

Walkart

Customer Analytics

Group Project Report

OPIM 5604 | Spring 2017 | Tuesday

Team 9

Aditya Rai, Gaurav Vijay, Namratha Kasineni, Rahul Kumar Thatikonda, Ruiqi Xing

Table of Contents

1. [**Executive Summary**](#_30j0zll)

[Problem Statement](#_3znysh7)

1. **Data Preprocessing**

Data Dictionary

1. **Data Visualization and Pattern Discovery**

Data Exploration

1. **Predictive Modeling**

Linear Regression

Regression Tree and Ensemble Techniques

Neural Network

1. **Model Validation and Implementation**
2. [**Plans for future upgrades**](#_2jxsxqh)

Market Basket Analysis

# **Executive Summary**

Team Members:

Aditya Rai, Gaurav Vijay, Namratha Kasineni, Rahul Kumar Thatikonda, Ruiqi Xing

## This project is conducted on the dataset taken from an ongoing data challenge on Analyticsvidhya.com and it consists of transactional data of purchases made at a retail store. Our objective is to predict the discounted price of the product purchased, based on customer demographics. Additionally, we also performed market basket analysis on the data. Any other observation, unrelated to the prediction goal, is also recorded and summarized in the paper.

## Problem Statement

A retail company “Walkart” wants to understand the customer purchase behavior (specifically, purchase amount) against various products of different categories. They have shared purchase summary of various customers for selected high volume products from last month. The data set also contains customer demographics (age, gender, marital status, city\_type, stay\_in\_current\_city), product details (product\_id and product\_category) and total purchase\_amount from last month. Now, they want to build a model to predict the purchase amount of customer against various products which will help them to create personalized offer for customers against different products.

# **Data Preprocessing**

The end goal of our project is to come up with personalized discounted pricing for users based on the customer demographics and shopping behavior. To achieve that, we performed pre-processing on various fields in the data before building visualizations and predictive models. We changed the data type and modeling type of certain variables as appropriate. We then performed outlier and missing value detection but the data did not contain any outliers or missing values. Also, most of the variables are categorical pertaining to customer demographic information and thus did not require any pre-processing.

## Data Dictionary

Below are the variables available in the data set:

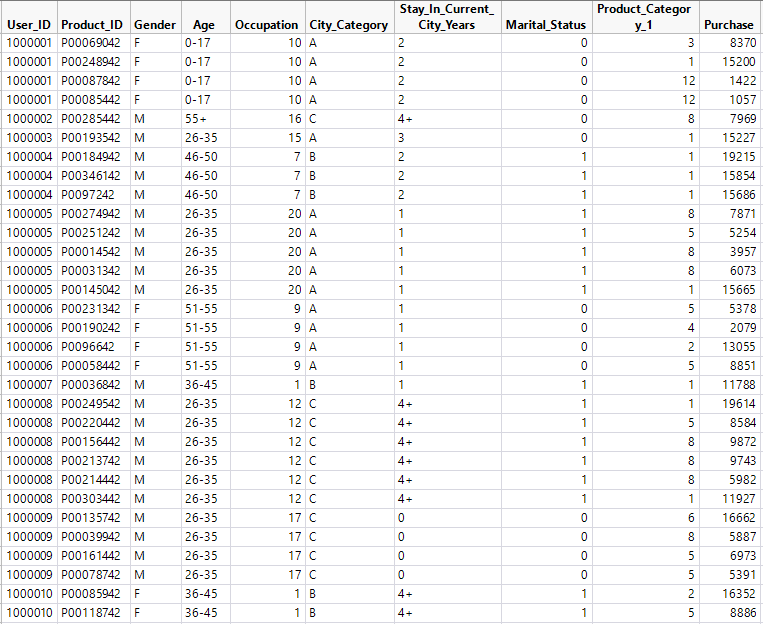
1. **User\_ID:** This variable contains unique customer id of each customer who performed the transaction in the store. Initially the variable was continuous numeric variable and we changed it to nominal character variable.
2. **Product\_ID:** This variable contains product id for each product that a customer bought in a transaction. This is a nominal character variable.
3. **Gender:** This variable contains gender information of the customer in the form of character ‘M’ or ‘F’ for Male or Female respectively. This is a nominal character variable.
4. **Age:** This variable contains age of the customer in the form of age bins such as ‘0-17’, ‘18-25’, ‘26-35’, ‘36-45’, ‘46-50’, ‘51-55’ and ‘55+’. Initially the variable was nominal character variable and we changed it to ordinal character variable.
5. **Occupation:** This variable contains different customer’s occupation categories which are masked and has values ranging from 0 to 20. Initially the variable was continuous numeric variable and we changed it to nominal numeric variable.
6. **City\_Category:** This variable contains the city category where the customer resides. The values are ‘A’, ‘B’ and ‘C’. Initially the variable was nominal character variable and we changed it to ordinal character variable.
7. **Stay\_In\_Current\_City:**  This variable contains the stay duration of customer in the current city and the value is ranging from 0 to 4+. Initially the variable was nominal character variable and we changed it to ordinal character variable.
8. **Marital Status:** This variable contains the marital status of the customer and has the values 0 and 1.Initially the variable was continuous numeric variable and we changed it to nominal numeric variable.
9. **Product\_Category\_1:** This variable contains the masked product categories represented by values ranging from 1 to 20. All the 3631 different products are categorized under these 20 product categories. Initially the variable was continuous numeric variable and we changed it to nominal numeric variable.
10. **Purchase:** This variable contains the billed price of the product. This is the variable we want to predict. This a continuous numeric variable.

