Covid-19 Data Analysis

Rashmi Jakkani

11445421

rj0233

INFO 5709

Data Visualization and Communication

Introduction:

I am going to analyze the Covid-19 data which includes the death count, confirmed cases, etc., and the dataset is downloaded from Kaggle. Covid-19 is caused by a coronavirus called SARS-CoV-2. It was first identified in Wuhan, China. This affects most old age people, especially those who have heart or lung or diabetes diseases will have a higher risk. Most people infected with this virus have experienced mild to moderate illness and also recover without requiring special treatment. The best way to prevent and slow down the spreading of this virus is to be well-informed by wearing face masks and sanitizing hands. Getting vaccinated is also so important due to which the covid cases have got reduced. This virus can spread from an infected person's mouth or nose in small liquid particles when they cough, sneeze, speak, etc.

As we all know there are some symptoms to know whether this virus is infected or not like fever, cough, shortness of breath, fatigue, body aches, headache, loss of smell or taste, sore throat, diarrhea, and vomiting. If one has an emergency, they need to immediately consult a doctor if they have trouble breathing, pressure in the chest or pain, pale, gray, or blue-colored skin, or new confusion. And now, everyone, everywhere we have access to covid 19 vaccines which have been declared by WHO.

Background:

In late December 2019, a new virus was identified in China causing respiratory diseases including pneumonia. It was originally named as Novel coronavirus and the World Health Organization (WHO) advised some language associated with the virus. And then named the virus causing the infection as severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2). The disease caused as a result of infection is named – coronavirus disease (COVID-19). It has been categorized as an airborne High consequence Infectious disease. Within a few days, SARS-CoV-2 is spreading

among people globally and can be seen on the WHO situation reports dashboard which is updated daily. As a viral infection, antibodies are not an effective treatment and there are vaccinations available.

About Dataset:

The covid-19 dataset is collected from the Kaggle website. It is a large dataset which 8 columns and 17 thousand rows collecting data on people infected with covid-19. This dataset consists of columns like country/region which has a list of many countries, state/province which has a list of many states, observation date which has the dates of covid-19 observed, confirm cases, count of deaths, recovered cases list, and the last date update of the information provided. This dataset contains two types of variables, our analysis is majorly dependent on the numeric type. The number of deaths from covid-19 from each region to each country can be found by doing an exploratory data analysis on this dataset so that I can know the trends of the number of deaths per area as it is important to know the trends of covid-19 in every corner of the world.

Data Cleaning:

It is a process of cleaning the dataset if the dataset contains any unwanted columns, null values, missing values, incorrect format, removing incorrect values, or any errors in structure.

Exploratory Data Analysis:

Exploratory data analysis is also called EDA. Using pandas, NumPy, statistical techniques, and data visualization tools to understand insights about the intrinsic characteristics of different entities of a dataset, such as columns and rows, is useful for data provided in different kinds of analyses. One of the most important processes when fine-tuning a set. You can identify trends, outliers, and theories based on your understanding of the data. Python was used for exploratory data analysis. Removed the 'inf' value in the column and replaced it with 0. I checked for null values and there were none.

Related work:

Description of previous papers on Covid-19 data analysis:

Paper-1:

According to the previous paper on Data Visualization and Analyzation of covid-19, Since December 2019, the deadly new coronavirus known as SARS CoV-2 has killed thousands of people worldwide. The WHO has advised that infection should be prevented by frequent hand washing, social isolation, keeping unwashed hands away from the face, and concealing coughs and sneezes with a tissue or inner elbow because there is currently no vaccine or antiviral medication. Since the number of confirmed cases and fatalities is rising daily and the virus hotspot has moved numerous times, it is very challenging to fully characterize COVID-19 using the data currently available. But this work still offers information data analytics and visualization to describe many disease-related characteristics using the datasets that are currently accessible. Based on the datasets used and the data analysis in this paper, we can see the combination with heat. Cough is one of the main indicators of wearing the virus. And the data visualizations are provided on the comparison of infections in females or males which shows that males are more prone to this disease and older people are more at risk. In this paper section 2, describes the different aspects of covid-19 using tabular data. Section 3 provides visualization of how the infection has spread across the world using pie charts and bar charts.

