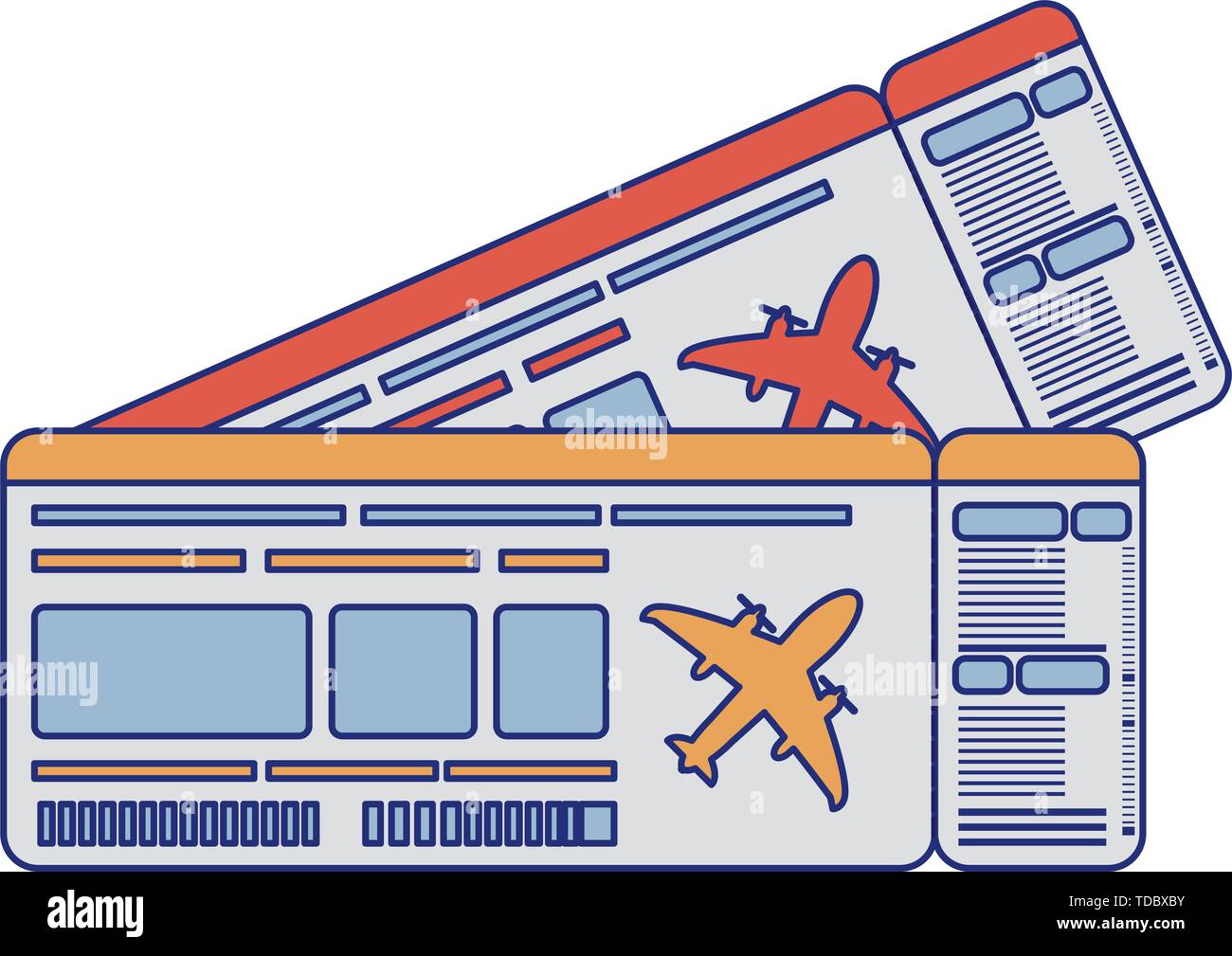
Blog on Flight Price Prediction

The tourism industry is changing fast and this is attracting a lot more travelers each year. The airline industry is considered as one of the most sophisticated industry in using complex pricing strategies. Now-a-days flight prices are quite unpredictable. The ticket prices change frequently. Customers are seeking to get the lowest price for their ticket, while airline companies are trying to keep their overall revenue as high as possible. Using technology it is actually possible to reduce the uncertainty of flight prices. So here we will be predicting the flight prices using efficient machine learning techniques..



