Project Report

**COVID-19 DATA INSIGHTS**

Ayush Maurya

E22CSEU1586

Ankita Gupta

E22CSEU1608

A report submitted in part fulfilment of the degree of

**BTech in Computer Science**

**Submitted To:** Dr. Satyam Omar



Department of Computer Science

Bennett University, Greater Noida

**Motivation**

The Covid-19 pandemic has been a really big and serious problem all around the world. It's not just about people getting sick; it's affected how we live, our communities, and even our economies. We're motivated to do this project because we realize it's super important to truly understand all the different ways the pandemic is impacting us. By looking closely at all the information, we have about Covid-19, we hope to uncover the various trends and patterns that show how the virus is spreading and affecting us.

Because the pandemic is so widespread, we know it's crucial to carefully look at all the data related to Covid-19. This includes lots of different information like how many people are getting sick, how it's affecting different areas, and more. By closely examining all these details, we want to find insights that go beyond just numbers. We're aiming for a complete picture of what's happening with the pandemic, and through our analysis, we want to help add to what we collectively known about the virus. Our hope is that by doing this, we can provide useful information to help make smart decisions and come up with effective strategies to deal with the challenges posed by the pandemic.

**Objective**

The primary objective of this project is to harness the power of probability and statistics as analytical tools to extract meaningful and actionable insights from the vast pool of Covid-19 data. The utilization of these statistical methodologies is motivated by the recognition that they offer a rigorous framework for understanding the inherent uncertainties and complexities associated with the pandemic.

1. **Understanding the Distribution of Cases:** Probability distributions will be employed to characterize the spread of Covid-19 cases over time and across different regions. This analysis aims to provide a comprehensive picture of how the virus has manifested, identifying peaks, troughs, and potential areas of concern.
2. **Exploring Correlations Between Different Variables:** Statistical methods will be applied to explore correlations between various variables, such as the relationship between testing rates and the number of confirmed cases or the impact of vaccination rates on the severity of outcomes. Uncovering these relationships can offer insights into the interconnected dynamics of the pandemic.
3. **Providing Valuable Information for Decision-Makers and the Public:** The ultimate goal of this project is to distil the analytical findings into valuable information that can inform decision-makers and empower the public. By presenting the insights in a comprehensible manner, we aim to contribute to evidence-based decision-making, public awareness, and collaborative efforts to mitigate the impact of the pandemic.

In summary, the motivation for this project is grounded in the profound impact of the Covid-19 pandemic, while the objective is to employ probability and statistics as powerful tools for gaining insights that can guide decision-making and public understanding in the face of this global crisis.

**Methodology**

1. **Data Collection and Pre-processing:** In our project, we gathered information about Covid-19 from reliable sources, such as government health agencies, the World Health Organization (WHO), and other trusted databases. The dataset we used contains a variety of information that gives us a comprehensive understanding of the Covid-19 situation.

In the process of preparing the data for analysis, we performed data cleaning to ensure its quality and reliability. This involved addressing missing values, correcting errors, and organizing the information in a way that makes it suitable for statistical analysis.

* **Steps:**

**Library Installation:** Installed required libraries using !pip install pandas and !pip install esda.

**Library Import:** Imported necessary libraries such as numpy, pandas, seaborn, matplotlib, plotly, and esda.

**Data Loading:** Loaded the COVID-19 dataset using pd.read\_csv('/content/worldwide covid data.csv').

**Handling Missing Values:** Checked for missing values using df.isnull().sum() to identify columns with null values. Imputed missing values using median for skewed data and mean for normally distributed data.

**Handling Duplicates:** Checked for and removed duplicates using df.duplicated() and df.drop\_duplicates().

1. **Data Visualization:** Data visualization serves as a crucial bridge between raw data and meaningful insights in the context of the "Covid-19 Data Insights" project. With the sheer complexity of Covid-19 data encompassing variables like cases, deaths, recoveries, testing rates, and vaccination rates, the role of visualization becomes paramount.

* **Achieved By:**

**Treemaps:** Calculated mortality and recovery rates for each country. Visualized rates using horizontal bar plots and treemaps with px.bar().

**Bar Plots:** Created bar plots for the top 20 countries with the most confirmed cases, deaths, active cases, recovered cases, and tests conducted using px.bar(). Visualized testing rates using a bar plot with px.bar().

**Choropleth Maps:** Created choropleth maps for the distribution of total cases, deaths, recovered cases, and active cases worldwide using px.choropleth().

1. **Descriptive Statistics:** In the realm of statistical analysis for the "Covid-19 Data Insights" project, descriptive statistics played a pivotal role in distilling the essential characteristics of the dataset. Measures of central tendency, including the mean, median, and mode, were computed to provide a sense of the typical or central values within the Covid-19 data. These central tendency metrics offer a consolidated understanding of the data's general trend, such as the average number of cases or deaths, providing a reference point for further analysis.

**Computed basic descriptive statistics using df.describe() to get insights into the central tendency and spread of numerical columns.**

1. **Probability Distributions:** In the "Covid-19 Data Insights" project, analysis through probability emerged as a key component in unravelling the inherent uncertainty and randomness associated with the spread and impact of the virus. Probability distributions were employed to model and understand the likelihood of different events occurring within the context of the Covid-19 dataset.

Probability analysis served as a foundational pillar in our exploration of the Covid-19 dataset, shedding light on the uncertainty inherent in the pandemic's trajectory. By employing probability tools, we not only quantified the likelihood of various events but also contributed to a more informed and probabilistically grounded understanding of the dynamics shaping the impact of Covid-19.

* **Steps:**

**Kernel Density Estimation** (KDE): Plotted probability distribution using (KDE) with sns.kdeplot().

1. **Hypothesis Testing:** Hypothesis testing is a fundamental statistical method used in the "Covid-19 Data Insights" project to draw meaningful inferences about the population based on the sample data. The objective is to evaluate whether observed patterns or differences in the data are statistically significant or if they could have occurred by random chance.

The process typically involves formulating a null hypothesis (H0) and an alternative hypothesis (H1). The null hypothesis represents a statement of no effect or no difference, while the alternative hypothesis posits the presence of a significant effect or difference. For instance, in the context of Covid-19 data, a null hypothesis might state that there is no difference in the average number of cases between two regions, while the alternative hypothesis suggests a significant difference.

* **Steps:**

**High vs. Low Population:** Conducted a **t-test** to explore the significant difference in mean total cases between countries with high and low populations using ttest\_ind().

**Total Cases vs. Total Tests:** Conducted a hypothesis test to explore the correlation between the total number of COVID-19 cases and the total number of tests conducted per million populations using pearsonr().

1. **Time-Series Analysis and Forecasting:** In the "Covid-19 Data Insights" project, time series analysis played a pivotal role in uncovering dynamic patterns, trends, and potential seasonality within the temporal dimension of the Covid-19 dataset. Time series analysis involved the exploration of overall trends in Covid-19 metrics such as cases, deaths, and recoveries.

Exploring trends and patterns in COVID-19 cases over time. Analysing daily increases in confirmed, recovered, and death cases and then Visualising weekly trends and progress of different case types using line plots.

**ARIMA Modelling:**

* **Goal:** Implement the ARIMA model for time-series forecasting.
* **Steps:**

Converting 'ObservationDate' to datetime and group data for training and testing.

Training the ARIMA model with a specified order.

Forecast future values and evaluate model performance using RMSE.

Visualize actual data, training data, and ARIMA predictions.

**Time-Series Modelling:**

* **Goal:** Predict future COVID-19 cases using Linear Regression and SVR models.
* **Steps:**

Split the dataset into training and validation sets.

Train Linear Regression and SVR models on the training data.

Evaluate model performance on the validation set.

Predict future cases for the next 30 days using both models.

1. **Corelation and Regression:** Correlation analysis was employed to quantify the strength and direction of relationships between pairs of variables. The Pearson correlation coefficient, for example, measures the linear relationship between two continuous variables. Correlation analysis in the project explored connections such as the relationship between testing rates and the number of confirmed cases, vaccination rates and the severity of outcomes, or any other associations present in the dataset.

Multiple regression extended this analysis to account for the influence of multiple independent variables simultaneously. In the context of the project, regression analysis could be applied to understand the impact of testing rates on the number of cases, the effect of vaccination rates on recovery rates, or any other relevant associations.

* **Steps:**

**Correlation Matrix:** Computed and visualized the correlation matrix between different variables using sns.heatmap().

**Loop for Analysis:** Utilized a loop to conduct linear regression analysis for selected predictors against the target variable 'Deaths/ 1M pop' using LinearRegression().

