**Hotel Recommendations**

**Problem Statement:**

Planning your dream vacation, or even a weekend escape, can be an overwhelming affair. With hundreds, even thousands, of hotels to choose from at every destination, it's difficult to know which will suit your personal preferences. Should you go with an old standby with those pillow mints you like, or risk a new hotel with a trendy pool bar?

Hotel Site wants to take the proverbial rabbit hole out of hotel site search by providing personalized hotel recommendations to their users. This is no small task for a site with hundreds of millions of visitors every month! Currently, Hotel Site uses search parameters to adjust their hotel recommendations, but there aren't enough customer specific data to personalize them for each user.

**Data Description:**

Hotel Site has provided us logs of customer behavior. These include what customers searched for, how they interacted with search results (click/book), whether or not the search result was a travel package.

Hotel Site is interested in predicting which hotel group a user is going to book. Hotel Site has inhouse algorithms to form hotel clusters, where similar hotels for a search (based on historical price, customer star ratings, geographical locations relative to city center, etc) are grouped together. These clusters serve as good identifiers to which types of hotels people are going to book, while avoiding outliers such as new hotels that don't have historical data.

Our goal is to predict the booking outcome (hotel cluster) for a user event, based on their search and other attributes associated with that user event.

**Aim of the Project:**

Aim In this project, It is challenging us to contextualize customer data and predict the likelihood a user will stay at 100 different hotel groups.

**Approach of the Project:**

**i) Exploratory Data Analysis**

Exploratory data analysis is the process of analysing the dataset to understand its characteristics. In this step, we will figure out the following.

a. days stayed by customer

b. Are there any NAN values?

c. Find outliers

d. Correlations

e. Check for duplicate values

**ii) Imputation**

Imputation is the process of filling the missing values in appropriate ways so that we don't lose much data. The detected missing values are filled appropriately

**iii) Further EDA and Visualizations to answer few questions**

Further Exploratory data analysis includes raising a few questions on the dataset and finding out answers for the same using various visualization and data manipulation techniques. This helps us to understand more about the dataset

**iv) Baseline accuracy calculation**

Baseline accuracy score provides the required point of comparison when evaluating all other machine learning algorithms.

**v) Model Building Evaluation:**

Model Building and Evaluation Since this is a classification problem various classification algorithms are tried and the one that is best suited for our data is considered. The models that we build for our dataset are

a. Random Forest classification

b. Guassian Naive Bayes classification

c. Logistic Regression

d. KNN classification

e. XGBoost classification

f. Decision Tree classification

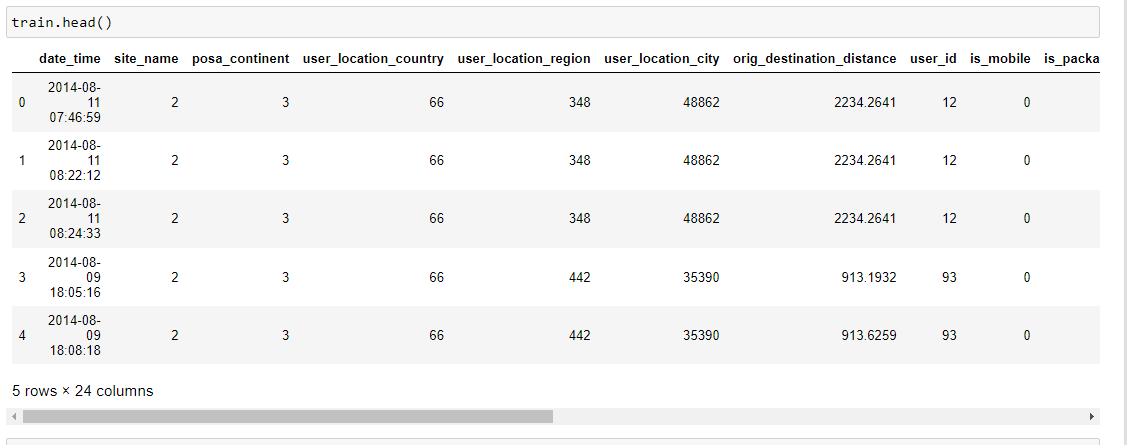
**vi) Solutionization**

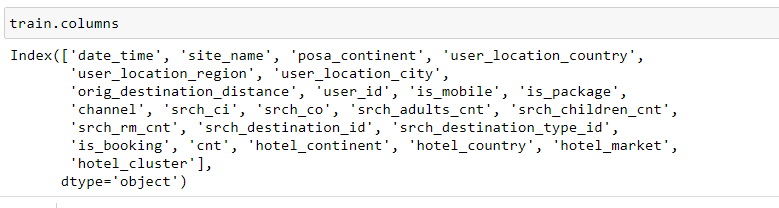
In this step, the result of the best suited model is mended a little further to produce better results.

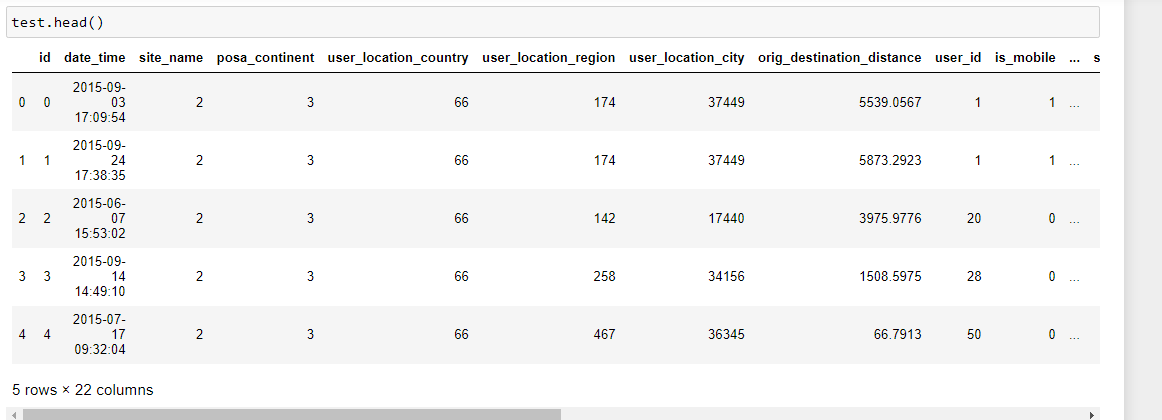
**2. Data Analysis.**

**The Dataset contains 37 million logs of customer behaviour, to avoid hang and lagging of system I have selected 100000 logs and entries from each of the dataset i.e train and test dataset.**

