**The Accidental Bot**

**About Blog**

My name is **Bilam Roy**. I am an aspiring Data Scientist and have 2.5 years of experience in IT under Banking Sector. I completed my B.Tech in Electronics and Instrumentation Engineering at the West Bengal University of Technology in West Bengal. Here you'll find resources about Data Science, Machine Learning and ML models to get an idea of what this Data Science is all about.

**Connect Over:**

**LinkedIn:** <https://www.linkedin.com/in/bilam-roy-022913187>

**Email:** [bilam95roy@gmail.com](mailto:bilam95roy@gmail.com)

**Welcome Aspiring Data Scientist!!**

It is never too late to learn new things, so if you are thinking of switching your job profile from any other field to become a Data Scientist then you are already a step closer to your goal.

Let us take some questionnaires for a better understanding:

1. **So what is basically Data Science and what Data Scientist do?**

-In simple words, Data Science is the extraction of knowledge and different insights from the raw data using scientific methods, processes, algorithms and systems.

Hence, Data Scientists are responsible for discovering insights from massive amounts of structured and unstructured data to help shape or meet specific business needs and goals.

1. **What are the objectives that a Data Scientist have to deliver?**

- Data science generally has a five-stage lifecycle that consists of:

**Data Acquiring**: Data acquisition, data entry, signal reception, data extraction.

**Data Wrangling**: Data warehousing, data cleansing, data staging, data processing, data architecture.

**Processing**: Data mining, clustering /classification, data modeling, data summarization.

**Reporting**: Data reporting, data visualization, business intelligence, decision making.

**Analyzing**: Data Exploratory/ confirmatory, predictive analysis, regression, text mining, qualitative analysis.

1. **What is Machine Learning?**

- Machine Learning is the process of teaching a computer system how to make accurate predictions when raw data is fed.

The key difference from traditional computer software is that a human developer hasn't written code; instead a machine-learning model has been taught how to reliably discriminate on a large amount of data.

1. **What are the main types of machine learning?**

- Machine learning is generally split into two main categories: supervised and unsupervised learning.

1. **How to evaluate machine-learning models?**

* Evaluating a model is a core part of building an effective machine learning model.
* There are several evaluation metrics, like confusion matrix, cross-validation, AUC-ROC curve, etc.
* Different evaluation metrics are used for different kinds of problems.

Once training of the model is complete, the model is evaluated using the remaining data that wasn't used during training, helping to gauge its real-world performance.

1. **What is machine learning used for?**

- Machine learning systems are used all around us.

For Example:

* Machine-learning systems are used to recommend which product we might want to buy next on Amazon or video we want to watch on Netflix.
* Every Google search uses multiple machine-learning systems, to understand our query to personalizing our results.

**Let us understand how to make Machine Learning Models**

**Project I**: **Bias correction of numerical prediction model temperature forecast Data Set**

**Dataset:** <https://archive.ics.uci.edu/ml/datasets/Bias+correction+of+numerical+prediction+model+temperature+forecast>

**Data Set Information:** This data is for the purpose of bias correction of next-day maximum and minimum air temperatures forecast of the LDAPS model operated by the Korea Meteorological Administration over Seoul, South Korea. This data consists of summer data from 2013 to 2017. The input data is largely composed of the LDAPS model's next-day forecast data, in-situ maximum and minimum temperatures of present-day and geographic auxiliary variables. There are two outputs (i.e. next-day maximum and minimum air temperatures) in this data. Hind cast validation was conducted for the period from 2015 to 2017.

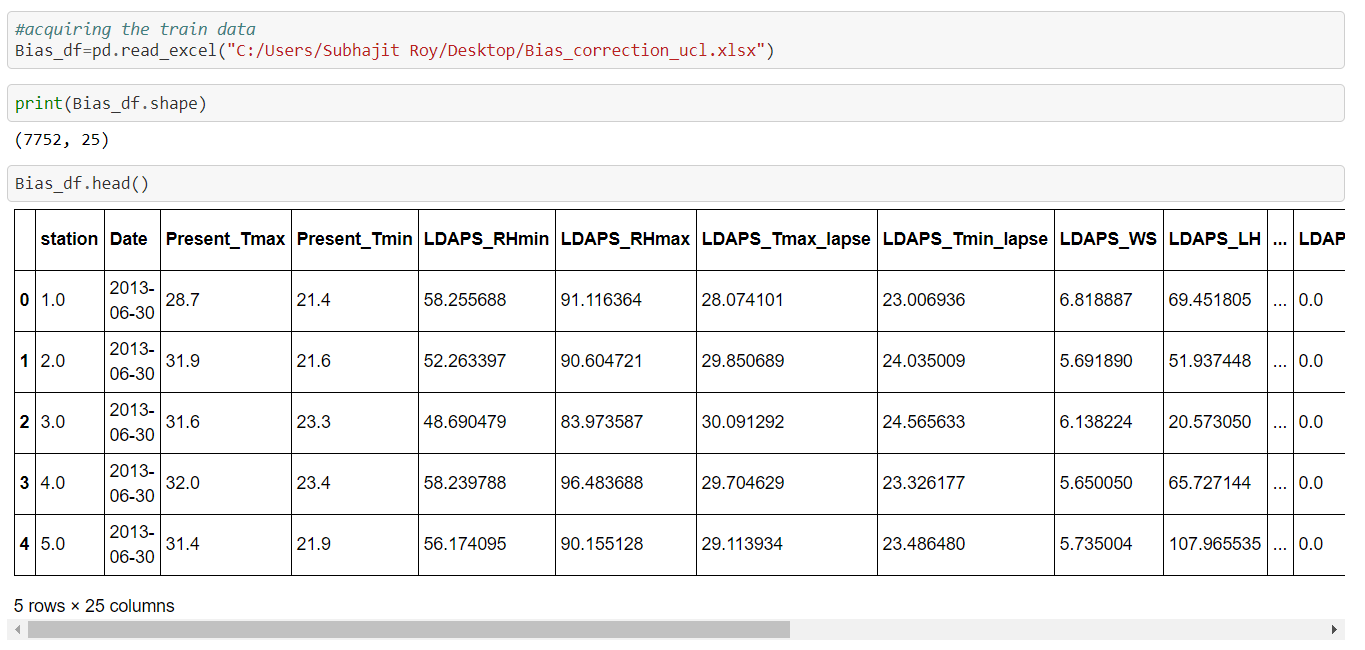
**Platform:** Jupyter Notebook (Anaconda FrameWork)

**Language:** Python

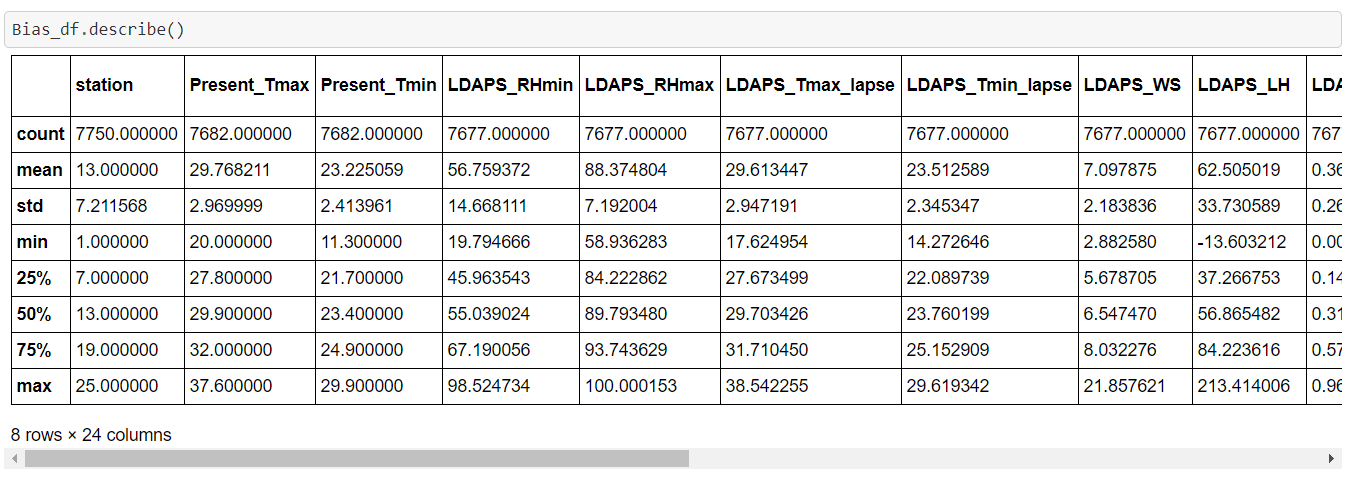
**Step 1:** Loading all the basic libraries for the manupulation of the data.



