University of New Haven

RETAIL SALES DATA ANALYSIS

ARIGA TEJASIMHA
SRIMANIKANTA ARJUN KARIMALAMMANAVAR
RAHUL AKKINENI

STATEMENT OF OBJECTIVE

The goal of our project is to provide the supermarket to improve their business with insights like

- which products to promote,
- which products to place near other products,
- which products to stock
- which products to recommend to specific customers

We will recommend the company throughout the year by considering the dataset of 3 months of sales at 3 different locations.By considering on time series modelling of sales which has patterns of sales based on time and location analysis based on the store present which will give insight to company to knew it's target audience or customers.

DATA AND METHODLOGIES

We have considered the data set from kaggle which is crisp enough and there is no need of data cleaning technique. Here is the link of the data set

https://www.kaggle.com/aungpyaeap/supermarket-sales

The data set consist of 1000 entries which we have used 80% of data for training the model and 30 % for the testing.which is perfect any kind of model building.we have models like

- KNN MODEL
- Linear regression
- Logistic regression
- Decision Tree
- Correlation Matrix
- XG Boost.

We use all the methods and models and find the best fit the our objective. We use the 30 % of data for testing in all kind of models and consider a model which has least RMSE value(Root Mean Square Error) and highest Accuracy rate. We use the Decision Tree for the further use if any extra row is added then the decision tree automatically arranges and does the required computation needed.

APPROACH

As we knew about the dimensions of the data set. All the dimensions are necessary and inter related to each other to find the relation between them and product line.first and fore most must consider some of the base line conditions like weather is being same as three months whole year ,number of holidays. As it is clearly understood from the data that it has three location once we perform the analysis for the one location we can repeat the analysis for remaining other two locations. But for time series modelling we have consider the three location at a time.

When we consider the time series modelling we have observed a patterns in the flow of purchases. We found from the data exploration technique i.e,when we take graph between the hours of the day to quantity of purchases .Most of the goods are purchased at 14:00 . we can see that food and beverages sales usually high in all three branches at evening especially around 19:00. We have interesting patterns between male and females which is females tends to shop more on Tuedays. Where is there is no much trends in male shopping but considerable sales on Friday.

We have taken correlation matrices for the products and locations to find which product being sold in which location and find the dependancy or most related. We found that in location 1 (yangon)most related is home and life style and next highest related is sports and travel.location 2(Naypyitw) food and beverages and next most related fashion and accessories. From location 3 (Mandalay)fashion accessories and sports & travel. From this location 2 has lowest rating. We have also taken the bar plots on the basis locations and product lines. From which we can analysis that which product to be promoted much which has less sales rate and products to be embed along with the other products based on the category the products.

We have taken the bar plot of purchases done males and females. We have found a very interesting analysis. Which is female seems giving most satisfying rating and interested in shopping when their payment method is cash. Where in cases of males they preferred E-wallets or credit cards.

VALIDATION AND MODEL BUILDING

As we have discussed in the methodologies we have used the all the models like KNN model, linear regression , logistic regression and XG boost. We have considered the RMSE(Root Means Square Error). We have used the model which has least RMSE. Which is the best fit for the prediction model.We have used the XG boost has as it has least RMSE value and it has enhanced the performance of the model. We have used the decision Tree for the further use of entries of data. We can validate the model by using the more of the testing data.

CONCLUSION

We have analysed the sales data and and developed the model to predict the sales report for whole year.We can say that based from analysis

- we can suggest the promotional recommendation on the product line with least sales report
- We can do the product embedding based on the sales which less than high sales and greater than lowest. We can also embed based on category
- We have done analysis completely on time series by considering all the scenarios like gender and payment methods, gender and shopping trends