**Problem Statement-**

Flight ticket prices can be something hard to guess, today we might see a price, check out the price of the same flight tomorrow, it will be a different story. We might have often heard travellers saying that flight ticket prices are so unpredictable. Here you will be provided with prices of flight tickets for various airlines between the months of March and June of 2019 and between various cities.

Size of training set: **10683** records

Size of test set: **2671** records

FEATURES:

**Airline**: The name of the airline.

**Date\_of\_Journey**: The date of the journey

**Source**: The source from which the service begins.

**Destination**: The destination where the service ends.

**Route**: The route taken by the flight to reach the destination.

**Dep\_Time**: The time when the journey starts from the source.

**Arrival\_Time**: Time of arrival at the destination.

**Duration**: Total duration of the flight.

**Total\_Stops**: Total stops between the source and destination.

**Additional\_Info**: Additional information about the flight

**Price**: The price of the ticket

**DataSet--**

* We will be using two datasets, train data and test data. You can download the data set using this link--<https://github.com/dsrscientist/Data-Science-ML-Capstone-Projects>

The train data comprises of 10683 rows and 11 attributes whereas the test data has 2671 rows.

**Exploratory Data Analysis—**

**What is data analysis?**

First of all, try to understand that what is data analysis. Wikipedia says “Data analysis is a process of inspecting, cleansing, transforming, and modeling data with the goal of discovering useful information, informing conclusions and supporting decision-making”. Okay but let’s understand in simpler words. We get the data from somewhere and it may be some dirty data and we first clean it and then we try to extract the meaningful things from the data by means of understanding it… The process of doing this is known as “Data Analysis”.

****

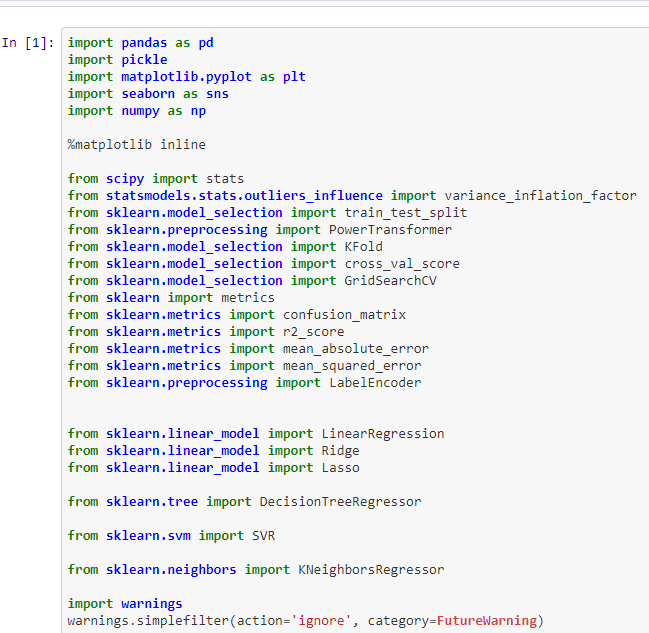
# But why do we need the data analysis?

Yes, we need to analyze the to understand how things going… Let’s understand it with an easy example... Do you remember when our mom dads leave all the works on the last Sunday of the month and calculate how much they spent that month and then worried about them again failing to save more money this month too? So why do they calculate? Yes, they calculate to understand how and where the money goes… To fix the unwanted money flow they need to know all the money spend data and they also priorities their works accounts to the needs.

I believe do you have ever seen some graphs that are representing the sales of the company. So why that graph was created? Yes, you got that to understand the sales of the company over time. So we can see how well the company is performing so users can choose that company for their benefits …

**Lets Start with a coding of flight price –**

1.Import necessary libraries-



# So what is the process to perform data analysis?

To perform the data analysis we need to understand the steps of data analysis...

