

**Port City International University**

**(PCIU)**

**Computer Science and Engineering**

**Course Title: Pattern Recognition Sessional**

**Course Code: CSE 332**

**Submitted To**

**Shafayet Nur**

**Teaching Assistant**

**Computer Science & Engineering**

**Port City International University**

**Submitted By**

**SM Shahrier Kawsar**

**ID: CSE 01806740**

**Batch: CSE 18(C)-Day**

**Date of Submission: 09-02-2022**

**Experiment No:**1

**Experiment Name:** Visualization, mean, median, & standard deviation.

**Objectives:** To visualize data & find mean, median & standard deviation from Melbourne Housing prices dataset.

**Theory:**

Statistics is a field of mathematics that pertains to data analysis. Statistical methods and equations can be applied to a data set in order to analyze and interpret results, explain variations in the data, or predict future data. A few examples of statistical information we can calculate are:

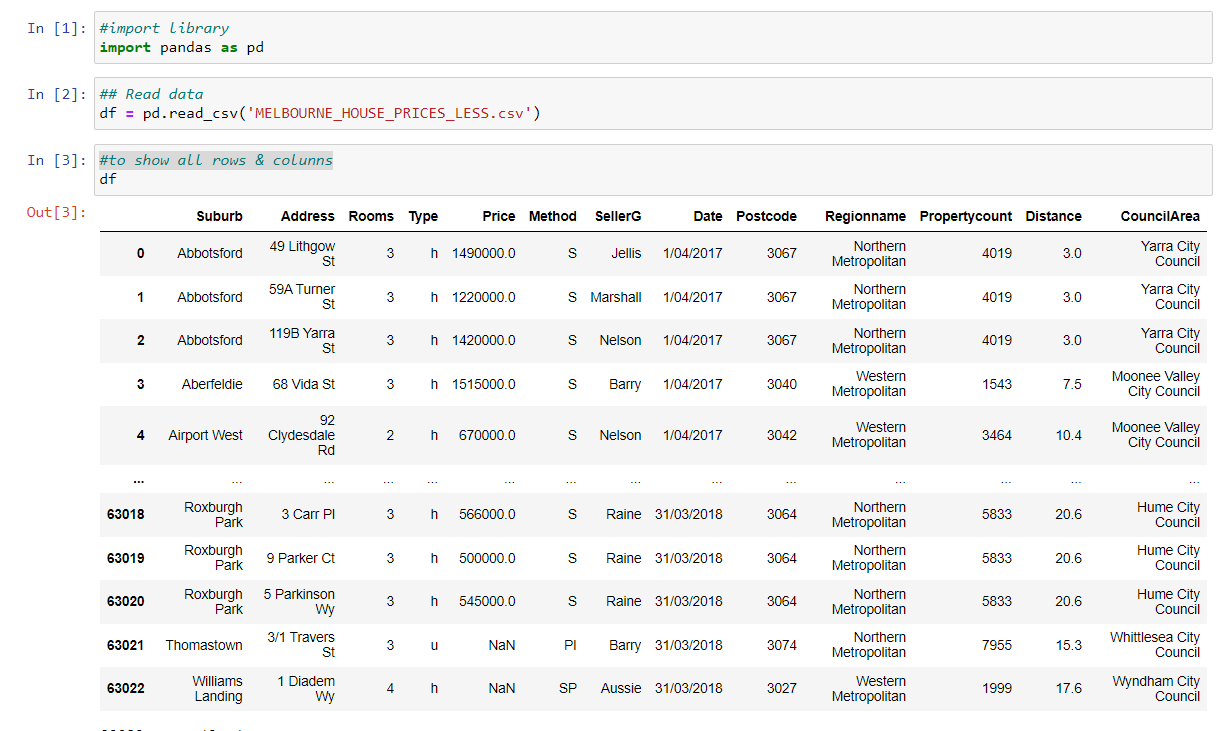
* Average value (mean)
* the middle value in a list ordered from smallest to largest. (median)
* On average, how much each measurement deviates from the mean (standard deviation of the mean)

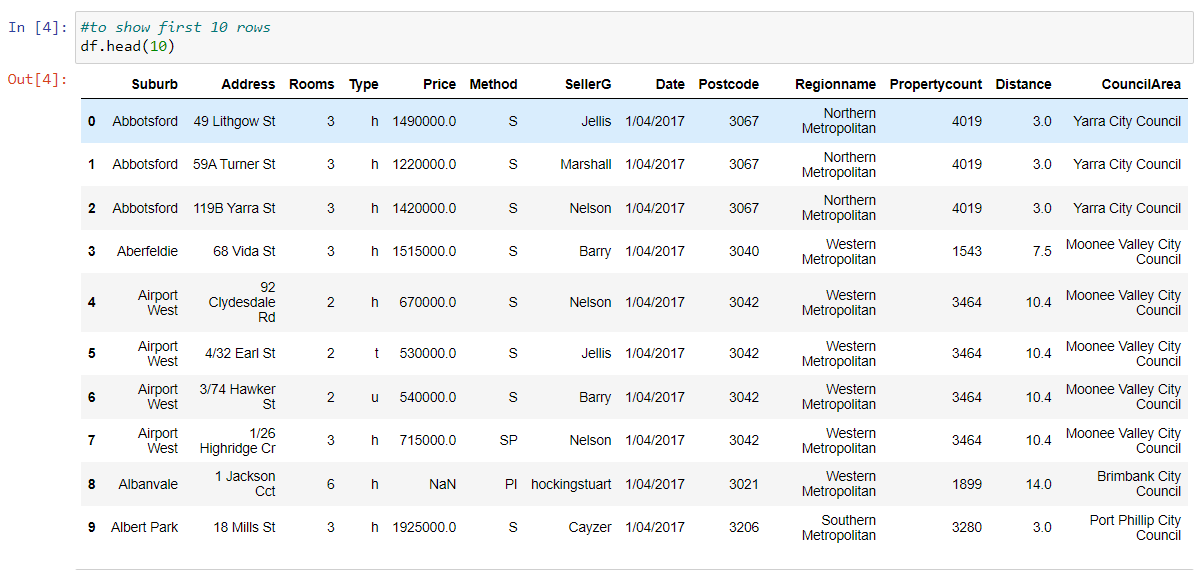
Suppose we draw a sequence of n samples of a random variable X: x1, x2, ..., xn. We can define mean (or arithmetic average) value of the sequence in the traditional way as (x1+x2+xn)/n. As we grow the size of the sample, we will obtain the mean (also called expectation) of the distribution. We will denote expectation by E(x).

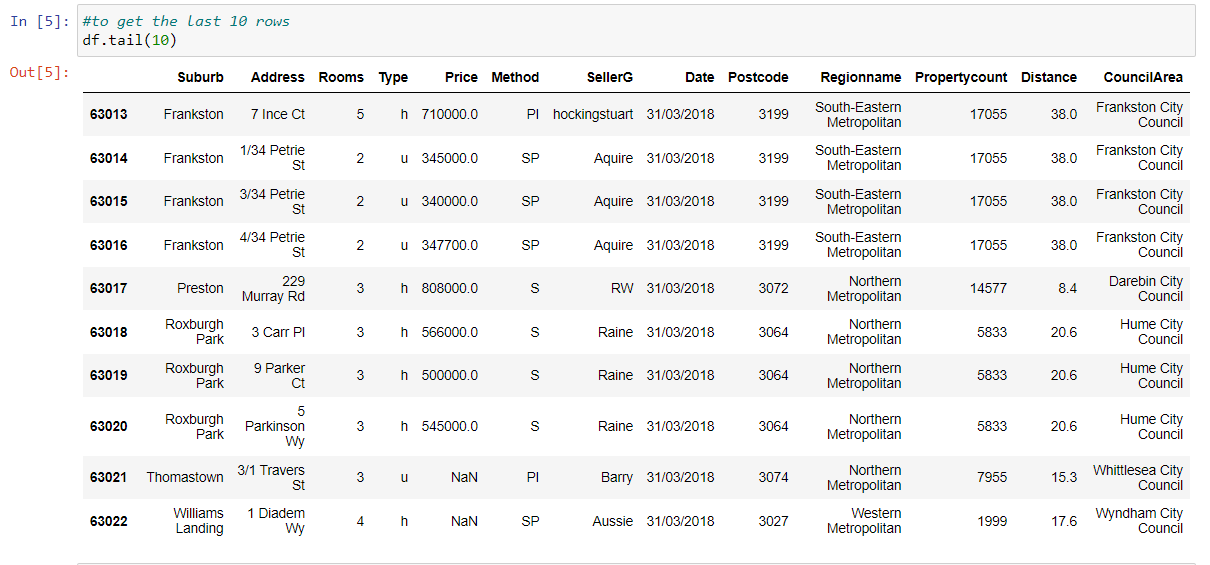
It can be demonstrated that for any discrete distribution with values {x1, x2, ..., xN} and corresponding probabilities p1, p2, ..., pN, the expectation would equal to E(X)=x1p1+x2p2+...+xNpN.

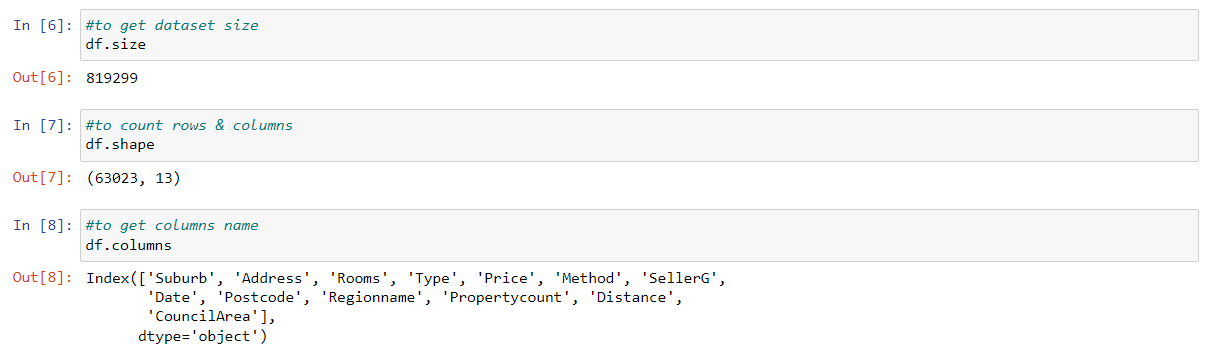
To identify how far the values are spread, we can compute the variance σ2 = ∑(xi - μ)2/n, where μ is the mean of the sequence. The value σ is called standard deviation, and σ2 is called a variance.