REFERENCES

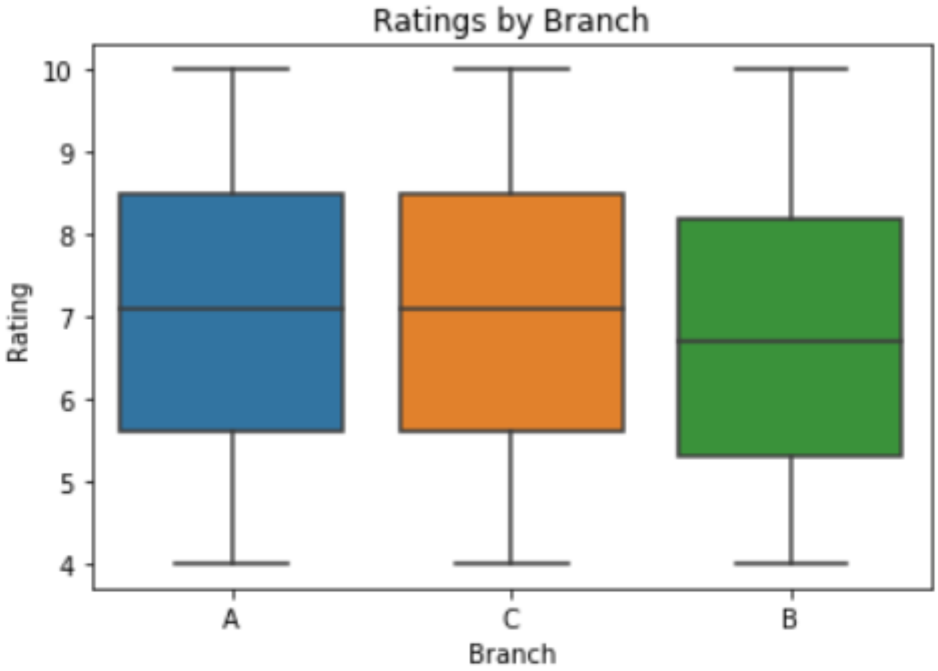
www.stackoverflow.com

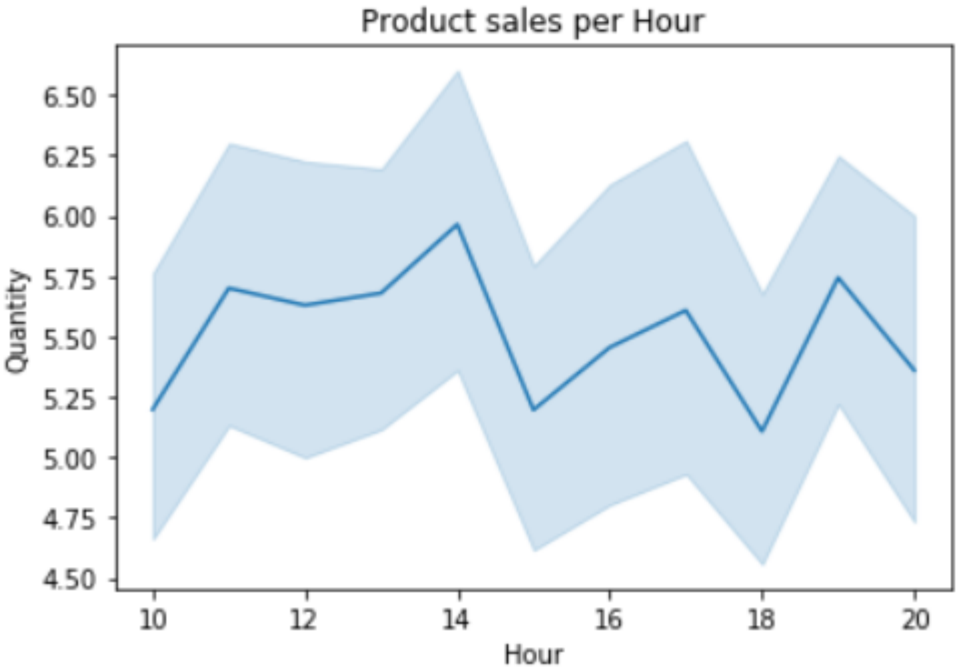
www.google.com

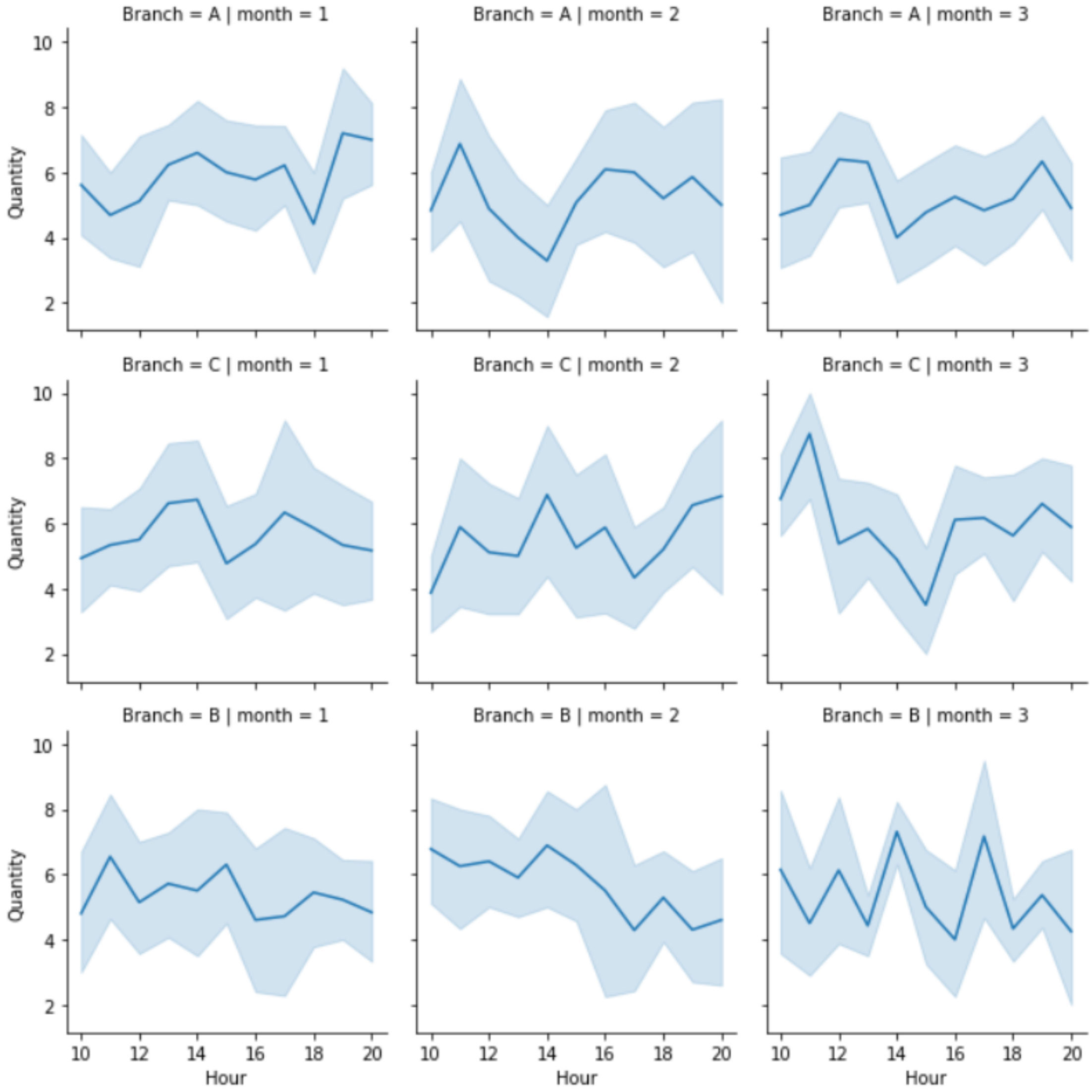
www.kaggle.com