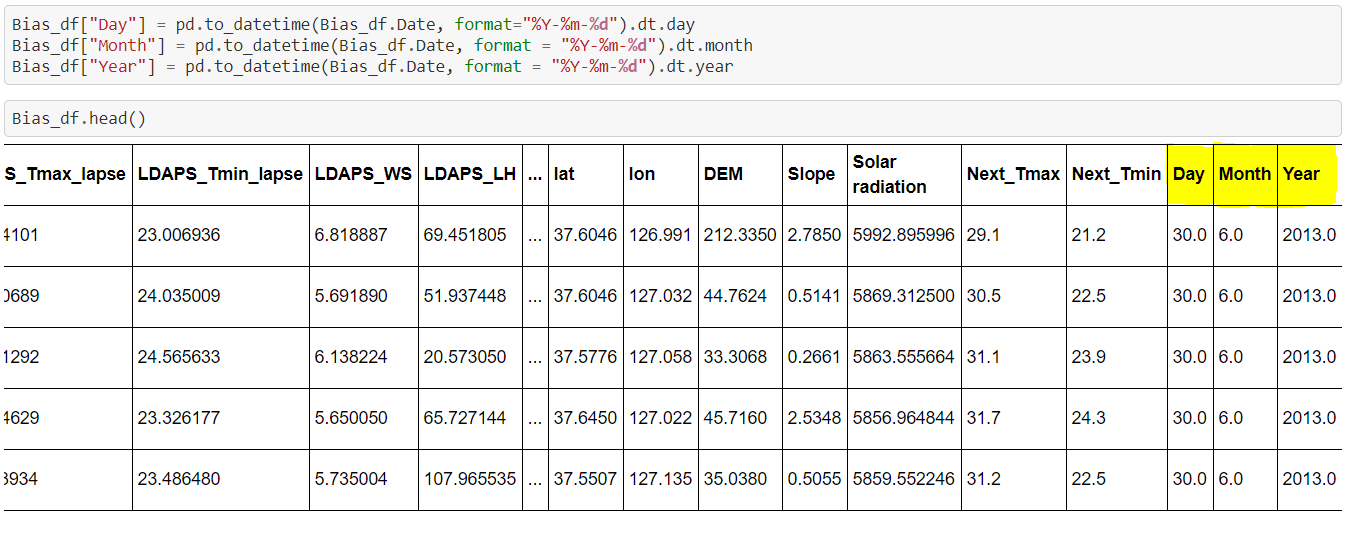
**Step 2:** Loading Dataset and validating the data.



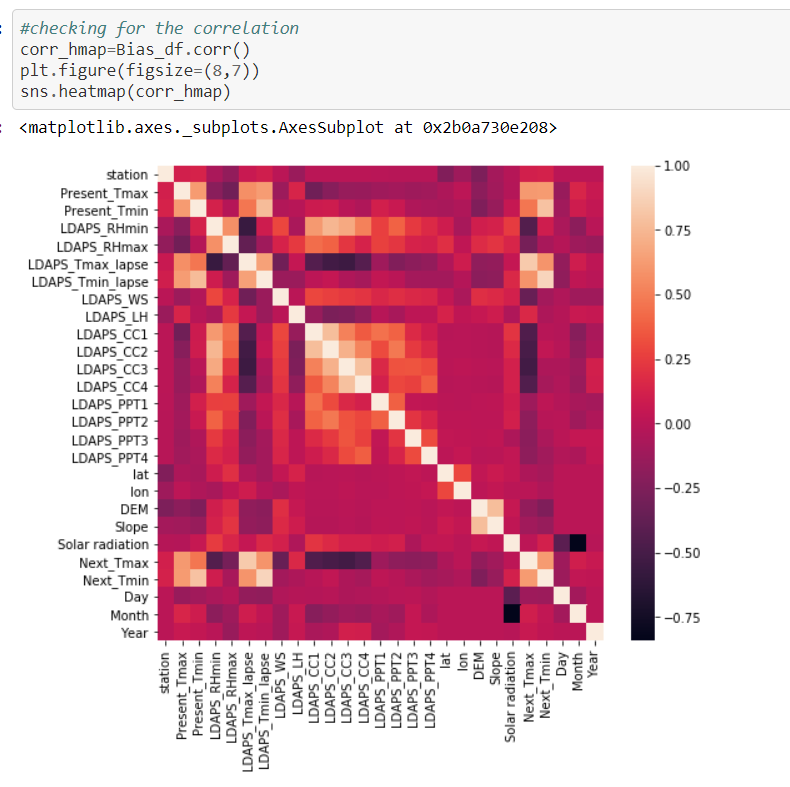
**Step 3:** Statistical Report Generation



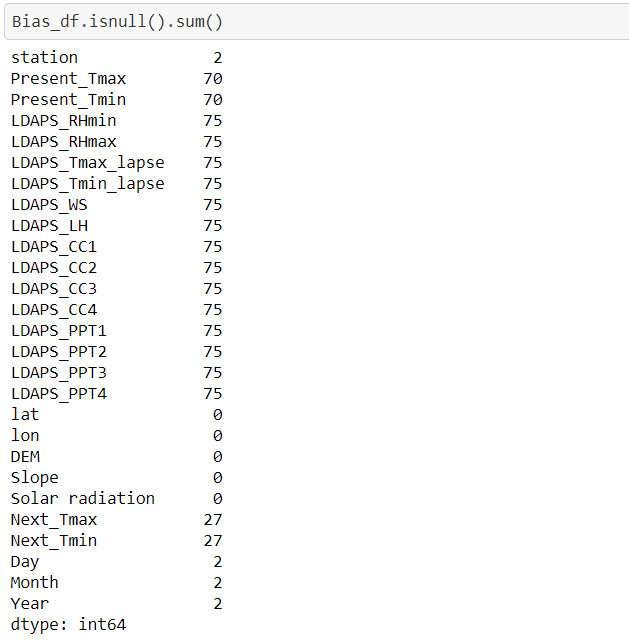
**Step 4:** Manipulating the dataset for better performance, hence converting all the fields into numerical data.