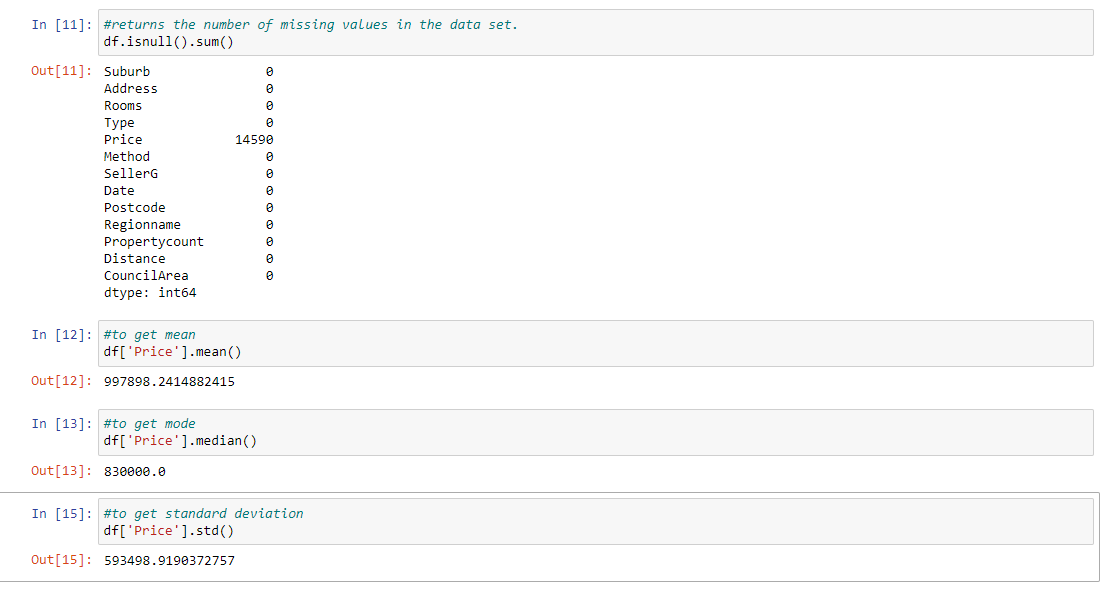
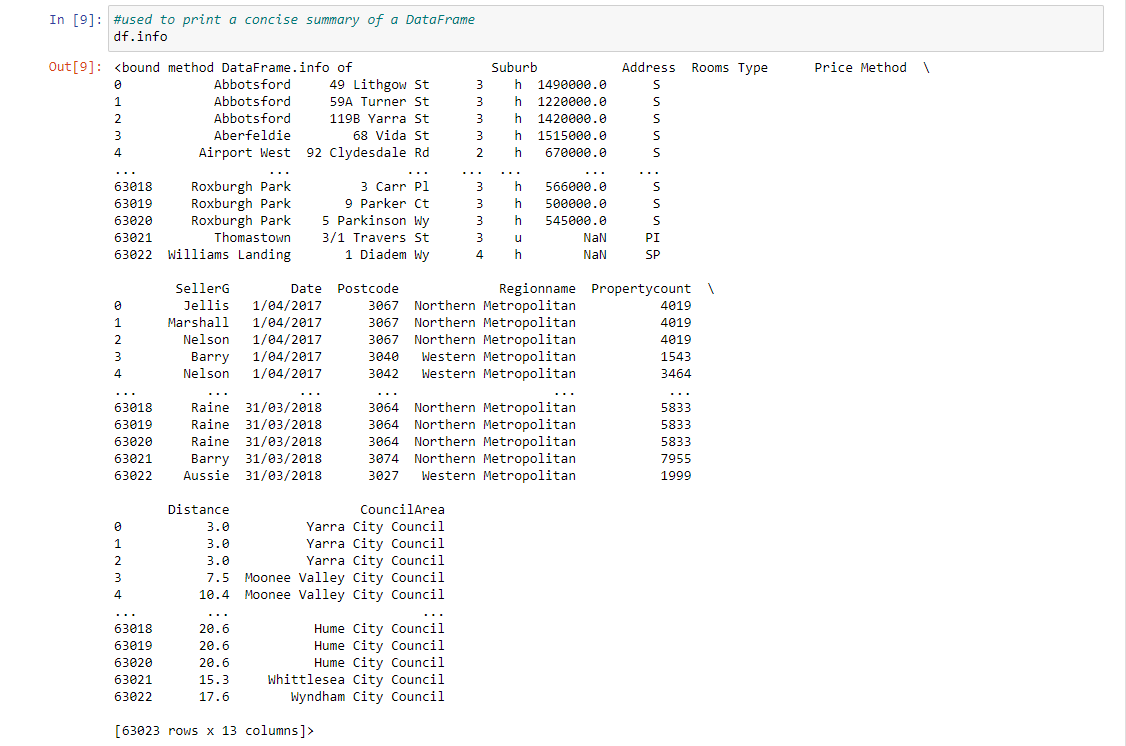
**Description:**

****

****

****

****

****

**Discussion:** In this lab, we have compute mean, median & standard deviation from a dataset using pandas’ library.

**Experiment No:**2

**Experiment Name:** House price prediction using simple linear regression.

**Objectives:** To predicate house price from”kc\_house\_price” using simple linear regression.

**Theory:**

Linear Regression is a supervised machine learning model that attempts to model a linear relationship between dependent variables (Y) and independent variables (X). Simple linear regression works by using the traditional slope-intercept form in which a and b are two coefficients that are elaborated “learn” and find the accurate predictions. In the below equation, X stands for input data, and Y stands for prediction.

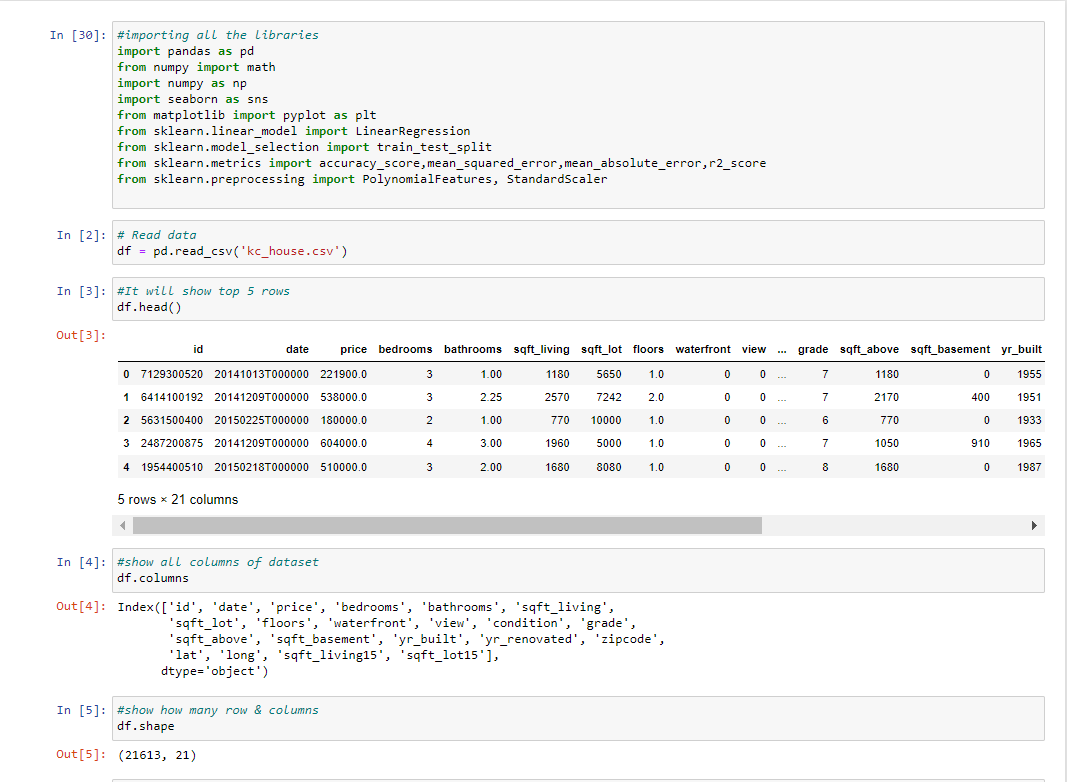
Y= bX + a

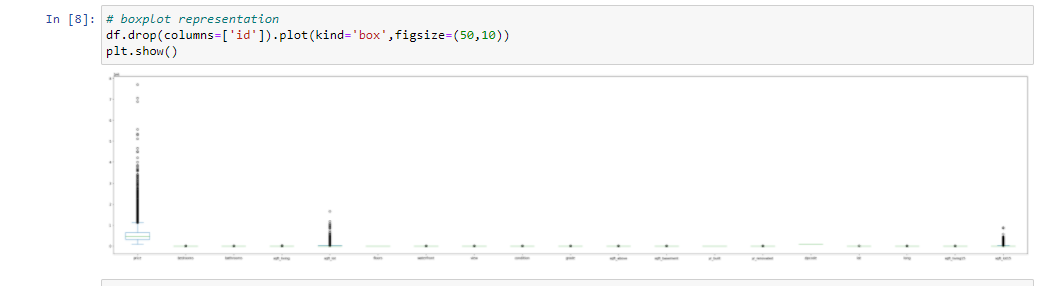
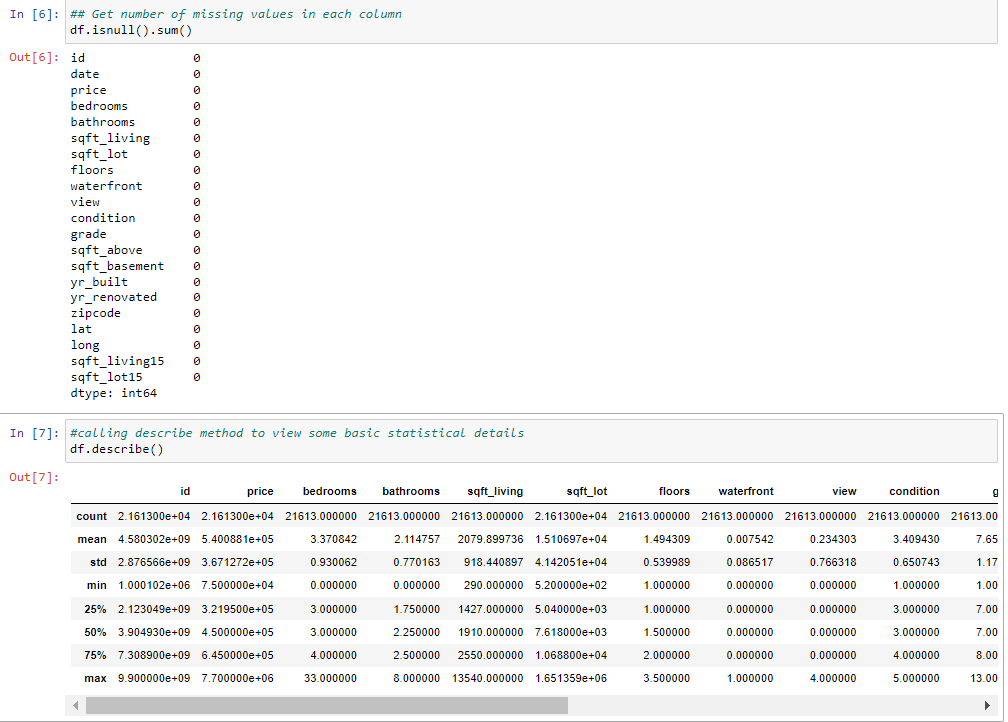
Now we need to consider every step for the house price prediction using simple linear regression. Consider a company of real estate with datasets containing the property prices of a specific region. The price of a property is based on essential factor like bedrooms.  Majorly, a real estate company requires:

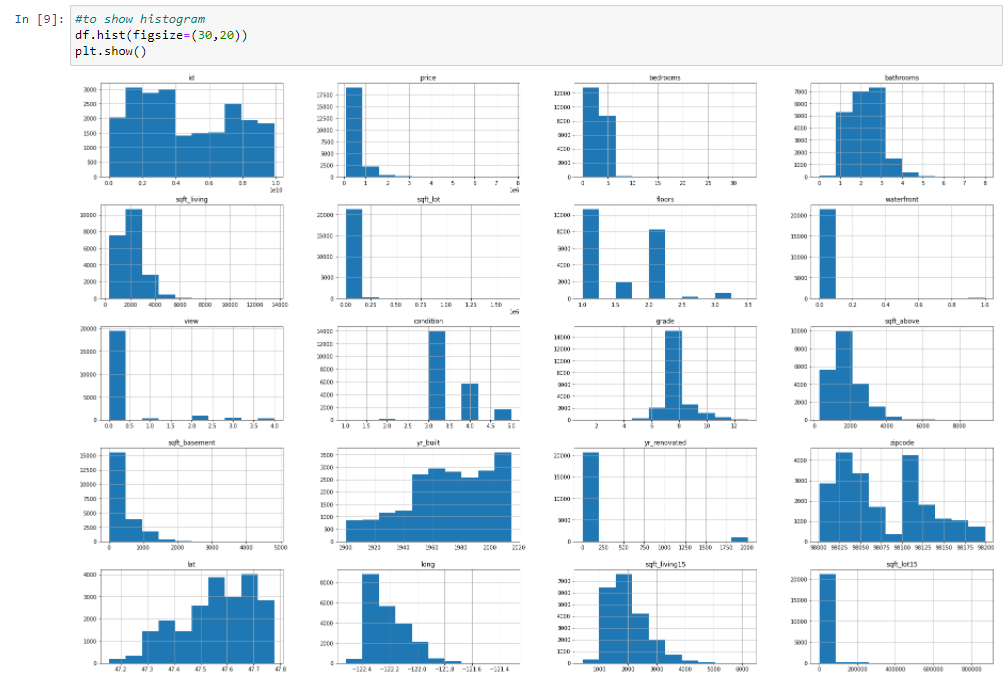
Find the variable that affects the price of a house.

For finding the accuracy of a model, that means how well the variables can predict the prices of a house.

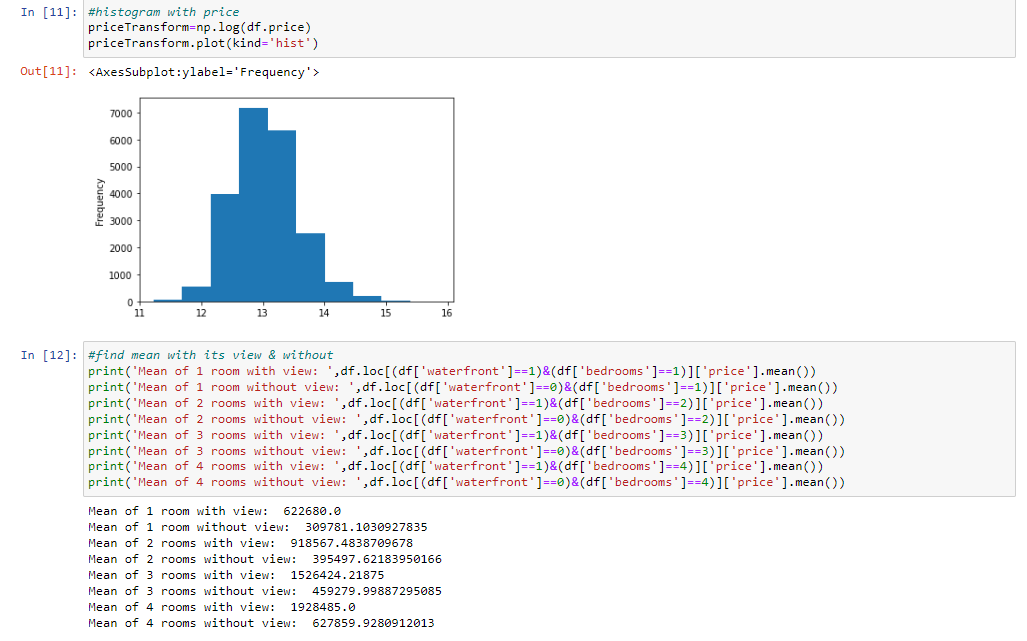
**Description:**

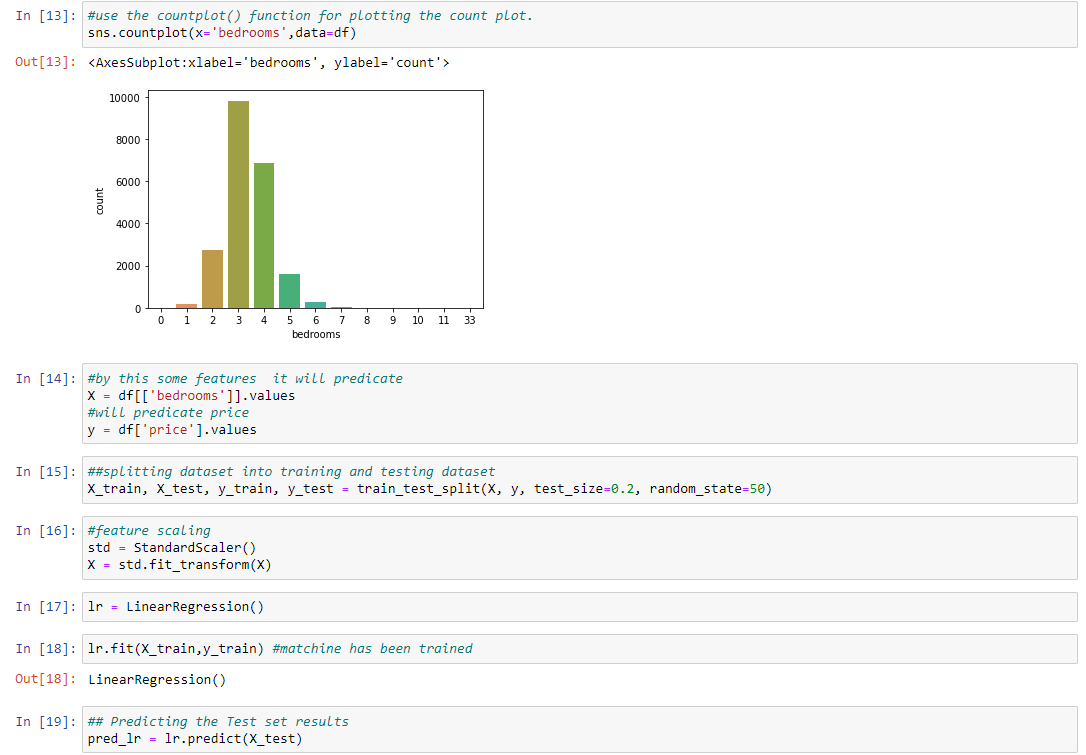
****

****

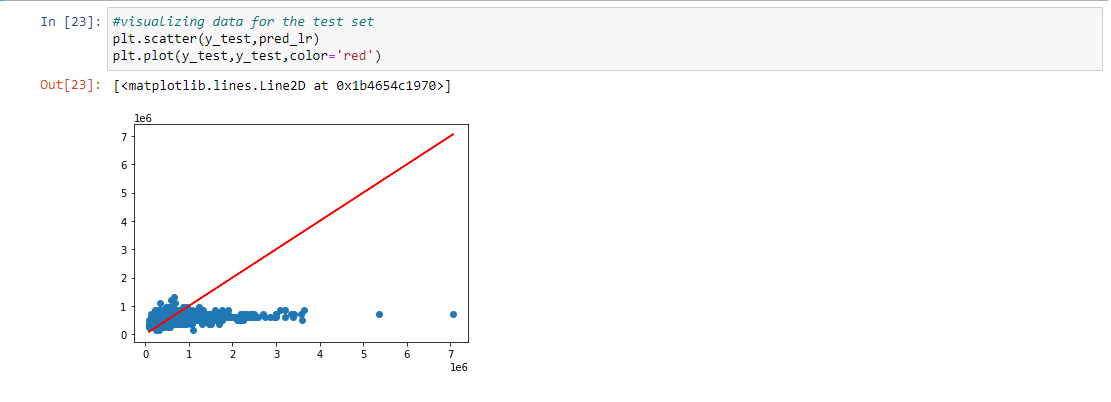
****

****

****

****



****

**Discussion:** We use simple linear regression for predicting house prices. At first, we have trained the model & test this. In the above scatter plot, we see data isn’t in line shape & also the R2 is 0.098(9.8%) only, which means our model hasn’t done good predictions.