Paper-2:

This paper focused on the covid-19 based on freely available datasets which include the Kaggle repository. They have mentioned that data analytics is provided based on the number of aspects of covid-19 including the symptoms of this disease, covid-19 difference with the diseases caused by severe acute respiratory syndrome (SARS), Middle east respiratory syndrome (MERS), and swine flu. Data visualization is provided based on the infections of males/females which shows that males are more prone to covid-19. It is also discussed with older people. They have used different visualization models to represent the data. They used a bar graph to show the number of infected patients based on age groups, the number of infected patients based on gender, and a graphical representation of the total case. The pie chart is used to covid-19 confirmed cases around the world till April 10, 2020, the percentage of deaths around the world, percentage of recovered cases around the world. They have mentioned that they need to do investigation more on this topic to understand more about the disease.

Methods:

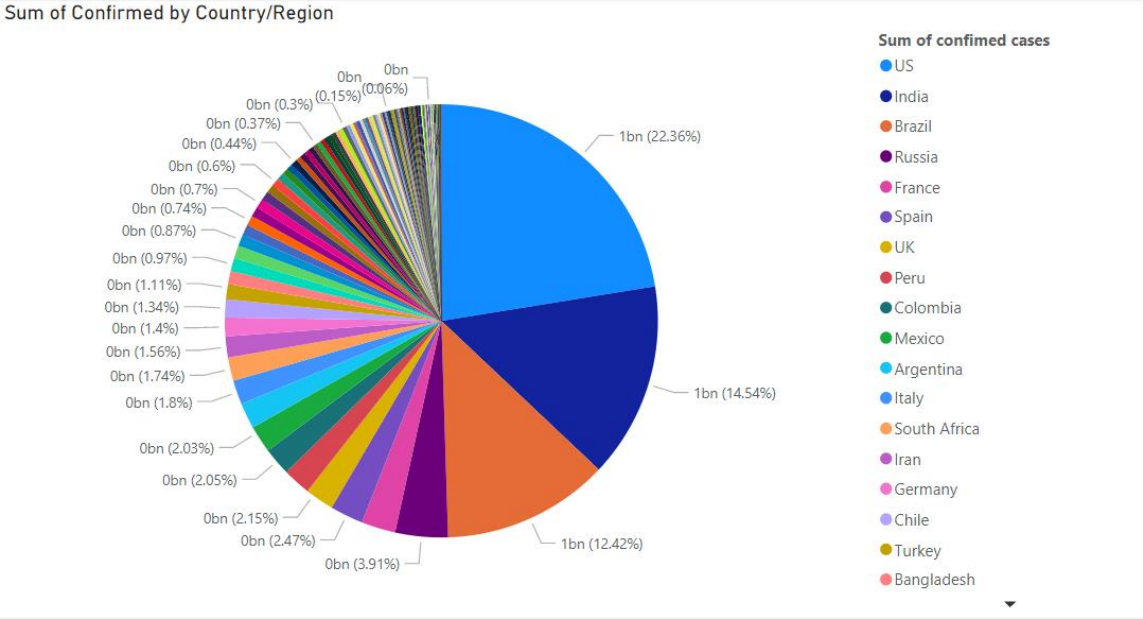
I have chosen the dataset from the Kaggle website that has the most valuable information.

Questions:

1) What is the sum of covid-19 cases by each country?
2) How was the overall trend of confirmed covid-19 cases throughout 2020?
3) What are the top 10 countries in recovery rates?
4) What are the top 5 countries with maximum deaths?
5) What are the states that were worst hit by covid-19 deaths in India?
6) Which country dealt with covid-19 in the most effective way?

