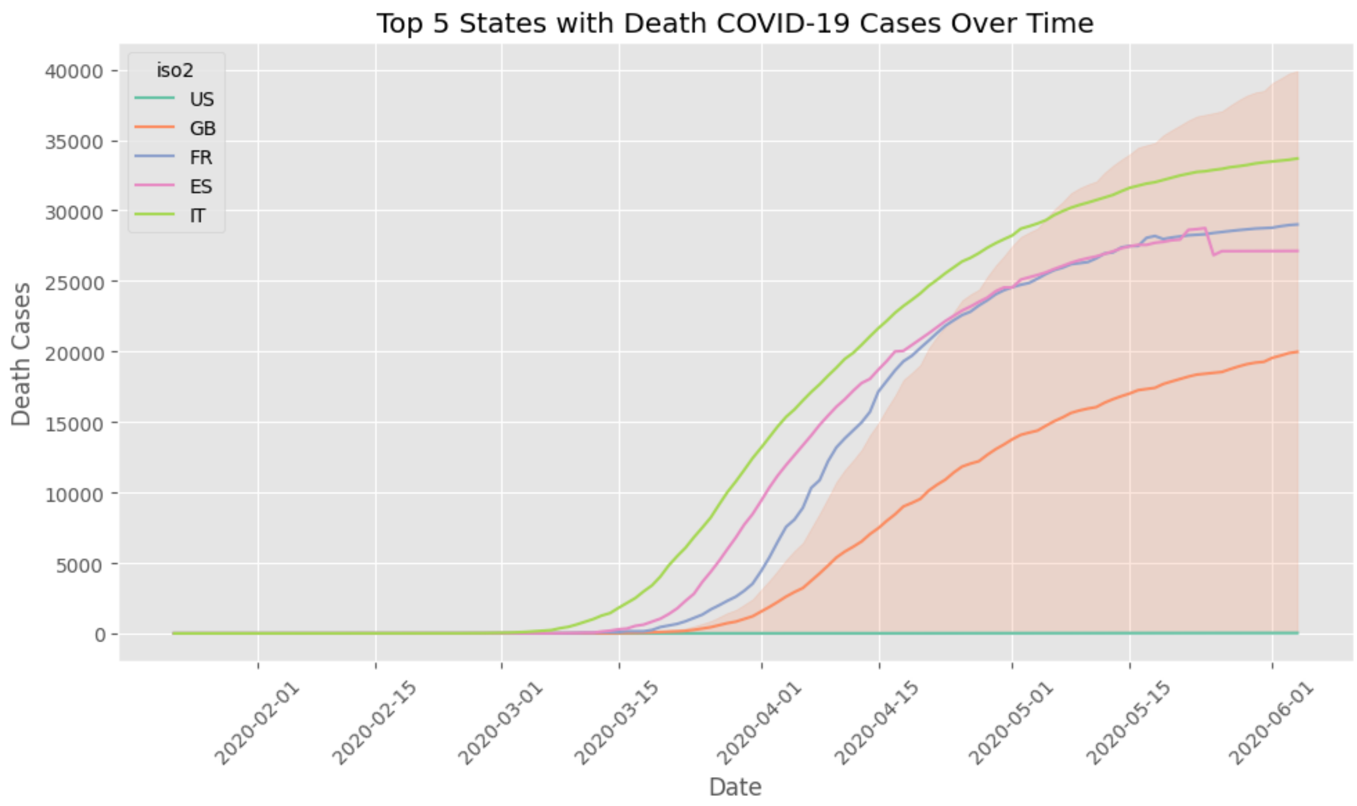
VISUALIZE DATA

Top 5 States with Deaths COVID-19 Cases Over Time

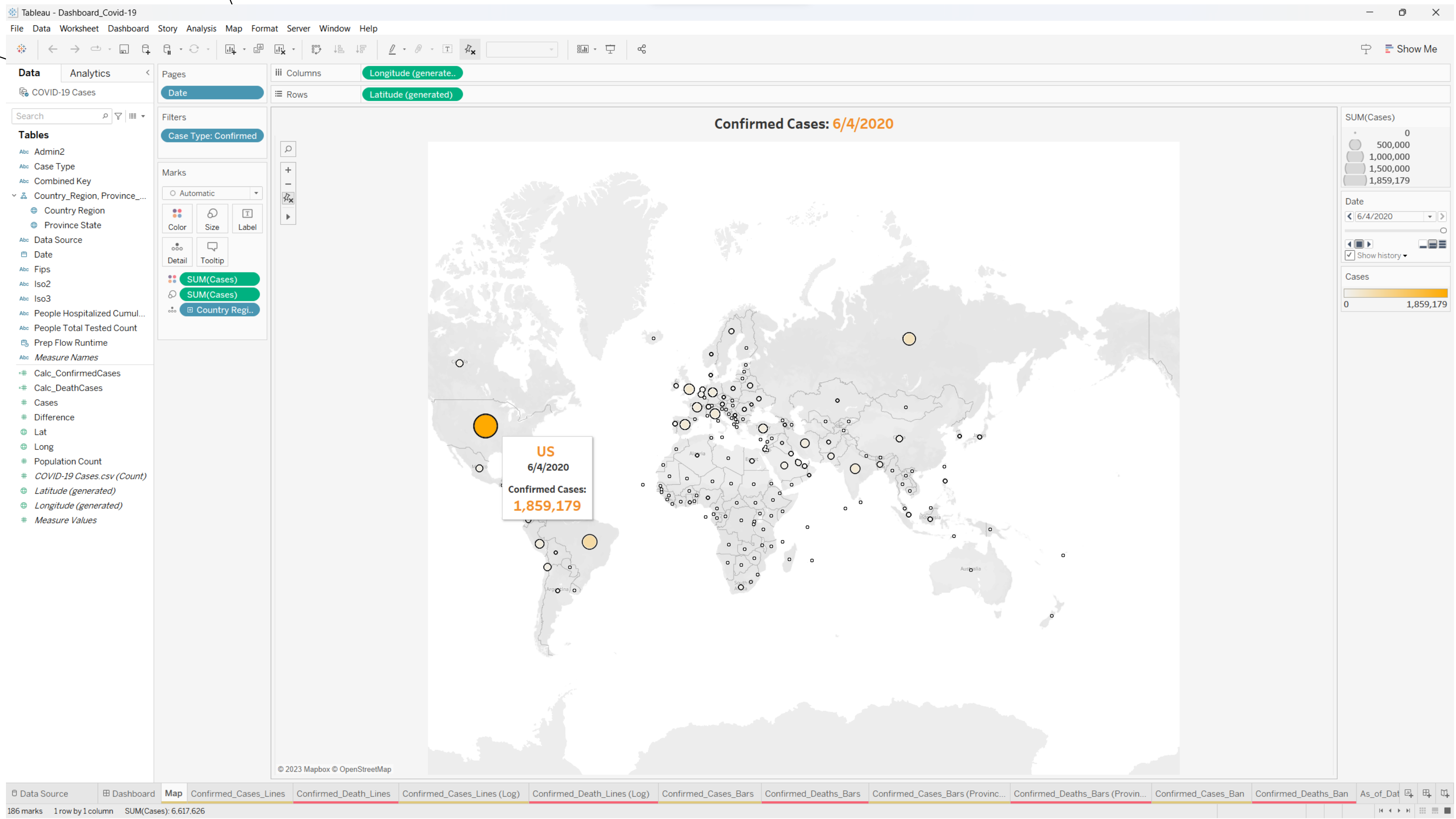
```
deaths_data = backup_df[backup_df['Case_Type'] == 'Deaths']
top5_state_deaths = deaths_data.groupby(
  ['iso2'])['Cases'].sum().sort_values(ascending=False).head(5)


# Filter the data to only include the top5 states
top5_state_deaths_data = deaths_data[deaths_data['iso2'].isin(
  top5_state_deaths.index)]

# Sort the data by date in ascending order
top5_state_deaths_data = top5_state_deaths_data.sort_values(by='Date')

# Create the subplot
fig, ax = plt.subplots(figsize=(10, 6))

# Plot the data using seaborn's lineplot by plot.express
sns.lineplot(x='Date', y='Cases', data=top5_state_deaths_data,
              hue='iso2', ax=ax, palette='Set2')

# Set labels and title
ax.set_xlabel('Date')
ax.set_ylabel('Death Cases')
ax.set_title('Top 5 States with Death COVID-19 Cases Over Time')

# Rotate x-axis labels for better visibility
plt.xticks(rotation=45)

# Show the plot
plt.tight_layout()
plt.show()
```