**Visualization:** Visualized the regression lines with plt.scatter() and plt.plot().

1. **Finding Outliers:**

* **Steps:**

**Z-Score Calculation:** Calculated Z-scores for selected columns using stats.zscore(df[columns\_of\_interest]).

**Threshold Definition:** Defined a threshold (e.g., 3 or -3) for considering Z-scores as outliers.

**Outlier Identification:** Created a DataFrame to store the outliers using boolean conditions on Z-scores.

**Outliers Count:** Counted the number of outliers in each column using outliers\_df.sum().

**Outliers Investigation:** Displayed rows with outliers for further investigation using df[outliers\_df.any(axis=1)].

**Outliers Visualization:**Visualized outliers using boxplots for selected columns with sns.boxplot().

1. **Lockdown Analysis:**

* **Steps:**

**Scatter Plot:** Created a scatter plot to categorize countries based on their need for lockdowns considering the ratio of active cases to population using px.scatter().

**10. Testing Rate Analysis:**

* **Steps:**

**Rate Categorization:** Categorized countries based on testing rates compared to the mean.

**Visualization:** Visualized testing rates using a bar plot with px.bar().

1. **COVID-19 Analysis in India:**

* **Steps:**

**Data Loading for India:** Loaded COVID-19 data specific to India from '/content/Latest Covid-19 India Status.csv'.

**Visualization by State/UTs:** Visualized COVID-19 statistics by state/UTs using bar plots and line plots with make\_subplots().

**Correlation Analysis:** Explored correlations between different COVID-19 variables using numerical\_columns.corr() and sns.heatmap().

1. **Vaccination Data Analysis in India:**

* **Steps:**

**Data Loading for Vaccination:** Loaded vaccination data for states/UTs in India from '/content/COVID-19 India Statewise Vaccine Data.csv'.

**Visualization:** Visualized total vaccination doses by state/UT using a bar plot with sns.barplot().

**Tools to Achieve the Goal**

1. **Programming Language: Python**

Python, a versatile and widely-used programming language, served as the primary tool for conducting the "Covid-19 Data Insights" project. Its simplicity, extensive libraries, and robust data analysis capabilities make it an excellent choice for data science and statistical analysis.

1. **Libraries:**

* **Pandas**: Pandas is a powerful library for data manipulation and analysis. It provided essential data structures, such as Data Frames and Series, allowing for efficient handling of the Covid-19 dataset. Operations like cleaning, filtering, and transforming data were streamlined using Pandas, contributing to the overall data preparation process.
* **NumPy**: NumPy is a fundamental library for numerical computing in Python. It played a crucial role in performing mathematical and statistical operations on the Covid-19 data. NumPy arrays facilitated efficient data handling and computation, enhancing the accuracy and speed of the analyses conducted in the project.
* **Matplotlib**: Matplotlib is a versatile plotting library in Python, enabling the creation of static, interactive, and publication-quality visualizations. It was employed to generate a variety of charts, graphs, and plots, such as histograms, box plots, and time series plots. Matplotlib's flexibility allowed for the effective communication of complex patterns within the Covid-19 dataset.
* **Seaborn**: Seaborn is a statistical data visualization library based on Matplotlib. It provided a high-level interface for creating aesthetically pleasing and informative statistical graphics. Seaborn was particularly useful for creating stylish and informative visualizations, enhancing the overall presentation of the project's findings.
* **Plotly**: Plotly is a library for interactive and dynamic visualizations. It was utilized to create interactive charts and plots that allow users to explore the Covid-19 data dynamically. Plotly's interactive features, such as zooming and hovering, provided an engaging way to present the results, enhancing user interaction and understanding.

1. **Google** **Colab:** Google Colab, a cloud-based platform provided by Google, served as the coding and documentation environment for the project. It offers a Jupyter Notebook environment with the added benefit of free access to GPU resources, making it ideal for data analysis and machine learning tasks. Google Colab facilitated collaborative coding and seamless integration with other Google services, enhancing the efficiency of the project's development and documentation processes.

**Model**

**Time Series Forecasting:**

1. Objective:

Predict the future number of confirmed cases, deaths, or recoveries for a specific region based on historical data.

1. Approach:

Utilize time series forecasting methods, such as ARIMA (AutoRegressive Integrated Moving Average), to make predictions based on the historical trend of the data.

**Steps:**

1. Data Preprocessing: Ensure your dataset is in a time series format, with a timestamp and corresponding numerical values for confirmed cases, deaths, or recoveries.
2. Choosing a Region: Identify the specific region or location for which you want to make predictions. Filter your dataset accordingly.
3. Time Series Analysis: Visualize the time series data to understand trends, seasonality, and any other patterns. This helps in choosing appropriate model parameters.
4. Stationarity: Check for stationarity in the time series. Stationarity is a key assumption for ARIMA. If the data is not stationary, perform differencing to make it stationary.
5. ARIMA Model Selection: Determine the order of the ARIMA model (p, d, q) based on:

Autocorrelation Function (ACF) and Partial Autocorrelation Function (PACF) plots.

Information criteria such as AIC (Akaike Information Criterion) or BIC (Bayesian Information Criterion).

1. Train-Test Split: Split your dataset into a training set and a testing set. The training set is used to train the ARIMA model, and the testing set is used to evaluate its performance.
2. ARIMA Model Training: Train the ARIMA model using the training set.
3. Model Evaluation: Evaluate the model's performance on the testing set. Common metrics include Mean Absolute Error (MAE), Mean Squared Error (MSE), and Root Mean Squared Error (RMSE).
4. Future Forecasting: Use the trained ARIMA model to forecast future values for the specified region.
5. Visualization: Visualize the predicted values alongside the actual values to assess the accuracy of the forecasting.

**Feature Importance Analysis:**

Feature Importance Analysis is a statistical technique that assigns a score to input features based on how useful they are at predicting a target variable. In the context of a pandemic, such as COVID-19, feature importance analysis can be used to determine which factors contribute most to the number of confirmed cases, deaths, or recoveries.

Data Collection: Gathering data related to the pandemic. This includes the number of confirmed cases, deaths, and recoveries.

Data Preprocessing: Cleaning the data by handling missing values and outliers.

Model Selection: Decision based tree model Random Forests is used for this purpose.

Model Training: Training the model using the pre-processed data.

Feature Importance Analysis: After the model has been trained, use it to score the importance of each feature. The importance of a feature is typically measured by the increase in the model’s prediction error after permuting the feature. A feature is considered “important” if shuffling its values increases the model error, because in this case the model relied on the feature for the prediction. A feature is considered “unimportant” if shuffling its values leaves the model error unchanged.

Interpretation: Interpret the results. The features with the highest scores are the ones that the model considers most important in predicting the target variable.

**Regression Model:**

**Linear Regression-** Linear regression is used in COVID-19 data analysis to explore relationships between variables and make predictions. Linear regression assumes a linear relationship between variables, and its effectiveness depends on the nature of the data. If relationships are nonlinear, you may need more complex models. Regression analysis is applied to understand the impact of testing rates on the number of cases, the effect of vaccination rates on recovery rates, or any other relevant associations.

**Support Vector Regression-** Support Vector Regression (SVR) is a type of Support Vector Machine (SVM) that uses the same principles but for regression problems instead of classification. It’s a powerful machine learning model capable of performing linear and non-linear regression.

The main idea behind SVR is to find a function that approximates the data in such a way that the error does not exceed a certain threshold. This is achieved by defining a decision boundary (or a margin) that includes as many data points as possible, and penalizing those that fall outside this margin.