So here are the 5 baby steps to perform Data analysis: -

1. Need of analysis
2. Collecting the data
3. Clean the Data
4. Analyze the data
5. Interpret and apply the result

# Collecting the data

Yes, the next step is collecting the data But where to go the collect the data. Don’t worry while understanding the need for the analysis you also get an idea about where data comes from. It may come from any API or any website any article or from any database or from many places



Basically, data can be divided into 2 types

1. Structured data
2. Un-structured data

Structured data are those data which is structured and well-formatted so we can use and consume the data very easily but on the other hand, unstructured data is very messy and needs a lot of time to extract the actual data and put into the structured form… because before doing anything we need the structured data that needs to be processed easily

We can also divide the data on the basis of form which could be

1. Audio
2. Video
3. Text
4. Image

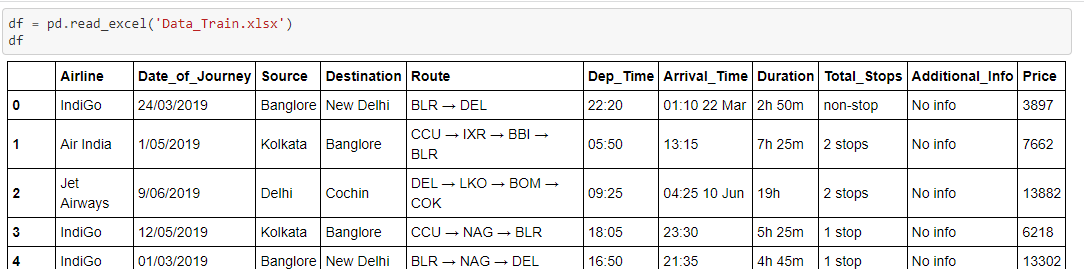
The concerned team can also provide you with any database, API, or text data to process.

