

**Using Machine Learning to Predict the Growth of COVID-19 in USA various Region**

**CONTENTS**

1. Problem Definition**…………………………………………………….**3
2. Exploratory Data Analysis **…………………………………………...**7
3. Visualization**…………………………………………………………..…**11
4. Preprocessing Pipeline**………………………………………………...**12
5. Building Machine learning model **……………………………….....**16
6. Concluding Remarks**…………………………………………………...**17
7. References **……………………………………………………………......**17
8. Problem Definition**:**

**Coronavirus** is a family of viruses that can cause illness, which can vary from **common cold** and **cough**to sometimes more severe disease.

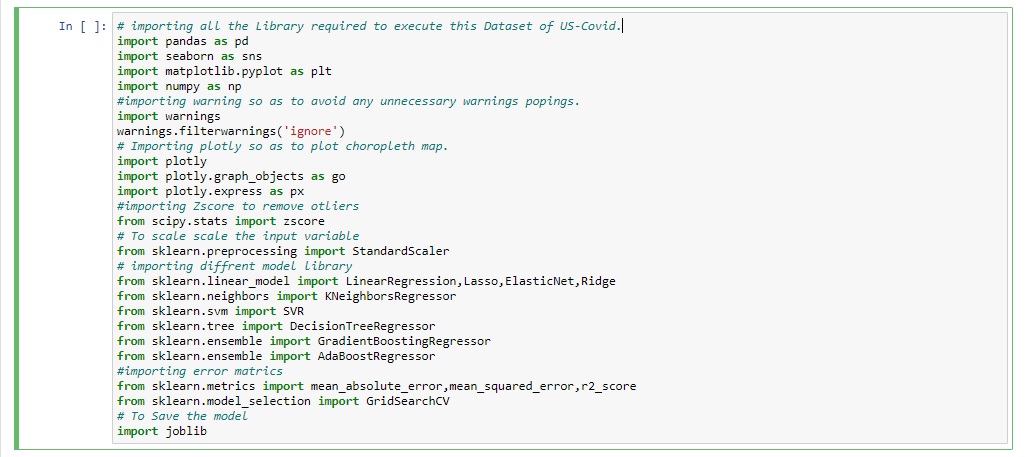
**Middle East Respiratory Syndrome (MERS-CoV)** and **Severe Acute Respiratory Syndrome (SARS-CoV)** were such severe cases with the world already has faced. **SARS-CoV-2 (n-coronavirus)** is the new virus of the coronavirus family, which first discovered in 2019, which has not been identified in humans 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.

There are already many different analyzes and dashboards describing this dangerous phenomenon of **Covid-19.** In this blog I am going to write about the **covid-19** spread In **USA** and his different **province region.**

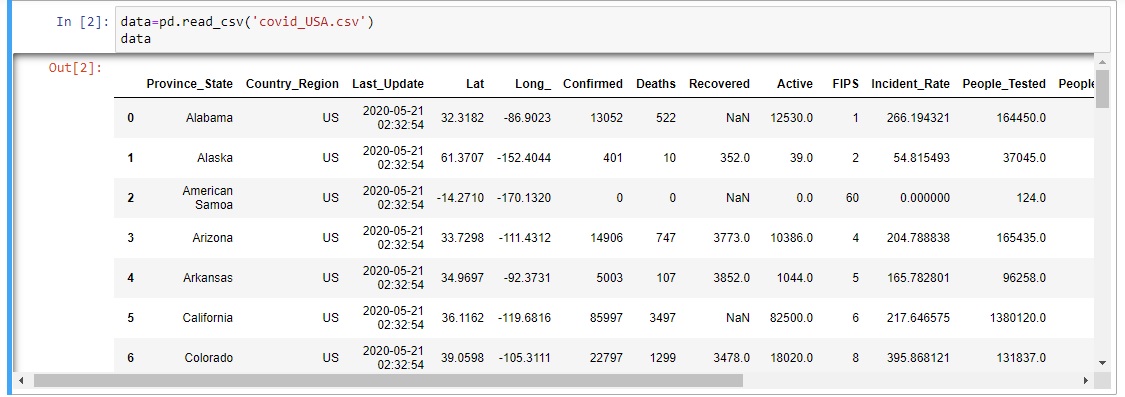
**The Goal is to predict the Total deaths which is continuous in nature.**

* **You can download the data set and code from this**[**Link**](https://github.com/khurram-DS/Covid_19_us_dataset)

