**Date: 01/08/2019**

Cab Fare

Prediction

Project

Report

**Submitted by Anushka Mishra**

**Contents**

1. **Introduction** 
   1. **Problem Statement**
   2. **Data**
   3. **Exploratory Data Analysis**

1. **Methodology** 
   1. **Pre-Processing** 
      1. **Missing Value Analysis**
      2. **Outlier Analysis**
      3. **Feature Selection**
      4. **Feature Scaling**

* 1. **Model Development** 
     1. **Choosing Model**
     2. **Baseline model**
     3. **Linear Regression**
     4. **Decision Tree**
     5. **Random Forest**

1. **Conclusion** 
   1. **Model Evaluation** 
      1. **Root Mean Square Error (RMSE) & R square**
   2. **Model Selection**
2. **Deployment** 
   1. **Python Deployment**
   2. **R Deployment**
3. **Summary**

**Appendix A – Extra Figures**

**Appendix B – R Script**

**References**

# Chapter 1

## 1. Introduction

### 1.1 Problem Statement

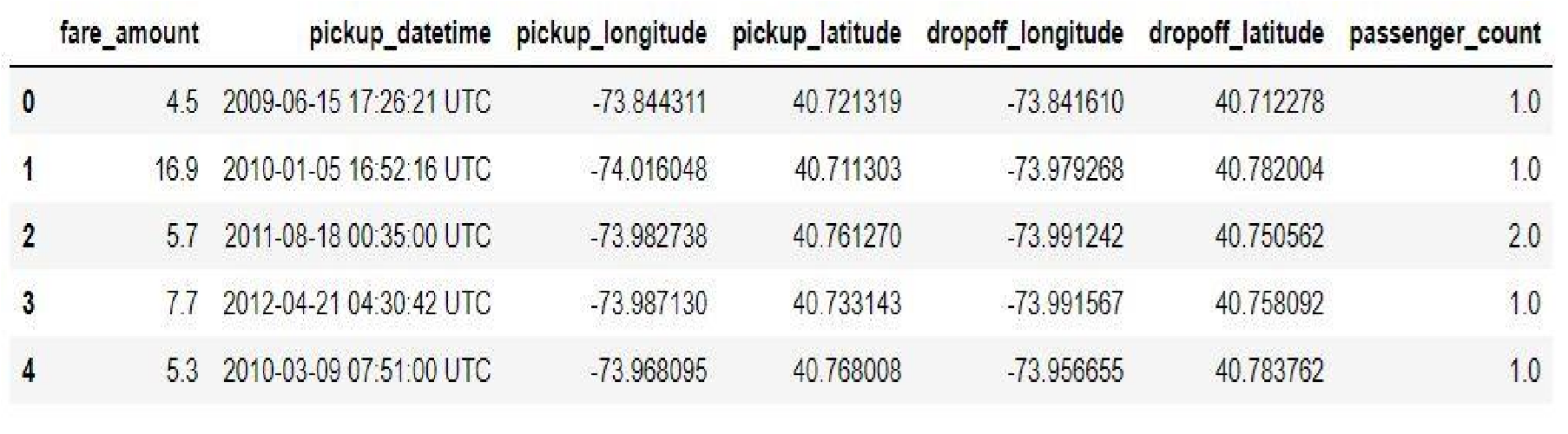
I am a cab rental start-up company. I have successfully run the pilot project and now want to launch My cab service across the country.

I have collected the historical data from My pilot project and now have a requirement to apply analytics for fam prediction.

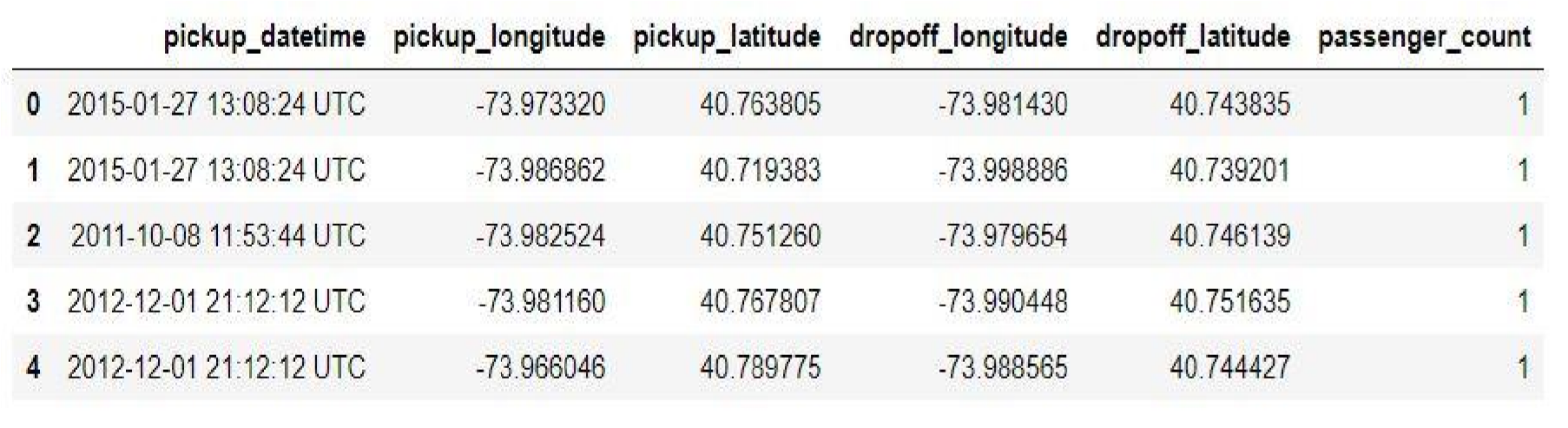
##### My objective of this Cab Fam Prediction Project is to design a system that predicts the fam amount for a cab ride taken in the city.

### 1.2 Dataset

Train Dataset



Test Dataset



In train set, I can see My historical data. It has 16067 observations and 7 features, including fare\_amount which is My target variable.

In test set, I can see data, whose target variable i.e. fare\_amount values to be predicted. It has 9914 observations and 6 features, where, fare\_amount is not available, as I predict those values later in my project later.

As said above, My primary focus is to determine the values for My target variable fare\_amount for future test cases using the above shown dataset in image 1.

### 1.3 Exploratory Data Analysis

In Exploratory Data Analysis, I go through different things, like:

* **Brainstorming** – I draw a rough sketch which talks about – the steps I am going to follow to achieve my given objective.

* **Defining Problem Statement** – In this project I am given with problem statement to work with.

* Knowing the types of variables in the given dataset, whether they are factor, character or numeric. In My project I changed the pickup\_datetime variable from object to datetime, the same way, I converted the datatype of fare\_amount from object to numeric.

* Changing and removing datatypes if required. In My project, I removed my variables i.e. pickup\_latitude, pickup\_longitude, dropoff\_latitude and dropoff\_longitude after I obtained distance from them.

# Chapter 2

## 2. Methodology

Now, I have the dataset and also, I discussed about Exploratory Data Analysis, let’s talk about the **methodology** I am going to follow to achieve my goal.

I will be going through:

* Pre-processing which includes missing value analysis, outlier analysis, feature selection and feature scaling.
* Model development, where I will choose what machine learning algorithms to apply.

### 2.1 Pre-processing

In pre-processing, I actually apply few techniques like missing value analysis, outlier analysis, feature selection, feature scaling.

Why I do that? Ill, actually I never get a structured data to work with. Always messy data is handed to us, and I need to clean that data.

The data may have many observations (rows in dataset), where values in few fields will be absent. I can also say, there may be some inconsistent values in a variable (column in dataset), when compared with other values.

When I go for model development, I should have a structured dataset. I can’t go forward for model development, if I don’t apply pre-processing techniques on data and convert it into structured format.

#### 2.1.1 Missing Value Analysis

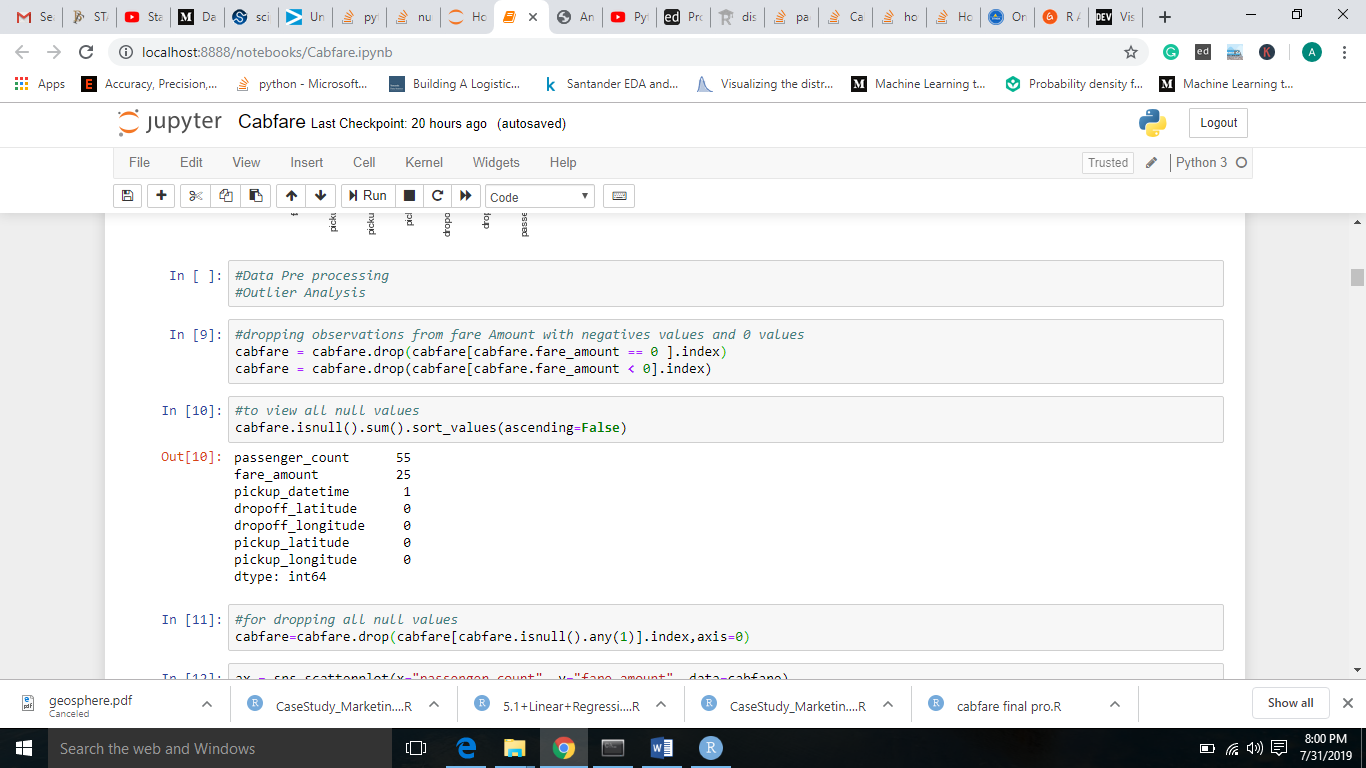
Missing value analysis, as the name suggests, I face with situations, where I am given with dataset, and I have missing values in the observations.

