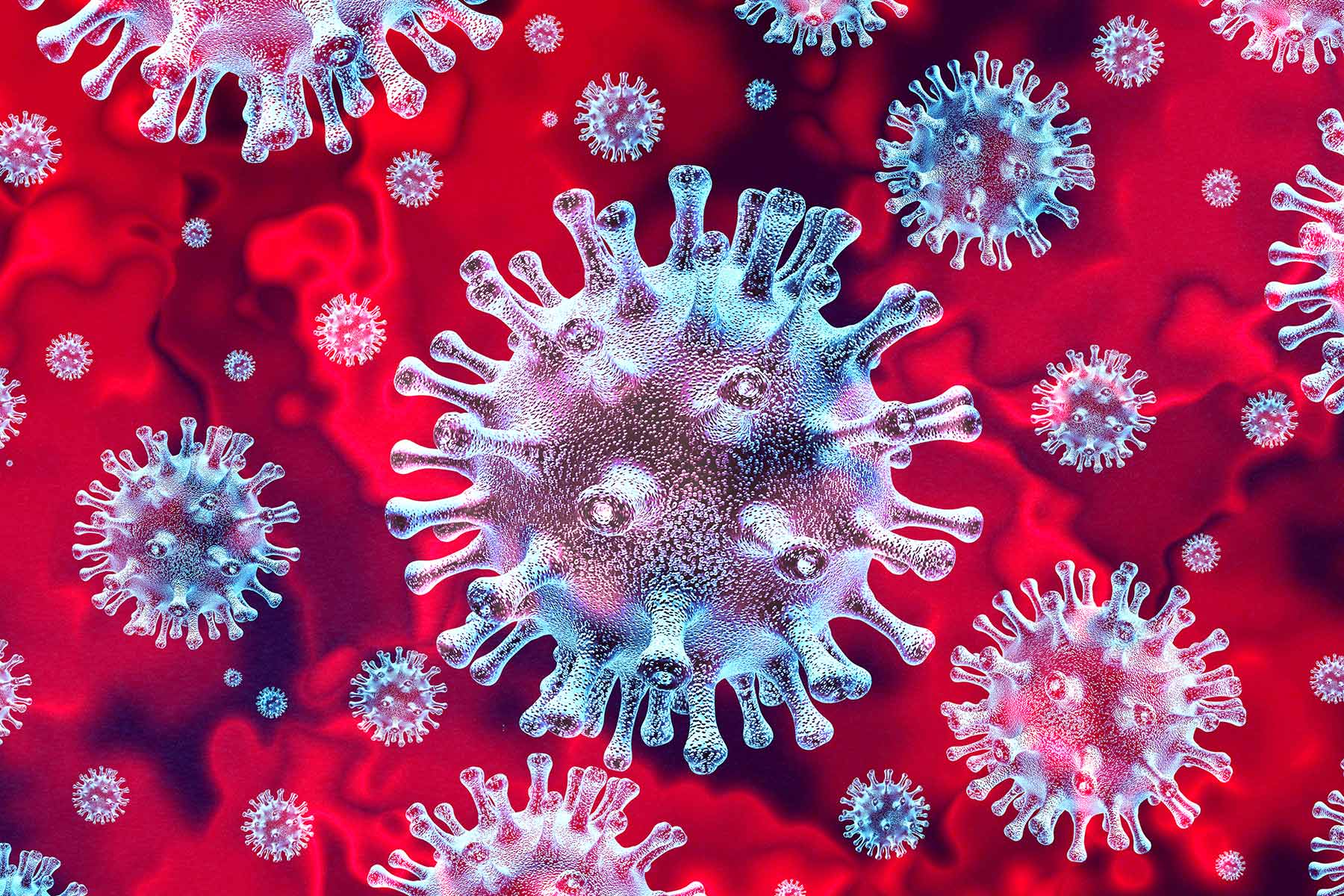
**Predicting the Deaths of Covid-19 Disease**

****

In this blog, I will explain the process of machine learning model, exploring data variables, Bivariate and univariate analysis on the covid-19 dataset, the recent pandemic declared by WHO. The dataset provided information of covid-19 cases of USA state. I will build machine learning model and predict deaths of covid-19 disease.

Covid-19 Dataset:

Below is the description of Covid-19 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. (According to map)

Long\_ - Longitude. (According to map)

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.

TYPE OF MACHINE LEARNING ALGORITHMS

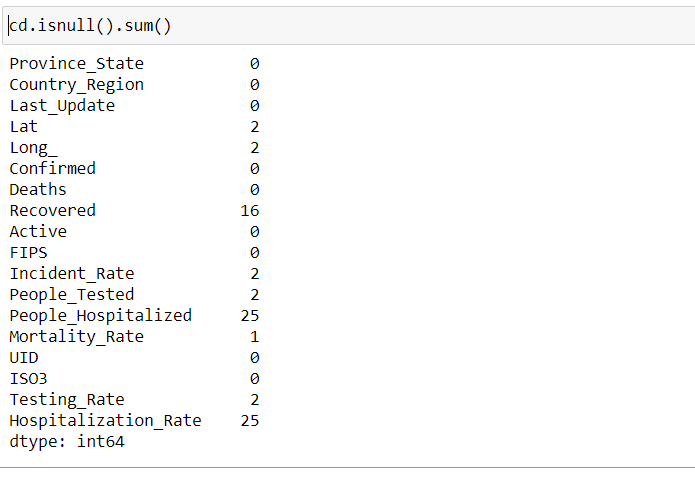
* The Target variable is deaths which is continuous value so I will use regression model.