Figure 1: Raw variables provided in the data set.

Below are the variables we created using feature engineering:

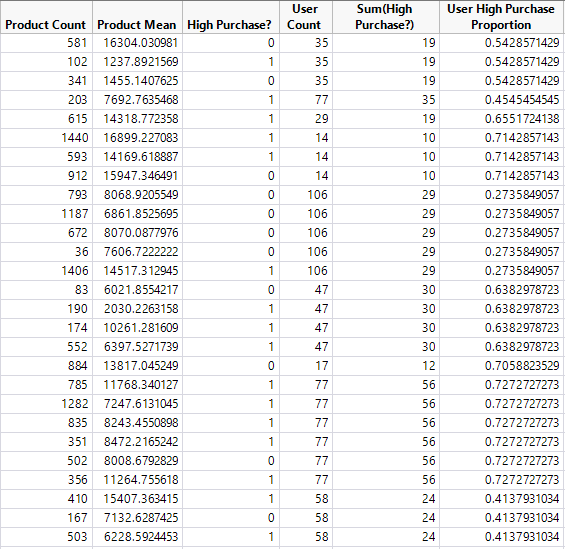
1. **User Count:** This variable contains the number of different products a customer has bought. This is a continuous numeric variable.
2. **Product Count:** This variable contains the number of times a product was bought by different customers.This is a continuous numeric variable.
3. **Product Mean:** This variable contains the average price of each product available in the store. Basically, each of the three thousand plus products was purchased multiple times by different users and at different prices depending upon user’s profile. We calculated the average price of each product and stored the value in this variable against each transactional row. This is a continuous numeric variable.
4. **High Purchase?:** This variable contains values 0 or 1 depending upon whether a product was sold to a particular user at a price lower or higher than the average price (represented by ‘Product Mean’ column) of that product. If the user was offered a price that was lower than the average price of that product, than a value of ‘0’ was assigned against that user and if the user was charged an amount that was higher than the average price of that product than a value of ‘1’ was assigned against that user. This is a nominal numeric variable.
5. **Sum (High Purchase?):** Thisvariable containsthe sum of total number of times a user has been charged an amount for a product which was more than the average price of that product. This is a continuous numeric variable.
6. **User High Purchase Proportion:** This variable containsthe ratio of the values of ‘Sum (High Purchase?) column and ‘User Count’ column. This column will provide information whether a user was charged more or less for the overall shopping done by him/her. If the value exceeds 0.5 then it means that the user was charged more than the product mean price for majority of the transactions and if the value is less than 0.5 then it means that the user was charged less than the product mean price for majority of the transactions. This is a continuous numeric variable.

Figure 2: New variables created using feature engineering.

# **Data Visualization and Pattern Discovery**

Before running the predictive model, we performed visualizations on different fields to understand the relationships between them. We extracted summarized information such as total number of distinct customers and total number of distinct products. To extract the aggregated information from the transactional data, we executed different data exploration techniques and identified distribution of data among different categorical variables such as city category, age, gender, marital status, and product categories.

## Data Exploration

As the first step of data exploration, we calculated the summary level statistics of user and product related variables.

Number of distinct users: 5891

Number of distinct products: 3631

Number of different product categories: 20 (1 to 20)

Number of occupation categories: 21 (0 to 20)

Number of age bins: 7 (‘0-17’, ‘18-25’, ‘26-35’, ‘36-45’, ‘46-50’, ‘51-55’ and ‘55+’)

Number of city categories: 3 (‘A’, ‘B’ and ‘C’)

Number of categories in Stay in Current City Years: 5 (‘0’, ‘1’, ‘2’, ‘3’, ‘4+’)

Below attached are the visualizations performed to extract useful insights and to identify relationships between different variables.