**Experiment No:**3

**Experiment Name:** House price prediction using multiple linear regression

**Objectives:** To predicate house price from”kc\_house\_price” using multiple linear regression.

**Theory:**

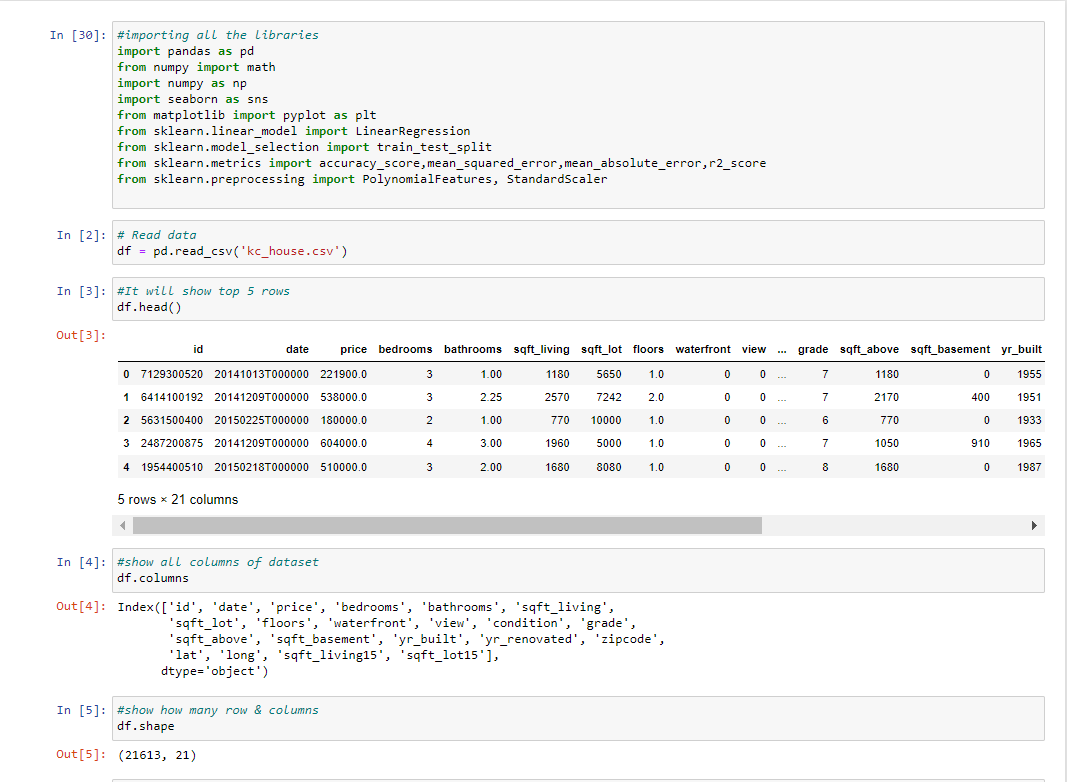
A multivariable regression is a bit more complex than other procedures. In the below equation, 𝒘 stands for the weights or coefficient which requires to be elaborated. All variables 𝑥1, 𝑥2, and 𝑥3 information attributes of the observations.

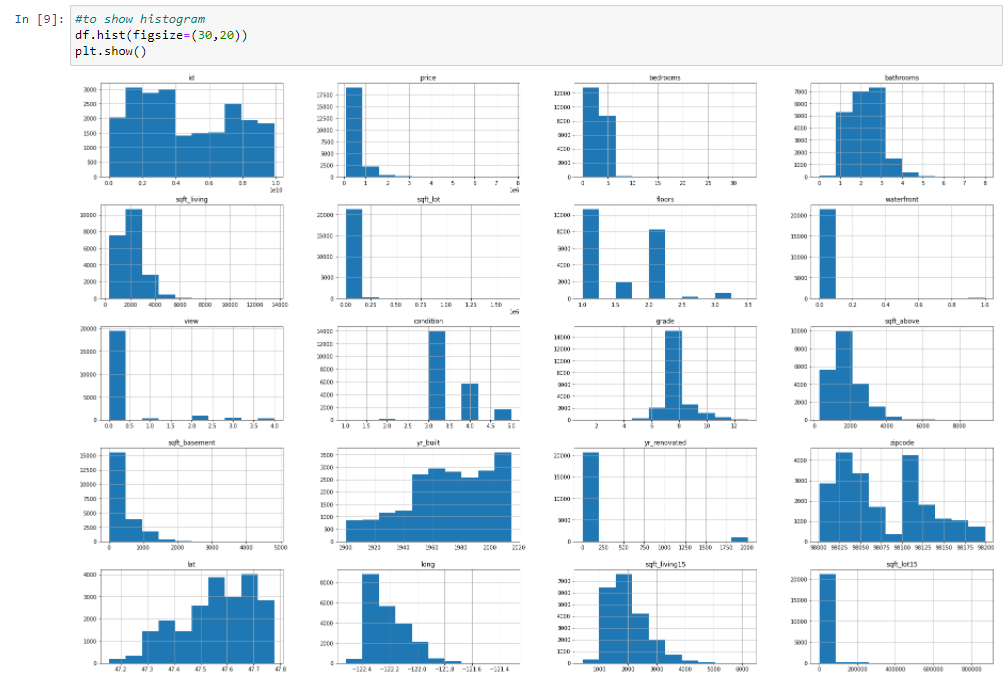
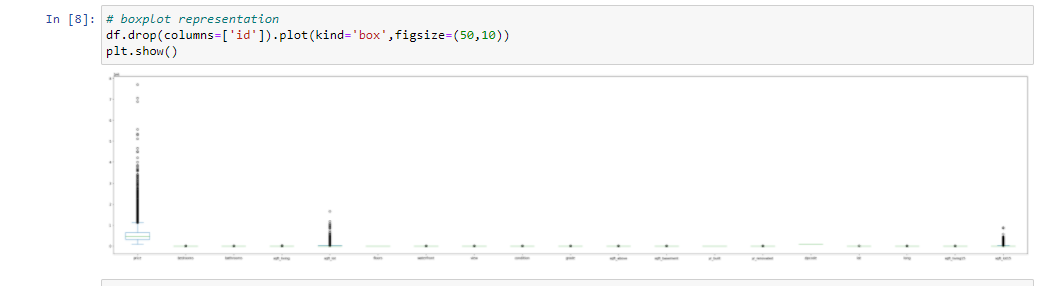
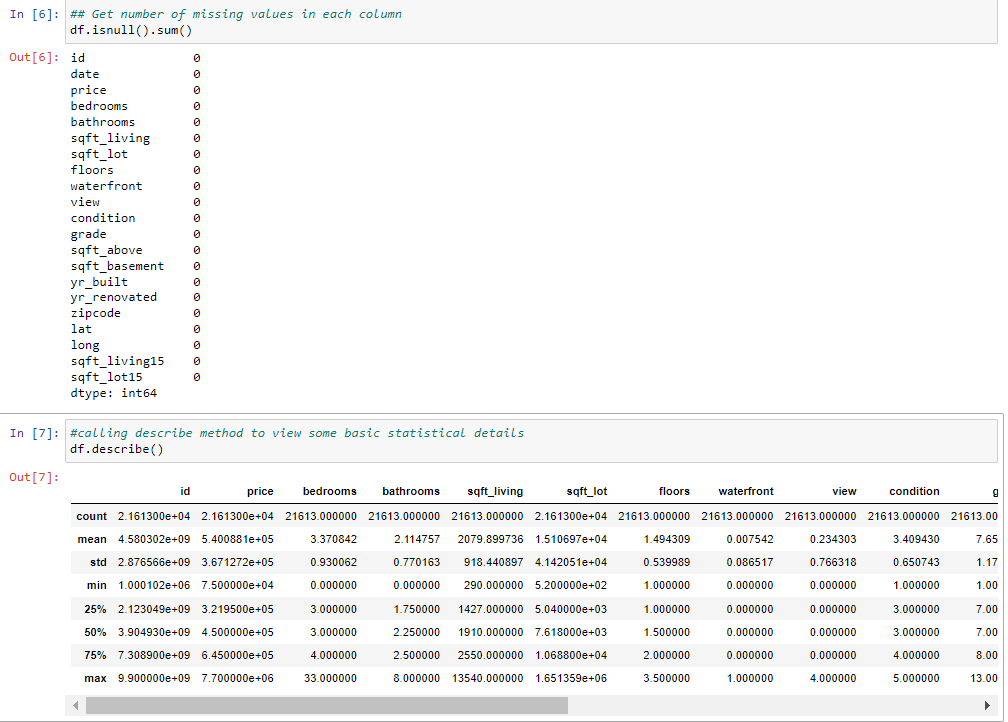
Y(X1,X2,X3) =W1X1+W2X2+W3X3+W0

Now let’s consider every step for the house price prediction using linear regression. Consider a company of real estate with datasets containing the property prices of a specific region. The price of a property is based on essential factors like bedrooms, areas, and parking.  Majorly, a real estate company requires:

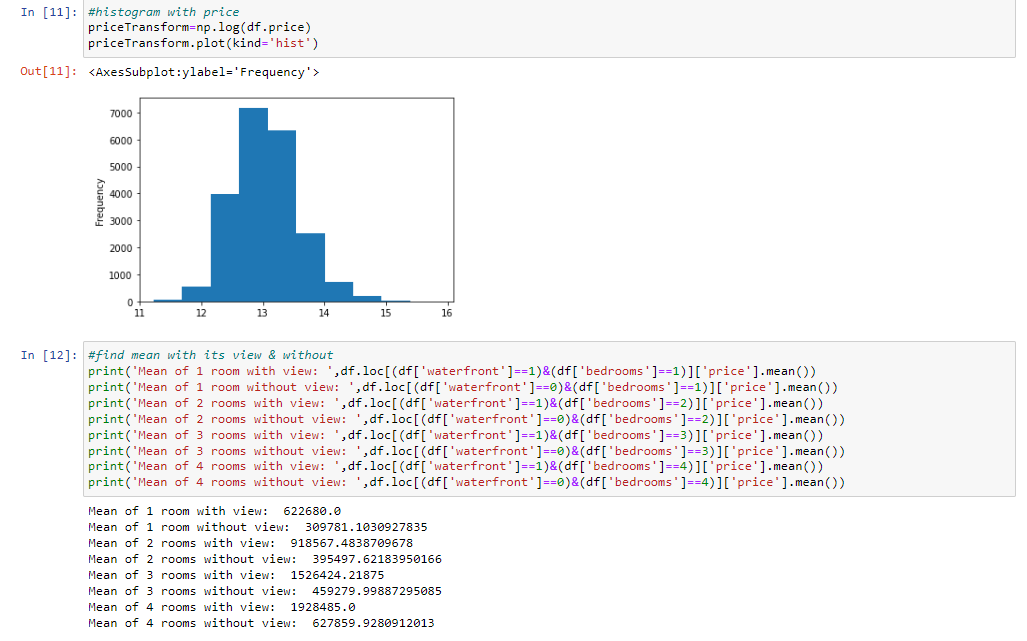
* Find the variable that affects the price of a house.
* Creating a linear model quantitatively related to the house price with variables like areas, number of rooms and bathroom, etc.
* For finding the accuracy of a model, that means how well the variables can predict the prices of a house.

