1. **Objective**

The growth of supermarkets in most populated cities is increasing and market competitions are also high. The Super Market Sales Analytics is used to analyse the sales of different products in a supermarket located in different cities. Based on the proposal and business requirement we need to build a model that will ensure the rating given by the customers of various products. Along with we need to answer the following questions:

● Which city has a better sale for products in the Electronic Accessories product line.

● Which payment method is used more often at a particular city, branch and for which product type

● Which Product type has been more purchased by female customers

● In which month does the highest number of home and lifestyle products have been sold

● At what time most of the female customers are purchasing products.

1. **Methodology**
   1. **Supervised method**: As the target variable (Rating) has been provided here, we can directly use some supervised learning methods like Logistic regression, Decision tree, Random Forest, Extended

gradient boosting algorithm etc. to learn the pattern and identify those products which are having great customer feedback.

* 1. **Unsupervised learning**: Again, we can use some clustering technique like K-means or K-NN or DBScan to find out the pattern which are most similar and create cluster based on them. Now we can identify the cluster where the most rated products are belonging and the cluster will be our targeted cluster because the basic definition of cluster is the same type of product having similar kind of rating should be grouped together in a cluster and their behavior will be similar. So, if a new product belongs to that cluster, then behavior of the product is highly similar to those who are belonging to that cluster.

• **The four machine learning techniques used**

1. K – Means Algorithm
2. Decision Tree
3. PCA
4. Logistic Regression

* Read the data: Read the data into Panda’s data frame
  + Size of the dataset - **1000**
  + Variable type– Categorical / Continuous

**c. Familiarize with the dataset and variables:**

The dataset consists of historical sales data of a supermarket company from 3 different branches over 3 months from Jan-March 2019

The dataset consists of data from 3 cities or 3 branches in India as given below-

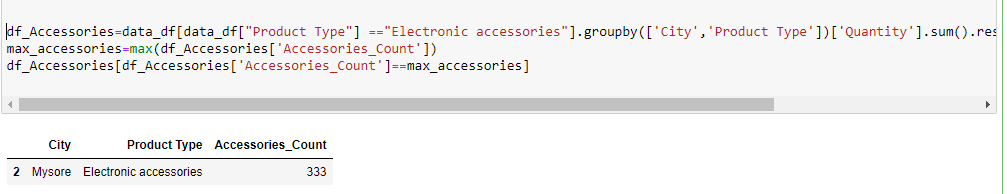
a) Branch A (Bangalore)

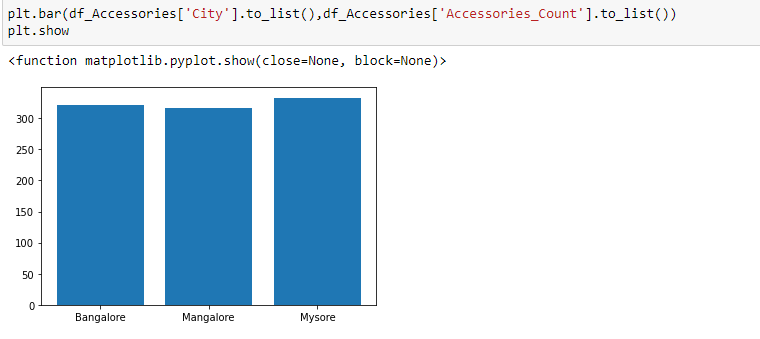
b) Branch B (Mysore)

c) Branch C (Mangalore)

