Grocery Store Customer Analysis

IST 687

12/21/2021

Jaclyn Karboski, Ash Wan, Joaquin Rodarte, Brad Engelbert

Dataset description

Title: Customer personality analysis: analysis of company's ideal customers

Source information:

Customer personality analysis: Analysis of company's ideal customers. (2021). [Data file]. Retrieved from https://www.kaggle.com/imakash3011/customer-personality-analysis/metadata

Description:

The dataset is a customer personality analysis that tracks customers' demographics, their purchase history, reactions to promotions, and place of purchase. The set consists of 29 variables and 2240 rows of data, for 64,960 objects.

Attributes:

Personal:
ID: Unique customer ID (num)
Year_Birth: Birth year (num)
Education: Highest level of education (chr)
Marital_Status: Marital Status (chr)
Income: Household income (num)
Kidhome: Number of children < 13 in the household (num)
Teenhome: Number of teens > 12 in the household (num)
Dt_Customer: Date of enrollment in system (chr)
Recency: Days since last purchase (num)
Complain: Made a complaint (num)

Product spending in last 2 years:
MntWines: Amount spent on wine (num)
MntFruits: Amount spent on fruits (num)
MntMeatProducts: Amount spent on meat (num)
MntFishProducts: Amount spent on fish (num)
MntSweetProducts: Amount spent on sweets (num)
MntGoldProds: Amount spent on gold (num)

Place of purchase:

NumWebPurchases: Number of web purchases (num)
NumCatalogPurchases: Number of catalog purchases(num)
NumStorePurchases: Number of in-store purchases (num)
NumWebVisitsMonth: Number of web visits per month (num)

Promotion reactions:
NumDealsPurchases: Number of purchases with a discount (num)
AcceptedCmp1: Accepted offer in 1st campaign (num)
AcceptedCmp2: Accepted offer in 2nd campaign (num)
AcceptedCmp3: Accepted offer in 3rd campaign (num)
AcceptedCmp4: Accepted offer in 4th campaign (num)
AcceptedCmp5: Accepted offer in 5th campaign (num)

Business questions

1. What are the demographics of our best customers and what variables predict a good customer?
2. Is there a correlation between promotions and place of purchase (in-store, website, catalog)?
3. Can we predict if a customer will accept a promotion offer and what variables will drive it?
4. How much of a factor is place of purchase on customer spending?

Data acquisition

For this project, we wanted to work with a dataset that a met a number of criteria. First, the dataset had to be accessible and ethically sourced. We were able to find datasets online through Kaggle.com and find a dataset that was available. The marketing data seemed to fit the purpose of our project and we wanted to work with a dataset that our group had familiarity with so we can understand the elements within and how they relate to one another.

Next, we needed the dataset to be a viable option with a limited number of at least 10,000 values. In deciding to utilize the “marketing campaign” dataset we had a good mixture of values in the form of characters, dates, and integers that would allow us a variety of analysis and to have a more complete dataset to work with.

The data set came in the form of an Excel spreadsheet, so the first step was to read it into R.

```
#Create new data frame called 'df' to store the data from the CSV file
df <- data.frame(read_excel("data/MarketingData.xlsx"))
```

Data cleansing, transformation, architecture description

NAs for Income

Some of the data in the column “Income” were NAs. In order to be able to use those rows, we replaced the NA with the average of the values surrounding the NA.

```
#How many missing/NA income values? 24 noted.
nrow(df[is.na(df$Income),])

#Replaced missing income na's with the average value of income before and after the na's
df$Income <- na_interpolation(df$Income)

#Validated that missing na's were replaced
nrow(df[is.na(df$Income),])
```

Date cleansing for Dates

The dataset in Excel had ‘Dt_Customer’ column under two different format which are ‘Date’ and ‘Character’. When we exported the Excel file into RStudios, RStudios converted the ‘Date’ format date as numeric character like so “41126”. The ‘character’ formatted dates stayed the same format like so “2014-12-12". Our goal was to have ‘Dt_Customer’ to be a Date format. The challenge we found was that due to the two different ways the date values were written, there wasn’t a step solution for formatting.

We converted ‘Dt_Customer’ values as numeric. The numeric character values returned as numbers while the character values returned null. Then, we converted the numeric numbers as Date and stored it in a dataframe called ‘Date_numeric’. The numbers were converted into Dates while the nulls stayed as null. We converted ‘Dt_Customer’ values again but this time as characters. The character values return the character values while the numeric character values returned null. The characters were then converted as Dates and stored in a dataframe called ‘Date_char’. We took the two dataframes and compared it to each other. Wherever a ‘Date_char’ row returns a null, we took the index of ‘Date_numeric’ and add that value into ‘Date_char’. From there, we replaced ‘Dt_Customer’ values of the main dataframe with the values of ‘Date_char’

```
#Changing the format of Dt_Customer so they are the same date format
cleanDate <- df$Dt_Customer #store date values in a vector

#Create a data frame containing Dates format converted from numeric format Date ie. '41126'
Date_numeric <- data.frame(as.Date(as.numeric(cleanDate), origin="1899-12-30"))
```

```
#Create a data frame containing Dates format converted from character format Date ie. '2014-12-12'
Date_char <- data.frame(as.Date(as.character(cleanDate), format="%d-%m-%Y"))

#Combine the values from Date_numeric into Date_char to keep it all in one data frame.
Date_char[is.na(Date_char)] <- Date_numeric[is.na(Date_char)]
names(Date_char) <- "Dt_Customer" #Name the column DT_Customer to keep the same naming as before
df$Dt_Customer <- Date_char$Dt_Customer #Add new Dates format into the df data frame to be used
```

Analysis, visualization, interpretation

#1 What are the demographics of our best customers and what variables predict a good customer?