**Code**

Follow the following link to access the code:

Colab 1 Link:

<https://colab.research.google.com/drive/1Jgm-pXeGM47LRzBYAcXLtiHP-CXnzsVw?usp=sharing>

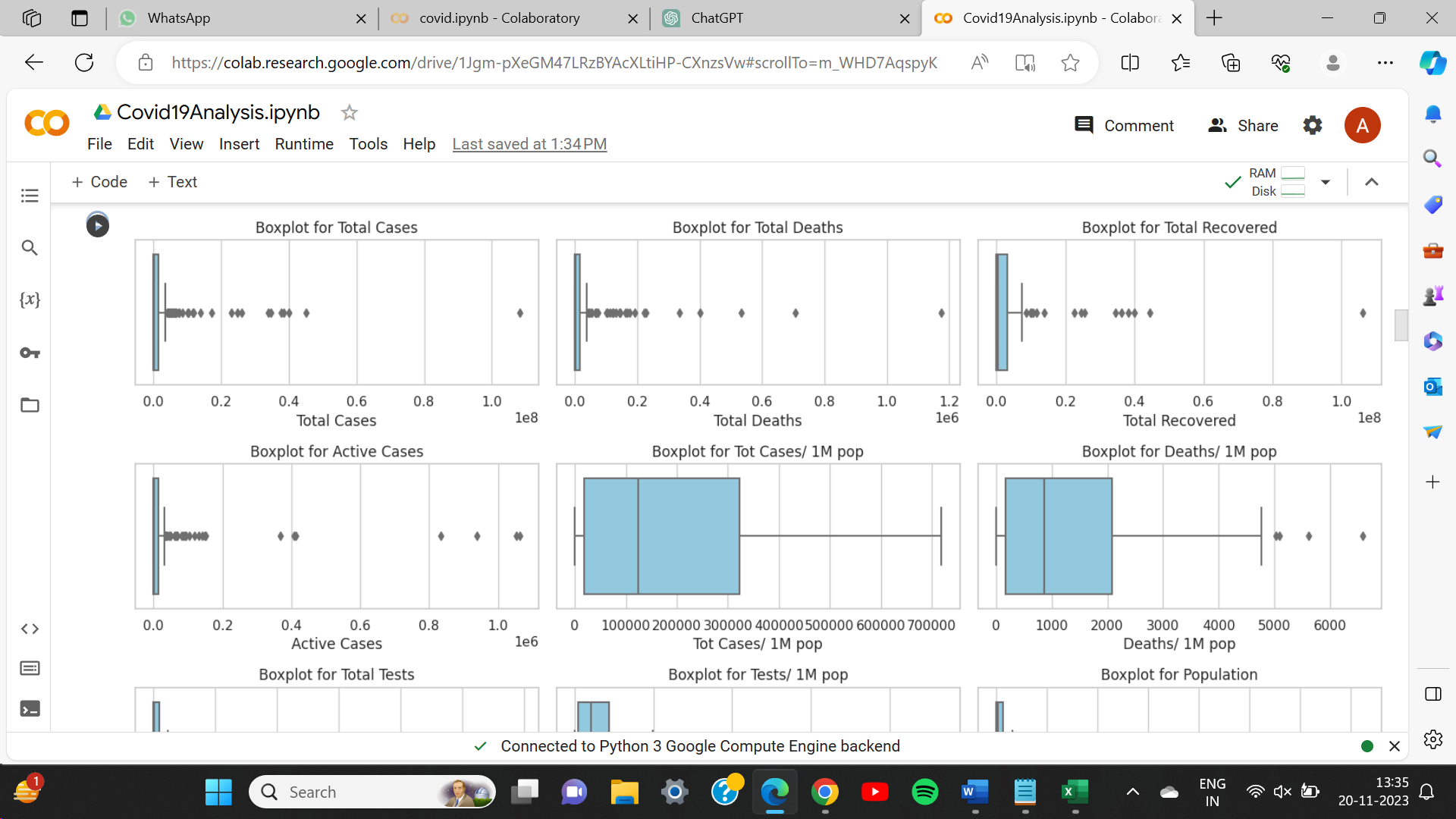
Colab 2 Link :

https://colab.research.google.com/drive/1-hTUSHkg13ZQeC641KDz8Ibd3z6UHPJb?usp=sharing

**Analysis and Graphical Representation**

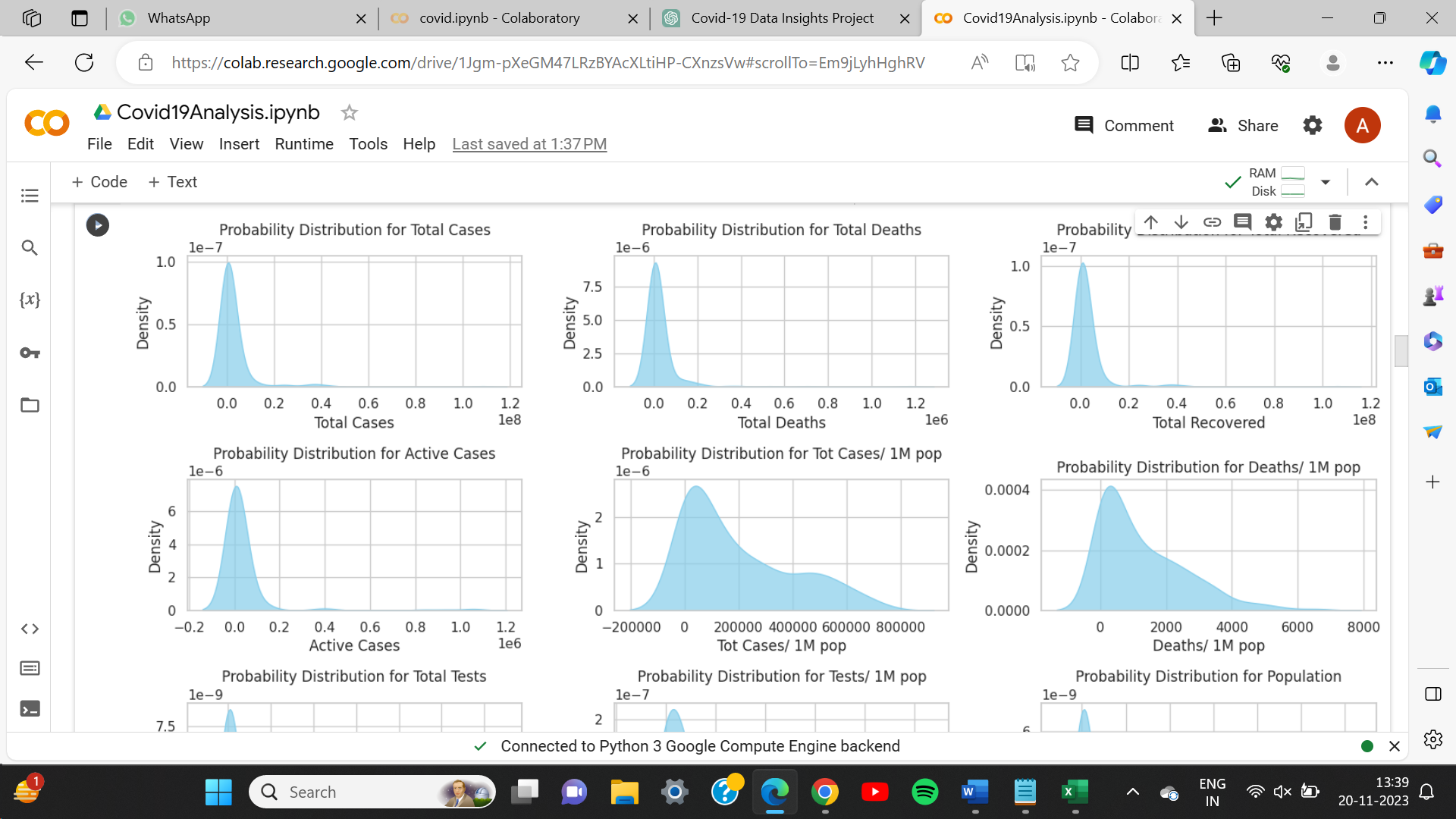
**Finding Outliers**

Outliers are data points that significantly differ from the majority of the observations in a dataset. They can distort statistical analyses and machine learning models, leading to inaccurate results or biased predictions. Identifying and handling outliers is an essential step in the data preprocessing phase.