www.towardsdatascience.com

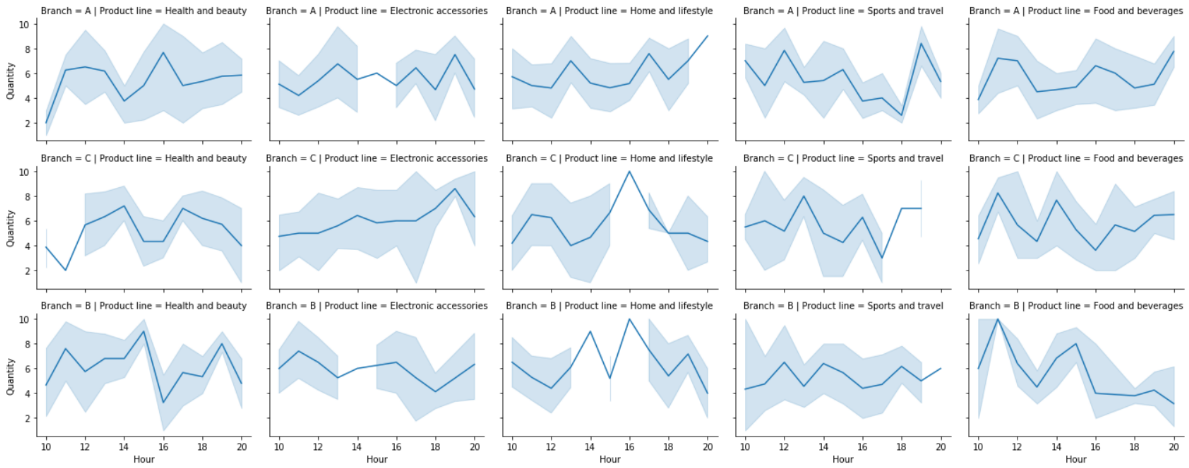
SCREENSHOOTS





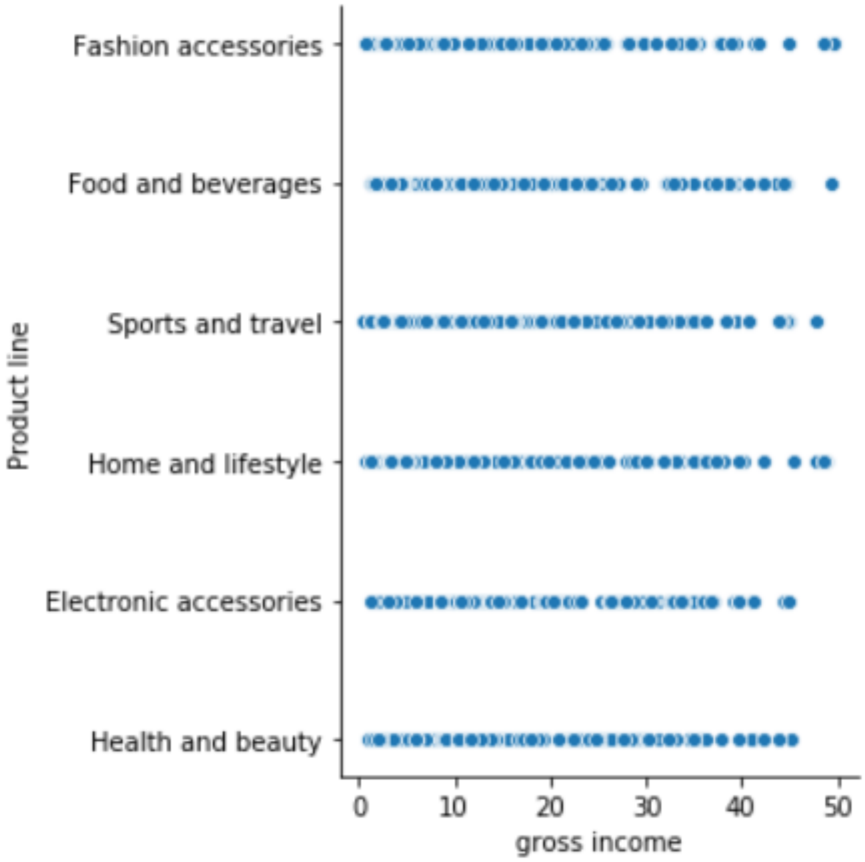


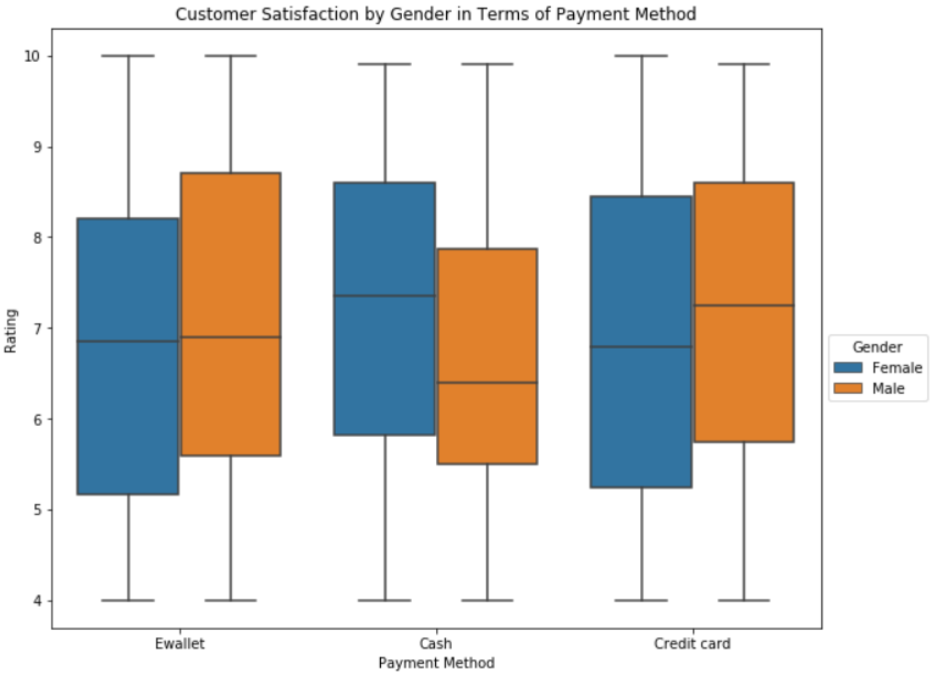
GRAPH BETWEEN THE QUANTITY AND BRANCHES AND MONTHS