VISUALIZE DATA

The line chart shows the number of deaths by COVID-19 in the top 5 countries (IT, FR, ES, GB, US) from February 1, 2020, to June 1, 2020. The chart shows that the IT had the highest number of deaths throughout the period, reaching over 30,000 deaths by June 1. FR had the second highest number of deaths, followed by ES, GR, and US. The chart also shows that the deaths increased sharply in March and April, especially in Italy and Spain.

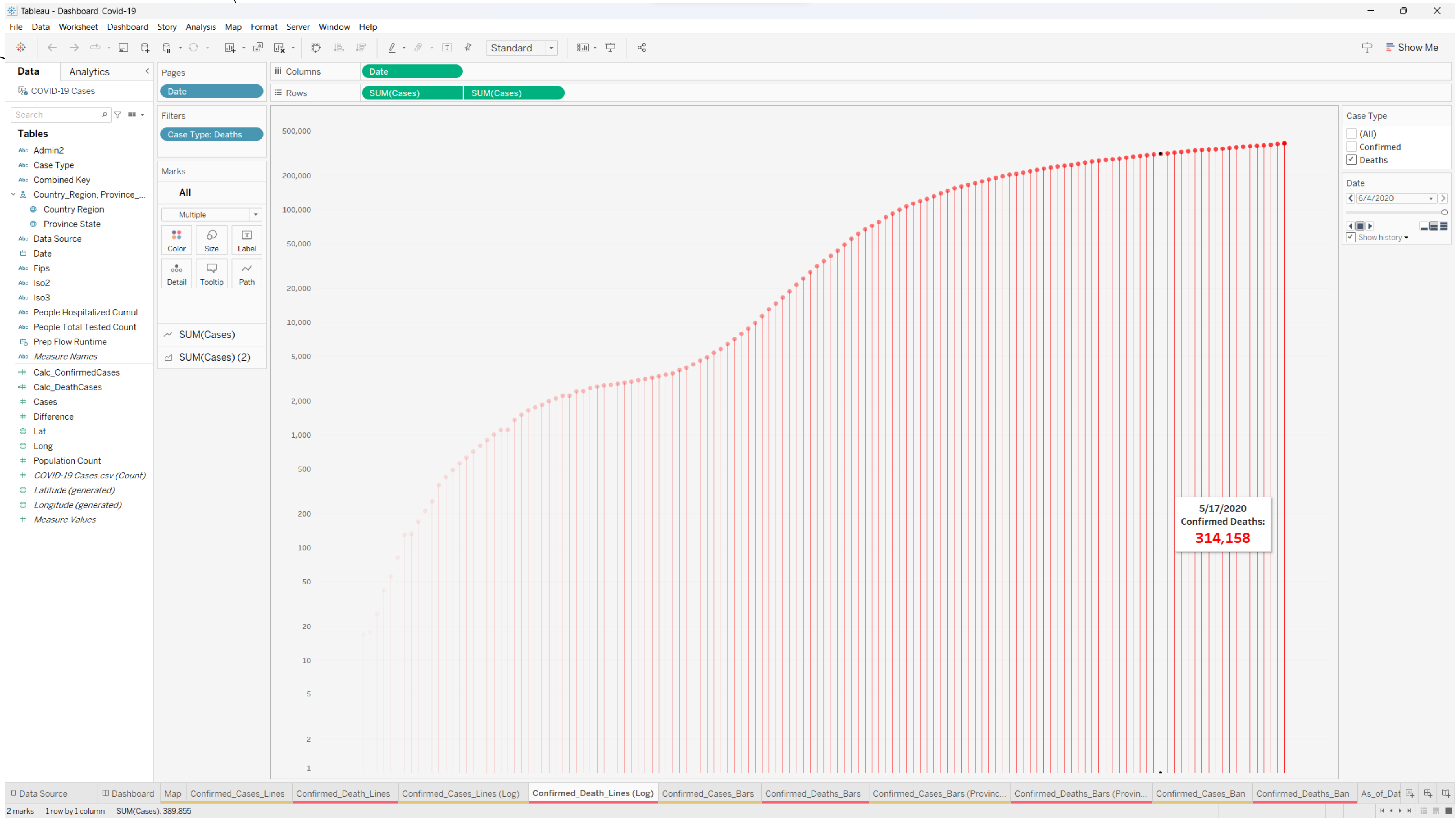