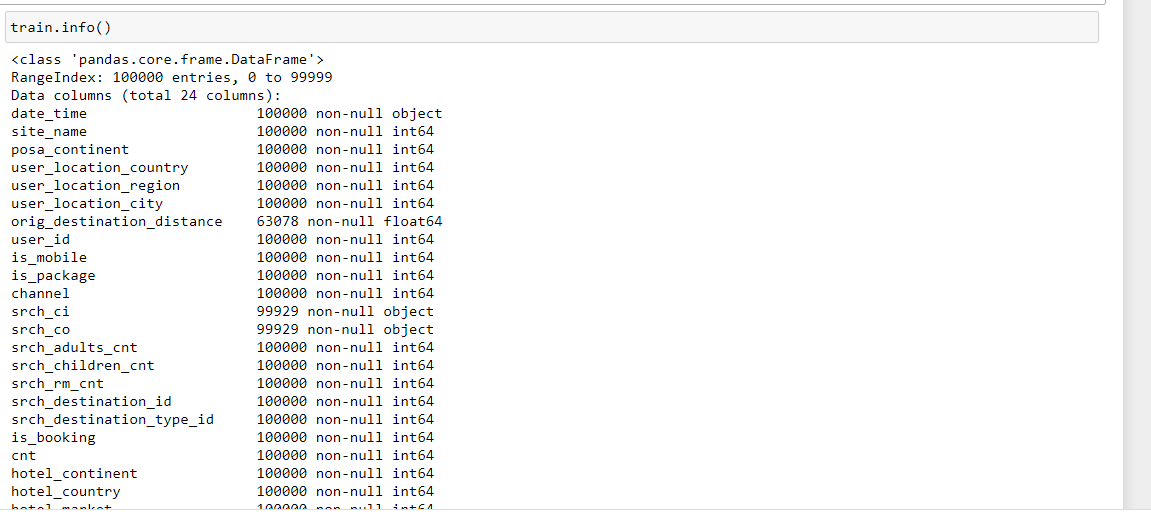
****

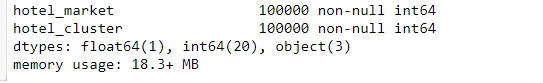
****

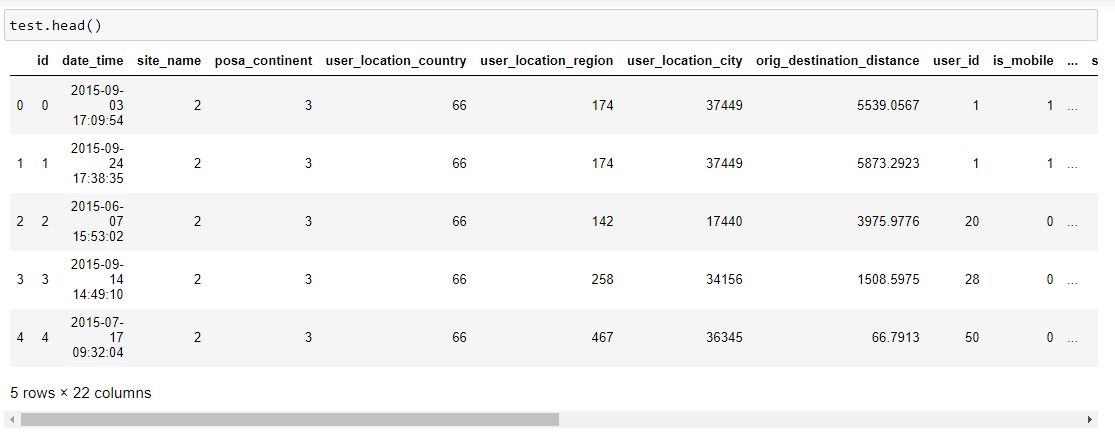
****

****

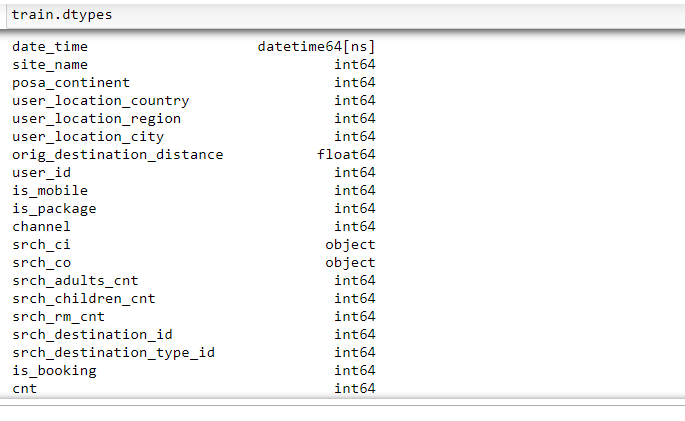
**All the dates in test.csv are 1 year later than the dates in train.csv .The testing set contains dates from 2015 ,and the training set contains dates from 2013 and 2014.**

