**COVID-19 (Regression)**

**Problem Statement:** This is the data repository for the 2019 Novel Coronavirus Visual Dashboard operated by the Johns Hopkins University Center for Systems Science and Engineering (JHU CSSE). Also, Supported by ESRI Living Atlas Team and the Johns Hopkins University Applied Physics Lab (JHU APL).

**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 .The dataset  is provided to identify the deaths and recovered cases.

This table contains an aggregation of each USA State level data.

File naming convention

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.

There are 18 columns in the Dataset. The target or y variable in this dataset is Deaths and others are x variable. Y variable is dependent variable and x variable is independent variable which means the value of y is always dependent on the value of x.

This is a Regression problem because the target variable in this dataset is a continue variable. Machine learning model for continues variable can be prepare by the regression process. In this dataset the target variable Deaths is a continue variable so we use the regression process for this dataset.

**DATA ANALYSIS:**

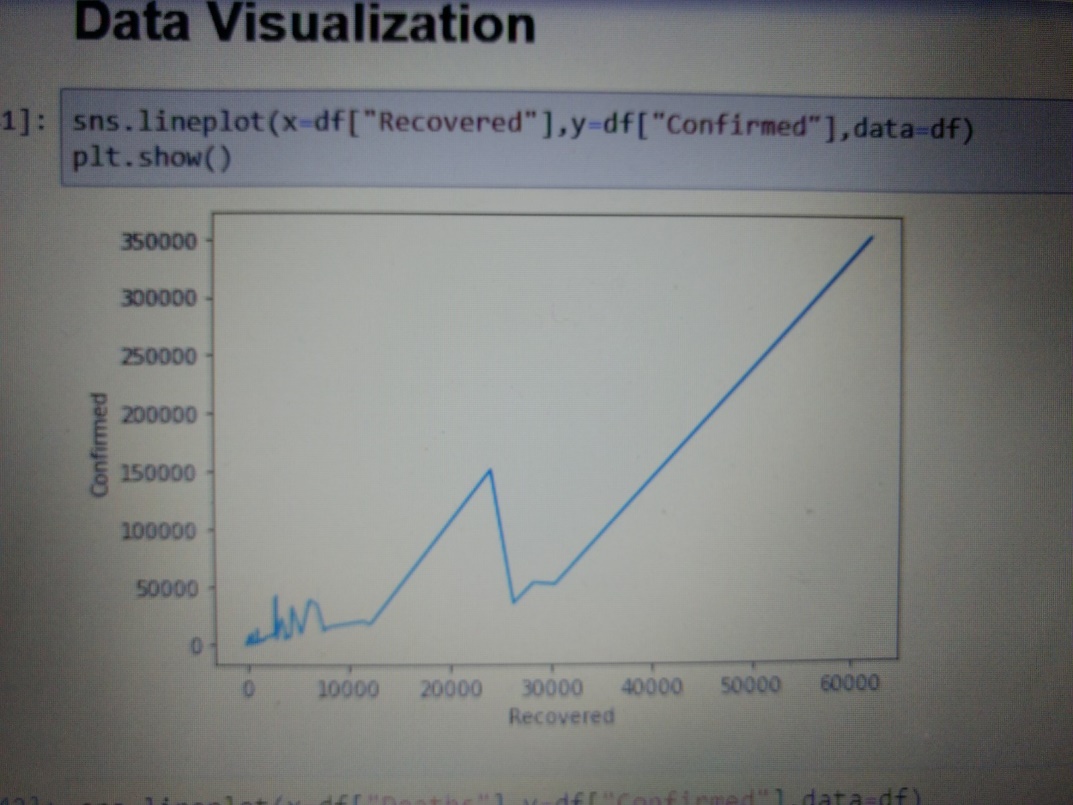
First we read the csv file into our jupyter notebook using pd.read\_csv method and check the columns of the data using df.columns() function. After checking the columns. We used df.shape for checking the shape of data then we used the df.info() function for getting the information about the datatype of this dataset. We found that there are some object type variables some int type and some float type. This step has given us the basic idea of our Dataset that how do we have to handle the Dataset.