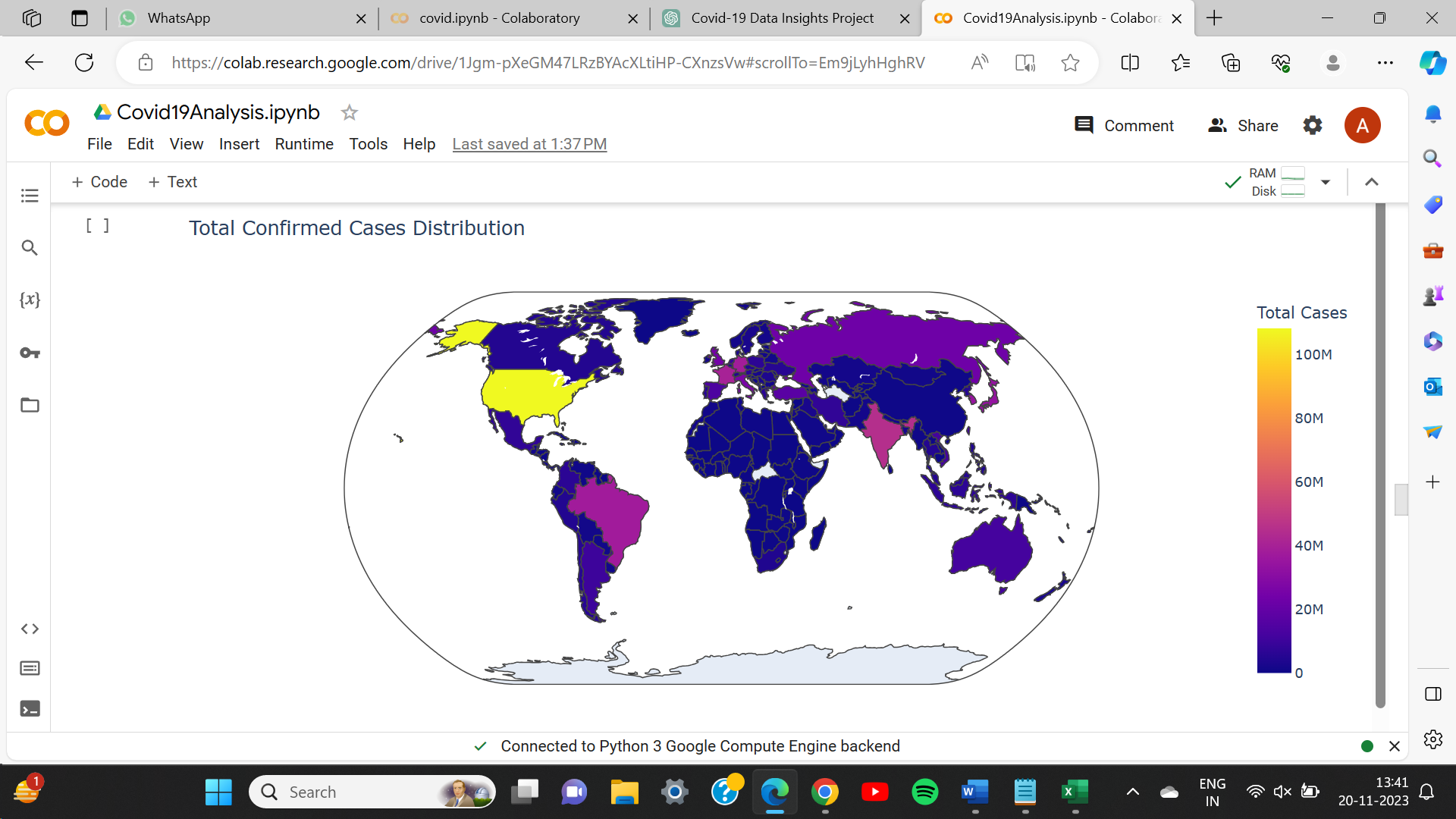


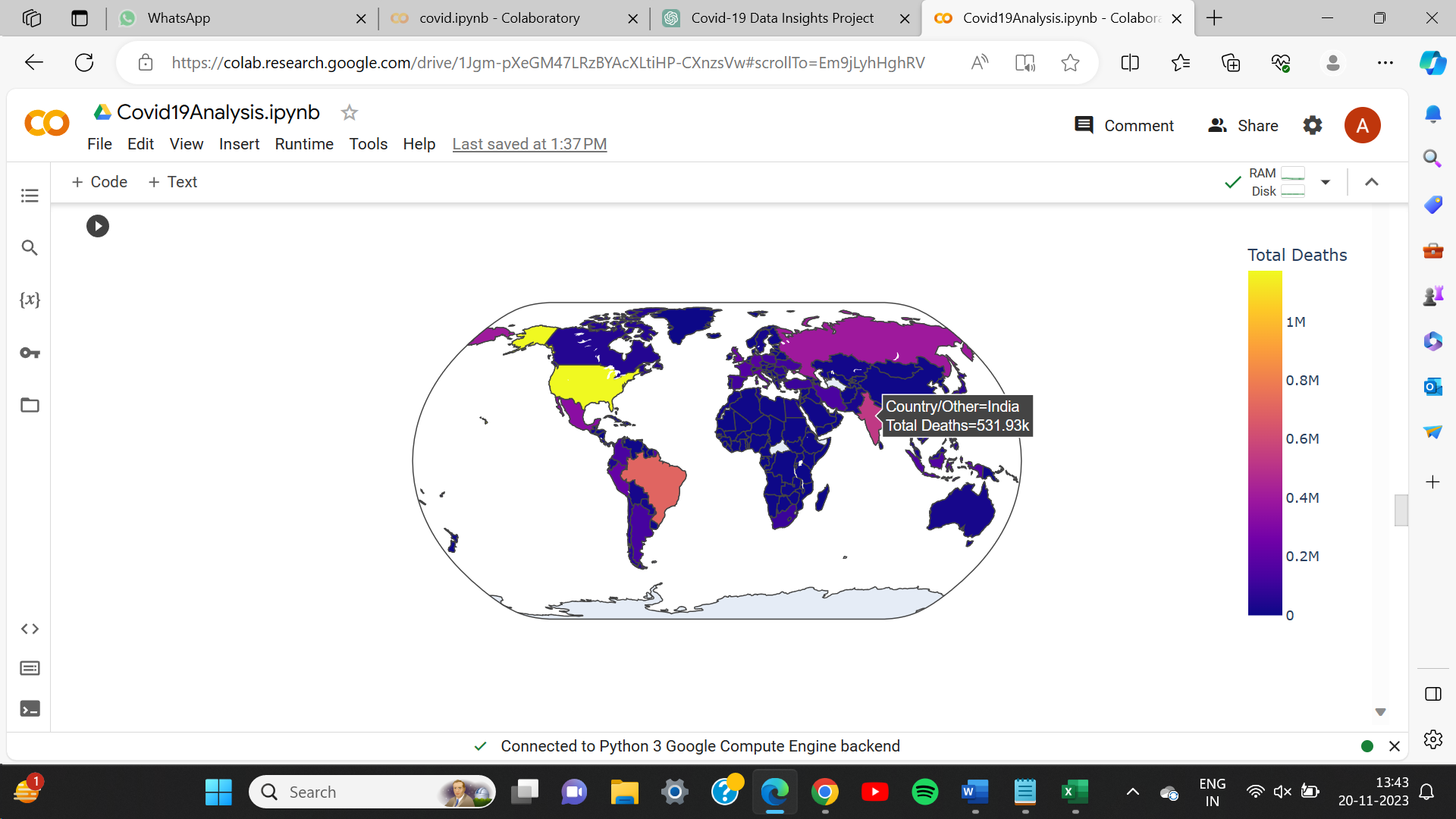
**Probability Distributions**

Probability distributions describe how the values of a random variable are spread or distributed across different outcomes. In statistics and probability theory, understanding and characterizing these distributions are fundamental for analysing data and making predictions.