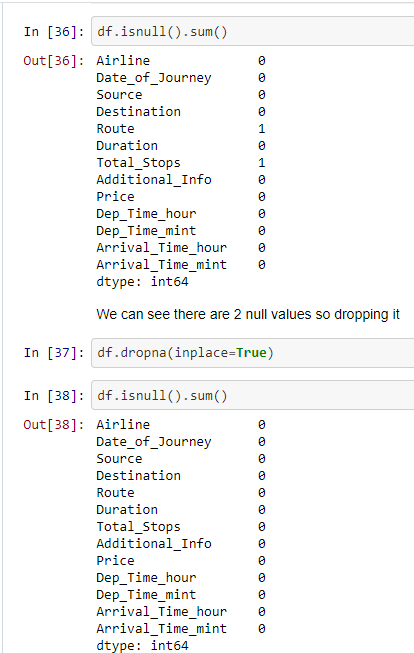
But the day could be

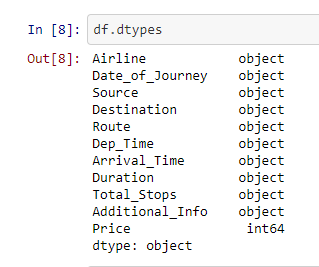
1. Internal Data
2. External Data

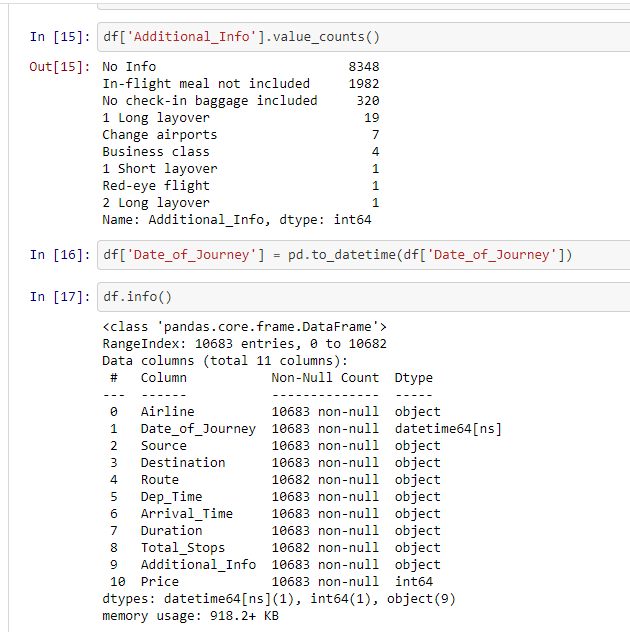
# 

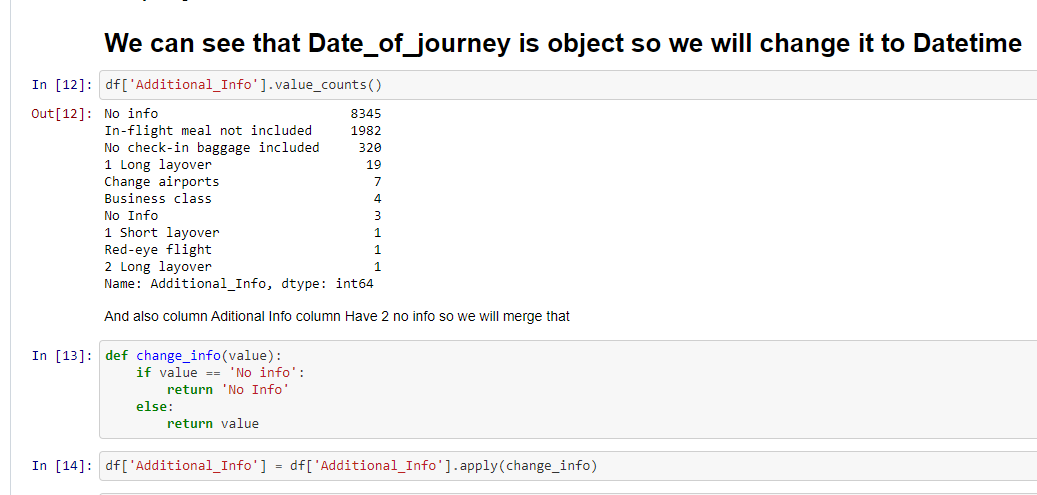
1. Import training data and display the first five rows to take the overall view of the data.



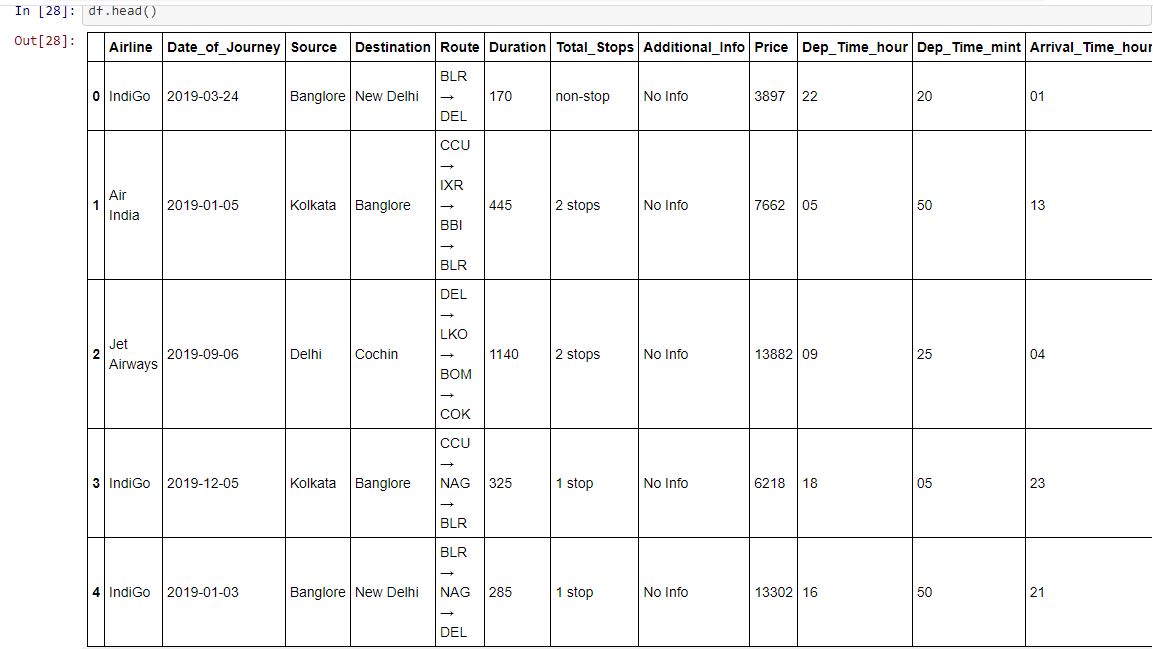
1. We will check whether the data contains any null values using isna() and sum() method and then drop the NaN values using dropna() method so that there are no discrepancies in our data by which we can predict precisely.
2. 
3. Now that our data is clean with no NaN values we will have a look at the datatypes.



