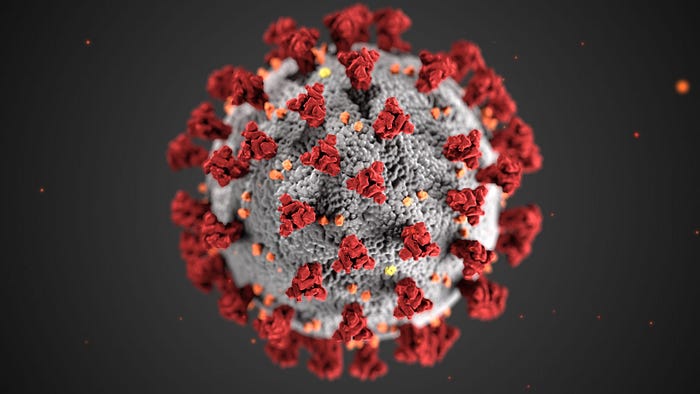
**COVID-19 Data Visualization Using Python**



1. **Introduction**

The outbreak of the Coronavirus pandemic at the wake of the year 2020 caught the entire world unawares and can be considered one of the greatest challenges the world has faced in the last century.

Coronavirus disease (COVID 19) is an infectious disease caused by a newly discovered coronavirus. Most of the countries in the world have been infected. The world grapples with this invisible, deadly enemy, trying to understand how to live with the threat posed by the coronavirus.

**“A picture is worth a thousand words”**

As the quote speaks, Data Visualization helps to derive information insights from data sources in a better way. This paper showcases visuals of how countries around the world are affected by this dreadful virus.

1. **Data Description**

In this paper of Data Visualization of COVID 19 data - [COVID 19 dataset](https://raw.githubusercontent.com/datasets/covid-19/master/data/countries-aggregated.csv) consisting of data related to country, population, the cumulative number of confirmed, recovered, death and tested cases from Kaggle is used.

1. **Objective**

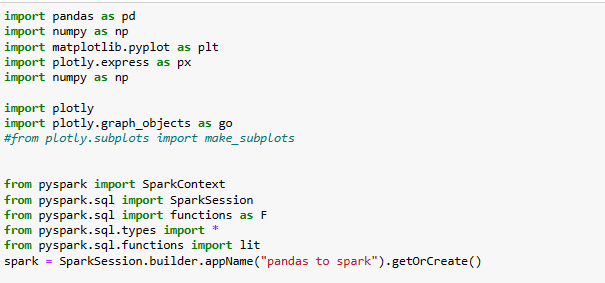
Using the variables; Total confirmed cases, Deaths cases, Recovered cases, Active cases, and Total test to:

* Analyze and display countries with higher number covid-19
* Analyze and display the impact of covid-19 on the population of the countries in the world
* Analyze and display the impact of covid-19 on the population of the world

1. **Data Analysis Model**
2. Data Collection
3. Data Cleaning and Preparation
4. Exploratory Data Analysis (EDA) and Visualization

**D.1** **Data Collection**

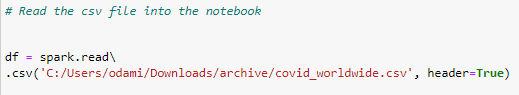
**1. Importing Required Packages**

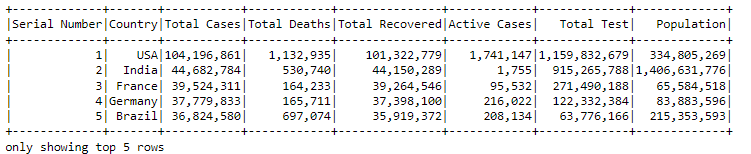


Tools used include: Python, PySpark, Pandas, Matplotlib and Plotly

**2. Data Gathering**

Data Source: [Covid Cases and Deaths WorldWide | Kaggle](https://www.kaggle.com/datasets/themrityunjaypathak/covid-cases-and-deaths-worldwide?resource=download)

