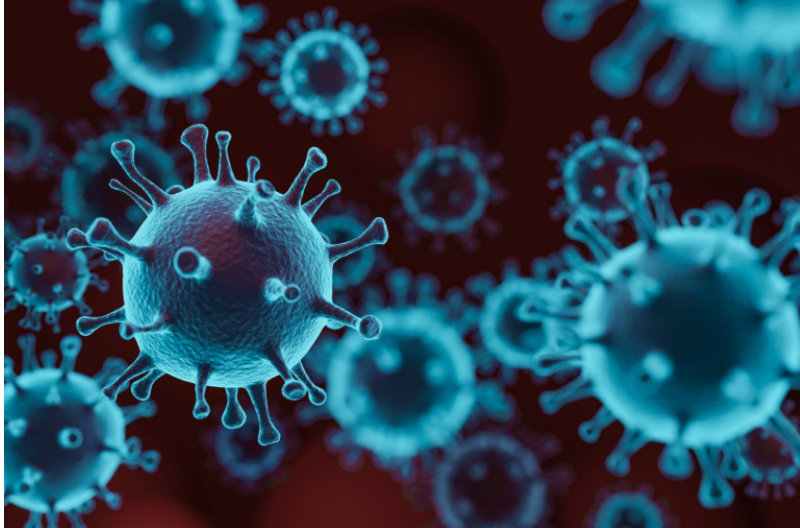
**PREDICTING DEATHS IN U.S DUE TO COVID-19**

****

**INTRODUCTION:**

Coronavirus disease (COVID-19) is an infectious disease caused by a newly discovered coronavirus. The first known infections from SARS-COVID 19 were discovered in Wuhan, China. Coronaviruses vary significantly in risk factor. Some can kill more than 30% of those infected, such as MERS -CoV, and some are relatively harmless, such as the common cold.

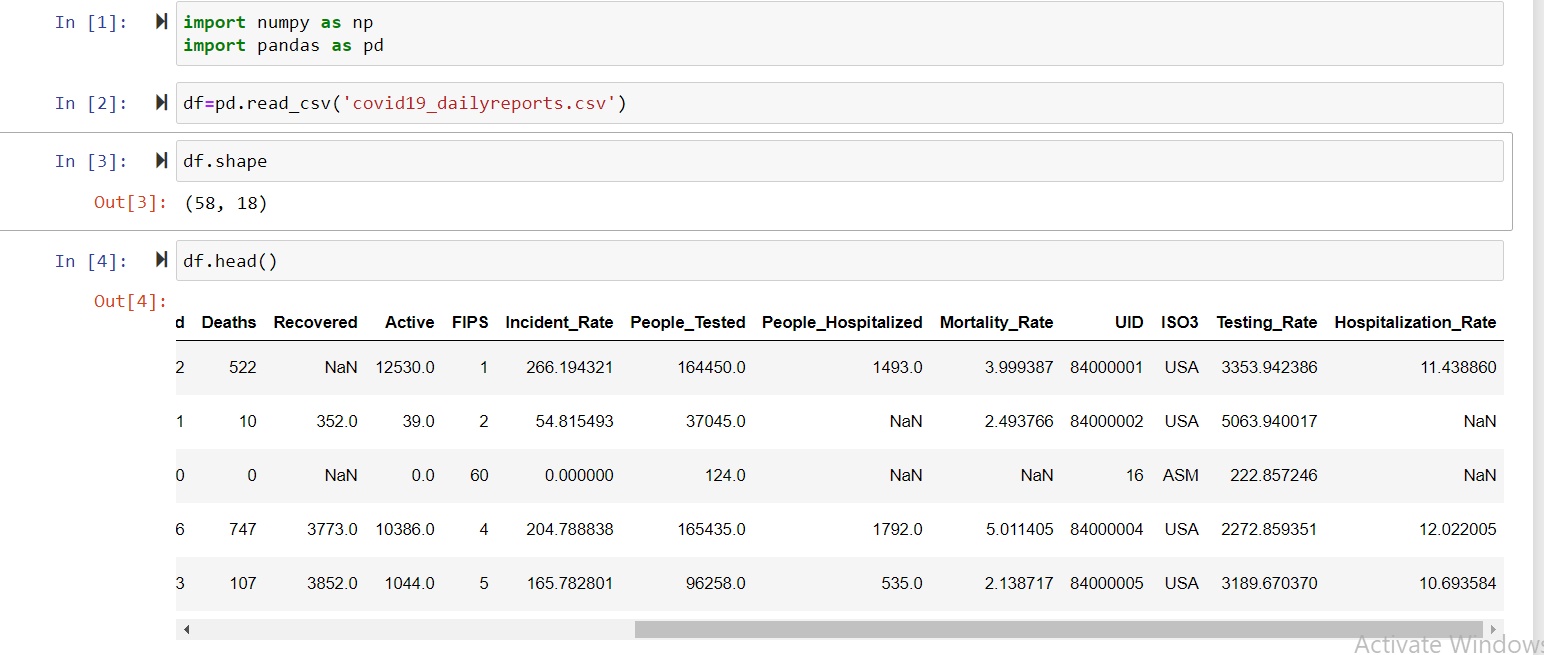
In December 2019, a pneumonia outbreak was reported in Wuhan, China. On 31 December 2019, the outbreak was traced to a novel strain of coronavirus which later declared as **Pandemic** by **WHO** due to high rate spreads throughout the world. As of 28 February 2021, there have been at least 2,524,133confirmed deaths and more than 113,745,002 confirmed cases in the COVID-19 pandemic.

**PROBLEM DEFINITION:**

This data set is retrieved from data repository for the 2019 Novel Coronavirus Visual Dashboard operated by the Johns Hopkins University Center for Systems Science and Engineering (JHU CSSE). This data set contains an aggregation of each USA State level data. We will be predicting the no.of deaths using machine learning techniques.

**DATA ANALYSIS:**

We will be importing the necessary libraries for performing further operations.



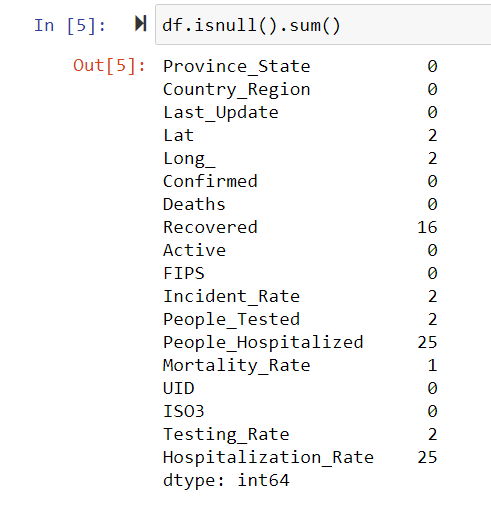
In the above code snippet firstly I have imported the numpy and pandas libraries as they are quite essential to perform operations on the data set.

Secondly, I have loaded the file(which was provided in the .csv format) into pandas dataframe.