The reason why I have missing values may be plenty, like human error, the user didn’t want to sham his complete information, the one who was supposed to fill the data may didn’t work properly.

But, as discussed earlier, I have to give structured data to the model. In order to achieve that, I perform missing value analysis on top of the data to clean the data, to transform the data from unstructured format to structured format.

I give a line of code and it gives us the total number of missing values in each column. Later, I impute those missing values using mean, median or KNN method for numerical variables or using mode for categorical variables. In some cases, I may delete the observations with missing values, only when I have a case where I got few observations with missing values.

After imputing the missing values, I proceed further with outlier analysis. In My project, I had missing values in fam\_amount i.e. 25 missing values, pickup\_datetime 1 missing value and passenger\_count i.e. 55 missing values. I dropped these observations from the dataset, as mentioned above, I may go for deletion of observations with missing values when I have a low number of it.



Another important thing is, I apply missing value techniques only on numerical variables.

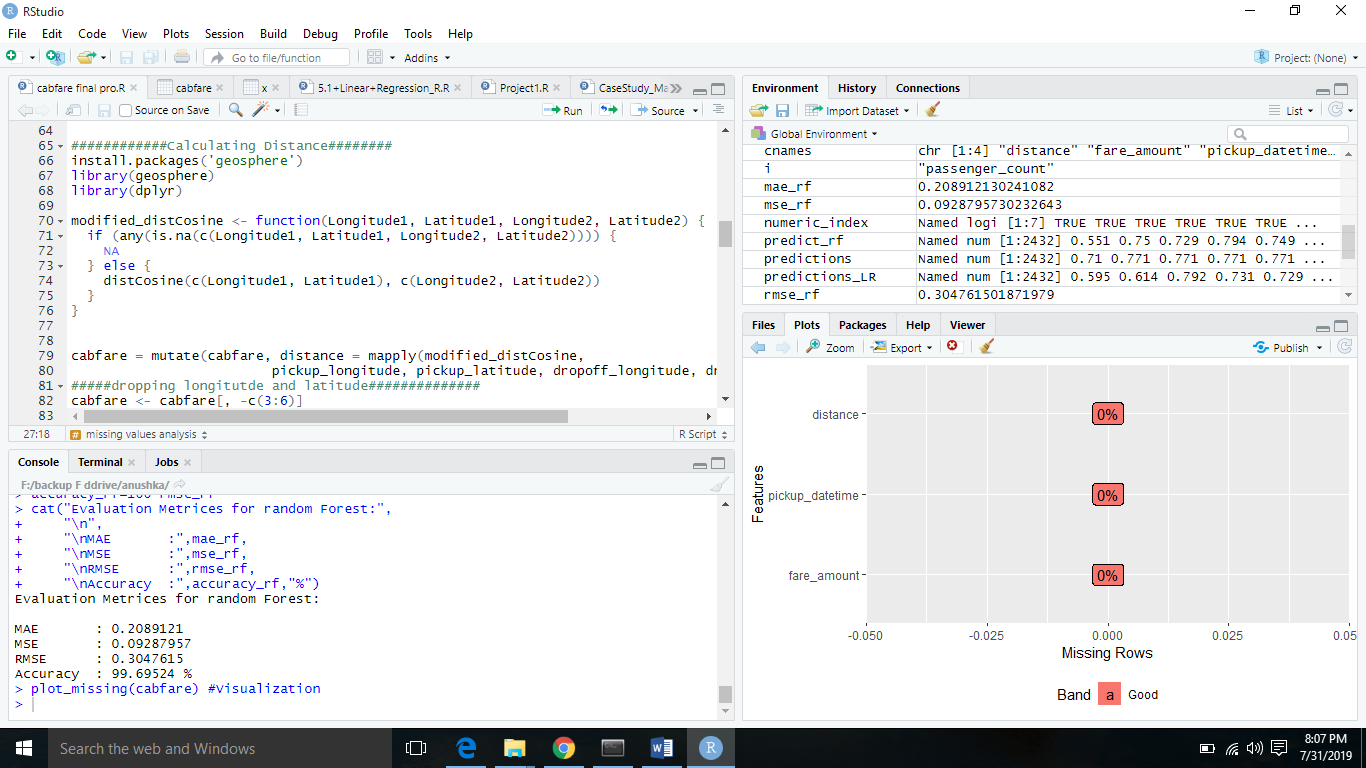
#### Calculating Distance

Obtaining Distance from pickup\_latitude, pickup\_longitude, dropoff\_latitude and dropoff\_longitude

In python

#### 

In R



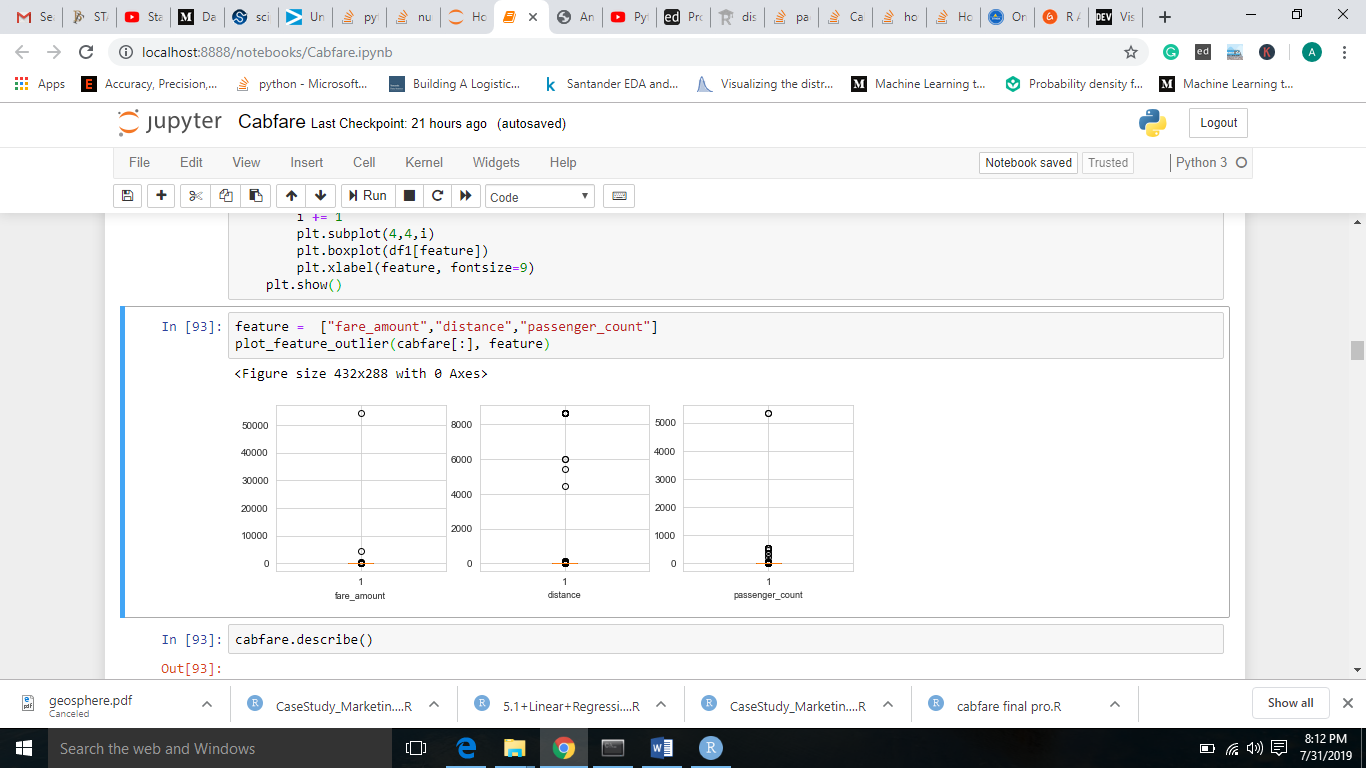
#### 2.1.2 Outlier Analysis

Outliers may be defined as the inconsistent values in a variable. For example, a = 1,2,3,4, 20. In object a, 20 is inconsistent, in terms of mean. Another important thing is, I apply outlier analysis only on numerical variables.