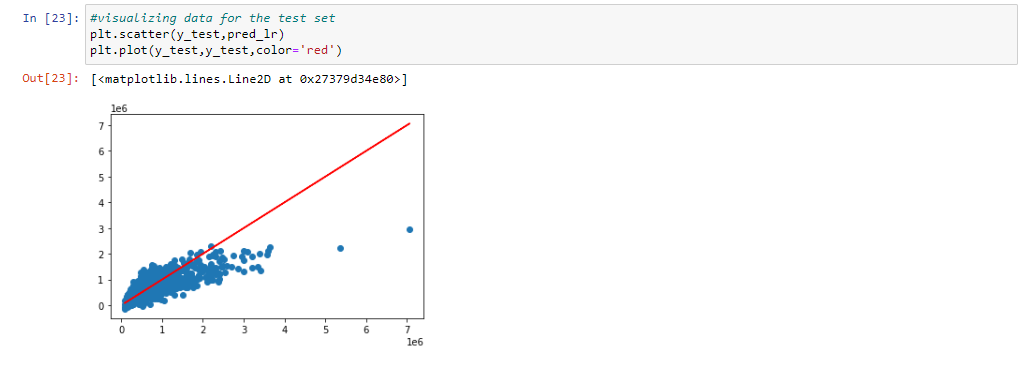
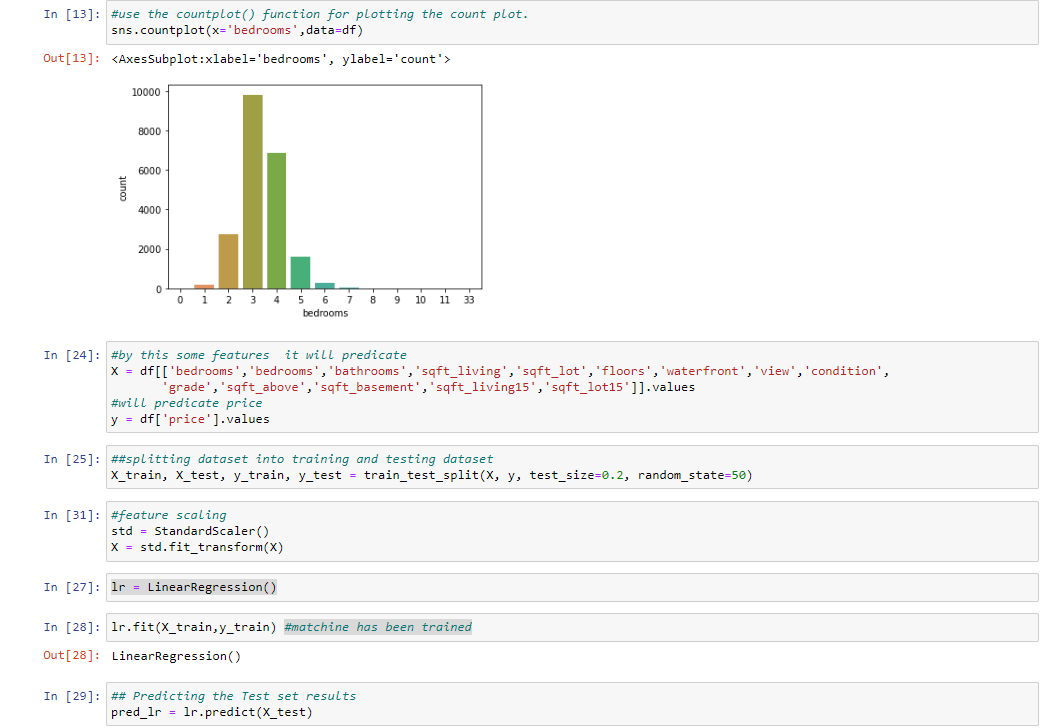
**Description:**











**Discussion:**

We have used multiple linear regression to predict the results more accurately using different variables. Apart from it, we have used a complete dataset that has accurate information regarding the houses. Majorly, all of the above codes and libraries we have used are not unique as there is a specific procedure to perform the house prediction procedure by linear regression.

At first, we have trained the model & test this. In the above scatter plot, we see data is in line shape & also the R2 score is 0.611 (61.1%), which means our model has done good predictions.

**Experiment No:**4

**Experiment Name:** Predict outcome of diabetes using multinomial naive bayes.

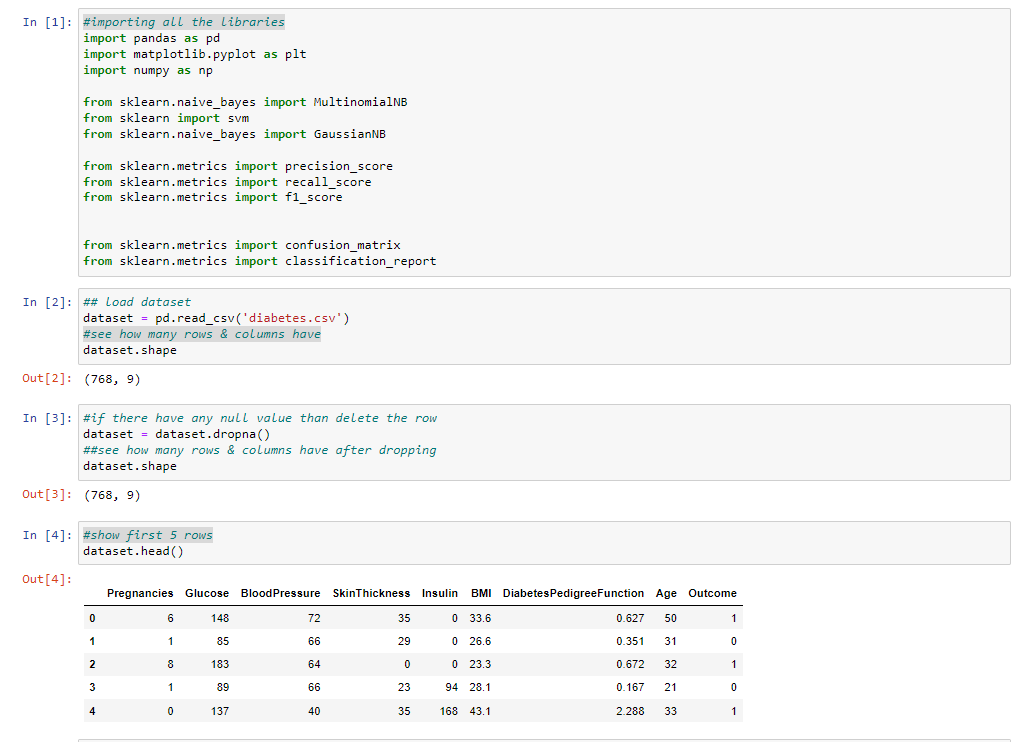
**Objectives:** To predict the outcome of diabetes, given some features of the patients. We will use Naive Bayes Model for this classification.

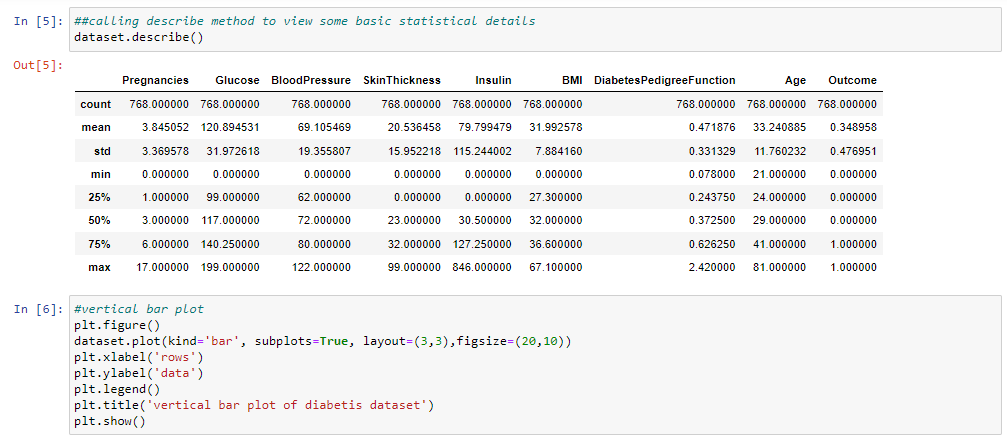
**Theory:**

Naive Bayes classifiers are a collection of classification algorithms based on **Bayes’ Theorem**. It is not a single algorithm but a family of algorithms where all of them share a common principle & every pair of features being classified is independent of each other.

