



SRI RAMACHANDRA

INSTITUTE OF HIGHER EDUCATION AND RESEARCH

(Category - I Deemed to be University) Porur, Chennai

SRI RAMACHANDRA FACULTY OF ENGINEERING AND TECHNOLOGY

INT-200 Internship FINAL REVIEW

Project Type: Internal (Data Analysis)

Project Title: Covid-19 Data Analysis

Presented By:

○ Rishitha Thoka- E0322026

Guided By:

Dr. ASHOK KUMAR

• PLAN OF THE TALK •

1. Introduction
2. Survey
3. Problem statement / Objectives
4. System requirements
5. Workflow / Methodology
6. Work Done
7. Timeline of the project
8. Implementation
9. Test Result
10. Conclusion
11. Future scope
12. Mentor's guidance
13. References
14. Sample code

INTRODUCTION

The COVID-19 pandemic has presented unprecedented challenges globally, requiring data-driven approaches to understand and combat the spread of the virus. In response, this COVID-19 data analytics project utilizes Python and its powerful data analysis capabilities to analyze and visualize the available COVID-19 datasets. By leveraging Python's libraries such as Pandas, NumPy, Matplotlib, and scikit-learn, the project aims to extract meaningful insights, track the progression of the pandemic, and develop predictive models. These insights and models will provide valuable information for decision-making, resource allocation, and public health responses, ultimately contributing to the global efforts in managing and controlling the impact of the COVID-19 pandemic.

PROBLEM

STATEMENT / OBJECTIVE

The objective of the COVID-19 data analysis project is to utilize Python and its data analysis libraries to gain insights from the available COVID-19 datasets, track the spread of the virus, and provide valuable information for decision-making and public health responses.

The project aims to address the following challenges:

- ☐ Data Exploration and Cleaning
- ☐ Descriptive Analysis
- ☐ Data Visualization
- ☐ Predictive Modelling
- ☐ Insights for Decision-Making

• RESEARCH / PRODUCT SURVEY •

- ❑ The research/product survey conducted as part of our COVID-19 data analysis project has provided valuable insights and perspectives to complement our data-driven analysis.
- ❑ The findings have enhanced our understanding of the impact of the pandemic, behavioral changes, perception of public health measures, emotional well-being, information sources, and user feedback on COVID-19 data and visualizations.
- ❑ These insights will further inform our data analysis, visualization, and decision-making processes as we continue to contribute to the global efforts in combating the COVID-19 pandemic.

WORK FLOW / METHODOLOGY

Here are the key points to follow a systematic workflow and methodology for COVID-19 data collection:

1. Identify reliable data sources:

- ☐ Look for reputable sources such as government health agencies (e.g., CDC, WHO), academic institutions, and reliable research publications.
- ☐ Check for data sources that provide regular updates and have a track record of accurate reporting.

2. Validate the data:

- ☐ Verify the accuracy and consistency of the data by cross-referencing multiple sources.
- ☐ Ensure that the data is in a standardized format, such as using common variables and units, to facilitate meaningful analysis and comparisons.

3. Document the process:

- ☐ Maintain a detailed record of the data collection process, including the sources accessed, data extraction methods, and any data transformations or cleaning steps performed.
- ☐ This documentation helps ensure transparency, reproducibility, and traceability of the data analysis.

4. Build a robust dataset:

- ☐ Continuously update the dataset with the latest available data to reflect the evolving nature of the COVID-19 pandemic.
- ☐ Regularly review and validate the data to maintain its accuracy and reliability for ongoing analysis and future research.