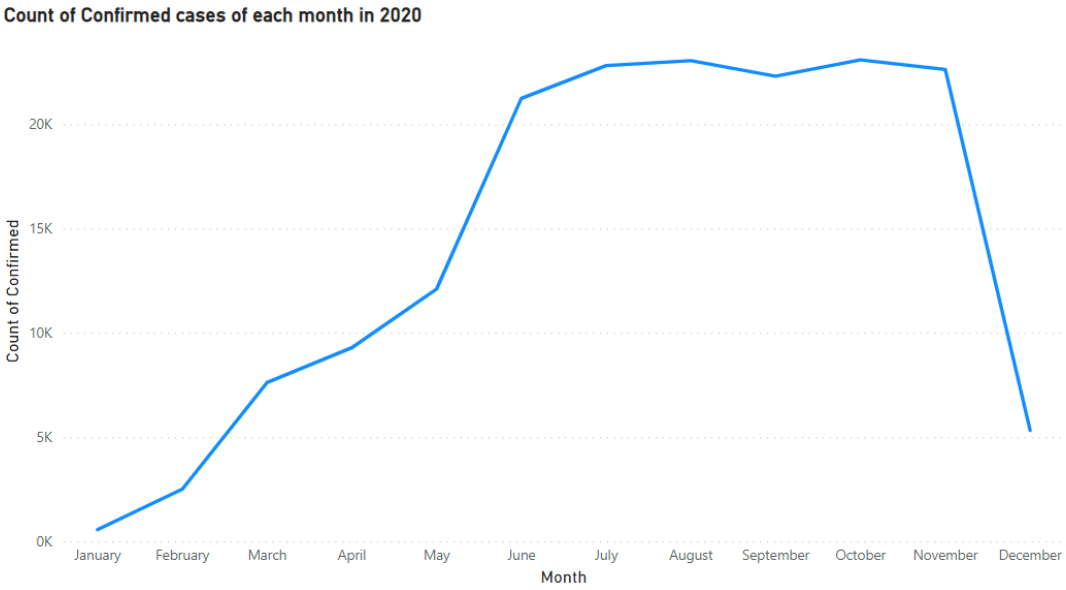
Solutions:

1. **What is the sum of covid-19 cases by each country?**



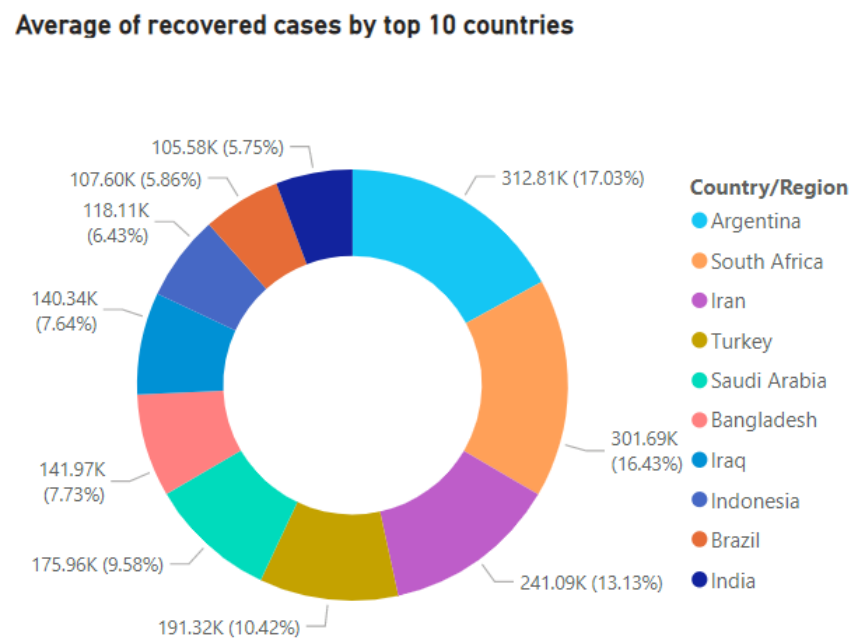
From the above visualization, I have found the sum of each country's confirmed covid-19 cases. The United States has the highest number of covid-19 cases compared to all the other countries. I have chosen the pie chart for this analysis because it represents each country's data clearly in a slice with a different color for each country and it can clearly say that it has shown us the highest to lowest cases list.