Outliers am used for fraud detection. Let’s say, in one bank account, consistently, I observe an amount which ranges from 50,000 to 1,00,000, but in one case 10 lakhs gets deposited. In that case, simply using outlier technique, I can get the inconsistent values.

In My project, I found outliers in fare\_amount, distance and passenger\_count variables. In other variables the values am consistent with each other.

I deleted the outliers in these variables as the number of outliers Ire minimum. I Int for ascending order and knew the inliers and Int for descending order and learnt about outliers and finally sometimes I used to describe function to know minimum and maximum.



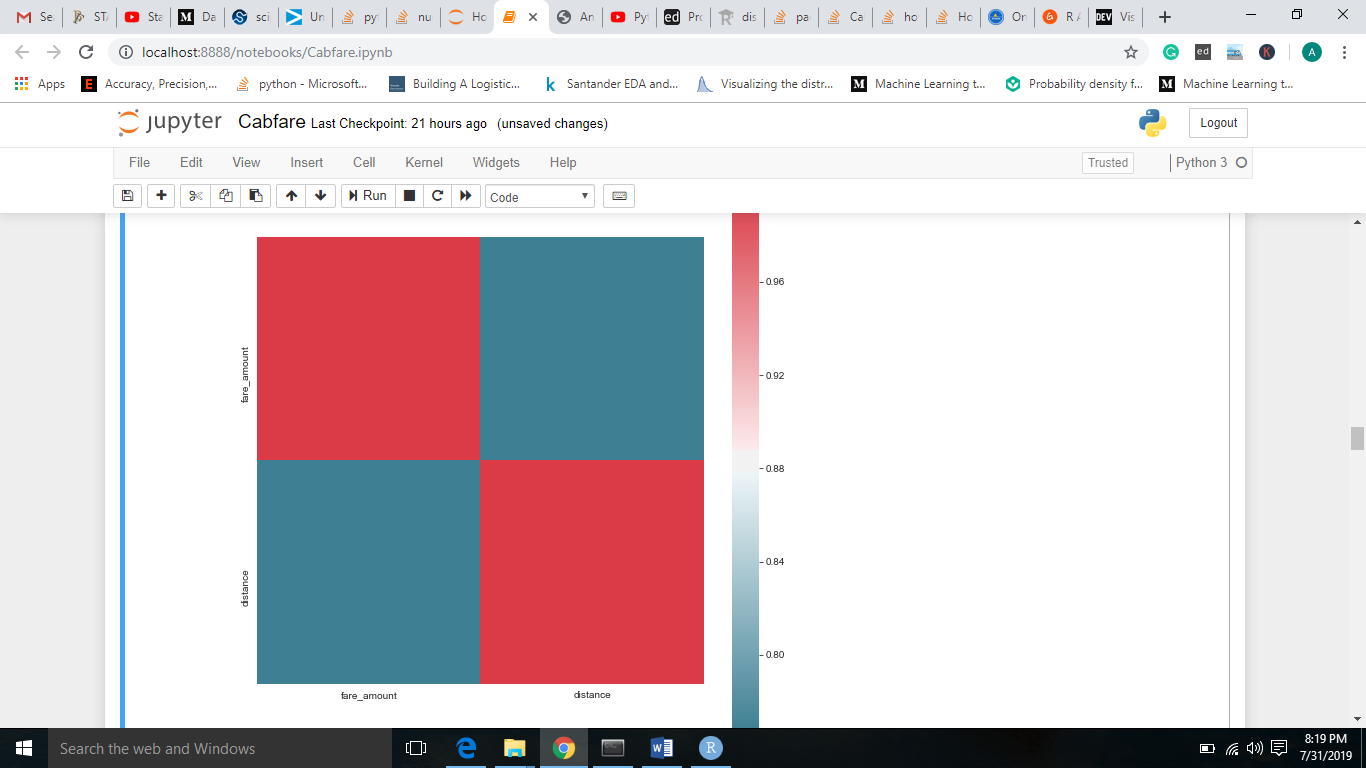
#### 2.1.3 Feature Selection

In Feature Selection, I use few techniques to know which variables am important to us, it’s all about selecting subset of variables from all variables. Actually, when I am given with a dataset, I perform exploratory data analysis, later data pre-processing to clean the data and transform it from unstructured to structured data to feed into model.

However, after this, I may face a situation where I may have variables which have same information with them about the target variable. Let’s talk about an example, in a situation where I am sending five people on a mission. Later, you came to know that, two individuals have the same exact information with them about the mission. Definitely you would drop one, in order to reduce infrastructure and complexity.

The same way, I also drop few variables if they have same information. I always aim that, there should be no independent variable which talks the same as other independent variables but, I appreciate those variables which talks more about the target variable.

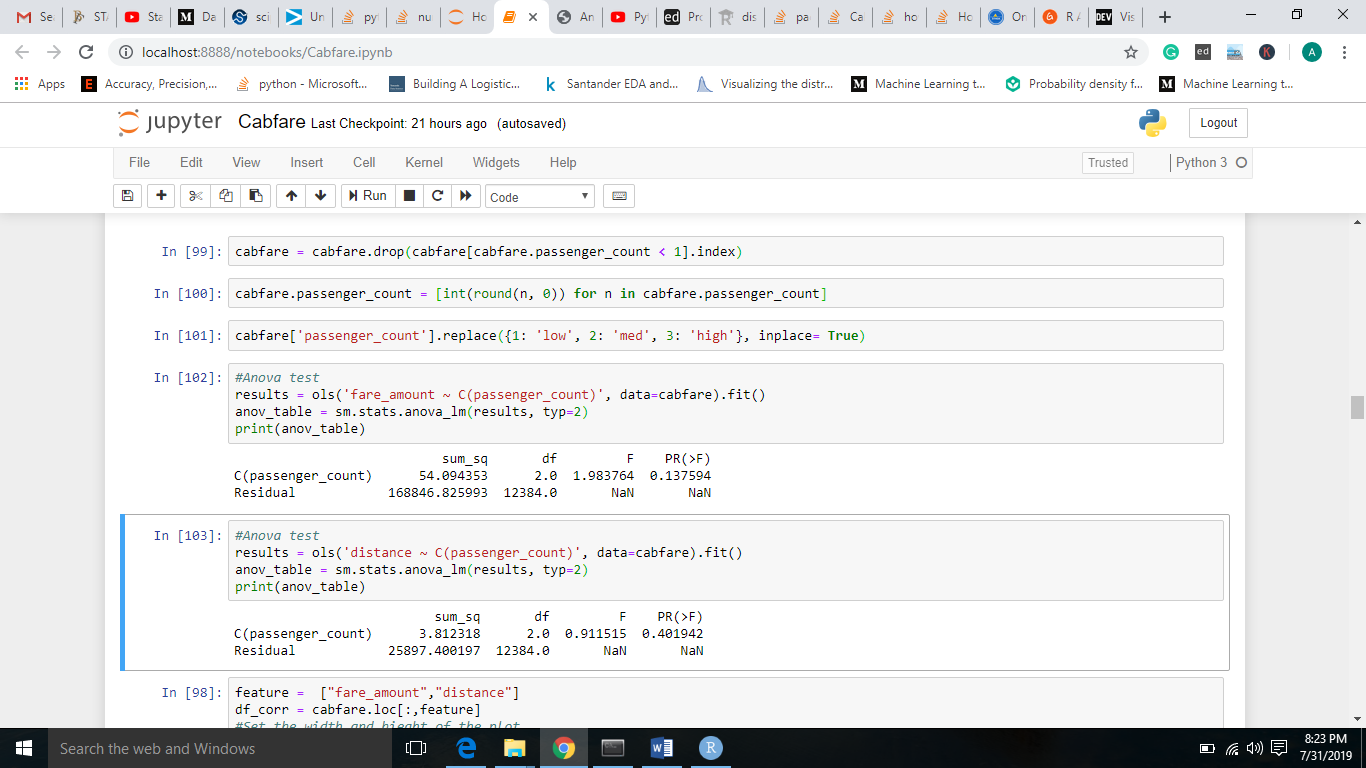
In My project, I did correlation analysis for target variable (continuous variable) and other continuous variable i.e. distance. Let’s check it:



Now, I can see, distance got pretty good information about fare\_amount. I am not going to drop it.

Next, I did Anova test between categorical variables and target variable (continuous variable)

Let’s check that:



It’s clear, that passenger\_count, p value > 0.05, by which I accept null hypothesis (all means am equal, they carry same information). So, I am going to drop it.

That’s all about Feature Selection.