1. Exploratory Data Analysis**:**

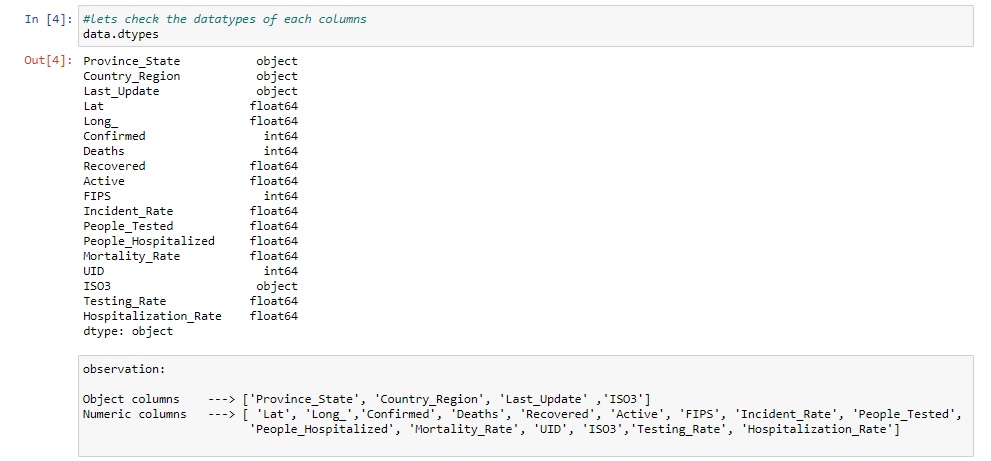
**I am importing all the necessary Library required to explore this USA covid-19 Dataset, with these libraries we can do EDA, Visualization, and Prediction of this Dataset, why I am doing this because you can navigate all the library at one place.**

* **Loading Datasets:**



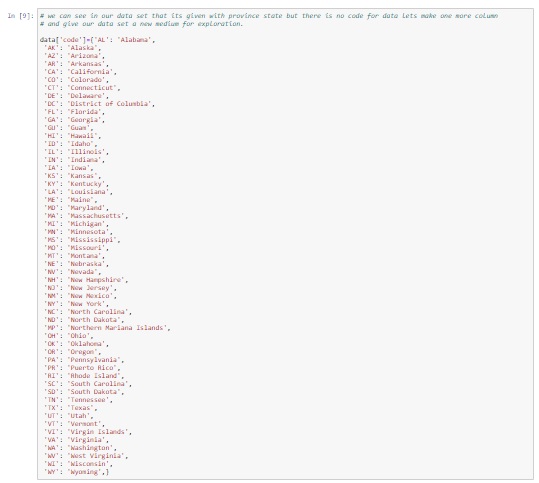
I have loaded the datasets with the name data. We can observe from the above Dataset that the data is been distributed over different Province state of US, Let’s see the data types first before moving further to explore this dataset.