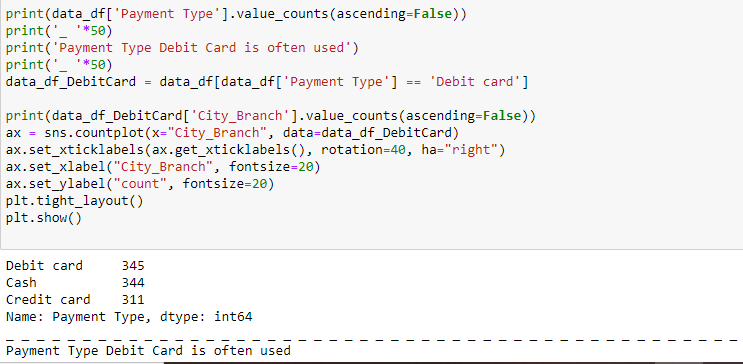
**Objective Questions and Answers Based on Data visualization:**

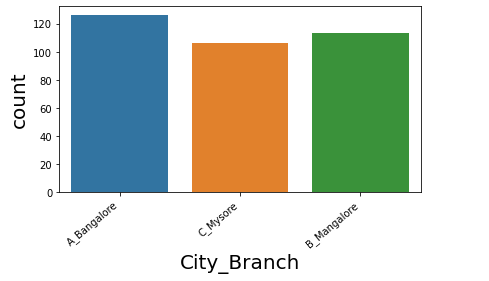
1. **Which city has a better sale for products in the Electronic Accessories product line: Mysore**



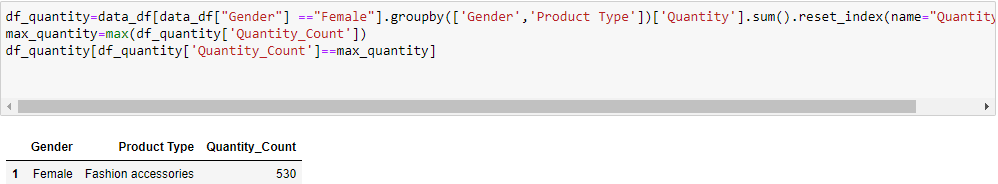


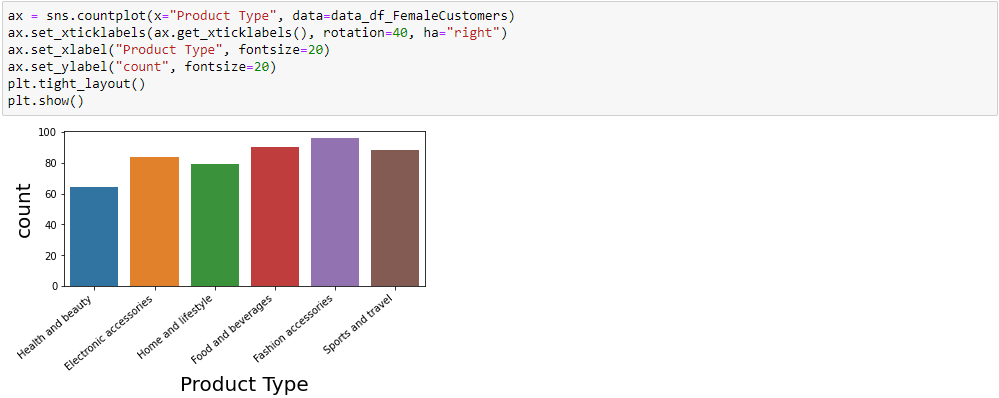
1. **Which payment method is used more often at a particular city, branch and for which product type: Debit Card**