****

****

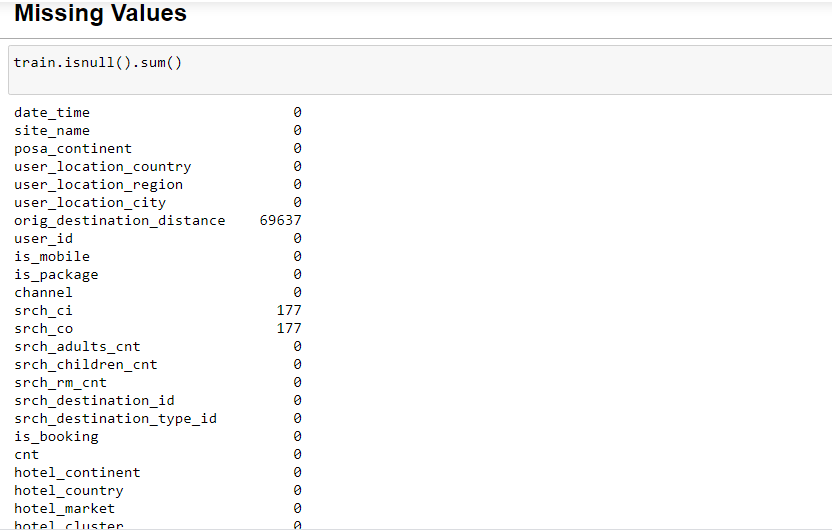
****

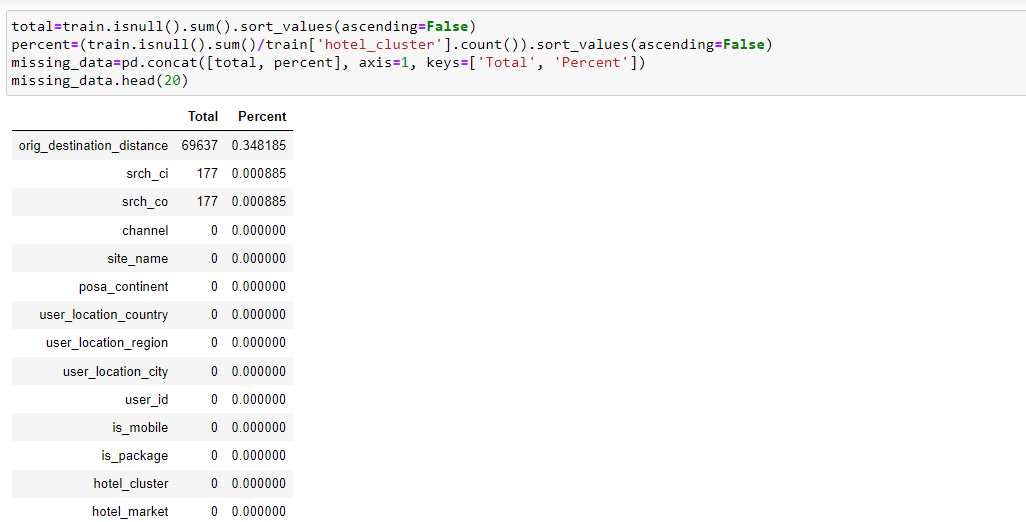
****

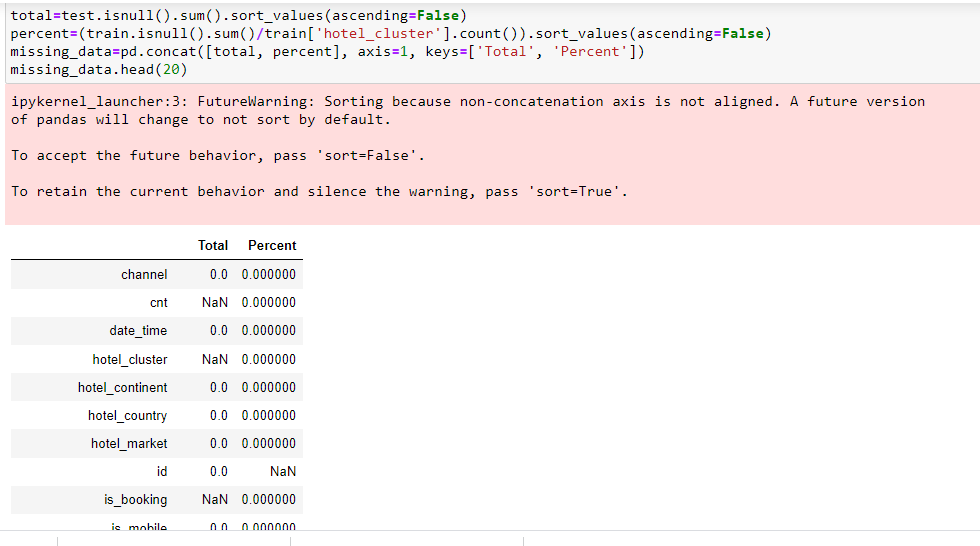
****

## What we're predicting?

**Our goal is to predict the booking outcome (hotel cluster) for a user event, based on their search and other attributes associated with that user event. According to the description given by Hotel, there are 100 clusters in total.**

****

****

****

****

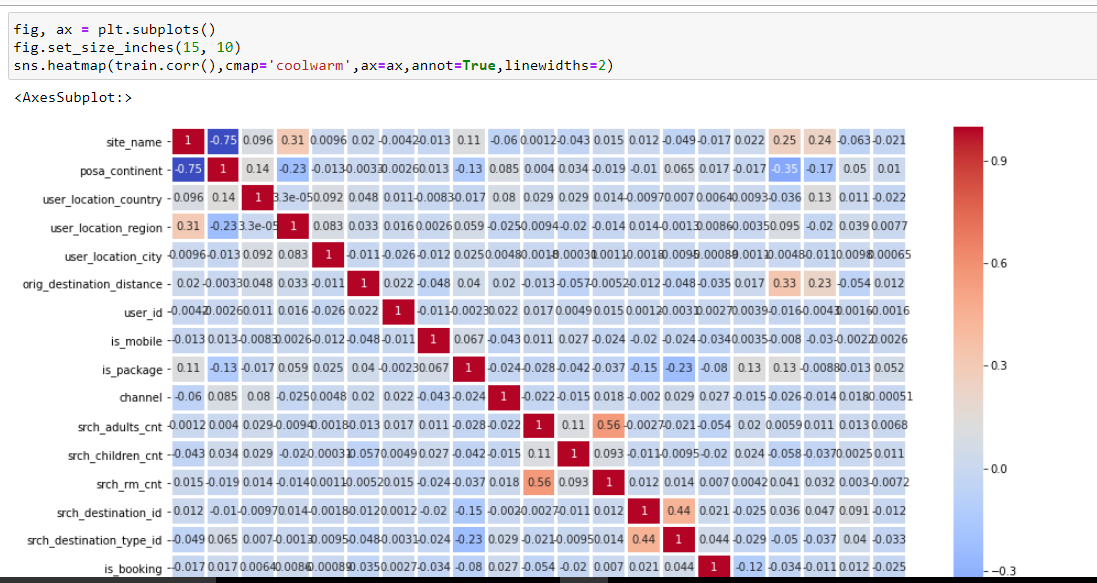
****

****

**3. Visualization:**

### 1. Corelation between the columns

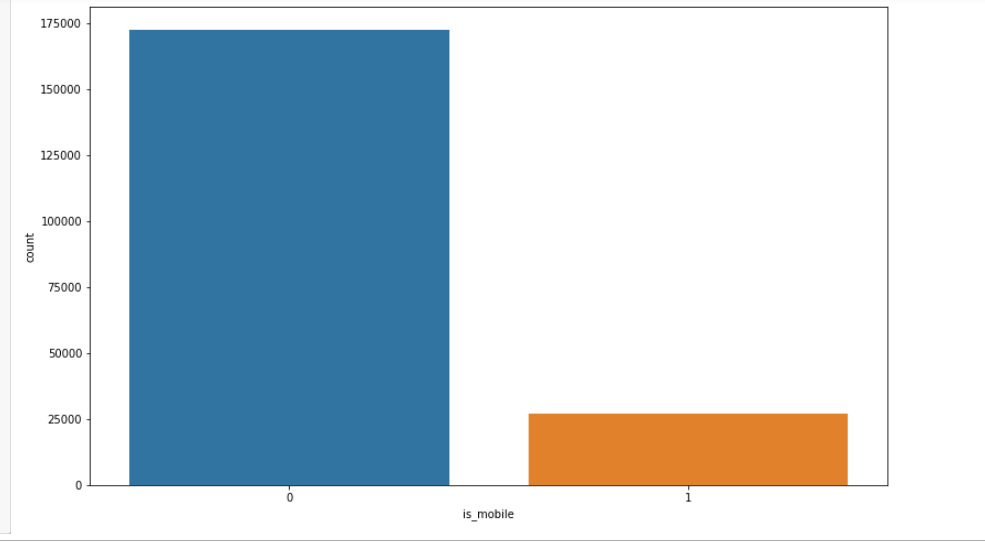
**A Heatmap to infer the correlation between different columns with each other and to check if there is any linear correlation between any columns.**