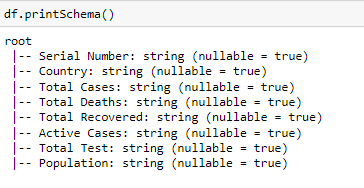




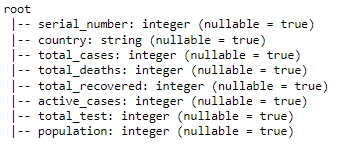
**A table view of the ingested dataset**

**D.2 Data Cleaning and Preparation**

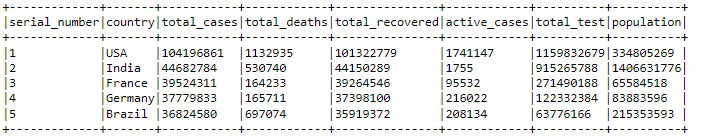
A check on the schema of the ingested dataset shows that some columns in the dataset are not in the correct format and hence needs to be cleaned and transformed.



**Schema of the ingested dataset before cleaning and preparation**

****

**Schema of the ingested dataset after cleaning and preparation**

****

**A table view of the ingested dataset after cleaning and preparation**

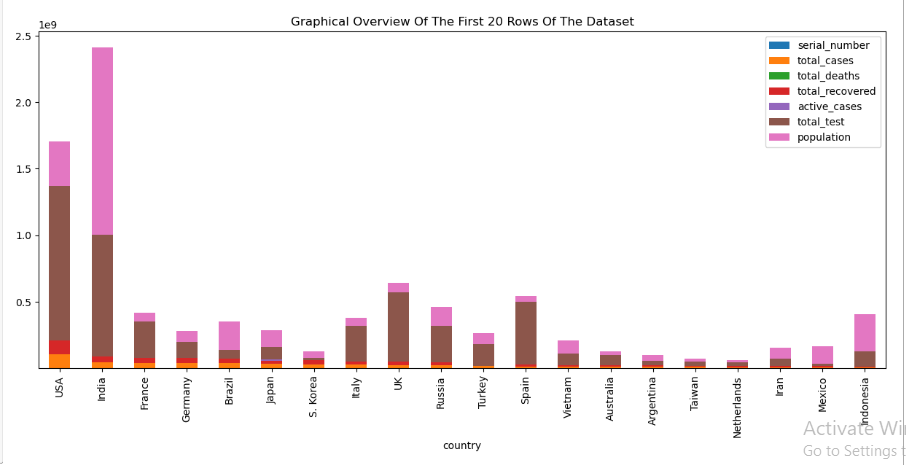
**D.3 Exploratory Data Analysis (EDA) and Visualization**

This will be limited to ranking of countries (based on COVID-19 variables available in the dataset) for the purpose of this paper.

From our previously extracted data, we are going to

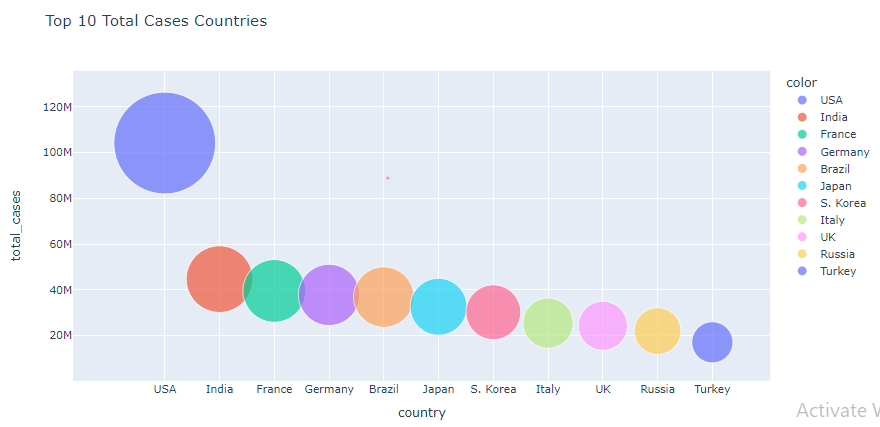
1. Rank countries
2. Rank the impact of covid-19 on population of countries
3. Analyze the overall Covid-19 cases in the world

based on confirmed total cases, deaths cases, recovered cases, active cases and total test, by doing some EDA and Visualization using python, matplotlib and plotly.