* **Data types:**

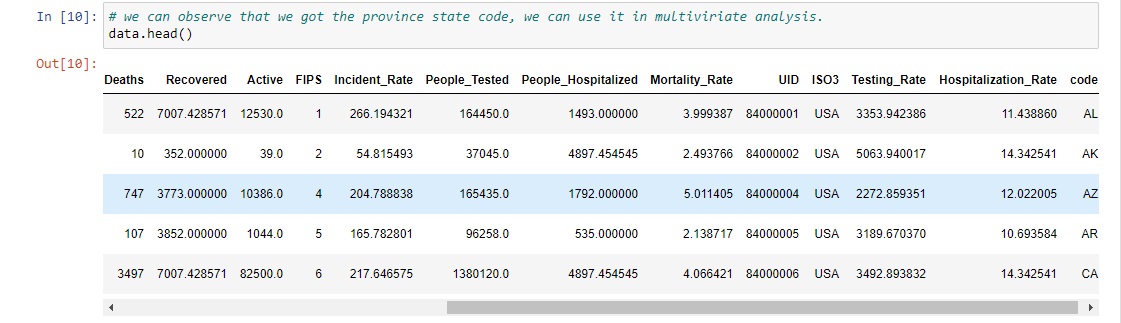


* **Checking Null values:**

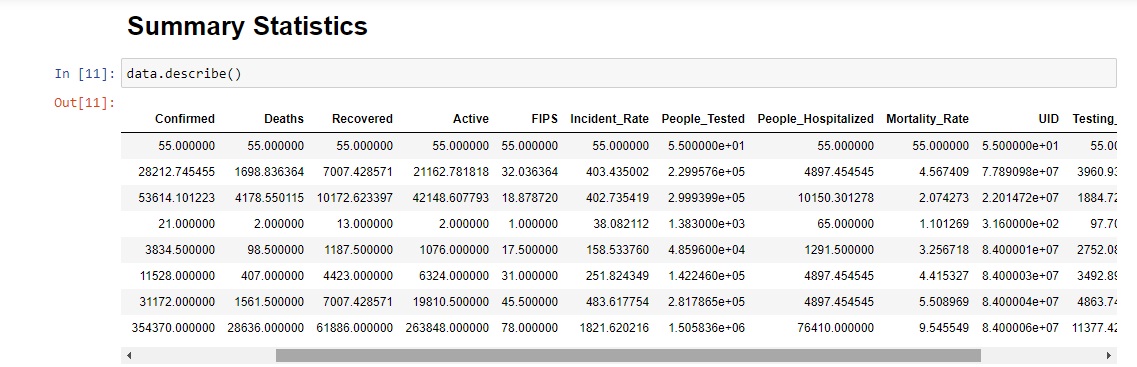
**Above we can observe that our dataset doesn’t contain any Null values in their columns now we can plot our data with different kind of plots, as we can observe that our data set belongs to different province state of US, so we can plot a choropleth map showing all the province state and the Data belongs to that. Before Plotting Choropleth Map we need to attach code corresponding to province state, you can do it using FuzzyWuzzy or manually like shown below.**



**Above you can observe I put the province code manually so that we can plot choropleth map, cause to plot choropleth map you must need to have long. And lat. Data with you or you can plot it with the CODE also. Like I did, you can witness that code column below.**



* **Summary Statistics:**



**Observation:**

1. We can observe in death column minimum no of death is 0 and maximum we can observe is 28636, we can also observe that mean is greater than median which means my data is Right Skewed, as we can also notice that the difference between 75% and max is quite large so we can say that some outliers might be there.

1. we can observe in confirmed column the minimum case is 0 and maximum confirmed case we ca notice is around 354370,we can also observe that the mean is greater than the median so my data is right skewed, there is also large difference in 75% and max so outliers must be there.

1. we can observe in the recovered case the minimum case is 13 and maximum case we can observed is 61886,the data is left skewed because mean is lesser than median, here also we can observe the difference in 75% and max we can say outliers must be present.

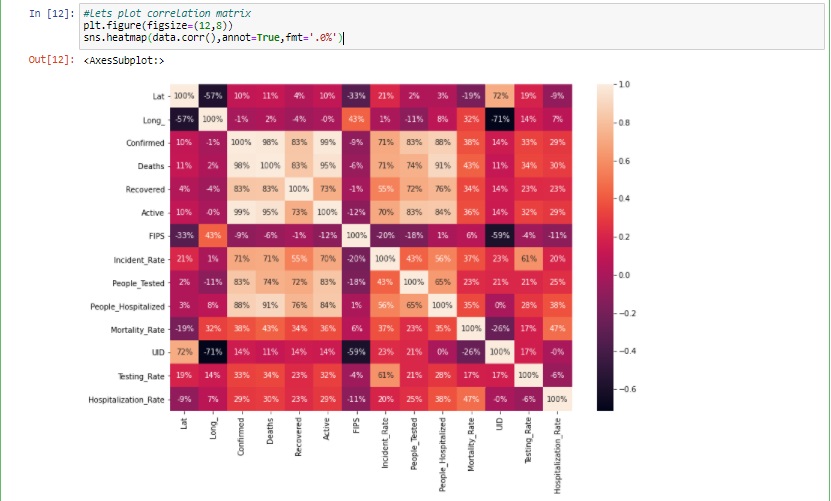
1. As we observed the Active case the minimum case is 0 and maximum case is around 263848, and the data is Right skewed, and the difference of 75% and Max is there so the outliers we can observe here.

1. Incident rate column we can observe that the minimum rate is 0 and maximum rate is 1821.62, the data is left skewed and we can observe outliers in this column.

1. As we can observe in the column of People Hospitalized the minimum people is 65 and the maximum we can observed is 76410,we can observe the difference of 75 and max so we ca say that the data contains outliers.

1. Mortality Rate we can observe with minimum rate of 1.10 and maximum rate of 9.54.

* **Correlation Matrix:**

****

**Observation:**

1. As you can get to notice above that the confirmed,Deaths,Recovered,Active,Incident\_rate,people\_tested,people hospitalized

Are strongly positively correlated with each other.