Later, checked for the shape of data frame, which gives us an idea of how many rows and columns are present in our data.

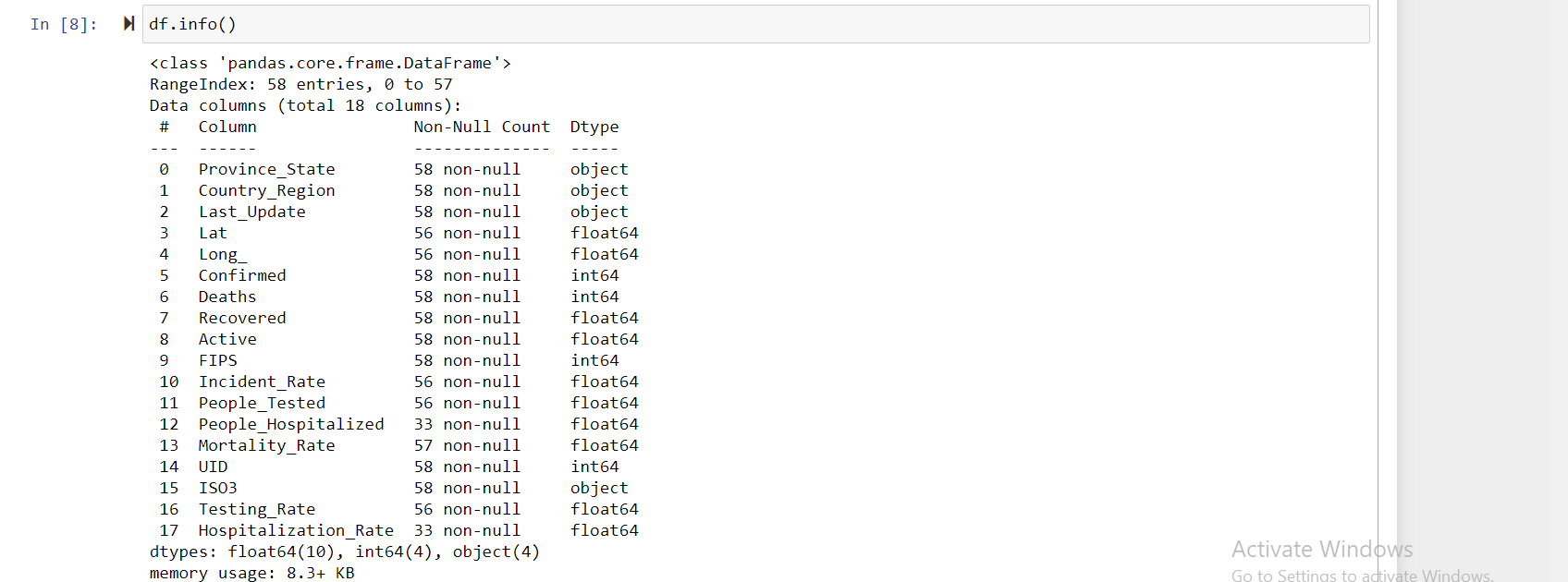
The head() function returns first 5 rows, which gives us the insight of how the data is present. we can notice that there are few NaN values which means null values are present in the data.

As its evident that there are null values, checking the presence of null values in the data frame,

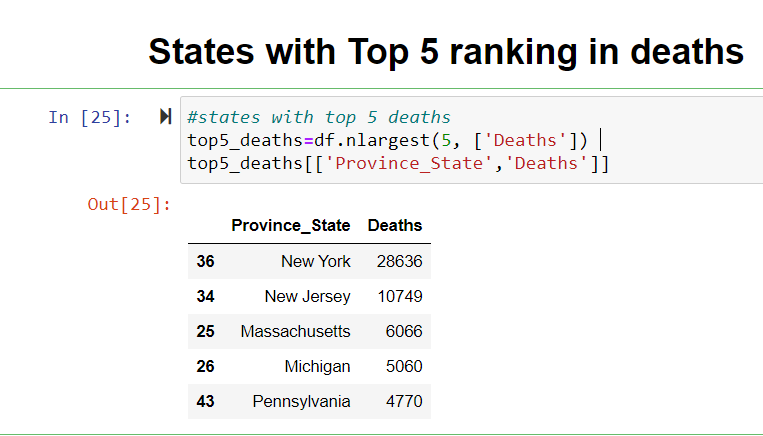


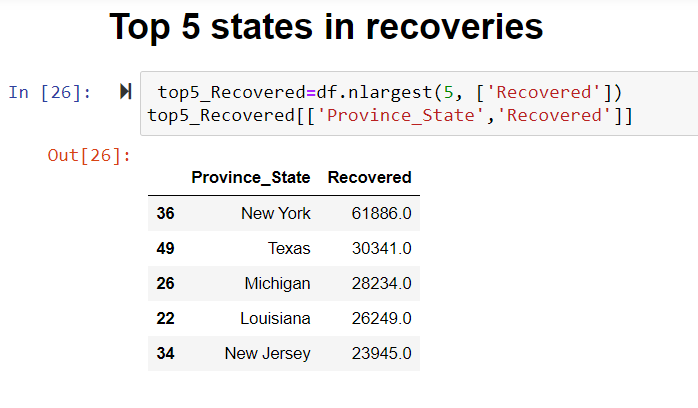
Isnull().sum() returns sum of null values in each column.If there are no null values,zero will be returned.These null values will be dealt later.

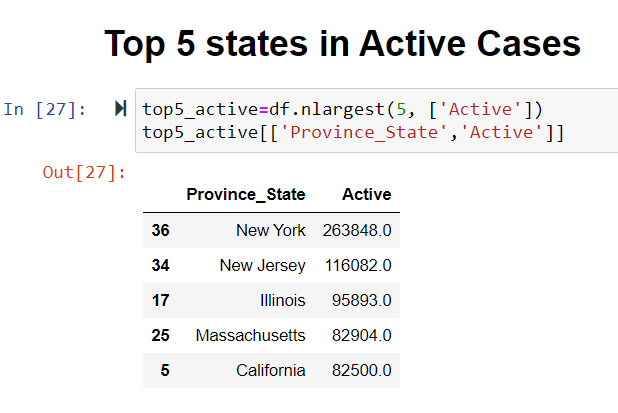
Lets check about the data types of each column