****

**This tell us that no column correlate linearly with any other column in the dataset.**

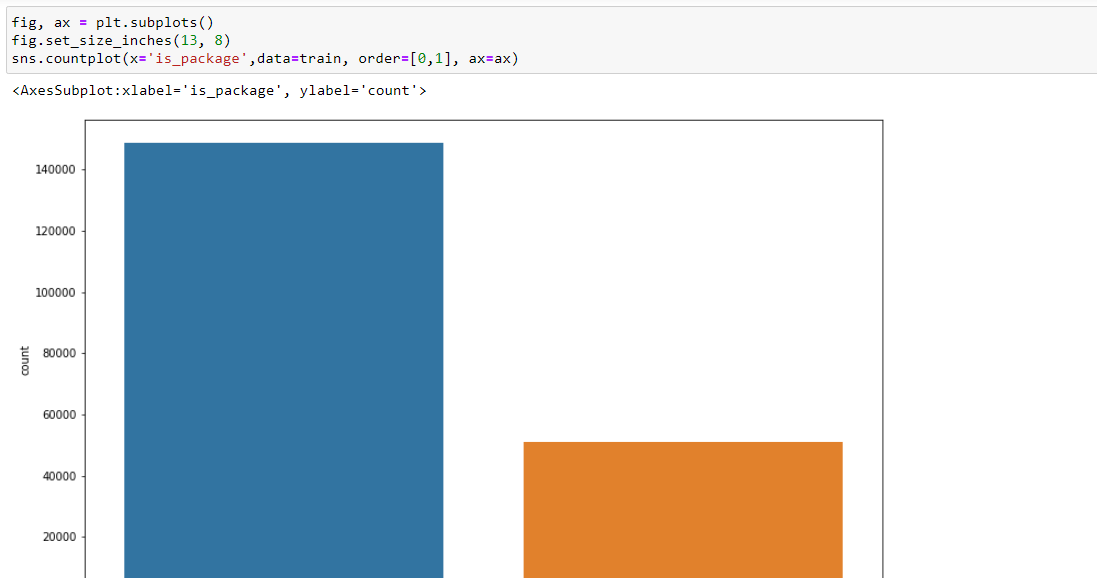
### 2. Frequency of bookings through mobile

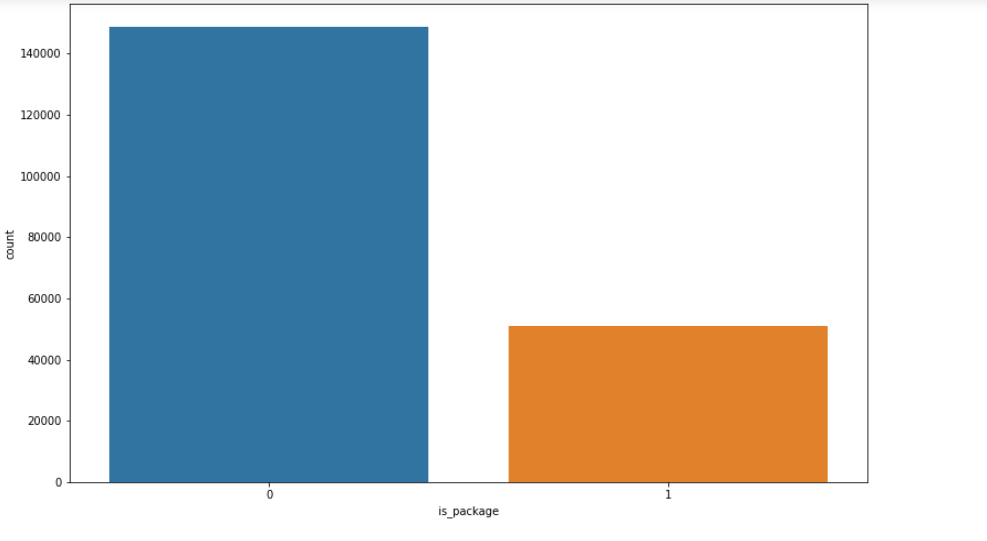
****

****

**Customers book their hotels from Hotel Website rather than mobile. So we can suggest Hotel to make more interactive, easy to use website so customers can easily book their hotels.**

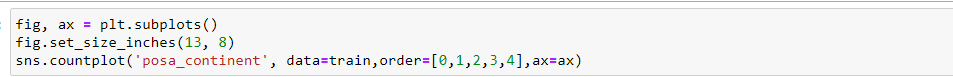
### 3. Frequency of bookings with total package(i.e with travel)

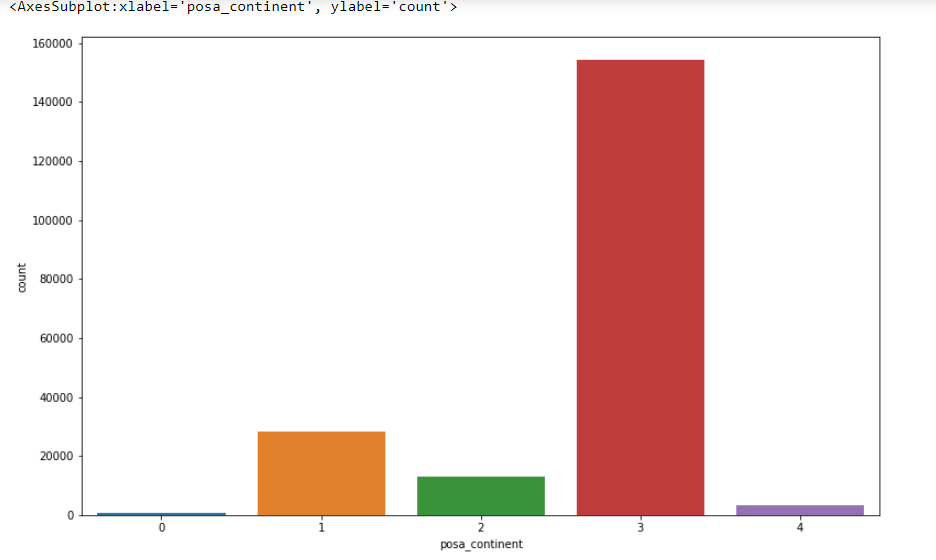




**The number of customers booking their hotels with total package i.e with travel are less. We can interpret this that there can be many reasons of this like booking with travel is getting very expensive or less range of hotels where the customer wants to travel to.**