1. Mortality Rate, Testing Rate, Hospitalization Rate they are just positively correlated with each other.
2. Visualization**:**

* As you can see I am doing bivariate analysis between people tested for covid and out of that how many got confirmation of getting affected with this disease, using plotly plot.
* I am attaching link of an html file below you can navigate and observe the plotly plot, I could have attached the Image file but you won’t get that much of clarity, so I will be plotting only plotly plot you have to navigate to explore more in depth.



**Observation:**

1. Above we can observe how many people gets affected with covid case on basis of testing, as per the data we have with us, only in New York city out of 105 Million of testing results 354.37k confirmed cases as a covid positive, which is most highest in entire New York.



* I am attaching link of an html file link below you can navigate and observe the plotly plot of people hospitalized and total deaths.

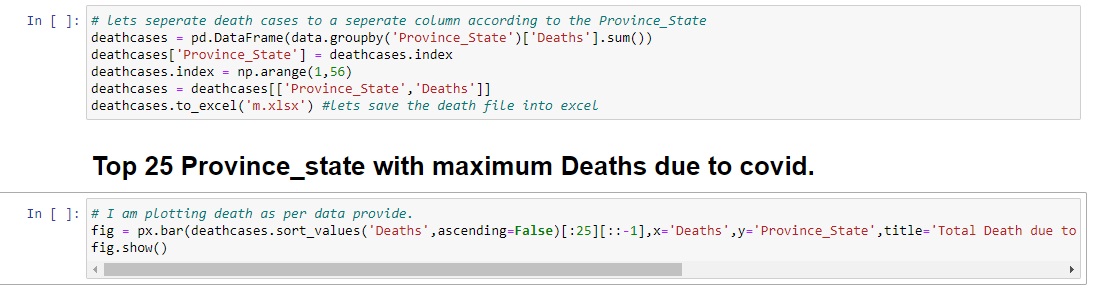


**Observation:**

1. As we have observed in previous plot that most number of confirmed cases in the city New York i.e. more number of people deaths are from the city New York, you can notice once you navigate the above HTML link, that out of 76.41k people Hospitalized 28.636k deaths we can witness only in the city New York.

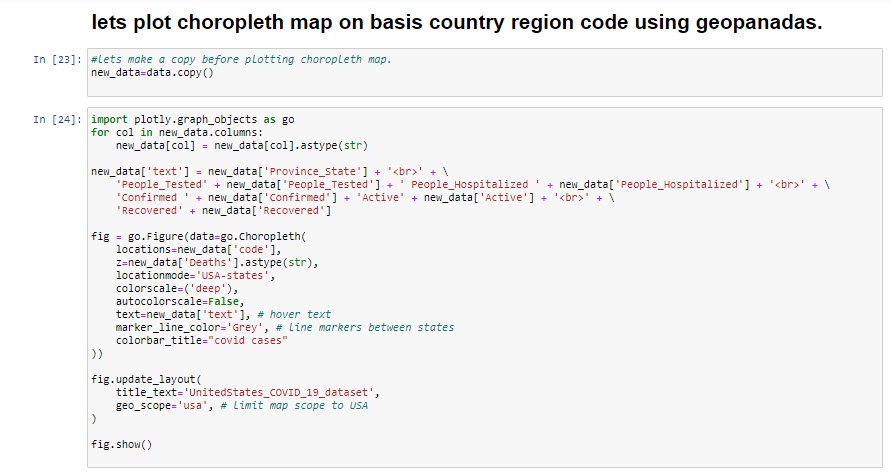
**Plotting Top 25 Province state with maximum Deaths due to covid-19.**

* I am going to plot top 25 province state of USA with highest number of death case noticed, same I am going to attach HTML link below.