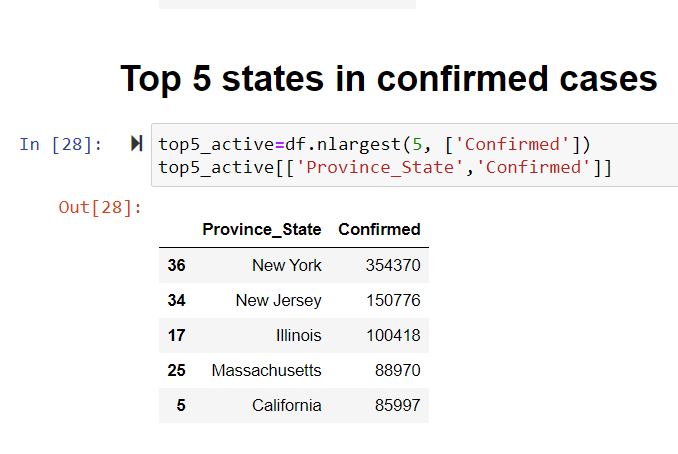


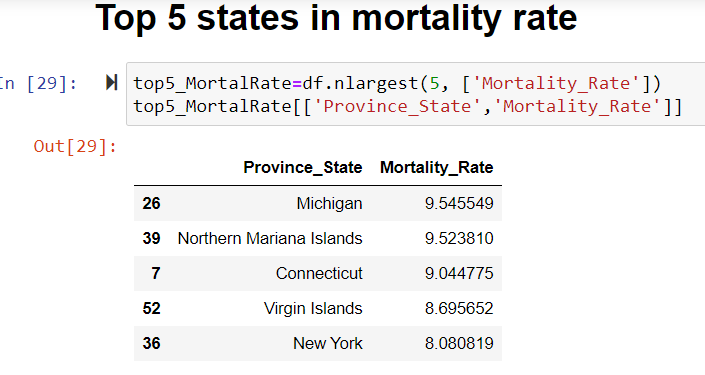
The columns of object type will be later converted to numeric through label encoding or one hot encoding.

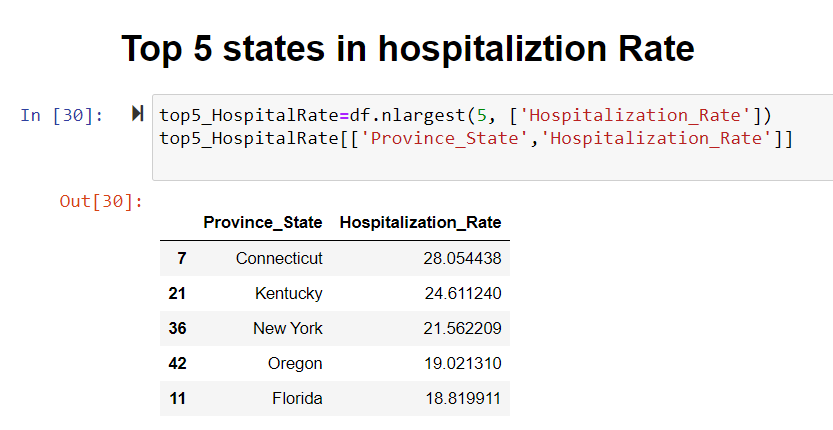








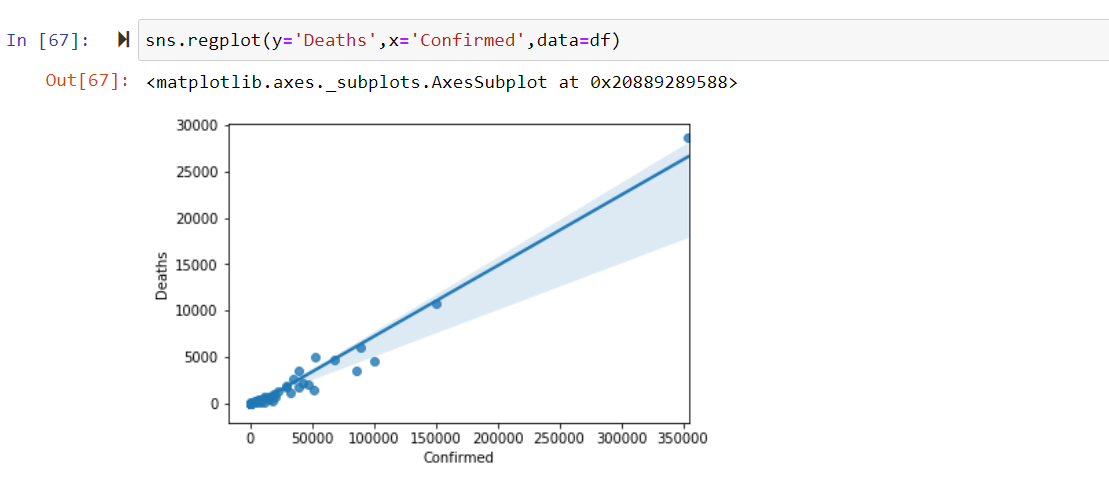




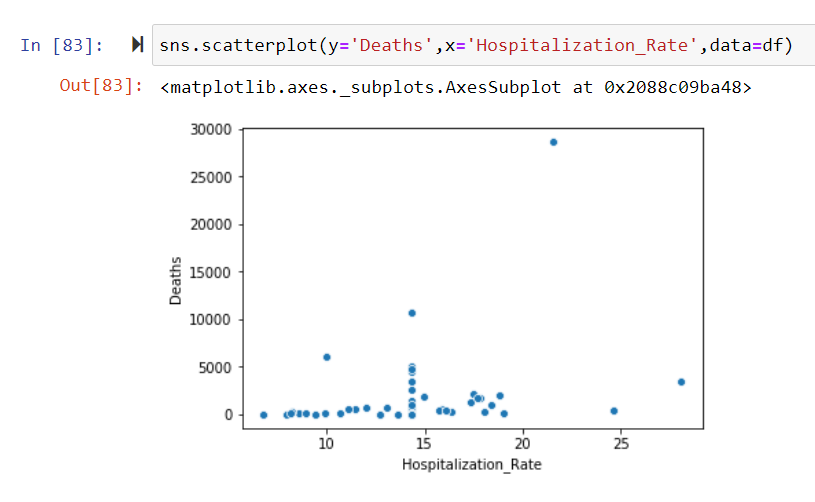




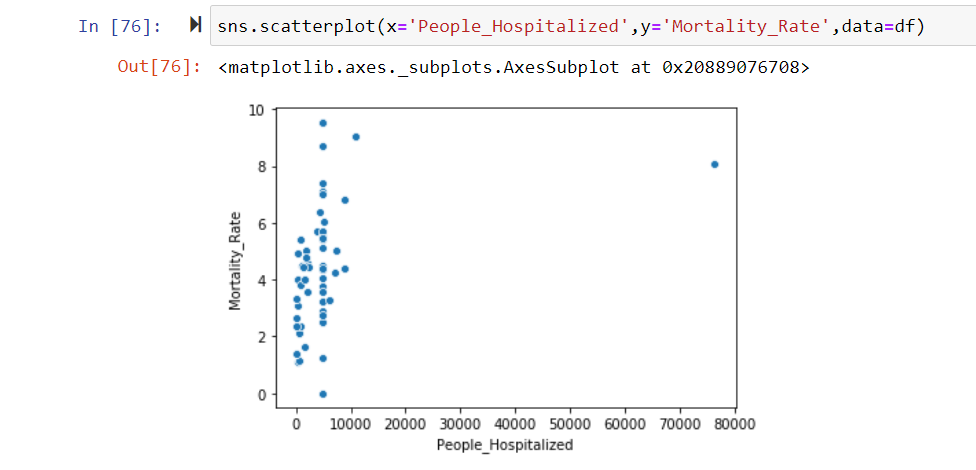




Plotted deaths with the confirmed cases, we can notice that death rate is 10% of confirmed cases. We can also notice that even though there are huge no. of confirmed cases, deaths are relatively low.