**Importing important Libraries and Covid-19 dataset-**



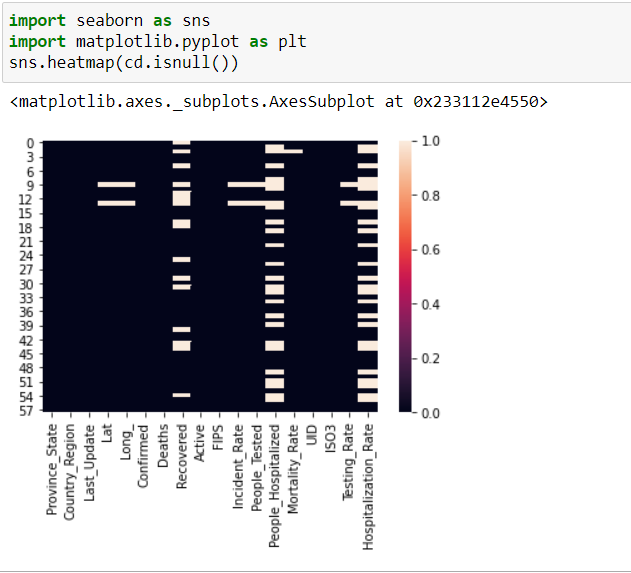
**Data exploring and analysis**

**Checking null values**

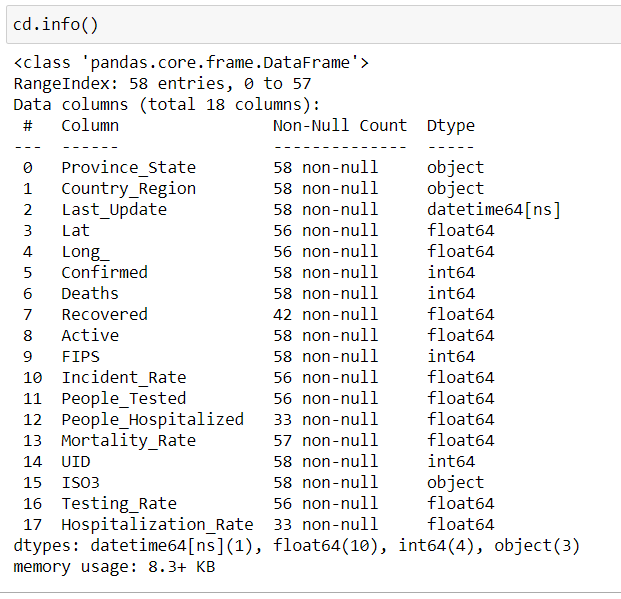


The above data shows missing/null values present in dataset.

Checking null values using heatmap for better understanding. And importing plot libraries.



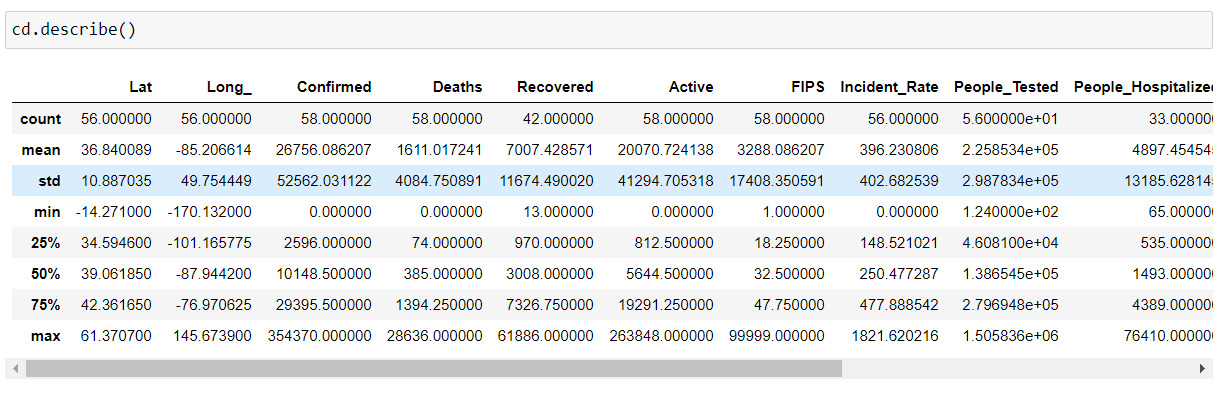
Checking data set info



The dataset have 58 rows and 18 columns.

Last\_Update column have datetime data type, Province\_state,Country\_region and ISO3 columns have object data type and all other columns have numeric data type. We need to convert object data type into numeric, so machine learning algorithms can process them.