Identifying best customers by demographics

The first step was to create a new column with sums of all spending by each customer and create a column that identified age by subtracting the current year from the date of birth in the data set.

```
dfGCust <- df
dfGCust$tSpent <- rowSums(dfGCust[c("MntWines", "MntFruits", "MntMeatProducts", "MntFishProducts", "MntSweetProducts", "MntGoldProds")])
dfGCust$age <- 2021-dfGCust$Year_Birth
```

The best customers were identified as being in the top 25% of spenders. Using the quantile fuction we see that they are the customers who spend over $1,045.

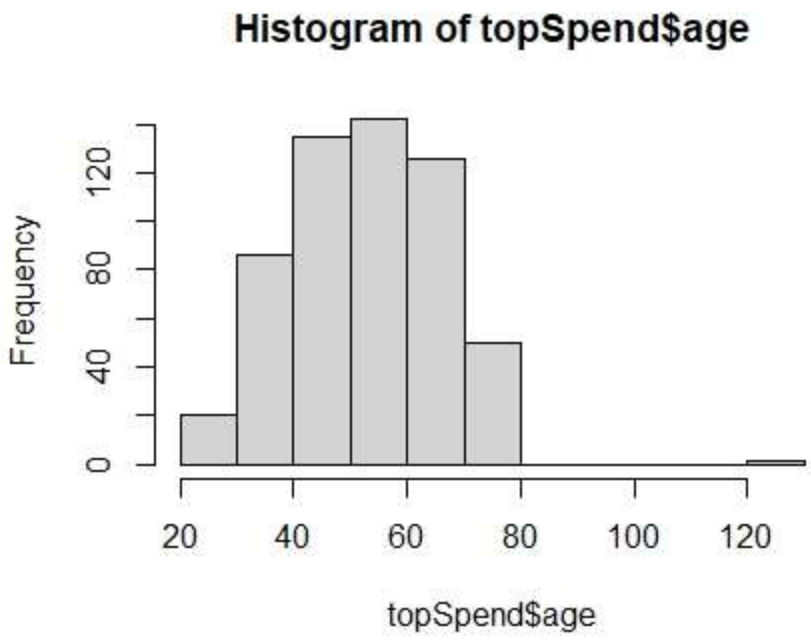
```
hSpend <- quantile(dfGCust$tSpent, prob=.75)
```

To work with only the best customers, we created a data frame of the top 25% of spenders.

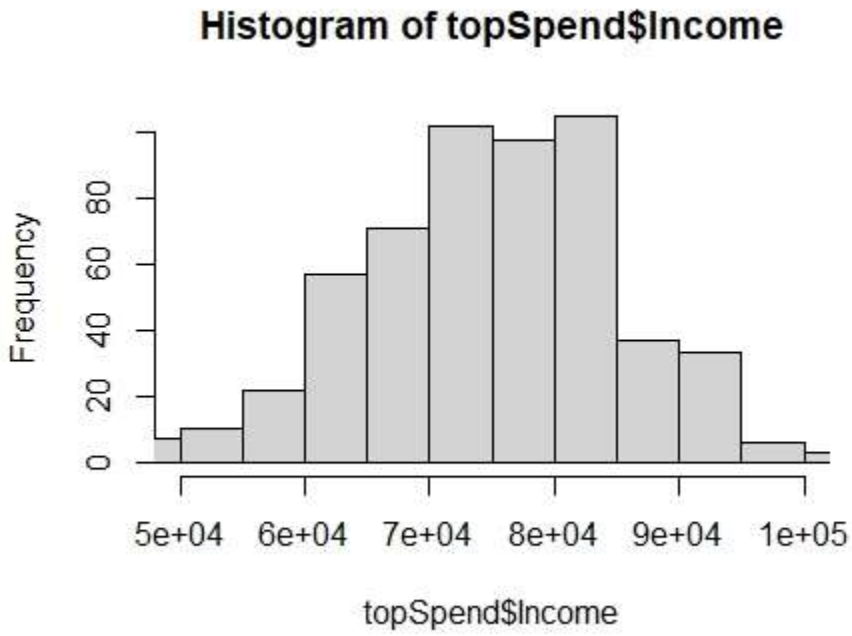
```
topSpend <- dfGCust[which(dfGCust$tSpent >= hSpend[[1]]),]
```

Visualize what age customers are who spend the most and the income of the top spenders. We see that the the top customers fall mostly within the age of 40-70 and have an income of $70,000-85,000.

```
hist(topSpend$age)
```




```
hist(topSpend$Income,breaks = 40, xlim = c(50000,100000))
```



We obtained a count of individuals in the categories Education and Marital Status to find further demographics of the top customers. The largest group of top customers graduated high school