**Map Visualizations**





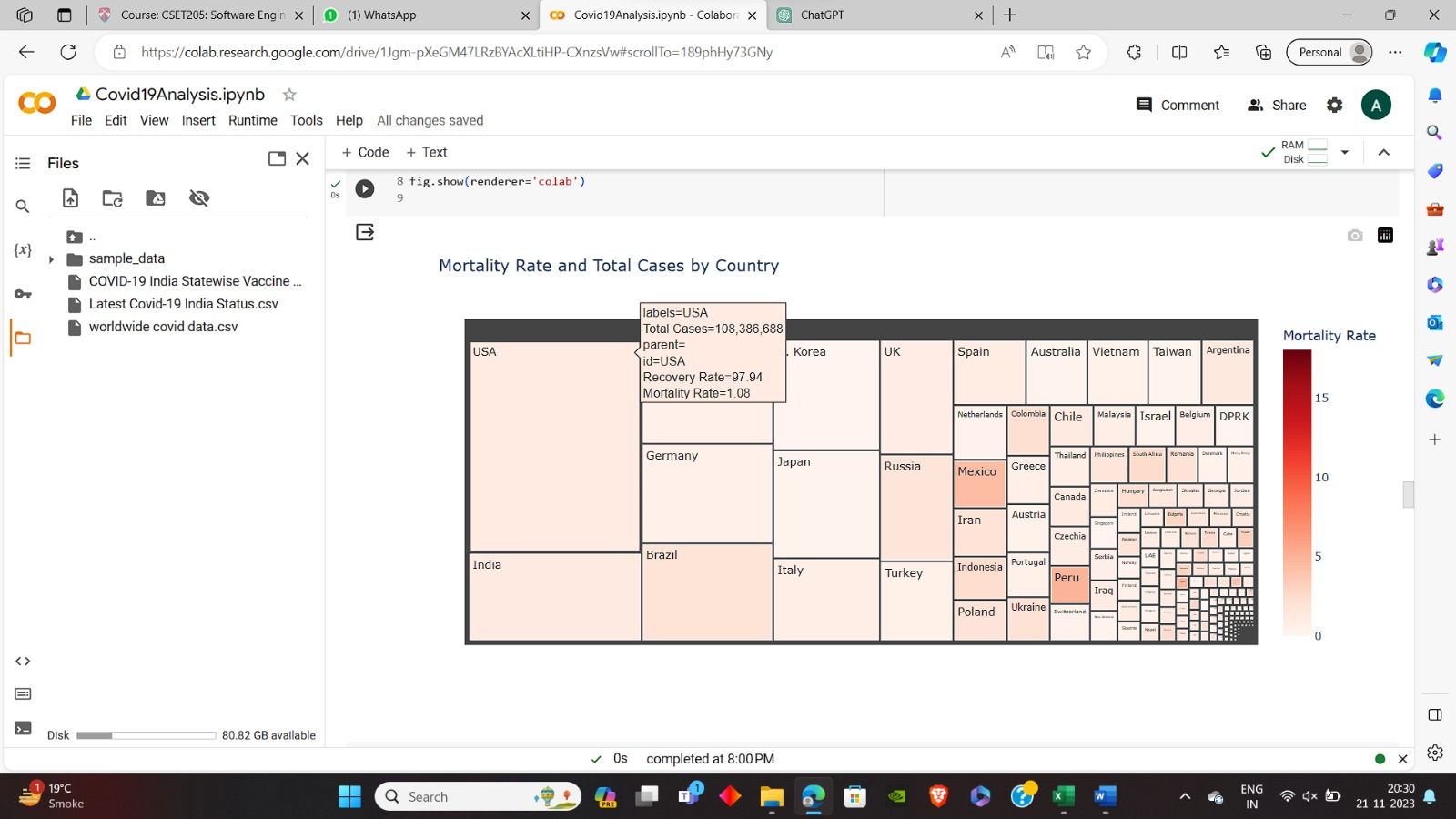
**Corelation**

Correlation is a statistical measure that describes the degree of association between two variables. It quantifies the strength and direction of a linear relationship between the variables such as, total cases, total deaths, total recovered, active cases. The correlation coefficient is a numerical value that ranges from -1 to 1.



**Mortality Rate and Recovery Rate**

Treemap depicting mortality rate and total cases by countries.



**Countries with Mortality Rate > 5%:** ['Sudan', 'Syria', 'Yemen']

**Countries with Recovery Rate < 80%:** ['Algeria', 'Brunei', 'Cayman Islands', 'Chad', 'China', 'Costa Rica', 'French Guiana', 'Greenland', 'Honduras', 'Jamaica', 'Japan', 'Kiribati', 'Lesotho', 'Macao', 'Nicaragua', 'Paraguay', 'Saint Martin', 'Saint Pierre Miquelon', 'Samoa', 'Somalia', 'Tunisia', 'Turkey', 'Uganda', 'Wallis and Futuna', 'Yemen']

**Countries with Both Conditions** (Mortality Rate > 5% and Recovery Rate < 80%): ['Yemen']

**What should be done by Government….**

Here are some general steps that could be advised to the government, WHO (World Health Organization), and UN (United Nations) to address the situation:

**1. Urgent Healthcare Capacity Building:**

* Emergency Medical Facilities: Establish or reinforce emergency medical facilities to handle the increasing number of severe cases.
* Medical Personnel: Deploy additional healthcare professionals to areas in need, providing training and support.

**2. Enhanced Testing and Surveillance:**

* Testing Infrastructure: Strengthen testing infrastructure to identify and isolate cases promptly.
* Surveillance Systems: Implement robust surveillance systems to monitor the spread of the virus and identify hotspots.

**3. Vaccination Acceleration:**

* Vaccination Prioritization: Prioritize vaccination efforts, ensuring that vulnerable populations and healthcare workers are vaccinated first.
* Public Awareness: Conduct targeted vaccination awareness campaigns to address vaccine hesitancy.

**4. Supply Chain Management:**

* Medical Supplies: Ensure a stable and sufficient supply of medical equipment, medications, and other essential supplies.
* Oxygen Supply: Strengthen oxygen supply chains to meet the increased demand for patients requiring respiratory support.

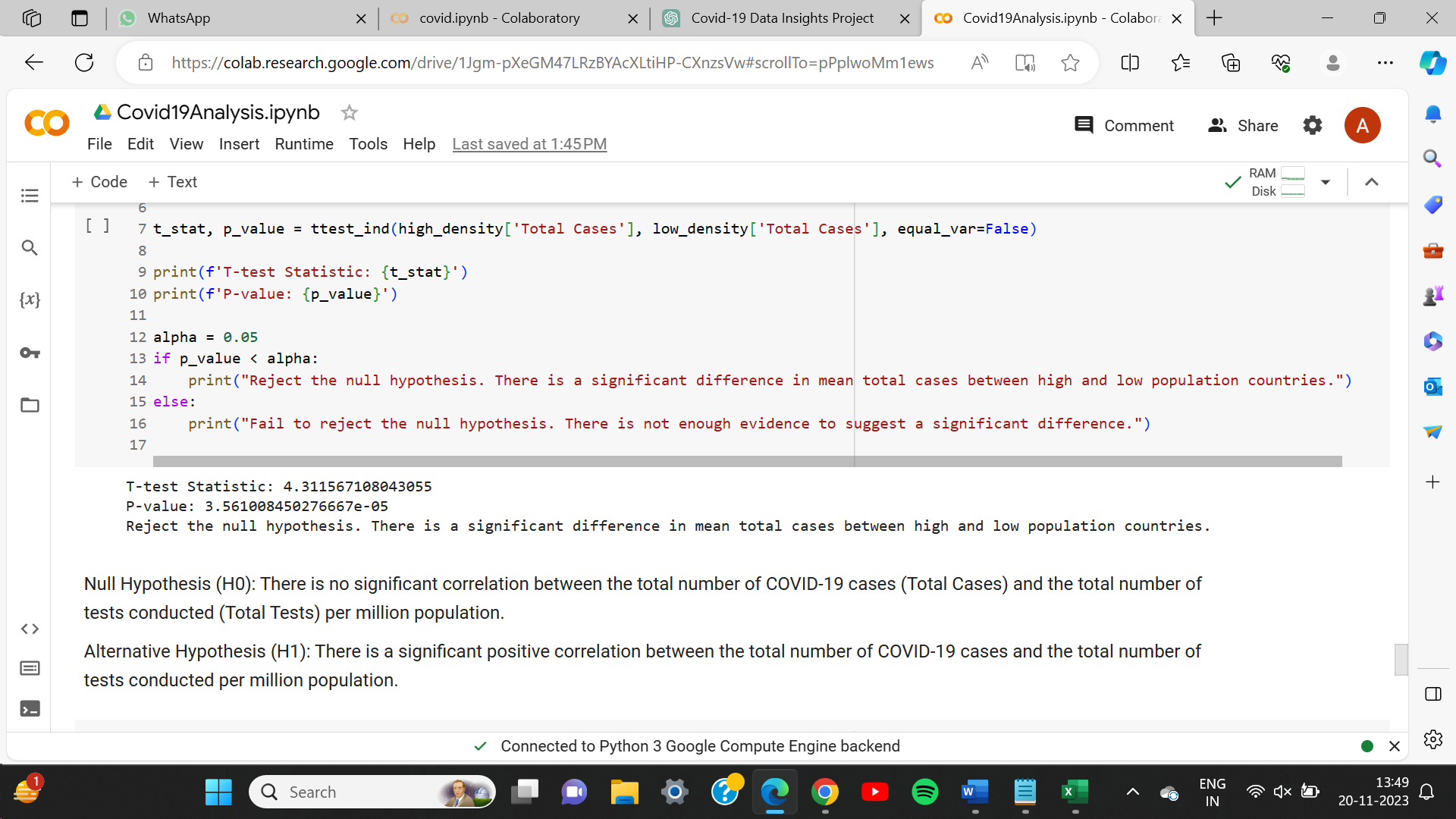
**Regression Analysis**

Linear regression is a statistical method used to model the relationship between a dependent variable and one or more independent variables. It assumes a linear relationship, meaning that a change in the independent variable(s) is associated with a constant change in the dependent variable. The goal of linear regression is to find the best-fit line that minimizes the difference between the observed values and the values predicted by the model.

It depicts deaths vs tests per population.