* **User level visualizations**

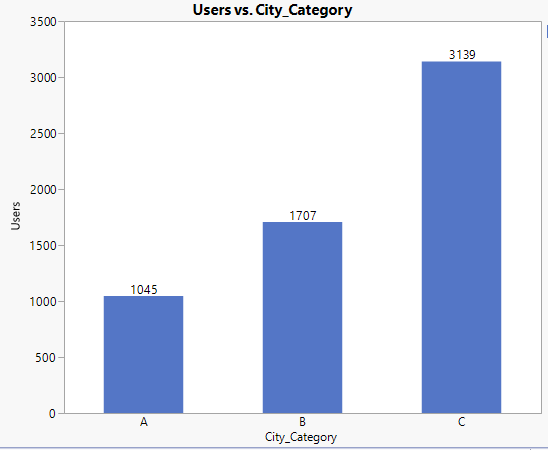
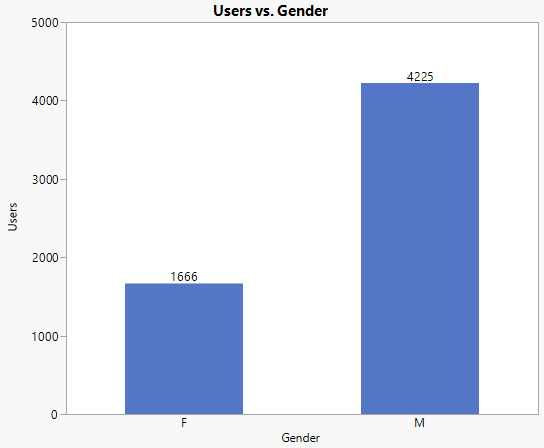
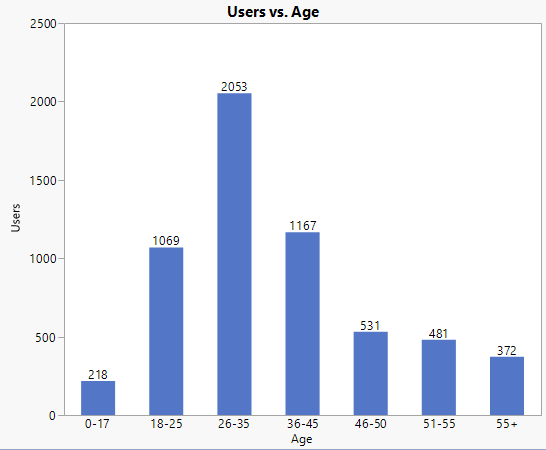
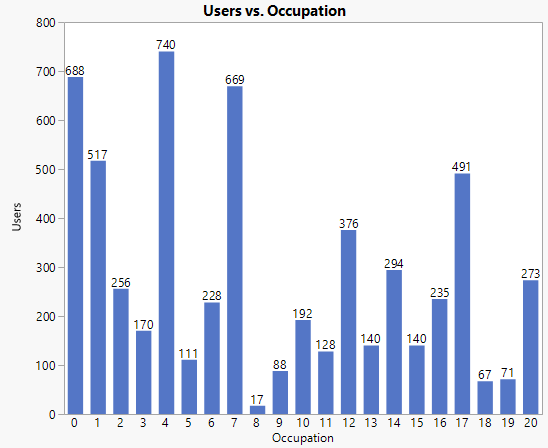
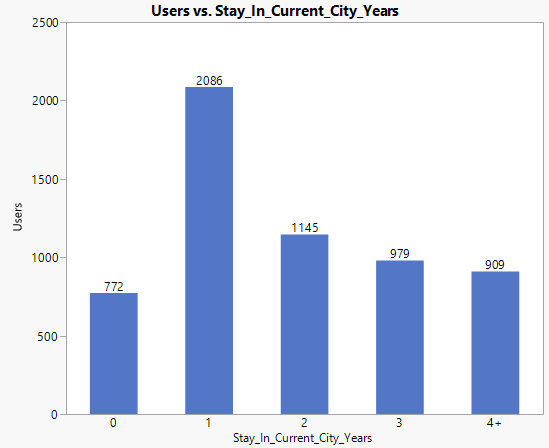
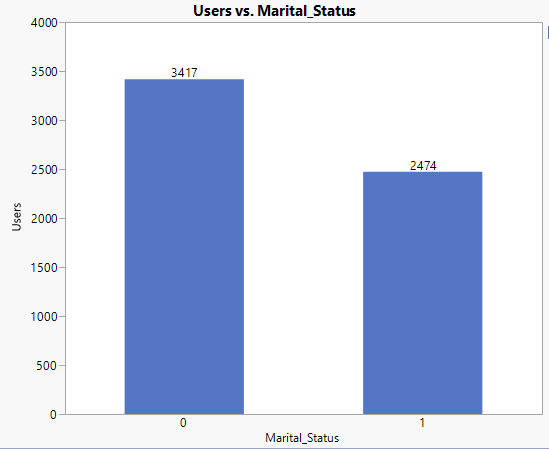


Figure 3: No. of users per city category. Figure 4: No. of users per gender category.

Figure 3 to Figure 10 show some insights on customer demographics. The major customer base for Walkart is from City Category C and among the existing customers - number of men customers exceeds number of women customers. The customers with age segment 26-35 correspond to around 33% and occupation categories ‘0’, ‘4’ and ‘7’ constitute around 33%. Around 50% of the customers are living in their current city since last 1 or 2 years. Also, most of the customers are single/unmarried.

Figure 5: No. of users per age bin category. Figure 6: No. of users per occupation category.