(293), followed by PhD holders (142) and MA (83). For marital status, 201 were married, 142 together, and 127 single.

```
summary(as.factor(topSpend$Education))
summary(as.factor(topSpend$Marital_Status))
```

Finding what variables predict a good customer

The first step for predicting what makes a good customer was to create and run an SVM model by: creating a new column assigning a 1 to customers whose spending is in the top 25% and a 0 for others, turning the numerical values into a factor, and creating a dataframe without the summed total spent column.

```
dfGCust$highSpend <- 8
for(i in 1:nrow(dfGCust)){
if(dfGCust$tSpent[i] >= hSpend[[1]]){
  dfGCust$highSpend[i] <- 1
}else{
  dfGCust$highSpend[i] <- 0
}}

dfGCust$highSpend <- as.factor(dfGCust$highSpend)

dfGCust1 <- dfGCust[,-31]
```

Training and testing data was then created

```
set.seed(111)
trainList<-createDataPartition(y=dfGCust1$highSpend,p=.70,list=FALSE)
trainData<-dfGCust1[trainList,]
testData<-dfGCust1[-trainList,]
```

We ran the SVM model with training data.

```
svm.model<-
train(highSpend~.,data=trainData,method="svmRadial",trControl=trainControl(method="none"),
preProcess=c("center","scale"))
```

We then tested prediction and ran a confusion matrix. This told us that the SVM model was pretty good, with an accuracy of 94% and a low p-value.
Accuracy : 0.9405
95% CI : (0.9198, 0.9571)
No Information Rate : 0.75
P-Value [Acc > NIR] : <2e-16

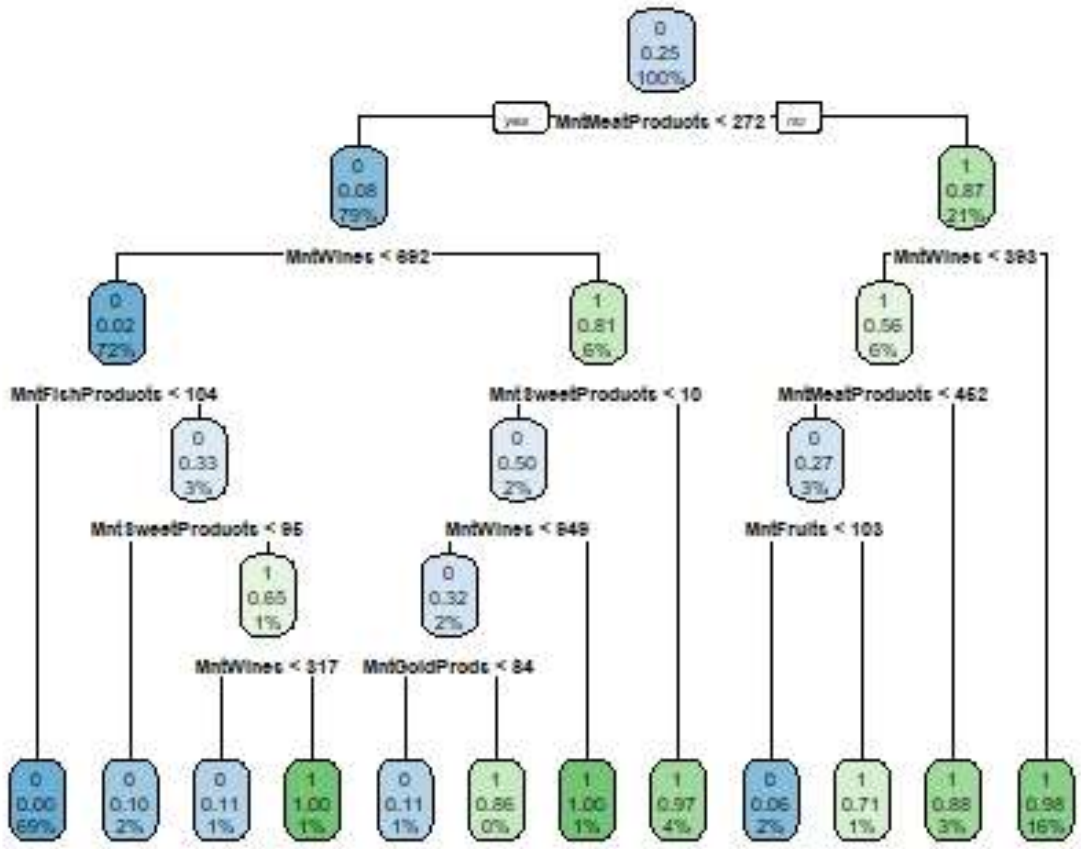
```
svm.model
svmPred<-predict(svm.model,newdata=testData)
confusionMatrix(svmPred,testData$highSpend)
```

We could then view the variables of importance. This shows that the top five in decreasing order were: MntWines, MntMeatProducts, Income, NumCatalogPurchases, and MntFruits. So those customers who more frequently buy wine and meat, and have a higher income, will spend the most at the store.

```
model.rpart<-rpart(highSpend~.,data=trainData,method="class")
varImp(model.rpart)
```

We then created a tree model to visualize the prediction of top customers

```
rpart.plot(model.rpart)
```