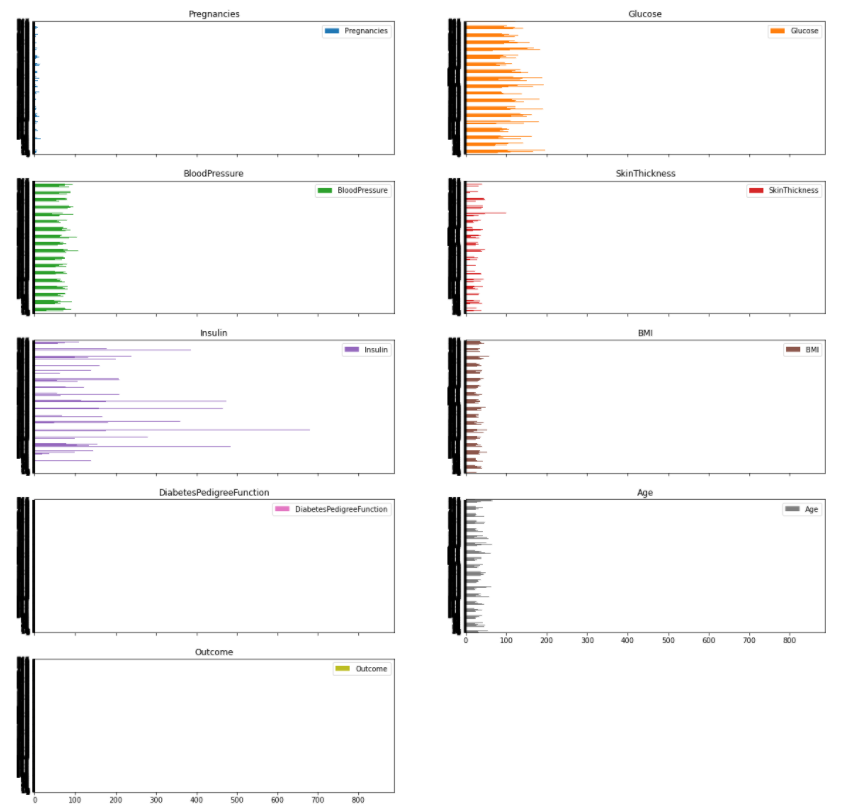
To start with, let us consider a dataset called ‘diabetes’. A prediction of diabetes at an early stage of the disease can lead to an improved treatment that doctors can use, this can be achieved with the help of machine learning and data analytics. The aim of this assessment is to build a machine learning model to predict whether or not a patient has diabetes based on their medical attributes. Since the diabetes predictor has only two outputs, it is considered as a binary classification, given that it involves deciding whether a sample is part of one class or another.

**Description:**

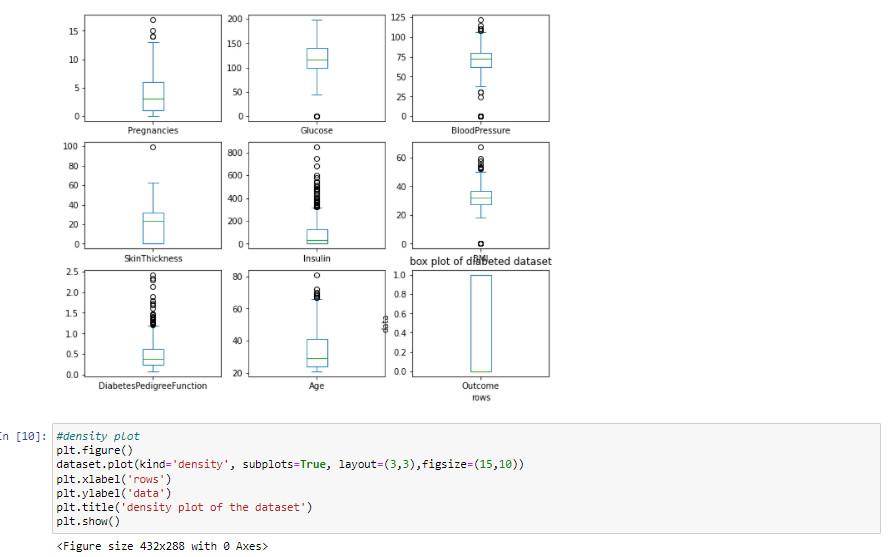
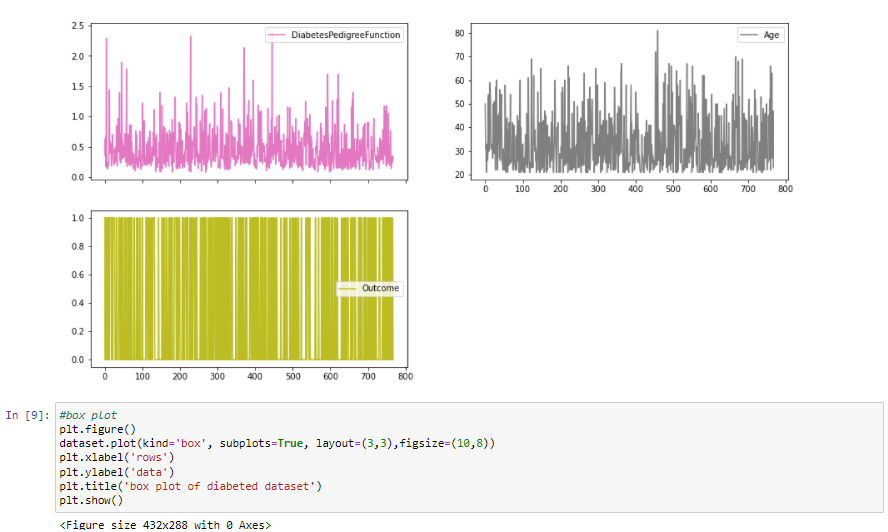
****

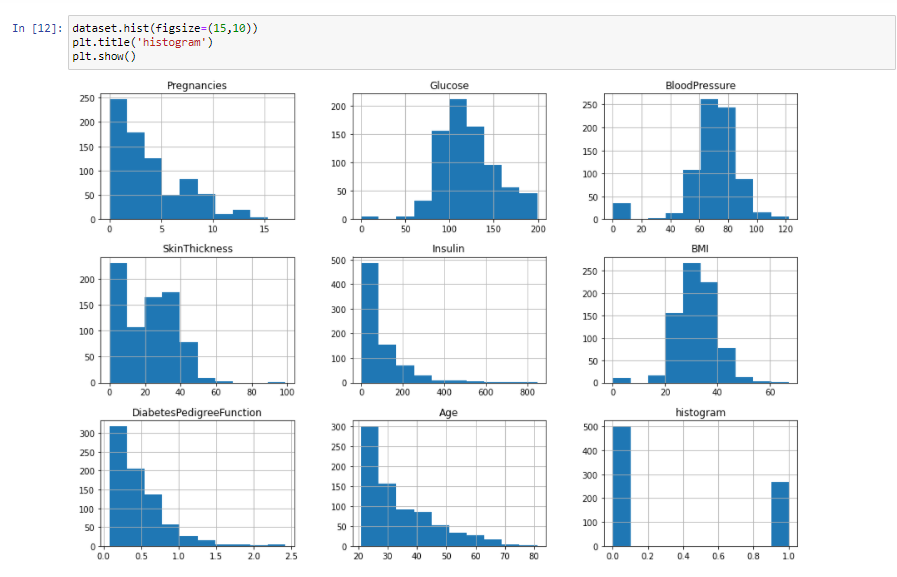
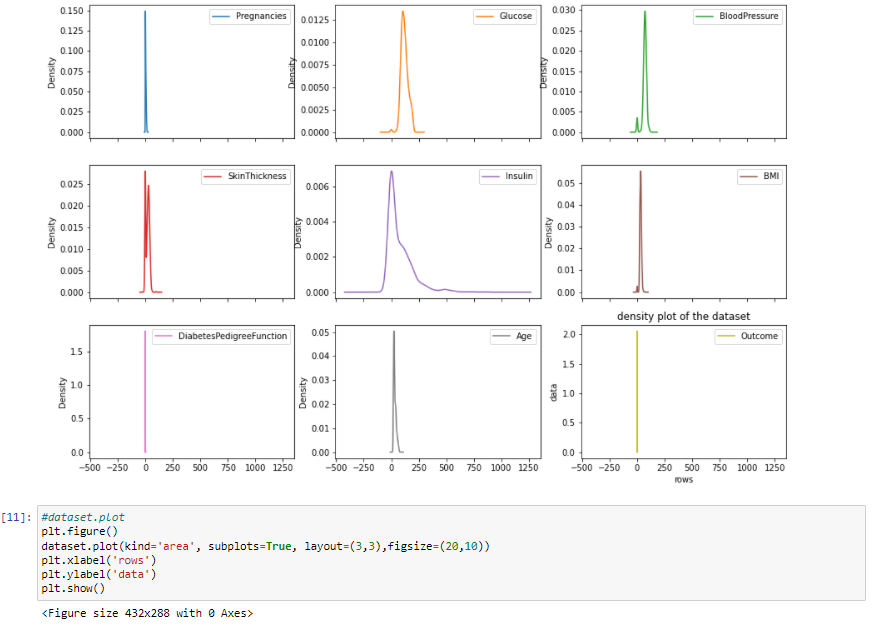
****