Even though hospitalization rate increased, Majority of deaths are less than 5000.

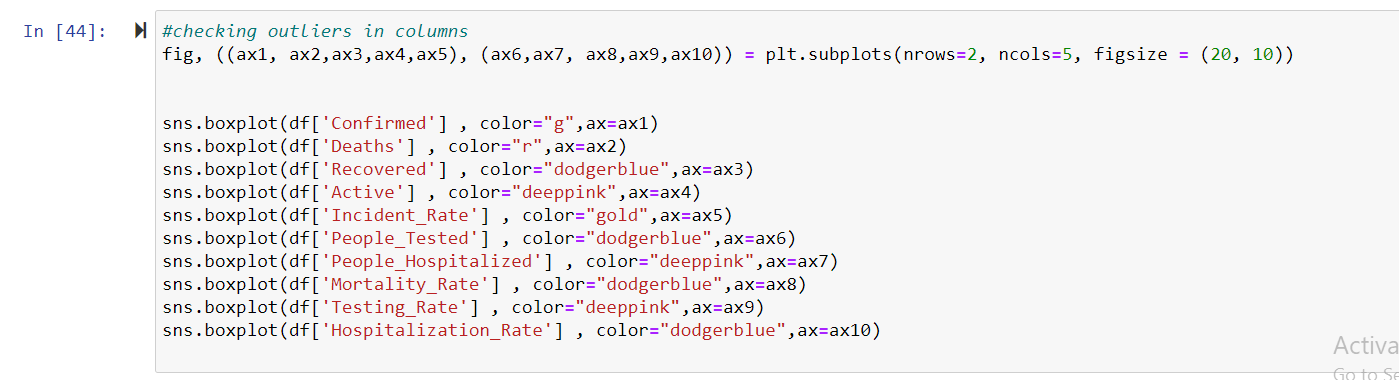


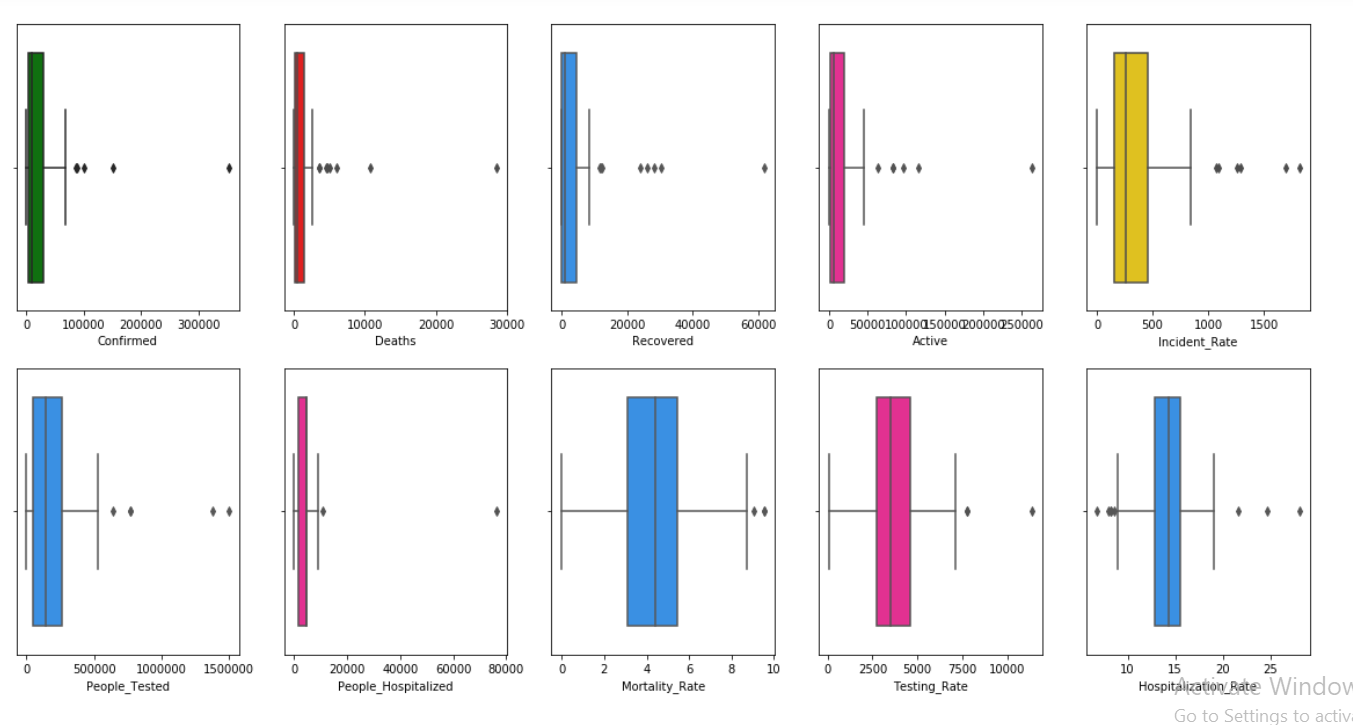
Out of all hospitalized people (80,000), higher mortality rate is in less than 10,000.

Outliers detection is an important aspect of data analysis. From statistical point of view, outliers are those values which are remarkably deviant from rest of the values in a sample.

An outlier may indicate an experimental error, or it may be due to variability in the measurement. By using box plots we can visualize the presence of outliers.

Pictorial representation of outliers for the given dataset.

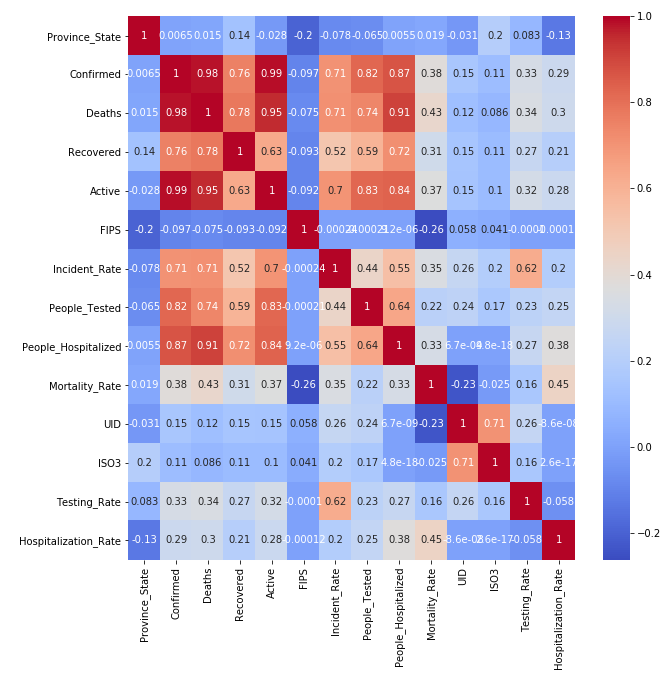




We can observe that there are outliers present in almost all the columns. These outliers has to be treated before feeding the data into model in order to achieve predictions with greater accuracy.