**Step 5:** Data Visualization, checking how the fields are correlated with each other.(Multivariate Plot)



**Step 6:** Checking if any Null values are present in the dataset then imputing them for better performance.

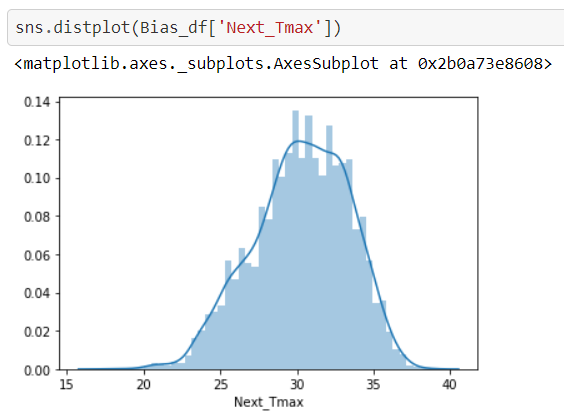


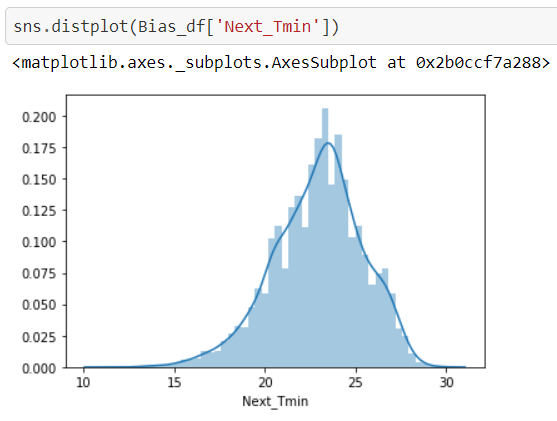


Dropping the NA values those are not required

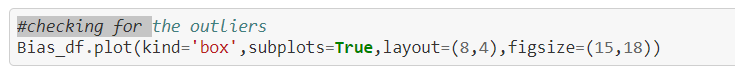


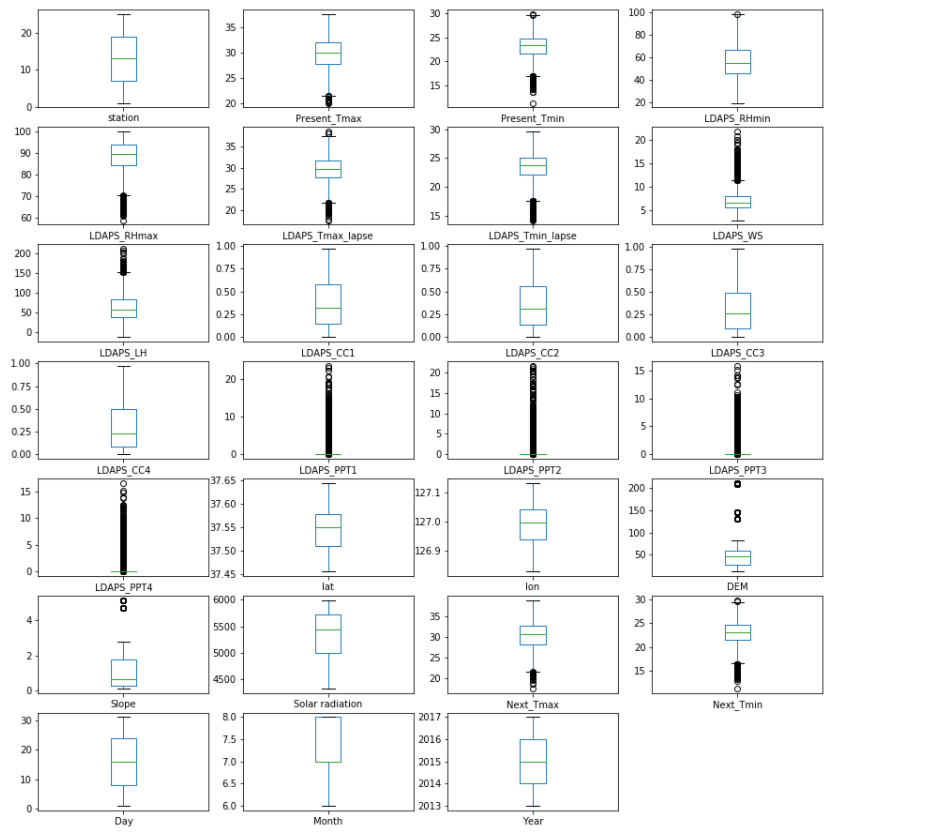
**Step 7:** Plotting univariate graph (distribution plot) for density estimate and histogram with bin size determined automatically with respect to the two Target vales Next\_Tmax and Next\_Tmin



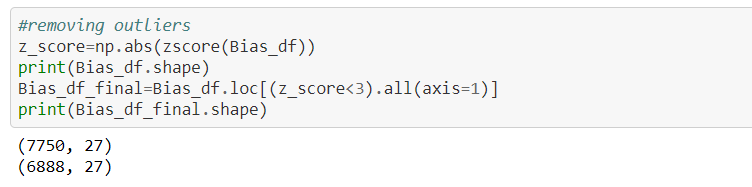


**Step 8:** Checking if any outliers are present or not, if yes then eliminate.

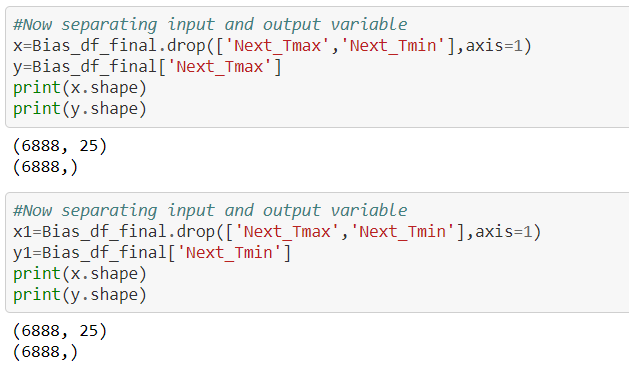




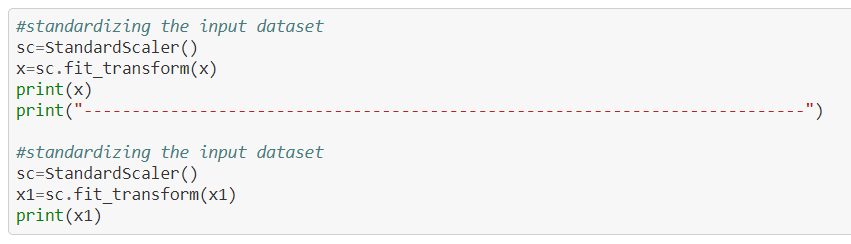
Removing all the data that have z-score greater than 3



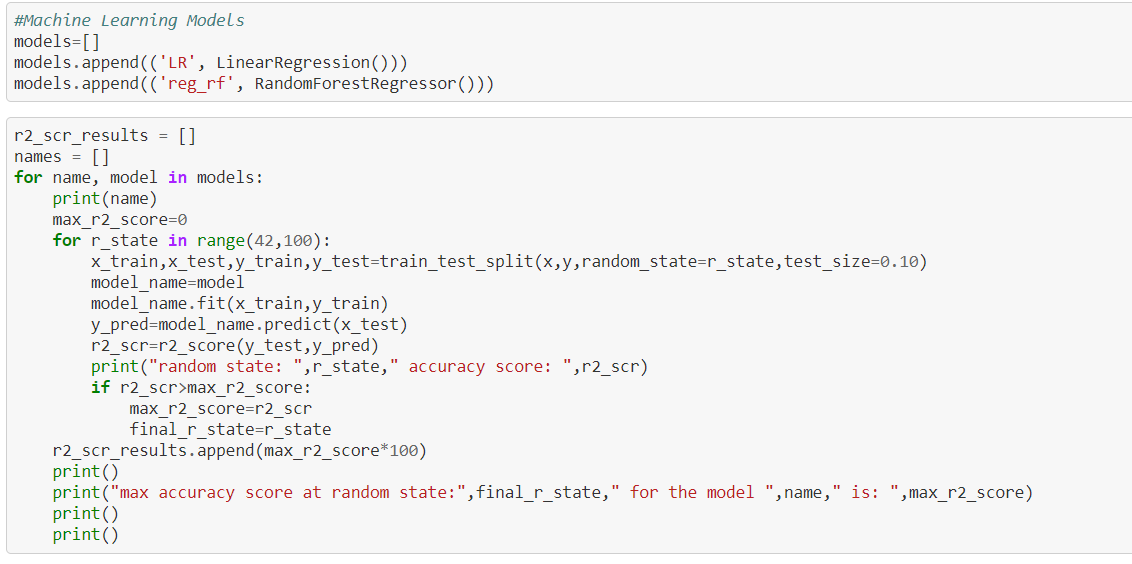
**Step 9:** Separating the input and output data



**Step 10:** Standardize the input features by removing the mean and scaling to unit variance



**Step 11:** Training the Models.



Linear Regression is used when variables should be measured at the continuous level (i.e., they are either interval or ratio variables).