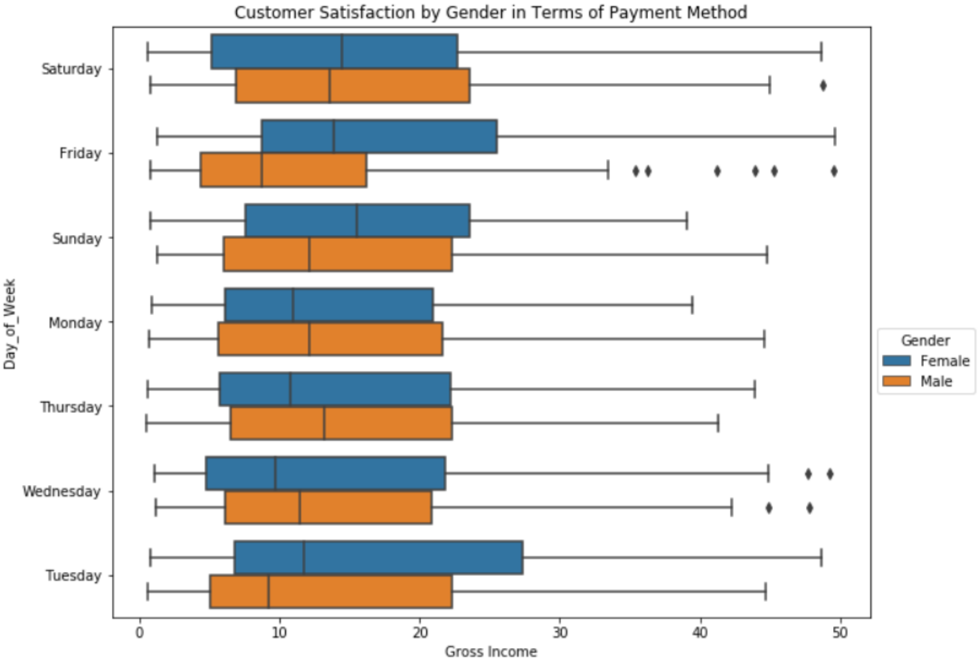


GRAPH BETWEEN THE QUANTITY AND BRANCHES AND MONTHS

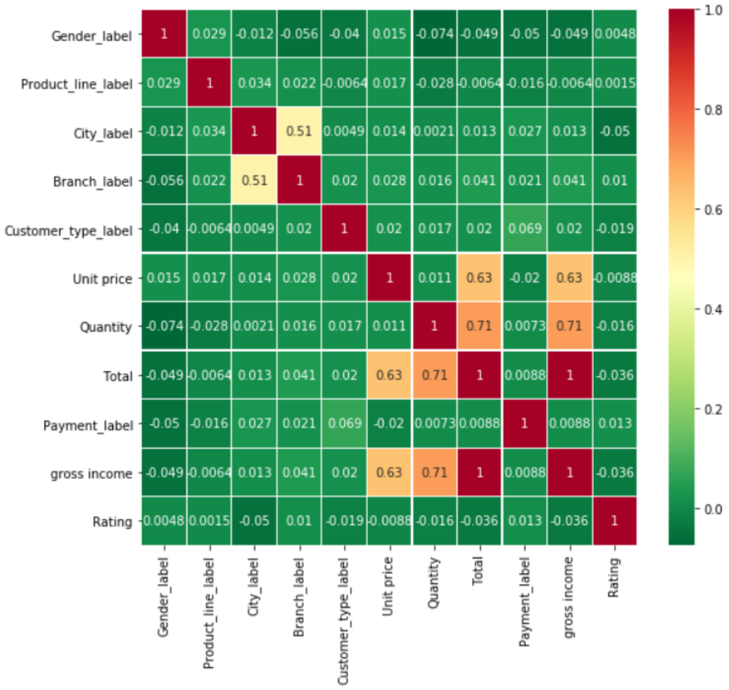
GRAPH BETWEEN THE PRODUCT LINE AND GROSS INCOME