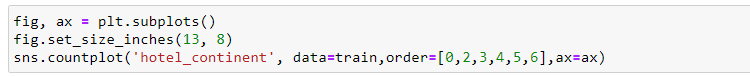
### 4. Frequency of Continent site used for booking

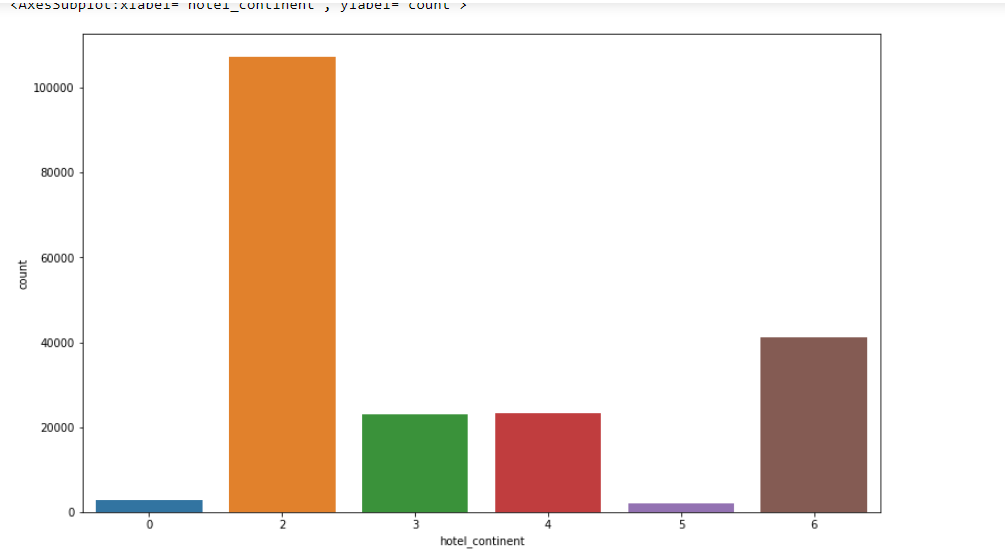




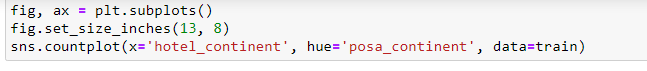
**Most of people booking are from continent 3. Maybe people there have more expending power. So Hotel can increase its business by increasing more hotel options, more variety, better user experience, etc. For other continents Hotel can lower its prices on hotel or give discounts, loyalty points.**

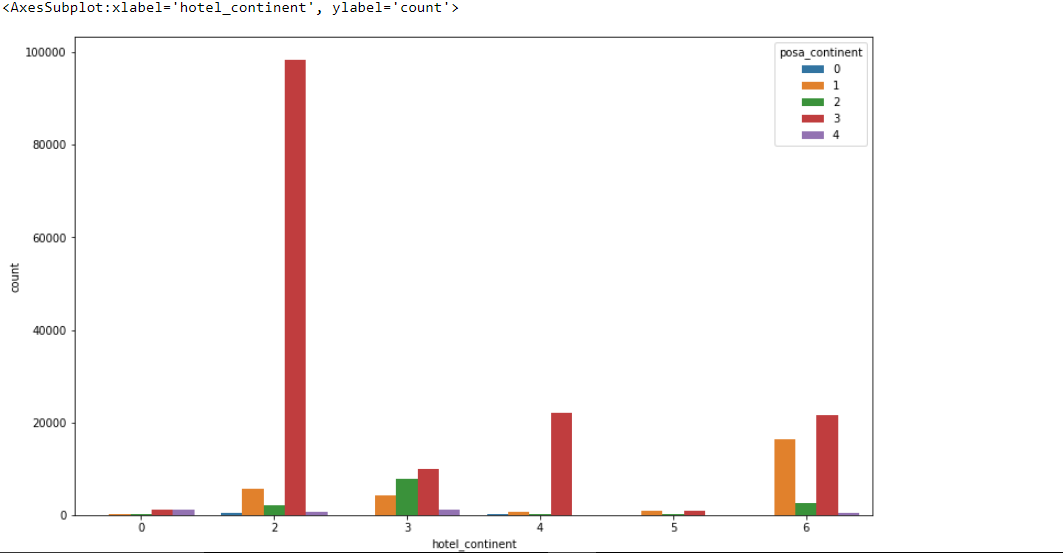
### 5. Which continent has the most number of hotels?