To see the details behind this model, we can then train rpart model, test prediction, and output a confusion matrix and variables of importance.
Accuracy : 0.936
95% CI : (0.9148, 0.9533)
No Information Rate : 0.75
P-Value [Acc > NIR] : < 2.2e-16
Variables of importance in descending order matched the last model: MntWines, MntMeatProducts, Income, NumCatalogPurchases, and MntFruits

```
model.rpart2<-train(highSpend~.,data=trainData,method="rpart")
rpartPred<-predict(model.rpart2,newdata=testData)
confusionMatrix(rpartPred,testData$highSpend)
varImp(model.rpart2)
```

Now that we know the variables of importance, we can create a multiple linear regression model to compare fits. This model results in an adjusted r-squared of 0.8183 and variables of significance as Income, MntWines, and MntMeatProducts, matching the other models.

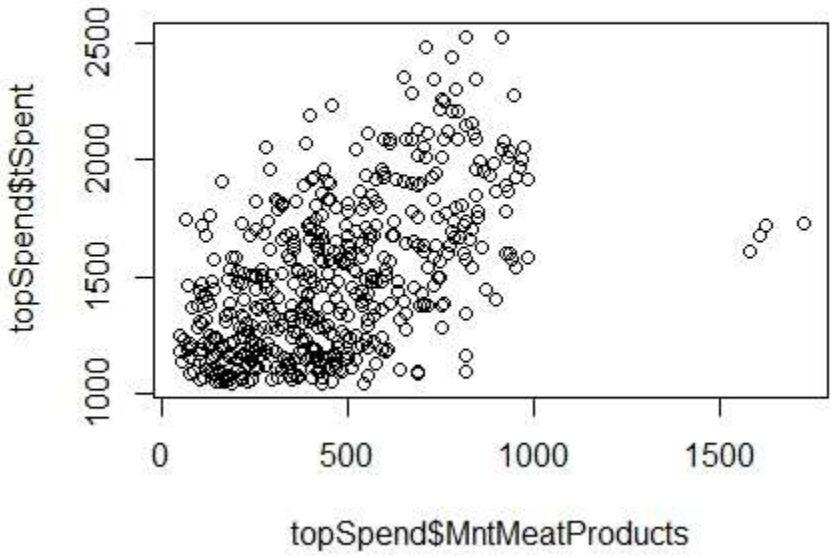
```
topModel<-lm(formula=tSpent~Income+MntWines+MntMeatProducts,data=topSpend)
summary(topModel)$adj.r.squared
summary(topModel)
```

```
Residuals:
    Min      1Q  Median      3Q     Max
-496.46  -97.99   -9.27   80.36  441.35

Coefficients:
                Estimate Std. Error t value Pr(>|t|)
(Intercept)    3.411e+02  3.796e+01   8.986  < 2e-16 ***
Income         1.412e-03  5.028e-04   2.808  0.00516 **
MntWines       8.161e-01  2.079e-02  39.253  < 2e-16 ***
MntMeatProducts 9.639e-01  2.634e-02  36.593  < 2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 140.4 on 556 degrees of freedom
Multiple R-squared:  0.8193,  Adjusted R-squared:  0.8183
F-statistic: 840.3 on 3 and 556 DF,  p-value: < 2.2e-16
```

We can then plot the regression model for a visualization, which shows a positive correlation between top spending customers and their meat purchasing habits.

```
plot(topSpend$MntMeatProducts,topSpend$tSpent)
```

```
abline(topModel)
```