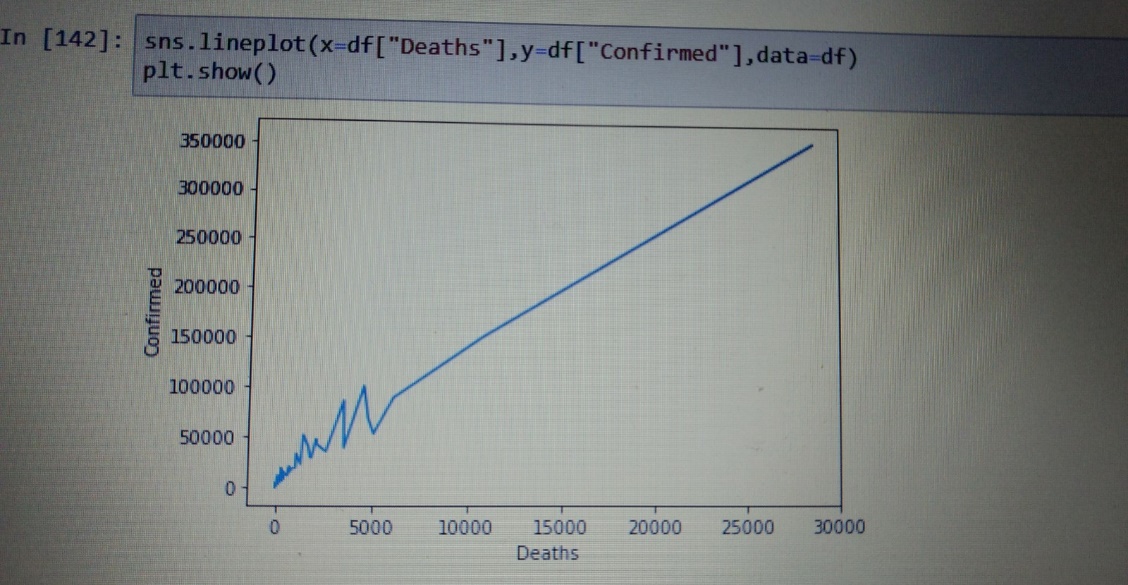
**Pre-processing pipeline:**

After that we have checked null values in our data using df.isnull().sum(). We have observed that there are 9 columns those have null values. Now we have to manage null values. We can use various methods for this. We used mean method. In this we replace null value by the mean of the Column. We used this method in six columns.

In next step we have dropped data which is not relevant for our model. We have dropped data by df.drop function. We have dropped 8 columns. And again we have checked the shape of our data by df.shape. we have total 10 columns.

Then for data visualization we have drawn line plots. With the help of data visualization we showed relation between various variables . We showed the surge in number of cases and deaths because of COVID.