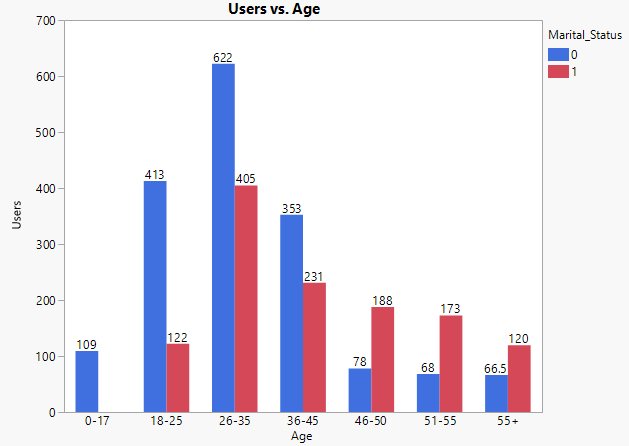
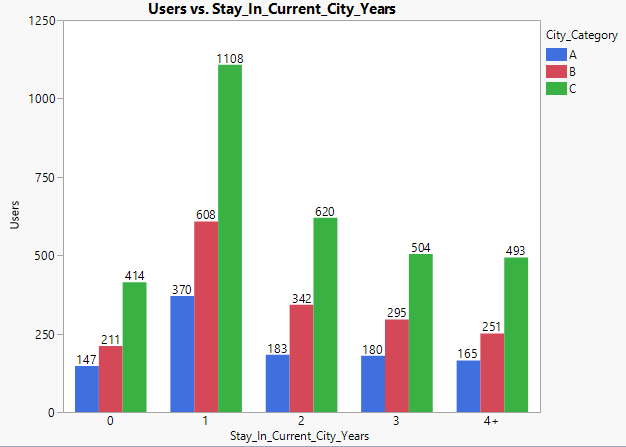
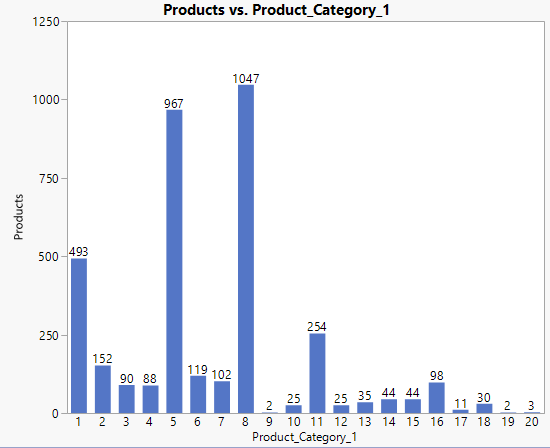
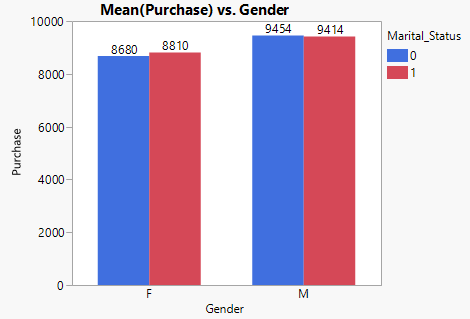
 Figure 7: No. of users per age stay duration category. Figure 8: No. of users per marital status category.

Figure 9: No. of users per age stay duration and city category. Figure 10: No. of users per age and marital status category.

* **Product and Price level visualizations**

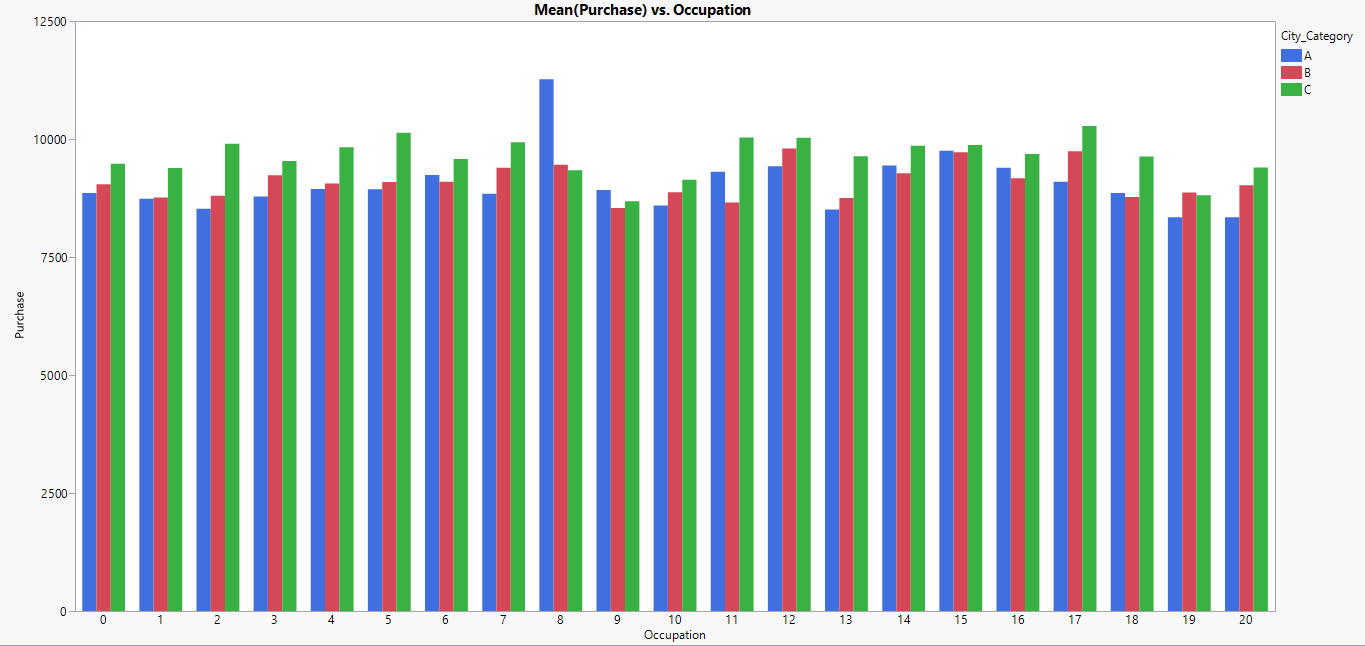
Figure 11: No. of products per product category. Figure 12: Mean purchase price by gender.