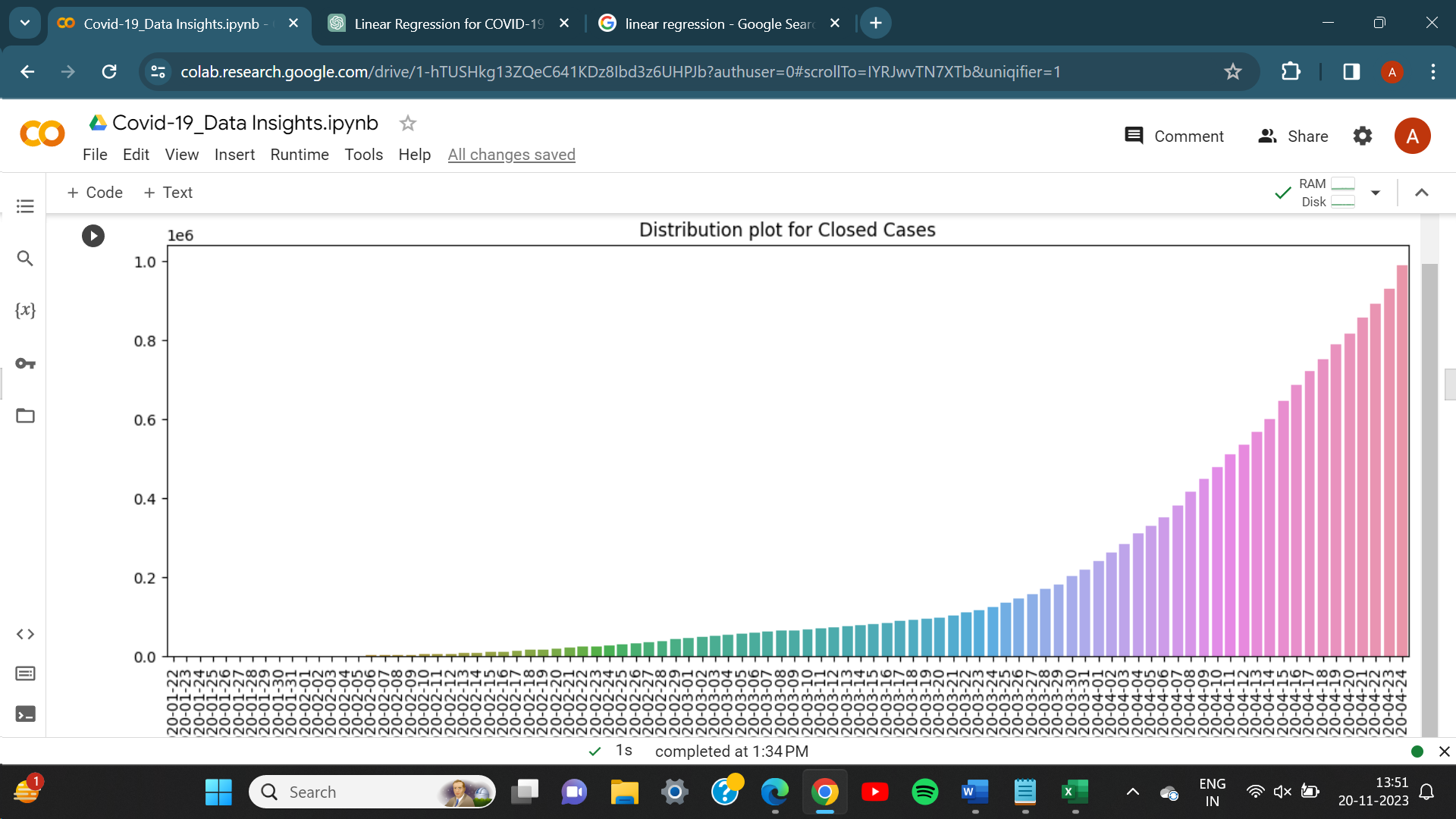


**Hypothesis Testing**

Hypothesis testing is a statistical method used to make inferences about population parameters based on a sample of data. It involves formulating a hypothesis, collecting data, and assessing whether the observed results provide enough evidence to reject or fail to reject the null hypothesis.

**Time-Series Analysis**

Time series analysis involved the exploration of overall trends in Covid-19 metrics such as cases, deaths, and recoveries. Techniques like moving averages or exponential smoothing were applied to smooth out noise and highlight underlying patterns. The analysis aimed to discern whether there was a consistent upward or downward trend over the course of the pandemic, providing insights into the trajectory of the virus.



**Which countries require more testing centres?**

As depicted by the image, where there are slower testing rates, those countries need more attention towards building new testing centres and buying more test-kits. And the countries having faster testing rates, are in good conditions.



**Countries with Slower Testing Rates:**

['Afghanistan', 'Albania', 'Algeria', 'Angola', 'Antigua and Barbuda', 'Argentina', 'Azerbaijan', 'Bahamas', 'Bangladesh', 'Benin', 'Bolivia', 'Bosnia and Herzegovina', 'Botswana', 'Brazil', 'Burkina Faso', 'Burundi', 'CAR', 'Cabo Verde', 'Cambodia', 'Cameroon', 'Chad', 'China', 'Colombia', 'Comoros', 'Congo', 'Costa Rica', 'DPRK', 'DRC', 'Djibouti', 'Dominican Republic', 'Ecuador', 'Egypt', 'El Salvador', 'Equatorial Guinea', 'Eritrea', 'Eswatini', 'Ethiopia', 'Fiji', 'French Polynesia', 'Gabon', 'Gambia', 'Ghana', 'Guatemala', 'Guinea', 'Guinea-Bissau', 'Guyana', 'Haiti', 'Honduras', 'India', 'Indonesia', 'Iran', 'Iraq', 'Ivory Coast', 'Jamaica', 'Japan', 'Kazakhstan', 'Kenya', 'Kiribati', 'Kyrgyzstan', 'Laos', 'Lebanon', 'Lesotho', 'Liberia', 'Libya', 'Macao', 'Madagascar', 'Malawi', 'Mali', 'Marshall Islands', 'Mauritania', 'Mauritius', 'Mayotte', 'Mexico', 'Micronesia', 'Moldova', 'Morocco', 'Mozambique', 'Myanmar', 'Namibia', 'Nepal', 'New Caledonia', 'Nicaragua', 'Niger', 'Nigeria', 'Pakistan', 'Palestine', 'Papua New Guinea', 'Paraguay', 'Philippines', 'Rwanda', 'S. Korea', 'Samoa', 'Sao Tome and Principe', 'Senegal', 'Seychelles', 'Sierra Leone', 'Solomon Islands', 'Somalia', 'South Africa', 'South Sudan', 'Sri Lanka', 'Sudan', 'Suriname', 'Syria', 'Tajikistan', 'Tanzania', 'Thailand', 'Timor-Leste', 'Togo', 'Trinidad and Tobago', 'Tunisia', 'Uganda', 'Ukraine', 'Uzbekistan', 'Vanuatu', 'Venezuela', 'Vietnam', 'Yemen', 'Zambia', 'Zimbabwe']

**What should be done by Government….**

Here are some steps that could be advised to the government, WHO (World Health Organization), and UN (United Nations) to improve testing rates:

**1. Strengthen Testing Infrastructure:**

* Testing Facilities: Increase the number of testing facilities, especially in underserved areas.
* Mobile Testing Units: Deploy mobile testing units to reach remote or high-risk communities.

**2. Testing Accessibility:**

* Community Testing Centers: Establish accessible and community-based testing centers.
* Free Testing: Ensure testing is widely available and, if possible, free of charge to encourage widespread participation.

**3. Public Awareness Campaigns:**

* Inform the Public: Launch public awareness campaigns to educate citizens about the importance of testing for early detection and control of the virus.
* Address Stigma: Address any stigma associated with testing and COVID-19 to encourage more people to get tested.

**4. Data Transparency and Reporting:**

* Timely Reporting: Ensure timely and accurate reporting of testing data to provide a clear picture of the COVID-19 situation.
* Data Sharing: Collaborate with international organizations to share testing data for a more comprehensive global understanding.

**5. International Collaboration:**

* WHO Support: Seek technical and financial support from the WHO to enhance testing capacities.
* Resource Sharing: Collaborate with countries with higher testing capacities for knowledge and resource sharing.

**6. Training Healthcare Workers:**

* Capacity Building: Train healthcare workers in efficient testing procedures and ensure they have the necessary equipment and protective gear.
* Remote Areas: Focus on training healthcare workers in remote or rural areas.