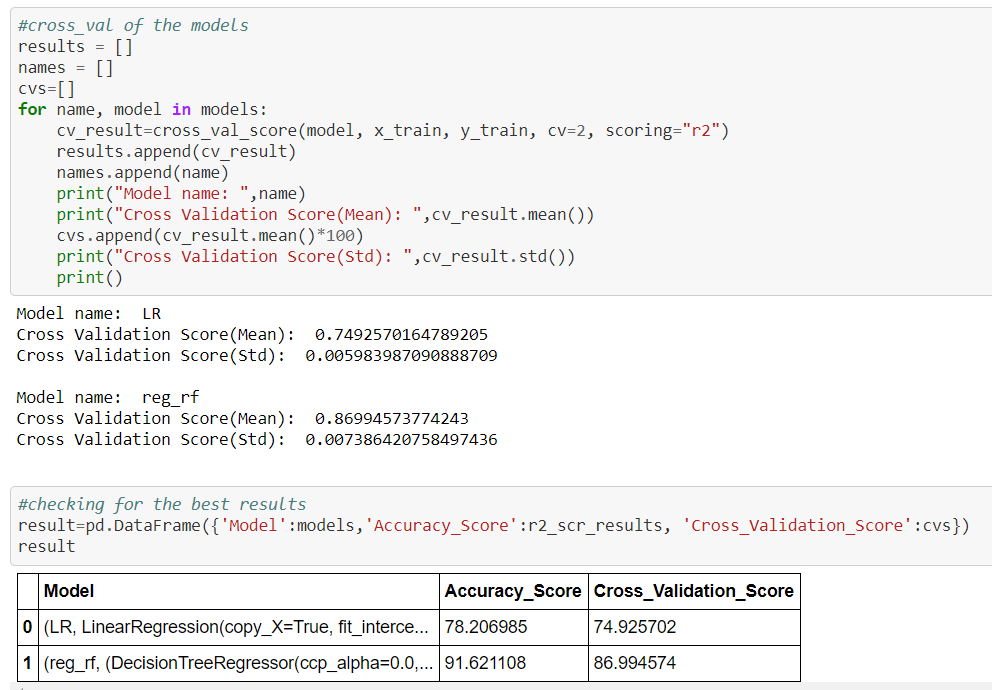
Random forest is a Supervised Learning algorithm which uses ensemble learning method for classification and regression.

An Ensemble method is a technique that combines the predictions from multiple machine learning algorithms together to make more accurate predictions than any individual model. A model comprised of many models is called an Ensemble model.

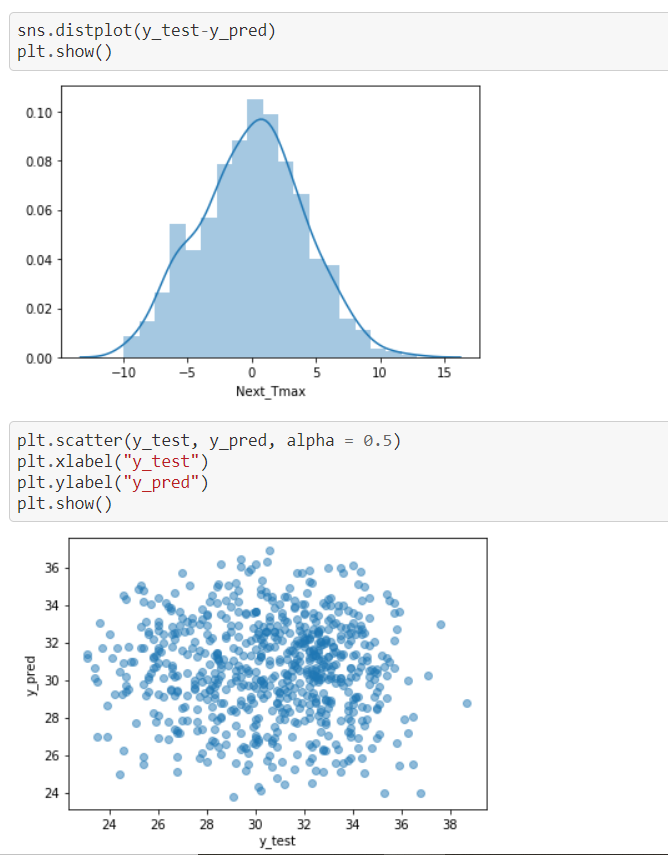
Random forest operates by constructing a multitude of decision trees at training time and outputting the class that is the mode of the classes (classification) or mean prediction (regression) of the individual trees.

A random forest is a meta-estimator (i.e. it combines the result of multiple predictions) which aggregates many decision trees.

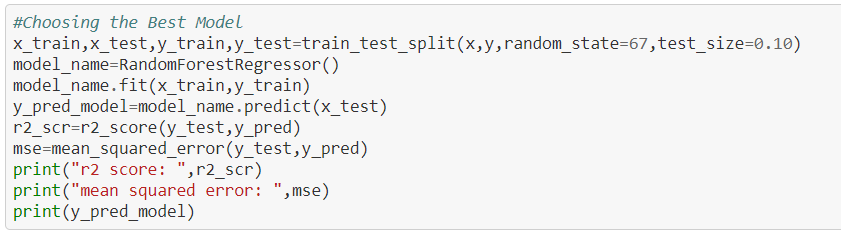
**Step 12:** Evaluating the Results.



**Step 13:** Evaluating the Results graphically between the test and predicted values.



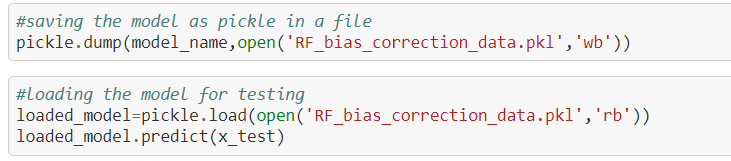
**Step 14:** Choosing the Best Model



**r2  Score:** Best possible score is 1.0 and it can be negative (because the model can be arbitrarily worse). A constant model that always predicts the expected value of y, disregarding the input features, would get a R^2 score of 0.0.

**MSE:** It is the sum, over all the data points, of the square of the difference between the predicted and actual target variables, divided by the number of data points.

**Step 15:** Saving/Pickling the Best Model for future use



Note: For more detailed information visit

https://github.com/bilamroy/datascientist/blob/main/Project%2017-Bias%20correction.ipynb

**Project II**: **Country wide COVID-19 Datasets**

**Dataset:** <https://github.com/dsrscientist/COVID_19_Datasets/blob/master/worldometer_snapshots_April18_to_May18.csv>

<https://github.com/dsrscientist/COVID_19_Datasets/blob/master/population_structure_by_age_per_contry.csv>

**Data Set Information:** The dataset contains data about the numbers of tests, cases, deaths, serious/critical cases, active cases and recovered cases in each country for every day since April 18, and also contains the population of each country to calculate per-capita penetration of the virus

Additionally, an auxiliary table with information about the fraction of the general population at different age groups for every country is added (taken from Wikipedia). This is specifically relevant since COVID-19 death rate is very much age dependent.