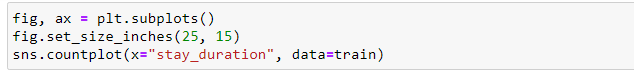


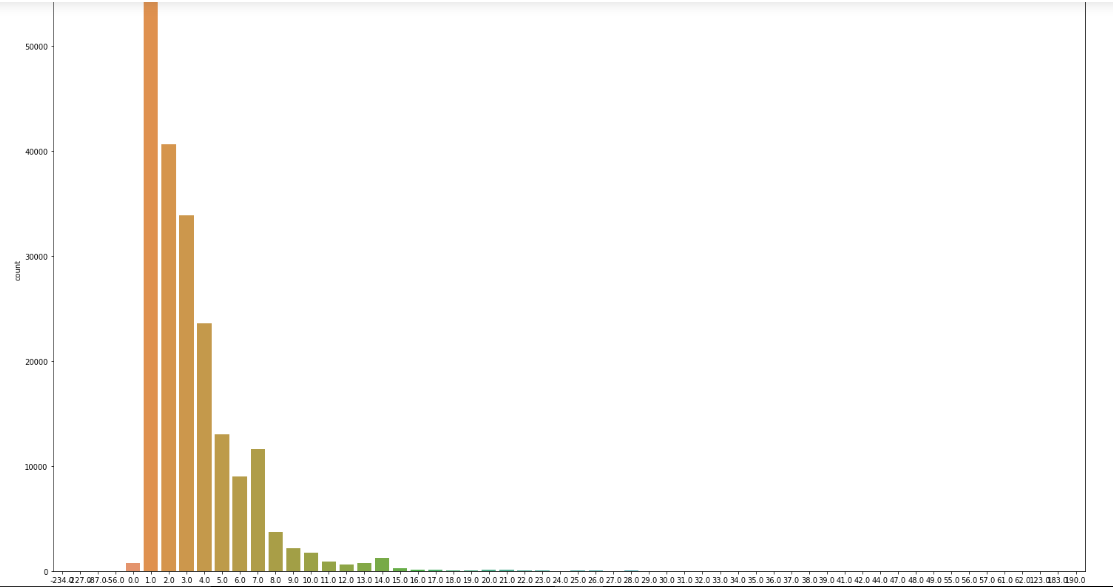
### 6. Putting the two above together

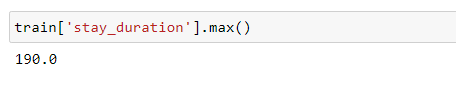




### 7. Frequency of stay duration booked by customers

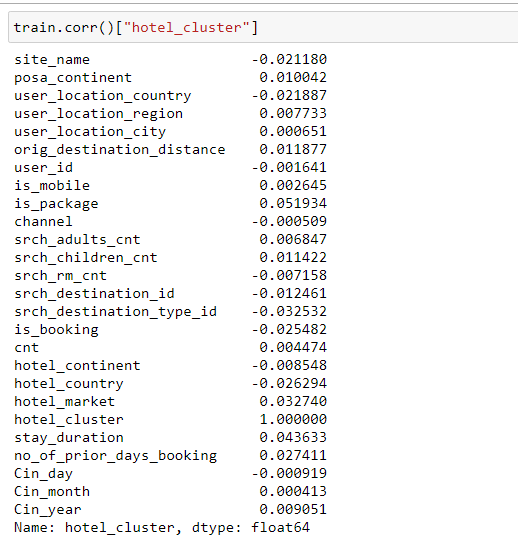




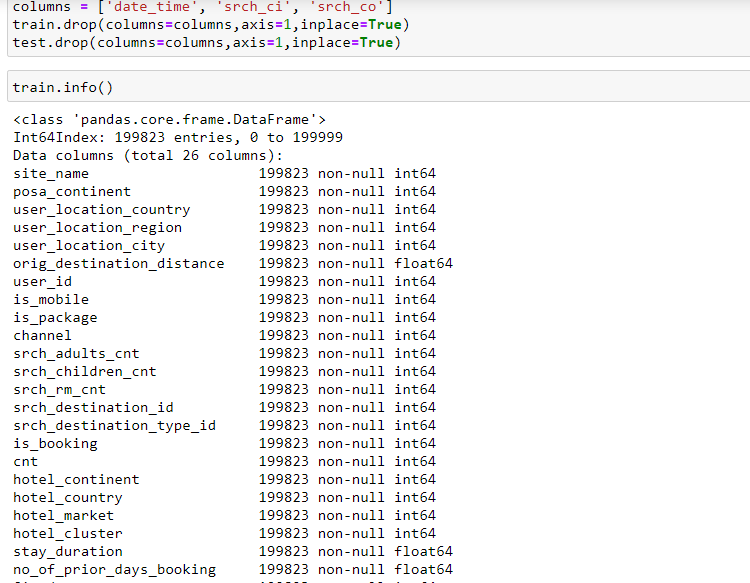


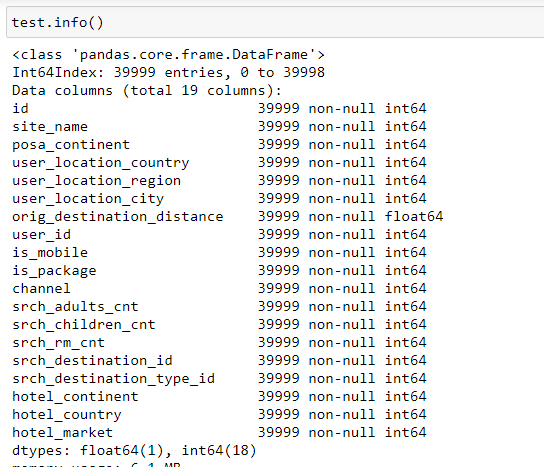
**4. Data Cleaning:**

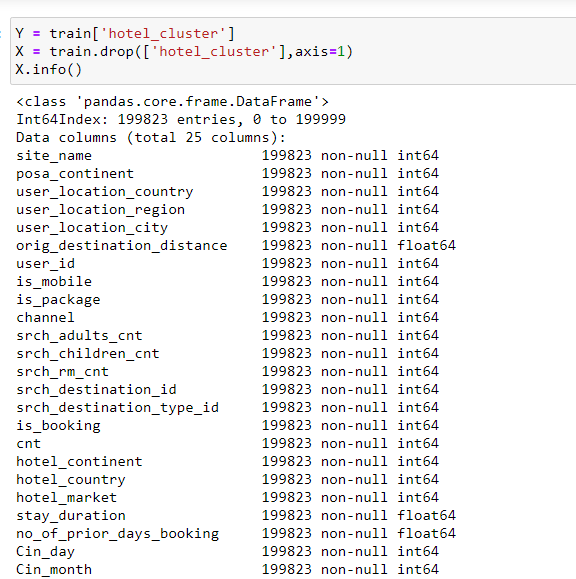
**Data cleaning is the process of cleaning / standardising the data to make it ready for analysis. Most of times, there will be discrepancies in the captured data such as incorrect data formats, missing data, errors while capturing the data.**