**Summary Statistics**

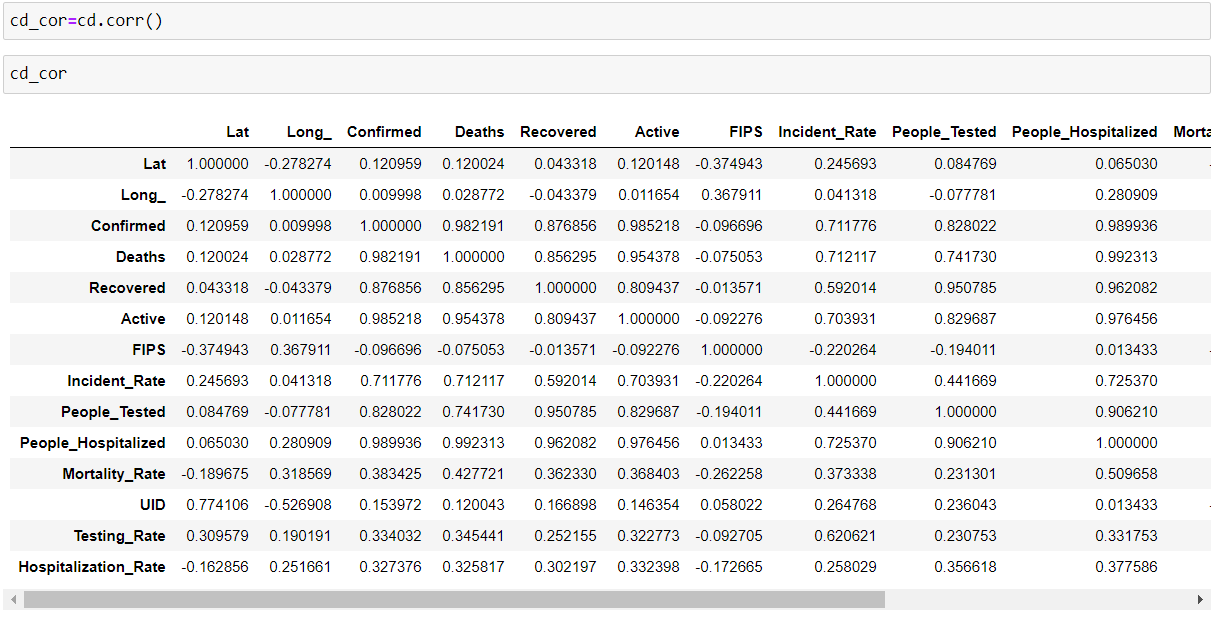


The above data shows minimum death is 0 and maximum is 28636.

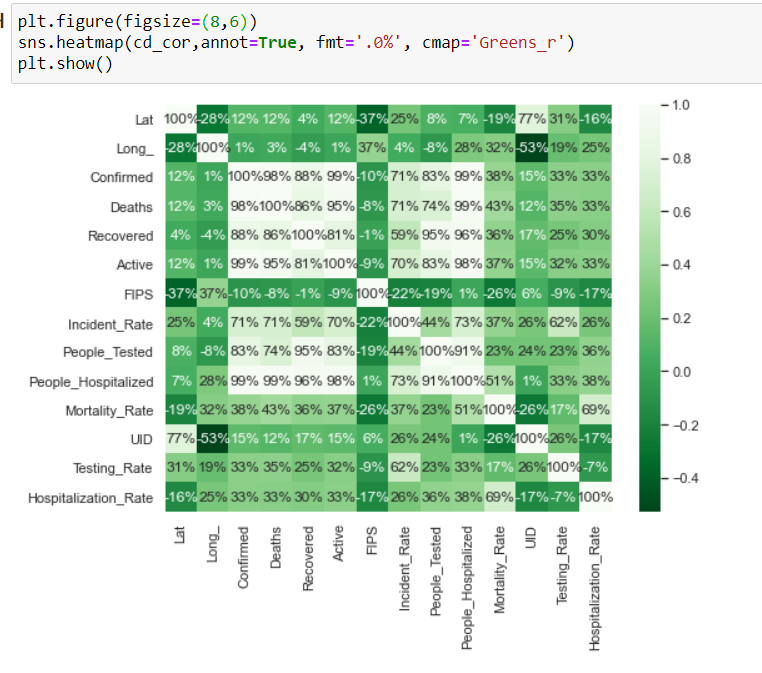
The minimum recovered is 13 and maximum is 61886

Unable to say outliers present in confirmed, deaths, recovered, active, incident\_rate, people\_hospitalized, testsing\_rate columns because in may month New York is epicentre of coronavirus disease in USA.

**Checking correlation between columns.**



Checking correlation using heatmap for better understanding.



Observation - Confirmed, recovered, active, people\_tested, incident\_rate, people\_hospitalized are highly correlated with target variable.