Work Done

- ❑ Data Sources: Identify reliable and authoritative sources of COVID-19 data. This may include national health agencies, international organizations (e.g., WHO, CDC), research institutions, and official government websites. Consider the availability of comprehensive and up-to-date datasets for various aspects of the pandemic, such as cases, deaths, recoveries, testing, hospitalizations, and vaccination.
- ❑ Data Collection: Develop a data collection plan to retrieve the necessary information from the identified sources. This can involve manual data entry from online platforms, web scraping techniques, or utilizing APIs (Application Programming Interfaces) provided by data sources. Ensure that the collected data is comprehensive and covers relevant geographical regions and time periods

Work Done

- ❑ **Data Validation:** Implement a data validation process to ensure data accuracy and reliability. Cross-reference the collected data from multiple sources to identify any inconsistencies or discrepancies. Verify data integrity by comparing it with official reports and published studies. Cleanse the data by removing duplicates, correcting errors, and addressing missing values.
- ❑ **Data Standardization:** Standardize the collected data to ensure consistency and compatibility for further analysis. This involves converting data into a common format, addressing differences in units of measurement, and harmonizing variable names and categories. Adopt standardized data formats such as CSV (Comma-Separated Values) or JSON (JavaScript Object Notation) for easy integration into analysis tools

Work Done

- ❑ Data Documentation: Maintain detailed documentation of the data collection process. This includes recording the sources, dates of data collection, data extraction methods, any modifications made to the original data, and relevant metadata. Documentation facilitates transparency, reproducibility, and allows for future reference or updates to the dataset.
- ❑ Linear Regression :
 - By applying linear regression to COVID-19 data, researchers can analyze the correlation between factors like infection rates, vaccination rates, socioeconomic indicators, and other relevant variables.
 - This method enables the estimation of trends, predictions, and the impact of various factors on the spread and mitigation of the virus.
 - The findings from linear regression models can contribute to formulating effective public health policies, improving resource allocation, and guiding decision-making processes during the ongoing fight against the COVID-19 pandemic.

Work Done



Data visualization:

- ☐ Data visualization has played a crucial role in COVID-19 data analysis projects, enabling researchers and the general public to comprehend complex information and trends with ease.
- ☐ By leveraging the power of visual representations, data visualization has played a pivotal role in increasing public awareness, facilitating informed decision-making, and supporting effective public health responses during the COVID-19 pandemic.

Implementation

- ❑ Reliable and up-to-date COVID-19 dataset was collected, including infection rates, mortality rates, and mitigation measures.
- ❑ Preprocessing steps were performed using the pandas library to clean the data, handle missing values, and transform variables as necessary.
- ❑ Exploratory data analysis techniques were employed to gain initial insights, identify trends, and patterns in the data.
- ❑ Statistical techniques like linear regression were utilized to analyze relationships between variables and assess the impact of factors on the spread and severity of the virus.
- ❑ Visualizations were created using the matplotlib library to present findings in a clear and interpretable manner, including line charts, bar graphs, and heatmaps.

Implementation

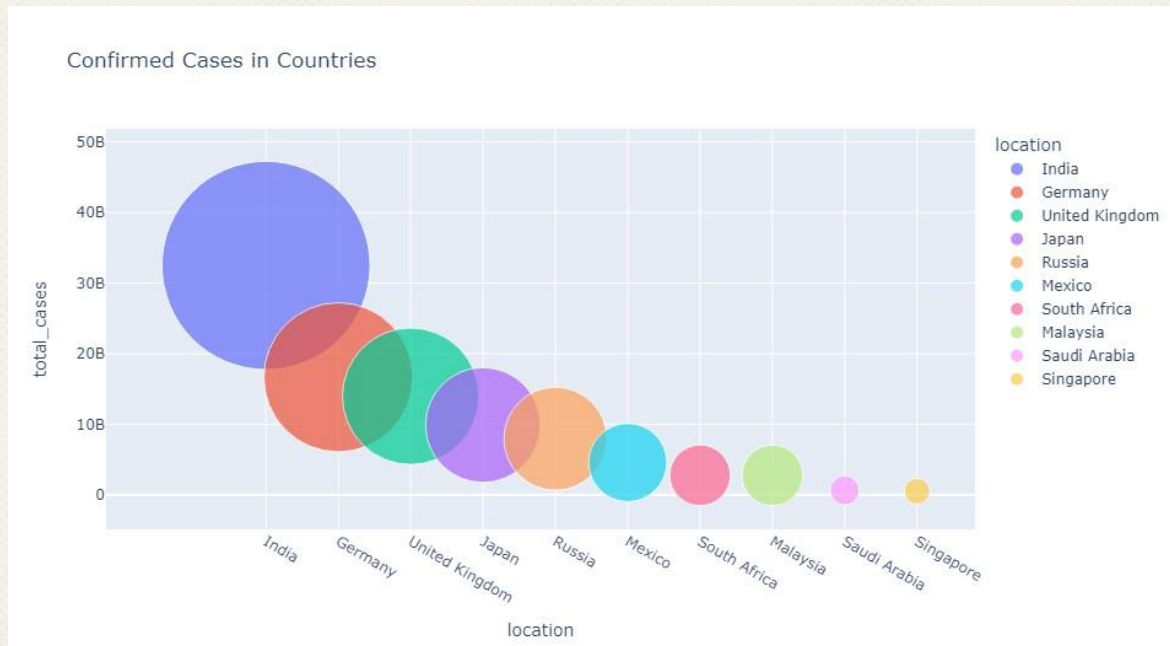
- ❑ The effectiveness of mitigation measures was assessed through visualizations and statistical analysis.
- ❑ A meta-analysis was conducted to evaluate the real-world effectiveness of COVID-19 vaccines, synthesizing data from multiple studies.
- ❑ The project aimed to provide evidence-based insights to support decision-making, inform public health interventions, and guide policy-making in the ongoing fight against COVID-19.