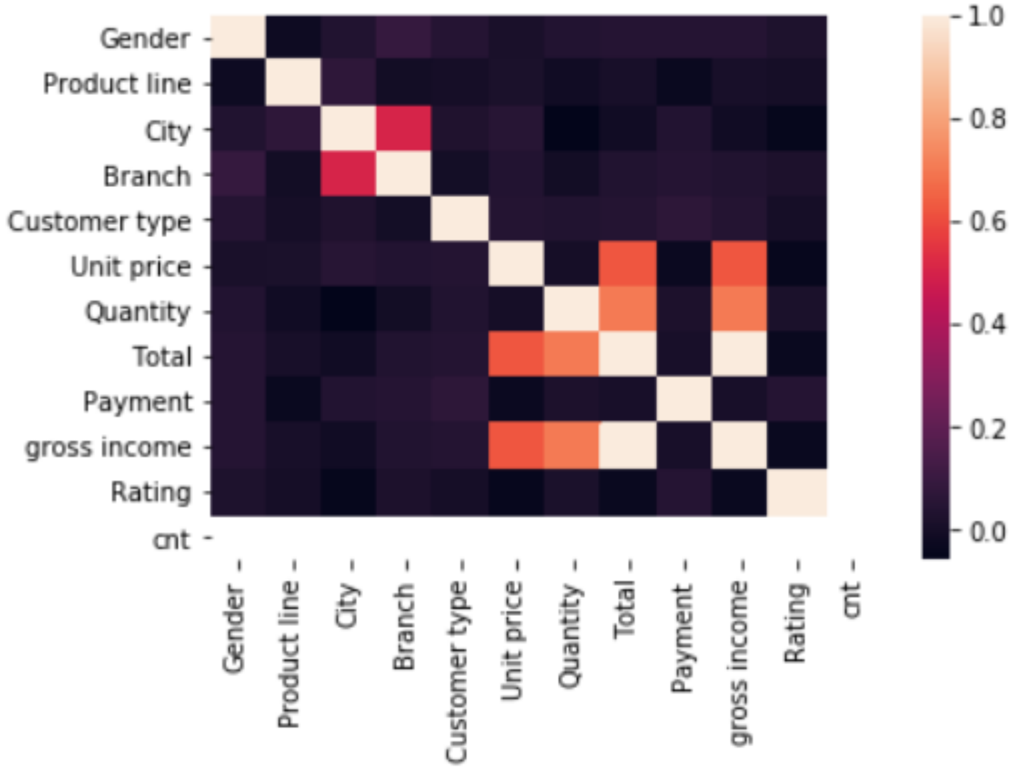


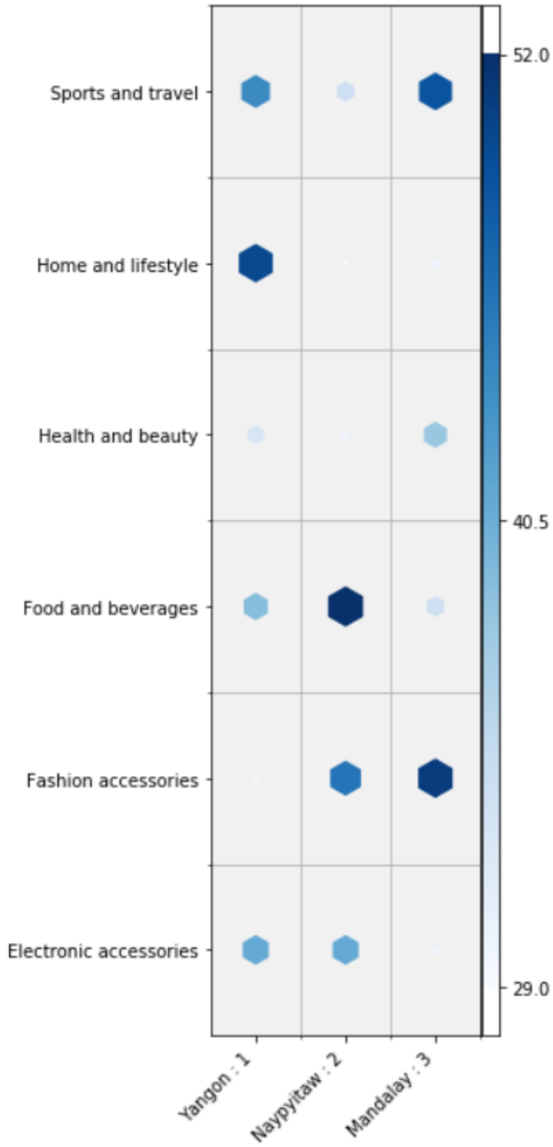

Customer Satisfaction by Gender in Terms of Payment Method
Rating
10
9
8
7
6
5
4
Ewallet
Cash
Credit card
Payment Method
Gender
Female
Male


