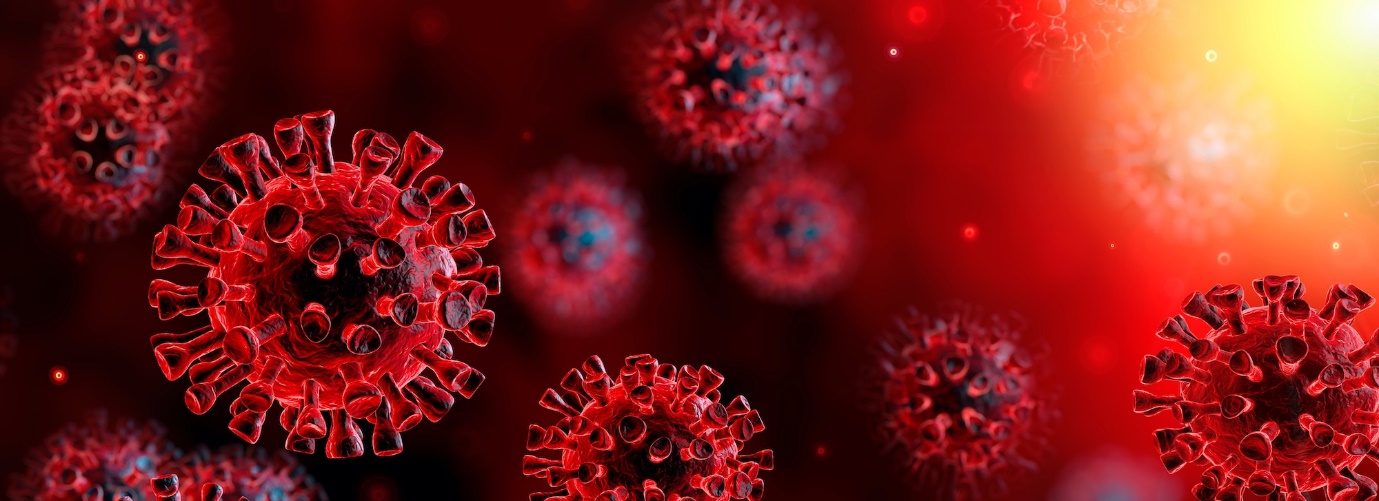
2019 Novel Coronavirus

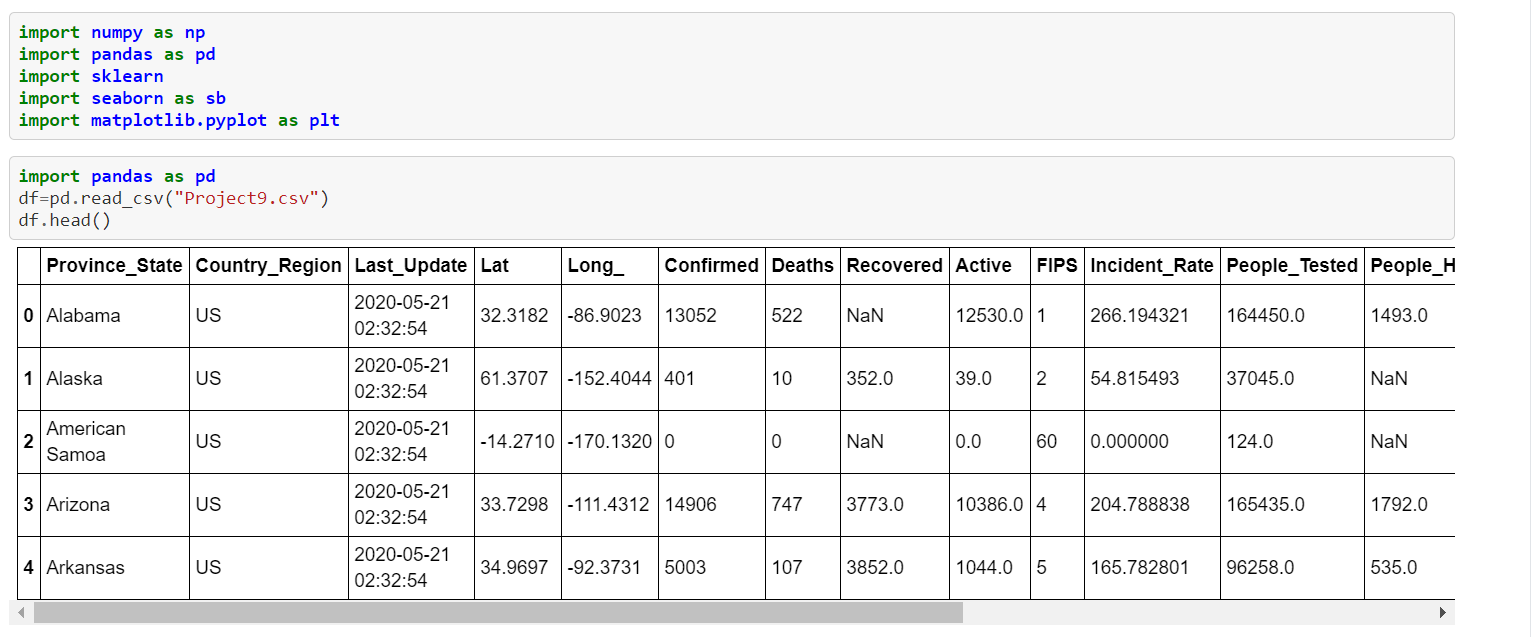
*  By Johns Hopkins University Center for Systems Science and Engineering (JHU CSSE)

**About Coronavirus:**

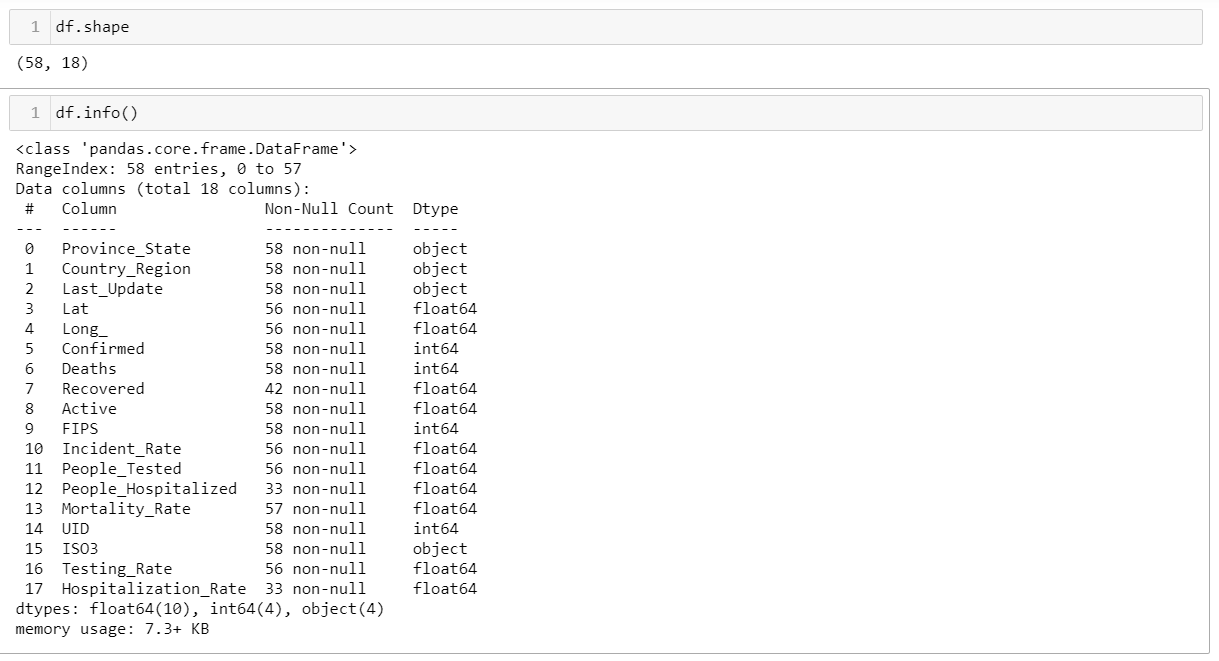
Coronavirus is a family of viruses that can cause illness, which can vary from common cold and cough to sometimes more severe disease. Middle East Respiratory Syndrome (MERS-CoV) and Severe Acute Respiratory Syndrome (SARS-CoV) were such severe cases with the world already has faced. SARS-CoV-2 (n-coronavirus) 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 USA.