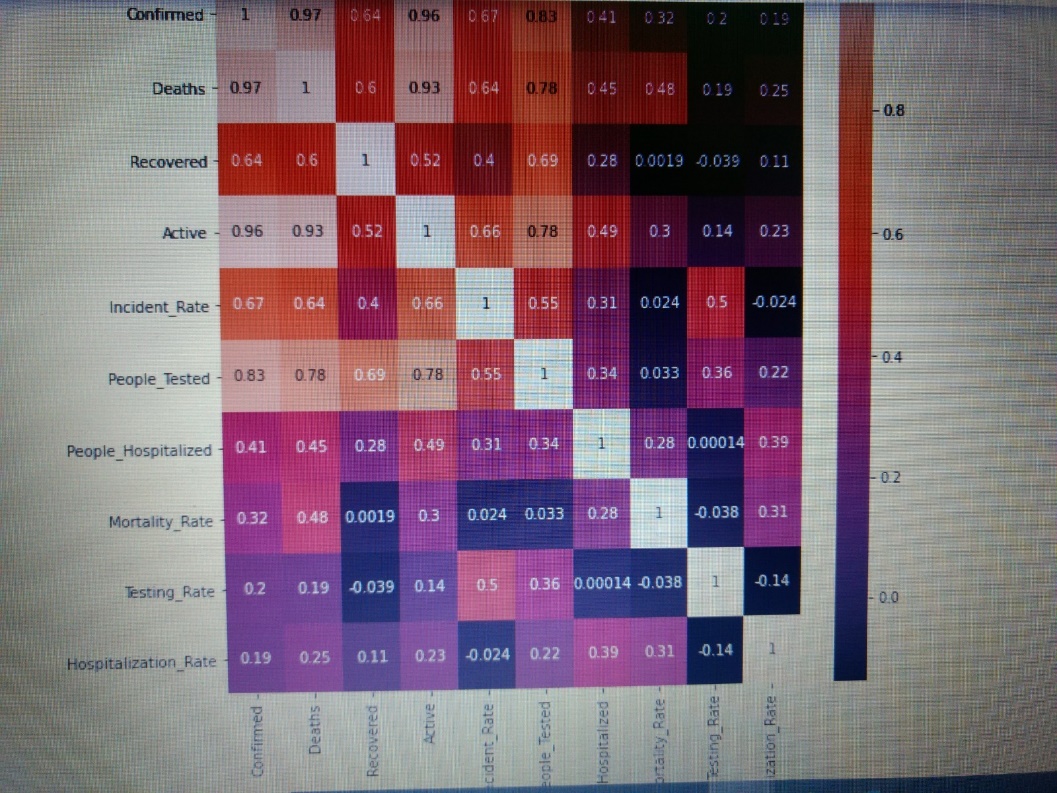


****

**EDA:**

EDA is known as the exploratory data analysis. In this analysis it is shown that which variable has the most impact on the target variable and how a column is performing in predicting the target variable.

In this section first we used heatmap correlation. For implementing this we imported Seaborn library. After applying heatmap correlation matrix. We can see the correlation between variables.