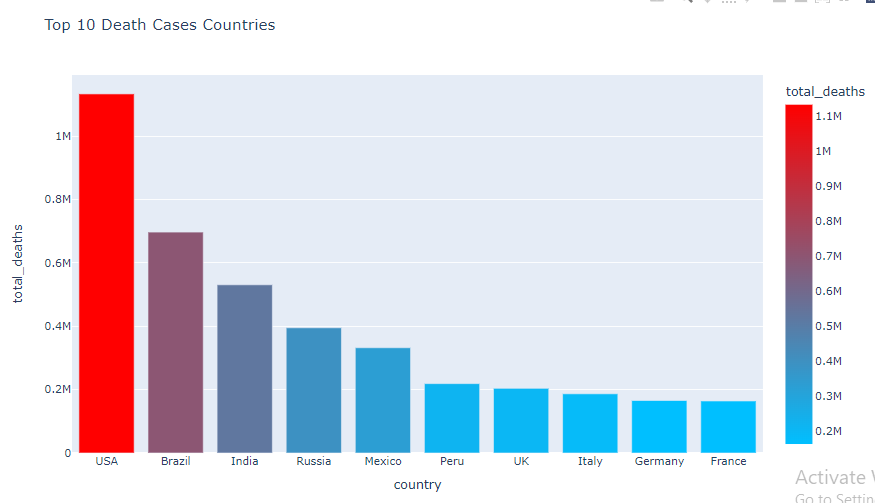


**A visual overview of the first 20 rows of the dataset**

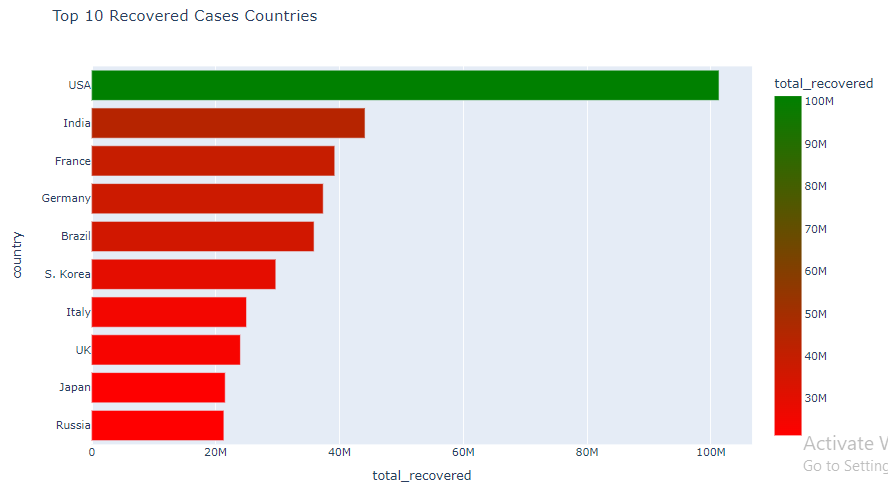
1. **Ranking Countries**
2. **Top 10 Confirmed Cases Countries**