DESIGNING DASHBOARD WITH INTERACTIVE CHARTS

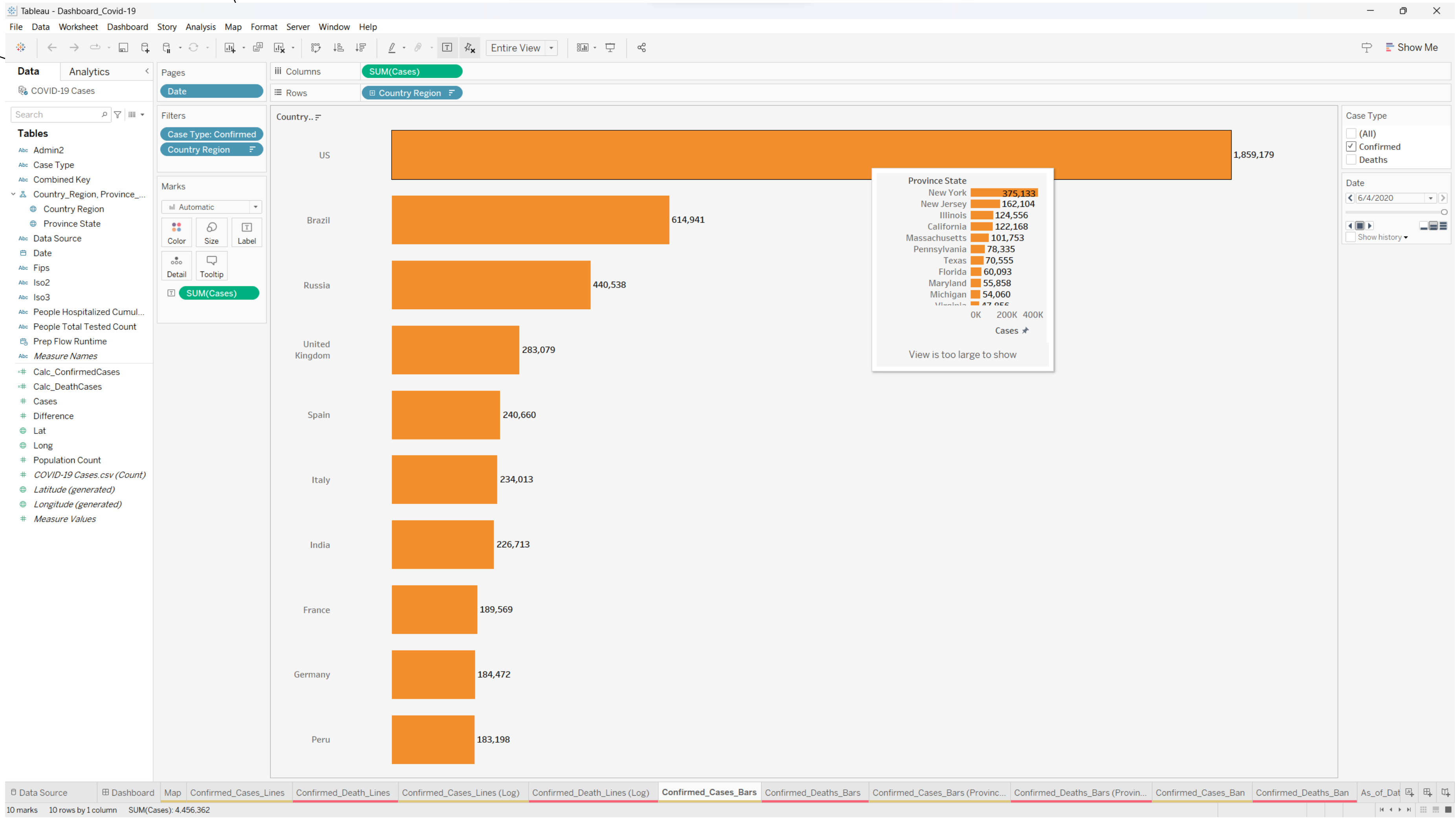


DESIGNING DASHBOARD WITH INTERACTIVE CHARTS

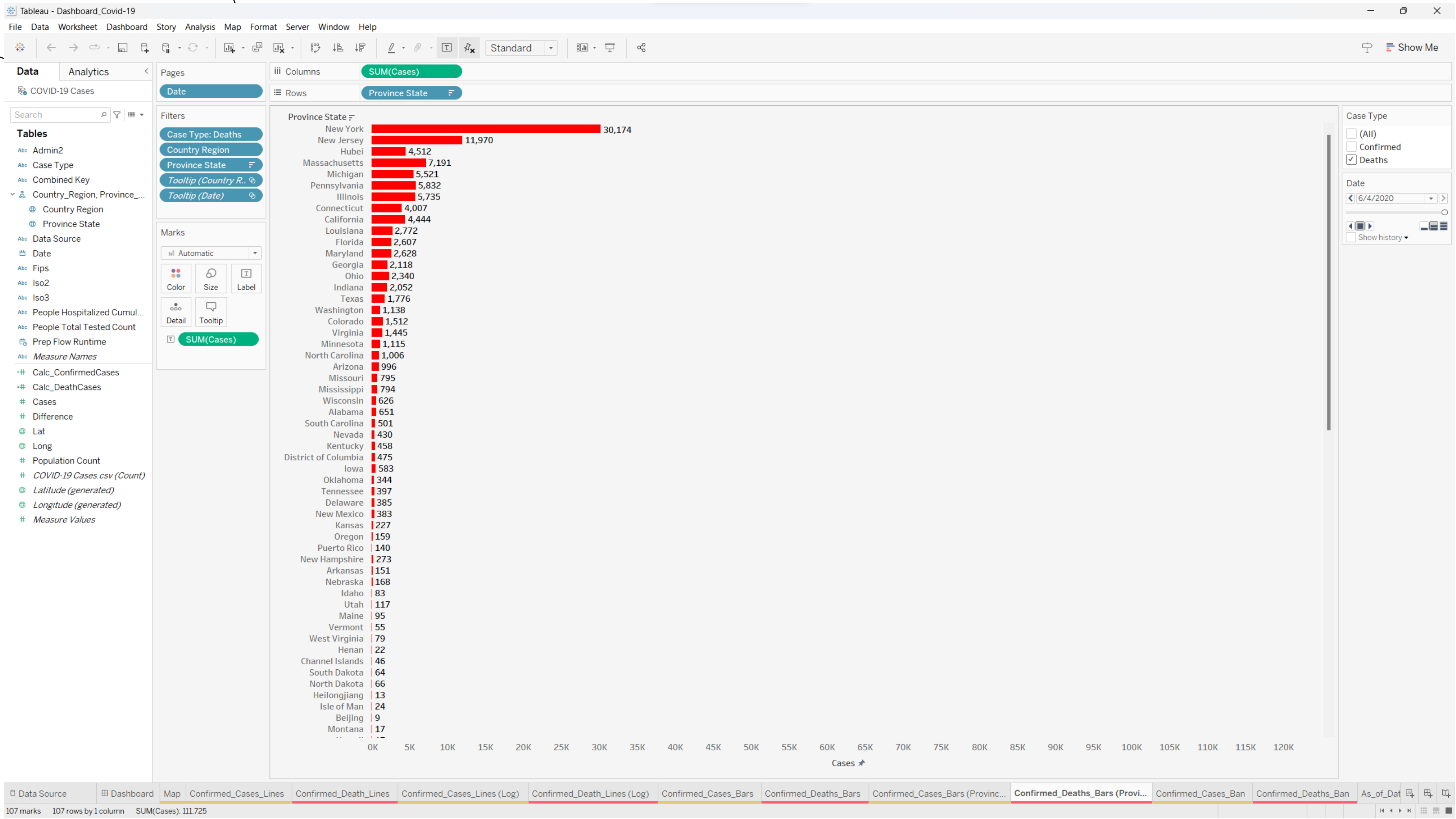
DESIGNING DASHBOARD WITH INTERACTIVE CHARTS