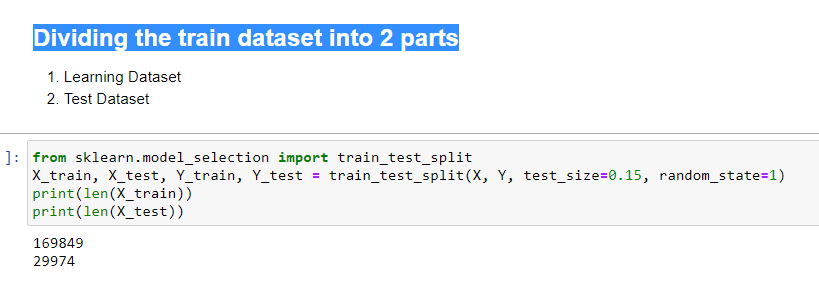
****

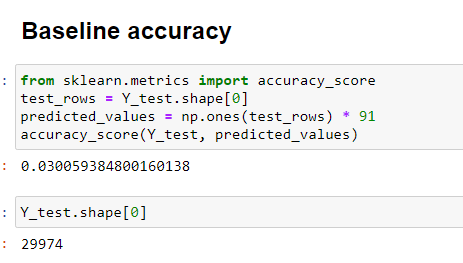
**Removing unnecessary columns**

****

****

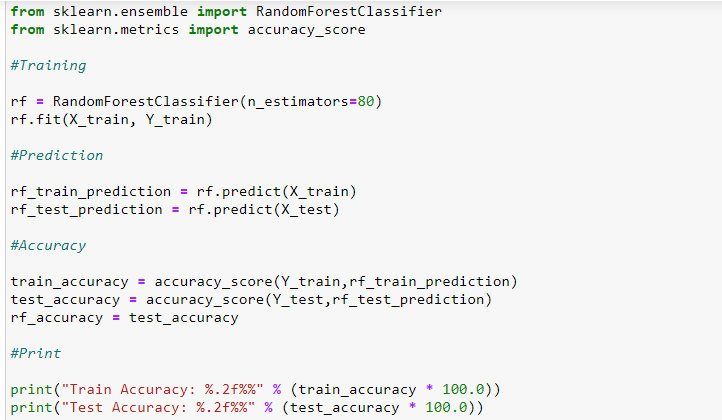
****

****

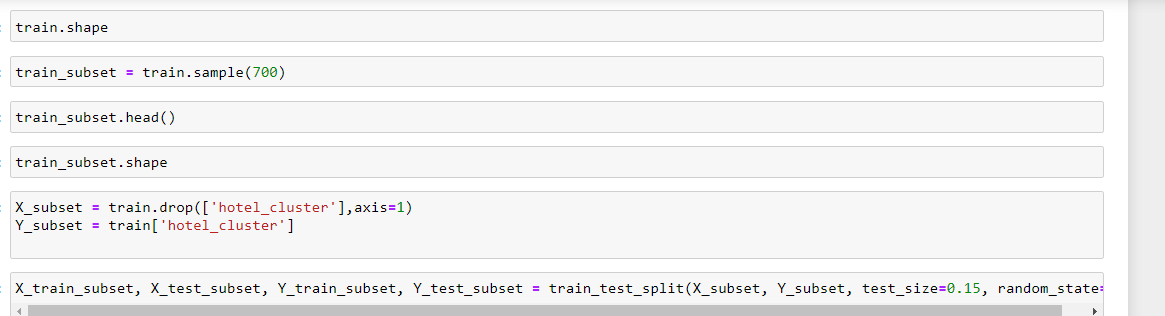
****

**5. Building Machine Learning Models.**

## 1. Random Forest

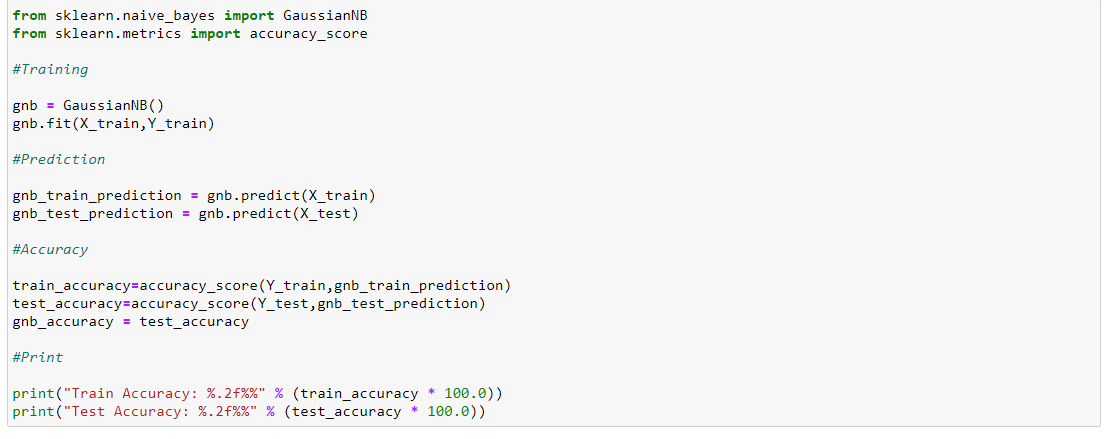
****

****

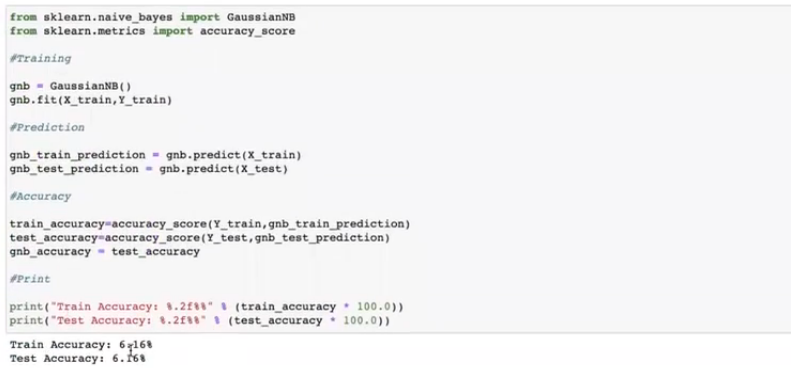
****

## 2. Second Algorithm - Guassian Naive Bayes

**In statistics, naive Bayes classifiers are a family of simple "probabilistic classifiers" based on applying Bayes' theorem with strong independence assumptions between the features.**

****

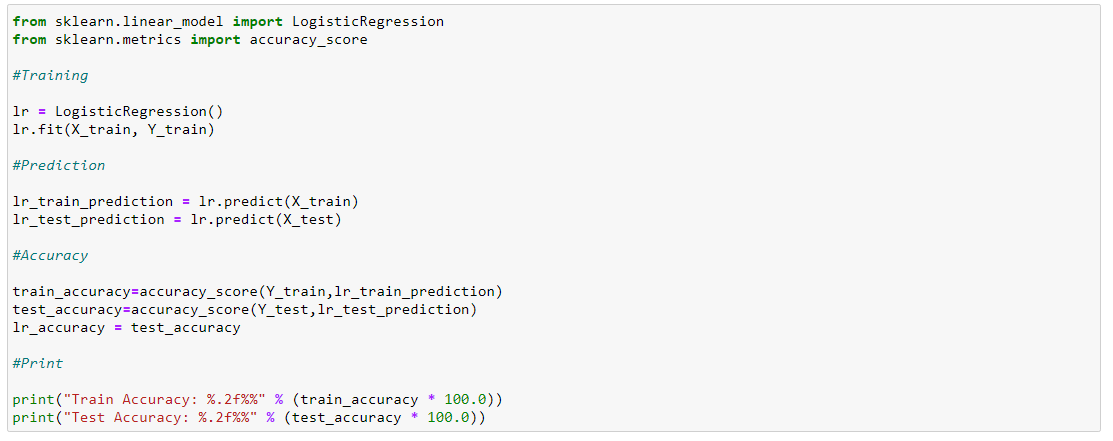
**Naive Bayes are not good enough to generalize over the entire data and produce very less results compared to other methods. This can be due to the fact that most features in the dataset are either one-hot vectors or have missing values, both of which become hindrances in achieving a good solution using this method.**

****

## 3. Third Algorithm - Logistic Regression

**Logistic regression is a statistical model that in its basic form uses a logistic function to model a binary dependent variable, although many more complex extensions exist. In regression analysis, logistic regression (or logit regression) is estimating the parameters of a logistic model (a form of binary regression).**

**The model is used to model the probability of a certain class or event existing such as pass/fail, win/lose, alive/dead or healthy/sick. This can be extended to model several classes of events such as determining whether an image contains a cat, dog, lion, etc.**