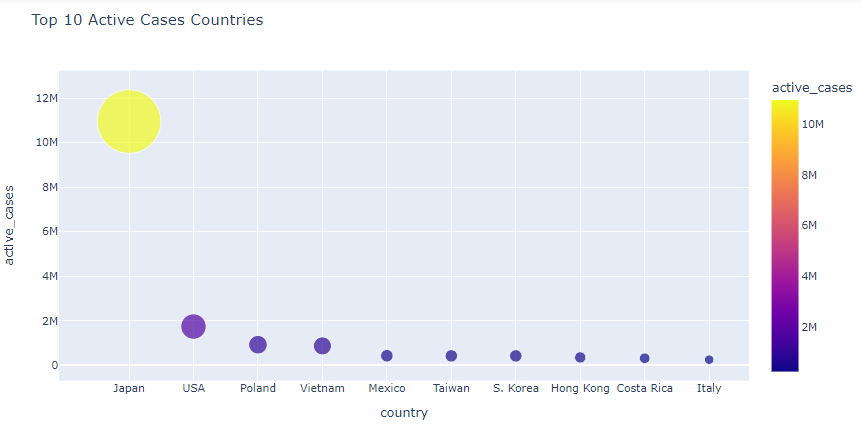
1. **Top 10 Deaths Cases Countries**



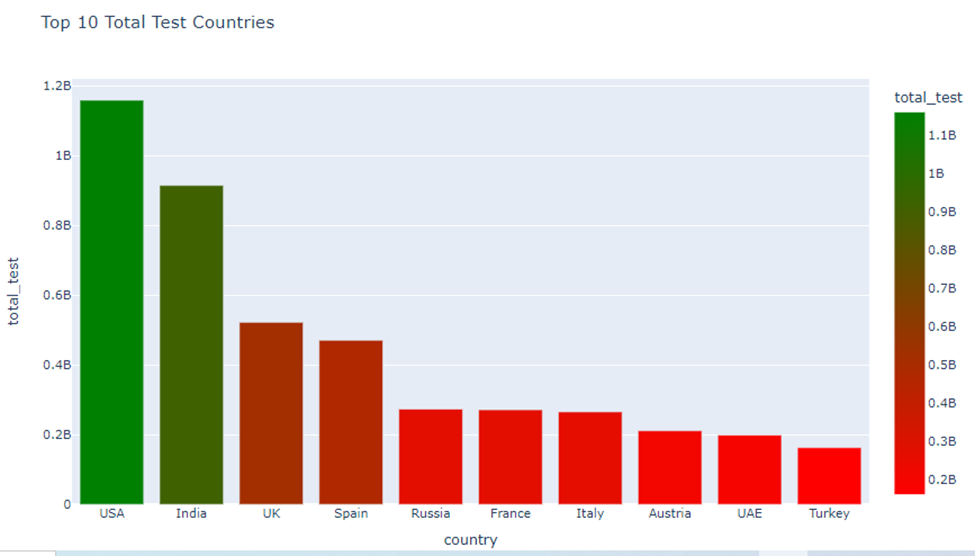
1. **Top 10 Recovered Cases Countries**



1. **Top 10 Active Cases Countries:**

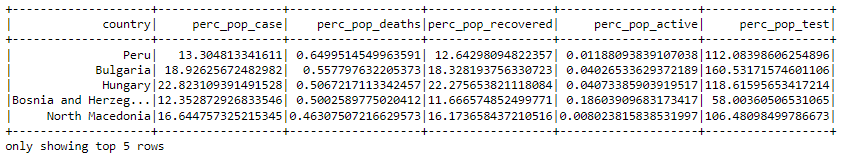


1. **Top 10 Total Test Countries:**

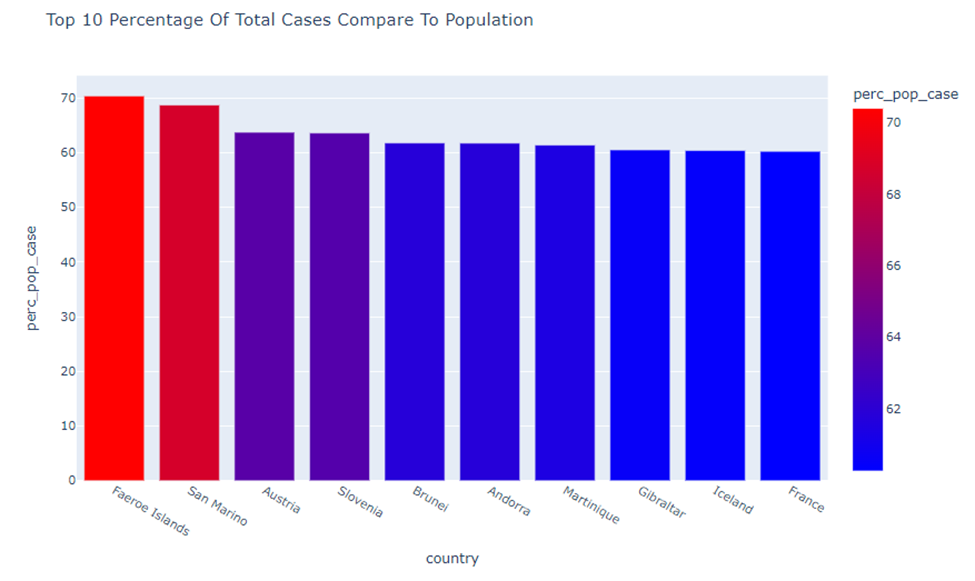