Customer Satisfaction by Gender in Terms of Payment Method
Day_of_Week
Saturday
Friday
Sunday
Monday
Thursday
Wednesday
Tuesday
0
10
20
30
40
50
Gross Income
Gender
Female
Male

CORRELATION MATRIX

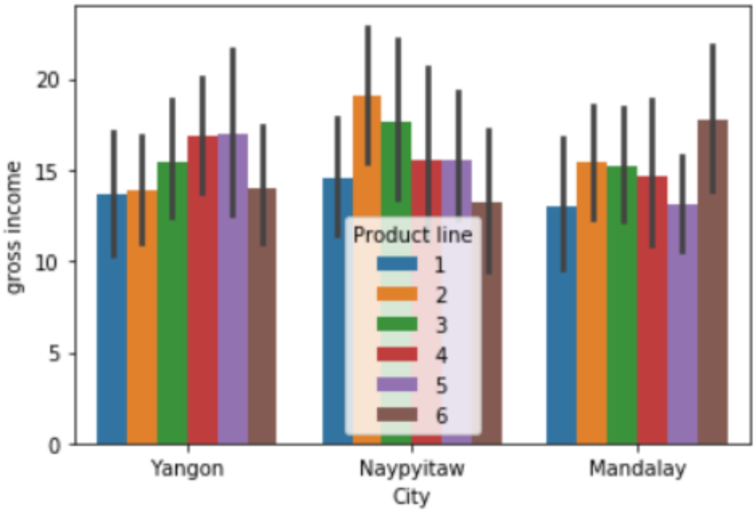




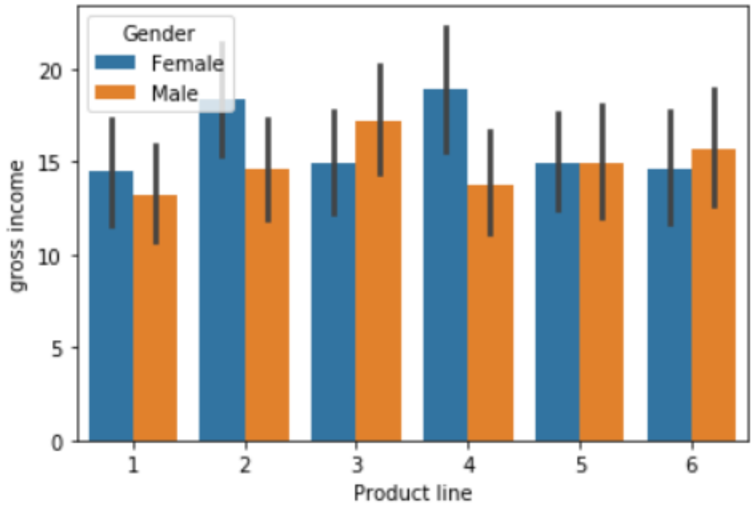


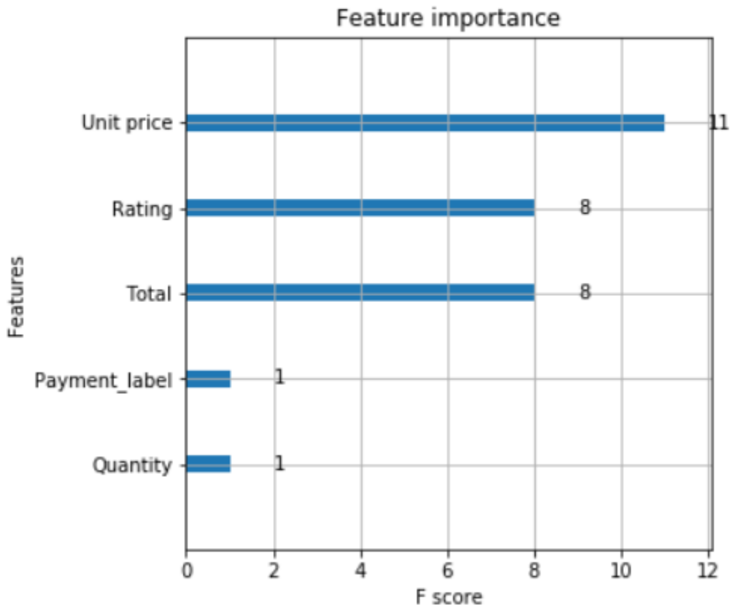
COORELATION MATRIX BETWEEN THE PRODDUCT LINE AND LOCATION

```
'Electronic accessories':1, 'Food and beverages':2, 'Sports and travel':3,'
' 'Home and lifestyle':4, 'Fashion accessories':5, 'Health and beauty':6
```