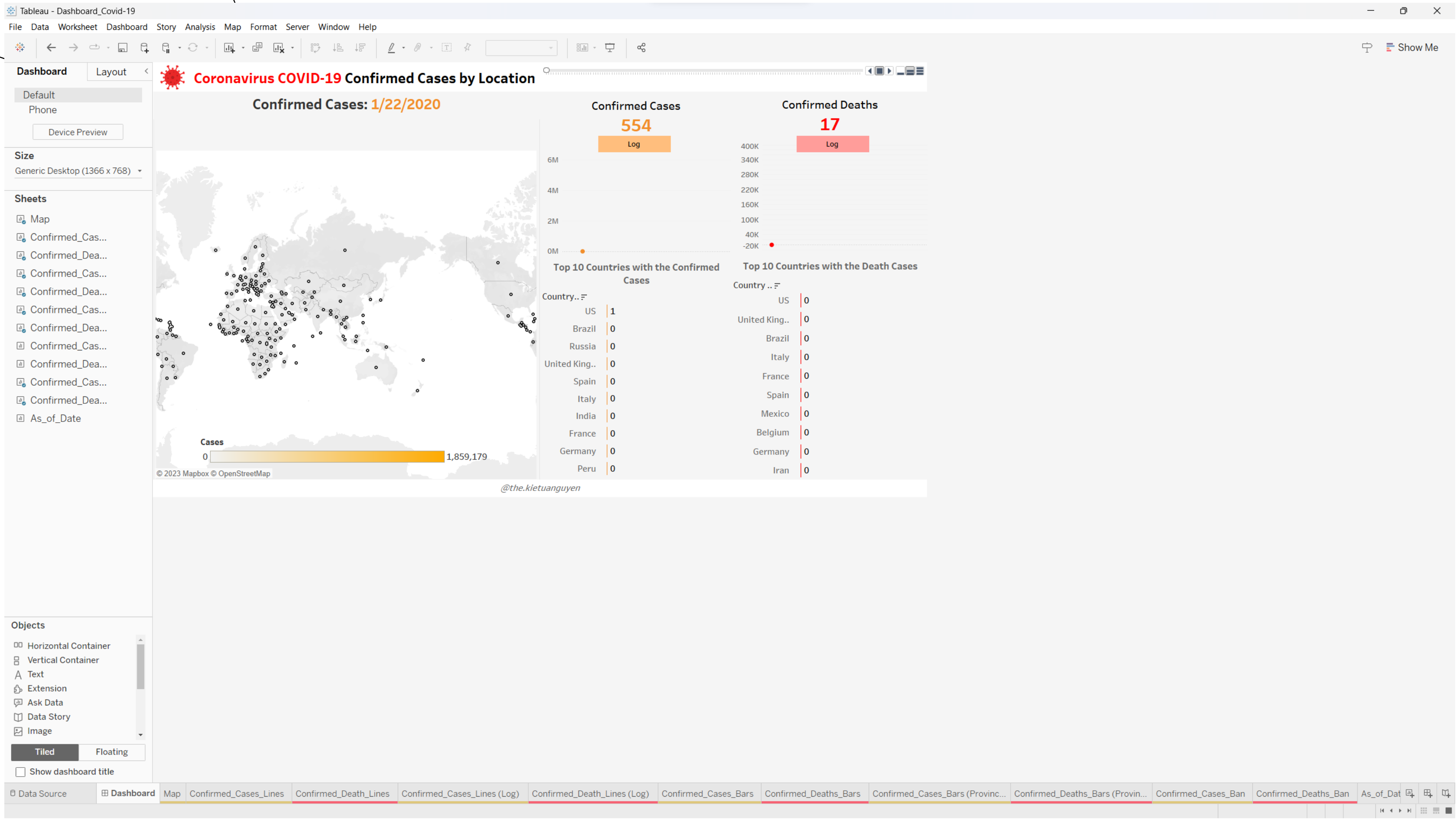
DESIGNING DASHBOARD WITH INTERACTIVE CHARTS



