**Problem Statement: To find the U.S. state with highest and lowest no. of deaths reported**.

**Coronavirus** is a family of viruses that can cause illness, which can vary from *common cold* and *cough* to sometimes more severe disease. **Middle East Respiratory Syndrome (MERS-CoV)** and **Severe Acute Respiratory Syndrome (SARS-CoV)** were such severe cases with the world already has faced. **SARS-CoV-2 (n-coronavirus)** is the new virus of the coronavirus family, which first *discovered* in 2019, which has not been identified in humans before. It is a *contiguous* virus which started from **Wuhan** in **December 2019**. Which later declared as **Pandemic** by **WHO** due to high rate spreads throughout the world. Currently (on the date 20 May 2020), this leads to a total of *300K+ Deaths* across the globe, including *90K+ deaths* alone in the USA.

**Data Analysis:**

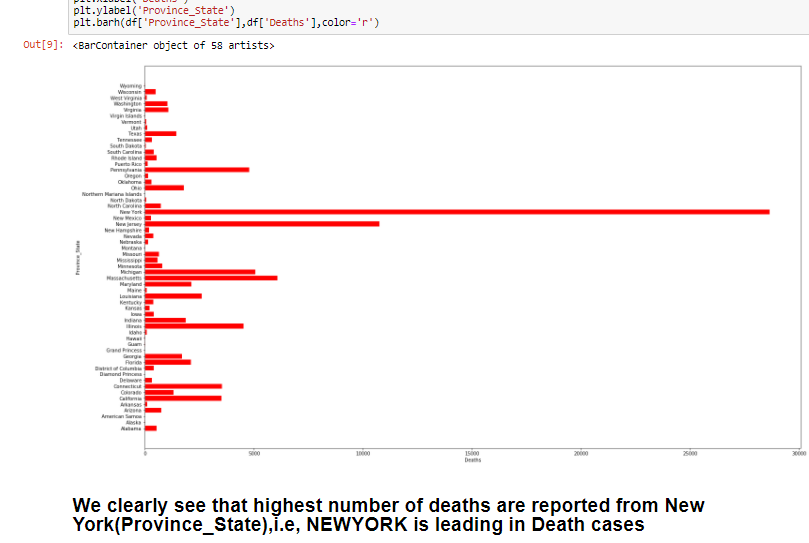
This Dataset contains an aggregation of each USA State-level data.