My goal will be to use the additional data about the number of tests performed in each country to estimate the true death rate of COVID-19 and examine the relationship with age .(Consider the both files for understanding the scenario.)

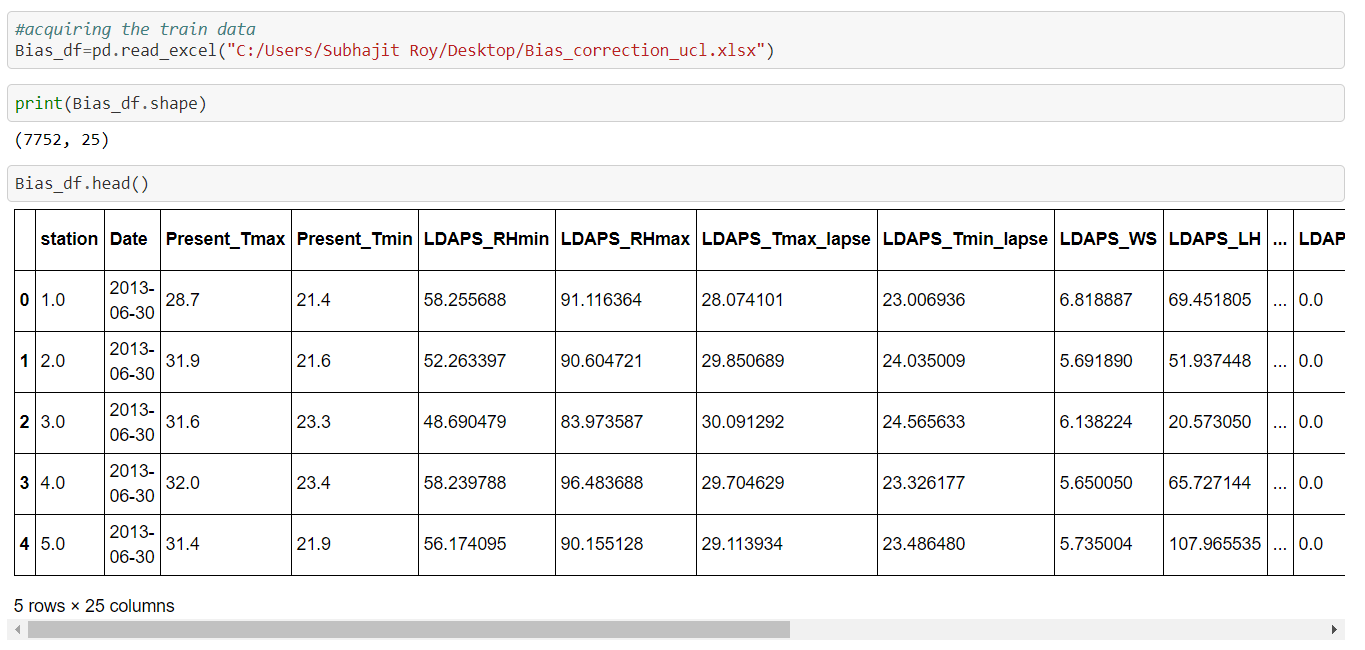
**Platform:** Jupyter Notebook (Anaconda FrameWork)