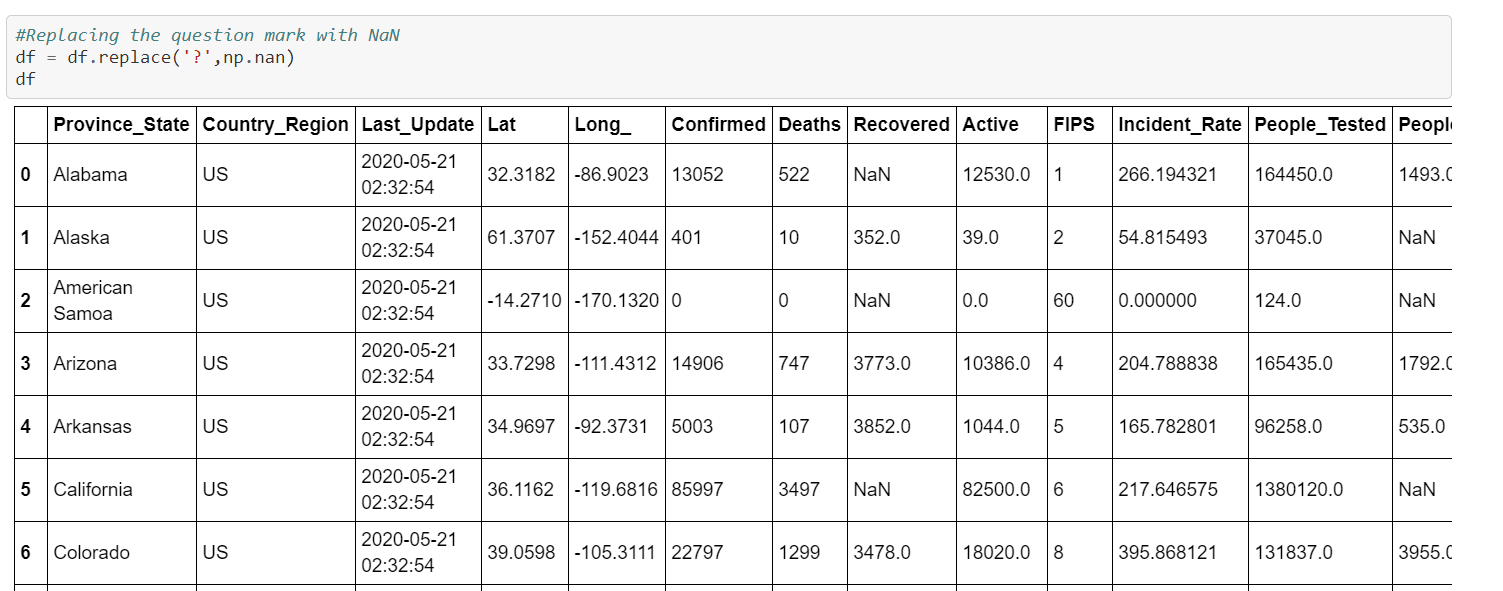
**Feature In Dataset:**

* 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 case count for the state.
* Deaths - Aggregated death toll for the state.
* Recovered - Aggregated Recovered case count for the state.
* Active - Aggregated confirmed cases that have not been resolved (Active cases = total cases - total recovered - total deaths).
* FIPS - Federal Information Processing Standards code that uniquely identifies counties within the USA.
* Incident\_Rate - cases per 100,000 persons.
* Total\_Test\_Results - Total number of people who have been tested.
* People\_Hospitalized - Total number of people hospitalized. (Nullified on Aug 31, see Issue #3083)
* Case\_Fatality\_Ratio - Number recorded deaths \* 100/ Number confirmed cases.
* UID - Unique Identifier for each row entry.
* ISO3 - Officialy assigned country code identifiers.
* Testing\_Rate - Total test results per 100,000 persons. The "total test results" are equal to "Total test results (Positive + Negative)" from COVID Tracking Project.
* Hospitalization\_Rate - US Hospitalization Rate (%): = Total number hospitalized / Number cases. The "Total number hospitalized" is the "Hospitalized – Cumulative" count from COVID Tracking Project. The "hospitalization rate" and "Total number hospitalized" is only presented for those states which provide cumulative hospital data. (Nullified on Aug 31, see Issue #3083)

**Let’s Import the Libraries / Dataset:**

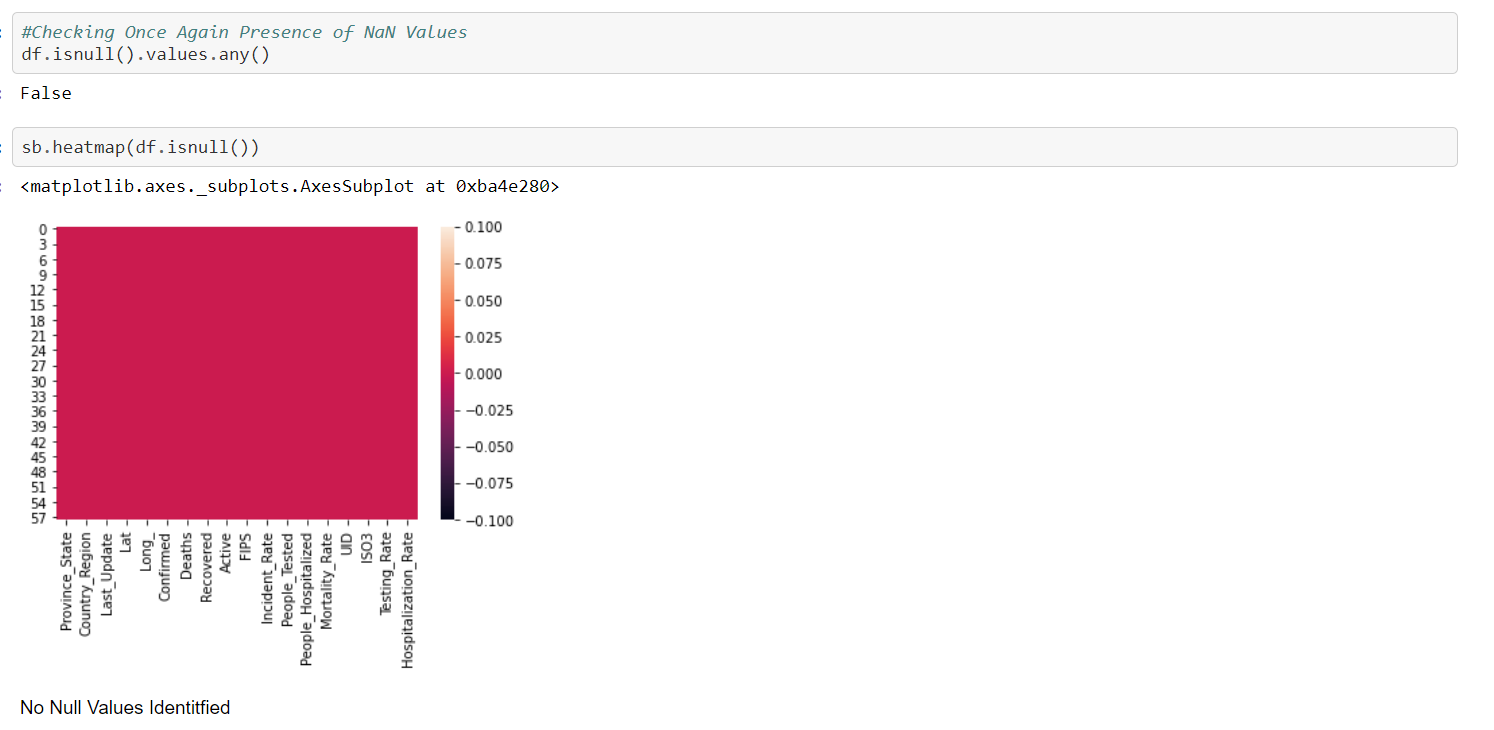
**Checking the dimensions of the dataset:**