DESIGNING DASHBOARD WITH INTERACTIVE CHARTS



K-BI DASHBOARD



K-BI DASHBOARD

Purpose of Dashboards in Decision-Making for K-BI:

- **Data Visualization and Communication**: Dashboards condense complex data into intuitive charts, graphs, and visualizations. For K-BI, this means presenting COVID-19 case statistics, trends, and patterns in a visually compelling way, making it easier for decision-makers to understand and communicate critical information.
- **Real-Time Monitoring**: Dashboards can provide real-time or near-real-time updates on COVID-19 cases globally. This enables health officials, policymakers, and other stakeholders to monitor the situation in real-time and respond promptly to emerging trends and developments.
- **Identification of Hotspots and Trends**: With dashboards, decision-makers can quickly identify regions with higher infection rates, emerging hotspots, and changing trends. This helps allocate resources and interventions where they are needed most.
- **Resource Allocation**: Dashboards allow decision-makers to allocate resources efficiently. By visualizing data on hospital capacities, medical supplies, and healthcare needs, K-BI can help health systems make informed decisions on resource distribution.
- **Scenario Planning**: Dashboards enable scenario analysis by allowing decision-makers to explore "what-if" scenarios based on different assumptions and inputs. This aids in anticipating potential outcomes of various strategies and interventions.
- **Global Collaboration**: Dashboards can be accessed by stakeholders worldwide, promoting global collaboration and knowledge sharing. By providing a common platform for data interpretation, dashboards facilitate joint efforts in tackling the pandemic.

K-BI DASHBOARD

Structure of Designed Dashboards for K-BI:

- **Summary Metrics: Start with a section displaying key summary metrics, such as total confirmed cases, total deaths, and total recovered. These metrics provide an immediate overview of the global COVID-19 situation.**
- **Global Map: Include a world map with color-coded regions indicating the severity of COVID-19 cases. Users can quickly identify areas with high infection rates or outbreaks.**
- **Time Series Trends: Display time series charts illustrating the progression of cases over time. Use line charts to show trends in confirmed cases, deaths, and recoveries. Users can identify spikes, flattening curves, and potential second waves.**
- **Regional Breakdown: Provide a breakdown of cases by country or region. Use bar charts or pie charts to display the distribution of cases, enabling users to compare different regions.**
- **Hospital Capacity: Include visuals on hospital bed occupancy, ventilator availability, and other healthcare resources. This helps decision-makers allocate resources strategically.**
- **Testing and Tracing: Show data on testing rates, contact tracing efforts, and testing positivity rates. These insights inform public health strategies and intervention effectiveness.**
- **Vaccination Progress: If relevant, incorporate a section on vaccination progress, displaying the number of vaccinated individuals by region and vaccine type.**
- **Interactive Filters: Enable users to filter data by region, time period, and specific metrics of interest. Interactive filters enhance user engagement and customization.**
- **Annotations and Insights: Provide annotations to explain notable data points, trends, and changes. Offer insights on policy implications and potential interventions.**
- **References and Sources: Clearly indicate the sources of data and references used in the dashboard. Transparency in data sources builds trust in the information presented.**

ENHANCING USER-FRIENDLINESS AND FUNCTIONALITY

Customization for User-Friendliness

User-Centric Features:

- Intuitive Navigation:
- Interactive Filters:
- Zoom and Pan on Maps:
- Tool Tips and Legends:
- Time Series Analysis:
- Responsive Design:

Functional Interface Considerations

- Comprehensive Data Exploration:
- User-Centric Navigation:
- Interactive Filters and Parameters:
- Contextual Tooltips and Explanations:
- Granular Time Series Analysis:
- Geographical Insights with Map Interaction:
- Responsive Design for Accessibility:

IMPACT OF BI TOOLS ON DECISION-MAKING

- Generate Reports with Valuable Insights:
- Data Visualization to Understand the Reports:
- Access to Real-Time Intelligence:
- Factual Decision-Making:
- Streamline Operations and Resource Optimization:
- Optimize Limited Resources:

THANK YOU

K-BI Project Team