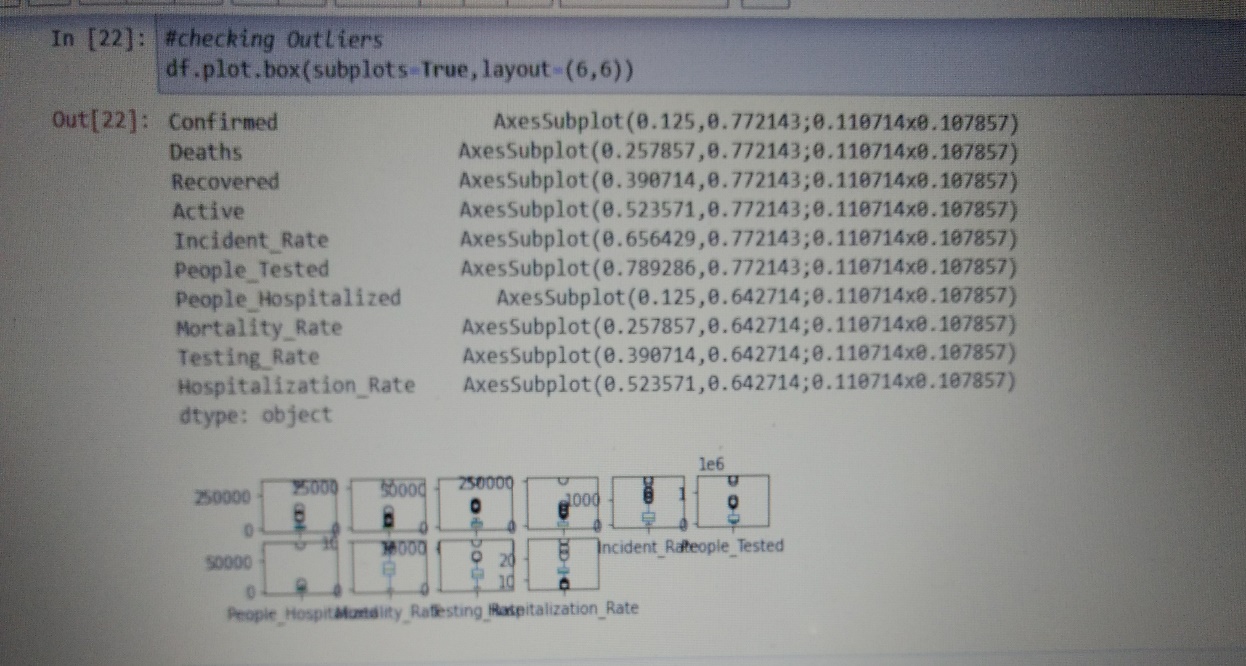


**Outliers**

An outlier is a data point in a data set which is distant or far from all other observations available. It is a data point which lies outside the overall distribution which is available in the dataset.

**Plotting outliers in the Dataset:**

For detecting outliers we have plotted box plot.



**Z – Score**

A Z-score is a numerical measurement that describes a value's relationship to the mean of a group of values in the dataset. Z-score is measured in terms of standard-deviation from the mean.

For removing outliers we have detected Z-score. We have removed data which has Z-score more than 3.

**To check distribution of Skewness**

**Skewness**

Skewness refers to distortion or asymmetry in a symmetrical bell curve, or [normal distribution](https://www.investopedia.com/terms/n/normaldistribution.asp) in a set of data. Besides positive and negative skew, distributions can also be said to have zero or undefined skew. The skewness value can be positive, zero, negative, or undefined.

We used df.skew() function for checking skewness. We have detected some skewness in data.

**Treating Skewness**

The log-transformed data follows a normal or near normal distribution while treating skewness. In this case, the square root transformation method removes or reduces skewness.

**Building Machine Learning Model:**

We are using the regression model for this dataset as target variable of this dataset is a continue variable. First we import all the necessary library for the machine learning model from the sklearn. After importing all the libraries we first split the data into train test split.

**Using Regressor models for model building.**

Different type of Regressor models used in model building.

**1. Linear Regression –** The linear regression is a linear approach to modelling the relationship between a scalar response and one or more explanatory variables (also known as dependent and independent variables). The case of one explanatory variable is called simple linear regression; for more than one variable, the process is known as multiple linear regression.

2. **Lasso** - Lasso regression is a linear regression that uses shrinkage. Shrinkage is where data values are shrunk towards a mid point, like the mean. The lasso procedure encourages simple and sparse models. In machine learning model, lasso is a regression analysis method that performs variable selection and regularization in order to bring out the prediction accuracy and interpretability of the statistical model it produces.

**3. Ridge** - Ridge regression is used to create a model when the number of predictor variables in a set exceeds the number of observations in the given dataset, or when a data set has multicollinearity (correlations between predictor variables) in it.

**4. Elastic-Net** - In statistics the fitting of linear or logistic regression models, the elastic net regularize regression method that linearly combines the L₁ and L₂ penalties of both lasso and ridge methods.

**5. SVR** - As in classification, support vector regression (SVR) is characterized by the use of kernels, sparse solution, and VC control of the margin and the number of support vectors. Support Vector Regression (SVR) is quite different than all other Regression models. The Support Vector Machine (SVM) algorithm, a classification category algorithm is used to predict a continuous variable.

**6. KNeighbors Regressor** – It is a Regression based on k-nearest neighbours. In the KNeighbours model target is predicted by local interpolation of the targets which associated to the nearest neighbours in the training set.

7. **Decision Tree Regressor** - Decision tree learning is one of the predictive modelling approaches used in statistics, data mining and machine learning. It uses a decision tree to go from observations about an item to conclusions about the item's target value.

**8. Random Forest Regressor -** Random Forest uses multiple decision trees as base learning models in the dataset. Random forest is a meta estimator that fits a number of classifying decision trees on various sub-samples of the dataset and uses averaging to improve the predictive accuracy and control over-fitting in the dataset. The main concept of Random Forest is to combine multiple decision trees in determining the final result rather than relying on individual decision trees.