****

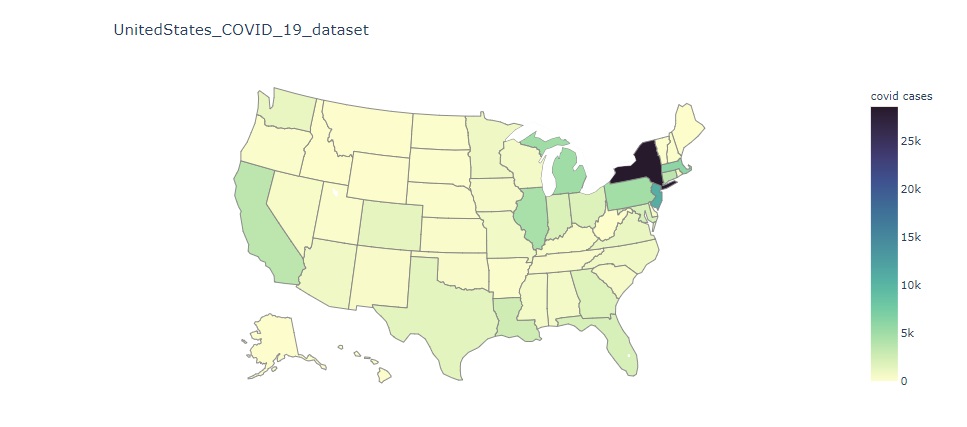
****

**Finally I am going to plot Choropleth map showing the data based on the different Province State.**

****

* Above code you can observe for choropleth map for USA states that is the reason I make a columns with code to identify the various region of USA.
* So I am attaching the link to navigate you to choropleth map of USA states.



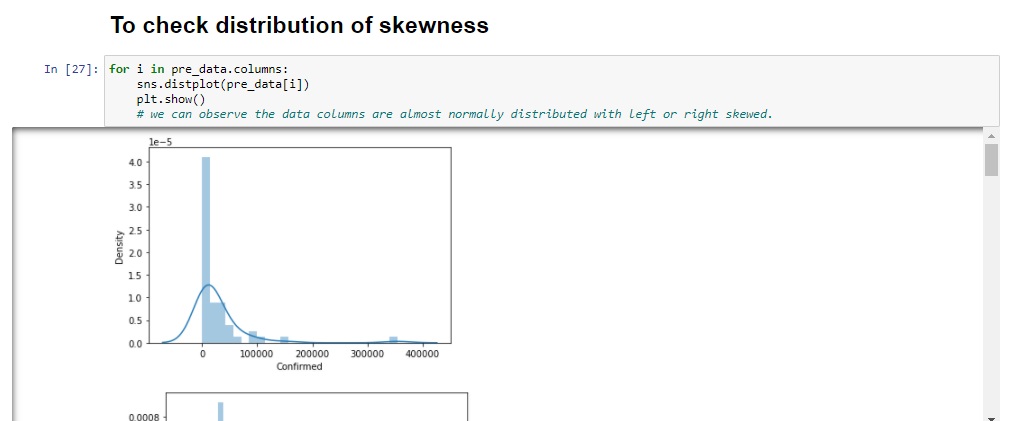


* If you navigate the link you are going to see something like this.

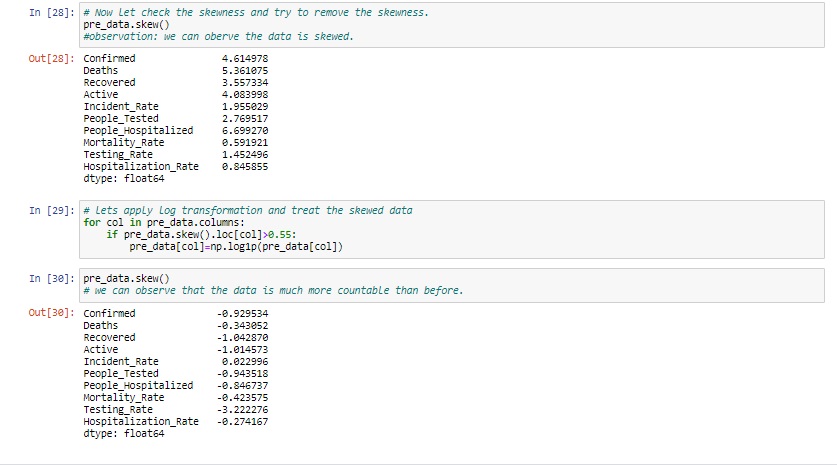
**Things to notice:**

I could have shown you all the plot but this is just a blog I can’t show you all if you want all the details and plot just navigate this [**Link**](https://github.com/khurram-DS/Covid_19_us_dataset) **, this will fetch you to my GitHub page.**