• TIMELINE OF THE WORK •

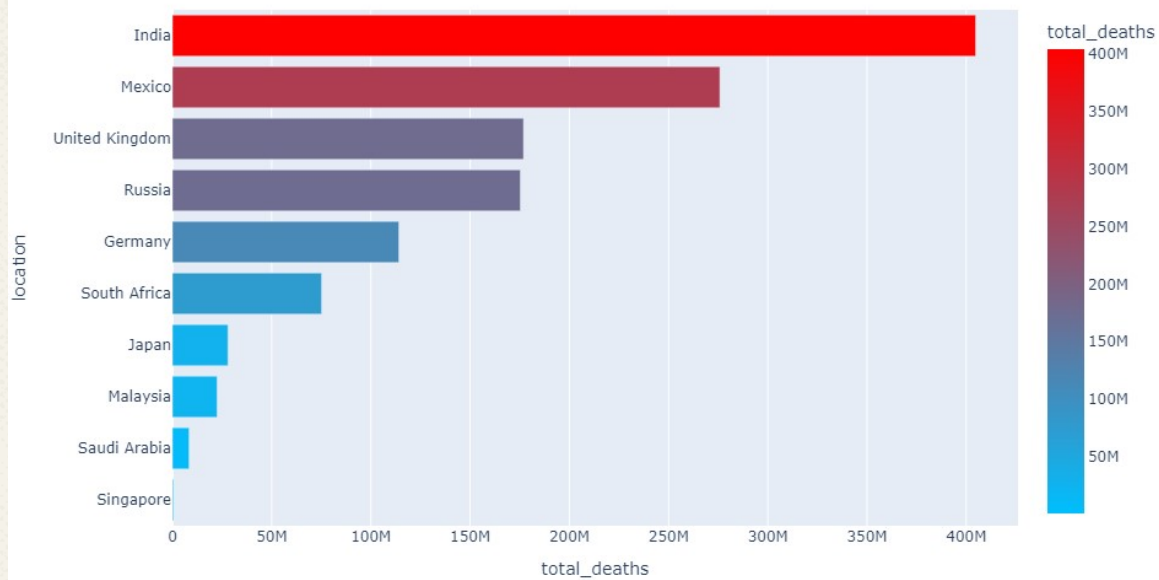
- 09/05/2023- Discussion of Project with team and project guide
- 11/05/2023- Finding Data sources on covid-19
- 16/05/2023- Data collection and Data gathering
- 18/05/2023- Data validation
- 23/05/2023- Data standardization
- 25/05/2023- Data documentation