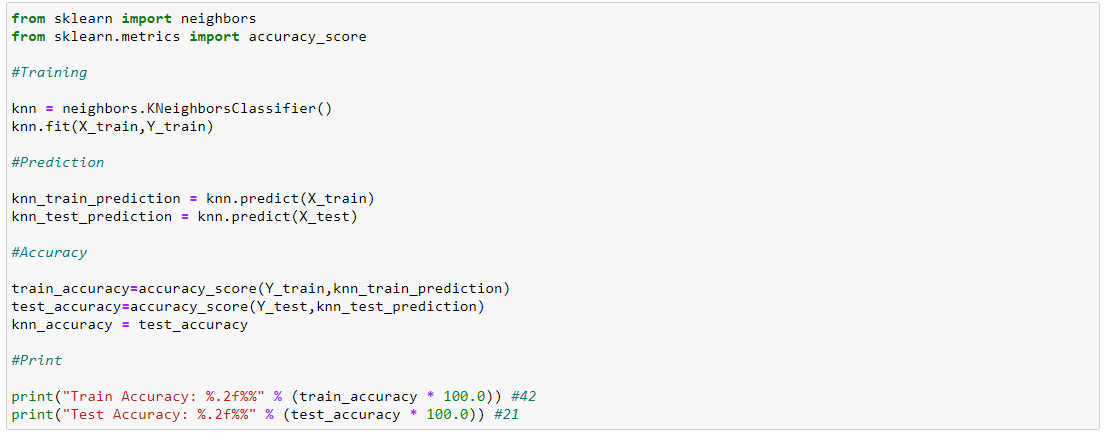
****

****

**The accuracy still goes bad for logistic regression.**

## 4. Fourth Algorithm - KNN

**K-Nearest Neighbors (KNN) is one of the simplest algorithms used in Machine Learning for regression and classification problem. KNN algorithms use data and classify new data points based on similarity measures (e.g. distance function). Classification is done by a majority vote to its neighbors.**

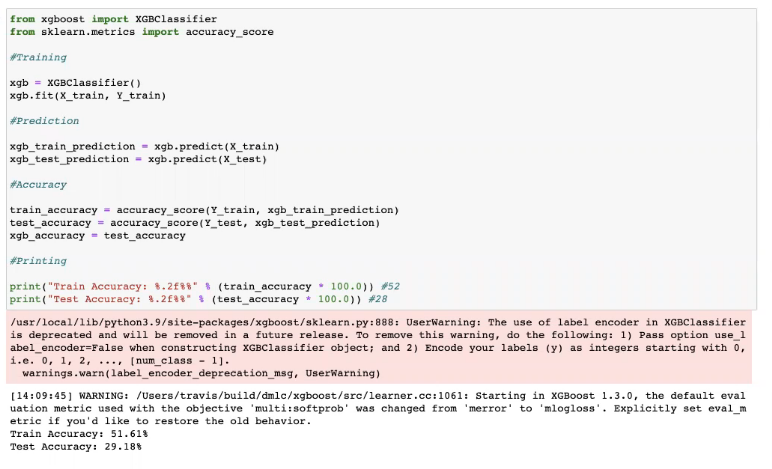
****

**Accuracy for KNN:**

****

## 5. Fifth Algorithm - XGBoost

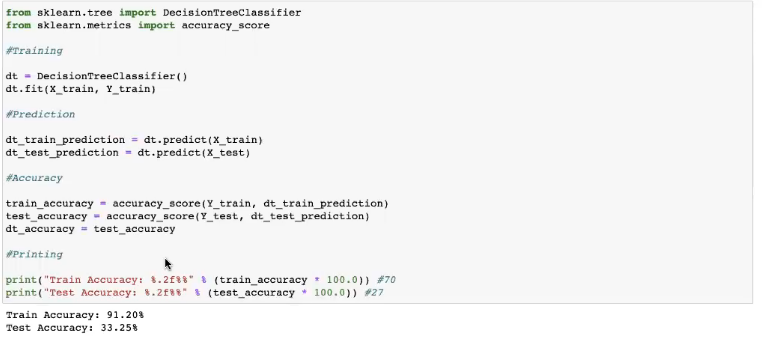
**XGBoost is a decision-tree-based ensemble Machine Learning algorithm that uses a gradient boosting framework.**

****

**XGBoost has a performance very slightly less to that of Random Forest. This is because both these methods work in a similar manner, handling both categorical features efficiently.**.

## 6. Sixth Algorithm - Decision Tree

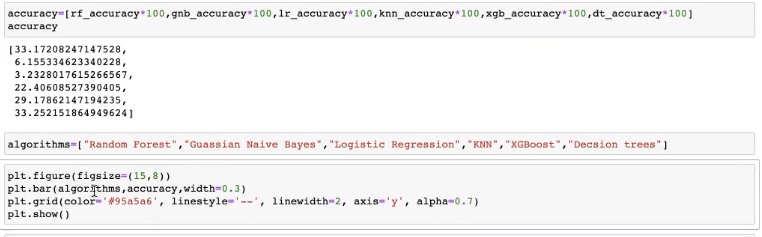
**A decision tree is a decision support tool that uses a tree-like model of decisions and their possible consequences, including chance event outcomes, resource costs, and utility. It is one way to display an algorithm that only contains conditional control statements.**

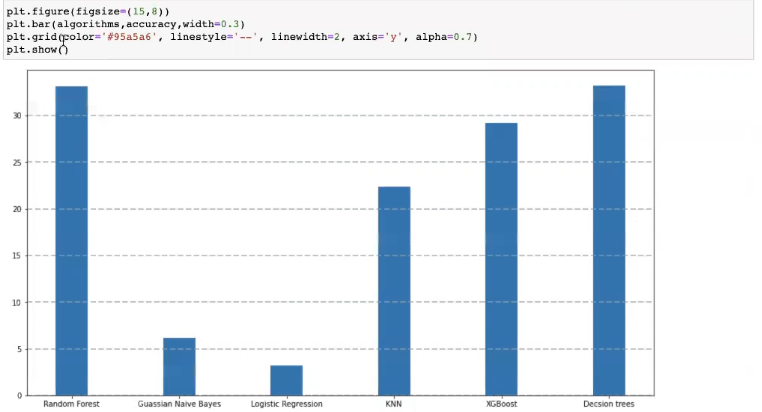
****

**Model Comparison:**

****

**Random Forest and XG Boost are very close.**

****

****

# Solutionising and Conclusion:

# Lets take our random forest model & explore the solutionising a bit more

# 

# 

# 

# 

# 

# 

# 

# 

# 

# 

# 

# 

# 

# 

# 

# Feature engineering which includes identification, creation and deletion of unwanted features is important. In this particular problem statement, the most important is the approach & outlook towards the problem statement.