Now filling null values

# filling Lat, Long\_ columns NaN values according to map

# filling recovered columns NaN values by Confirmed-active-deaths.



# Filling Incident\_rate column NaN values, use infomration from Wikipedia

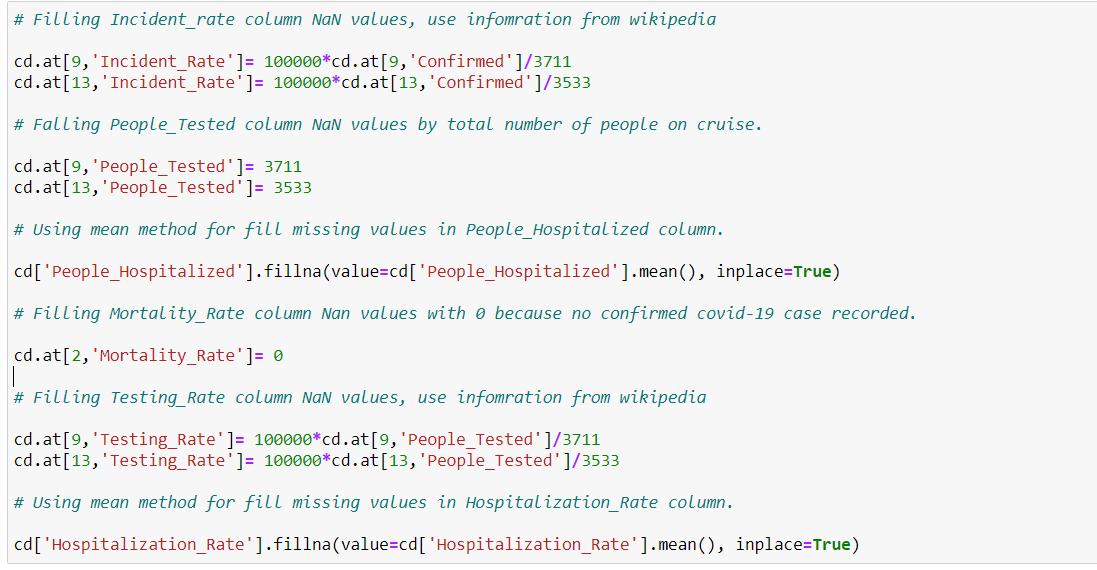
# Falling People\_Tested column NaN values by total number of people on cruise.

# Using mean method for fill missing values in People\_Hospitalized column.

# Filling Mortality\_Rate column Nan values with 0 because no confirmed covid-19 case recorded.

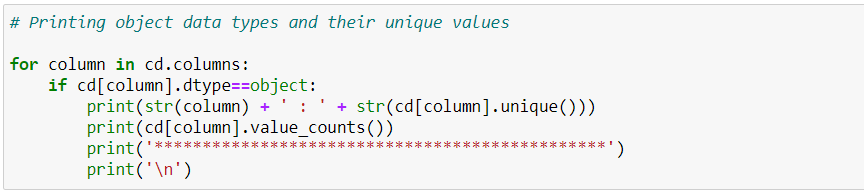
# Filling Testing\_Rate column NaN values, use infomration from Wikipedia

# Using mean method for fill missing values in Hospitalization\_Rate column.



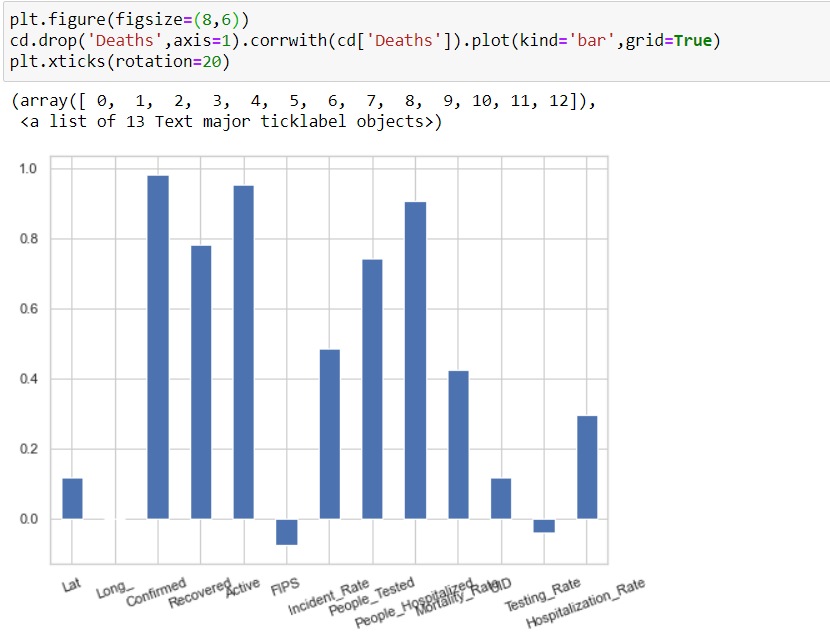
Note – Unable to figure out approx. number of people hospitalized using other variables also 25 null/missing values in this column so used mean method to fill data.

# Exploring Data Variables



There is 3 columns which has object data type. Will convert into numeric later on.

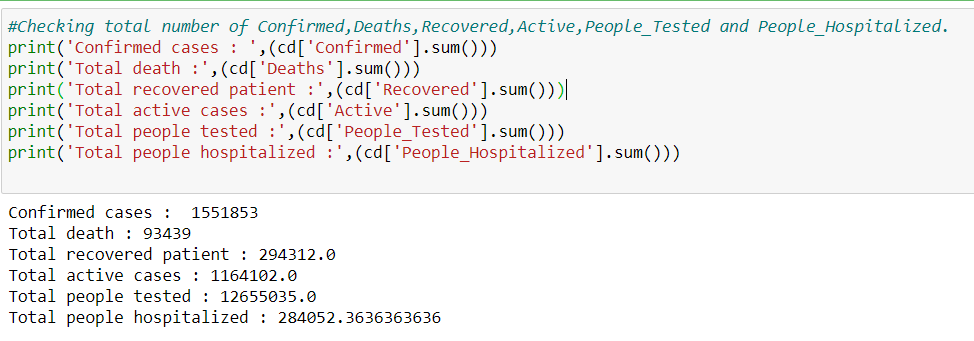
Checking correlation with target variable using bar plot.



Observation - Fips and testing\_rate are negatively correlated with target variable. confirmed,recovered,active and people\_hospitalized highly correlated with target variable.