- 30/05/2023- Preparing ppt for review
- 01/06/2023- First review
- 06/06/2023- Revising Linear regression
- 08/06/2023- Revising Linear regression
- 13/06/2023- Coding
- 15/06/2023- Coding
- 20/06/2023- Coding
- 22/06/2023- Preparing ppt for review

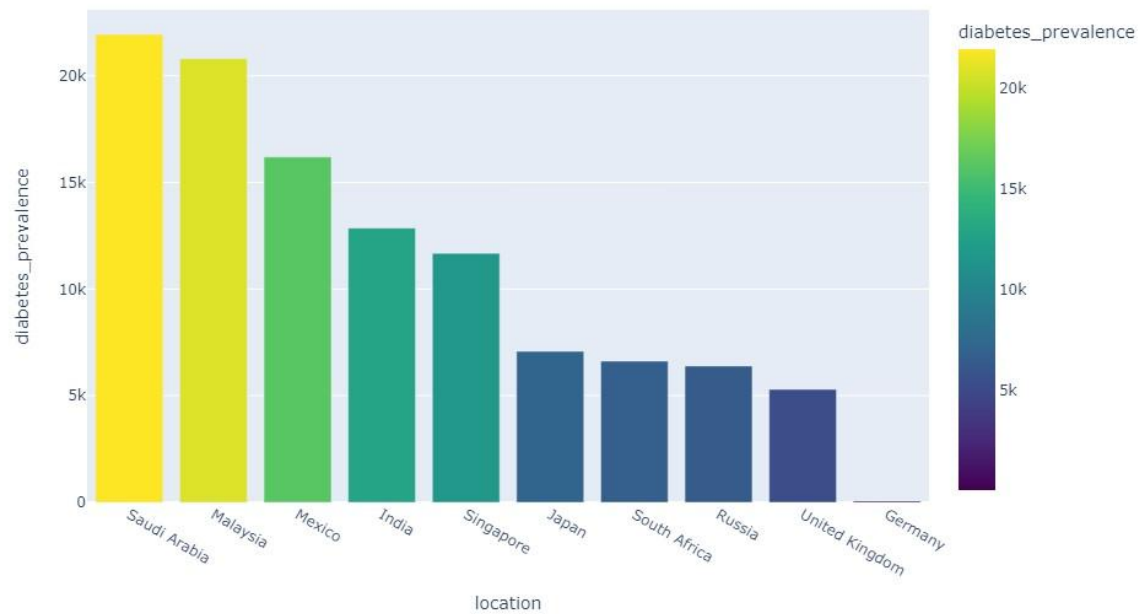
Test Results