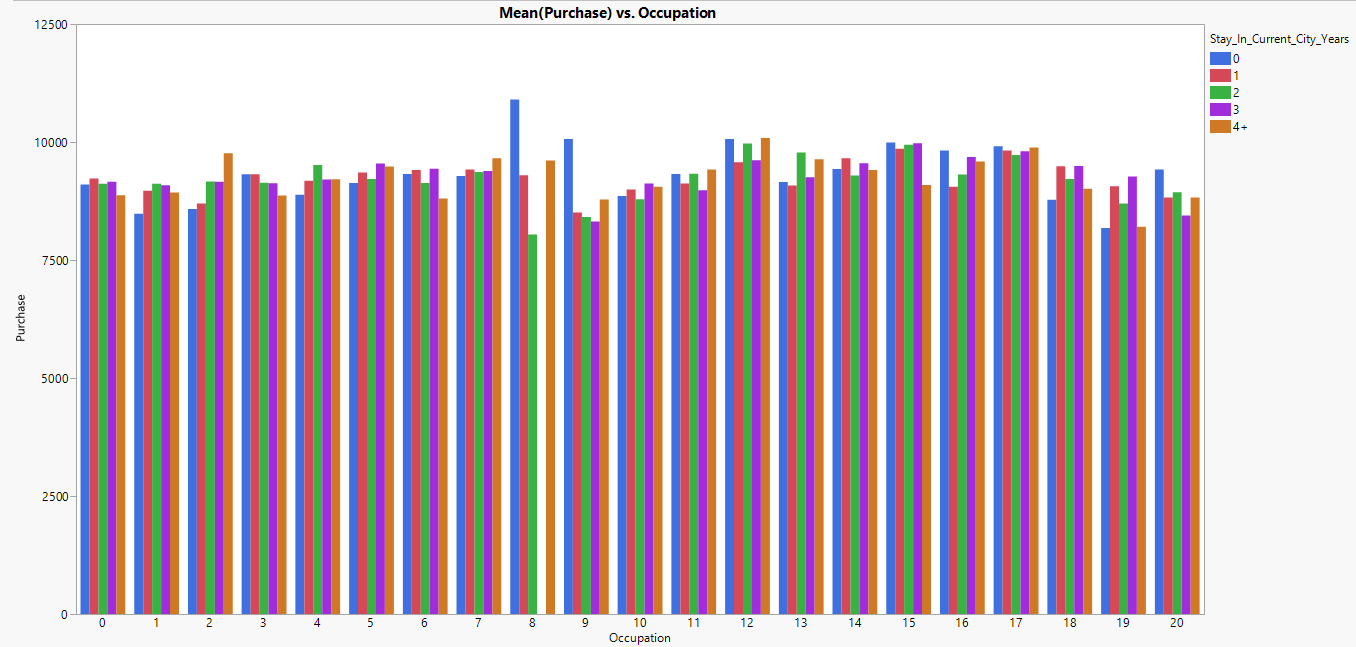
Figure 13: Mean purchase price by occupation and city category.

Figure 14: Mean purchase price by occupation and years of stay in current city.

