Proposal

DataSet- Novel Coronavirus (COVID-19)
Source- https://github.com/CSSEGISandData/COVID-19/tree/master/csse_covid_19_data

We will be working with the publicly available dataset from JHU CSSE's COVID-19 data repository. The data repository consists of different datasets related to covid-19.

What is the problem you are solving?

Data in this dataset is a collection of the covid-19 cases in various countries, in which various locations are taken into consideration. The aim of the project is to provide data analysis of covid-19 pandemic, various cases have been studied like most affected countries due to this pandemic. Study of data from various countries is combined to show the growth of cases and recovery. Through this project we are trying to achieve a sustainable solution and understanding of the behavioral trend of the crisis and, a step towards helping people to understand the spread and predict the raise in cases for a country.

Who benefits from this project?

By the analysis of the dataset the results help any Government or national body to examine the current situation and establish clear protocols to help the stop of the virus. The measures collected in the dataset will provide an accurate statistic to take critical actions in order to contain the spread of the virus. The analysis of the tends in the dataset will be beneficial to have a better idea in the behavioral patterns of the cases and help to mitigate the spike. With the help of the results determined by the analysis, the local bodies can help to suppress any serious conditions that may arise, as the result would give us clear directions of what precautionary measures to be implemented. Hence by carrying out any kind of analytics to provide insights about the dataset will help a number of different sectors and avoid similar situation in future.

Why is the project important?

The pandemic has already taken grip over life of people. Since the start of the pandemic, some countries are facing problem of ever-increasing cases. Through the data analysis of cases one can analyze how countries all over the world are doing in terms of controlling the pandemic. Analyzing data leads to adapt the prevention model of the countries that are doing great in terms of lowering the cases, Predictions can be made with the dataset available to the individual/country/organizations, thus helping to decide how far we are able to control the pandemic or up to what extent we should guide preventive measures.

Analysis is more important than ever during these unprecedented times. As a society, we've seen how important even basic line graphs, bar charts, and heat maps are to understanding the spread of the virus. We've heard a lot about various models in terms of predicting deaths associated with COVID-19. Many people want to see the data and understand the facts in this rapidly changing environment.

What are the business questions you are looking to answer or objectives you are looking to achieve from this project?

Business questions-

Which countries are the most affected from covid-19?
In which countries covid-19 cases were growing fast?
Where was the Covid-19 pandemic having the highest spike?
Which country had the highest death rate?
Where was the virus contained quickly?
How would the analysis help to answer any business question?
What conclusion, we get after performing analysis?

Objectives-

With the help of the data we can estimate the cases and requirements that are essential for the treatment of patients effected by COVID 19. Improve on distribution and control the tools that are necessary in this critical times for example:- If a country is suffering from high amount of covid cases then the essentials are more required there and with the project we can understand how much tools are required for the stability of the situation.

Describe the data at a high level, explain the data collection process, source of data, etc

The dataset represents the detailed characteristics of the 2019 Novel Coronavirus, which provides information about the different countries and provinces with the number of confirmed, deaths and recovered cases. This list includes a complete list of all sources ever used in the data set since January 21, 2010. Some sources listed here (e.g. ECDC, US CDC, BNO News) are not currently relied upon as a source of data. The data is collected from multiple sources such as World Health Organization (WHO), European Centre for Disease Prevention and Control (ECDC) and different US data sources at the state (Admin1) or county/city (Admin2) level.

What insights does this data give and how can it be used In future?

The dataset provides the insights of the day-to-day measures of increase and decrease of the cases, deaths, and recovery for each country. This kind of data is very useful in the current situation as it gives the government, the Ngo's, and other medical and social organization to analyze the covid affected areas for a better decision making.
This data is also future proof as it shows the pattern of transmission of the covid 19 and this kind of dataset can help the countries more resilient to any kind of pandemic in future.

What type of problems are solved with this approach?