# Univariate Anaylsis

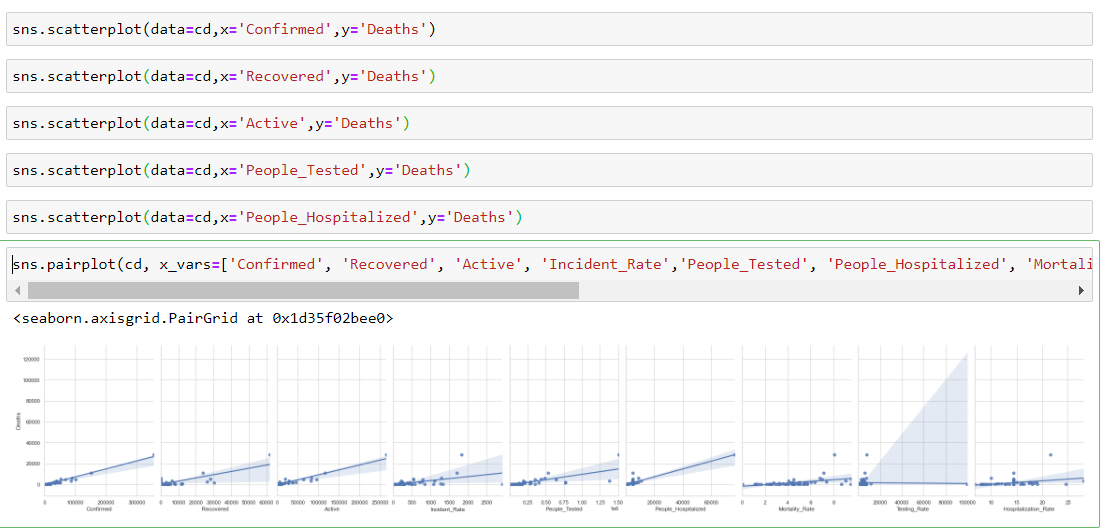
Checking total number of Confirmed, Deaths, Recovered, Active, People\_Tested and People\_Hospitalized.



Used violin plot to represent a variables distribution.

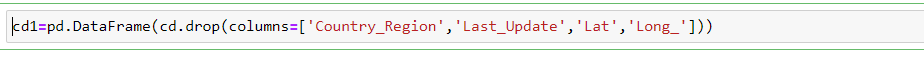


# Bivariate Analysis



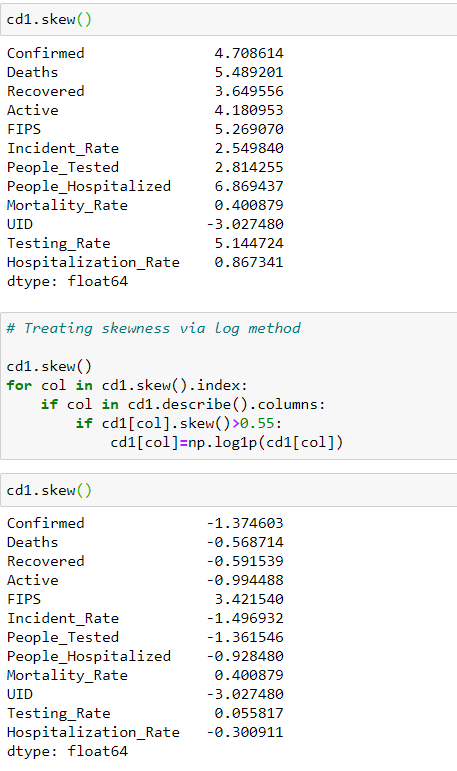
The above graph shows positive correlation with target variable.

Removing columns Country\_Region, Last\_Update, Lat, Long\_ columns as the columns does not contribute to a death probability.



# Checking skewness

Now checking skewness of dataset and treating skewness via log method.



Now skewness removed from the dataset.

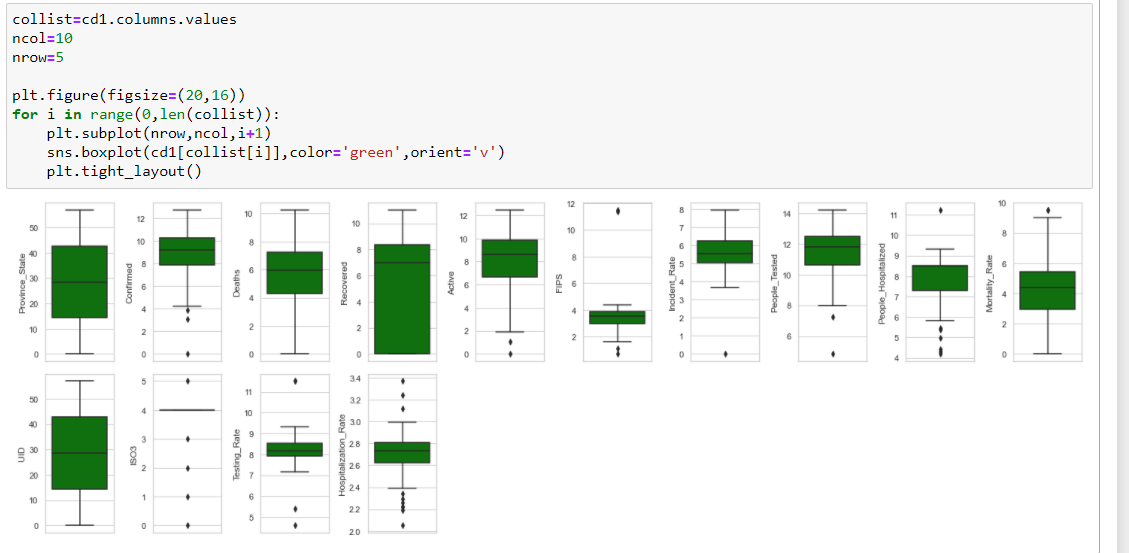
# Convert categorical data into numeric variable