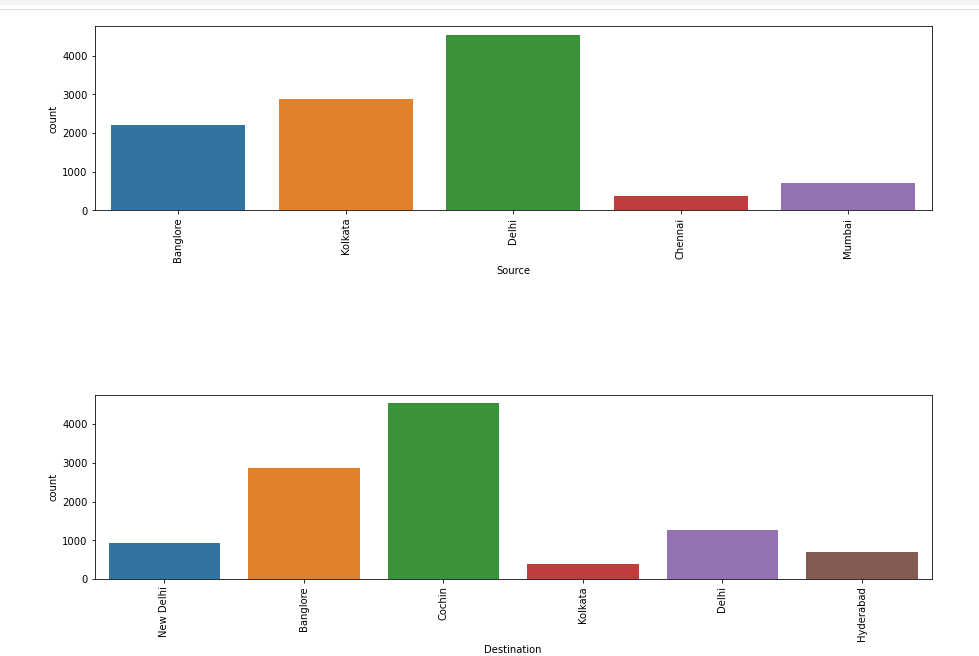
It can be observed that Date\_of\_Journey , Dep\_Time and Arrival\_Time have been assigned as object by default. We will convert this datatype into timestamp to use this column for prediction. 



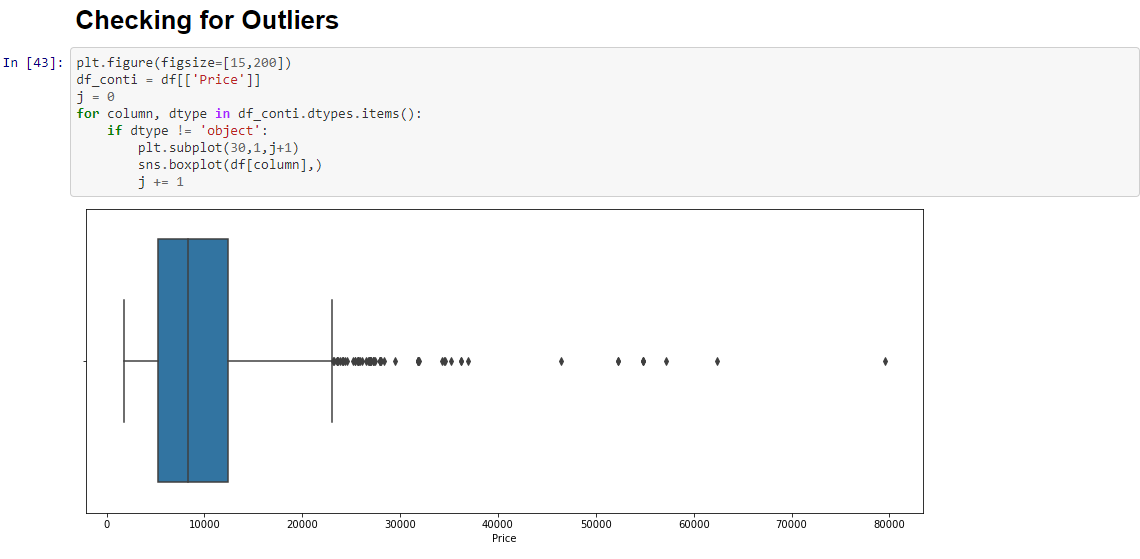
Separate the column “Date\_of\_Journey” into “journey\_day” and “journey\_month” to help our machine learning model understand and use the column for prediction. After doing so we will delete the “Date\_of\_Journey” column which will no longer be useful to us.