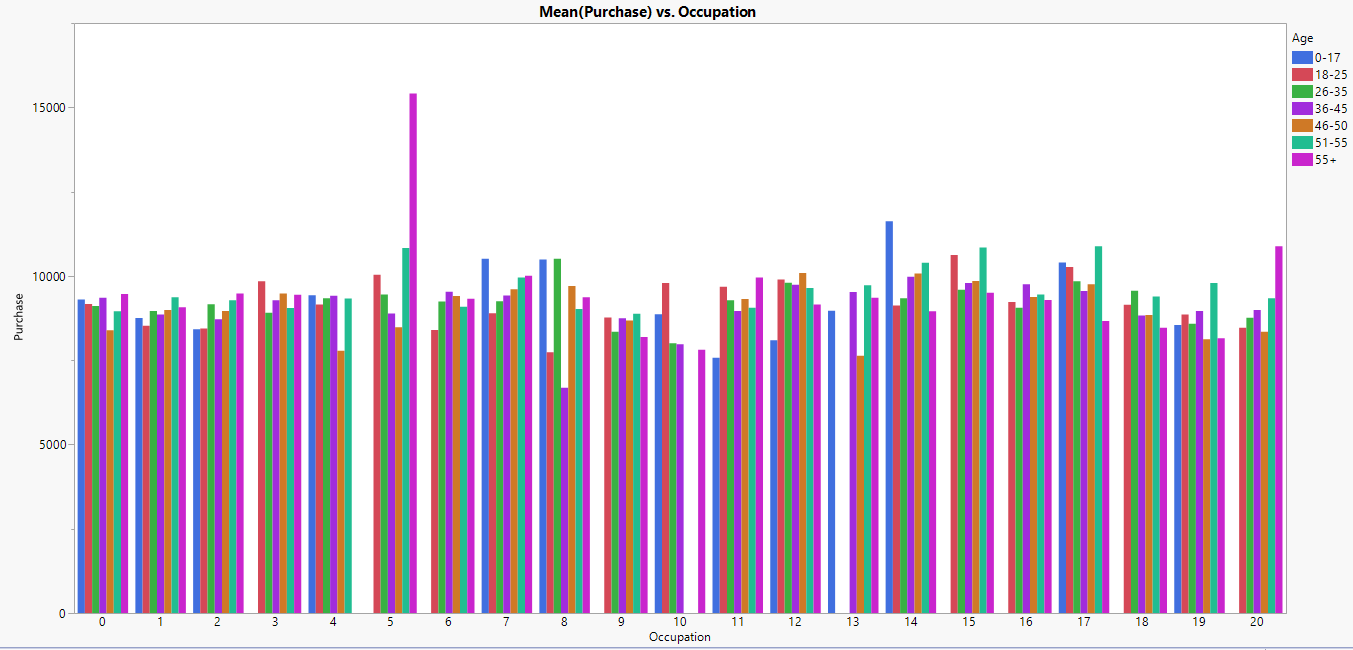
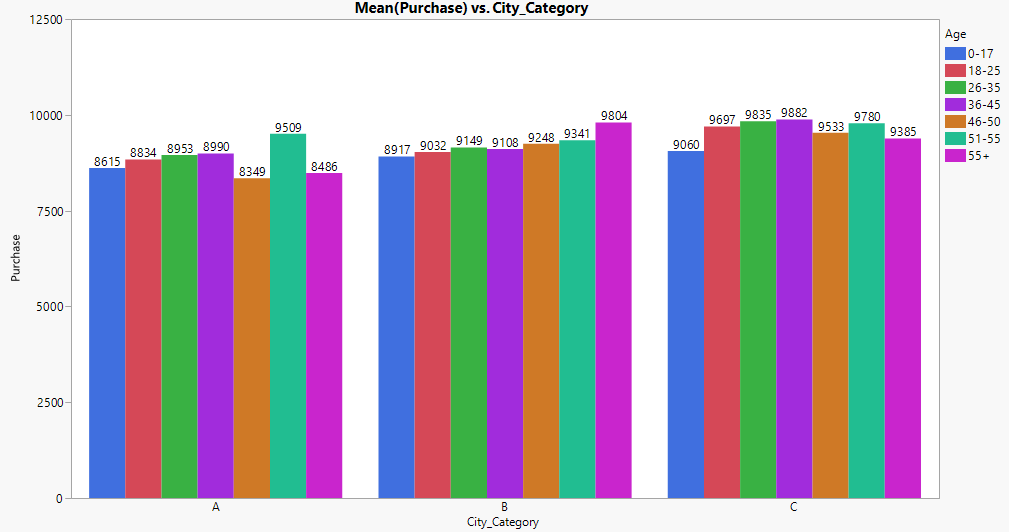
 Figure 15: Mean purchase price by occupation and age bins.

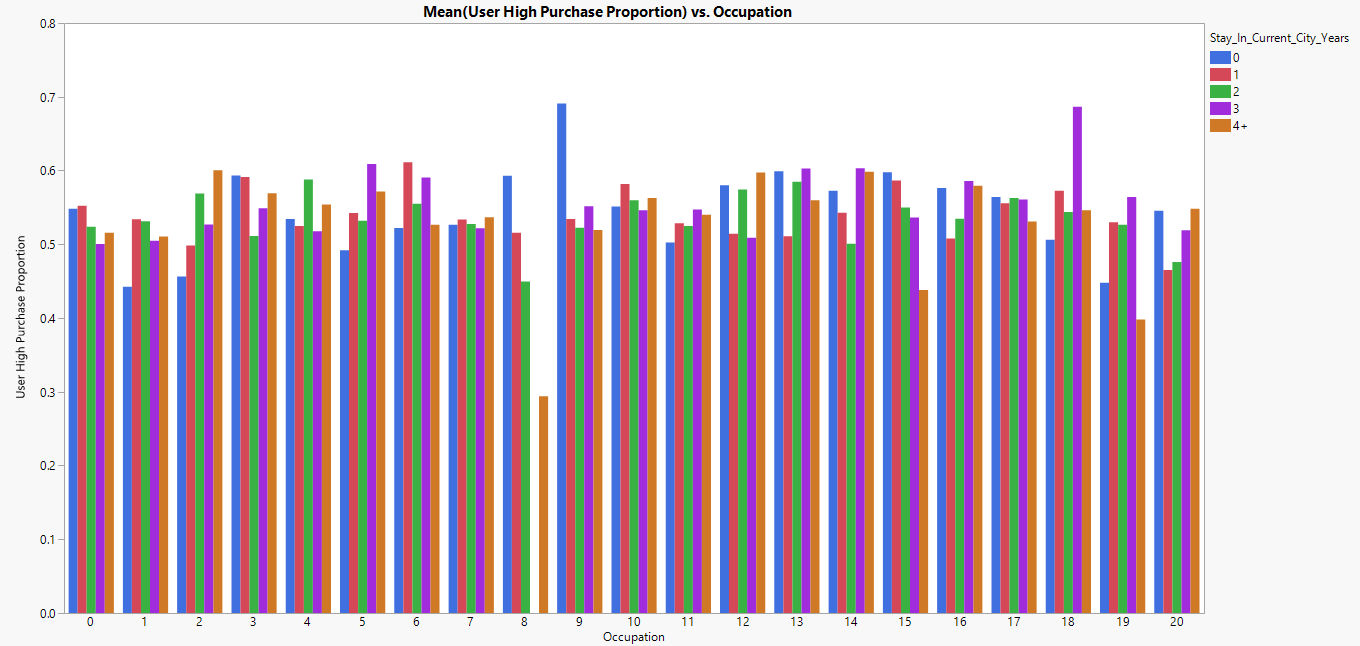
Figure 16: Mean purchase price by city category and age bins.