1. Preprocessing Pipeline**:**

* **Let’s check Distribution skewness**

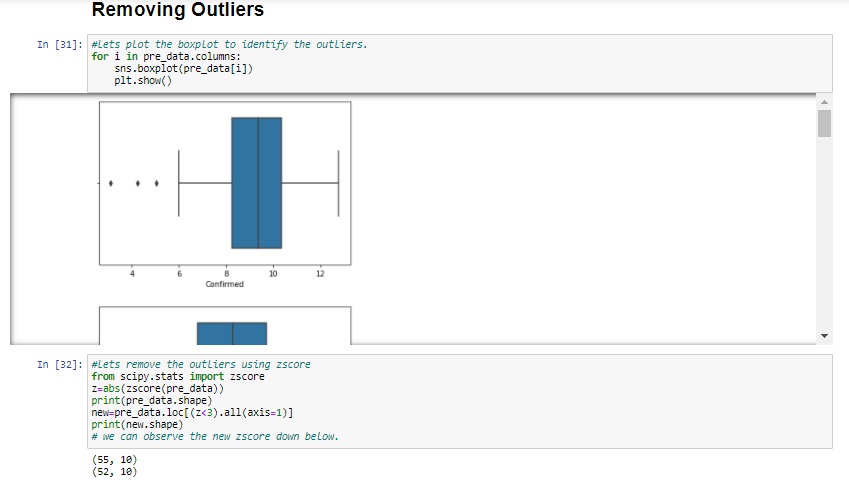
If you are going to scroll down you are going to observe the data columns are almost normally distributed with some of the columns with left or right skewed from axis.



You can notice above that my data is skewed let’s just treat our skewed data with log transformation, you can use boxcox, square root transformation also to treat the skewness of data I am using log transformation you can see above.

You can notice in out [30] that my data is much more skewed now.

* **Removing outliers:**



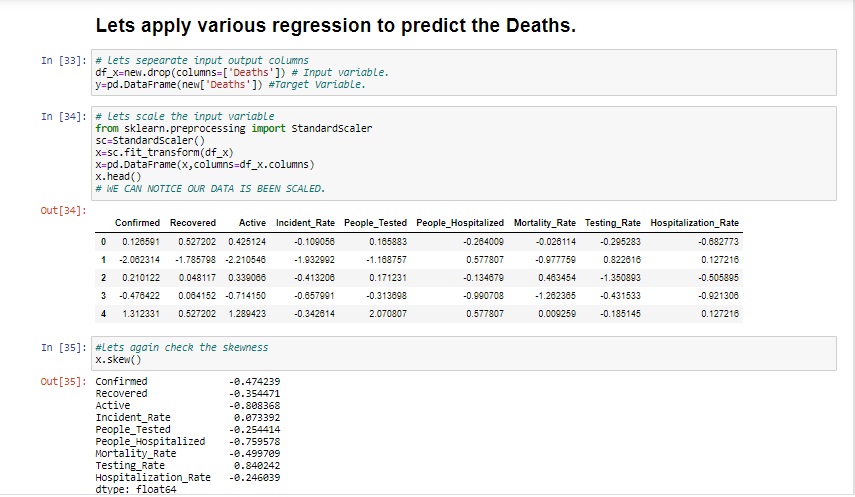
**Observation:**

We can observe that my we removed the outliers using zscore, you can notice that previously my data was 55 row with 10 columns but after applying zscore the data is decreased to 52 row and 10 columns.

This complete all the preprocessing pipelines.

1. Building Machine Learning model**:**

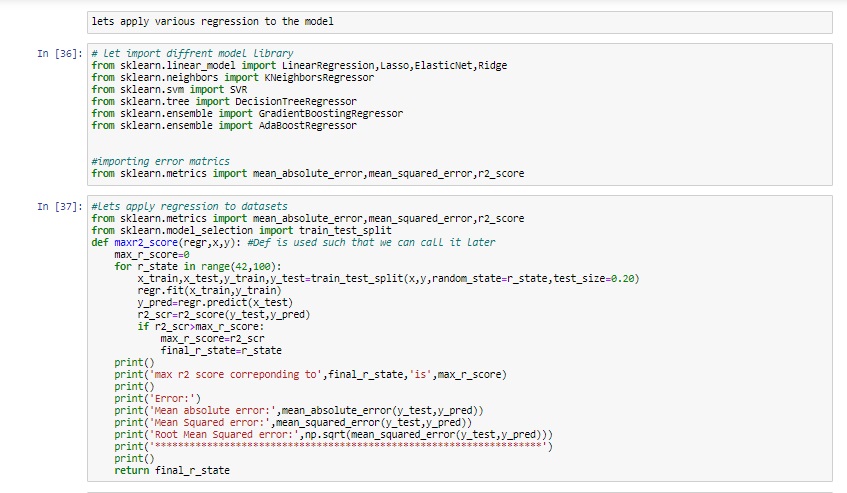
* **As we have done the preprocessing let’s start with building model.**