# 

# 

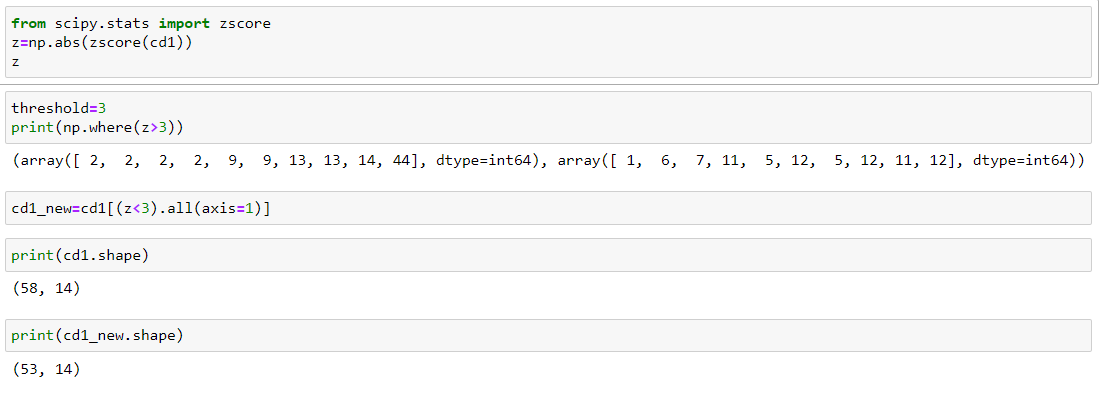
# Above you can see all data converted into numerical data type.

# Plotting outliers



Above graph shows outliers present in most of columns except Province\_state and Recovered columns.

Now we will remove outliers using Zscore.



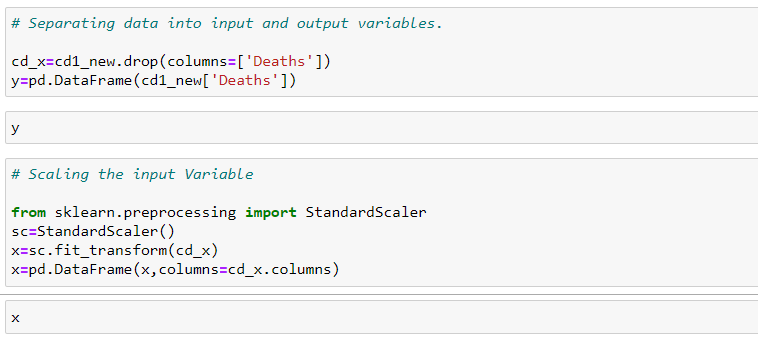
Now outliers removed from dataset. 5 rows removed from dataset due to outliers present.

Now our dataset is ready for build machine learning model.

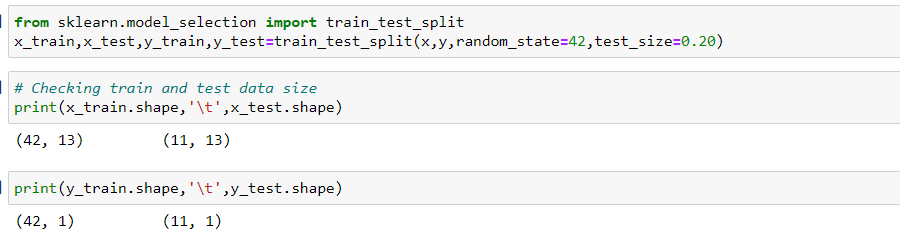
# Model Training

Now we will train several machine learning models and compare their results. Note that because the dataset does not provide labels for their testing-set, we need to use the predictions on the training set to compare the algorithms with each other. Later on, we will use regularization.

Now separating data into input and output variables. And scaling the input variable with standard scaler. It transform the data in such a manner that is has mean as 0 and standard deviation as 1. It arranges the data in a standard normal distribution.



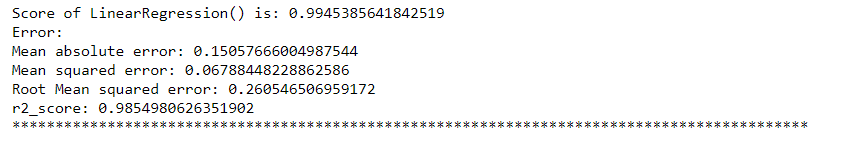
Now make input and output variables into train and test data and checking size of train and test data set. Will use 80% data in training and 20% data in testing.