Figure 17: Mean User High Purchase Proportion by and occupation and years of stay in current city.

Figure 11 shows the distribution of products among different product categories and we can see that product categories 5 and 8 constitute around 66% of the products. Through figure 12, we found that mean purchase price for single male customers is highest and the same for single female customers is lowest. In figure 13, its clearly visible that customer with occupation category ‘8’ and city category ‘A’ combined, has the highest mean purchase value. A customer from city category ‘A’ and occupation category ‘19’ or ‘20’ has the lowest mean purchase price. Relationship between mean purchase price, occupation category and stay in city in years is shown in figure 14. The lowest mean purchase price is for customers who are staying in the current city since last 2 years and are from occupation category ‘8’. Customers with age more than 55 years and occupation category ‘8’ has the lowest mean purchase price. Moreover, the customers with age more than 55 years but from the occupation category ‘5’ has the highest mean purchase price. This contrast between the relationship of age and occupation is shown in figure 15. City category and Age are also related in deciding the mean purchase price for a customer. Figure 16 shows that customers with age between 36 and 45 years and from city category ‘C’ have highest mean purchase price. The relationship between user high purchase proportion, occupation and stay in current city in years is well explained by figure 17. The customers from occupation category 9 and staying in current city since less than a year or those from occupation category 18 and staying in current city since last 3 years have paid the purchase amount higher than the product mean for highest number of transactions. Interestingly, the customers from occupation category 8 and staying in current city since more than 4 years have the lowest user high purchase proportion of 0.3 but the mean purchase price for the same customer segment is very high as shown in figure 14. This mean that the products bought by these customers are from category with high mean purchase price but these customers are getting higher discount as compared to other customers.

# **Predictive Modeling**

Through visualizations, we identified the columns that looked important as predictor variables and would play a vital role in predicting the discounted price. Next, we created predictive models to predict the personalized discounted price of the products for the customers based on the customer profile and his historical purchase behavior. We used different algorithms and identified the best model out of all the models based on the model results.

## Linear Regression

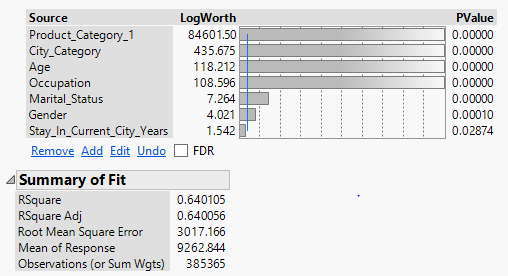
As the next step of the group project, we ran linear regression model with only the initial variables present in the data set. The model did not perform well as the RSquare was just 0.64 and RMSE was 3017.166, which seemed very high. We realized that we need to use some other technique to improve our model.

Figure 18: Linear Regression model with variables present in data.

We used the technique called feature engineering to introduce new features or variables based on the functional knowledge about retail transactional data. As the User ID and Product ID were not giving significant information regarding a product or a user, we calculated some aggregated variables such as Product Mean, User Count and Product Count. We also calculated a proportion variable User High Purchase Proportion, which gives the percentage of number of times a user bought any product at a price higher than the mean product price. We ran the linear regression model including new fields and as a result the RSquare jumped to 0.75 and the RMSE dropped to 2482.70. The RMSE and RSquare for validation set came out as slightly less than the RMSE and RSquare of training set.

We observed that the new variables - Product Mean and User High Purchase Proportion were identified as the strongest predictors by linear regression model. The intercept coefficient for User High Purchase Proportion came out as 5655.58 which clearly shows that User High Purchase Proportion is the strong contributor in increasing the purchase price. Among the variables that were present in the data set, occupation and gender were identified as the most important contributing variables by the linear regression model.

## Figure 19: Linear Regression model with all the variables.

## Figure 20: Linear Regression model result.

## Regression Tree and Ensemble Techniques

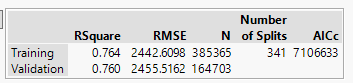
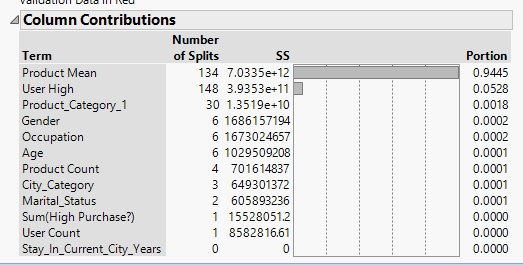
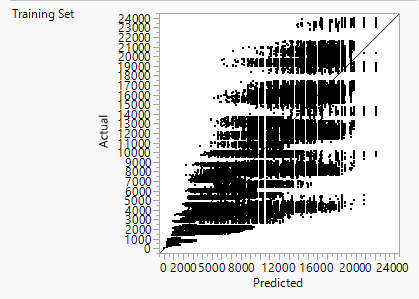
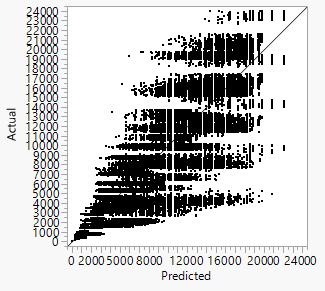
 After Linear Regression, we performed Regression Tree on the data set. The model gave the RSquare of 0.764 and RMSE of 2442.60 for training data set. The validation data set did not outperform the training data set. On looking at the column contribution and leaf report, we observed that Product Mean and User High were again the two most important variables in predicting the purchase price for a customer.