#### 2.1.4 Feature Scaling

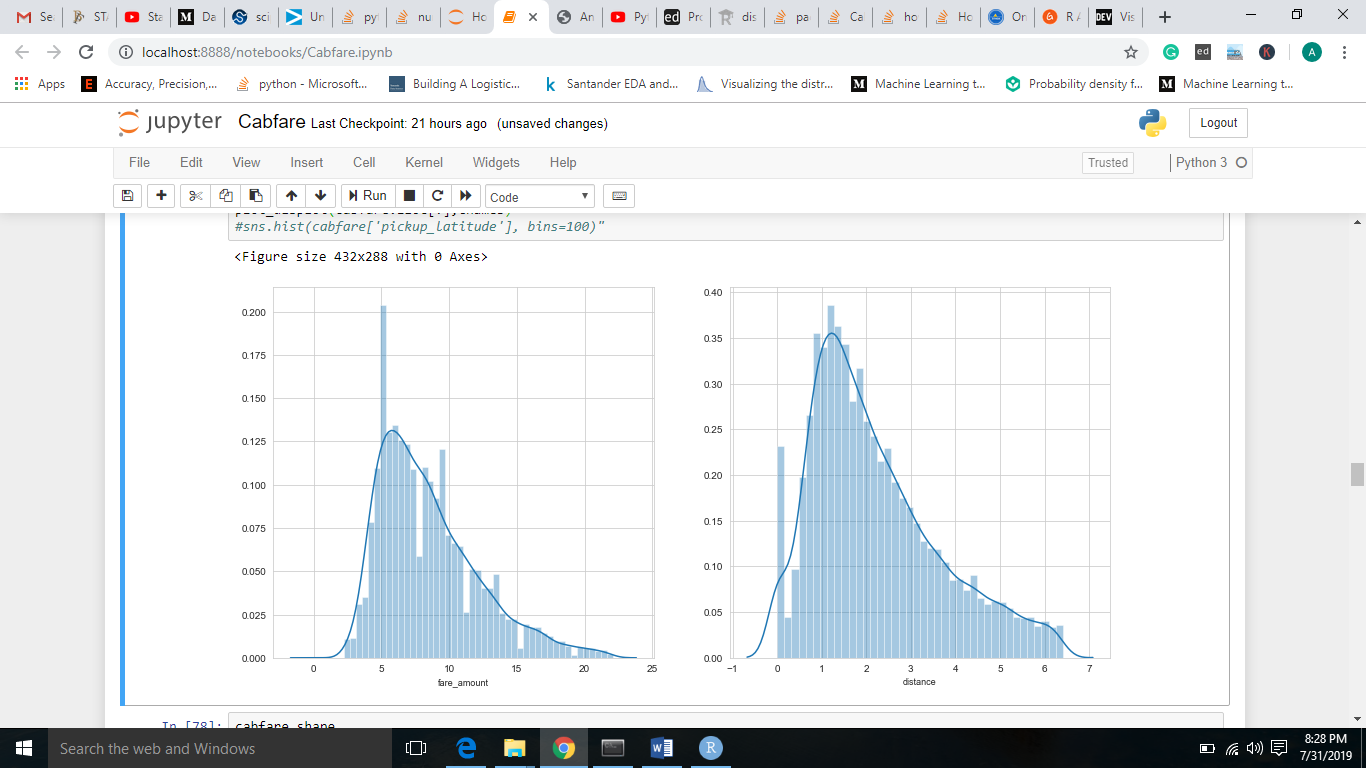
In Feature Scaling, I try to limit the ranges of variables, so that they can be compared on the same ground.

Let’s talk about one example, consider, two variables, age and income. Age is varied from 1 to 100 but coming to income, it ranges in a large scale.

At this situation, the higher values bias the result towards themselves, in order to overcome a situation like this, I use Feature Scaling, where I limit the ranges of variables.

In My project, I did go for feature scaling for distance variable as the given data was left skeId. I also went for distribution check, so let’s have look on that.

I



Now, as observed, the data points of distance variable are left skewed. Now, I am going to apply normalization.

Now, let’s step into Model Development.

### 2.2 Model Development

Model Development, is the phase which comes after I am done with applying the exploratory data analysis, data pre-processing techniques, on the top of data.

The data, will be in structured format, which was my goal, is now ready to develop model.

After I defined my objective and received the data, I transformed it into my required form, I enter into model development, but before that, let’s discuss about model selection.

#### 2.2.1 Model Selection

Model Selection particularly depends on the objective, the problem statement. I have to know at first hand, that, under which category, the problem statement falls.

I have four categories:

* Forecasting
* Classification
* Optimization
* Unsupervised Learning

##### My problem statement is to design a system that predicts the fare amount for a cab ride taken in the city.

My problem statement is a regression problem (target variable is a continuous variable) and it falls under Forecasting category.

In my project I decided to go with, Baseline, Linear Regression, Decision Tree and Random Forest.

#### 2.2.2 Base Line Model

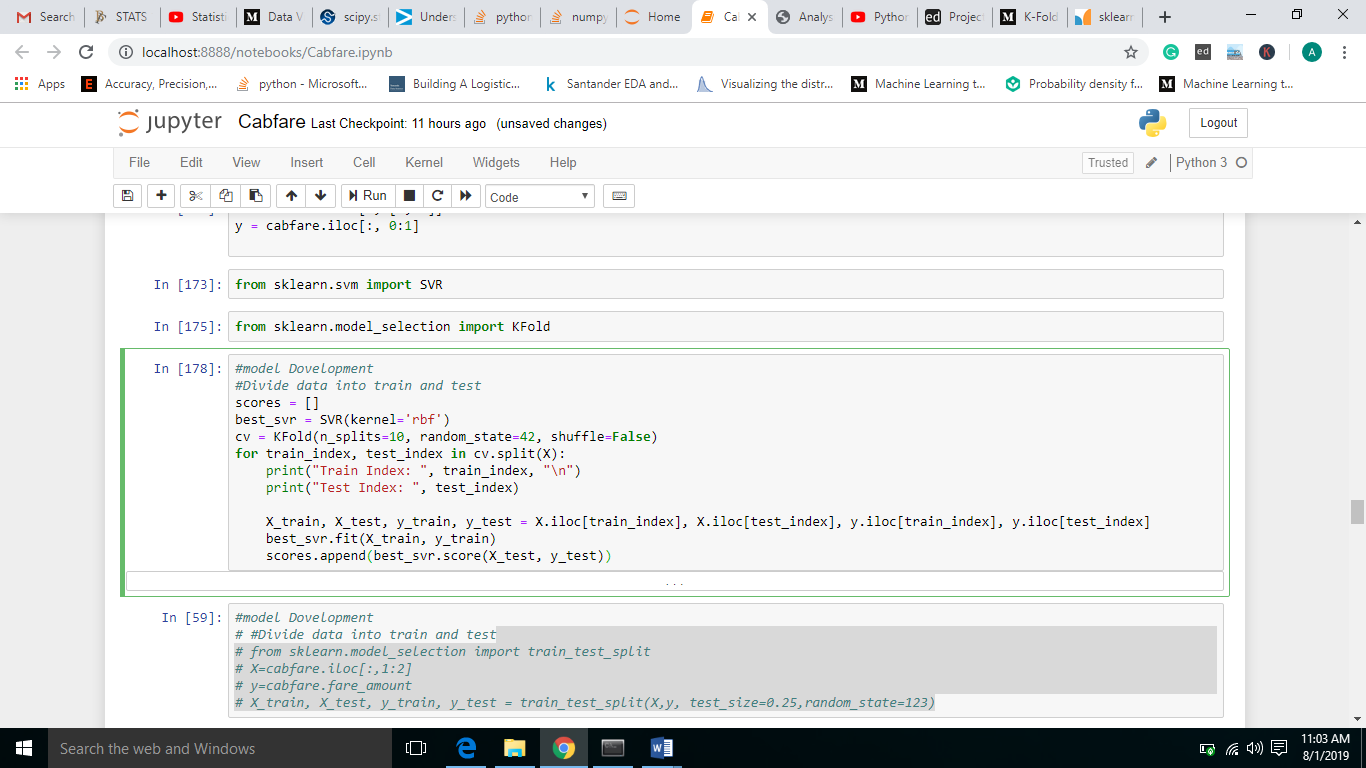
A **baseline model** is the result of basic statistics without applying any machine learning techniques. **Any model we build must improve upon this solution.** Some ways of building a baseline model am taking the most common value in case of classification, and calculating the average in a regression problem. In this analysis, since we am predicting fare amount (which is a quantitative variable).We will predict the average fare amount. RMSE of this model is 17.44.