Correlation:

Correlation determines how variables depend on the target variable. Correlation is an important factor in building machine learning models.



Deaths has high correlation with :

Hospitalised people, Active Cases and confirmed cases.

**DATA ANALYSIS CONCLUSION:**

* NewYork tops in Confirmed cases,Active cases,Deaths and Recoveries
* Zero deaths are recorded in American Samoa and Diamond Princess provinces.
* Connecticut province tops in hospitalization rate,New York ranks third in hospitalization rate.
* North Dakota has very less hospitalization rate which is 6.7%.
* Out of all confirmed cases, hospitalization rate is very less(10%).
* There a possibility of 30% deaths in hospitalized people.

**DATA PRE PROCESSING:**

Pre processing involves cleaning the data into a suitable form such that it can be used by a machine learning model to perform predictions. For this data set, will be performing the below

1) Dealing with missing values

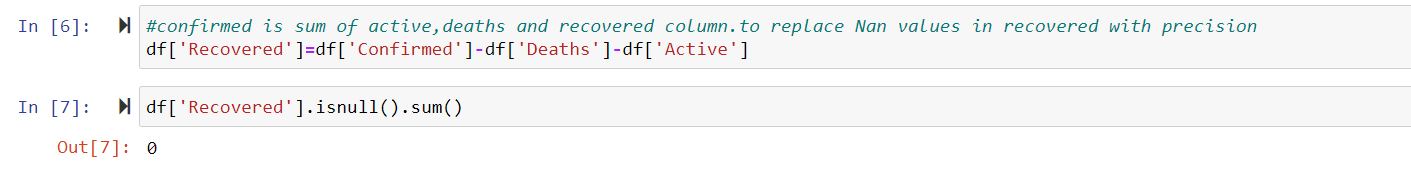
2) Conversion of object data type column into numeric through variety of techniques like one hot encoding or label encoding

3) Outlier Removal

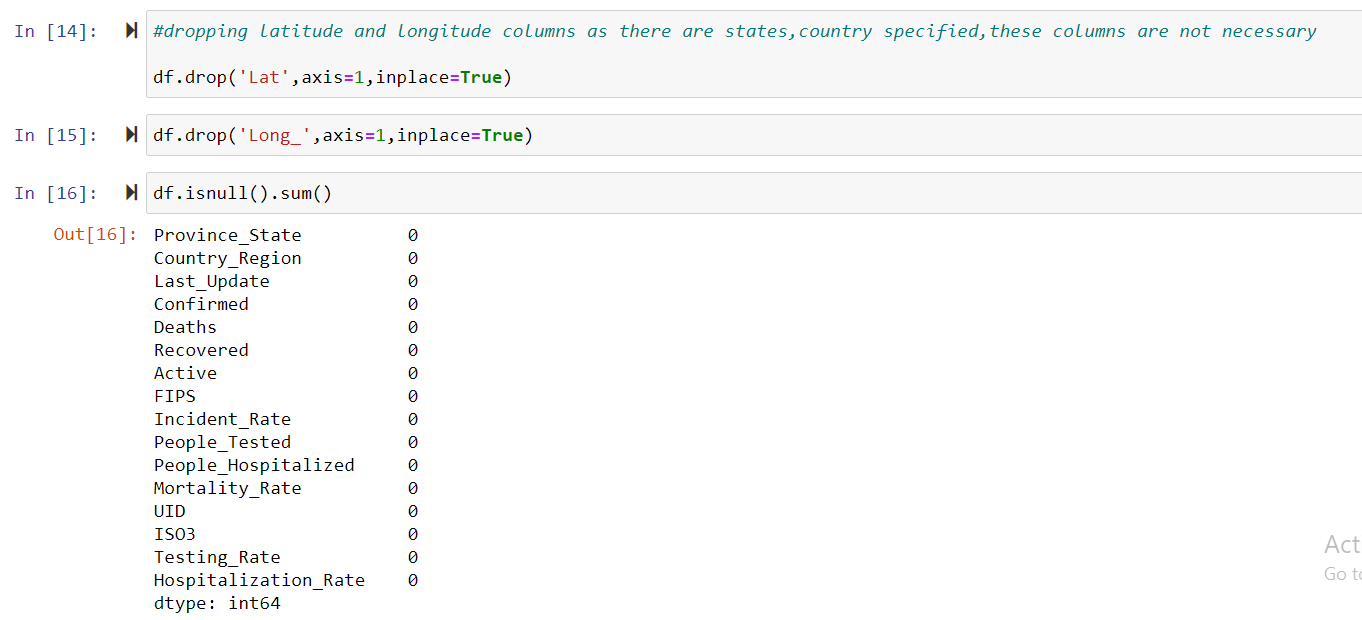
4) Splitting data into training and testing sets.