****

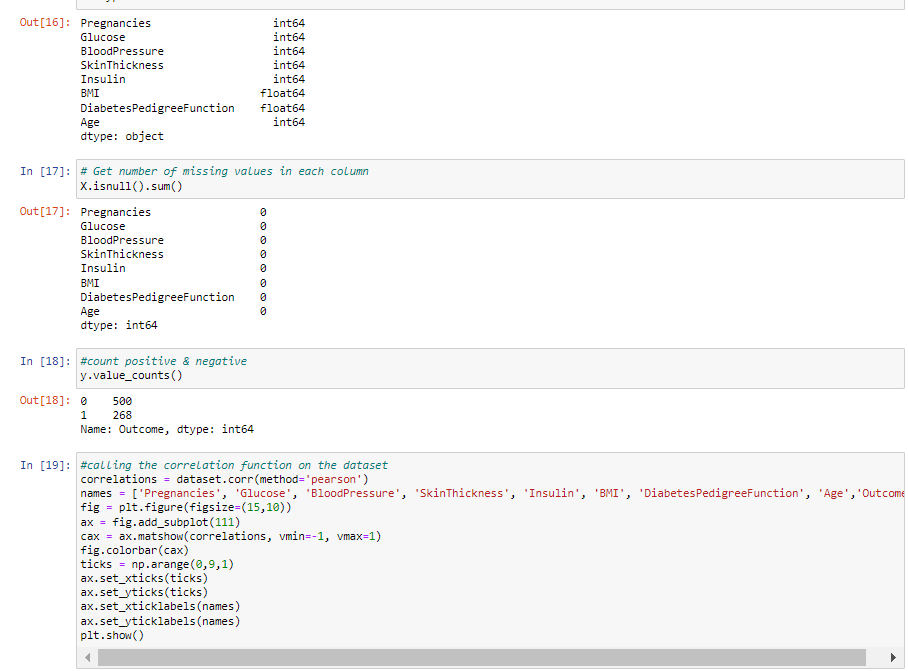
****

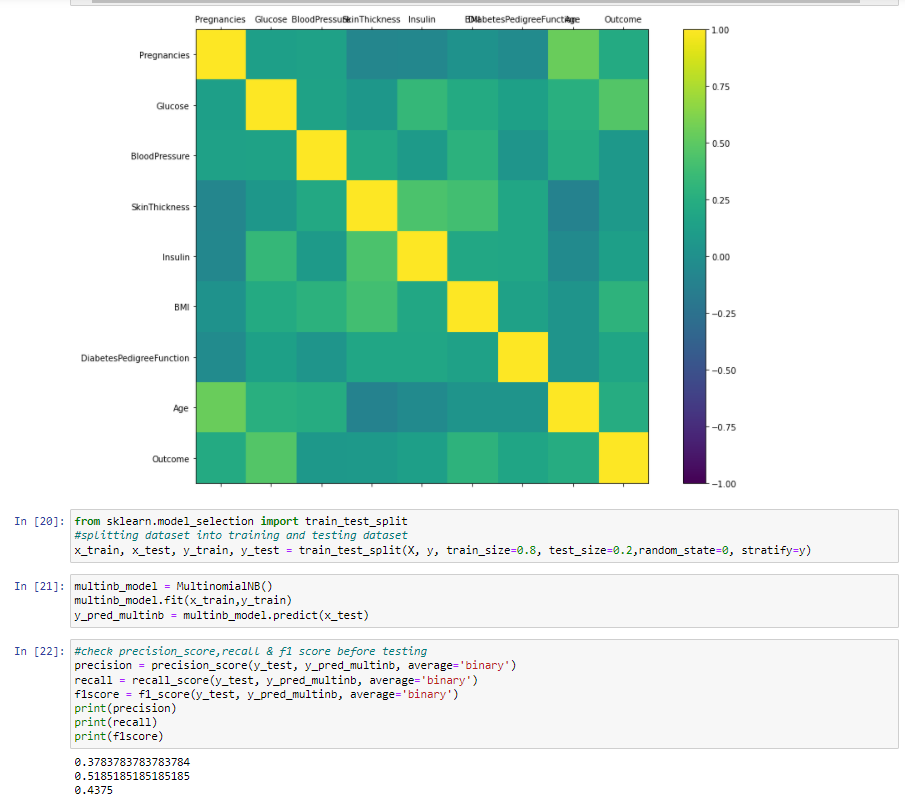
****

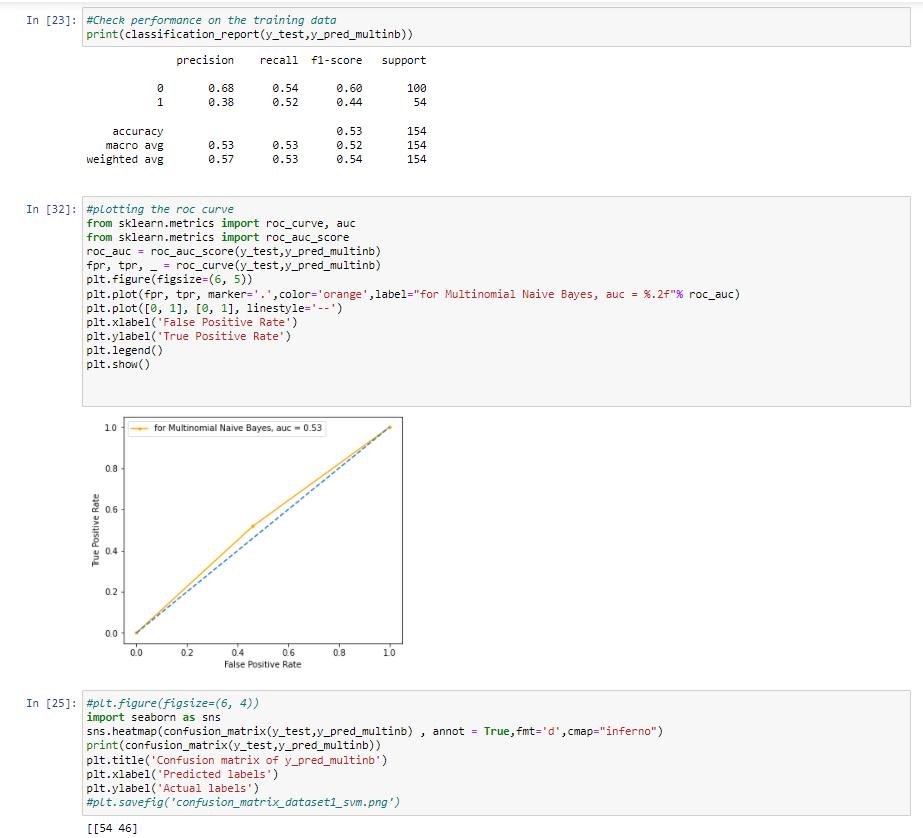
****

****

****

****

****

****

**Discussion:**

When we trained the machine, it had given only 53% accuracy. After training, I think that the best accuracy I’ve seen on this dataset was an optimized gradient boosting classifier with approximately 77%. On the other hand, this classifier could handle missing values and used 20% of the data as a testing set. So, we can say the machine has done a good prediction.