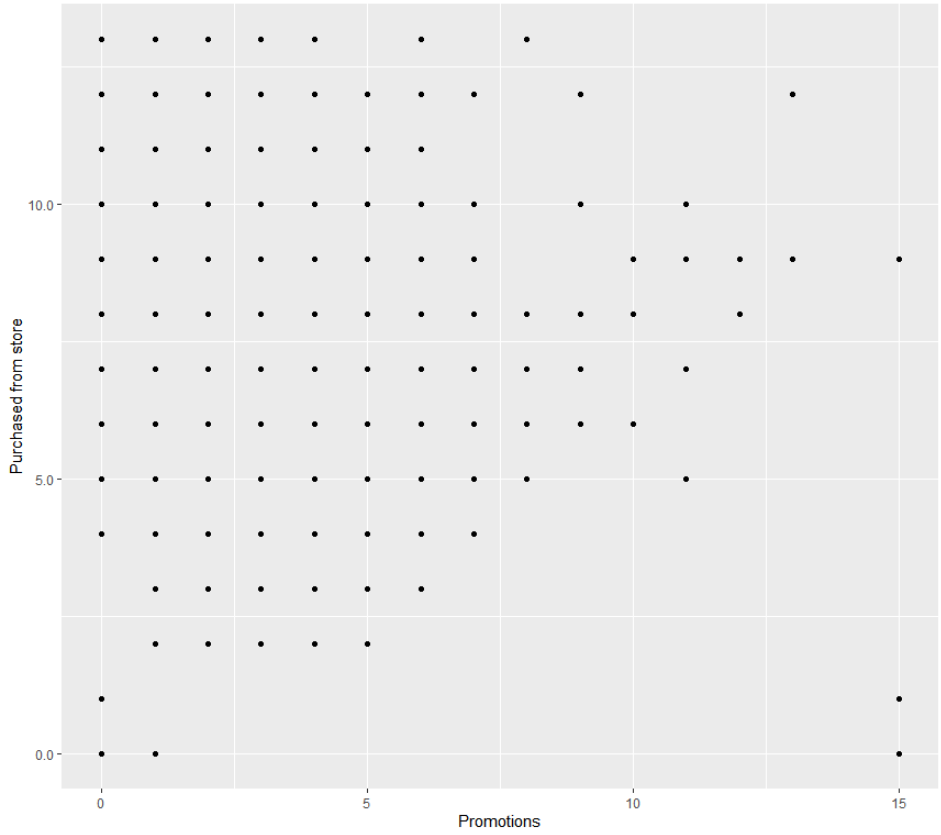


#2 Is there a correlation between promotions and place of purchase (in-store, website, catalog)?

We ran three scatterplots to find out if there's correlation between promotions and place of purchase.

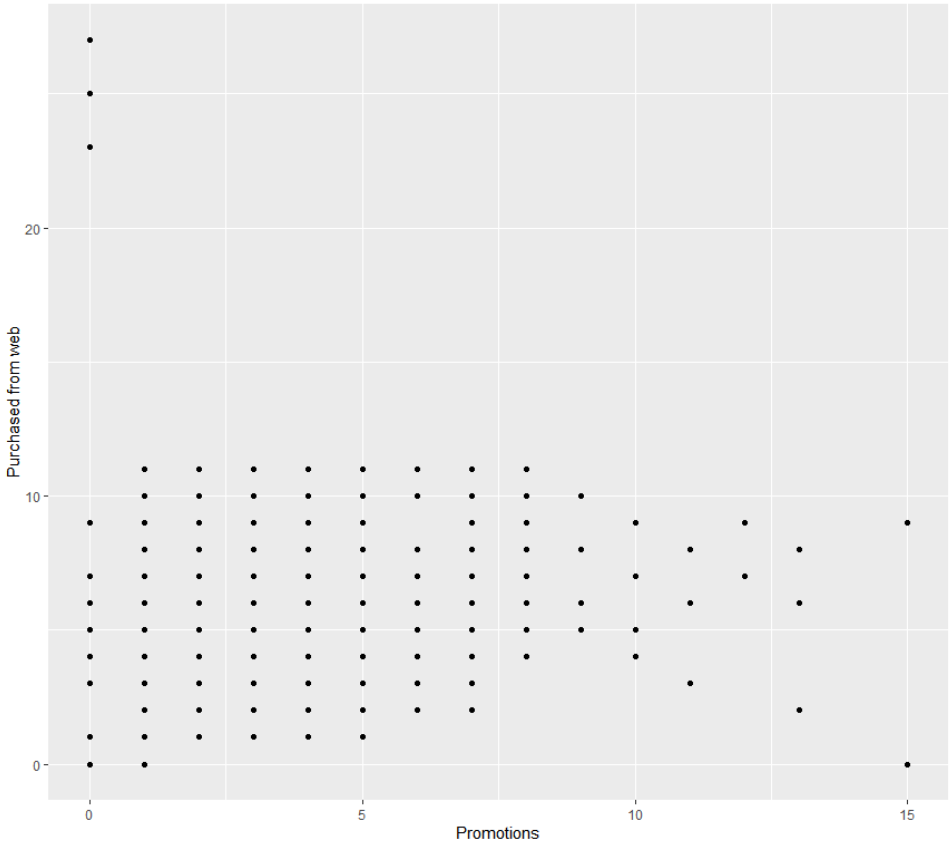
The scatterplot between promotions and in-store purchases shown below indicates that there's a weak positive correlation, with a correlation value of 0.069.

```
cor(df$NumDealsPurchases, df$NumStorePurchases)
0.06887883

ggplot(data=df, aes(x=NumDealsPurchases, y=NumStorePurchases)) + geom_point() +
scale_y_continuous(name="Purchased from store", labels = scales::comma) +
scale_x_continuous(name="Promotions")
```