5) Feature Scaling

**Dealing with missing values:**

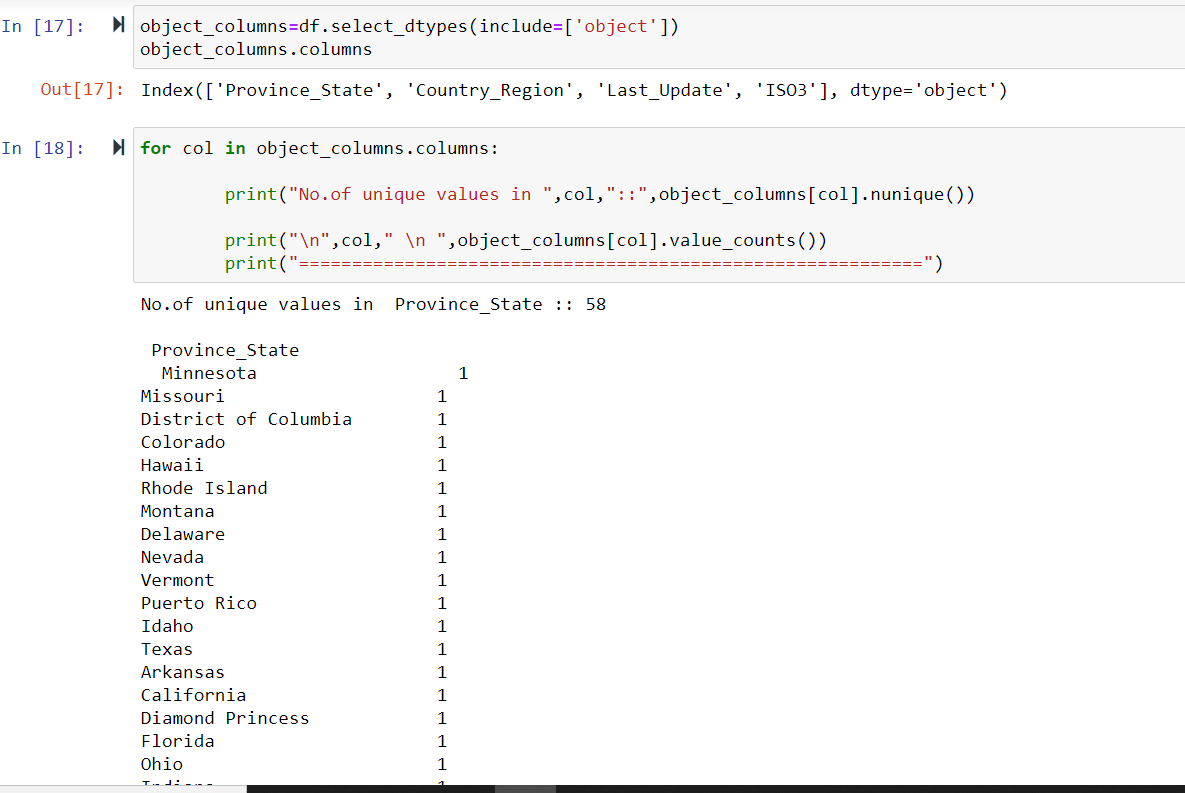
****

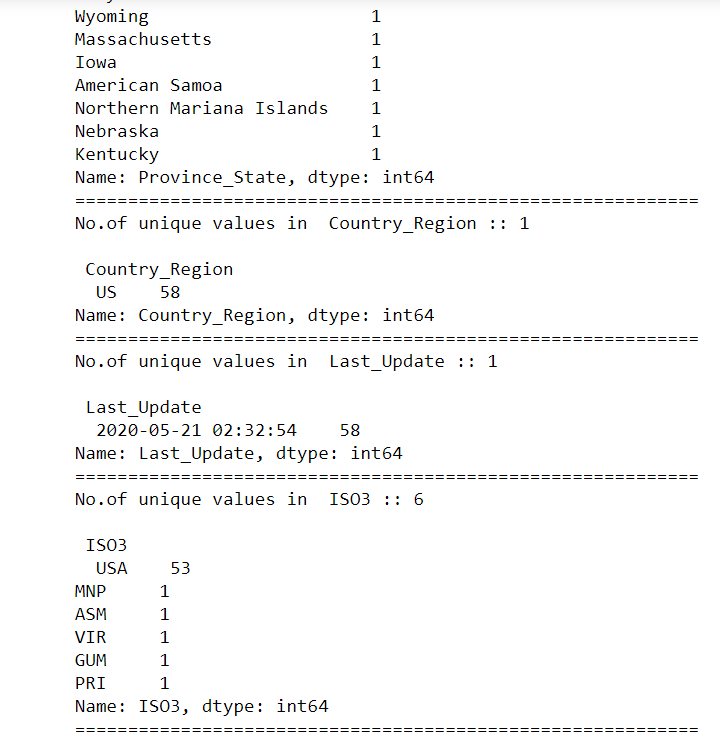
****

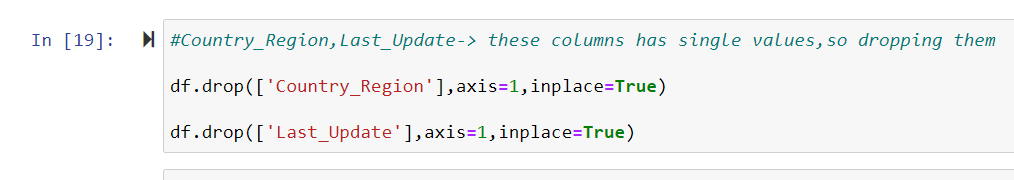
****

**Conversion of object data type column:**

Checking the entries of object data type columns:

****

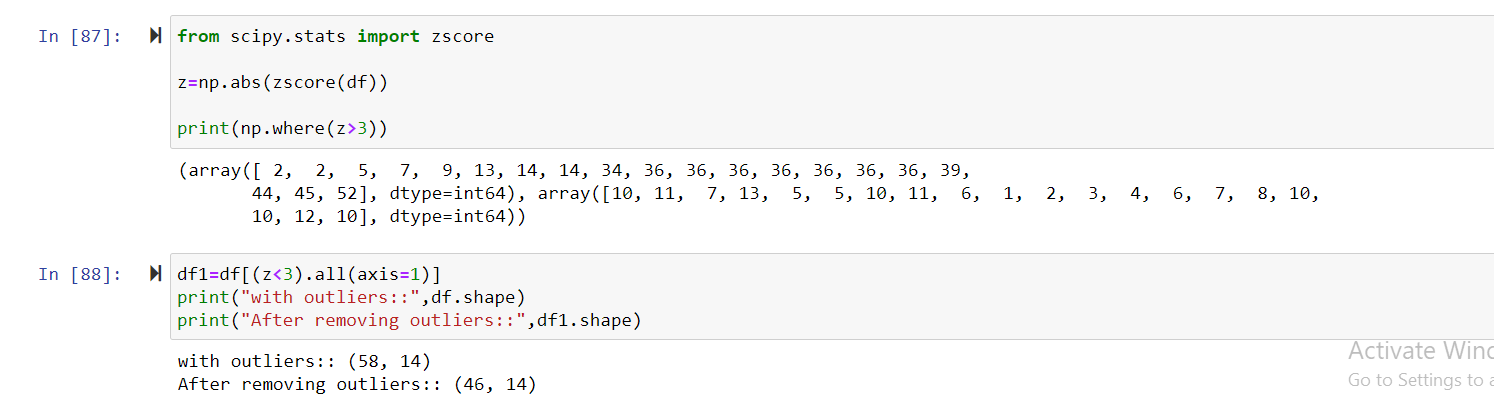
****

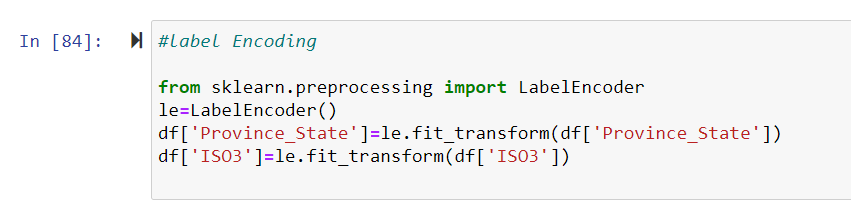
****

As there are same values in all the rows of country\_region and last\_update columns,it wont have an impact on predicting deaths,so dropped those 2 columns.

**OUTLIER REMOVAL:**

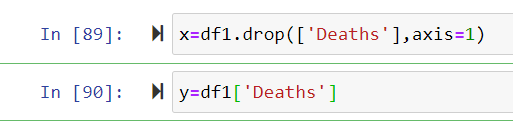
Outliers can be detected through Z Score or IQR(inter quartile Range),I opted for Z Score.