After importing required libraries we have imported all the models and detected Score of every model, mean square error, mean absolute error, root mean square error and r2 score.

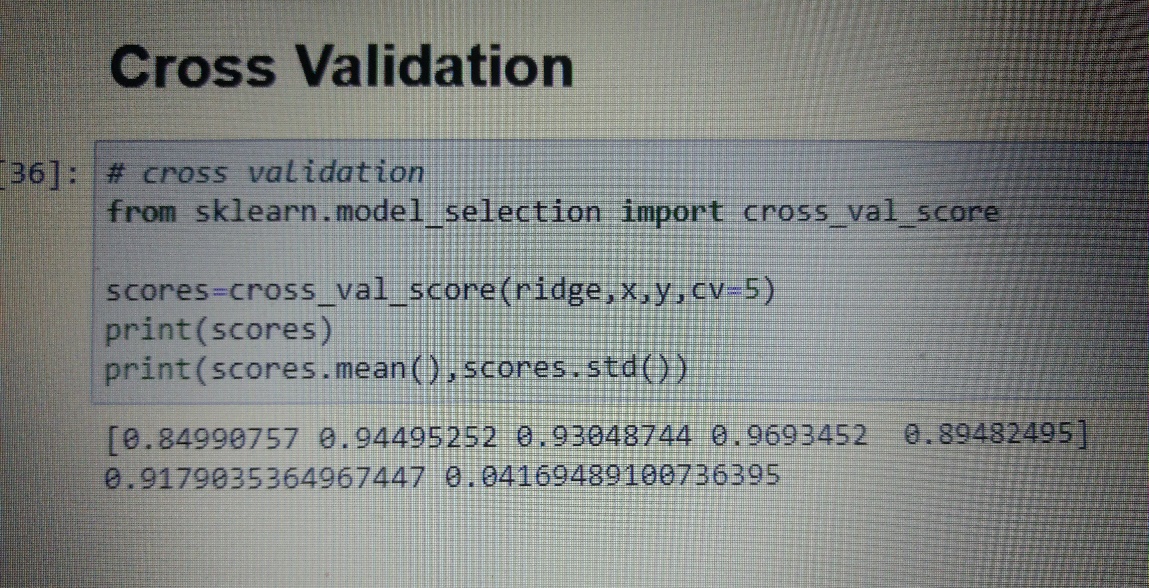
For increasing accuracy we have imported Random Forest and Gradient boost methods. These are known as ensemble techniques. After using these we have observed accuracy and error in both of the techniques.

We observed that we got best parameter using Ridge regression. Now we have used GridSearch CV for getting the best parameters for Ridge.

**Cross Validation**

Cross validation helps to find out the over fitting and under fitting of the model.In the cross validation the model is made to run on different subsets of the dataset which will get multiple measures of the model. If we take 5 folds, the data will be divided into 5 pieces where each part being 20% of full dataset. While running the Cross validation the 1st part (20%) of the 5 parts will be kept out as a hold out set for validation and everything else is used for training data. This way we will get the first estimate of the model quality of the dataset. In the similar way further iterations are made for the second 20% of the dataset is held as a hold out set and remaining 4 parts are used for training data during process. This way we will get the second estimate of the model quality of the dataset. These steps are repeated during the cross validation process to get the remaining estimate of the model quality.

We have applied Cross validation and get the estimate.



**Saving the model:**

After cross validation we have saved our model by importing joblib library and used dump method. We have saved our model by ‘COVID19.obj’ name.

**Conclusion:**

We conclude that dataset contains null values. While doing EDA we have removed outliers and also treated the skewness than sent our data to models also used ensemble techniques to get the best accuracy score and find that logistic gave high accuracy so we found best parameters for both and again checked the accuracy score using best parameters and get to know that the best model for prediction is Ridge which we have saved by using joblib for future use.