2. **How was the overall trend of confirmed covid-19 cases throughout 2020?**



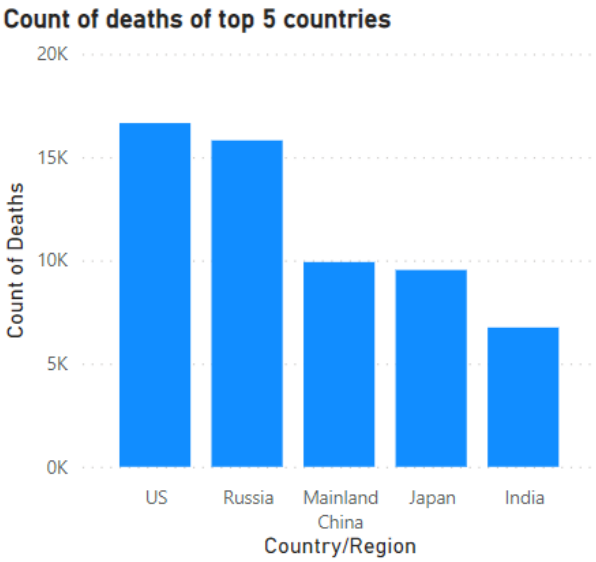
In this question, I have analyzed the overall trend of confirmed cases for each month in 2020 around the world. On the X-axis months are represented and on the Y-axis the count of confirmed cases is arranged from this visualization, we can observe that there is a gradual increase in covid cases using a line chart. Initially, in the month of January, there are very few positive cases but gradually it got increased and from the month of July to November there were huge confirmed cases and again in the month of December it decreased in 2020.

3. **What are the top 10 countries in recovery rates?**



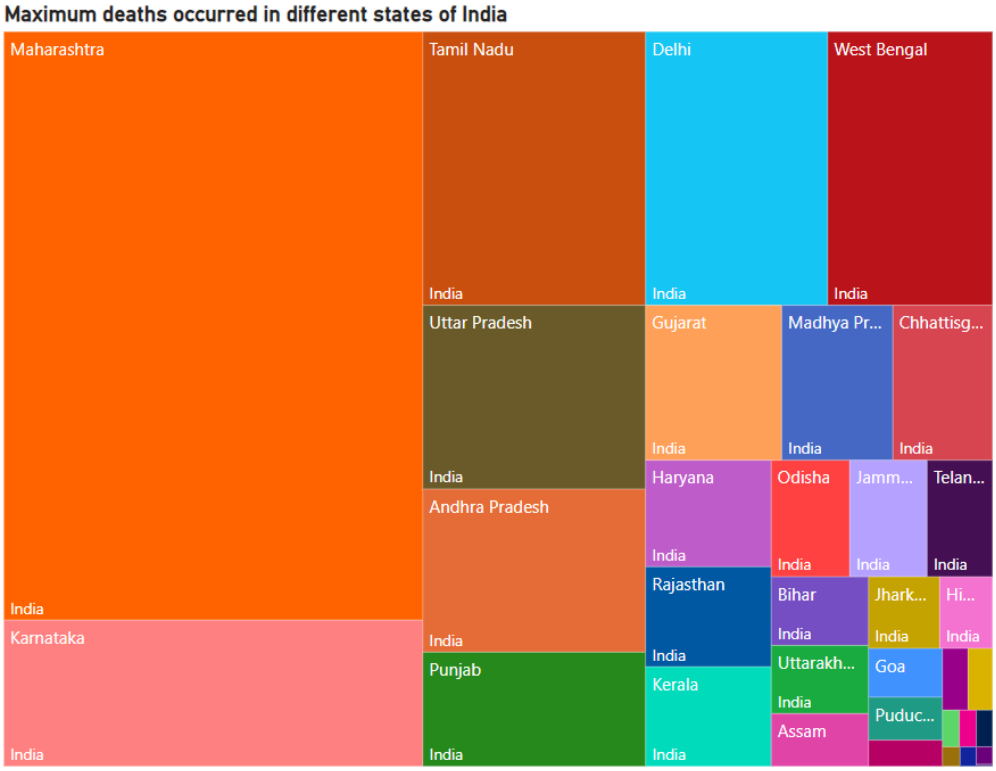
For this question, I have analyzed the data of the average of recovered cases by the top 10 countries using a Donut chart with a different color for each country. From the above visualization, we can observe that Argentina is the highest-recovered country with 17% and clearly mentioned the top 10 countries that recovered data.