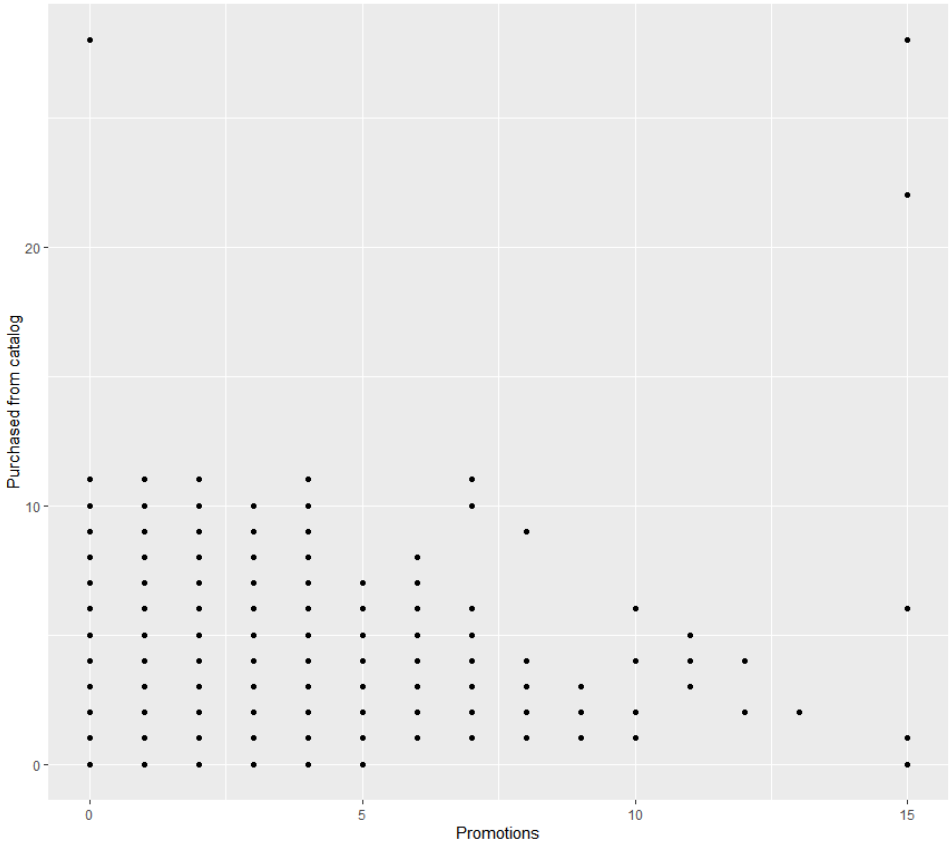
The scatterplot between promotions and website purchases shown below indicates that there's a weak positive correlation, with a correlation value of 0.234.

```
cor(df$NumDealsPurchases, df$NumWebPurchases)
0.2341847

ggplot(data=df, aes(x=NumDealsPurchases, y=NumWebPurchases)) + geom_point() +
scale_y_continuous(name="Purchased from web", labels = scales::comma) +
scale_x_continuous(name="Promotions")
```



The scatterplot between promotions and catalog purchases shown below indicates that there's a weak negative correlation, with a correlation value of –0.009.

```
 cor(df$NumDealsPurchases, df$NumCatalogPurchases)
-0.008617246

ggplot(data=df, aes(x=NumDealsPurchases, y=NumCatalogPurchases)) + geom_point() +
scale_y_continuous(name="Purchased from catalog", labels = scales::comma) +
scale_x_continuous(name="Promotions")
```



#3 Can we predict if a customer will accept a promotion offer and what variables will drive it?

We first ran a multiple regression model, including the inventory of primary data attributes, against the population of customers who had historically accepted an offer on either the first promotion or the second promotion received. To do this, we created two separate dataframes, one for the customers who accepted the promotion on the first offer and then a second dataframe for the customers who accepted the offer on the second promotion cycle. We then merged the two separate dataframes together; this gave us the population of customers (174 in total) who accepted on either the first or second promotion cycle. Refer to code below:

```
dfPromo2 <- df[df$AcceptedCmp2==1, ]
dfPromo1 <- df[df$AcceptedCmp1==1, ]

dfPromos <- rbind(dfPromo1, dfPromo2)
```

Next, we built the multiple regression equation based on the primary data attributes known about the group of customers. Refer to code below:

```
PromoAcceptv1 <-lm(formula = AcceptedCmp1 ~ Year_Birth + Education + Marital_Status +
                   Income + Kidhome + Teenhome + Dt_Customer + Recency + MntWines +
                   MntFruits + MntMeatProducts + MntFishProducts + MntSweetProducts +
```

```
            MntGoldProds + NumDealsPurchases + NumWebPurchases +
NumCatalogPurchases +
            NumStorePurchases, data = dfPromos)
```