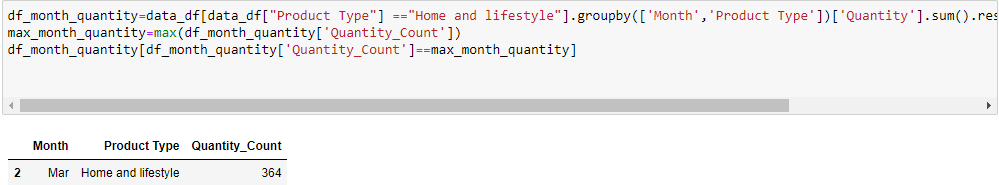


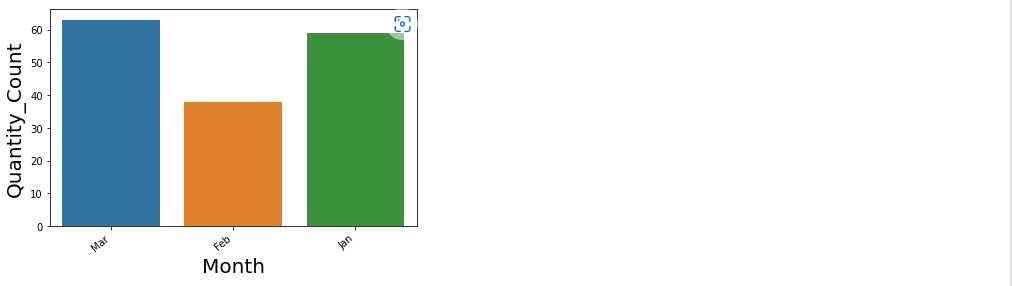
1. **Which Product type has been more purchased by female customers: Fashion Accessories**



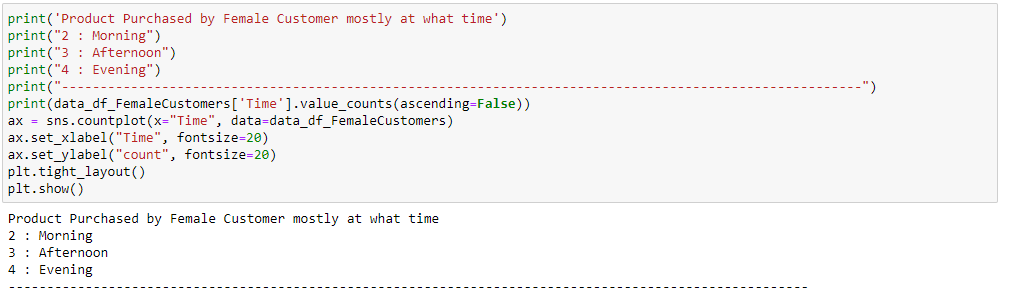


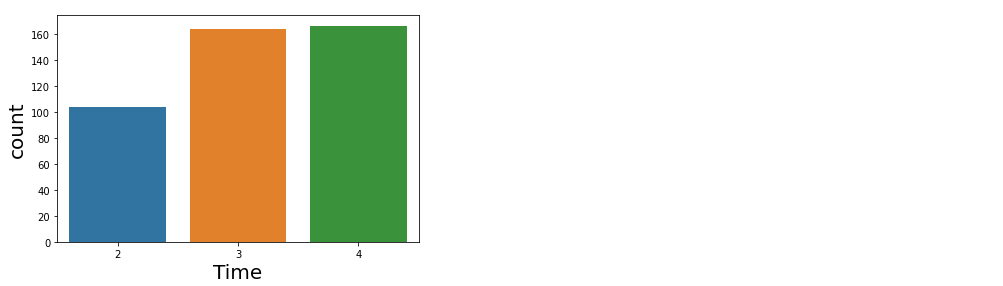
1. **In which month does the highest number of home and lifestyle products have been sold: March**





1. **At what time most of the female customers are purchasing products: Time wise purchased count of female customers and graphical representation**