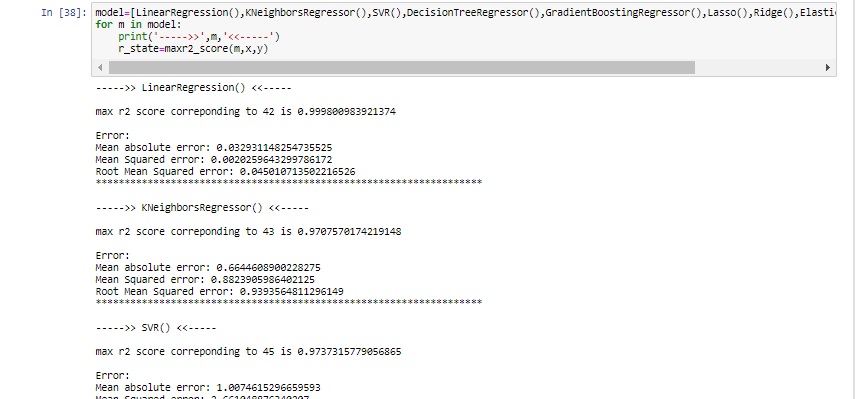
Let’s separate the Input and output first, after splitting the Input and output will scale the input df\_x with **standard scaler,** the idea behind **StandardScaler** is that it will transform your data such that its distribution will have a mean value 0 and **standard** deviation of 1. In case of multivariate data.

So after scaling the data will check the skewness again and see the limit, to be extra confirm that my data is well within the skewness limit.

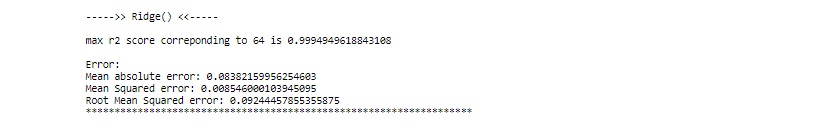
* **Now we are going import various regression algorithms.**



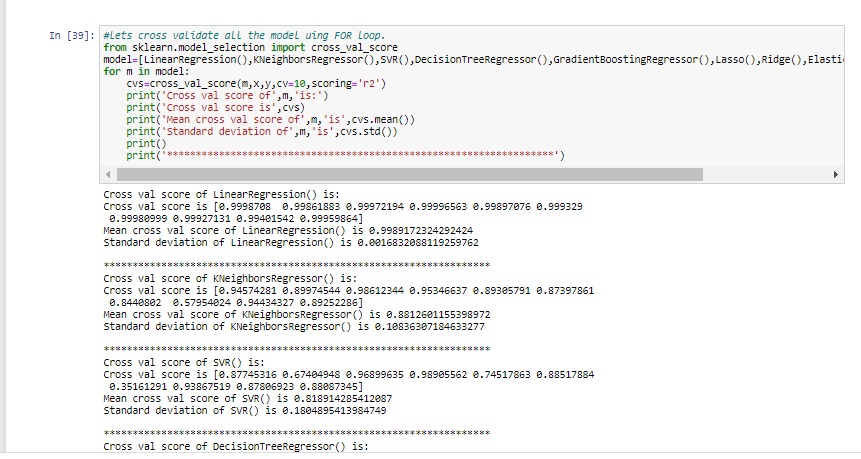
Above I am using the for loop which help me to provide the R2 score at each random state and for the best state where R2 score is maximum we are going to set the parameters.

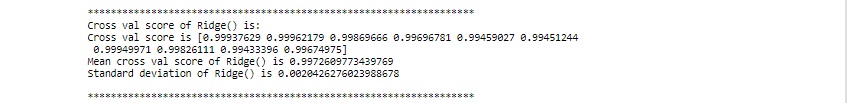


Folks just call the algorithms you must be so eagerly waiting to see the results.