The result of this formula left us with a low adjust R-squared value (0.2087) and very few low p-values.

```
Call:
lm(formula = AcceptedCmp1 ~ Year_Birth + Education + Marital_Status +
    Income + Kidhome + Teenhome + Dt_Customer + Recency + MntWines +
    MntFruits + MntMeatProducts + MntFishProducts + MntSweetProducts +
    MntGoldProds + NumDealsPurchases + NumWebPurchases + NumCatalogPurchases +
    NumStorePurchases, data = dfPromos)

Residuals:
     Min       1Q   Median       3Q      Max
-0.86787 -0.06422  0.04981  0.14844  0.40908

Coefficients:
                          Estimate Std. Error t value Pr(>|t|)
(Intercept)             -1.825e+00  3.713e+00  -0.492   0.6237
Year_Birth              -4.062e-04  1.706e-03  -0.238   0.8121
EducationGraduation      9.620e-02  8.081e-02   1.190   0.2358
EducationMaster          9.547e-02  9.834e-02   0.971   0.3332
EducationPhD             6.983e-02  8.955e-02   0.780   0.4368
Marital_StatusDivorced  -1.856e-01  2.901e-01  -0.640   0.5233
Marital_StatusMarried   -4.550e-02  2.841e-01  -0.160   0.8730
Marital_StatusSingle    -6.816e-02  2.881e-01  -0.237   0.8133
Marital_StatusTogether  -2.372e-01  2.854e-01  -0.831   0.4072
Marital_StatusWidow     -1.452e-01  3.063e-01  -0.474   0.6361
Income                   5.686e-06  2.257e-06   2.520   0.0128 *
Kidhome                  6.439e-02  7.872e-02   0.818   0.4147
Teenhome                -3.152e-02  6.621e-02  -0.476   0.6347
Dt_Customer              1.825e-04  1.015e-04   1.798   0.0742 .
Recency                 -3.557e-04  7.682e-04  -0.463   0.6440
MntWines                -7.266e-05  6.234e-05  -1.165   0.2457
MntFruits                1.097e-04  4.987e-04   0.220   0.8261
MntMeatProducts          1.054e-04  9.700e-05   1.087   0.2790
MntFishProducts          4.987e-04  3.682e-04   1.354   0.1777
MntSweetProducts         7.655e-04  4.891e-04   1.565   0.1197
MntGoldProds            -6.585e-04  3.782e-04  -1.741   0.0837 .
NumDealsPurchases        2.633e-02  2.431e-02   1.083   0.2806
NumWebPurchases          1.631e-02  9.500e-03   1.716   0.0882 .
NumCatalogPurchases      2.133e-02  1.014e-02   2.103   0.0371 *
NumStorePurchases       -5.980e-03  7.911e-03  -0.756   0.4509
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.2649 on 149 degrees of freedom
Multiple R-squared:  0.3184,    Adjusted R-squared:  0.2087
F-statistic: 2.901 on 24 and 149 DF,  p-value: 4.392e-05
```

In an attempt to refine the equation and increase our adjusted R-squared value, we ran a second regression equation including only those variables which are determined to have low p-values (<0.1). Refer to code below:

```
PromoAcceptv2 <-lm(formula = AcceptedCmp1~Income + Dt_Customer + NumWebPurchases
+ NumCatalogPurchases,
            data=dfPromos)
```

As a result, our updated equation produces an adjusted R-square value that is even lower (0.1397).

```
Call:
lm(formula = AcceptedCmp1 ~ Income + Dt_Customer + NumWebPurchases +
    NumCatalogPurchases, data = dfPromos)

Residuals:
     Min       1Q   Median       3Q      Max
-0.95414 -0.01844  0.07428  0.12066  0.44312

Coefficients:
                     Estimate Std. Error t value Pr(>|t|)
(Intercept)         -2.868e+00  1.542e+00  -1.859 0.064737 .
Income               6.623e-06  1.817e-06   3.644 0.000356 ***
Dt_Customer          1.941e-04  9.557e-05   2.031 0.043803 *
NumWebPurchases      1.278e-02  9.181e-03   1.392 0.165601
NumCatalogPurchases  1.636e-02  9.300e-03   1.760 0.080287 .
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.2762 on 169 degrees of freedom
Multiple R-squared:  0.1596,    Adjusted R-squared:  0.1397
F-statistic: 8.022 on 4 and 169 DF,  p-value: 6.039e-06
```

To build a stronger equation with a higher likely hood for prediction, we decided to revisit our source data. Instead of including customers who accepted promotions on a second cycle, we decided to only focus on the customers who accept promotions the first time. We re-ran our initial model, with all data attributes, against the dataframe with just the population of customers who accepted the first time (144 customers in total).

```
PromoAcceptv1 <-lm(formula = AcceptedCmp1 ~ Year_Birth + Education + Marital_Status +
                Income + Kidhome + Teenhome + Dt_Customer + Recency + MntWines +
                MntFruits + MntMeatProducts + MntFishProducts + MntSweetProducts +
                MntGoldProds + NumDealsPurchases + NumWebPurchases +
NumCatalogPurchases +
                NumStorePurchases, data = dfPromo1)

summary(PromoAcceptv1)
```