1. **Ranking the impact of covid-19 on population of countries**

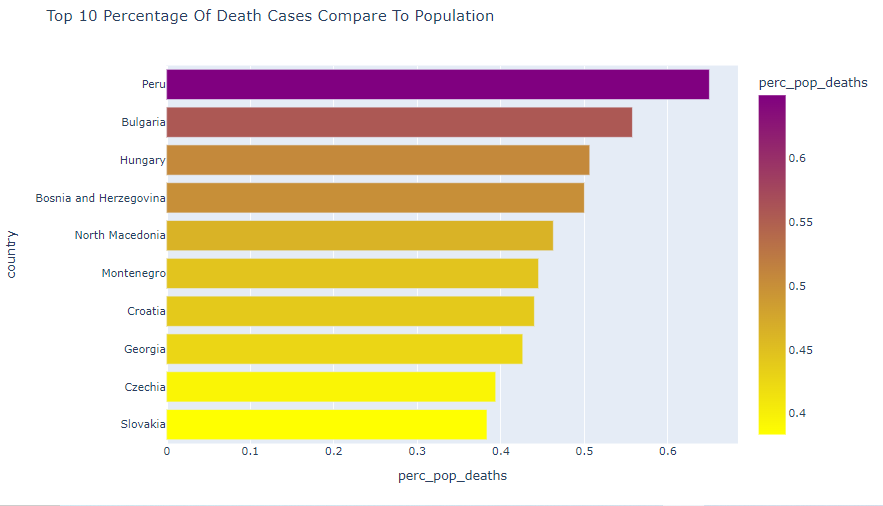
**A table view of the percentage population of covid-19 impact on countries**

****

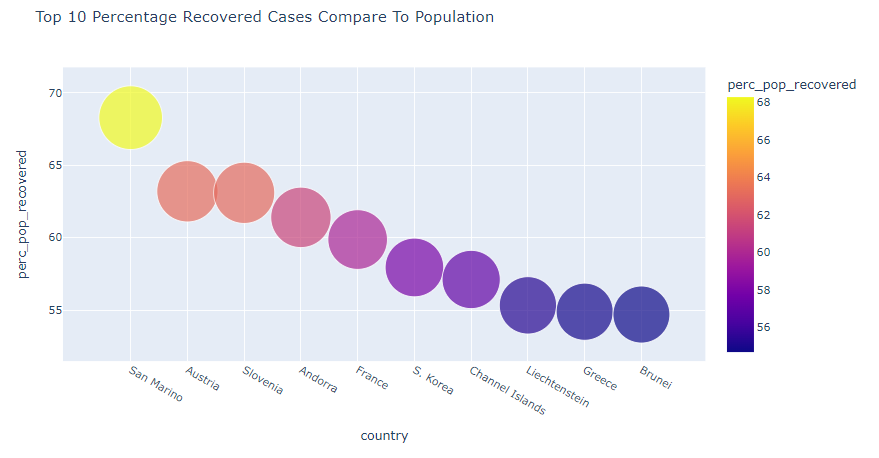
1. **Top 10 Percentage of Total Cases Compared To Population Countries**