Figure 21: Regression Tree model result.

Figure 22: Actual v/s Predicted Purchase price for training and validation data set 

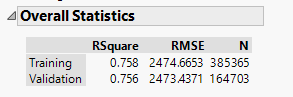
 We also ran Bootstrap Forest and Boosted Tree algorithms on our data but the results were not significantly better in comparison to the results from Regression Tree.

Figure 23: Bootstrap Forest model result. Figure 24: Boosted Tree model result.

## Neural Network

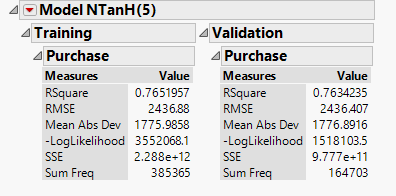
 After Regression Tree and Ensemble Techniques, we performed Neural Network. Parameters used were 1 layer with 5 nodes. The RSquare value for training as well as validation was almost same as 0.76, with the value for validation set being slightly lower. The RMSE for validation set was slightly lower than that of training set.

Figure 25: Neural Network model result.

# **Model Validation and Implementation**

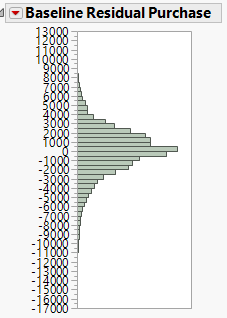
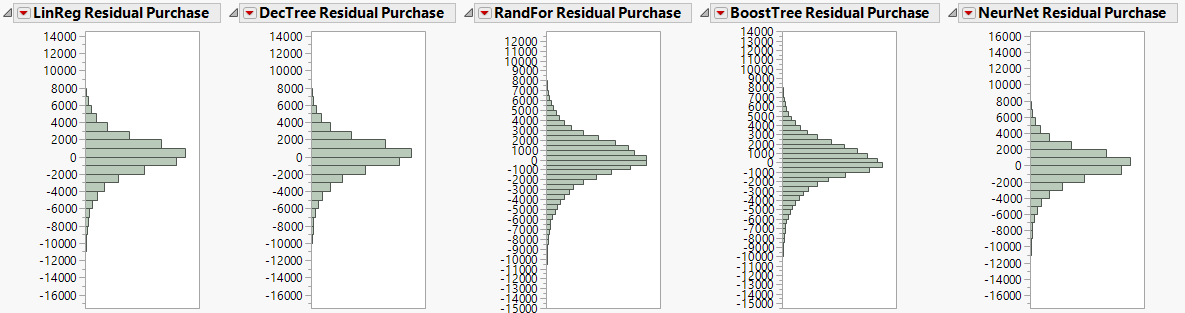
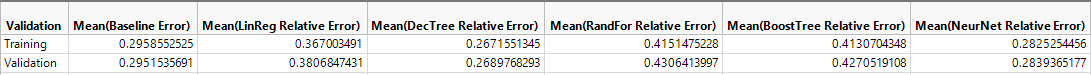
After running the models, we calculated the residual errors and relative errors for all the models and baseline case, and here are the comparisons. The residual distribution for all the models seem to follow normal distribution. We can clearly see that the relative errors for regression tree model and neural network model are lower than baseline relative error and the one for regression tree is the lowest between the two models. Thus, we choose to implement regression tree model for our project.

Figure 26: Comparison of residual distribution among different predictive models and baseline case.

Figure 27: Comparison of relative error among different predictive models and baseline case.

# **Plan for future upgrades**

The objective of this project was to predict the discounted price of a product based on customer demographics and purchase behavior. As a next step, to help Walkart better suggest recommendations to customers on previous purchase transactions, we performed market basket analysis. With the limitations of current software capabilities, we performed this analysis on a sample data, extracted one rule and calculated the values corresponding to this rule from the entire dataset.

## Market Basket Analysis

Association rule for (P00265242 & P00025442 & P00110742) ► P00112142

Total transactions - 5891

No. of times product P00112142 is bought - 1562

No. of times products (P00265242 & P00025442 & P00110742) are bought together – 248

No. of times products (P00265242 & P00025442 & P00110742) & P00112142 are bought – 112

Support – 112/5891 = 0.019; Confidence – 112/248 = 0.45

Expected – 1562/5891 = 0.26; Lift Ratio – 0.45/0.26 = 1.73

As the lift ratio between (P00265242 & P00025442 & P00110742) and P00112142 is greater than 1, there is an association between LHS and RHS.

By identifying the association rule between the products, we can create a recommender system for customer which would recommend customers the products they can buy based on the products already placed in their shopping cart. In addition to this, we can create a model that would provide a better personalized offer for combo deals and bulk orders.