4. **What are the top 5 countries with maximum deaths?**



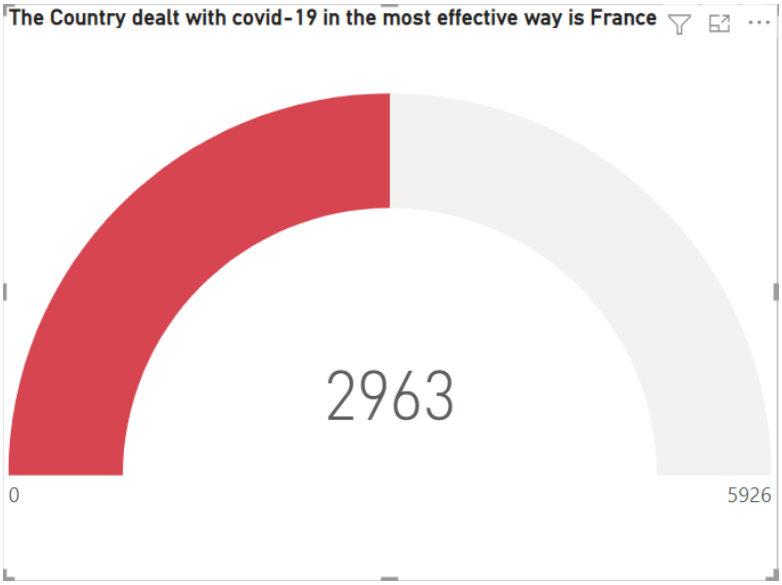
In this analysis, I have visualized the maximum number of deaths in the top 5 countries. I have done this using a bar graph with countries' names on the x-axis and the count of deaths on the y-axis. According to the above visualization, the United States has the maximum number of deaths using the maximum function on deaths count and followed by Russia, Mainland China, Japan and India.

5. **What are the states that were worst hit by covid-19 deaths in India?**



For this analysis, I have chosen a Treemap with different colors representing the different states in India. From the above visualization, we can observe that Maharashtra has the highest number of death cases compared to other states in India.

6. **Which country dealt with covid-19 in the most effective way?**



In this analysis, I have observed that the most effective country by covid-19 using the Gauge chart with red color indication. From the above visualization, we can observe that there were 2963 counts of 54804 covid deaths, and the earliest observation date is January 24th, 2020, Friday covid-19 cases and the most effective country was France.

Tools Used:

Microsoft Power BI :

Power BI is a data visualization software product which is developed by Microsoft. It is an interactive data product with a focus on business intelligence and is a collection of software services, apps, and connectors that will work together. In power bi, data can be imported and read by database, webpage, or structured files like spreadsheets, CSV, XML, and JSON. Generally, it provides a cloud-based service with a desktop-based interface. It offers data warehouse capabilities like data discovery, data interpretation, and some dashboards. And on the Azure cloud platform, Microsoft has released an additional service called power bi embedded.

Pros:

1. Cost, Power BI is a free tool that everyone can easily use it
2. Learning Curve, it is easy to use and it basically excels pivot tables and visualizations
3. Constant updates and Innovations, there will be frequent updates and suggest more improvements

4. Data Sources, where power bi connects to many data sources
5. Excel Integration, this is a very nice feature of power bi where it has the ability to save data to excel
6. Custom Visualizations has many different types of tools, and wedges to make data visualizations

Cons:

1. The User Interface, the user interface is very bulky
2. Rigid Formulas, power bi has some shortcomings
3. Data handling capacity, it has a limit on the size of the data that can ingest.
4. Visuals Configurability, the custom visuals are not comfortable in this.
5. Table Relationships, while joining tables together we get a little difficulty.

Discussion:

From all the visualizations and observations, the United States has the highest number of covid cases, Argentina has the highest recovery rate, the US has the highest count of deaths, and France is the most affected country. Initially, there is a graduate raise in covid cases but in between it has become constant with a high number of confirmed cases and gradually decreased by the end of the year throughout 2020. In particular, there are some states in India that are deadly affected by covid-19.

Future Work:

- By enhancing the data set we can develop the data far better by checking the increase in cases, and we can deploy and manufacture great information in the required countries.
- We can also check the most affected areas by observation dates using pivot tables

References:

https://www.kaggle.com/sudalairajkumar/novel-corona-virus-2019-dataset