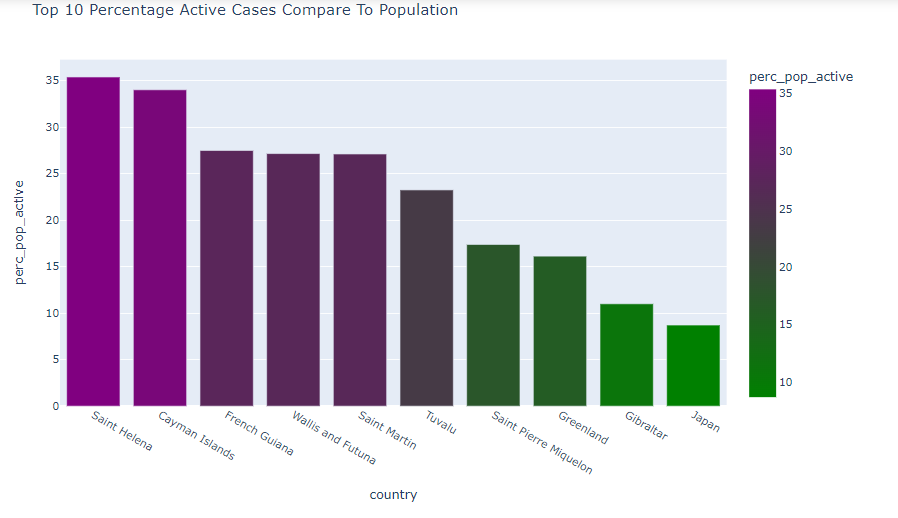
1. **Top 10 Percentage of Deaths Cases Compared To Population Countries**