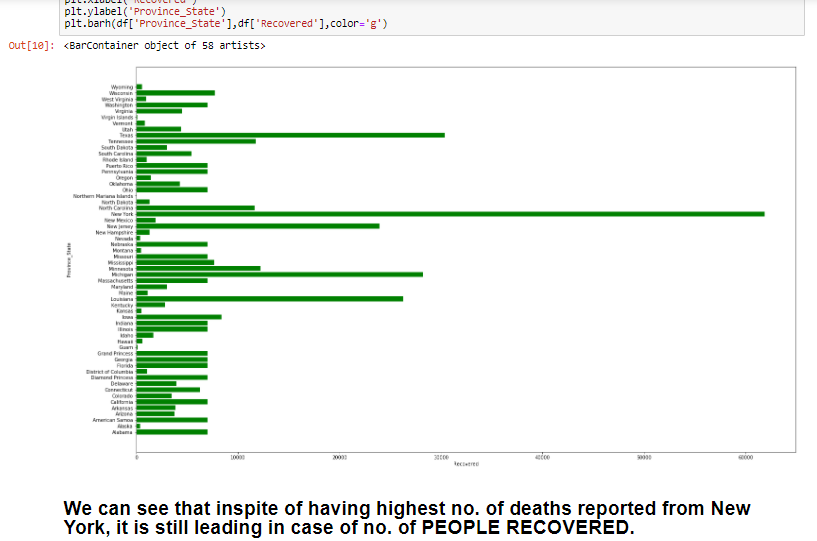
**Field description:-**

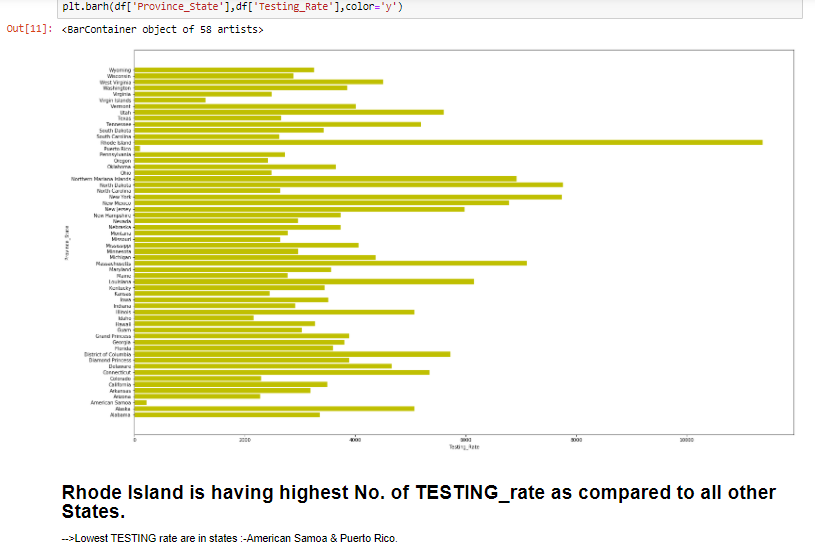
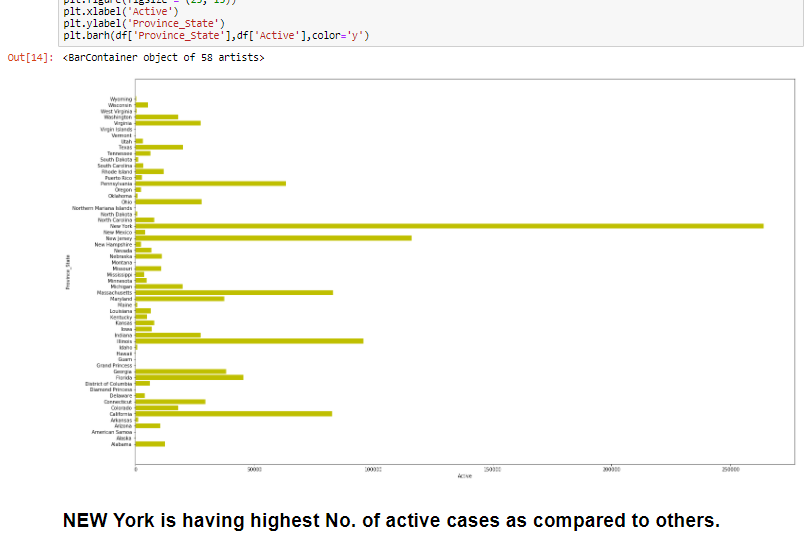
* **Province\_State** - The name of the State within the USA.
* **Country\_Region** - The name of the Country (US).
* **Last\_Update** - The most recent date the file was pushed.
* **Lat** - Latitude.
* **Long\_** - Longitude.
* **Confirmed** - Aggregated confirmed case count for the state.
* **Deaths** - Aggregated Death case count for the state.
* **Recovered** - Aggregated Recovered case count for the state.
* **Active** - Aggregated confirmed cases that have not been resolved (Active = Confirmed - Recovered - Deaths).
* **FIPS** - Federal Information Processing Standards code that uniquely identifies counties within the USA.
* **Incident\_Rate** - confirmed cases per 100,000 persons.
* **People\_Tested** - Total number of people who have been tested.
* **People\_Hospitalized** - Total number of people hospitalized.
* **Mortality\_Rate** - Number recorded deaths \* 100/ Number confirmed cases.
* **UID** - Unique Identifier for each row entry.
* **ISO3** - Officialy assigned country code identifiers.
* **Testing\_Rate** - Total number of people tested per 100,000 persons.
* **Hospitalization\_Rate** - Total number of people hospitalized \* 100/ Number of confirmed cases.

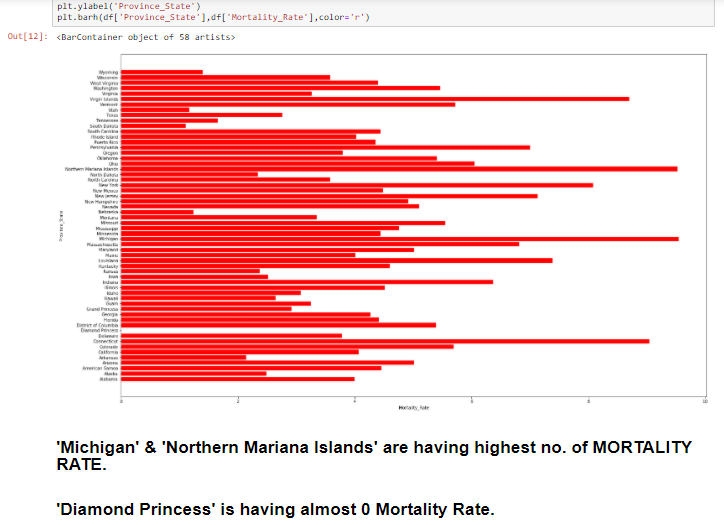
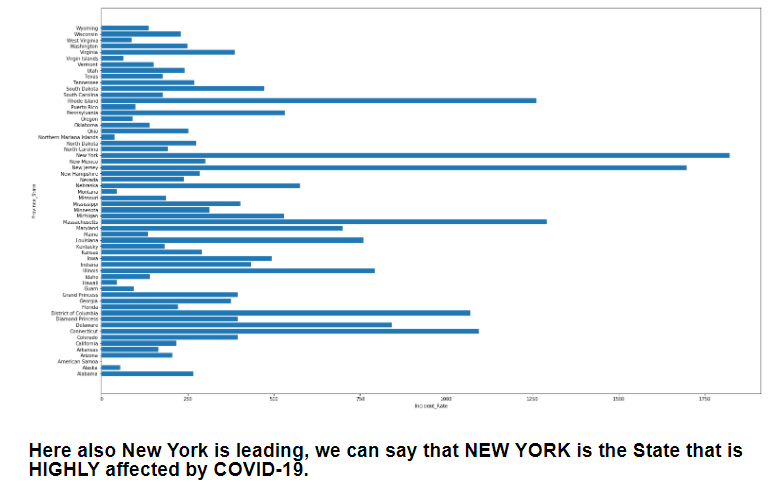
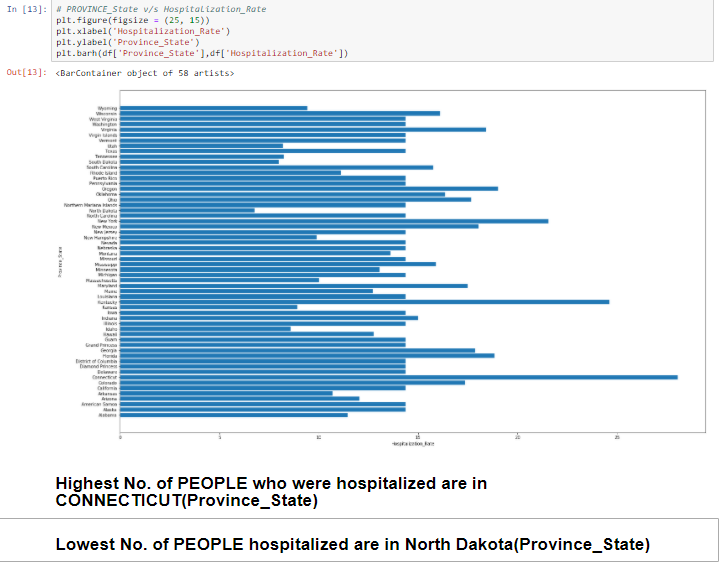
**EDA Concluding Remark:**