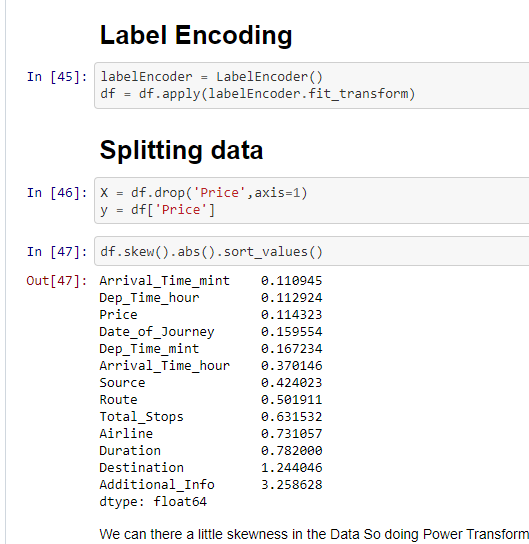
**Univariate analysis** is **the simplest form of analyzing data**. “Uni” means “one”, so in other words your data has only one variable. It doesn't deal with causes or relationships (unlike regression ) and it's major purpose is to describe; It takes data, summarizes that data and finds patterns in the data.



Now i am going to checking the Outliers-



 To get rid of high dimensionality issue we will be using label encoder class.



# Clean the Data

Cleaning the data is one of the most challenging and time-consuming parts of the process which needs lots of time and effort to clean Data. According to ldodds.com, 80% of the time goes into cleaning the data.  
But why do you need the clean the data though? Because the data is collected by and from human and human makes mistakes and that mistake needs to be removed to get the insight from the data. Let’s understand by an example.

In your school, you got a job to collect the name, phone no, hobby, favorite sports, favorite actor name, and marks in the last exam from all the students of your school. You have collected all the in a sheet but at the end of the submission, you realized that you have missed some data like a hobby and favorite actor of some students or they refused to share with you. Also, you noticed you have entered more than 100 marks for some students which is technically impossible So that data needs to be handled to get the insight.

You must have to process

1. Null values
2. Duplicate values
3. Wrong data (Like the length of cars can’t 0 marks can’t be more than 100)

For cleaning the data in python if you are using pandas to store data then pandas have lots of functions to clean the data and restructure the data if required. You can also use the manual way or use other tools like OpenRefine and Drake

# Analyze the data

Analyzing the data is the process of finding the pattern in the data to extract meaningful information from the data. We also use visualization tools to analyze the data and to understand it and in the end, we generate Dashboard, Graphs, or scoreboards to represent the insight to the user. To analyze the data you can use

* Cluster analysis
* Anomaly deduction
* Associate rule

In clustering, we group similar data into different groups so it will be easy to understand the data for example HR group the employee in two or more parts depending on the dedication of his/her work…

We also use algorithms and models to understand the data and to find the patterns in that.

For analyzing the data Python provides lots of function and library which is very flexible to use depending on requirement which includes

* Matplotlib
* Seaborn
* Plotly
* Geoplotlib
* ggplot

#### Matplotlib

It is the first python data visualization library and it’s the core of many libraries and that are built on top of matplotlib. It’s extensively used in python because it’s the oldest one. Pandas also used the matplotlib lib as the default backend.

Seaborn

#### Seaborn

Seaborn is built on the top of matplotlib and is closely integrated with a pandas data structure to privet easy implementation with pandas The main difference between matplotlib and seaborn is seaborn creates beautiful charts in few lines of code Because seaborn uses default style and color palettes.

#### Plotly

Plotly is an online graph tool that also provides offline interactive graph but when you go online you have to pay some sorts of money. We can create advanced graphs in Plotly very easily

#### Geoplotlib

It is used for creating maps and plotting geographical data. Anyone can use this library to create a variety of maps like heatmaps don’t density map and choropleths graph easily It is used for maps because other libraries don’t support maps

#### ggplot

ggplot is the python version of the famous ggplot2 of R and Grammar of Graphics language. ggplot is a declarative style library which tightly coupled with pandas So better to use pandas when you want to go with ggplot

# Interpret and apply the result

This is the last step of the cycle which depends on the Organisation and their teams after an analyst submits his report, insight, and parts they find the data to the organization then the organization finally understand the problem and try to fix this by taking action, for example, an analyzed submitted the employee efficiency and dedication report based on their habits then the organization will thing about it what organization needs to do with the good or bad employee should they need to fire them or should organization warn them and again analyze the result or they should give bounces to the good employee that work harder and work to grow the organsiaton

**Now , Building the model--**

Where we build the model to find the best suited model for the dataset with the highest accuracy so that we can confirm that our data is perfectly analyze and cleaned .