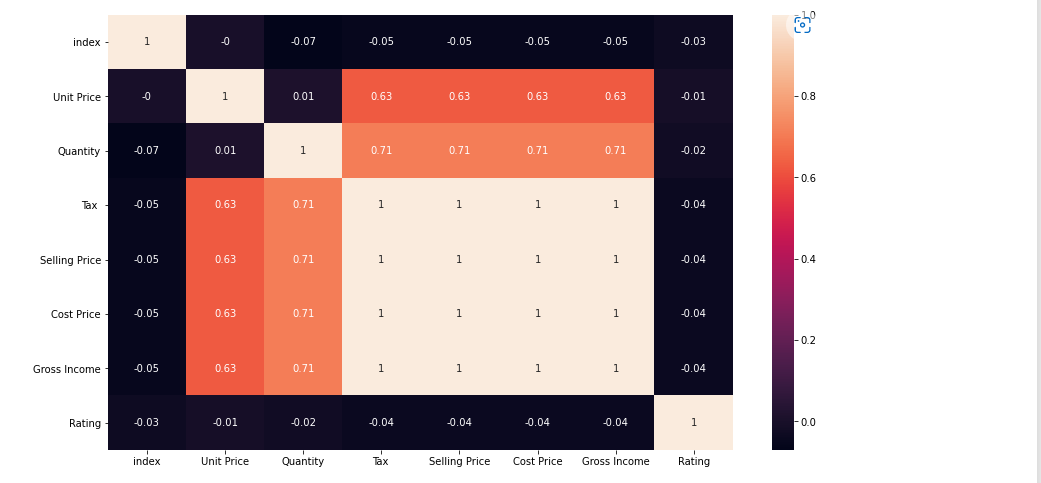




**Correlation analysis:**

Correlation analysis indicates the positive or negative or zero correlation between the variables. A positive correlation means that the values of the 2 variables increase together and negative correlation means the values of one variable decrease with the other. Zero correlation implies no correlation.

The seaborn heat-map can be used to visualize the correlations among variables.



As you can see, unit price is positively correlated to cogs with 63% correlation. Another interesting observation is that ‘Ratings’ hardly has any correlation with any other variables.

As is obvious, Quantity and gross income has very high correlation of 71%

1. **Business Understanding:**

A Supermarket is a self-service shop offering a wide variety of food, beverages and household products, organized into sections. The growth of supermarkets in most populated cities is increasing and market competitions are also high. The Super Market Sales Analytics is used to analyse the sales of different products in a supermarket located in different cities.

Here we are recommending the business model based on machine learning technique about the rating of each product sold in different cities based on their product type.

Our system will predict the rating for new instances and intimate business about the future goal of the product and rating based on past data.

1. **Ideas**

We have used below ideas to implement the business objectives and model:

1.Exploratory data analysis in pre-processing phase to understand the relationships among the variables and also provide the solid graphical representation to show the mostly dependent features to predict our target variables.

2.We have done heat map to get to know the correlation among the variables.

3.Feature selection techniques used here to determine the top 10 feature.

4.Done Principal component analysis to make the dimensionality reduction along with the graphical view.

5.K-means algorithm to check the cluster and also showed through elbow method.

6.Model build: Decision tree, random forest and logistic regression to predict rating.

1. **Selected Idea**
2. **Data Acquisition**

How many files:1

Read the data: Read the data into Pandas data frame Dataset

Size of the dataset - 1000

Variable type– Categorical / Continuous

1. **Understanding**

<How many features?> There are around 16 features.

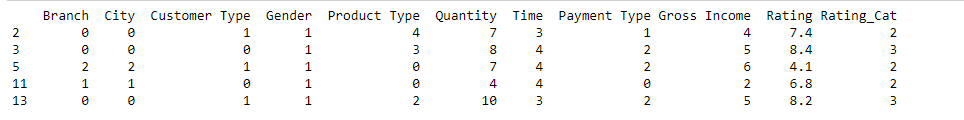
< In what way can the data be visualized to get to the answer that is required?>Data exploration and different plotting like Bar plot.