1. **Top 10 Percentage of Recovered Cases Compared To Population Countries**

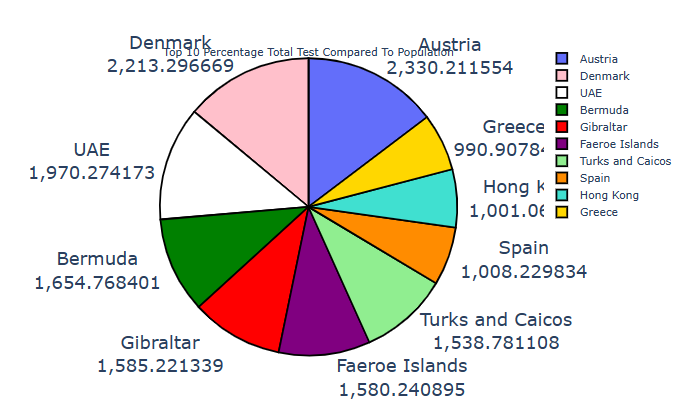


1. **Top 10 Percentage of Active Cases Compared To Population Countries**

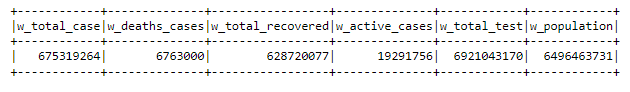


1. **Top 10 Percentage of Total Test Compared To Population Countries**

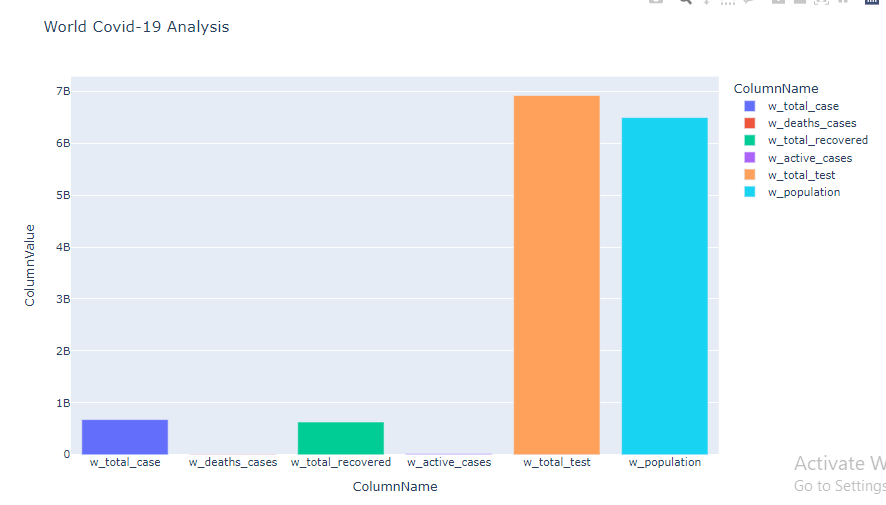
**Top 10 Percentage of Total Test Compared To Population Countries**



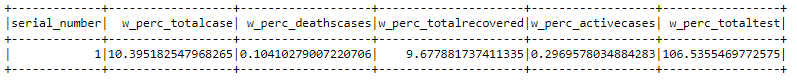
1. **Analysis of the overall Covid-19 cases in the world**
2. **World Covid-19 Analysis**

****

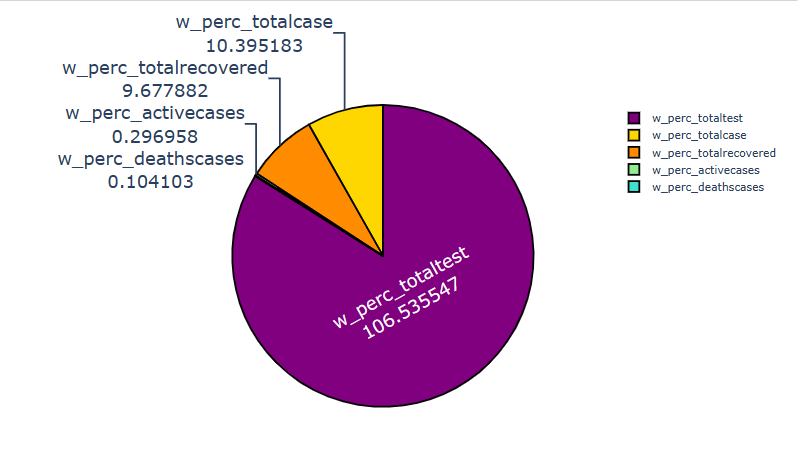
***A table view of the world covid-19 analysis***



1. **World Covid-19 Variable Compared To The World’s Population**

****

**World Covid-19 Analysis Variables Compared To Population**



**Conclusion**

This work analyzed and visualized the COVID-19 dataset published by **Kaggle** considering the following steps:

* Analysis of the countries with higher numbers of cases, deaths, recovery, still active and test of COVID-19 and use of bar and bubble chart to visualize the data.
* Analysis of the countries with higher percentage per population of cases, deaths, recovery, still active and test of COVID-19 and use of bar and bubble chart to visualize the data.
* Analysis of the COVID-19 scenario in the world: plot of the total number of cases, deaths, recovery, still active and test in the world.

This work gave me a great opportunity to practice and learn more about Pandas, Pyspark, Plotly and Matplotlib.

There are still more insights to be derived from the dataset like:

* Analysis of deaths and recovered cases compared to the confirmed cases for countries to show the response of patients to covid-19 treatments
* Analysis of deaths and recovered cases to the confirmed cases for the world to show the response of patients to covid-19 treatments

Lastly, the dataset can also be improved on to include information on

* Vaccination
* Monthly or yearly breakdown of records to be able to see the trend of the virus