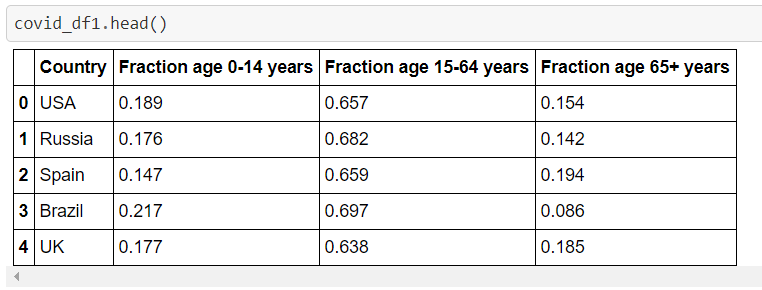
**Language:** Python

**Step 1:** Loading all the basic libraries for the manupulation of the data.

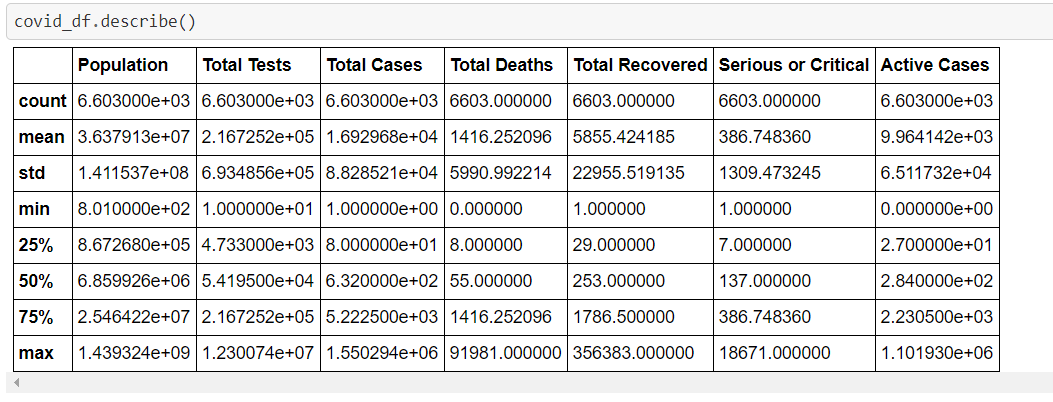


**Step 2:** Loading Dataset and validating the data.

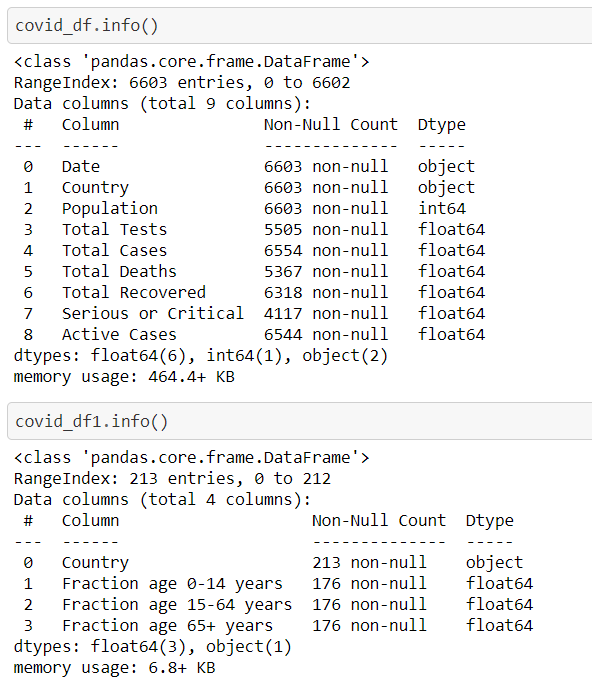




**Step 3:** Statistical Report Generation



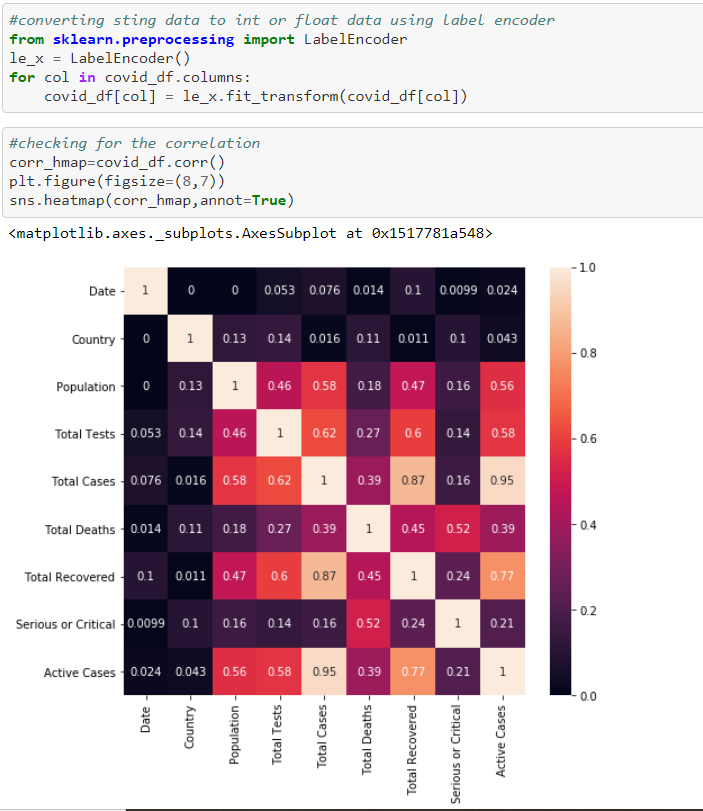
**Step 4:** Gathering information about the dataset



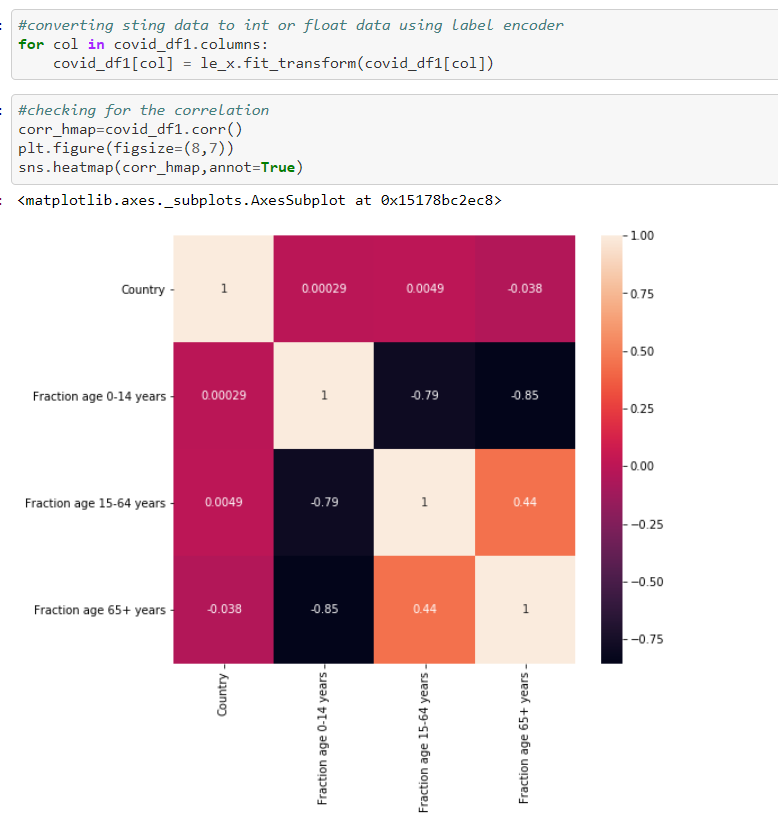
**Step 5:** Data Visualization, checking how the fields are correlated with each other.(Multivariate Plot)

Label Encoder is used to covert string data to int or float

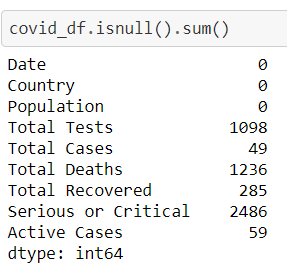
Dataset 1:

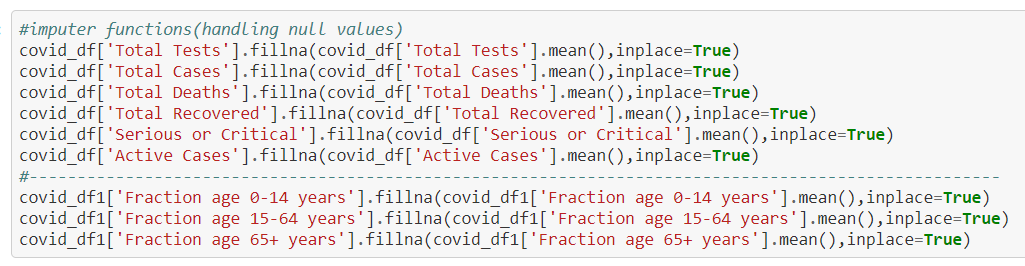


Dataset 2:

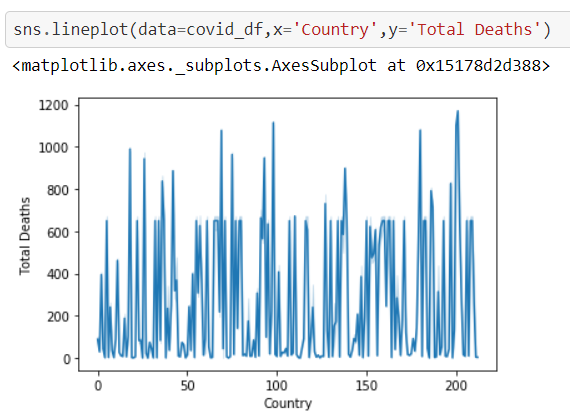


**Step 6:** Checking if any Null values are present in the dataset then imputing them for better performance.

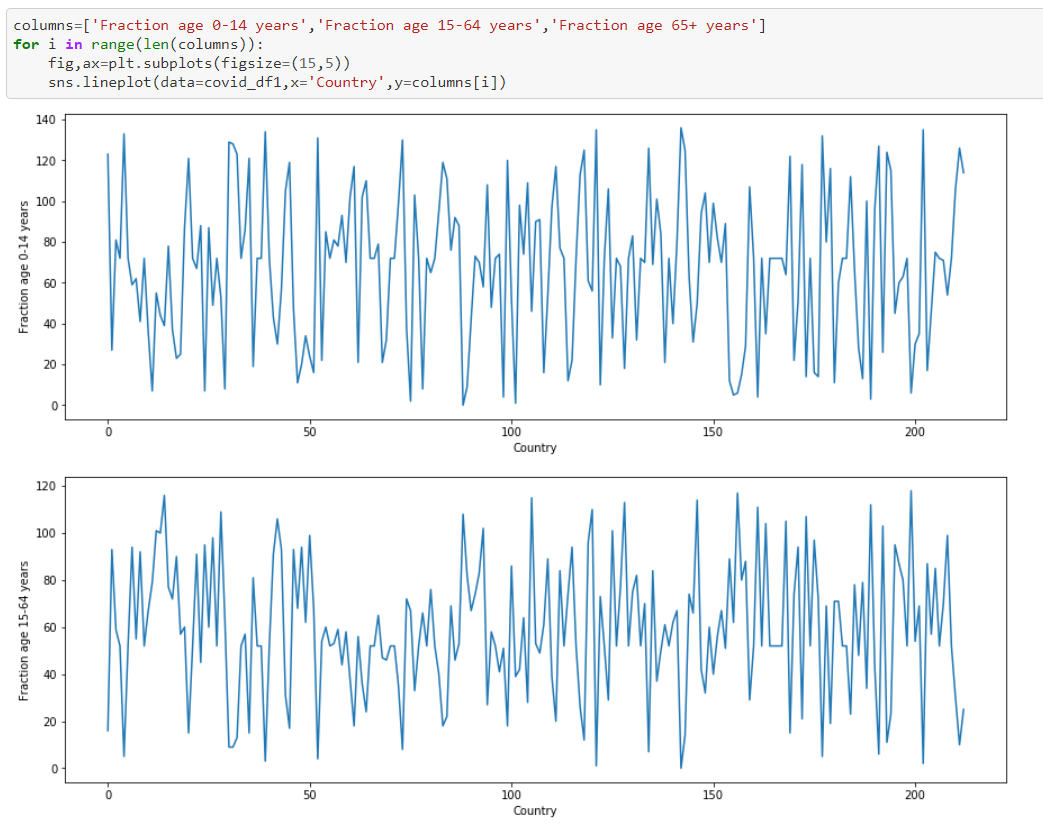


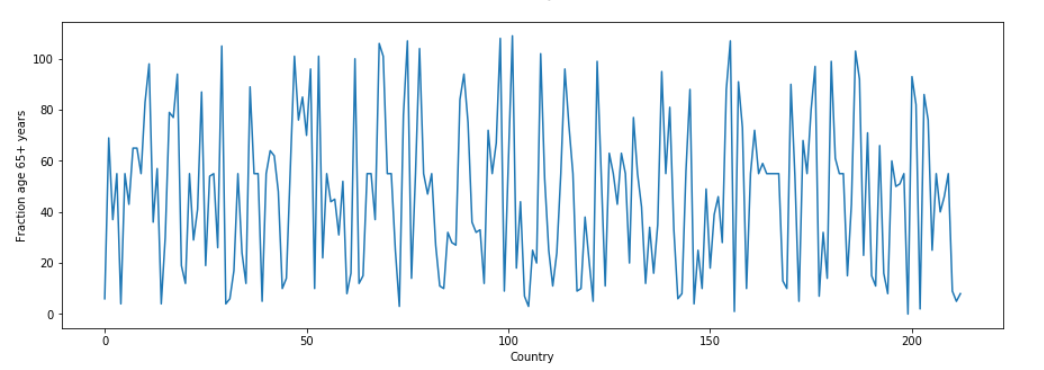


**Step 7:** Plotting univariate graph(line plot) Total Deaths vs Country graph

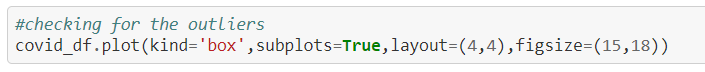


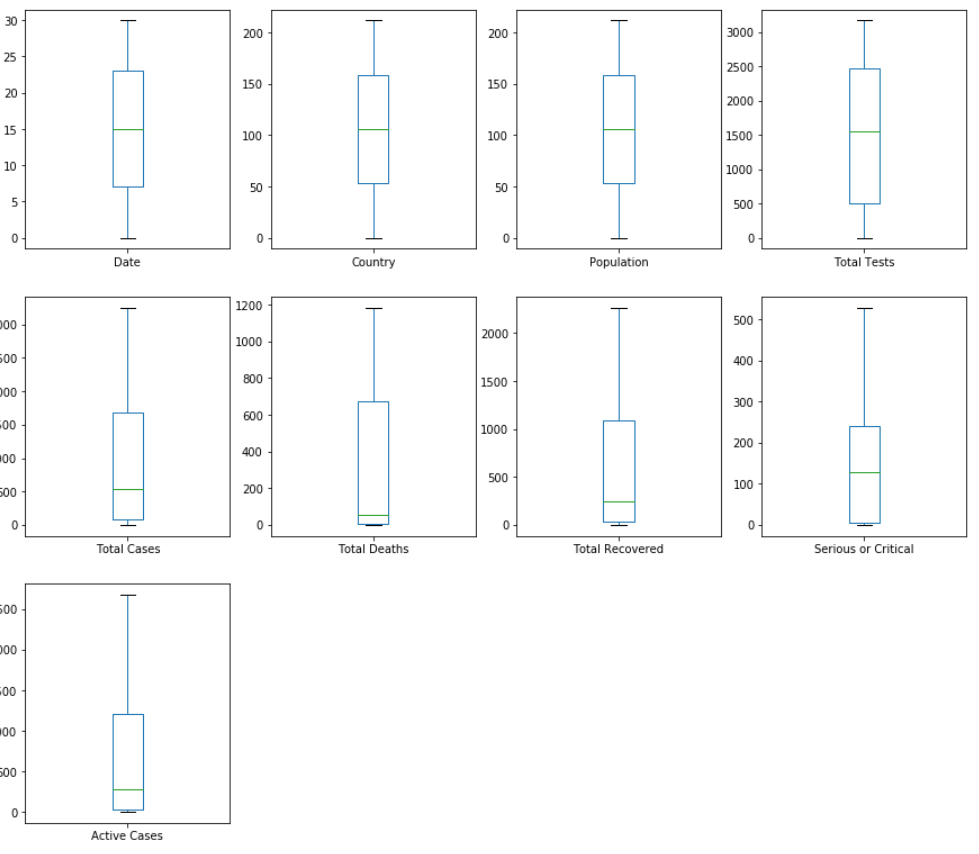
Validating different age group affected vs country graph in dataset 2



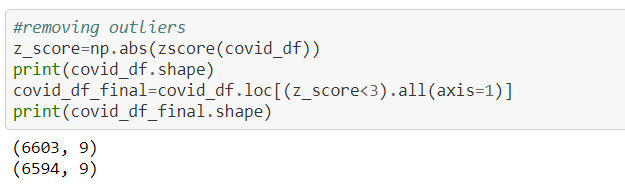


**Step 8:** Checking if any outliers are present or not, if yes then eliminate.

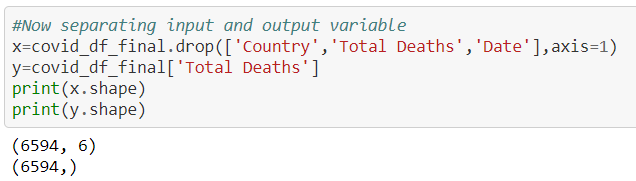




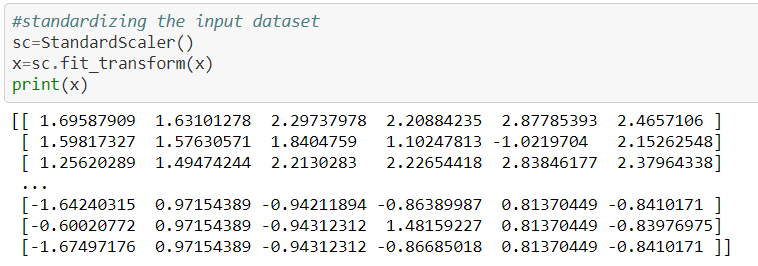
Removing all the data that have z-score greater than 3