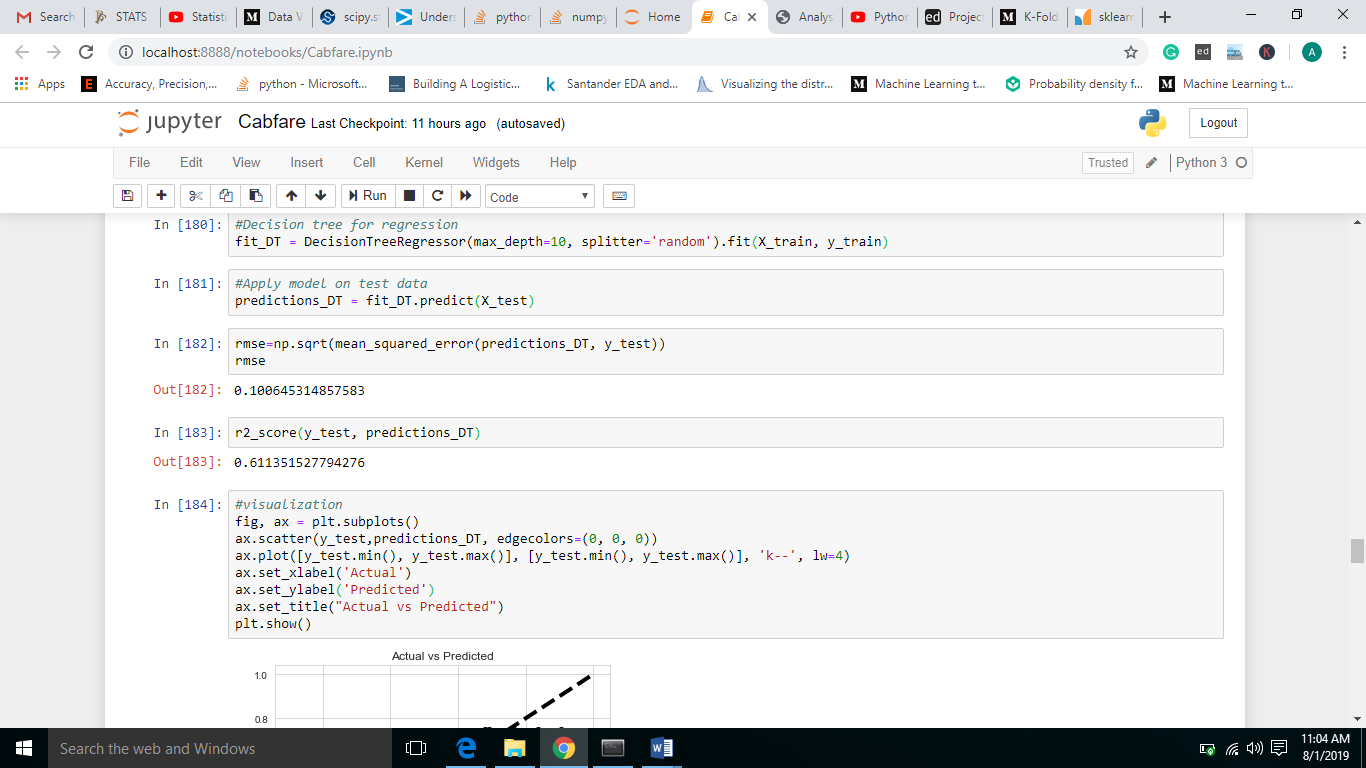
#### 2.2.3 Decision Tree

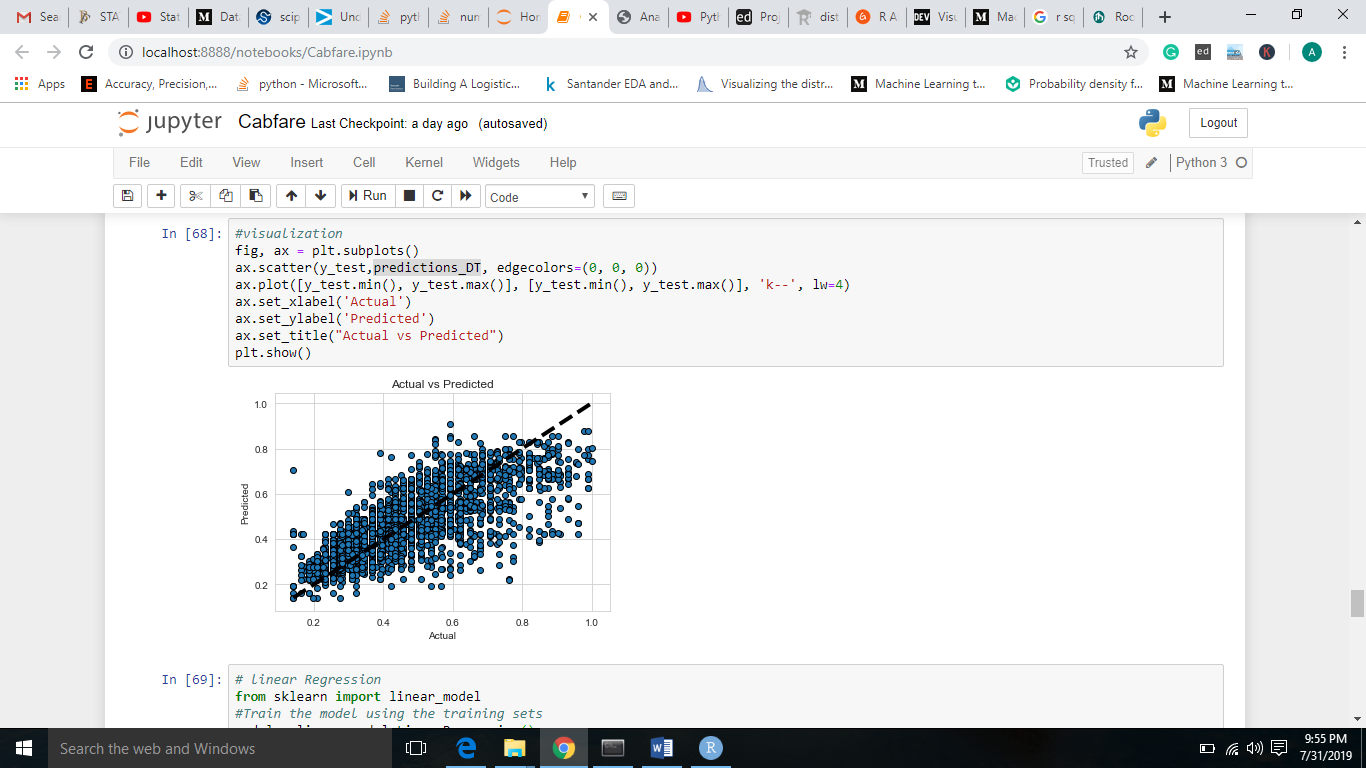
Decision Trees am a type of Supervised Machine Learning (that is you explain what the input is and what the corresponding output is in the training data) where the data is continuously split according to a certain parameter.

The tree can be explained by two entities, namely decision nodes and leaves. The leaves am the decisions or the final outcomes. And the decision nodes am where the data is split.

In My project, I get RMSE as 10.06453 and R square as 0.611351. My aim is – I always want a model with low RMSE value i.e. minimum calculated errors and high R square value i.e. the independent variables should have maximum potential to explain about the target variable.