****

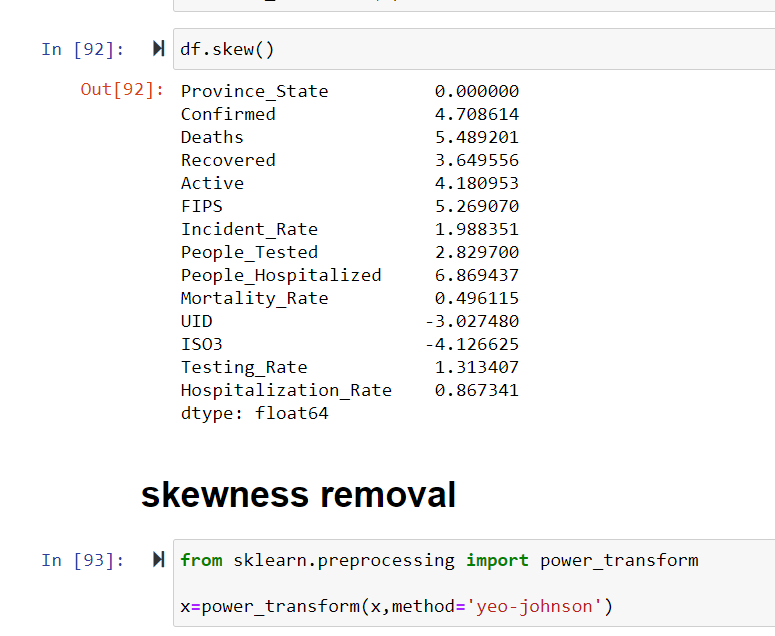
**Splitting data into training and testing sets:**

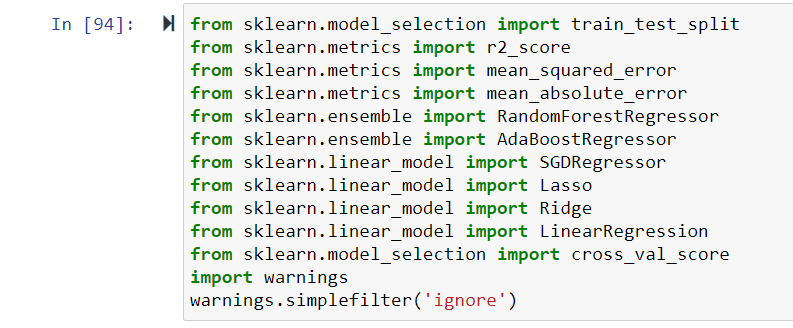
Before splitting the data, defining x and y. Y is the variable that is predicting. Rest of the variables will be stored into X.

****

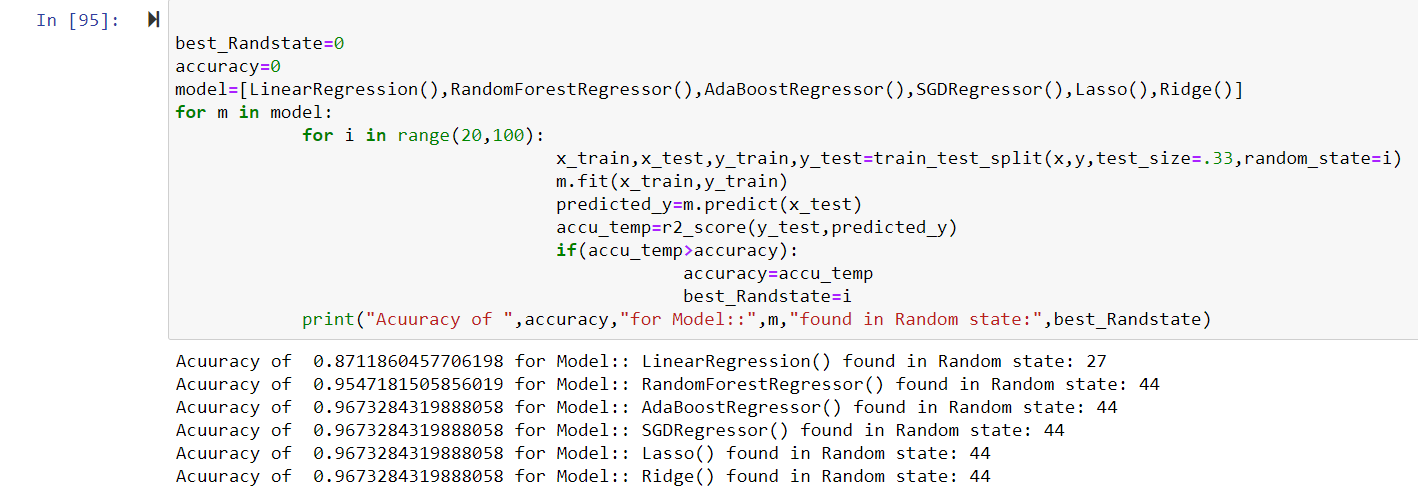
**Removing the skewness of data:**

Skewness is a measure of the symmetry in a distribution. A symmetrical dataset will have a skewness equal to 0.A normal distribution will have a skewness of 0.If distribution is skewed to the left it’s called as negative skew, If distribution is skewed to the right it’s called as positive skew. Presence of skewness is checked through skew() function. Skewness removal can be done through variety of techniques like box-cox, yeo-johnson, log transformation. I preferred yeo-johnson because it works for both positive and negatively skewed data.

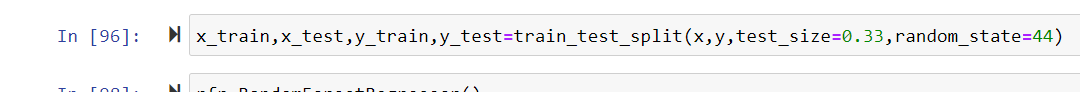


Importing the necessary libraries: 

Before splitting the data,finding the best random state for the possible models to achieve higher accuracy.I have opted r2 score as the performance metric.



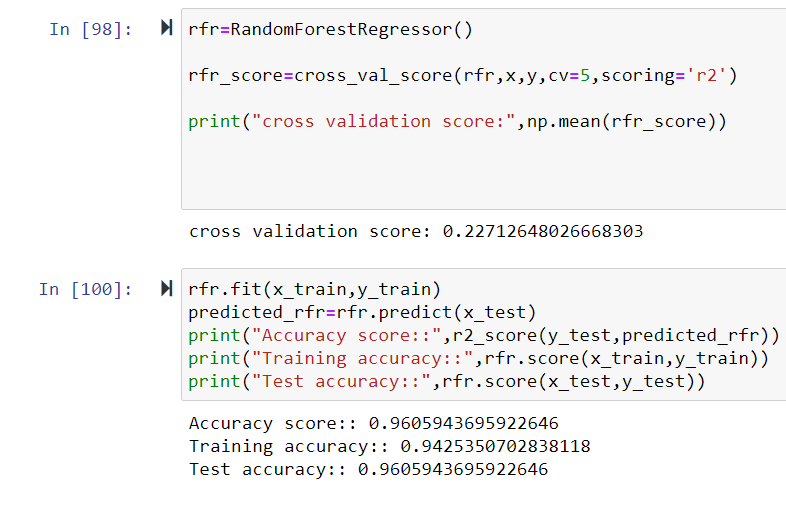
Note that highest accuracies are recorded for state 44.hence opting for it to split.