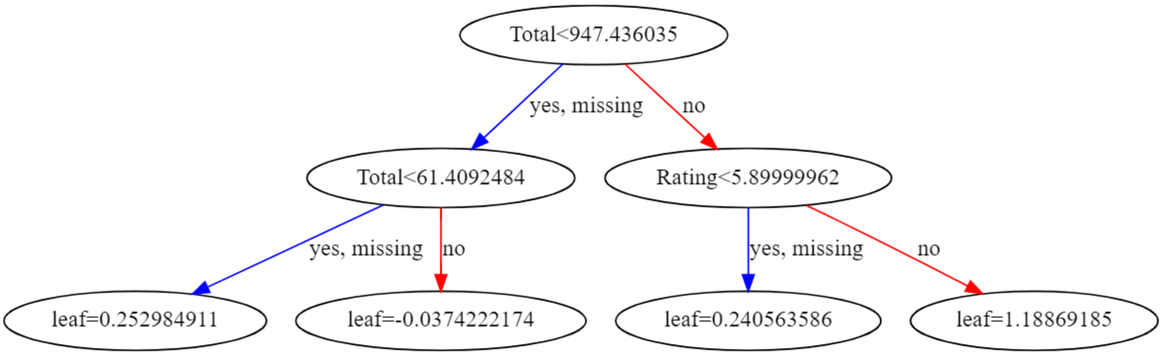


```
'Electronic accessories':1, 'Food and beverages':2, 'Sports and travel':3,'
' 'Home and lifestyle':4, 'Fashion accessories':5, 'Health and beauty':6
```

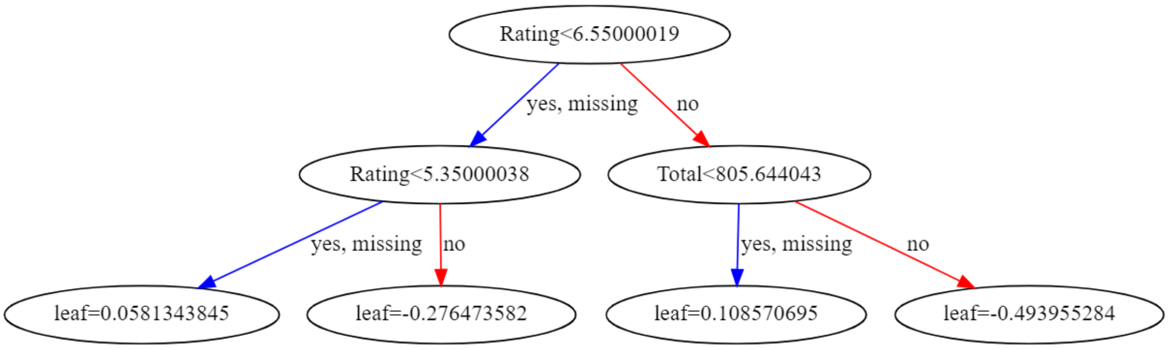


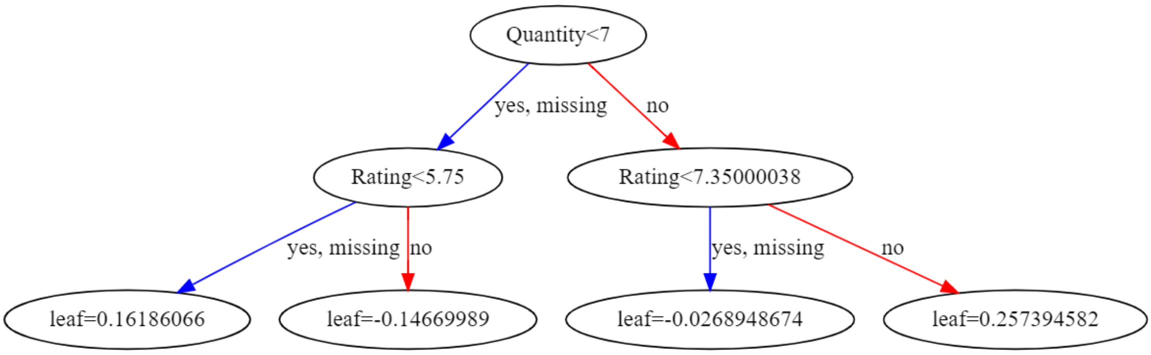


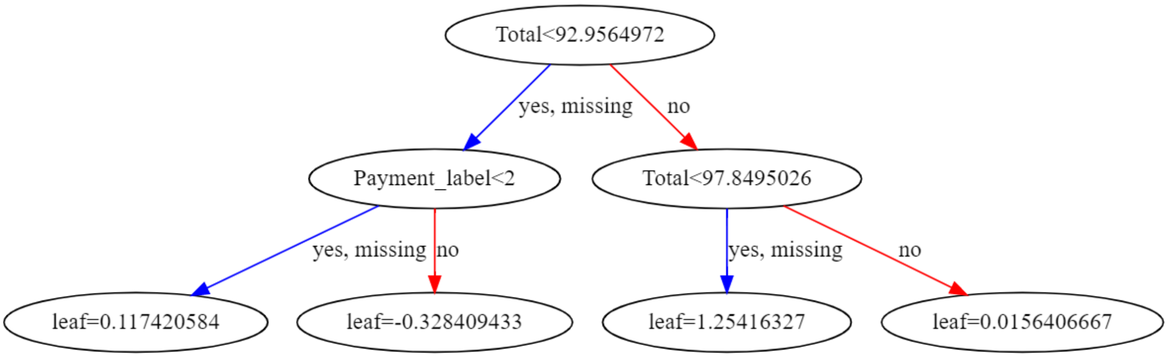
Feature importance
Unit price
Rating
Total
Payment_label
Quantity
Features
11
8
8
1
1
0
2
4
6
8
10
12
F score