**7. Ramp Up Contact Tracing:**

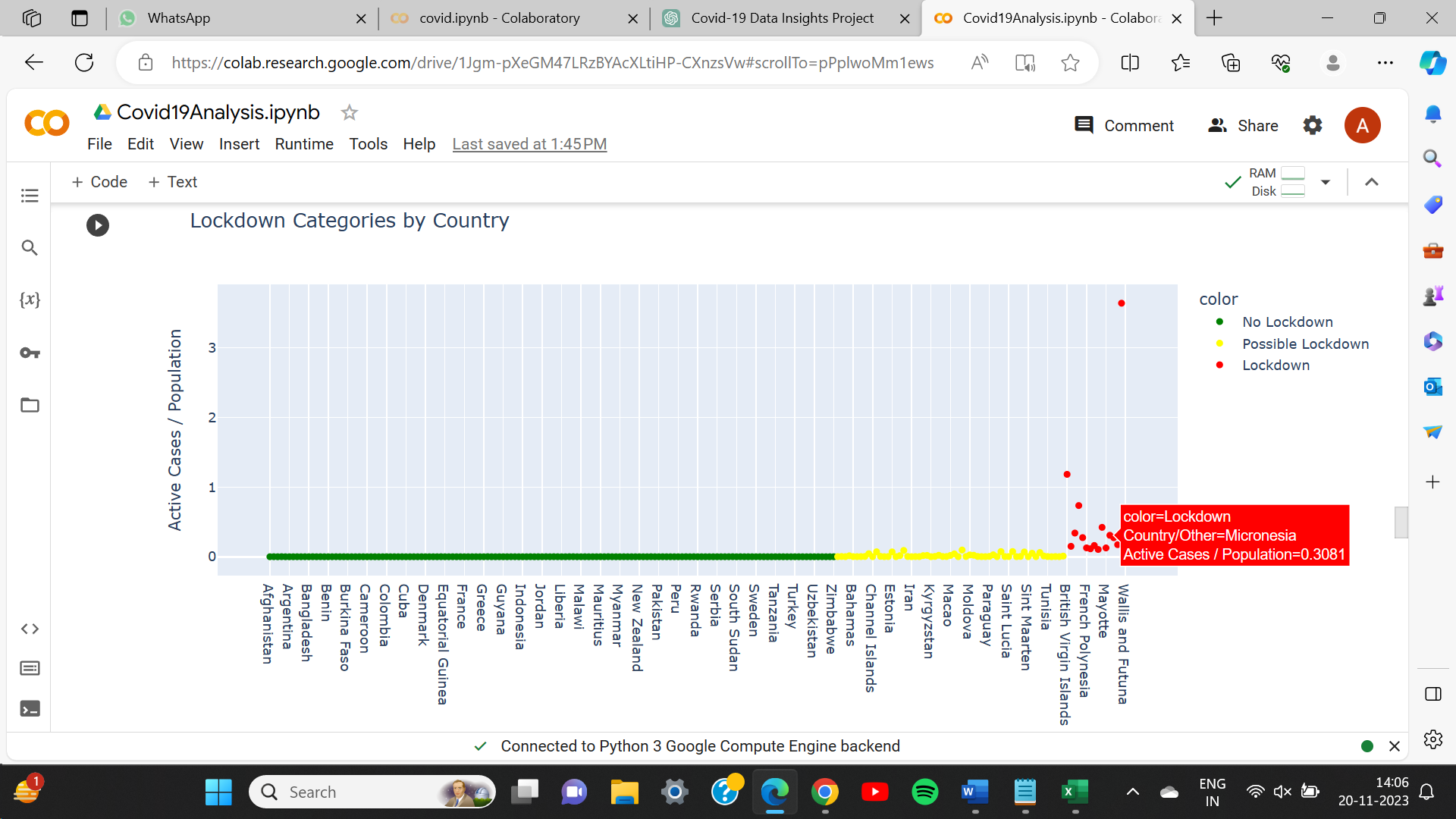
* Contact Tracing Teams: Increase the number of contact tracing teams to identify and isolate cases promptly.
* Technology Utilization: Utilize technology for more efficient contact tracing where feasible.

**8. Mobile Health Apps:**

* Use of Apps: Implement mobile health apps for easy scheduling of tests, result notifications, and contact tracing.
* Digital Solutions: Leverage technology for efficient data collection and reporting.

**Lockdowns?**

After analysing, countries shown in green colour requires no lockdown, countries in yellow colour requires possible lockdown and the countries with red colour are those where it is necessary for a lockdown.



**Countries for Lockdown:**

['Brunei', 'Cayman Islands', 'French Guiana', 'Gibraltar', 'Greenland', 'Saint Martin', 'Saint Pierre Miquelon', 'Wallis and Futuna']

**Countries for Possible Lockdown:**

['Algeria', 'Andorra', 'Armenia', 'Aruba', 'Bahamas', 'Bhutan', 'Bosnia and Herzegovina', 'Brazil', 'British Virgin Islands', 'Caribbean Netherlands', 'Channel Islands', 'Costa Rica', 'Curaçao', 'Dominican Republic', 'El Salvador', 'Estonia', 'Faeroe Islands', 'Fiji', 'French Polynesia', 'Georgia', 'Guadeloupe', 'Iceland', 'Iran', 'Isle of Man', 'Israel', 'Italy', 'Jamaica', 'Kiribati', 'Kyrgyzstan', 'Lebanon', 'Lesotho', 'Luxembourg', 'Macao', 'Maldives', 'Malta', 'Martinique', 'Mayotte', 'Mexico', 'Micronesia', 'Moldova', 'Monaco', 'Nauru', 'Nicaragua', 'Oman', 'Poland', 'Russia', 'Réunion', 'S. Korea', 'Saint Lucia', 'Samoa', 'San Marino', 'Singapore', 'Sint Maarten', 'Solomon Islands', 'South Africa', 'St. Barth', 'Suriname', 'Tonga', 'UAE', 'USA', 'Uganda', 'Vietnam']

**In these countries the Government, UN and WHO should:**

**Implement Strict Lockdown Measures:** Enforce a strict lockdown in the listed countries to control the spread of the virus. This may include stay-at-home orders, closure of non-essential businesses, and restrictions on public gatherings.

**Enhance Testing and Contact Tracing:** Increase testing capacity to identify and isolate cases promptly. Strengthen contact tracing efforts to break the chains of transmission.

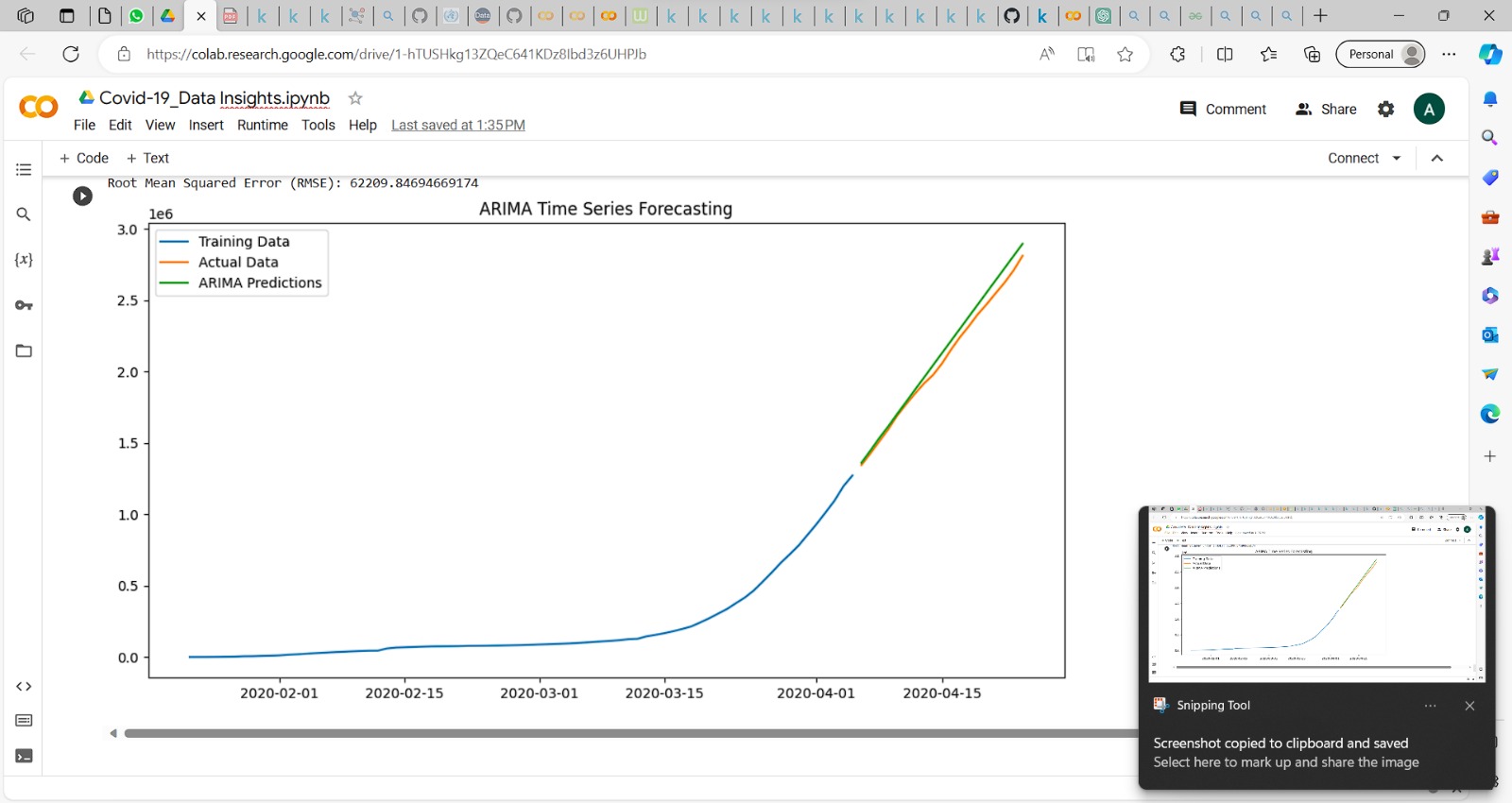
**Healthcare Capacity:** Strengthen healthcare infrastructure to handle an increased number of cases.Ensure an adequate supply of medical equipment, medications, and healthcare personnel.