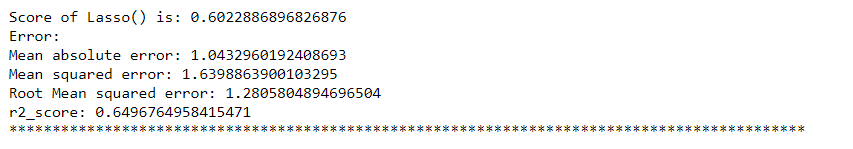
Now importing models library and using algorithms via for loop



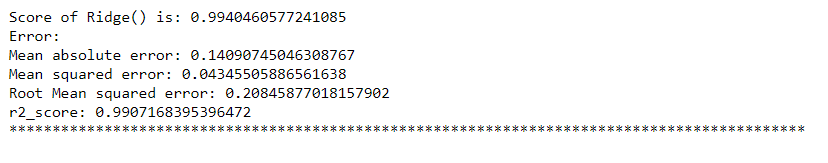
LinearRegression



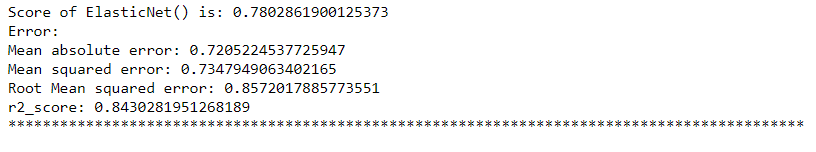
Lasso



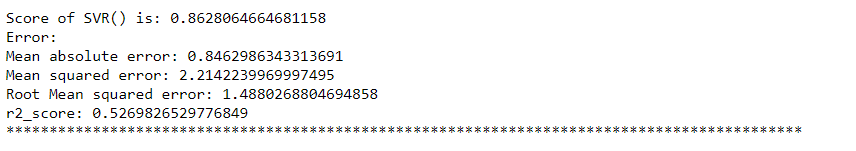
Ridge



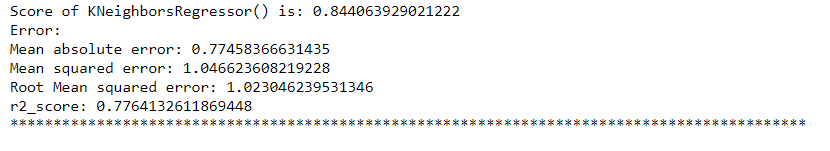
ElasticNet



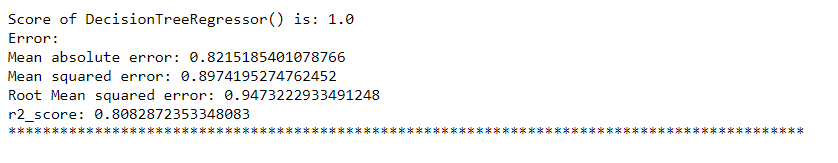
SVR



KNeighborsRegressor



DecisionTreeRegressor



Linear regression and Ridge working fine but Ridge provide best r2 score therefore will use ridge for hyperparameter tuning

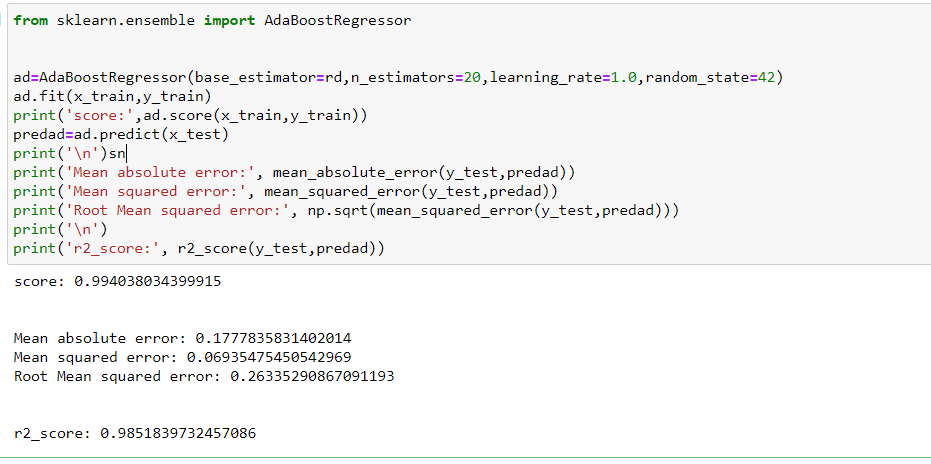
Ridge

What is ridge?

Ridge regression is a way to create a parsimonious model when the number of predictor variables in a set exceeds the number of observations, or when a data set has multicollinearity (correlations between predictor variables)

# Using Ensemble Technique to boostup score

Now will use ensemble technique to boostup score. Will use randomforestregressor and AdaBoostRegressor.