* After removing all the Null/Missing values, we will check which Features are relevant for our analysis.
* We will drop the columns that are not necessary for our analysis.
* Now we will check different Features v/s States.





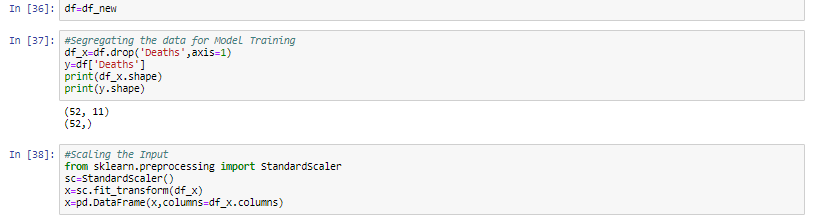




* 'Confirmed' is having highest Std.Deviation among all other features.
* Average 'Deaths' cases are 1161.
* Average 'Recovered' cases are 7007.
* Average No. of 'Active' cases are 20,070.

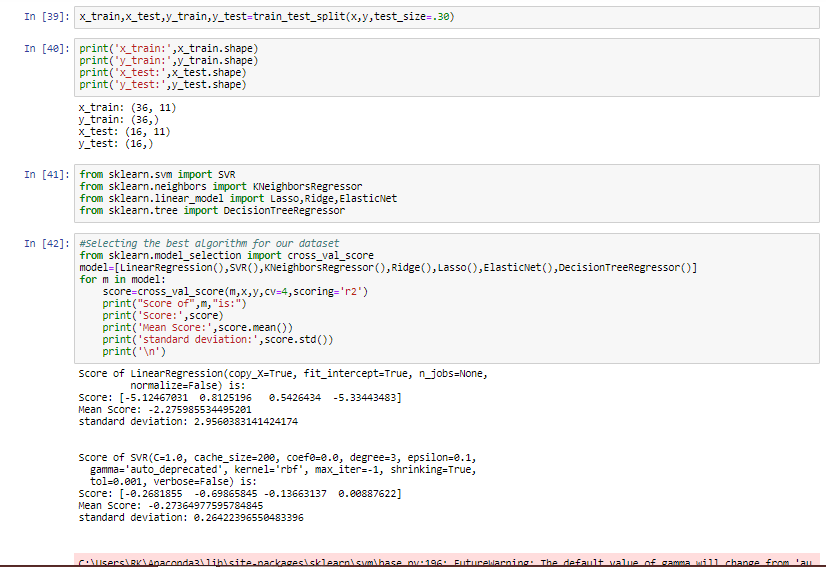
**Pre-Processing Pipeline:**

Scaling the Input data i.e, X.



**Building Machine Learning Models:**

* Now we will divide our dataset into train and test.
* After segregating the dataset, we will find the best model for the dataset.

****

* Once we have got the best algorithm for our analysis, we will find the best parameters for the same.



**CONCLUSION:**

* **The State which is most affected by COVID-19 is NEW YORK.**
* **New York is leading in no. of DEATHS cases inspite of having high recovery rate.**
* **Highest No. of People were hospitalized in ‘Connecticut’.**