1. **Wrangling 1**

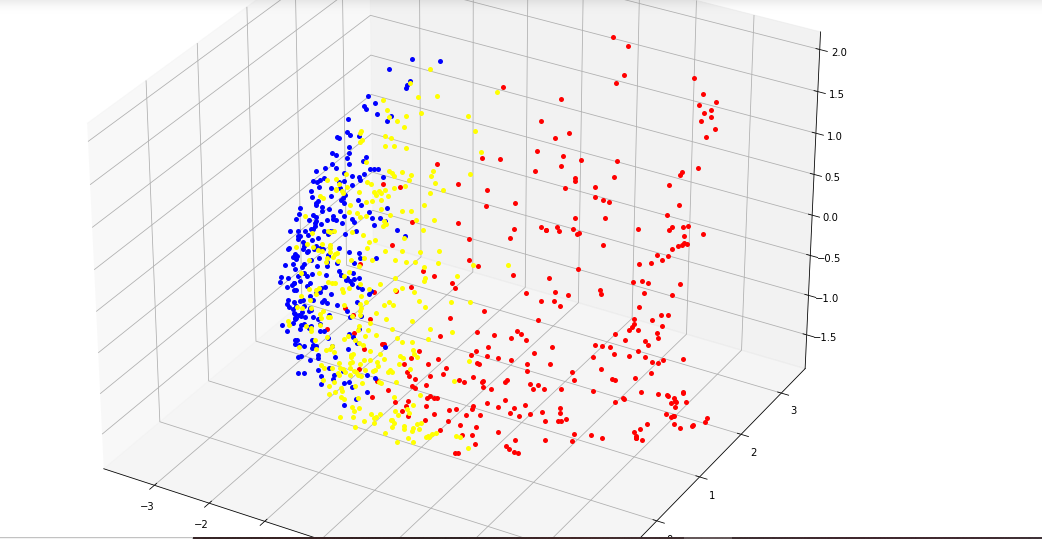
Missing data, which imputation technique> No missing data is there in the dataset.



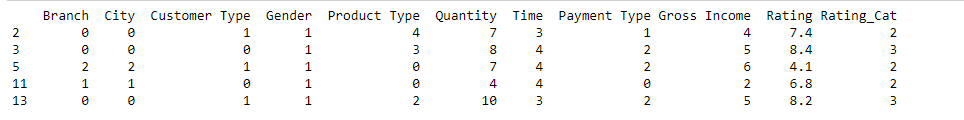
<discretization, which features and method> We used level encoder to discretize the data.



<reduction, sampling/PCA? >We have done principal component analysis and also showed the graphical view.

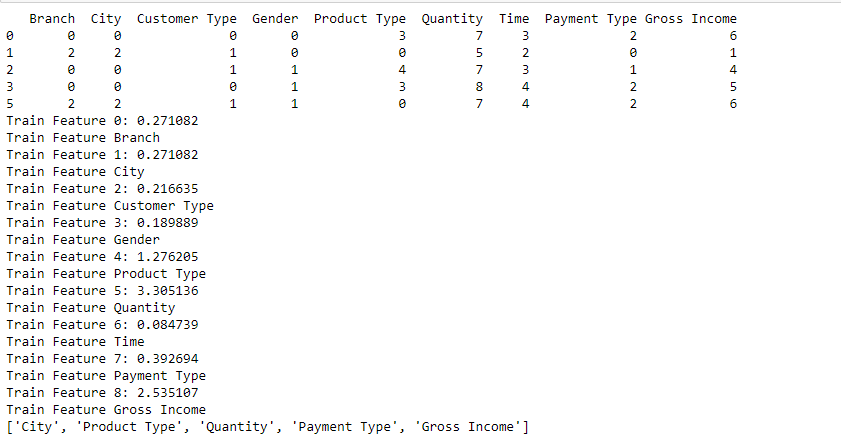


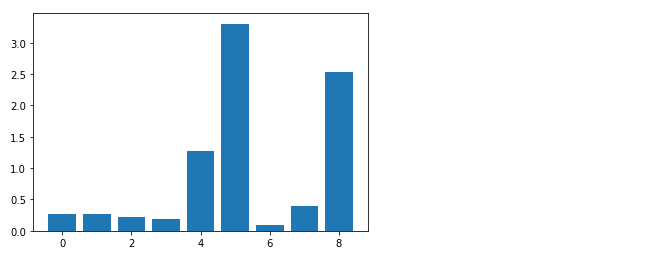
<Normalization?>



1. **Feature Selection:** CHI2 and Mutual information.
2. **Results of FE technique 1: CHI2**