As we used---

1.Linear regression Model-LinearRegression fits a linear model with coefficients w = (w1, …, wp) to minimize the residual sum of squares between the observed targets in the dataset, and the targets predicted by the linear approximation.

2.Decision Tree Model-- **Decision Trees (DTs)** are a non-parametric supervised learning method used for [classification](https://scikit-learn.org/stable/modules/tree.html#tree-classification) and [regression](https://scikit-learn.org/stable/modules/tree.html#tree-regression). The goal is to create a model that predicts the value of a target variable by learning simple decision rules inferred from the data features. A tree can be seen as a piecewise constant approximation.

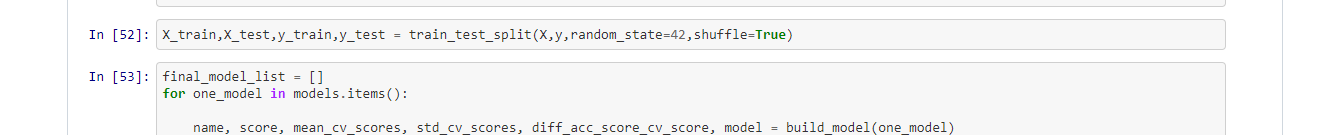
3.kneighbor Regression Model-The target is predicted by local interpolation of the targets associated of the nearest neighbors in the training set.

4.SVM Model- **Support vector machines (SVMs)** are a set of supervised learning methods used for [classification](https://scikit-learn.org/stable/modules/svm.html#svm-classification), [regression](https://scikit-learn.org/stable/modules/svm.html#svm-regression) and [outliers detection](https://scikit-learn.org/stable/modules/svm.html#svm-outlier-detection)

5.Lasso Model-- The [**Lasso**](https://scikit-learn.org/stable/modules/generated/sklearn.linear_model.Lasso.html#sklearn.linear_model.Lasso) is a linear model that estimates sparse coefficients. It is useful in some contexts due to its tendency to prefer solutions with fewer non-zero coefficients, effectively reducing the number of features upon which the given solution is dependent. For this reason, Lasso and its variants are fundamental to the field of compressed sensing

6-Ridge Model- [**Ridge**](https://scikit-learn.org/stable/modules/generated/sklearn.linear_model.Ridge.html#sklearn.linear_model.Ridge) regression addresses some of the problems of [Ordinary Least Squares](https://scikit-learn.org/stable/modules/linear_model.html#ordinary-least-squares) by imposing a penalty on the size of the coefficients.

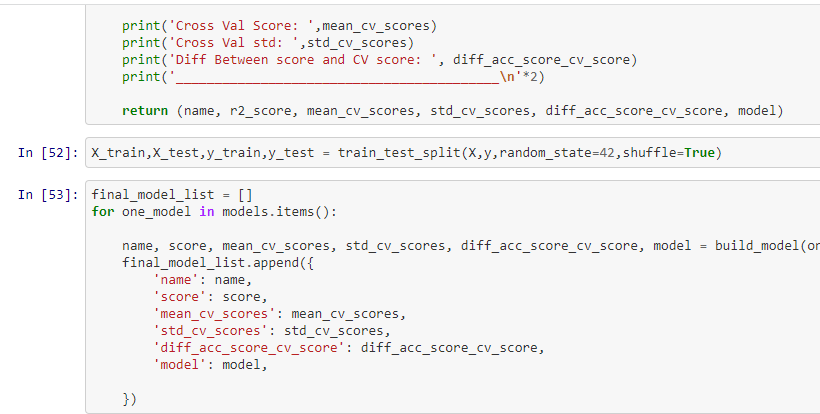
Splitting the data into train and test data. We taken 80% data for training and remaining 20% for testing.

****

Let us now define a function to predict the Training score, Predictions, r2 score, Mean absolute error(MAE), Mean square error(MSE) and Root mean square error (RMSE).

And crossvalidating model with the all the machine learning techniques--





Here we are getting best Output with the Kneighbor Regression with the hyperparameter tuning with the Grid search CV



We have achieved approximately 79% accuracy when we used KNeighbor and difference of the score is high with this .

Here is the end of article where we know that how to analyze the data get the best model to perform for the all the data’s