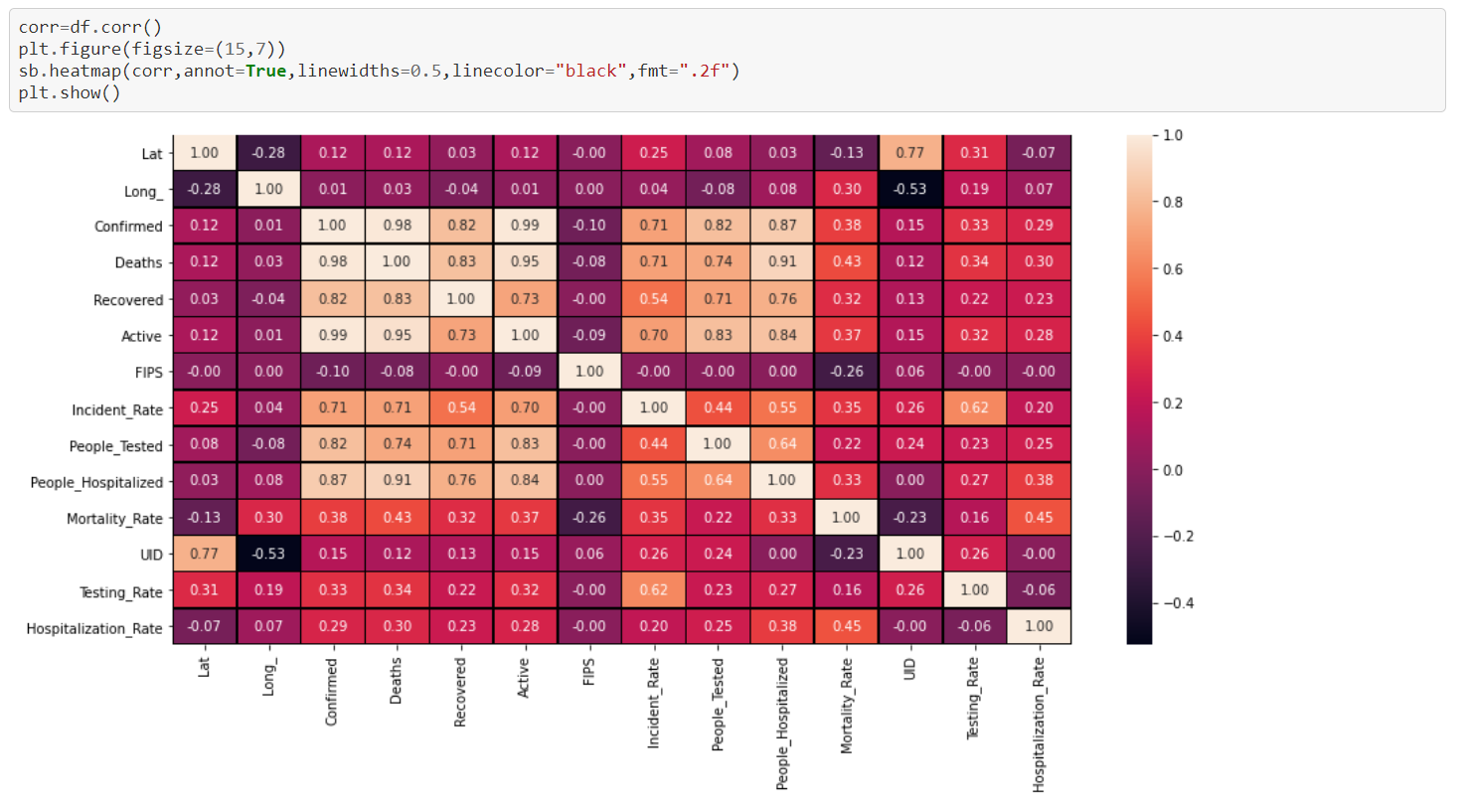
We had found out data with missing value and we replace with NaN:

**Checking the NaN Value and Treating them:**



We had replaced all the NaN values with Mean of respective attributes.

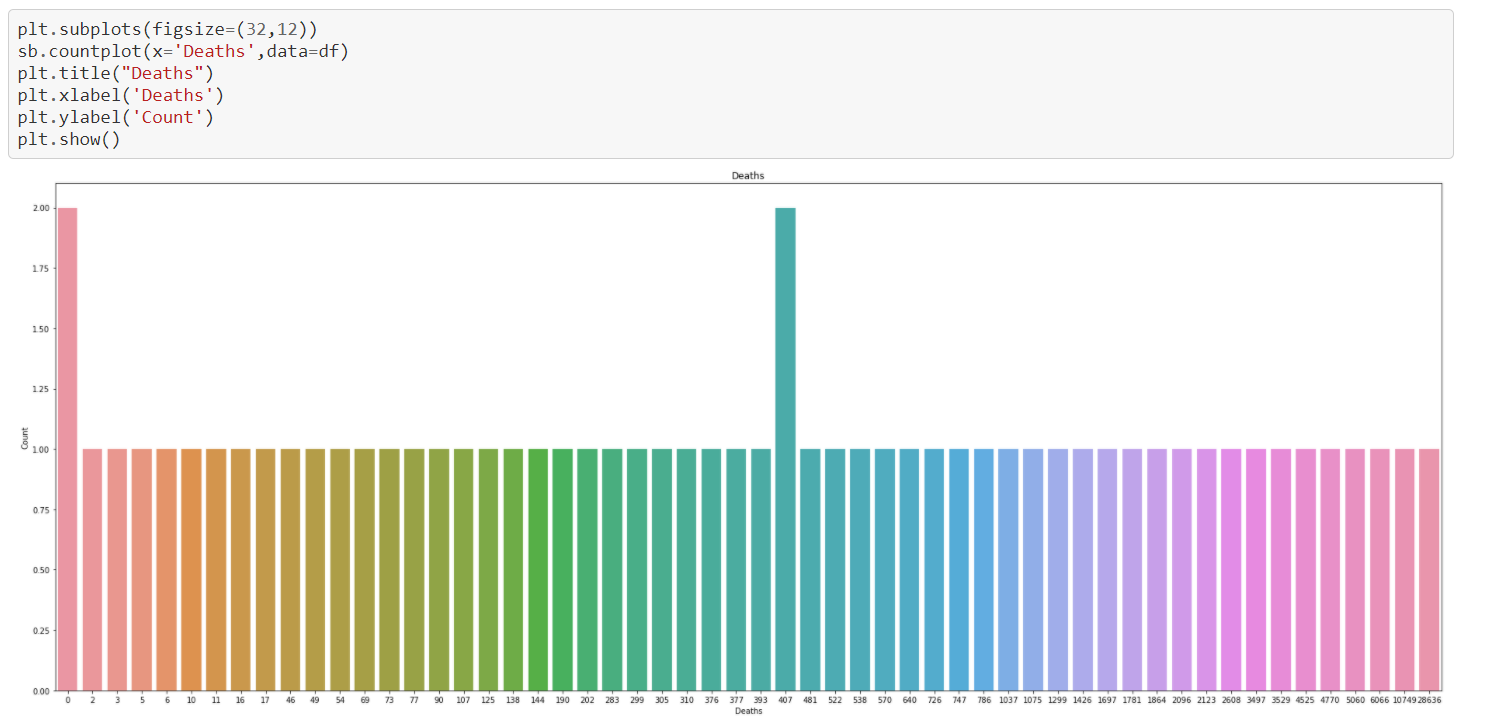
Correlation:



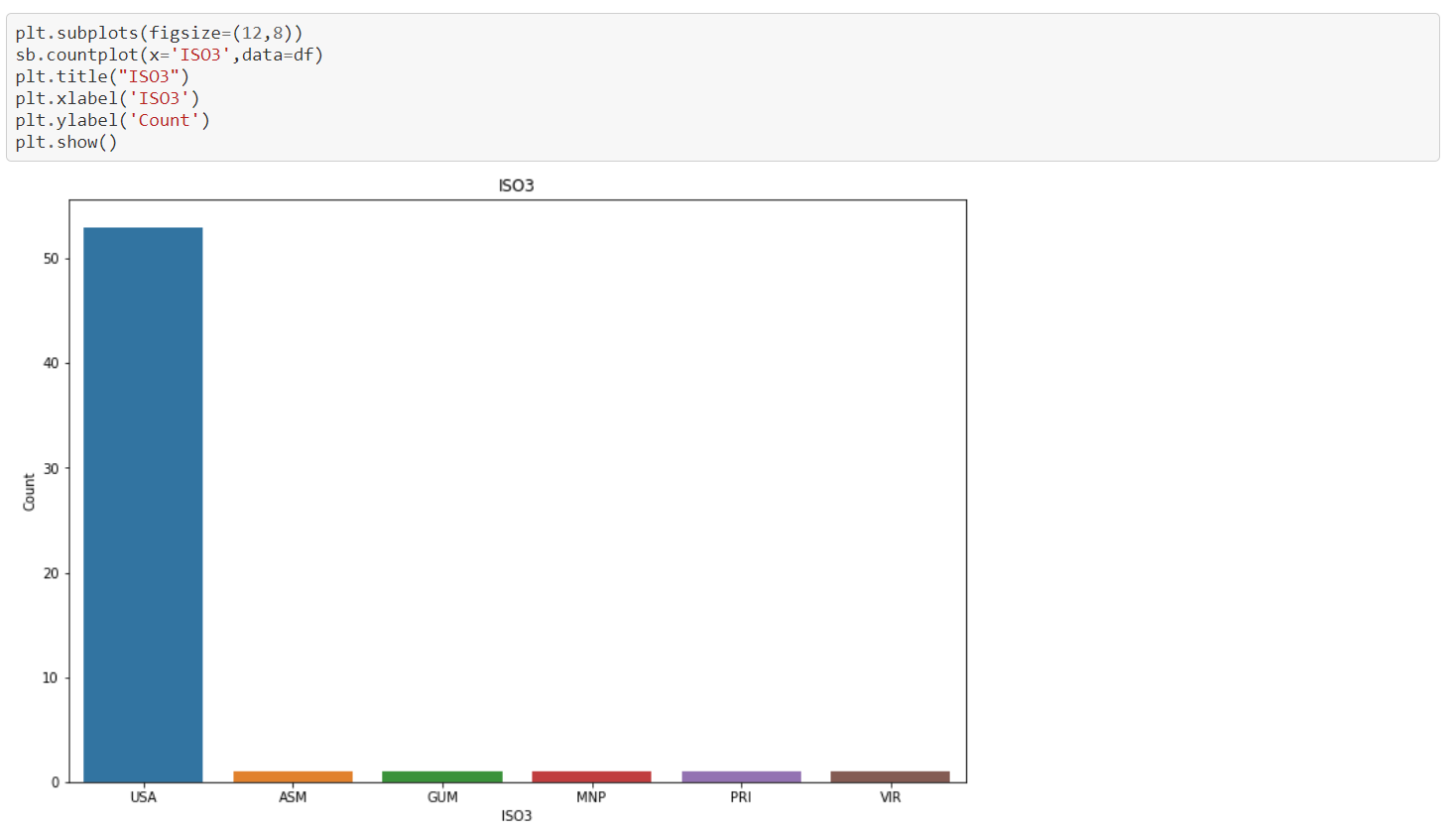
Except test rate, UID, FIPS, Lat other attributes have some positive correlation with Hospitalisation rate.

**Univariant Analysis:**

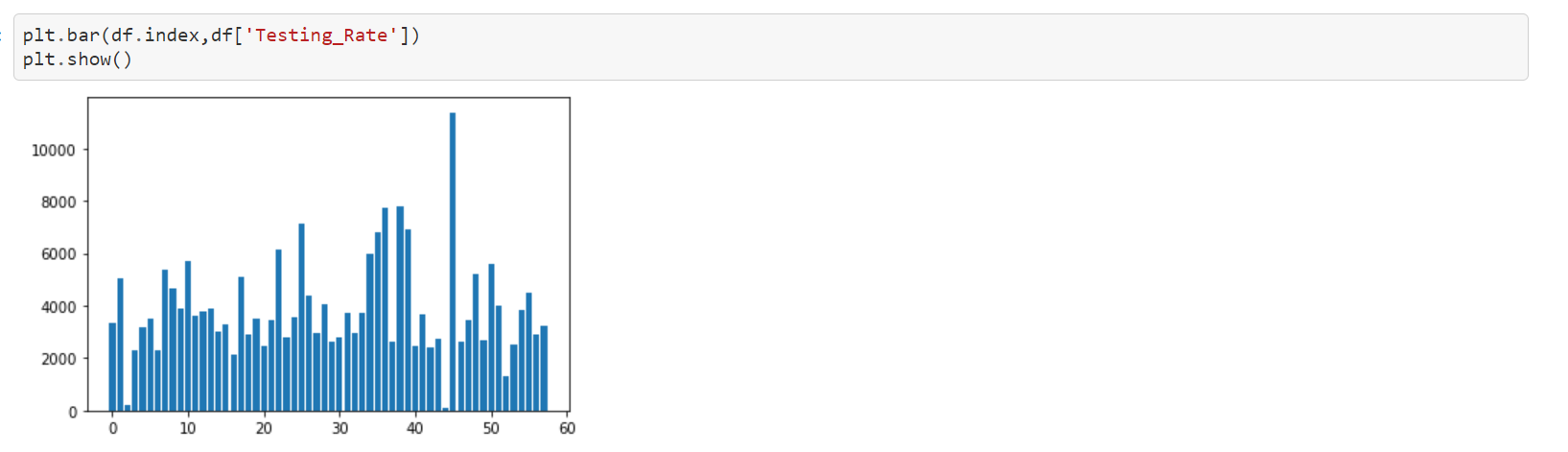
Ploting the death frequency for Covid 19 Dataset



Country Codes having highest frequency:

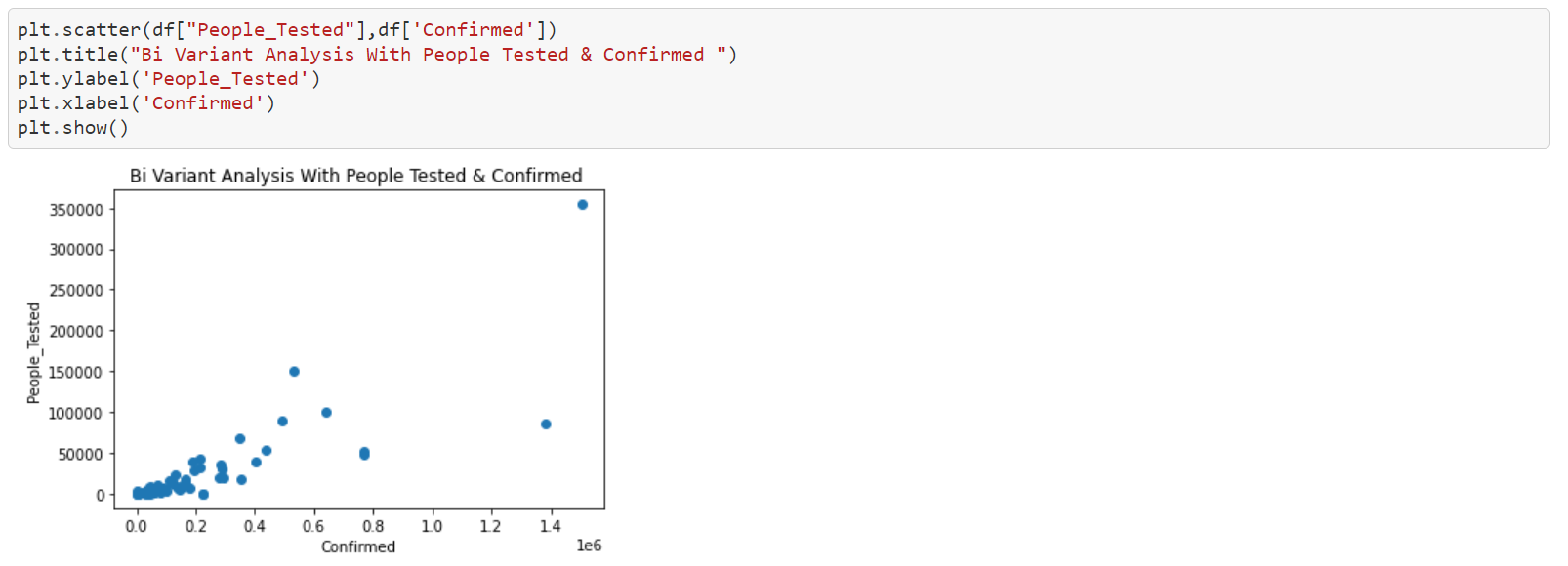
USA is having the highest country count in Covid-19 cases.

Plotting the Testing Rate:

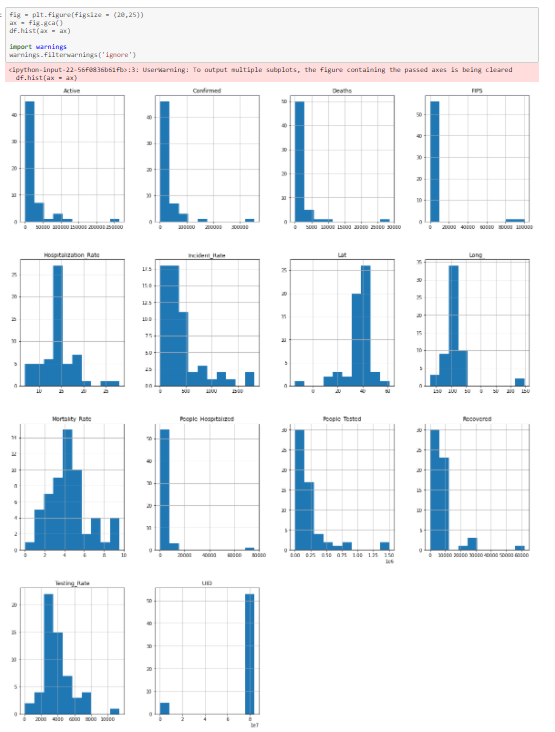


40 -50 Bins having the highest frequency of Testing Rate.

**Bi Variant Analysis:**

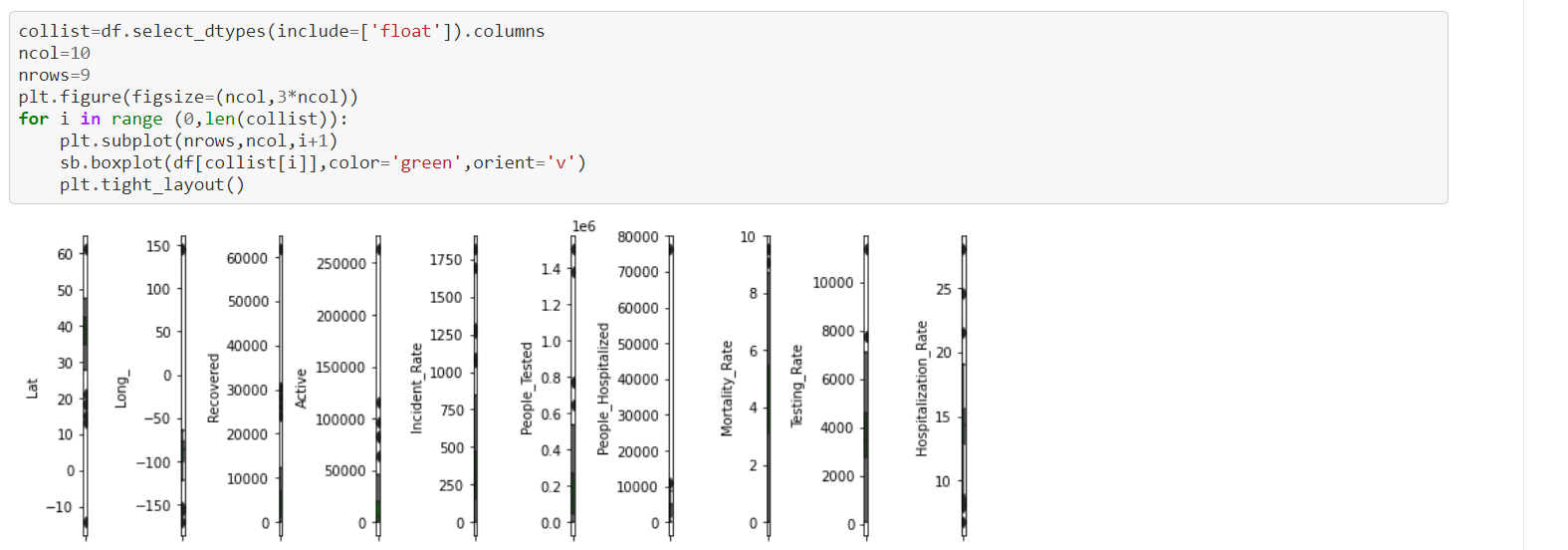


People tested with between 0 – 50000 having dense confirm cases.



Graphical Representation of Attributes of in histogram.

**Checking the outliers:**



We can find outlier Lat, Long, Recovered, Active, People Tested, Mortility Rate , Hospatility Rate.

