**MARKET ANALYSIS REPORT FOR FRESH DIRECT RETAILER**

Abstract

This report will look into the sample dataset known as ‘SAMPLE\_FD\_DATA.csv’ which contains information about the customer’s age, income geographical location, orders, and delivery details. The data contains the sales made from customers and the percentage of certain products in every order. In this report different qualitative and quantitative techniques will be used to make a regression model and a correlation analysis. The regression model will be used to predict sales and the correlation analysis will be used to group customers in based on their similarities.

**Introduction**

This dataset shows records of different features of a customer’s order and also takes into account the proportion of products of certain products in every order. It mostly contains continuous variables and is most useful for correlation and regression. However frequency analysis is useful to know the proportion of gender, a customer’s location, and loyalty segment.

**Exploratory data analysis**

This is where the data is summarized using graphs, frequency tables and the measures of central tendency are calculated and the distribution of the target variable is determined. This will assist in choosing the best regression model for predicting sales. This part of the report shows which variables correlated with the target variable but not with each other. This is to ensure all the assumptions are met for the model to perform better at predicting sales.

**Qualitative analysis**

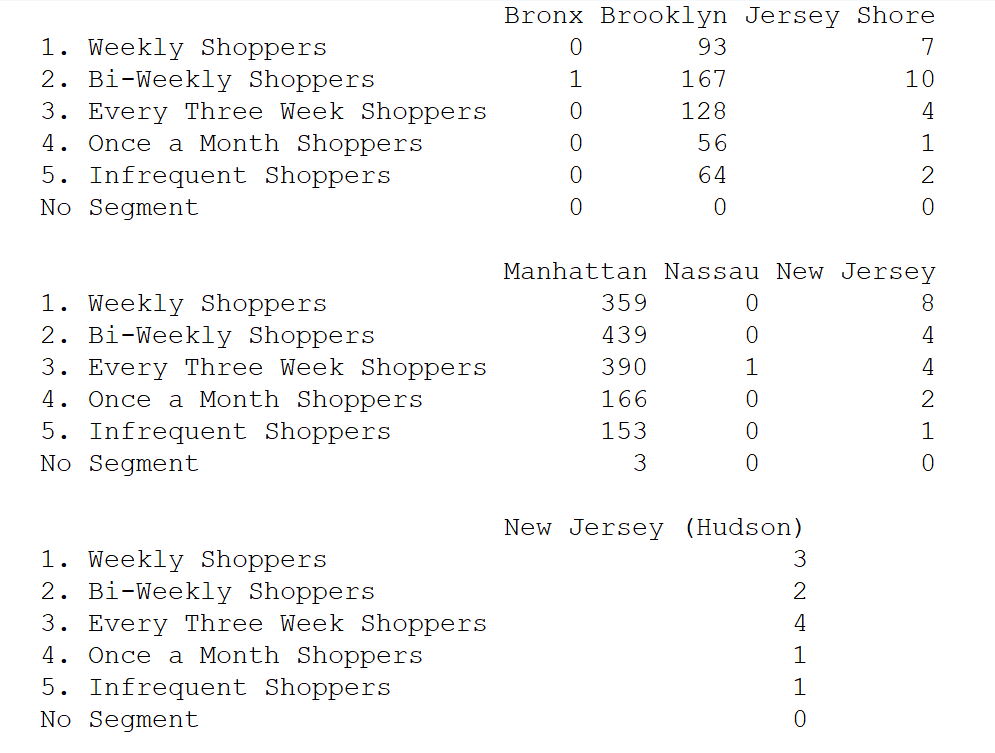
Table 1. Gender distribution

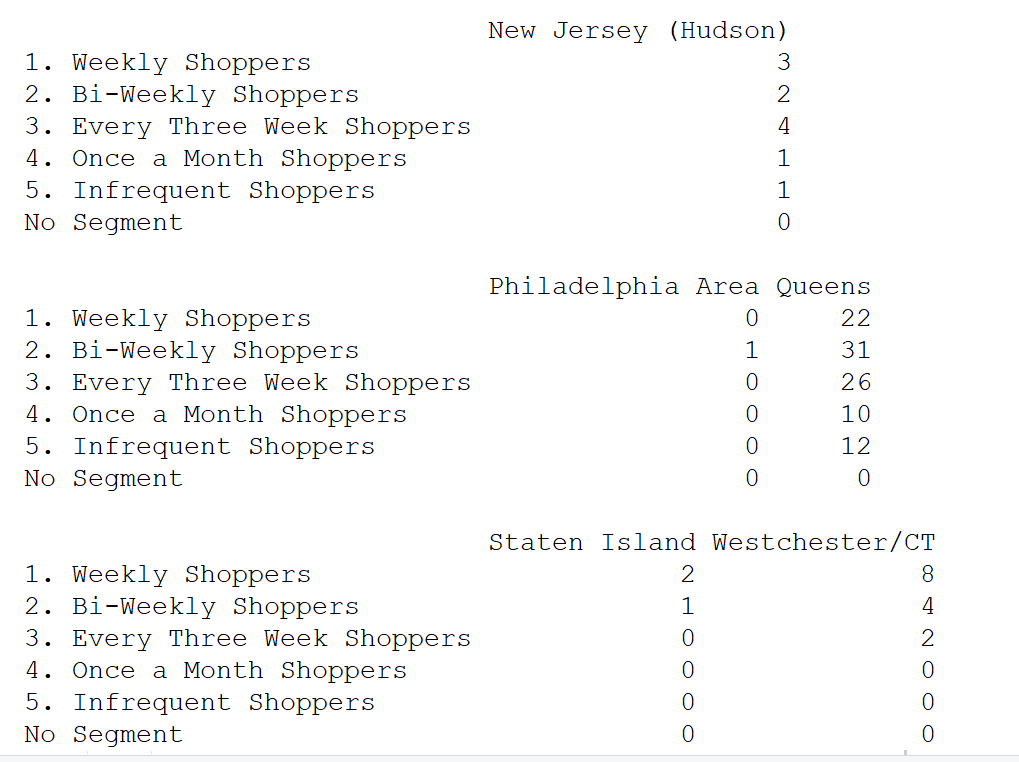
|  |  |  |
| --- | --- | --- |
| Female | Male | Missing |
| 1596 | 745 | 659 |