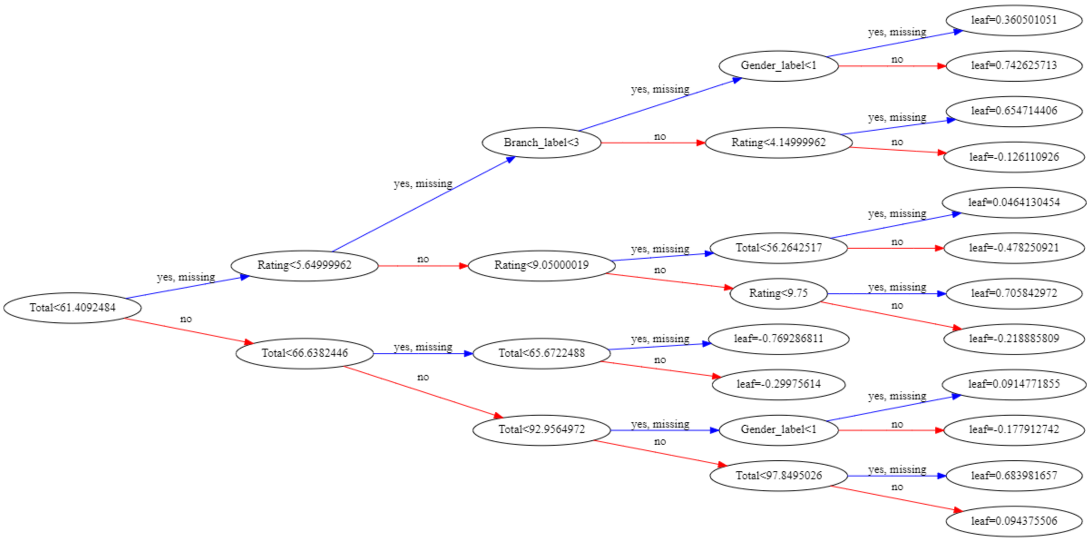


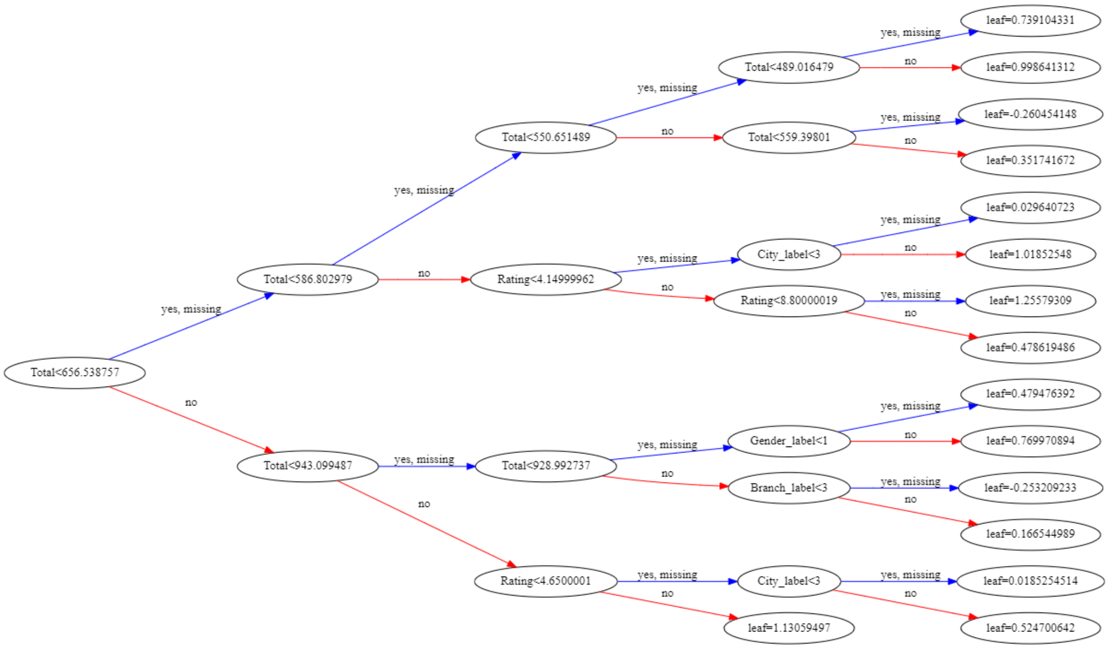
Total<947.436035
yes, missing
no
Total<61.4092484
Rating<5.89999962
yes, missing
no
yes, missing
no
leaf=0.252984911
leaf=-0.0374222174
leaf=0.240563586
leaf=1.18869185


Rating<6.55000019
yes, missing
no
Rating<5.35000038
Total<805.644043
yes, missing
no
yes, missing
no
leaf=0.0581343845
leaf=-0.276473582
leaf=0.108570695
leaf=-0.493955284


Quantity<7
yes, missing
no
Rating<5.75
Rating<7.35000038
yes, missing
no
yes, missing
no
leaf=0.16186066
leaf=-0.14669989
leaf=-0.0268948674
leaf=0.257394582


Total<92.9564972
yes, missing
no
Payment_label<2
Total<97.8495026
yes, missing
no
yes, missing
no
leaf=0.117420584
leaf=-0.328409433
leaf=1.25416327
leaf=0.0156406667





THE ABOVE ARE THE DECISION TREE WHICH WILL BE AUTOMATED FOR THE

CODE LINK :

https://colab.research.google.com/drive/1fkFiq1VXVkRBWRAukj8larJXPX69suVr#scrollTo=DZUP9SDe9PU5

https://colab.research.google.com/drive/1iB9jz7v7ZlWjjP2rl6BUqPUeKql9ZROn#scrollTo=bJUvstsoxY5W