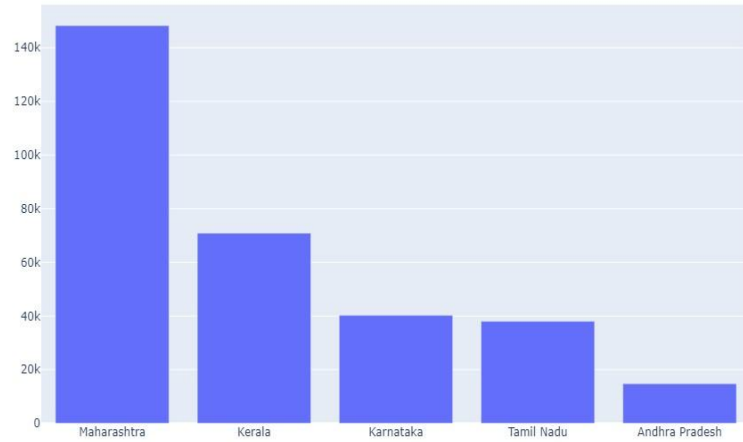
Death Cases in Countries



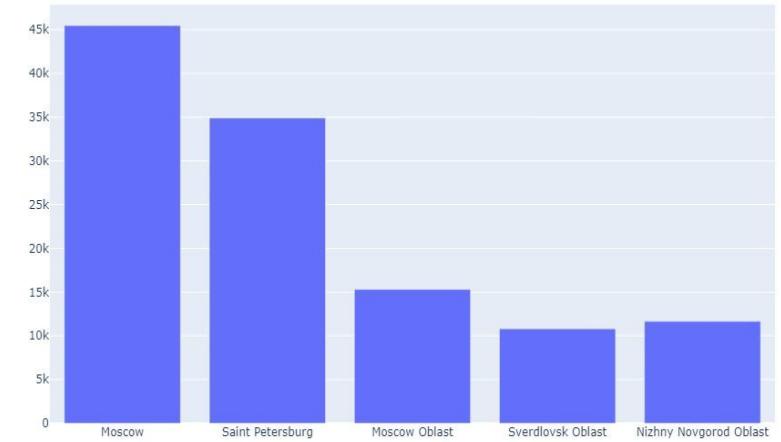
Diabetes prevalence in the countries



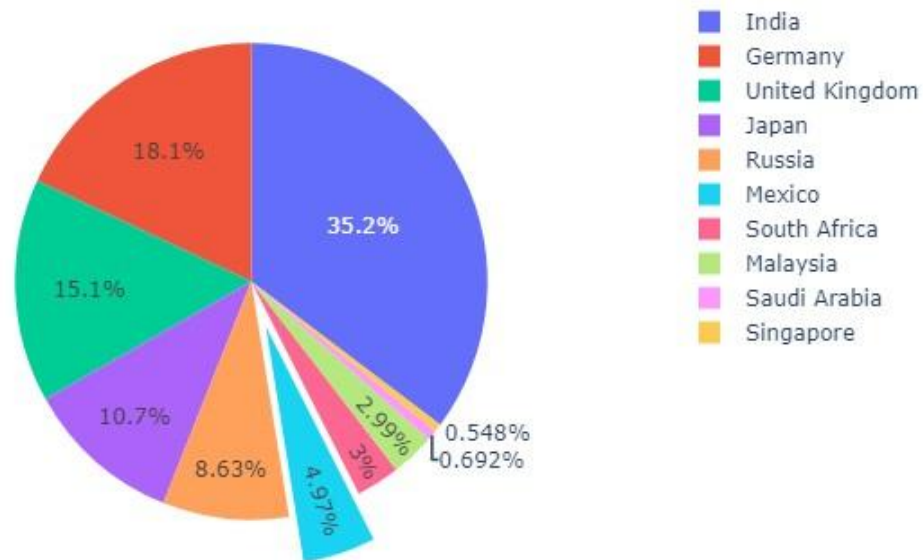
Death rates in Most Affected States in India



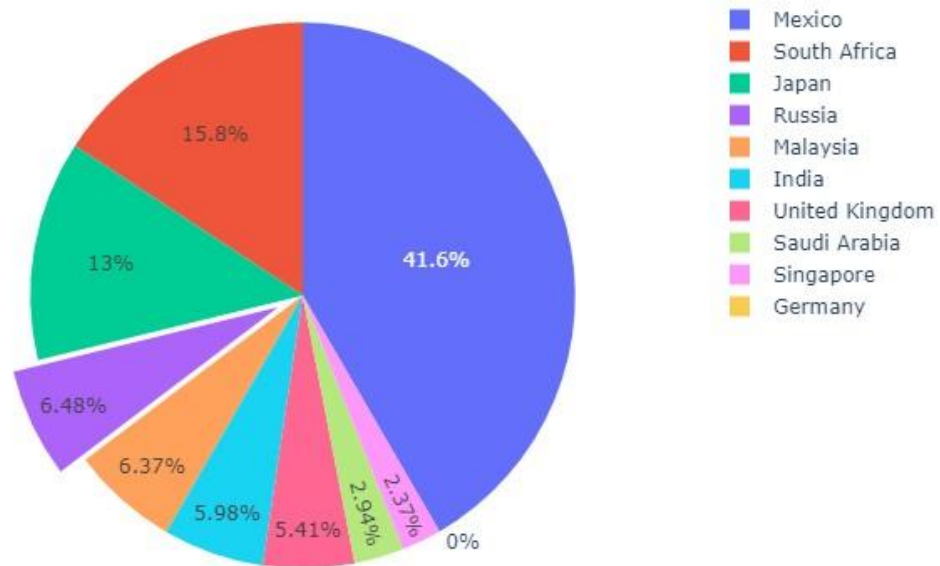
Death rates in Most Affected States in Russia