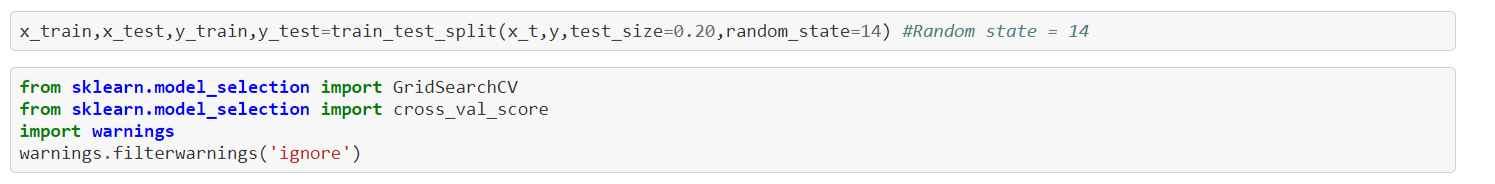
Checking Skewness:

Using Standard Scaler





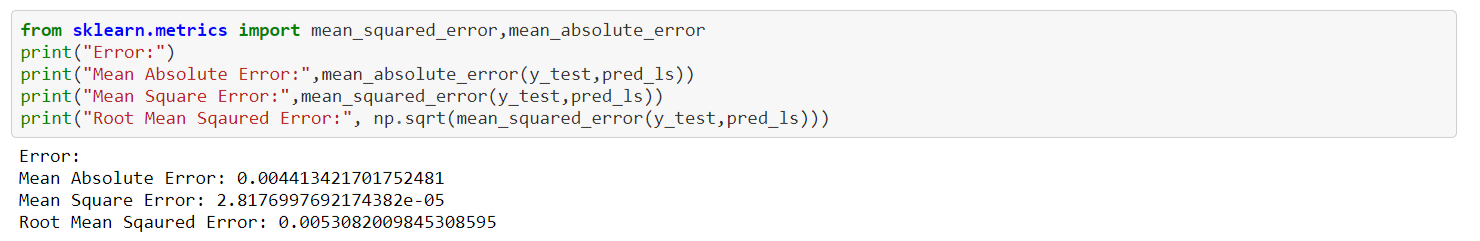


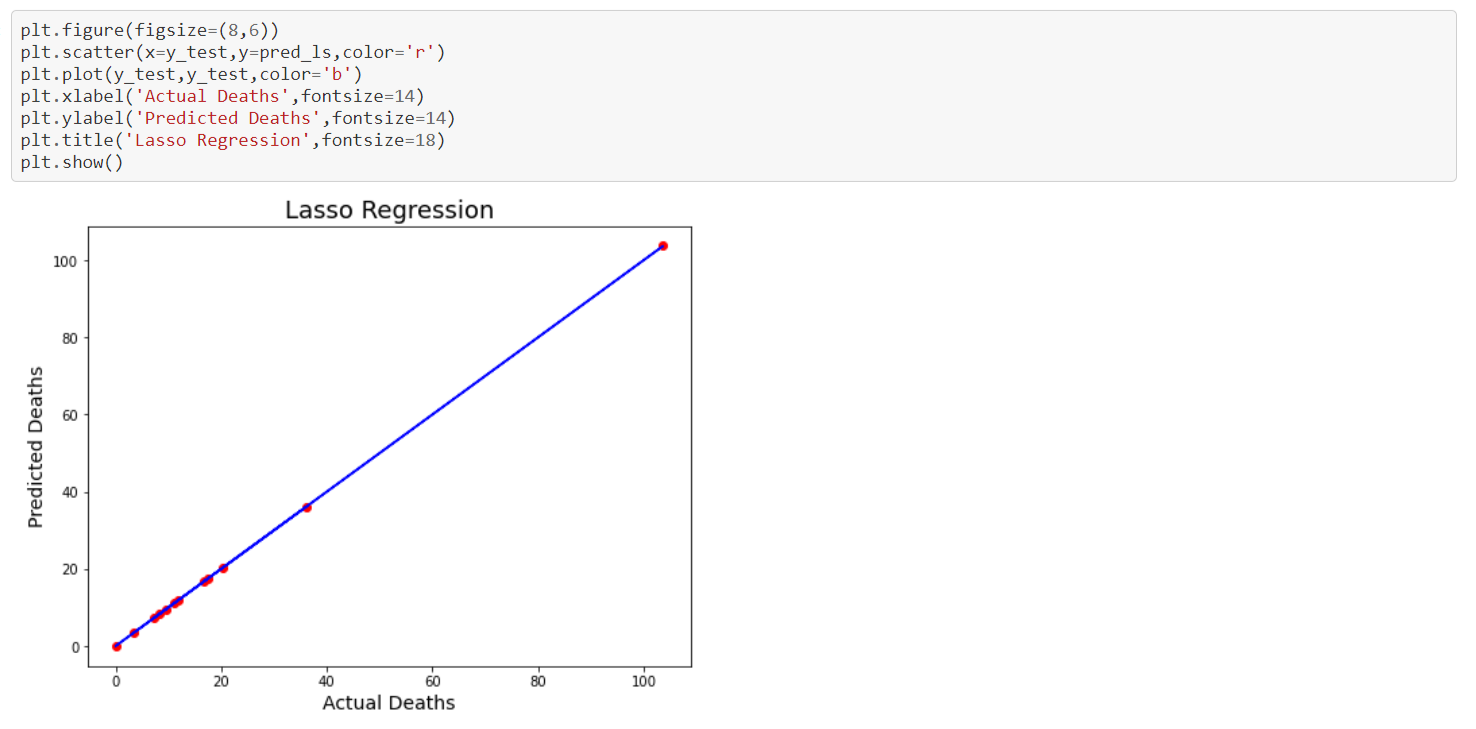


Using Regularisation:

Lasso Regression:



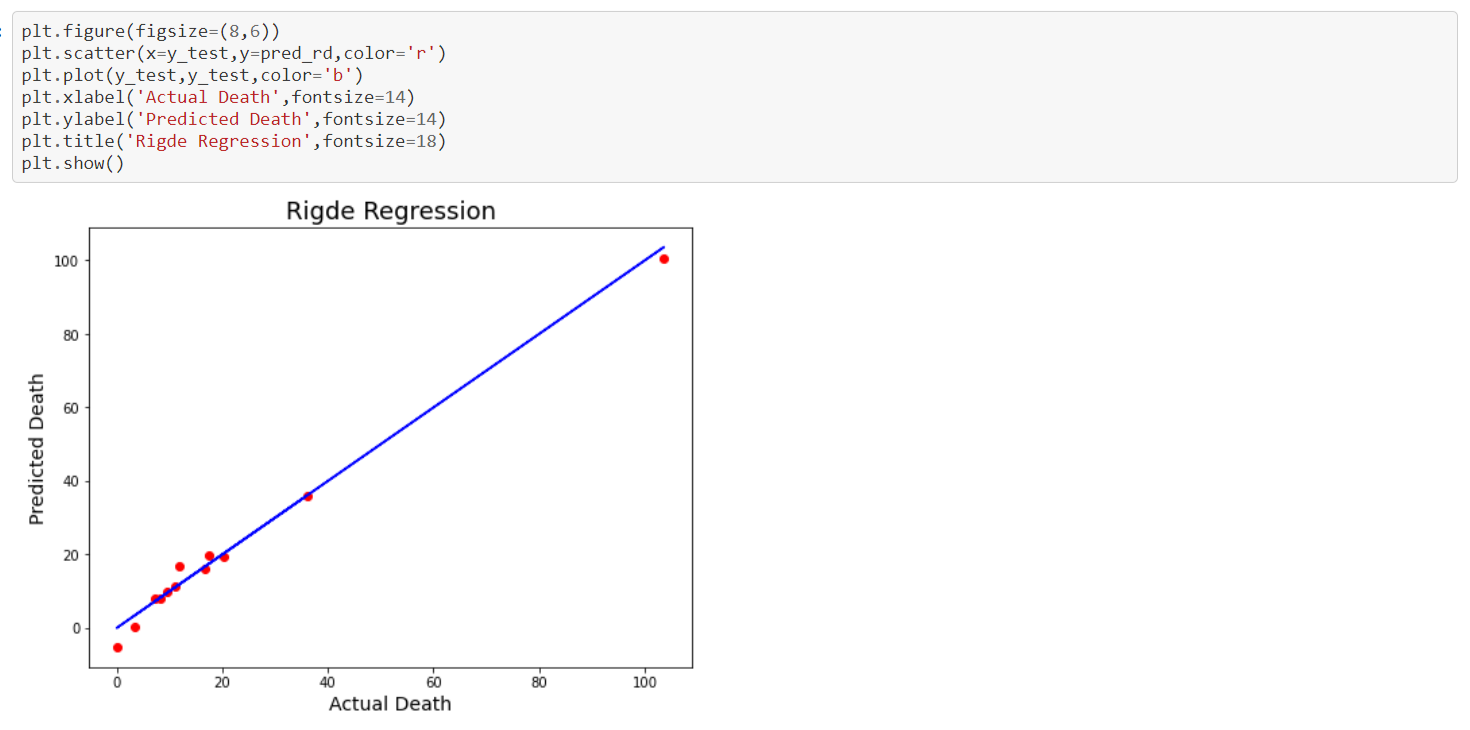




Datapoints are lying on the predicted lines.