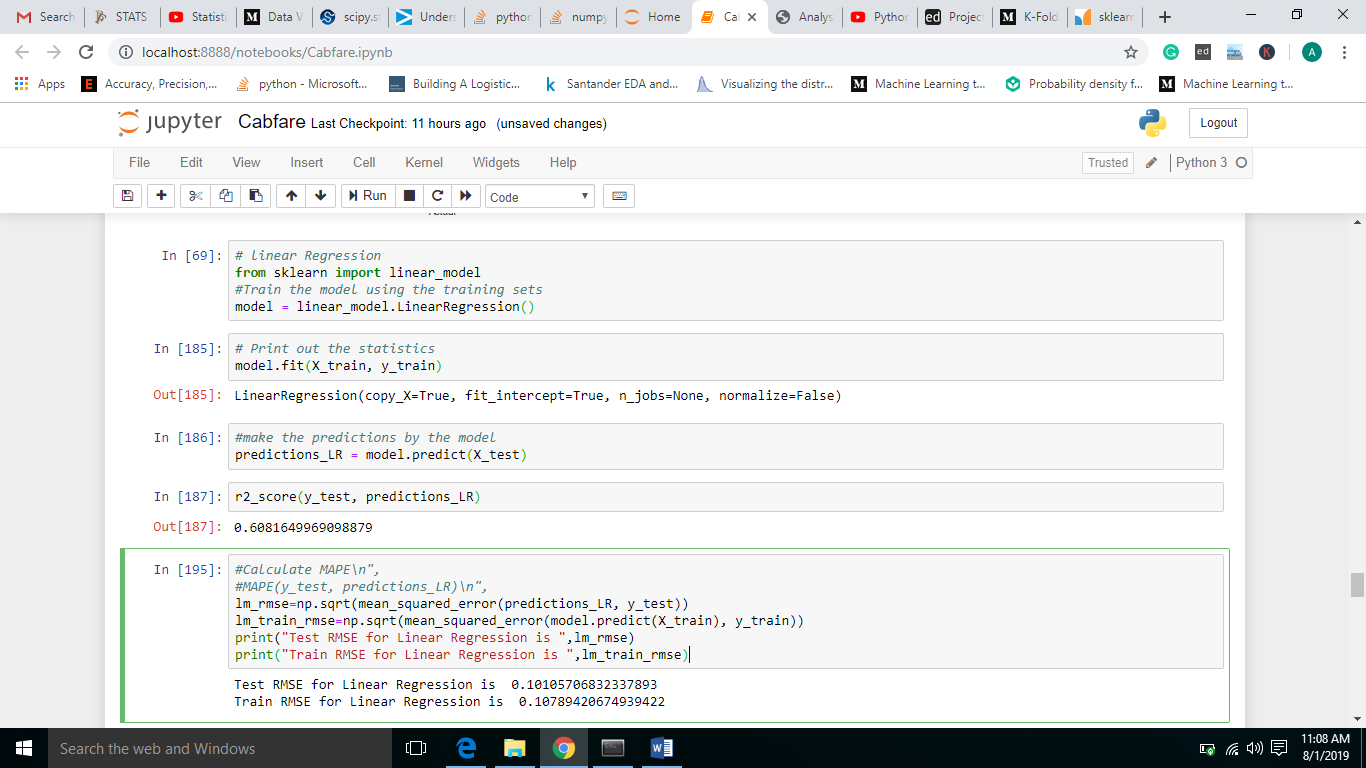
#### 2.2.4 Linear Regression

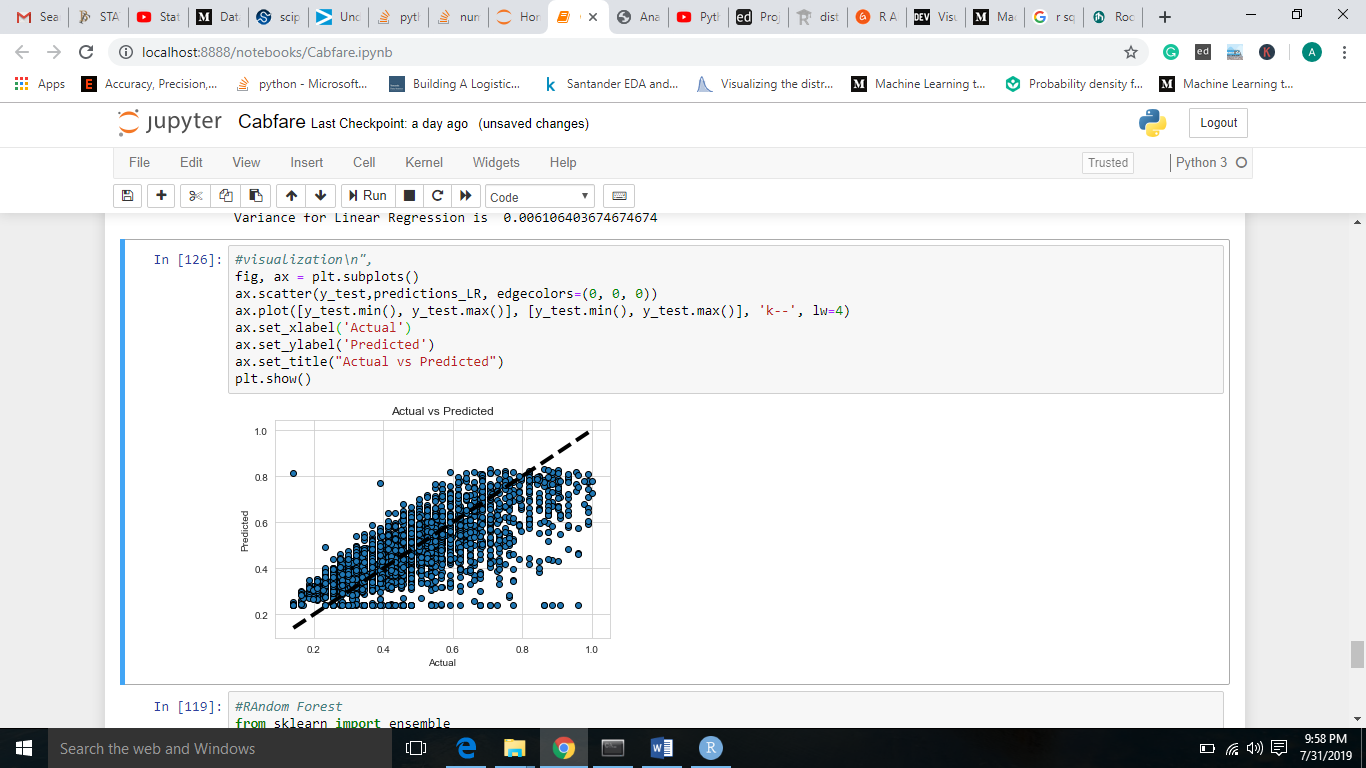
Linear Regression is a supervised machine learning algorithm where the predicted output is continuous and has a constant slope. It’s used to predict values within a continuous range, (e.g.

sales, price) rather than trying to classify them into categories (e.g. cat, dog).

Linear Regression, unlike other algorithms, stores information in terms of coefficients. It is a statistical model. I cannot use this for classification. It describes relationship among variables.

In My project, I get RMSE as 10.10570 and R square as 0.608164. I am rejecting this model as RMSE is high and R square is low when compared with all other models. My aim is – I always want a model with low RMSE value i.e. minimum calculated errors and high R square value i.e. the independent variables should have maximum potential to explain about the target variable.





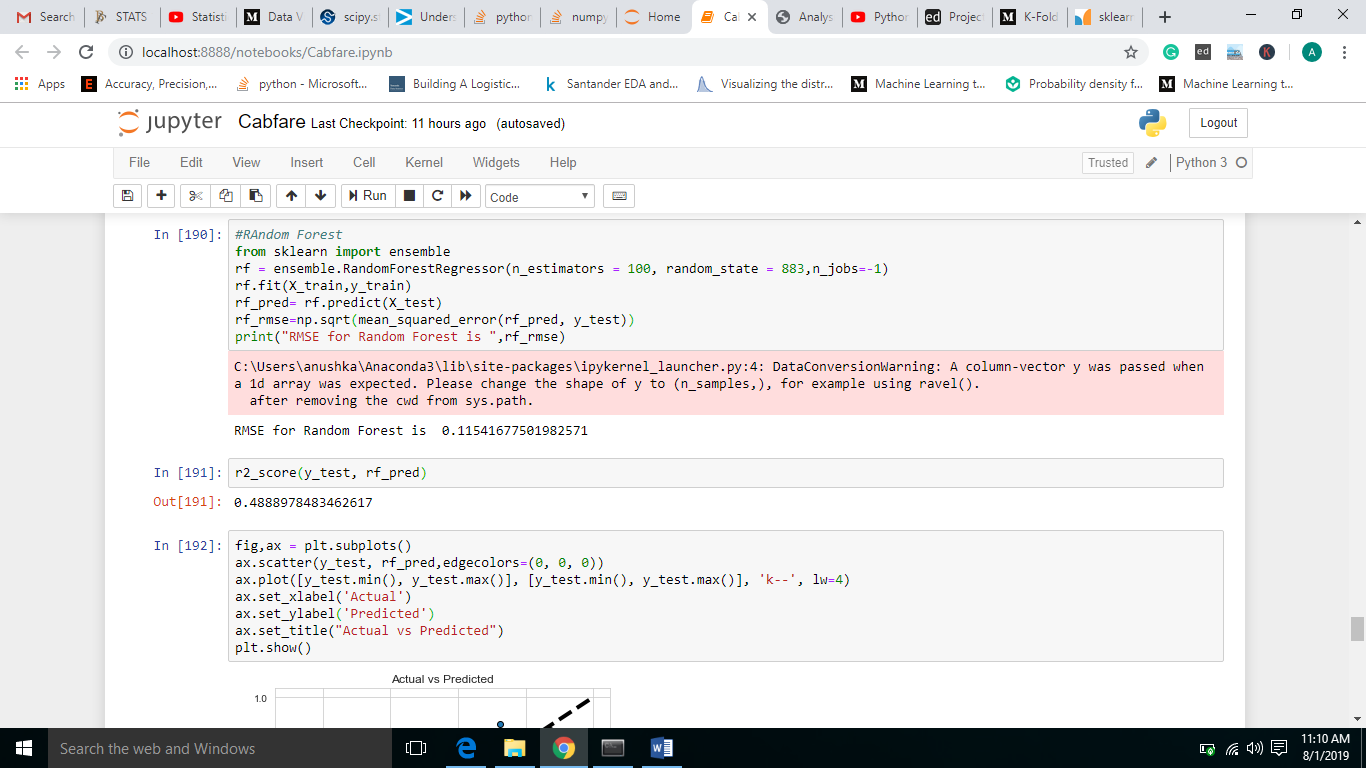
#### 2.2.5 Random Forest

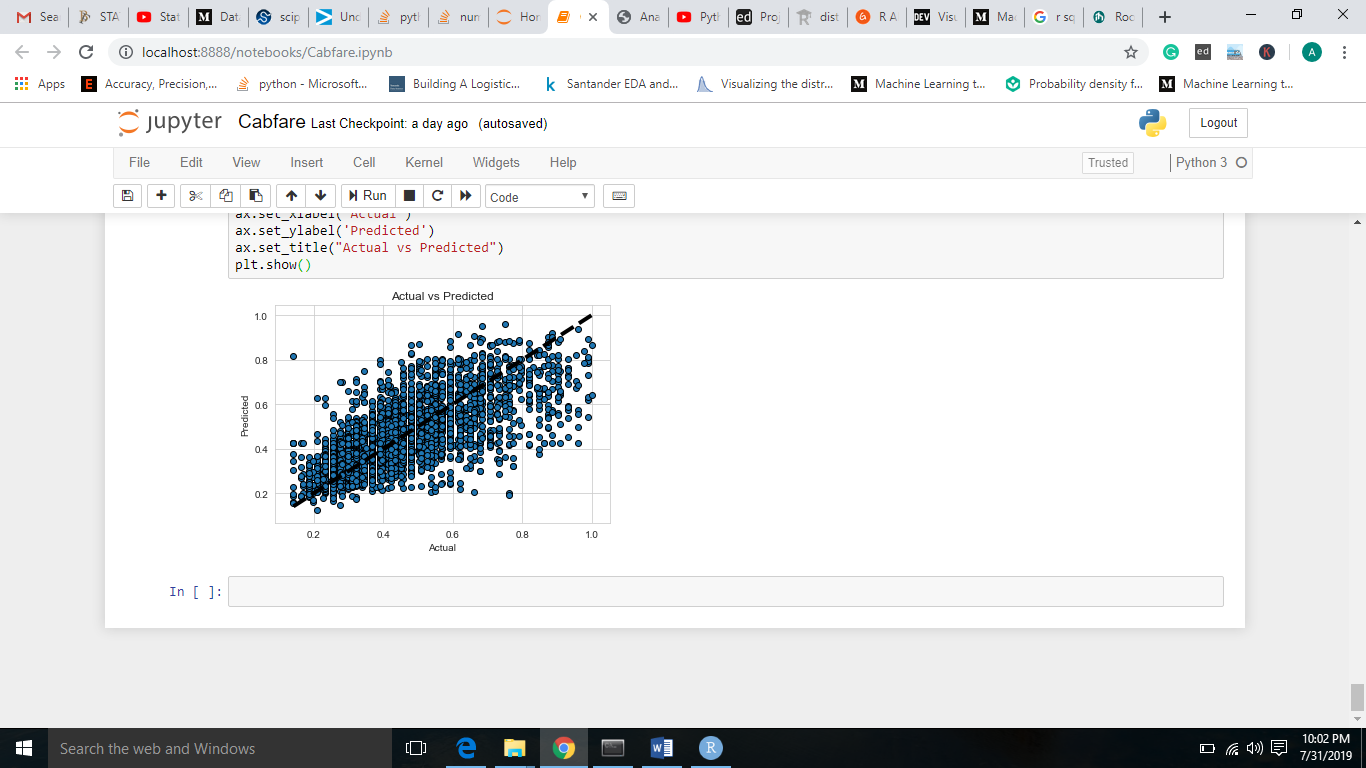
The Random Forest is a model made up of many decision trees. Rather than just simply averaging the prediction of trees (which I could call a “forest”), this model uses two key concepts that gives it the name *random*:

* Random sampling of training data points when building trees
* Random subsets of features considered when splitting nodes

The random forest combines hundreds or thousands of decision trees, trains each one on a slightly different set of the observations, splitting nodes in each tree considering a limited number of the features. The final predictions of the random forest am made by averaging the predictions of each individual tree.