**Experiment No:**5

**Experiment Name:** Predicate class of mushrooms from “mushrooms” dataset.

**Objectives:**

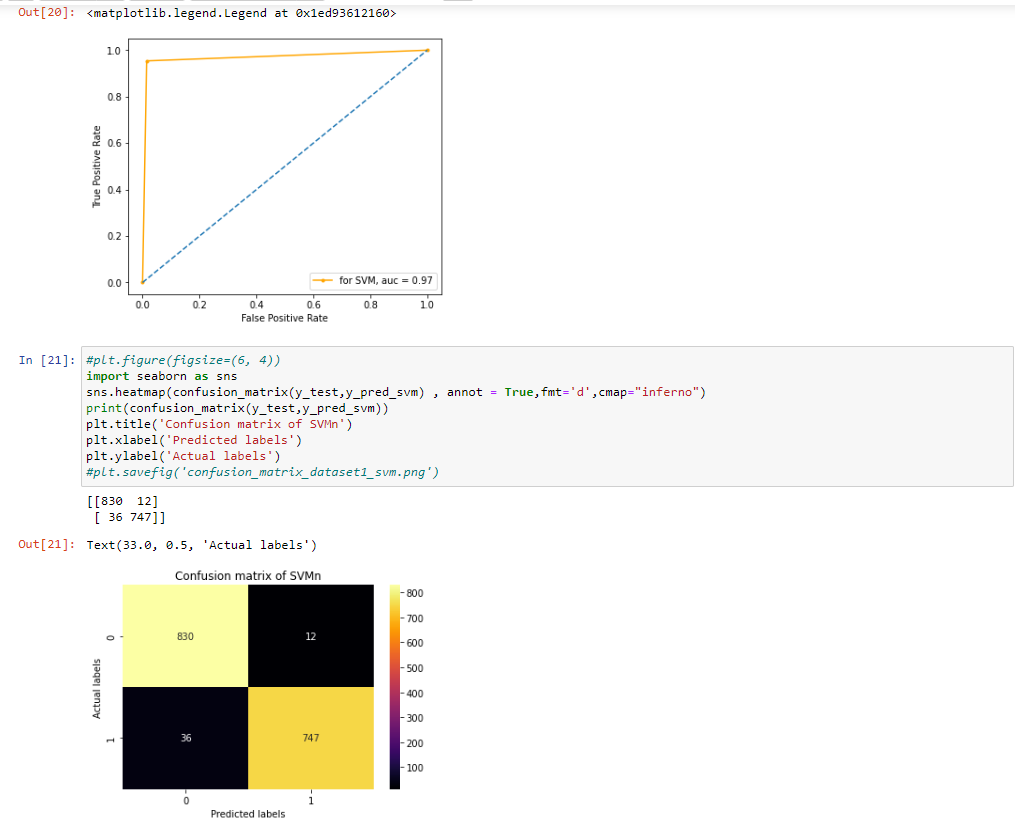
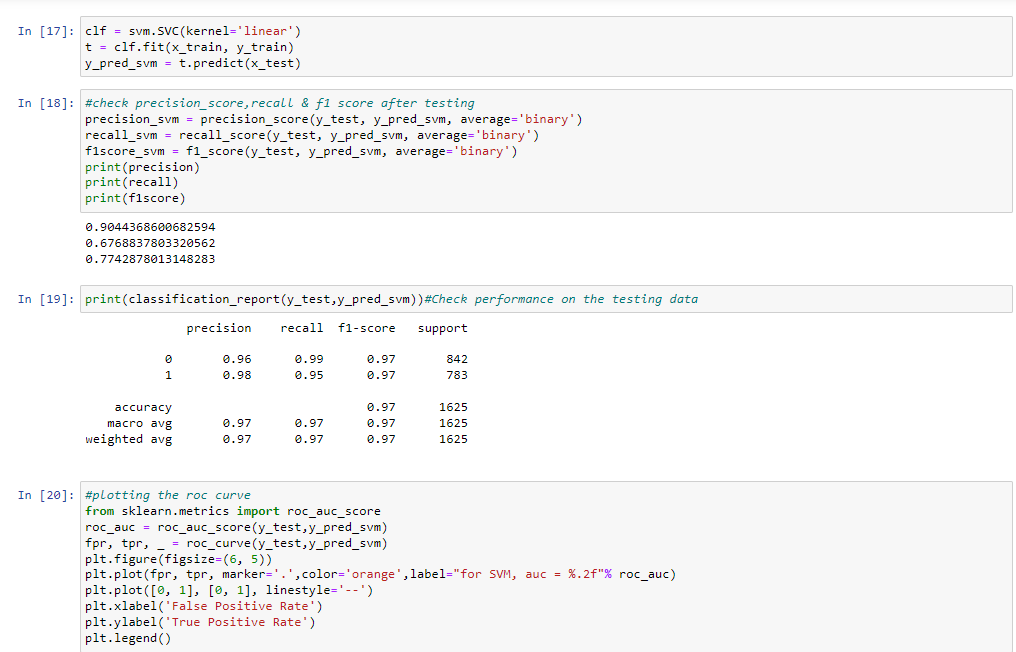
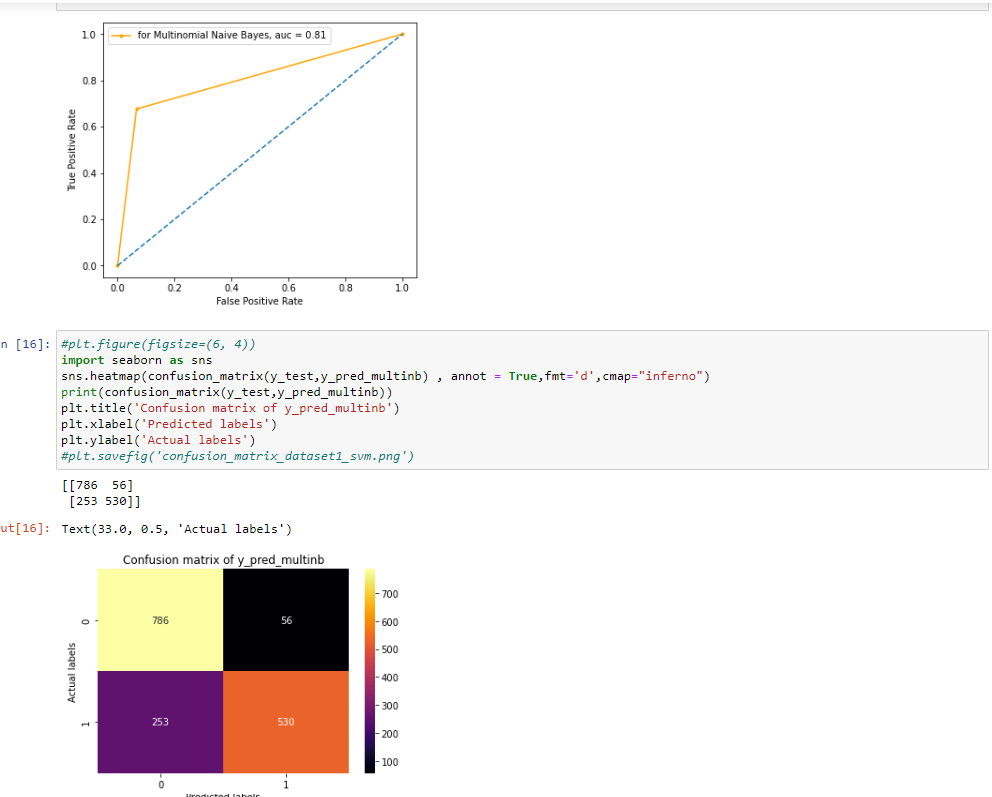
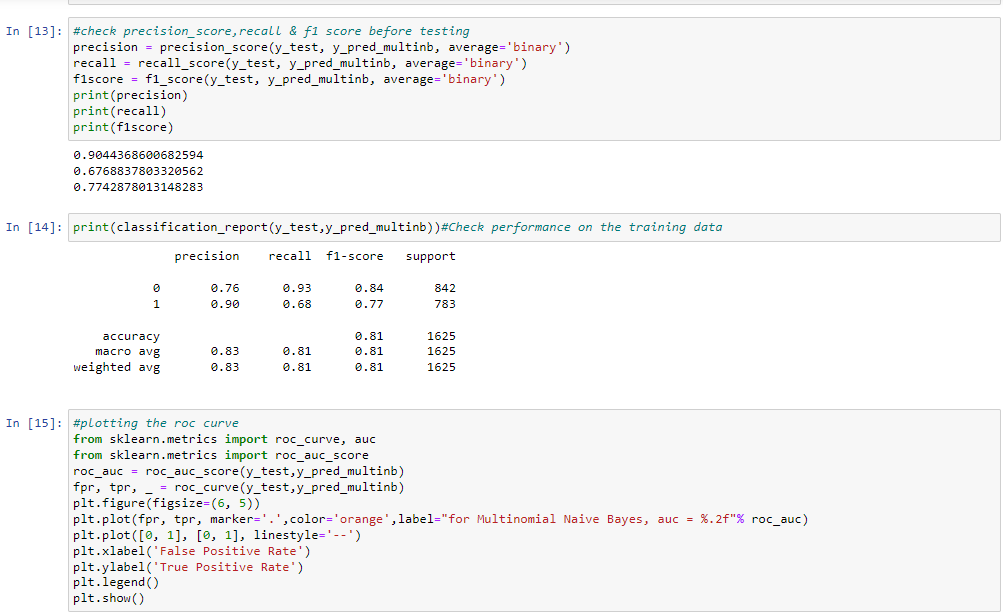
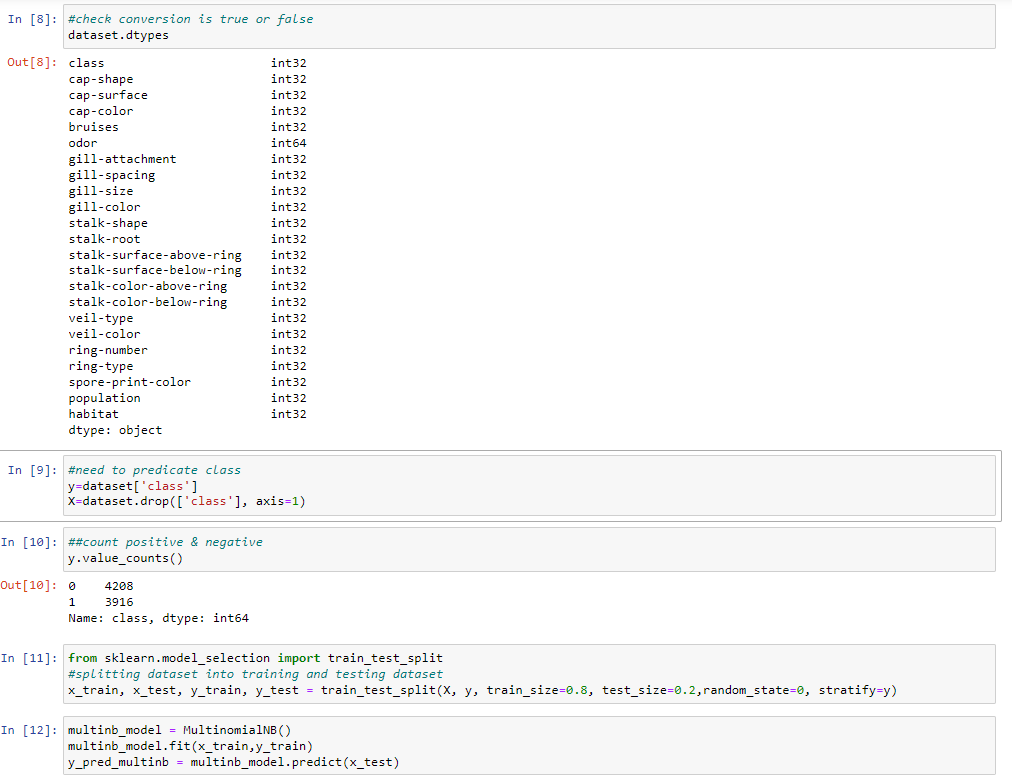
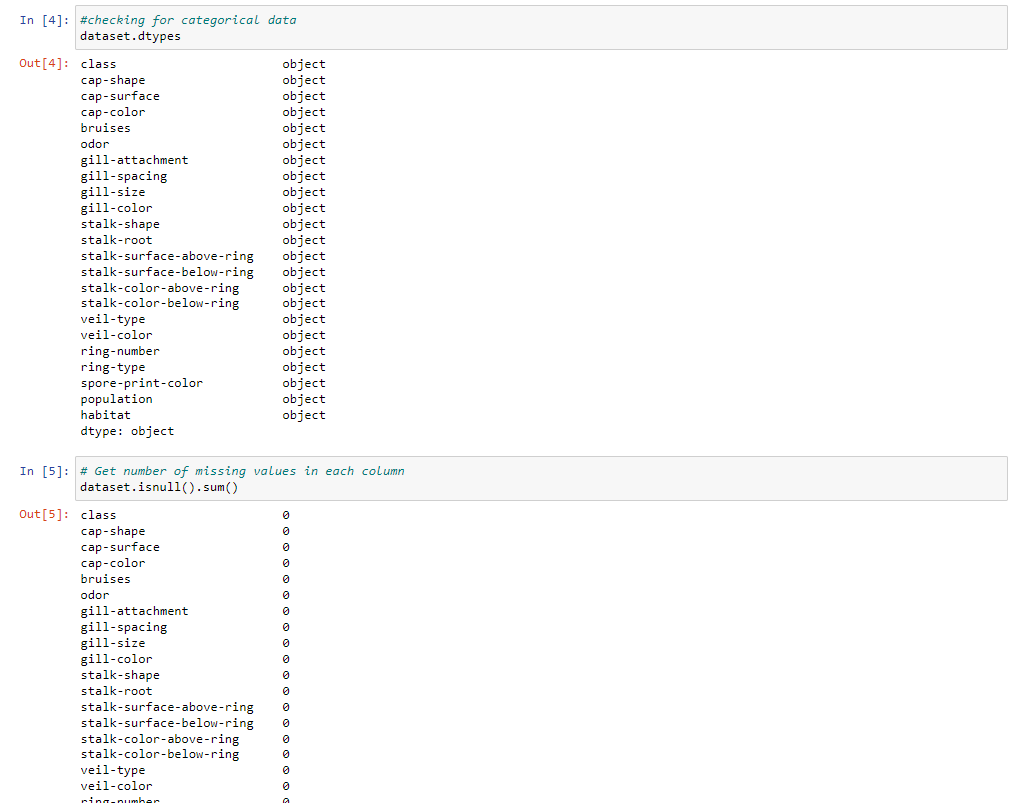
To predict the class of mushrooms, given some features of the mushrooms. We will use Naive Bayes Model for this classification.

**Theory:**

Consider a dataset called “mushrooms” with Mushroom Classification. This dataset includes descriptions of samples corresponding to 23 species of gilled mushrooms. Each species can be classified as definitely edible, definitely poisonous, or of unknown edibility and not recommended. Learning which features could spell certain death and which are most palatable is the goal for this dataset of mushroom characteristics.

**Description:**

****

****

**Discussion:**

We will use the accuracy score to quantify the performance of our model. The accuracy will tell us what percentage of our test data was classified correctly. The accuracy is a good metric choice because it will be easy to compare our model’s performance to that of the benchmark as it uses the same metric. Using Naive Bayes classifier, we get an accuracy of 97%, which is a very good prediction.