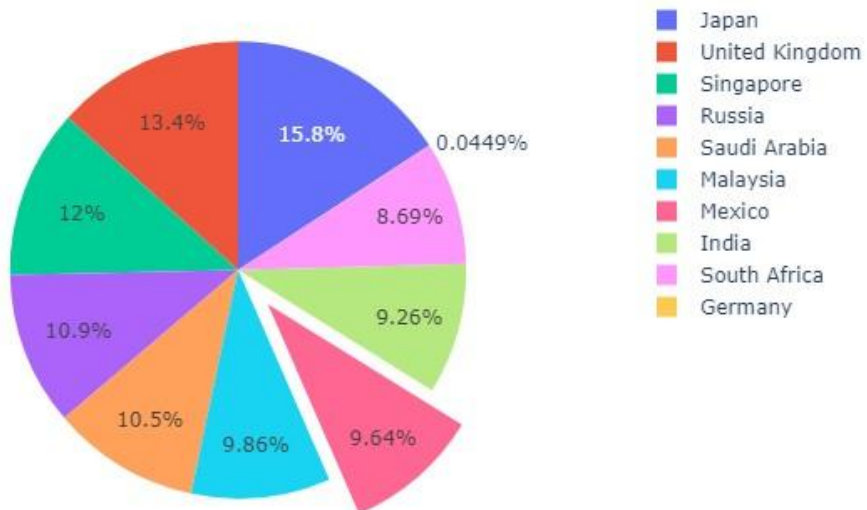
WHO Region-Wise Case Distribution



Positive Rate



Median age



Conclusion

- ❑ In conclusion, this COVID-19 data analysis project provided valuable insights into the dynamics of the pandemic.
- ❑ The findings underscored the importance of proactive measures such as vaccination, social distancing, and mask-wearing in controlling the spread of the virus.
- ❑ The results can inform public health policies and interventions, aiding in the development of targeted strategies to mitigate the impact of COVID-19.
- ❑ The study emphasized the need for continued vigilance, adherence to preventive measures, and widespread vaccination to overcome the challenges posed by the ongoing pandemic.








Future Scope

- ❑ Data analysis coupled with effective visualizations can help policymakers, healthcare professionals, and the general public better understand the patterns and trends of the virus's transmission.
- ❑ By visualizing data such as infection rates, mortality rates, vaccination coverage, and hotspot areas, decision-makers can make informed choices regarding public health measures, resource allocation, and containment strategies.
- ❑ Moreover, interactive visualizations can empower individuals to engage with the data, enabling them to comprehend the impact of the virus on their communities and make informed decisions about their own health and safety.
- ❑ With the availability of real-time data and advancements in visualization technologies, the potential for leveraging data analysis to combat COVID-19 and future pandemics is promising.

Mentor's guidance

- Our work was monitored and was updated to our mentor from time to time.
- We clarified our doubts regarding the project, its use, functioning, and its report

Reference

-  <https://ourworldindata.org/covid-deaths>
-  https://services1.arcgis.com/0MSEUqKaxRIEPj5g/arcgis/rest/services/Coronavirus_2019_nCoV_Cases/FeatureServer/1/query?where=1%3D1&outFields=* &outSR=4326&f=json
-  <https://medium.com/swlh/worldwide-covid-19-analysis-visualization-339002a821fe>
-  <https://covid19.who.int/data>
-  Dash, S., Chakraborty, C., Giri, S. K., & Pani, S. K. (2021). Intelligent computing on time-series data analysis and prediction of COVID-19 pandemics. Pattern Recognition Letters, 151, 69-75.
-  Mittal, S. (2020). An exploratory data analysis of COVID-19 in India. International Journal of Engineering and Technical Research, 9(4).
-  Nair, R., Soni, M., Bajpai, B., Dhiman, G., & Sagayam, K. M. (2022). Predicting the death rate around the world due to COVID-19 using regression analysis. International Journal of Swarm Intelligence Research (IJSIR), 13(2), 1-13.