In My project, I get RMSE as 12.92710 and R square as 0.4498216. Although, Random Forest performed better than Decision Tree, I am rejecting this model as RMSE is high at a small margin and R square is low at a small margin when compared with other models. My aim is – I always want a model with low RMSE value i.e. minimum calculated errors and high R square value i.e. the independent variables should have maximum potential to explain about the target variable.





# Chapter 3

## Conclusion

### 3.1 Model Evaluation

I always need a metric to evaluate the work I did. So, the same way, after I developed my models, I need a metric to validate the model I developed.

There am many metrics to evaluate, even, I have different metrics for classification problem and different metrics for regression problems.

For classification problems, I have metrics like:

* Confusion Matrix.
* Accuracy.
* Recall.
* Specificity.

For regression problems, I have metrics like:

* MSE.
* RMSE.
* MAPE.
* R square.

I am choosing RMSE and R square for my project. **Why RMSE over MAPE?**

Because, in RMSE, as the errors are squared before they are averaged, the RMSE gives a relatively high weightage to large errors, another reason is, RMSE penalizes large errors. For the above-mentioned reasons, I choose RMSE over MAPE.

#### 3.1.1 Root Mean Square Error (RMSE) & R square

I am going to use RMSE and R square as my error metrics to evaluate my models.

RMSE – Simply said, it is the sum of calculated errors.

R square – Simply defined, correlation of original and predicted values.

### 3.2 Model Selection

Finally, it’s My Model selection time. I developed three models. Linear Regression, Decision Tree and Random Forest

I am going to freeze Decision Tree with RMSE as 11.7 and R square as 0.57, even though Linear Regression has almost same RMSE value than Random Forest, because R square is My priority, it’s because, high R square is equivalent to minimizing sum of squared errors, on the other hand, minimizing RMSE may yield biased point predictions.

# Chapter 4

## Deployment

## For Python

* In anaconda pip install this

pip install pyinstaller

* After this type

pyinstaller --noconsole script.py

script is project name

* Then you can find your .exe(Window) under dist folder of your working directory.

Run Python script

Copy python installed directory path

Right Click on MY Computer

Click on Advance System settings

Click on Environment Variables

Path-> edit and paste the python code

Run CMD

Write python and load

If it shows you the version than run the Script.py

## For R

# If on R Server 9.0, load mrsdeploy package now

library(mrsdeploy)

# Create glm model with `mtcars` dataset

carsModel <- glm(formula = am ~ hp + wt, data = mtcars, family = binomial)

# Produce a prediction function that can use the model

manualTransmission <- function(hp, wt) {

newdata <- data.frame(hp = hp, wt = wt)

predict(carsModel, newdata, type = "response")

}

# test function locally by printing results

print(manualTransmission(120, 2.8)) # 0.6418125

##########################################################

# Log into Server #

##########################################################

# Use `remoteLogin` to authenticate with Server using

# the local admin account. Use session = false so no

# remote R session started

remoteLogin("http://localhost:12800",

username = "admin",

password = "{{YOUR\_PASSWORD}}",

session = FALSE)

##########################################################

# Publish Model as a Service #

##########################################################

# Generate a unique serviceName for demos

# and assign to variable serviceName

serviceName <- paste0("mtService", round(as.numeric(Sys.time()), 0))

# Publish as service using publishService() function from

# mrsdeploy package. Name service "mtService" and provide

# unique version number. Assign service to the variable `api`

api <- publishService(

serviceName,

code = manualTransmission,

model = carsModel,

inputs = list(hp = "numeric", wt = "numeric"),

outputs = list(answer = "numeric"),

v = "v1.0.0"

)

##########################################################

# Consume Service in R #

##########################################################

# Print capabilities that define the service holdings: service

# name, version, descriptions, inputs, outputs, and the

# name of the function to be consumed

print(api$capabilities())

# Consume service by calling function, `manualTransmission`

# contained in this service

result <- api$manualTransmission(120, 2.8)

# Print response output named `answer`

print(result$output("answer")) # 0.6418125

##########################################################

# Get Service-specific Swagger File in R #

##########################################################

# During this authenticated session, download the

# Swagger-based JSON file that defines this service

swagger <- api$swagger()

cat(swagger, file = "swagger.json", append = FALSE)

* # Now share this Swagger-based JSON so others can consume it

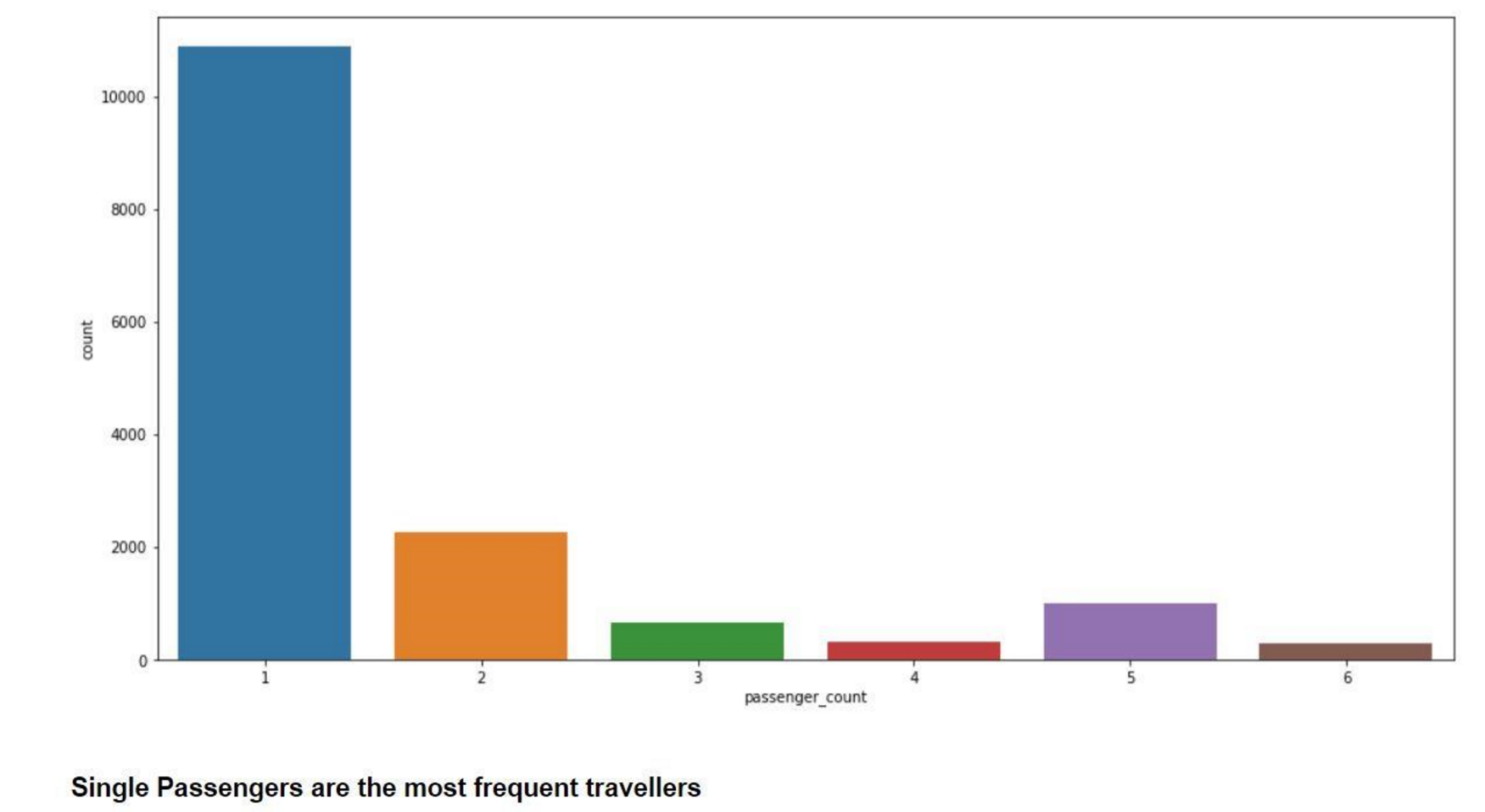
# Chapter 5

## **Summarize**

This project can help the business in achieving the strategic goals by predicting the fare amount of customers who will ride their cab rental services. It will give an idea to company about the fare amount.

They can use this to analyze their business.

## Appendix A – Extra Figures



## Appendix B – R Script

# Objective of My PROJECT - You am a cab rental start-up company. You have successfully run the pilot project and now want to

# launch My cab service across the country. You have collected the historical data from my pilot project and now have a requi

# -rement to apply analytics for fare prediction. You need to design a system that predicts fare amount for a cab ride in city.

# First step, as always, let's clean the environment rm(list = ls())

# Setting the working directory

setwd("C:/Users/Purushottam/Desktop/Data Science/Cab Fare Prediction Project")

# checking the working directory getwd()

# Loading the data into My R environment

train\_cab = read.csv("train\_cab.csv")

#~~~~~~~~~~~~~~~~Let's interact with My data and perform Exploratory Data

Analysis~~~~~~~~~~~~~~~~~~~~~~~~~~

class(train\_cab) # Its DataFrame

head(train\_cab) # Let's have a look on first 6 observations dim(train\_cab) # 16067 observations & 7 variables

str(train\_cab) # Have a look on the structure, fam\_amount and pickup\_datetime am as factors summary(train\_cab) # With a glance, I can get that, pickup\_latitude and passenger\_count have outliers names(train\_cab) # In names, I can get to see as perfect naming for respective variables.