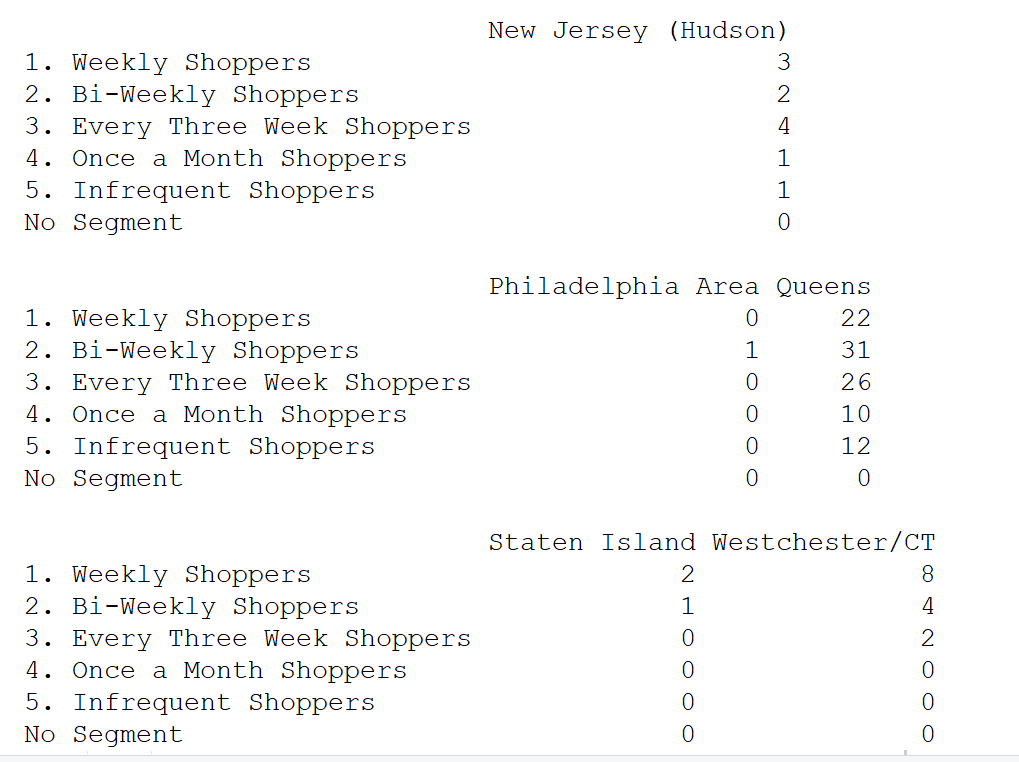
Table 2. Loyalty segment analysis

|  |  |
| --- | --- |
| **Segment** | **Count** |
| 1. Weekly Shoppers | 631 |
| 3. Every Three Week Shoppers | 783 |
| 5. Infrequent Shoppers | 413 |
| No Segment | 4 |
| 2. Bi-Weekly Shoppers | 826 |
| 4. Once a Month Shoppers | 342 |
| 7. 90 Day New Shoppers | 1 |

**Geographical segmentation**







After summing up all the orders per geographical area and loyalty segment the finding is that:

* Weekly shoppers are mostly in Brooklyn and least in Philadelphia, Nassau and Bronx.
* Bi-weekly shoppers are mostly in Manhattan and least in Nassau
* Every three week shoppers are mostly in Manhattan and least in Staten Island.
* Once a month shoppers are mostly in Manhattan and least in Staten Island, Westchester/CT, Philadelphia Nassau and Bronx.
* Infrequent shoppers are mostly in Manhattan.

**Quantitative analysis**

**Correlation Analysis**

Age with income =0.0997

Income with sales=0.0645

Sales with age = 0.1854

Sales with 12\_month delivery fee paid=0.1854

Multiple Linear regression

A multiple linear regression model is model used to explain a relationship between one dependent variable and two or more independent variable. In this dataset Sales is the dependent variable and the following are the independent variables:

Age

Gender

Income

12 month. Delivery\_fee\_paid

12 month. Orders\_w\_promo

12 month. Discount\_amount

12 month Delivery pass used

Loyalty segement

12 month. Orders

|  |  |  |
| --- | --- | --- |
| R- squared | P- value | Adjusted R squared |
| 0.5869 | p-value: < 2.2e-16 | 0.5845 |

According to the R –squared value, it shows that the independent variables explain 59% variation in the Sales. The Adjusted variable indicates that each independent variable explains 58% variation in Sales. The p-value justifies the model because it is less than the allowed 0.05 mark, therefore the model is statistically significant in predicting sales.

