# Experiment. No. 07

Title : Predict the price of the Uber ride from a given pickup point to the agreed drop-off location.

1. Pre-process the dataset
2. Identify outliers
3. check the correlation
4. Implement linear regression and random forest regression models.
5. Evaluate the models and compare their respective scores like R2, RMSE, etc

Objective : To preprocess dataset and identify outliers, to check correlation and implement linear regression and random forest regression models. Evaluate them with respective scores like R2, RMSE, etc

Theory :

Data Preprocessing :

Data preprocessing is a process of preparing the raw data and making it suitable for a machine learning model. It is first and crucial step while creating a machine learning model.

When creating a ML project, it is not always a case that we come across the clean and formatted data. And while doing any operation with data, it is mandatory to clean it and put in a formatted way so for this, we use data preprocessing task

A real world data generally contains noises, missing values & maybe in an unusable format which cannot be directly used for machine learning mode

It involves below steps

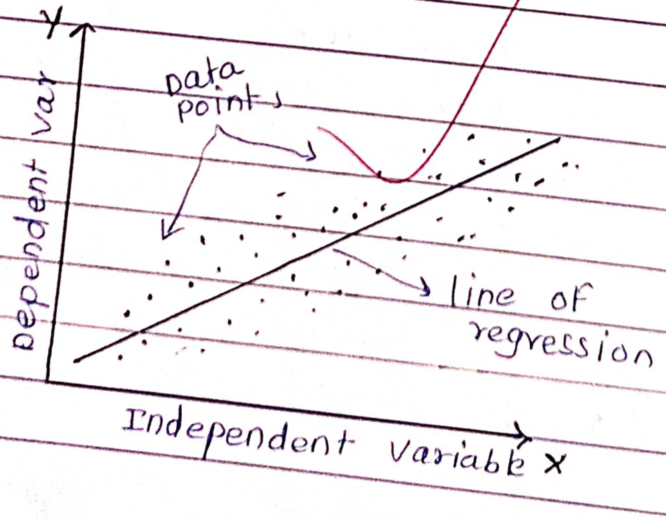
1. Getting the dataset
2. Importing libraries
3. Importing dataset
4. Finding missing Data
5. Encoding categorical Data
6. splitting dataset into training and test set
7. Feature scalling

Linear Regression:

Linear Regression is a statistical method that is used for predictive analysis. Linear regression makes predictions for continuous/re or numeric variables such as sales, salary, ag product price etc

Linear regression algorithm shows a linear relationship between a dependent (y) and one or more independent (x) variables, hence are calle as Linear Regression.

The linear regression model provides a sloped straight line representing the relationsh between the variables.



## Random Forest Regression models :

Random forest is based on the concept of ensemble learning, which is a process of combining multiple classifiers to solve a complex problem and to improve the performance of the model

Random forest is a classifier that contains a number of decision trees on various subsets of the given datasets and takes the average to improve the predictive accuracy of that dataset.

The greater number of trees in the forest leads to higher accuracy and prevents the problems of overfitting .

## Boxplot :

They are a measure of how well data is distributed across a dataset. This divides the dataset into three quartiles. This graph represents the minimum, maximum, average, first quartile & the third quartile in the dataset

R provides a boxplot() function to create a boxplot. There is following syntax of boxplot() fn-

```
boxplot (x, data, notch ; varwidth, names, main)
```

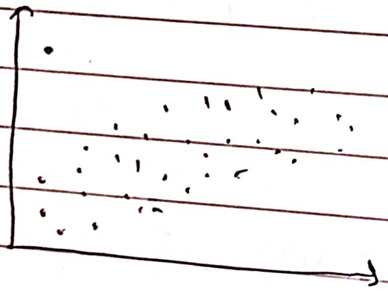
## Outliers :

It refers to the data points that exists outside of what is to be expected. The major thing about the outliers is what do you think. If you are going to analyze any task to analyze datasets, you will always have some assumption based on how this data is generated.

If you find some datapoints that are likely contain some form of error, then these are definitely outliers.

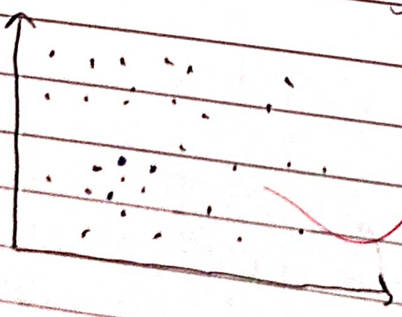
global outliers :

They are also called point outliers. are taken as simplest form of outliers. data points deviate form all the rest of data points in a given dataset, it is known as global outlier.



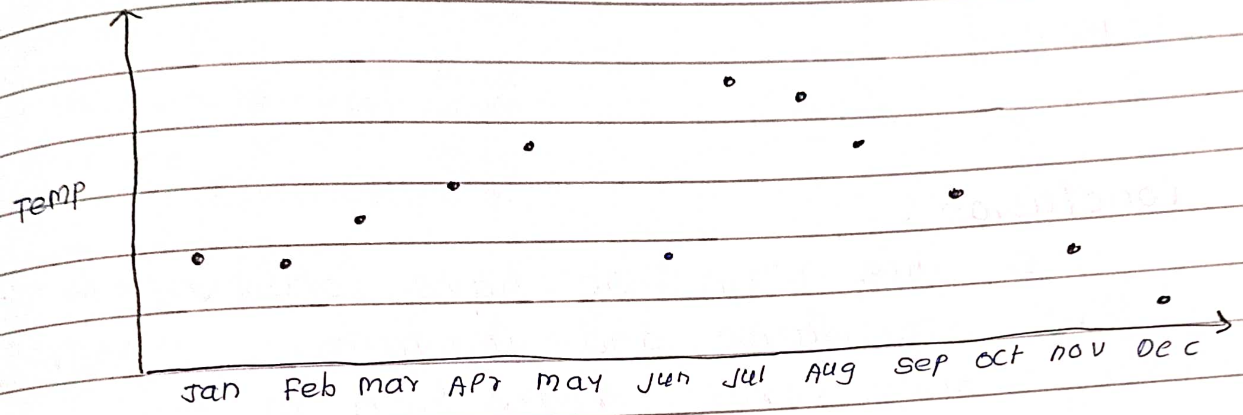
Collective outliers :

when a group of datapoints deviates f the rest of the dataset is called collecti outliers, but when you consider the data obje as a whole, they may behave as outliers.



contextual outliers :

It means this outlier introduced with a content These types of outliers happen if a data object deviates from the other data points because of any specific condition in a given dataset



Haversine:

The Haversine formula can be calculates the shortest distance between two points on a sphere losing their latitude and longitudes measured along the surface. It is important for use in navigation

matplotlib:

It is an amazing visualization library in python for 2D in arrays. matplotlib is a multi-platform data visualization library built on Numpy arrays and designed to work with the broader scipy stack It consists of several plots like line, bar, scatter, histogram, etc

mean squared error-

The mean squared error (MSE) or mean squar Deviation (MSD) of an estimator measure the averag of error squares. It is a risk function, correspo -ding to the expected value of the squared error loss. It is always non-negative and values close to zero are better. The MSE is the secon moment of the error and thus incorporates both the varience of the estimator and bias.

Conclusion :

In this way, we have studied & implem concept correlation and implement linear regre and random forest regression models.