# As observed, I have to change fare\_amount from factor to numeric train\_cab$fam\_amount = as.numeric(train\_cab$fare\_amount)

# Similarly, I am going to change pickup\_datetime from factor to datetime

# But first, let's replace UTC in pickup\_datetime variable with '' train\_cab$pickup\_datetime = gsub('\\ UTC', '', train\_cab$pickup\_datetime) train\_cab$Date = as.Date(train\_cab$pickup\_datetime)

#~~~~~~~~~~~~~~~~Here comes Missing Values~~~~~~~~~~~~~~~~~~~~~~~~~~~~~

# Line of code to know the sum of missing values in dataset sum(is.na(train\_cab))

# Line of code to know missing values in respective columns apply(train\_cab, 2, function(x) {sum(is.na(x))})

# Now, moving to calculate % of missing values

missing\_values$Missing\_values\_percentage =

(missing\_values$Missing\_values\_percentage/nrow(train\_cab)) \* 100

# I got idea that the missing values percentage is negligible, so I am dropping the observations with missing values

train\_cab = na.omit(train\_cab)

# Line of code to know the sum of missing values in dataset sum(is.na(train\_cab)) # Total number of missing values am 0

#~~~~~~~~~~~~~~~~~~Its time for OUTLIER analysis~~~~~~~~~~~~~~~~~~~~~

#deleting with boxplot

cnames = colnames(cabfare[,1:7])

for(i in cnames){

val = cabfare[,i][cabfare[,i] %in% boxplot.stats(cabfare[,i])$out]

cabfare= cabfare[which(!cabfare[,i] %in% val),]

}

# Next coming to latitudes and longitudes

# According to My learning, latitudes must range from -90 to +90 and longitudes must range from -180 to +180

# There am no outliers in pickup\_longitude, pickup\_latitude, dropoff\_longitude, dropoff\_latitude

# I have longitudes and latitudes. With the concept of feature engineering, I decided to create distance variable

# Using distCosine Formula, I can create distance variable.

# Calculates the geodesic distance betIen two points specified by radian latitude/longitude using the

modified\_distCosine <- function(Longitude1, Latitude1, Longitude2, Latitude2) {

if (any(is.na(c(Longitude1, Latitude1, Longitude2, Latitude2)))) {

NA

} else {

distCosine(c(Longitude1, Latitude1), c(Longitude2, Latitude2))

}

}

cabfare = mutate(cabfare, distance = mapply(modified\_distCosine,

pickup\_longitude, pickup\_latitude, dropoff\_longitude, dropoff\_latitude))

# Dimension Reduction - I am going to delete latitude and longitude variables as I obtained distance from these fMy variables

train\_cab = subset(train\_cab, select = -c(pickup\_latitude, pickup\_longitude, dropoff\_latitude, dropoff\_longitude))

#~~~~~~~~~~~~~~~~~~Time for Feature Selection~~~~~~~~~~~~~~~~~

# In Feature Selection, I perform Correlation Analysis and Anova test to find out the varaibles which am to be excluded # before feeding to the model

# Correlation Analysis is performed better num\_var (continuous independent variables) & fare\_amount (continuous target variable)

library(corrgram)

corrgram(train\_cab[,num\_var],order=FALSE,upper.panel = panel.pie, corrgram(cabfare[1:12156,1:4], order = F,

upper.panel=panel.pie, text.panel=panel.txt, main = "Correlation Plot")

#####Anova testing##########

anova\_one\_way = aov(distance~passenger\_count, data =cabfare)

summary(anova\_one\_way)

anova\_one\_way\_fare = aov(fare\_amount~passenger\_count, data =cabfare)

summary(anova\_one\_way)

#-->> From the result, I can observe, passenger\_count has p value > 0.05, by which, I accept null hypothesis.

# Deleting the below given continuous and categorical variables from day, as I found out they won't add any value to the model. train\_cab = subset(train\_cab, select = -c(passenger\_count))

#~~~~~~~~~~~~~~~~Let's jump to FEATURE SCALING~~~~~~~~~~~~~~~~~~

# Checking distance variable distribution using histogram hist(train\_cab$distance)

#--> The diagram represents, it is left skeId

Checking distance distribution

I will go for Normalization

#Normality check

qqnorm(cabfare$fare\_amount)

qq\_data = cabfare[, c("fare\_amount","distance")]

plot\_qq(qq\_data, sampled\_rows = 1000L)

#Normalisation

cnames = c("distance","fare\_amount","pickup\_datetime")

for(i in cnames){

print(i)

cabfare[,i] = (cabfare[,i] - min(cabfare[,i]))/

(max(cabfare[,i] - min(cabfare[,i])))

}

# Till now, I completed Exploratory Data Analysis, plotted visualizations, worked on Data Pre Processing, defined error metrics

#~~~~~~~~~~~~~~Now, I can go for MODEL DEVELOPMENT~~~~~~~~~~~~~~~~~

# Let's clean R Environment, as it uses RAM which is limited

library(DataCombine)

rmExcept("train\_cab")

# Defining Error Metric, which am used to evaluate the model

# Defining R Squam function - Correlation of original and predicted values.

Rsquam = function(y,y1){

cor(y,y1)^2

}

# Defining RMSE function - Root Mean Squamd Errors (Calculated Errors)

# RMSE over MAPE, Why? - Because, RMSE gives high weight to large errors during calculation and it penalizes large errors

# Let's create copy of data for further reference

train\_cab2 = train\_cab train\_cab = train\_cab2

# Now, as I know, I have to divide the data into train and test. So, let's go for that.

**I used Kfold in Python**

set.seed(123)

train\_index = sample(1:nrow(train\_cab),0.8\*nrow(train\_cab)) # Using Simple Random Sampling Technique train= train\_cab[train\_index,]

test= train\_cab[-train\_index,]

#--> I can see My train and test data am available in the environment

#~~~~~~~~~~~~~~It's time for DECISION TREE~~~~~~~~~~~~~~~~~~~~~~~~~~~

# Code to build Decision Tree

library(rpart)

####r part for regression###

fit=rpart(fare\_amount~.,data=cabfare,method="anova")

###prediction test

predictions=predict(fit,test[,-1])

############################ERROR MATRICS#################

library(DMwR)

mape=function (y,yhat){

mean(abs((y-yhat)/y))\*100

}

library(DMwR)

regr.eval(test[,1],predictions,stats=c('mae','rmse'))

# Calculating Rsquare for Test Data

DT = Rsquare(test[,1], predictions)

DT

#~~~~~~~~~~~~~~I am going for LINEAR REGRESSION MODEL~~~~~~~~~~~~~~~~

# Code for development of model

#run regression model

lm\_model = lm(fare\_amount ~., data = train)

#Summary of the model

summary(lm\_model)

#Predict

predictions\_LR = predict(lm\_model, test[,2:3])

#Calculate rSq.

LR= Rsquare(test[,1], predictions\_LR)

LR#~~~~~~~~~~~~~~Here I go for RANDOM FOREST~~~~~~~~~~~~~~~~~~~~~~~~

# Code for development of model

library(randomForest)

#model

rf\_model = randomForest(x=train[,-1],y=train$fare\_amount, importance = TRUE, ntree = 100)

#summary

summary(rf\_model)

#predictions

predict\_rf=predict(rf\_model,test[,-1])

#Error Matrics

rmse\_rf=measureRMSE(as.numeric(test$fare\_amount),as.numeric(predict\_rf))

RF= Rsquare(test[,1], predict\_rf)

RF

#~~~~~~~~~~~~~~~Let's work on test data~~~~~~~~~~~~~~~~~~~~~

# Now, it's time to bring the test data and perform the steps as performed on train\_cab and predict the fare\_amount

# Loading the data into My R environment

test\_cab = read.csv("test.csv")

cab\_test = read.csv('F:/test.csv', header = T)

str(cab\_test)

###############Converting Data types######################

cab\_test$pickup\_datetime = gsub( " UTC", "", as.character(cab\_test$pickup\_datetime))

cab\_test$pickup\_datetime <- as.numeric(as.POSIXct(cab\_test$pickup\_datetime,format="%Y-%m-%d %H:%M:%S"))

###############missing values analysis###########################

missing\_val = data.frame(apply(cab\_test,2,function(x){sum(is.na(x))}))

View(missing\_val)

###################################################################

#########BoxPLOT###################################

#####################################

################BOX PLOT###############

cnames = colnames(cab\_test[,1:6])

for(i in cnames){

val = cab\_test[,i][cab\_test[,i] %in% boxplot.stats(cab\_test[,i])$out]

#print(length(val))

cab\_test= cab\_test[which(!cab\_test[,i] %in% val),]

}

############Calculating Distance########

cab\_test = mutate(cab\_test, distance = mapply(modified\_distCosine,

pickup\_longitude, pickup\_latitude, dropoff\_longitude, dropoff\_latitude))

#####dropping longitutde and latitude##############

View(cab\_test)

cab\_test <- cab\_test[, -c(2:5)]

cab\_test <- cab\_test[, -c(2:2)]

#Normalisation

cnames = c("distance", "pickup\_datetime")

for(i in cnames){

print(i)

cabfare[,i] = (cabfare[,i] - min(cabfare[,i]))/

(max(cabfare[,i] - min(cabfare[,i])))

}

#Predicting Actual Test Data

predictions=predict(fit,cab\_test[,])

predictions

# Exporting the output to hard disk for further use

write.csv(test\_cab, "C:/Users/Purushottam/Desktop/Data Science/Cab Fare Prediction

Project/Cab Fare Prediction Output 2.csv", row.names = FALSE)

#--> Finally, I designed a model, which predicts the cab fare. My objective achieved.

**References**

* Edwisor.com
* Edwisor Community.
* Google
* Stackoverflows