**Public Communication:** Communicate clearly and transparently with the public about the severity of the situation and the importance of adhering to lockdown measures.Provide regular updates and information on vaccination opportunities.

**International Collaboration:** Seek international support and collaboration for resource-sharing, including medical supplies, expertise, and financial assistance.

**Time-Series Forecasting**

The training data, actual data, and ARIMA predictions are plotted for visualization. This helps to visually assess how well the model is capturing the patterns in the data. Adjusting the ARIMA order and other parameters may be necessary based on the characteristics of your specific dataset. The goal is to find a model that provides accurate predictions for your time series data.



**Time-Series Modelling:**

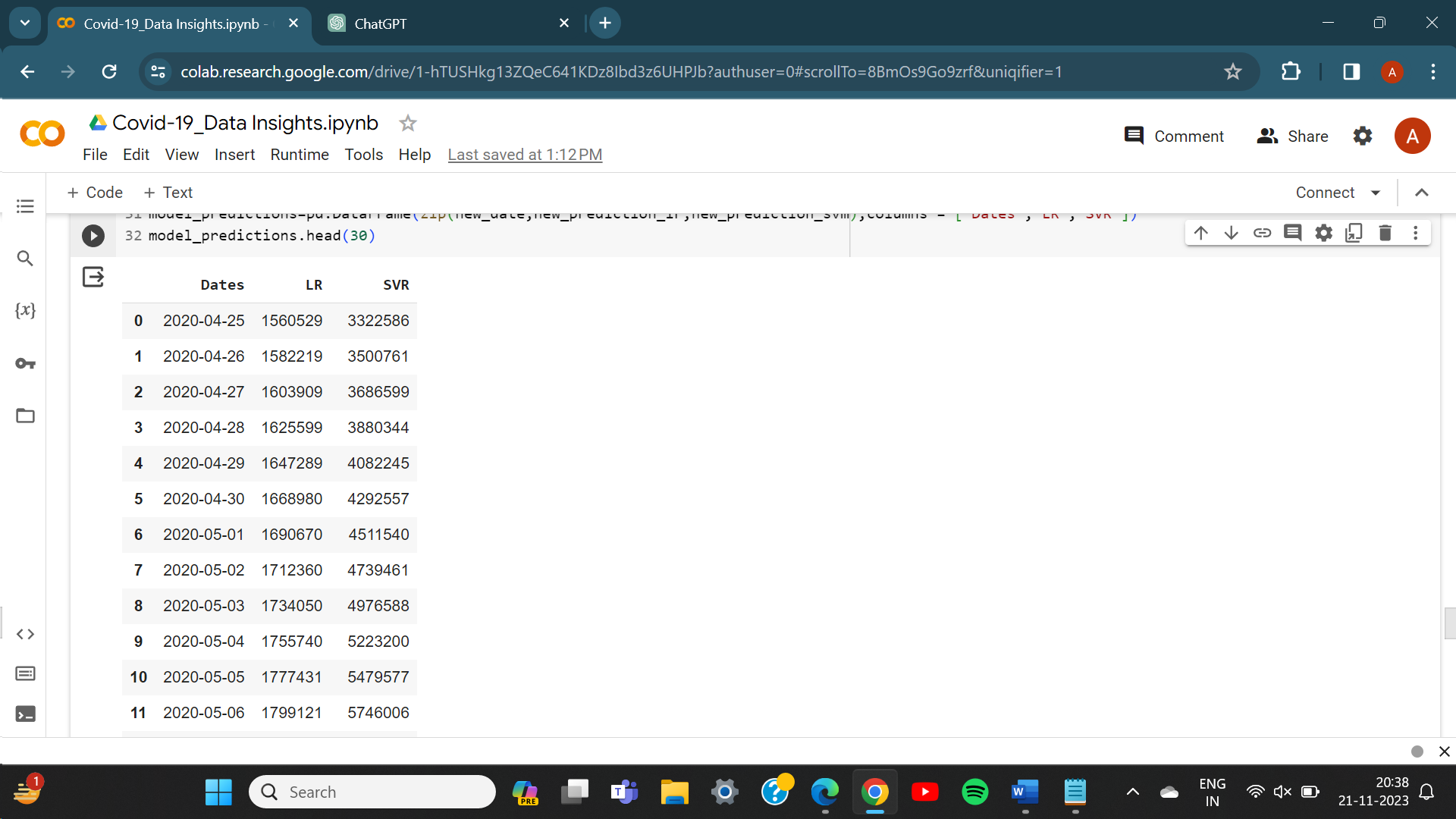
**Goal:** Predict future COVID-19 cases using Linear Regression and SVR models.

The dataset is split into training and validation sets.

Train Linear Regression and SVR models on the training data.

Evaluate model performance on the validation set.

Predict future cases for the next 30 days using both models.



**Limitations**

In conducting the "Covid-19 Data Insights" project, it is imperative to recognize and address several limitations that may impact the accuracy and generalizability of the findings. One notable limitation revolves around the quality and completeness of the Covid-19 data utilized in the analysis. The reliability of the insights heavily relies on the accuracy of reported cases, deaths, and recoveries. The variability in healthcare infrastructure among different regions also poses a challenge. Disparities in testing capabilities, reporting practices, and healthcare resources can impact the detection and reporting of cases. The demographic composition of populations, including age distribution, comorbidities, and population density, is another factor influencing the spread and severity of the virus. Failing to account for these demographic factors may limit the generalizability of the findings across diverse populations.

Model simplifications, such as the use of linear regression, may oversimplify the complex dynamics of the pandemic. The chosen model may not fully capture all contributing factors, potentially leading to inaccuracies in predictions. The chosen model may not fully capture all contributing factors, potentially leading to inaccuracies in predictions. Additionally, the impact of Covid-19 exhibits regional variability due to factors such as government interventions, public compliance with safety measures, and cultural practices. Generalizing findings from one region to another may be challenging and should be approached with caution.

Finally, ethical considerations related to data privacy and informed consent should be acknowledged. These considerations may impact the availability and usability of certain data sources, influencing the scope and depth of the analysis. Addressing these limitations transparently is essential for a comprehensive understanding of the project's results and provides a foundation for future research and improvements.

**Conclusion**

In conclusion, this comprehensive project titled 'Covid-19 Data Insights' has been instrumental in offering valuable perspectives on the ongoing global pandemic through the lens of probability and statistics. The exploration of Covid-19 data has provided illuminating insights into the intricate trends, correlations, and influential factors that contribute to the spread and impact of the virus.

The examination of trends over time, correlations between various variables, and the identification of patterns in the data contribute significantly to our comprehension of the pandemic's complexities. These insights not only serve as a reflective mirror of past occurrences but also provide a foundation for anticipating future trajectories. The ability to discern key factors influencing the virus's dissemination is pivotal for shaping informed decision-making processes and implementing effective public health interventions. Moreover, the utilization of probability and statistics has not only allowed for a descriptive overview of the data but has also facilitated predictive modelling, offering a glimpse into potential future scenarios. This forward-looking aspect of the analysis adds a proactive dimension to the project, enabling stakeholders to anticipate challenges and allocate resources strategically.

In essence, this project serves as a significant contribution to the ongoing collective effort to comprehend, manage, and eventually overcome the challenges posed by the Covid-19 pandemic. By harnessing the power of probability and statistics, it offers a data-driven narrative that not only informs current decision-making processes but also lays the groundwork for future research endeavours in public health and epidemiology.