Sample code

Regression:

```
from sklearn.model_selection import train_test_split
from sklearn.linear_model import LinearRegression
X_train, X_test, y_train, y_test = train_test_split(x, y, test_size = 0.25)
# Splitting the data into training and testing data
regr = LinearRegression()
regr.fit(X_train, y_train)
print(regr.score(X_test, y_test))
data=pd.read_excel(r"C:\Users\RISHITHA\OneDrive\Desktop\INT200\Covid_dataset.xlsx",sheet_name="I
NDIA")
```

```
d=data.loc[:,["total_cases","total_deaths"]]
d=d.dropna()
d["id"]=range(1,len(d)+1)
x=d.iloc[:,2:].values
y=d.iloc[:,0].values
X_train, X_test, y_train, y_test = train_test_split(x, y, test_size = 0.25)
# Splitting the data into training and testing data
regr = LinearRegression()
regr.fit(X_train, y_train)
print(regr.score(X_test, y_test))
```

```
val1=[]
for names in ["INDIA","GERMANY","JAPAN","UK","GHANA"]:
    data=pd.read_excel(r"C:\Users\RISHITHA\OneDrive\Desktop\INT200\Covid_dataset.xlsx",sheet_name=names)
    index=data.index
    d=data.loc[:,["total_cases","total_deaths"]]
    d=d.dropna()
```



```
d["id"]=range(1,len(d)+1)
x=d.iloc[:,2:].values
y=d.iloc[:,0].values
X_train, X_test, y_train, y_test = train_test_split(x, y, test_size = 0.2)
regr = LinearRegression()
regr.fit(X_train, y_train)
print(names)
v=regr.score(X_test, y_test)
val1=val1+[v]
print(v)
print("Total Cases = {1}*X+{0}".format(regr.intercept_,regr.coef_))
a2=800*regr.coef_+regr.intercept_
print(a2)
```

Data Visualization:

```
top10_confirmed =  
pd.DataFrame(data.groupby('location')['total_cases'].sum().nlargest(10).sort_values(ascending =  
False))  
fig1 = px.scatter(top10_confirmed, x = top10_confirmed.index, y = 'total_cases', size = 'total_cases',  
size_max = 120, color = top10_confirmed.index, title = 'Confirmed Cases in Countries')  
fig1.show()
```

```
top10_deaths =  
pd.DataFrame(data.groupby('location')['total_deaths'].sum().nlargest(10).sort_values(ascending =  
True)) fig2 = px.bar(top10_deaths, x = 'total_deaths', y = top10_deaths.index, height = 600, color =  
'total_deaths', orientation = 'h', color_continuous_scale = ['deepskyblue','red'], title = 'Death Cases in  
Countries') fig2.show()
```

```
data= pd.read_excel(r"C:\Users\RISHITHA\OneDrive\Desktop\INT200\covid19.xlsx")
data_location= pd.DataFrame(data.groupby('location')['total_cases'].sum())
labels = data_location.index values = data_location['total_cases']
```

```
fig9 = go.Figure(data=[go.Pie(labels = labels, values = values, pull=[0, 0, 0, 0, 0.2, 0])])
```

```
fig9.update_layout(title = 'WHO Region-Wise Case Distribution', width = 700, height = 400, margin =
dict(t = 0, l = 0, r = 0, b = 0))
```

```
data=pd.read_excel(r"C:\Users\RISHITHA\OneDrive\Desktop\INT200\covid19.xlsx")
data_location= pd.DataFrame(data.groupby('location')['median_age'].sum())
labels = data_location.index values = data_location['median_age']
```

```
fig9 = go.Figure(data=[go.Pie(labels = labels, values = values, pull=[0, 0, 0, 0, 0.2, 0])])
```

```
fig9.update_layout(title = 'Median age', width = 700, height = 400, margin = dict(t = 0, l = 0, r = 0, b = 0))
fig9.show()
```

THANK YOU