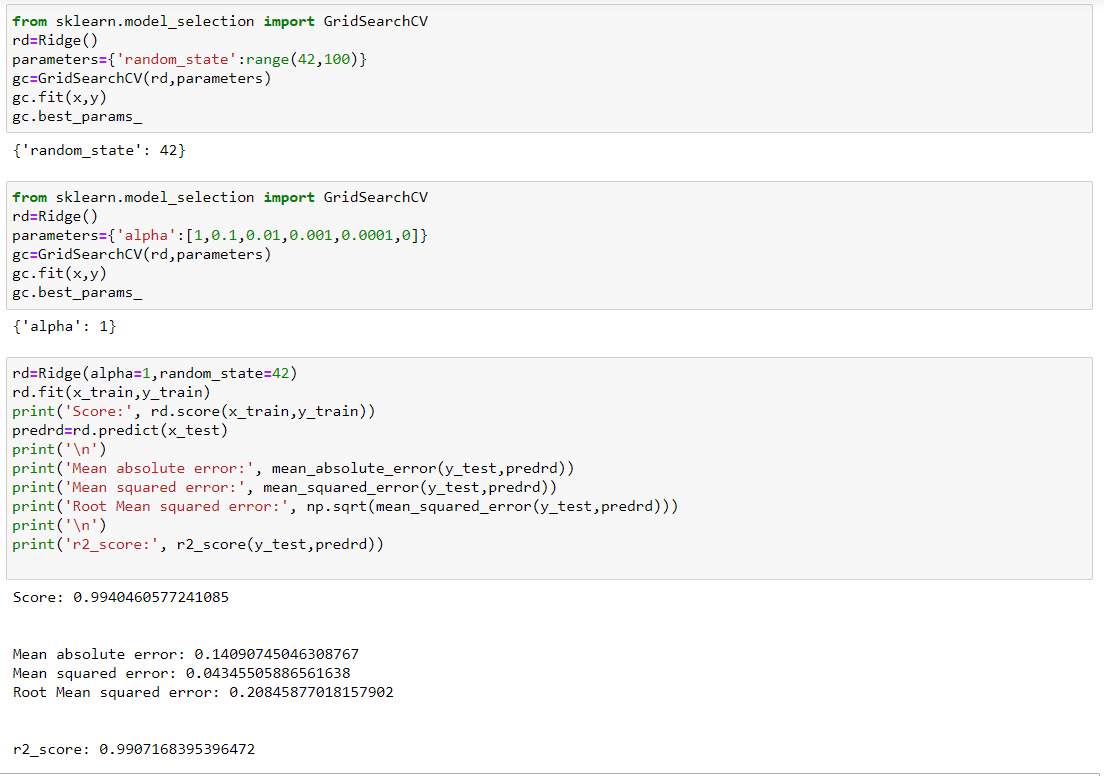
AdaBoostRegressor also provide high score and r2 score but Ridge provided better result than Adaboost.



AdaBoostRegressor also provide good score and r2 score but is less than ridge so will use ridge regression as final model.

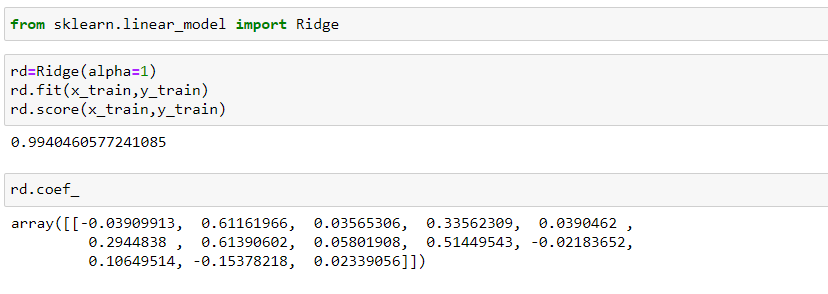
Now use hyperparameter tuning and regularization technics on ridge regression.

a hyperparameter is a parameter whose value is used to control the learning process. By contrast, the values of other parameters are derived via training

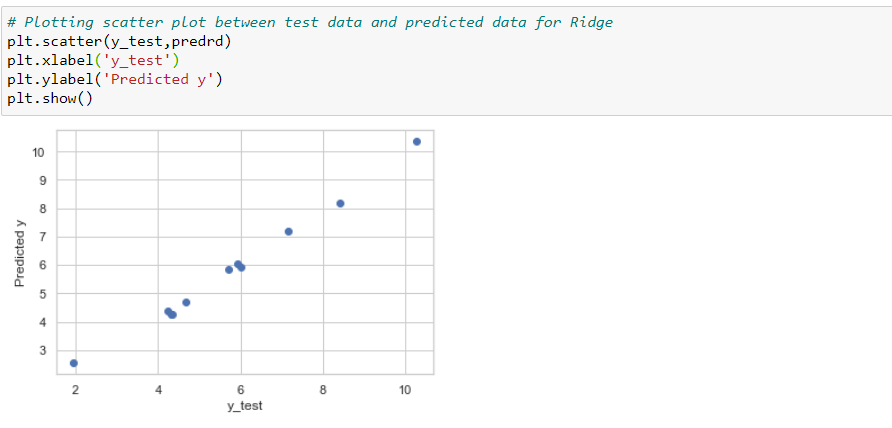


# Regularization

# Regularization is a technique used for tuning the function by adding an additional penalty term in the error function. The additional term controls the excessively fluctuating function such that the coefficients don't take extreme values.



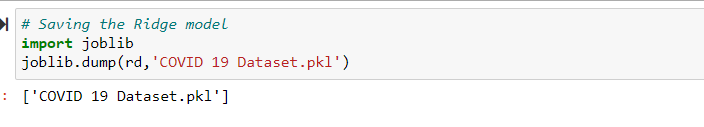
Plotting scatter plot between test data and predicted data for Ridge.



Above graph shows ridge model working fine and it is not under fitted and over fitted.

Now saving the ridge model.

Will use joblib library to save best model.



**Summary**

We started with importing libraries and load dataset. After that check null values, check correlation between two columns, performed univariate and bivariate analysis. During this process used seaborn and matplotlib to do best visualizations. During the data pre-processing part, we computed missing values, converted features into numeric ones, treated skewness via log method, removed outliers using zscore and scaling the input variables. Afterwards we started training 9 different machine learning models, picked one of them (Ridge) and applied hyperparameter tuning and Regularization on it. Then plot test data and predicted data and saved best model using joblib library.