Ridge Regression:

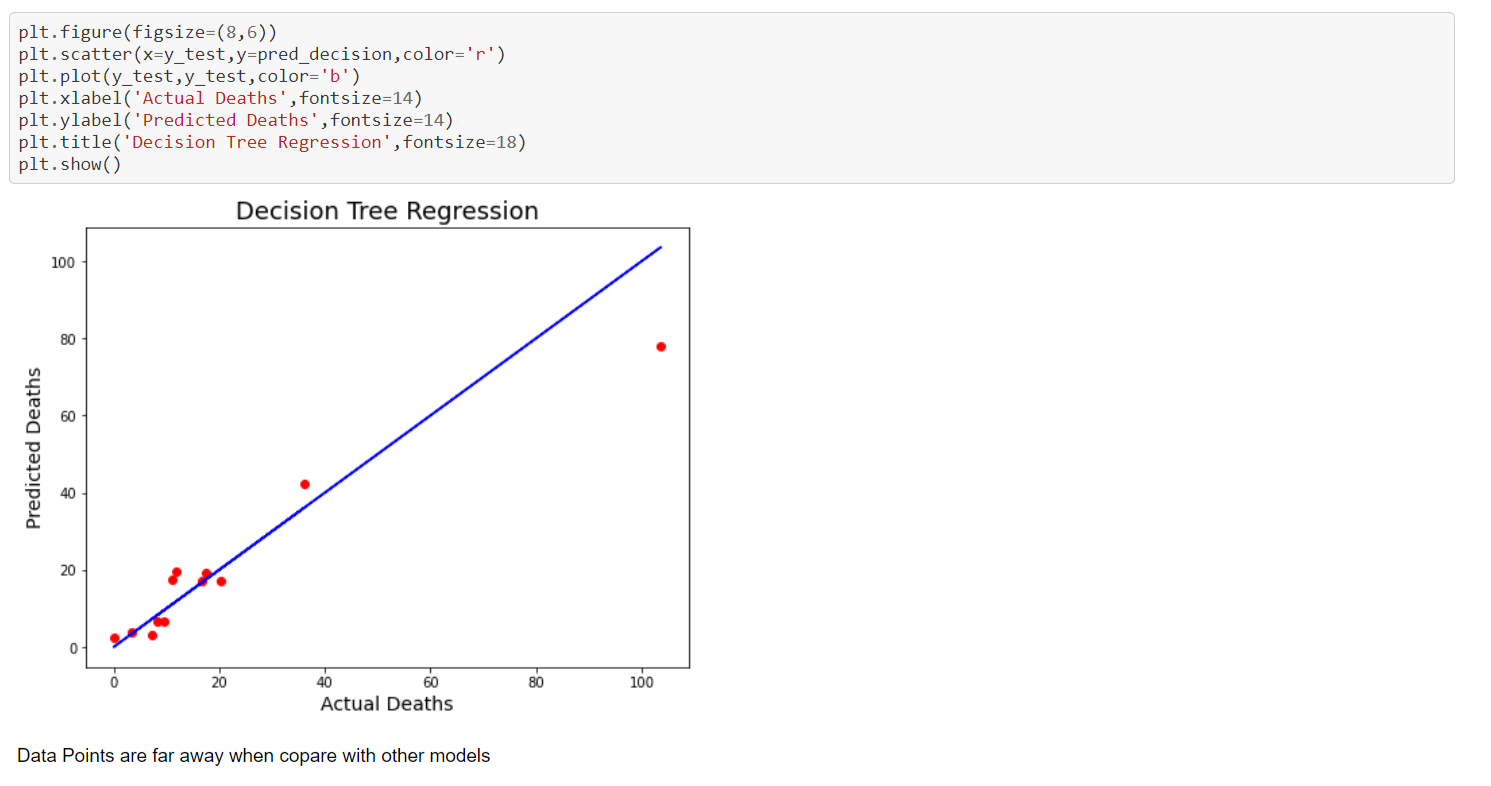




Few Data Points are lying outside the predicted point, Accuracy is only 97%

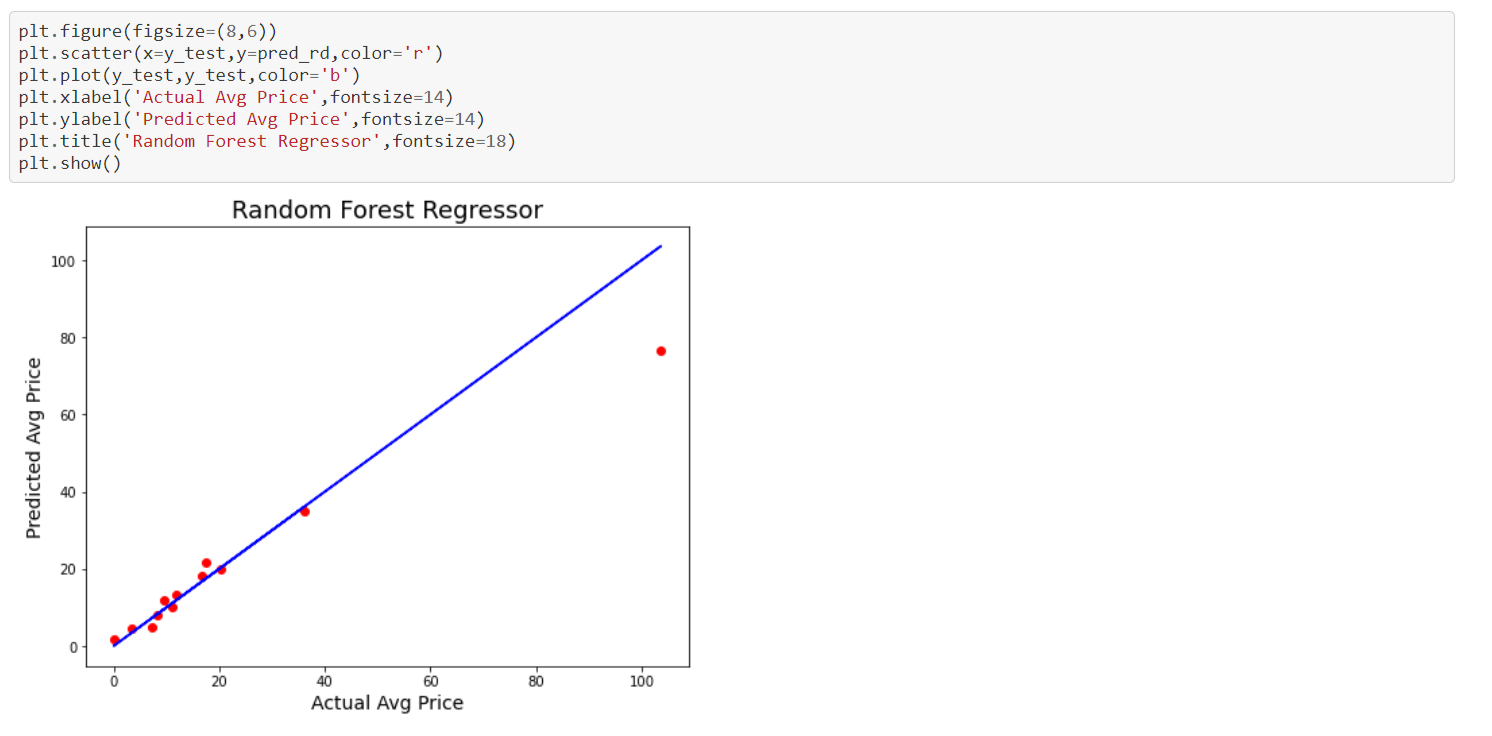
Ensemblier Technique:





Random Forest Regressor:

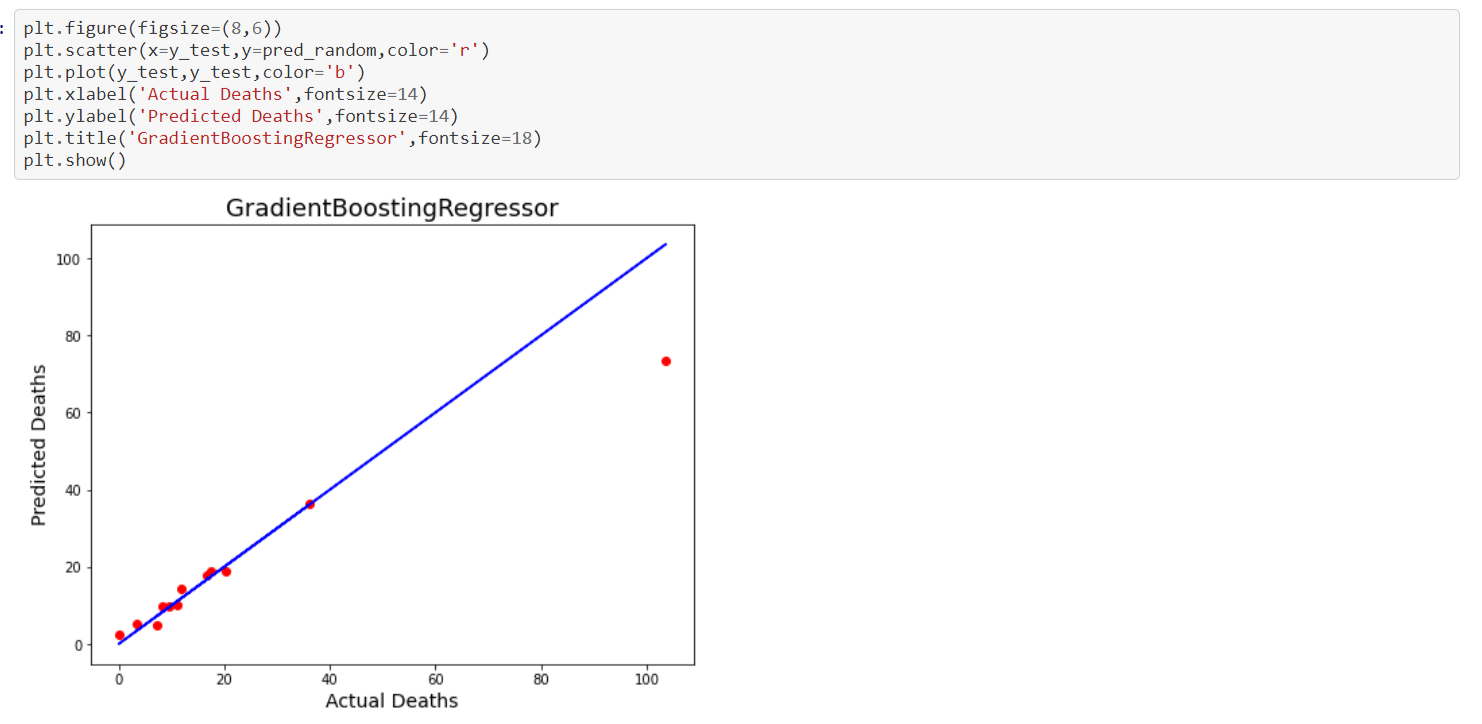




Datapoints are not inline with the predicted Line.

Gradient Boosting Regressor:

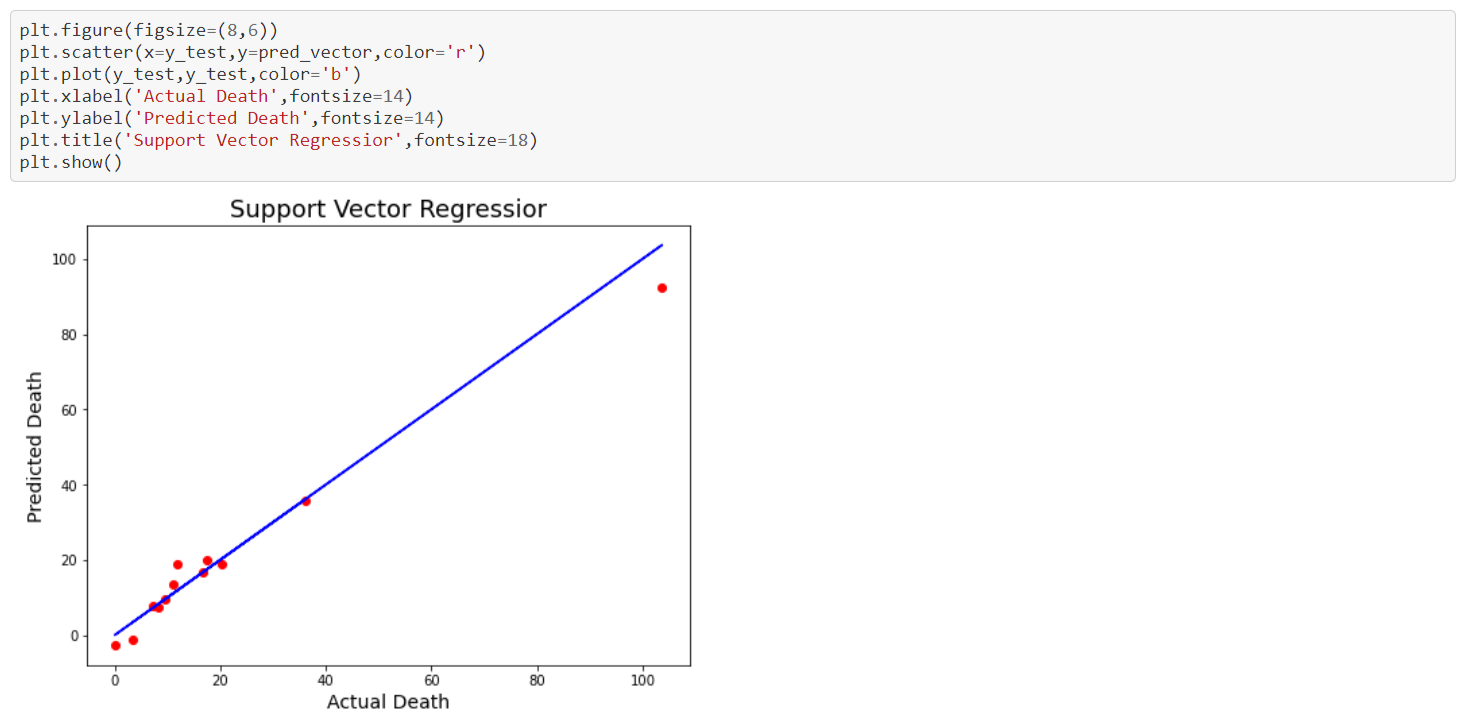




Few Datapoints are far from dataline, Accuarcy is 84%

Support Vector Regressor:





Few datapoint are in line with datapoints. 97% Accuracy.

Predictions:



Observations:

* Lasso at CV 7 is having 98.23 % accuracy when compare with other models.
* Next best model would be Support Vector Regression

Saving the best model:

Lasso is the best model 98.23%