With analytics, we are trying to acknowledge the general public about the pandemic and providing solutions to overcome this critical situation. In this project we are approaching with an open mindset and providing solution for general public rather than businesses.

Descriptive Statistics and Exploratory Data Analysis (EDA)

Step 1: Confirm the data is correctly loaded

Before loading the dataset to a dataframe, we need to verify and make sure if the dataset has consistent column names across all the files.

Importing the dataset can be achieved with the pandas packages in python, following is the Code to import a single file to dataframe-

```
import pandas as pd
covid=pd.read_csv("https://raw.githubusercontent.com/CSSEGISandData/COVID-19/master/csse_covid_19_data/csse_covid_19_daily_reports/01-01-2021.csv")
```

Step 2: Describe the data

Describe the data: The dataframe has 18 columns which gives information about the Province and country name, the number of Confirmed cases, death toll and Recovered cases. It also provides statistical information which are derived columns (calculated columns) like Active cases, Incident_Rate and Case_Fatality_Ratio which will be helpful to perform more EDA on the dataset.

Summary statistics: Summary statistics can be performed in python using dataframe.describe() function. The describe function will provide the measures to describe the dataframe like- count, mean, std, min, max and much more.

Breakdown of variables (numeric, ordinal, categorical)

Variable Type	Column Names
Categorical	Province_State, Country_Region, ISO3
Numerical	Lat, Long, Confirmed, Deaths, Recovered, Active, Incident_Rate, Total_Test_Results, People_Hospitalized,Case_Fatality_Ratio, UID, Testing_Rate, Hospitalization_Rate
Date	Last_Update
Ordinal	FIPS

Step 3: Check the validity of data

Define schema: With the help of df.dtype function from the pandas package, we can determine the schema of the dataframe. Most of the variables are objects, float, or integer types.

Variable Type	Column Names
Object	Province_State, Country_Region, ISO3
Integer	Confirmed, Deaths, Recovered, Active, FIPS, Total_Test_Results, People_Hospitalized, UID
Float	Last_Update, Lat, Long, Incident_Rate, Case_Fatality_Ratio, Testing_Rate, Hospitalization_Rate
Date	Last_Update

Understand the data: The primary information from the dataset is the daily cases of COVID-19 from US and different countries of the world. Which provides statistics about the number of confirmed, death and recovered based on the country and province.

Step 4: Answer the following questions:

1. Does the data include missing, incomplete, or invalid records?
Yes, our dataset has missing and incomplete records.

2. Does your data include outliers?
Yes, our dataset has outliers.

3. Is the data segmented into groups?
Yes, the dataset is divided in groups based on the country, province, and timeframe.

4. Is the data imbalanced (a large number of the records represent a majority class and very few records represent the minority class)?
No, the dataset is not largely concentrated or representing a majority class.

5. Are some data elements highly correlated with each other?
Yes, the dataset has correlation with each other, as few columns are dependent and helps us to determine the trend of another column and also derive a calculated column.

6. How was the data collected?
The data is collected from multiple sources such as World Health Organization (WHO), European Centre for Disease Prevention and Control (ECDC) and different US data sources at the state or county/city level. The data was loaded using pandas library into a dataframe from github.

7. What are the inclusion criteria for your data?
Inclusion criteria are defined on the key target feature which includes analysis on demographic, and geographic characteristics.

8. Can you generate preliminary visualizations for individual features?
Using the dataset we can induce visualization to check which province has the highest number of deaths or highest number of confirmed cases. By the visualization created we can have more data insights about the dataset.

Step 5: Use visualization to understand and explore, but not to explain
By using Matplotlib package we can produce visualization to understand the behavior of the dataset.
The following plot is a scatter plot displaying the relation between the number of confirmed and deaths. As we can see the number of confirmed and deaths are directly proportional.

```
plt.scatter(covid.Confirmed, covid.Deaths)
```

```
<matplotlib.collections.PathCollection at 0x2472b6d1bb0>
```