According to the p-value of each independent variable only income, delivery fee, 12 month. Discount amount, 12 month. Orders, 12 month. Orders with promotion turned out to be significant in predicting sales.

**Recommendations the CFO:**

* According to the analysis attention should be given to younger customers because age has an inverse proportional relationship with sales, as age reduces sales increases.
* The analysis recommends reduction to orders with promotion attached to them because as the number of these orders increase the sales reduce.
* Also the report recommends focus be given to delivery fee paying customers, delivery pass using customers and also customers with discount on their products.

**Recommendations to the COO:**

Focus delivery in Manhattan and Brooklyn because this is where most customers are geographically located.

**Key findings for CEO:**

* To increase sales delivery could marketing activities could target young customers with high income.
* Products with promotion attached to them do not increase sales.
* Discount on products increases sales and also delivery fee has a significant positive impact on sales. Most customers come from Manhattan and Brooklyn.

Appendix

R code:

#Data processing

FD=read.csv("SAMPLE\_FD\_DATA.csv",header = T)

str(FD)

#descriptive statistics

attach(FD)

#Descriptive statistics

#Qualitative analysis

table(GENDER)

table(LOYALTY\_SEGMENT)

#Quantitative analysis

#Correlation analysis

cor(FD$X12\_Mo\_Sales)

#SEGMENTATION OF CUSTOMERS

table(FD$LOYALTY\_SEGMENT,FD$GEOGRAPHY)

#MODEL

model=lm(FD$X12\_Mo\_Sales~FD$X12.Mo..DELIVERYPASS\_USED+FD$X12.Mo..ORDERS\_W\_PROMO+FD$GENDER+FD$AGE+FD$INCOME+FD$GENDER+

FD$X12.Mo..DELIVERY\_FEE\_PAID+FD$X12.Mo..DISCOUNT\_AMOUNT+

FD$X12.Mo..Orders)

summary(model)