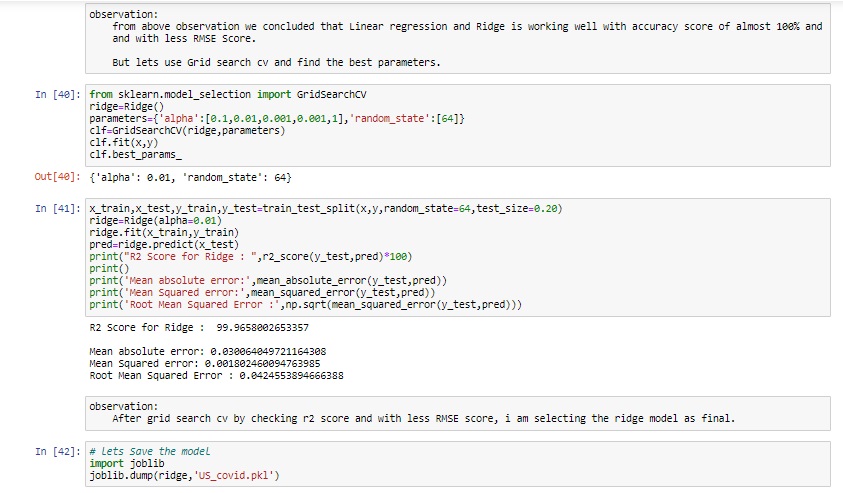
Folks let’s just check the cross validation score.





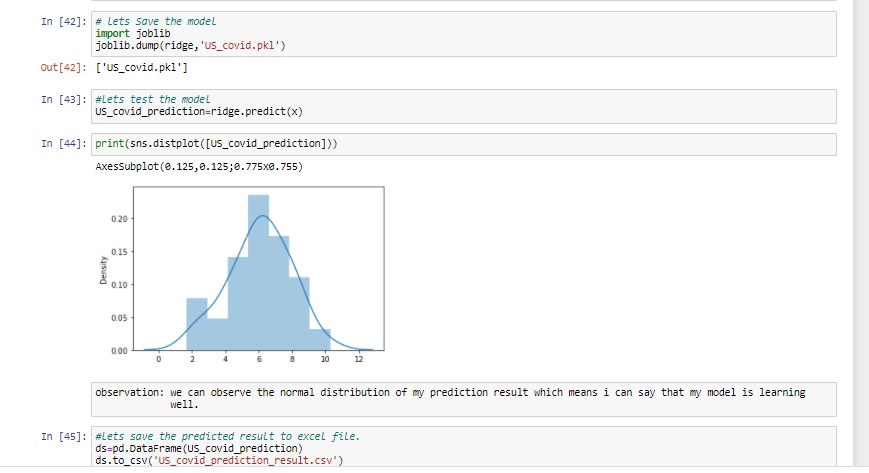
**Observation:**

From above observation we concluded that linear regression and Ridge is working well with accuracy score of almost 100% and Cross validation score of 99%

* But let’s use Grid search CV and find the best parameters.

After applying grid search CV to Ridge () regression we can say that the ridge is working well with accuracy score of 100% with less RMSE score of 0.042, as you know the less RMSE score gives the best model.

After grid search CV I am saving the model with Ridge () regression, and saving the pkl file by importing joblib.



As we have save the model let’s just predict the same input using ridge.predict(x)

And just plot the distplot of my predicted result, you can see the predicted result is normally distributed which means our data is well processed, later you can save the data to excel file.

1. Concluding remarks**:**

* After getting insight of the data we saw various covid-19 ups and down in USA.
* I had done this prediction by taking Deaths as an output variable which is continuity in nature so that why I’m using the regression technique.
* We observe that the maximum cases has been observed in New York City.
* We also observed that the almost 50% of people died if he or she admitted to the hospital.
* As we observed that the 30% of people is affected if he or she gone under testing.
* We also observed that almost 70% of people recovered if he or she has confirmed as covid positive.
* Overall it was great to visualize the covid -19 data set of US using Plotly,choropleth.
* While calculating the best random state the 64 is best state which providing the highest R2 score value for this model.
* I used Grid search cv to get the best params with ridge and I saved the model as ridge algorithms
* I plotted the distplot of predicted result and I found that the distplot of predicted result were quite normally distributed.
* Further correction is possible because I am not bound say that this is enough, lots of people have lots of Idea to visualize and implement this dataset.

1. Refrences**:**

* [**https://www.datatrainedacademy.com**](https://www.datatrainedacademy.com)
* [**https://medium.com**](https://medium.com)
* [**https://github.com/khurram-DS/Covid\_19\_us\_dataset**](https://github.com/khurram-DS/Covid_19_us_dataset)
* [**https://www.kaggle.com**](https://www.kaggle.com)
* [**https://towardsdatascience.com**](https://towardsdatascience.com)