For splitting the data the usual test:train ratio is 30:70 or 20:80.I have chosen 33 as the test size.

I have chosen random forest Regression model as it has recorded 95% accuracy,

**MODEL BUILDING:**



Cross Validationisa very useful technique for assessing the effectiveness of your model, particularly in cases where you need to mitigate overfitting**.**

The problem with general evaluation technique is that it does not give an indication of how well the built model will predict on an independent or unseen data set. Cross validation gives us this idea about our model.

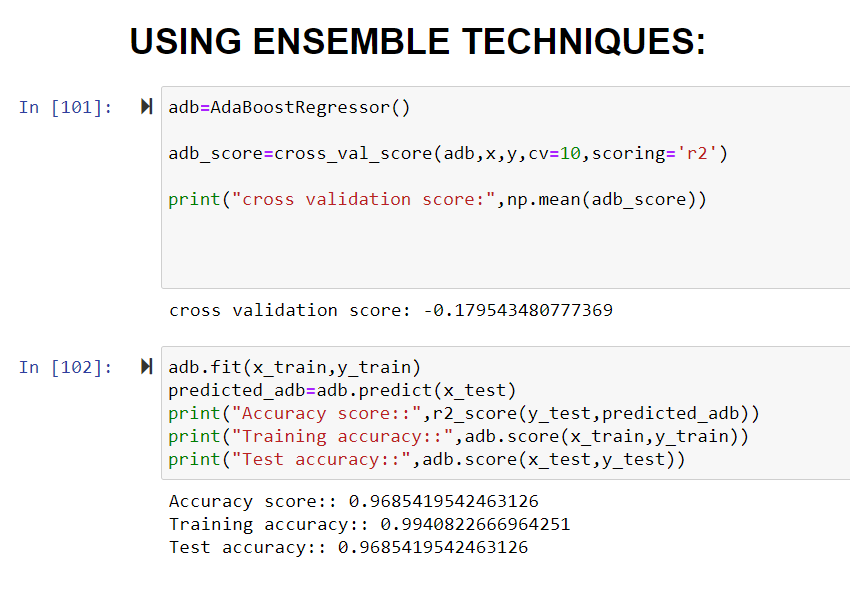
There are different kinds of cross validation techniques. I have used K-Fold cross validation.

In **K Fold cross validation**, the data is divided into k subsets. In general k value is taken as 5 or 10.

I selected k value as 5. Each time, one of the k subsets is used as the test set/ validation set and the other k-1 subsets are put together to form a training set*.*

**ENSEMBLE TECHNIQUES:**

Ensemble methods are algorithms that combine several machine learning techniques into one predictive model in order to **decrease variance**, **bias**, and thus**improves accuracy of predictions.**



**Hyper parameter Tuning:**

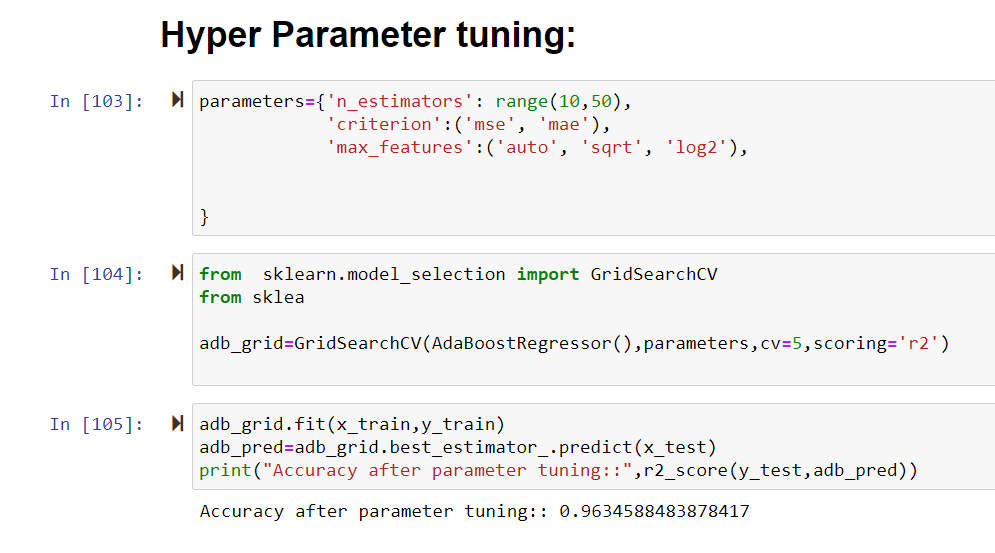
A hyperparameter is a parameter whose value is used to control the learning process.

Two strategies for Hyperparameter tuning are:

1)Grid Search CV

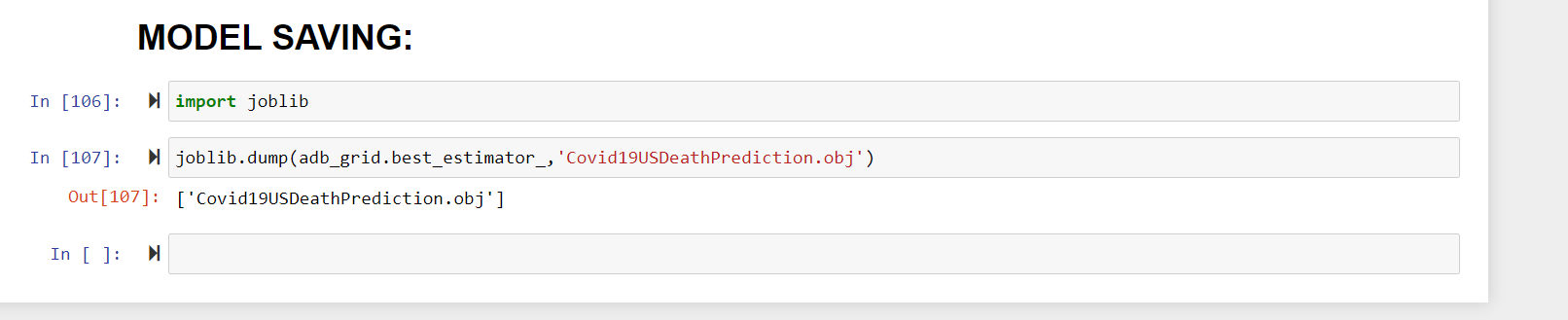
2)Randomized Search CV

In GridSearchCV, machine learning model is evaluated for a range of hyperparameter values. From a grid of hyperparameters values, GridSearchCV searches for best set of hyperparameters.



Hyper parameter tuning reduces the over fitting of the model.

**Saving the model:**

****

**Concluding Remarks:**

The built machine learning model predicts the deaths due to covid 19 with 96% accuracy.