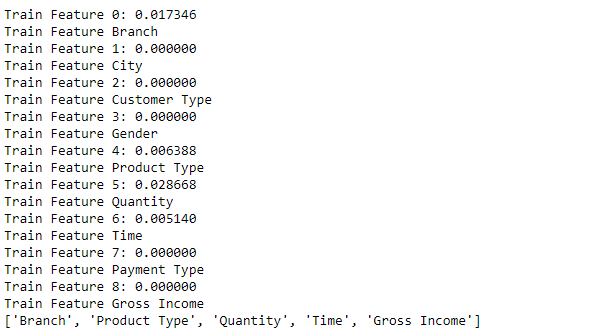
**CHI2 Method**: Chi-square is a statistical test used to examine the differences between categorical variables from a random sample in order to judge goodness of fit between expected and observed results. Using this test we picked top 5 features of the given dataset.

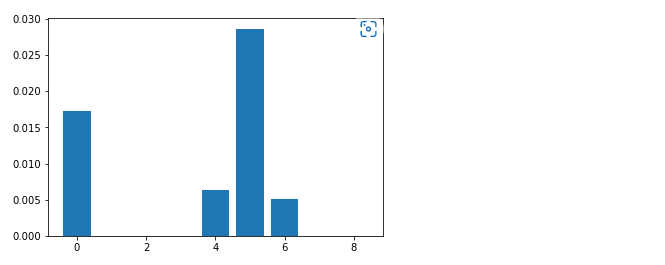




1. **Results of FE technique 1:** Mutual Information

**Mutual Info Method:** Mutual information is calculated between two variables and measures the reduction in uncertainty for one variable given a known value of the other variable. A quantity called mutual information measures the amount of information one can obtain from one random variable given another. Using this test, we picked top 5 features of the given dataset.





1. **Modelling:**

Sales forecasting is the prerequisite for a lot of managerial decisions such as production planning, material resource planning and budgeting in the supply chain. Promotions are one of the most important business strategies that are often used to boost sales. While promotions are attractive for generating demand, it is often difficult to forecast demand in their presence. In the past few decades, several quantitative models have been developed to forecast sales including statistical and machine learning models. However, these methods may not be adequate to account for all the internal and external factors that may impact sales. As a result, qualitative models have been adopted along with quantitative methods as consulting experts has been proven to improve forecast accuracy by providing contextual information. Such models are being used extensively to account for factors that can lead to a rapid change in sales.

Here we are using different machine learning algorithm to build our model and predict the rating which is our target variables.

1.Decision tree.