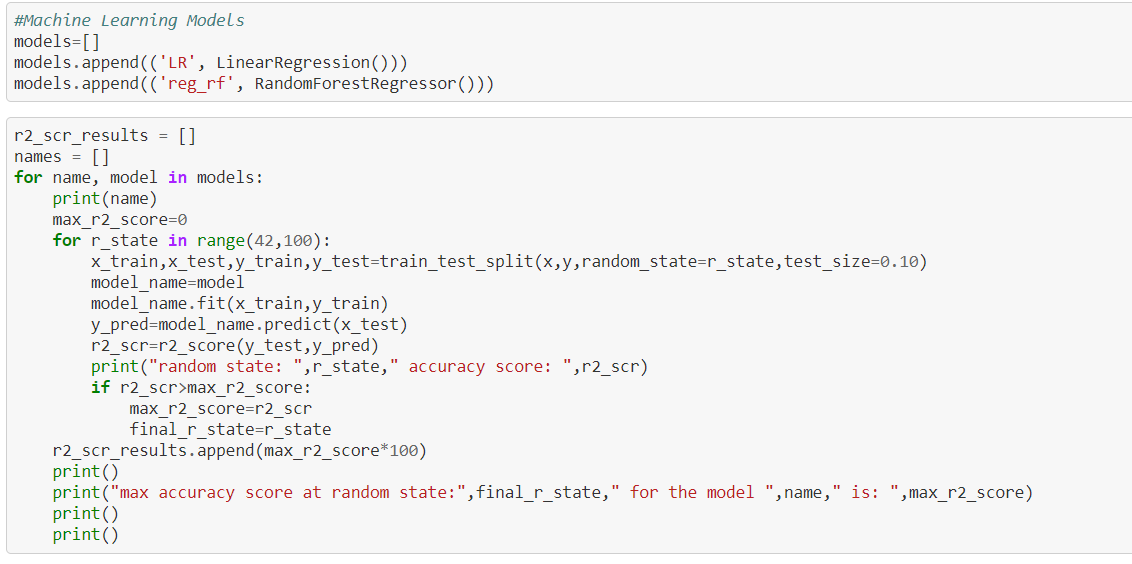
**Step 9:** Separating the input and output data



**Step 10:** Standardize the input features by removing the mean and scaling to unit variance



**Step 11:** Training the Models.



Linear Regression is used when variables should be measured at the continuous level (i.e., they are either interval or ratio variables).

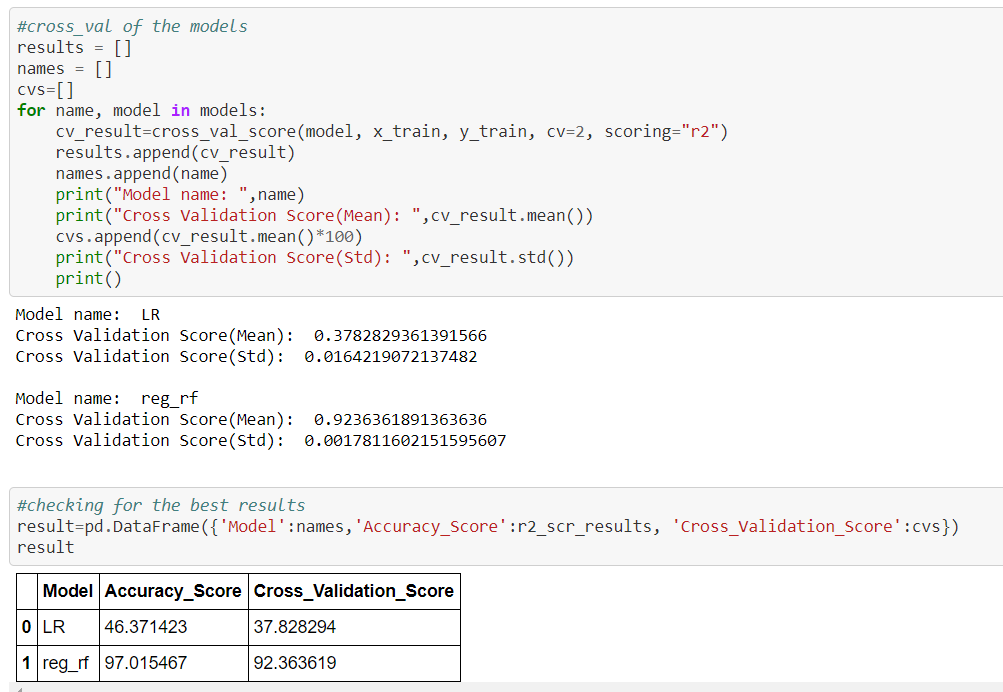
Random forest is a Supervised Learning algorithm which uses ensemble learning method for classification and regression.

An Ensemble method is a technique that combines the predictions from multiple machine learning algorithms together to make more accurate predictions than any individual model. A model comprised of many models is called an Ensemble model.

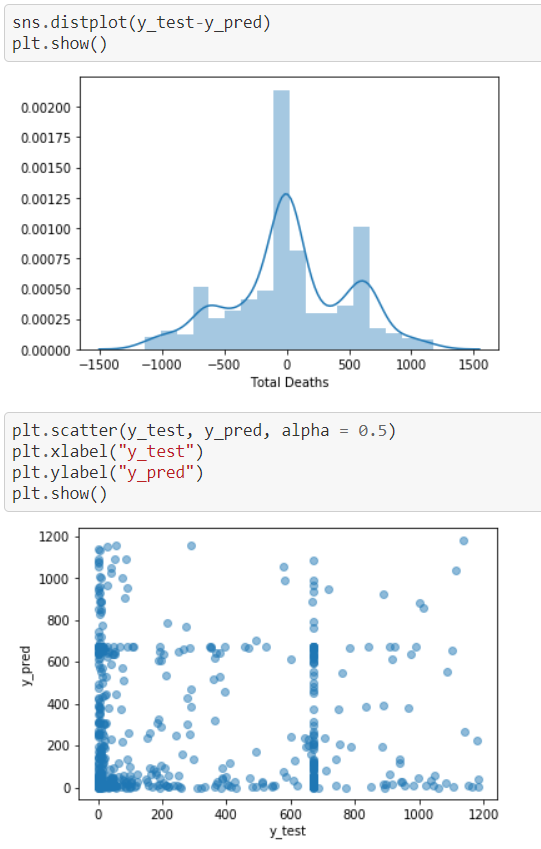
Random forest operates by constructing a multitude of decision trees at training time and outputting the class that is the mode of the classes (classification) or mean prediction (regression) of the individual trees.

A random forest is a meta-estimator (i.e. it combines the result of multiple predictions) which aggregates many decision trees.

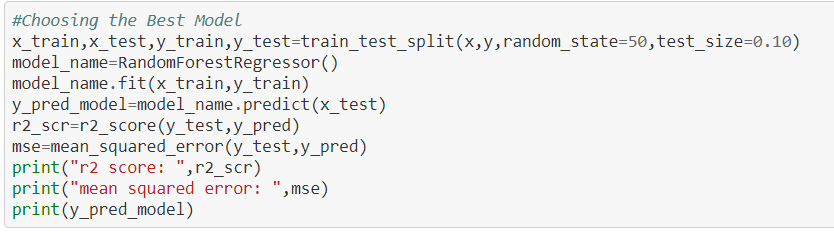
**Step 12:** Evaluating the Results.



**Step 13:** Evaluating the Results graphically between the test and predicted values.



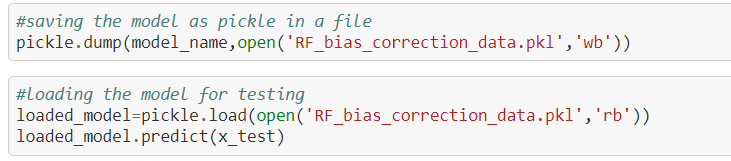
**Step 14:** Choosing the Best Model



**r2  Score:** Best possible score is 1.0 and it can be negative (because the model can be arbitrarily worse). A constant model that always predicts the expected value of y, disregarding the input features, would get a R^2 score of 0.0.

**MSE:** It is the sum, over all the data points, of the square of the difference between the predicted and actual target variables, divided by the number of data points.

**Step 15:** Saving/Pickling the Best Model for future use



Note: For more detailed information visit

https://github.com/bilamroy/datascientist/blob/main/Project%2018-Country\_wide\_COVID%2019\_Datasets%20.ipynb