2.Logistic regression

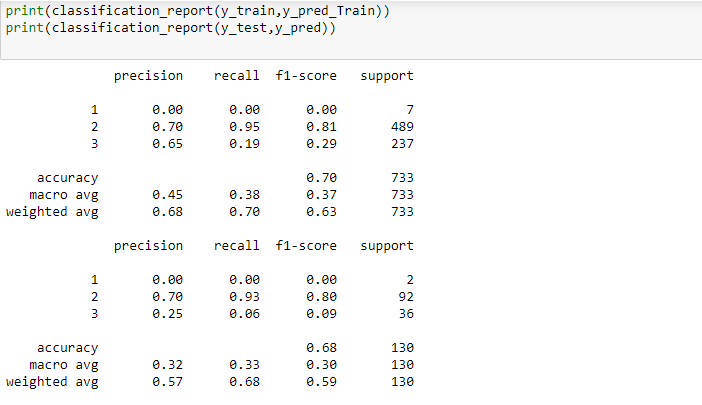
3.Random forecast

1. **Result of ML Technique 1:** Decision tree

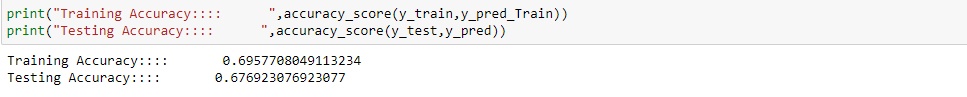
A decision tree is a tree whose internal nodes can be taken as tests (on input data patterns) and whose leaf nodes can be taken as categories (of these patterns). These tests are filtered down through the tree to get the right output to the input pattern. Decision Tree algorithms can be applied and used in various different fields. It can be used as a replacement for statistical procedures to find data, to extract text, to find missing data in a class, to improve search engines and it also finds various applications in medical fields

Here we are using decision tree to predict our target variable “Rating”.

Classification report:



Training and testing accuracy:

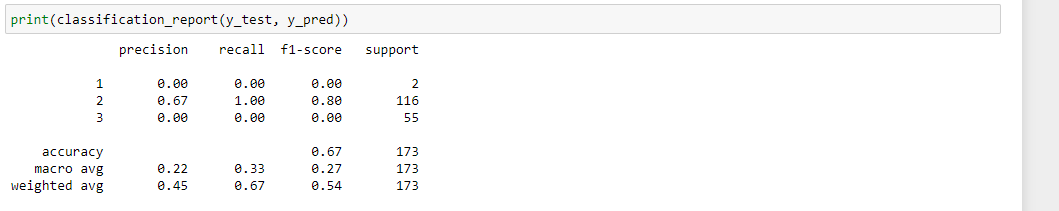


1. **Result of ML Technique 2 :** Logistic Regression

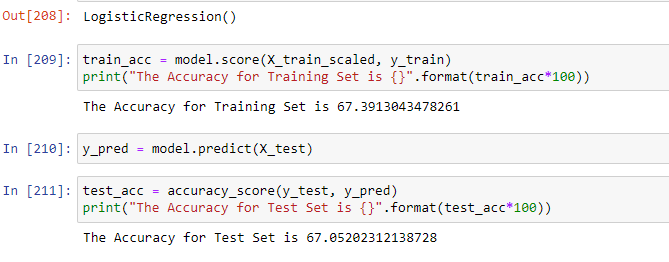
logistic regression permits the use of continuous or categorical predictors and provides the ability to adjust for multiple predictors. This makes logistic regression especially useful for analysis of observational data when adjustment is needed to reduce the potential bias resulting from differences in the groups being compared.

Here we are using decision tree to predict our continuous target variable “Rating”.

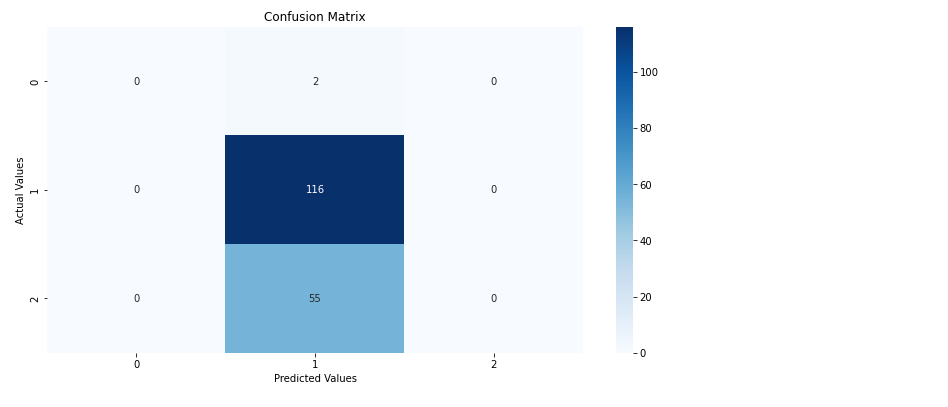
Classification report of logistic regression:



Training and testing accuracy:



Graphical view of confusion matrix:

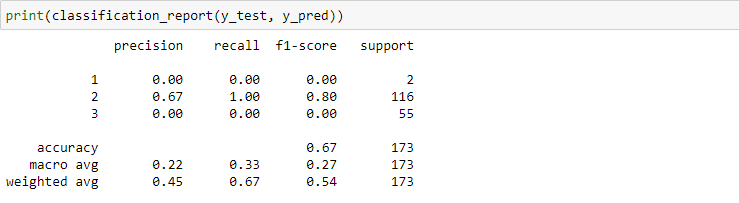


1. **Comparison**

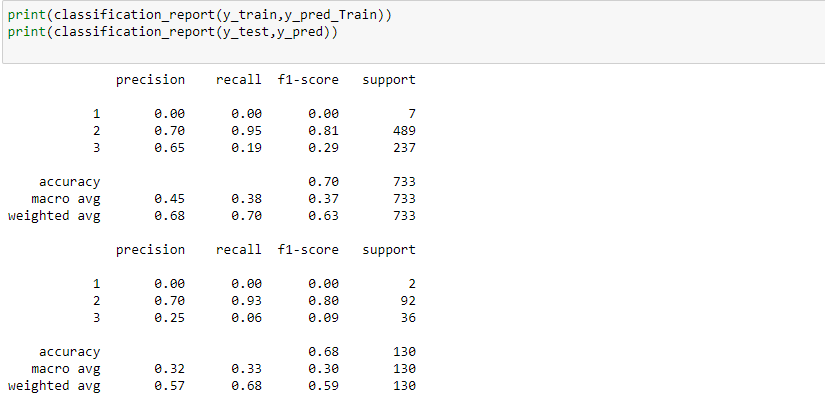
As per classification and accuracy report, our both the models are not overfitting and also accuracy is acceptable.

**Classification report comparison:**

**Logistic regression:**



**Decision tree:**



1. **Conclusion**
   1. We used uni-variate, bi-variate and correlation analysis to perform basic EDA on the supermarket sales data.
   2. To summarize below are some of the findings/observations from the data:
   3. The customer rating is more or less uniform with the mean rating being around 7 and there is no relationship between gross income and customer ratings.
   4. The data consists of 3 cities/branches.
   5. Fashion accessories and food and beverages are the most sold product and these products should be focused on along with electronic accessories.
   6. The most popular payment method is Debit card and cash payment is also on the higher side.
   7. There is no particular time trend that can be observed in gross income.
   8. At an overall level, ‘Sports and Travel’ generates highest gross income.
   9. Gross income is similar for both male and female, though female customers spend a bit higher at the 75th percentile. Females spend on ‘fashion accessories’ the most and for males surprisingly it is ‘Health and beauty’. Females also spend more on ‘Sports and travel’ which generates highest income overall.
   10. Using the correlation analysis, one interesting observation has emerged that customer ratings is not related to any variable.
   11. Most of the customers buy 10 quantities and busiest time of the day is afternoon i.e. around 2 pm which records highest sales. Sales is higher on Tuesdays and Saturdays compared to the rest of the week.
   12. Though the rating for ‘fashion accessories’ and ‘food and beverages’ is high but the quantity purchased is low. Hence, supply for these products need to be increased.

Conclusion on prediction:

* 1. Based on the available data and important features, we got training accuracy of 72% (80% of total data) and testing accuracy of 71.6% (20% of total data). So, we can say that model is not overfitting.
  2. Though here we are assuming training and testing distribution of independent variables are more or less similar. There are certain processes which can check the distribution similarity between
  3. straining and testing datasets. For an example, KL-Convergence, Violin plot etc. Due to limited amount of time, we were not able to implement those processes.

1. **Recommendation**