Based on the summary results below, the adjusted R-squared value is 0.3991. There are also a greater number of variables identified with even lower p-values.

```
Call:
lm(formula = AcceptedCmp1 ~ Year_Birth + Education + Marital_Status +
    Income + Kidhome + Teenhome + Dt_Customer + Recency + MntWines +
    MntFruits + MntMeatProducts + MntFishProducts + MntSweetProducts +
    MntGoldProds + NumDealsPurchases + NumWebPurchases + NumCatalogPurchases +
    NumStorePurchases, data = dfPromo1)

Residuals:
      Min        1Q    Median        3Q       Max
-2.875e-15 -6.559e-16 -5.280e-17  4.530e-16  1.498e-14

Coefficients:
                       Estimate Std. Error    t value Pr(>|t|)
(Intercept)           1.000e+00  2.483e-14  4.028e+13  < 2e-16 ***
Year_Birth            1.964e-17  1.140e-17  1.724e+00  0.08739 .
EducationGraduation  -1.538e-15  5.358e-16 -2.870e+00  0.00485 **
EducationMaster      -1.310e-15  6.441e-16 -2.034e+00  0.04419 *
EducationPhD         -1.318e-15  6.001e-16 -2.196e+00  0.03005 *
Marital_StatusDivorced  6.468e-16  1.818e-15  3.560e-01  0.72269
Marital_StatusMarried   4.097e-16  1.769e-15  2.320e-01  0.81727
Marital_StatusSingle    1.203e-15  1.798e-15  6.690e-01  0.50465
Marital_StatusTogether  3.168e-16  1.782e-15  1.780e-01  0.85922
Marital_StatusWidow     1.155e-15  1.937e-15  5.960e-01  0.55232
Income                  1.231e-20  1.626e-20  7.570e-01  0.45054
Kidhome                 2.709e-16  5.514e-16  4.910e-01  0.62412
Teenhome                1.474e-15  4.878e-16  3.022e+00  0.00308 **
Dt_Customer             1.206e-18  6.839e-19  1.763e+00  0.08050 .
Recency                -1.177e-18  5.427e-18 -2.170e-01  0.82861
MntWines               -3.111e-19  4.645e-19 -6.700e-01  0.50439
MntFruits              -5.379e-19  3.142e-18 -1.710e-01  0.86437
MntMeatProducts        -2.723e-19  6.440e-19 -4.230e-01  0.67315
MntFishProducts         7.120e-18  2.361e-18  3.016e+00  0.00314 **
MntSweetProducts       -5.520e-18  3.155e-18 -1.750e+00  0.08277 .
MntGoldProds           -3.291e-18  2.524e-18 -1.304e+00  0.19483
NumDealsPurchases      -2.811e-16  1.958e-16 -1.436e+00  0.15374
NumWebPurchases         1.219e-16  7.198e-17  1.694e+00  0.09294 .
NumCatalogPurchases     1.233e-16  6.929e-17  1.780e+00  0.07759 .
NumStorePurchases       3.271e-17  5.430e-17  6.020e-01  0.54811
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 1.638e-15 on 119 degrees of freedom
Multiple R-squared:    0.5,     Adjusted R-squared:  0.3991
F-statistic: 4.958 on 24 and 119 DF,  p-value: 1.719e-09
```

To further strengthen our new equation, we re-ran the regression model with the variables that have a low p-value (< 0.1, indicating a relationship):

- Year_Birth
- Education
- Teenhome
- Dt_Customer
- MntFishProducts
- MntSweetProds
- NumWebPurchases
- NumCatalogPurchases

Refer below for code:

```
PromoAcceptv2 <-lm(formula = AcceptedCmp1~Year_Birth +
                   Education +
                   Teenhome +
                   Dt_Customer +
                   MntFishProducts +
                   MntSweetProducts +
```

```
NumWebPurchases +
NumCatalogPurchases,
data=dfPromo1)
```

The summary output for the updated equation gave us an adjust R-squared value or 0.4606, all p-values are low, and the sum of the residuals is very close to zero.

```
Call:
lm(formula = AcceptedCmp1 ~ Year_Birth + Education + Teenhome +
    Dt_Customer + MntFishProducts + MntSweetProducts + NumWebPurchases +
    NumCatalogPurchases, data = dfPromo1)

Residuals:
       Min         1Q     Median         3Q        Max
-2.766e-15 -5.500e-16 -9.770e-17  5.224e-16  1.671e-14

Coefficients:
                     Estimate Std. Error    t value Pr(>|t|)
(Intercept)          1.000e+00  2.247e-14  4.450e+13  < 2e-16 ***
Year_Birth           2.116e-17  1.059e-17  1.999e+00 0.047669 *
EducationGraduation -1.694e-15  4.939e-16 -3.430e+00 0.000804 ***
EducationMaster     -1.350e-15  5.950e-16 -2.269e+00 0.024872 *
EducationPhD        -1.277e-15  5.520e-16 -2.313e+00 0.022265 *
Teenhome             8.641e-16  3.617e-16  2.389e+00 0.018305 *
Dt_Customer          1.235e-18  6.541e-19  1.888e+00 0.061177 .
MntFishProducts      5.382e-18  2.230e-18  2.413e+00 0.017183 *
MntSweetProducts    -3.479e-18  2.793e-18 -1.246e+00 0.215127
NumWebPurchases      1.260e-16  6.745e-17  1.868e+00 0.064025 .
NumCatalogPurchases  1.274e-16  6.054e-17  2.105e+00 0.037211 *
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 1.636e-15 on 133 degrees of freedom
Multiple R-squared:  0.4983,    Adjusted R-squared:  0.4606
F-statistic: 13.21 on 10 and 133 DF,  p-value: 7.407e-16
```

Based on this data and the multiple tests that we have conducted, we conclude that this has provided the best model for predicting what if a customer will likely accept a promotion on the first offer. When populating the equation with the data captured for each customer, against the identified coefficients, our model will predict if a customer will accept the promotion with 46% variability.

Some interesting points to note, the high level of education the customer has, the least likely they are to accept the promotion (as evidenced by the negative coefficients for Graduation, Master and PhD). The number of teens in the home and the amount spent on fish has a positive correlation with whether or not the promotion is accepted on the first offer (as evidenced by the large positive coefficients).

#4 How much of a factor is place of purchase on customer spending?

We extracted the necessary values from place of purchase, website, store, catalog and ran another series of multiple regression models to find the likelihood that a customer's place of purchase would affect how much customers are spending. At times customers purchase from the website, catalog or in store, does that affect customer spend?

In order to better understand the data, we can begin by visualizing the data.

```
#capture total customer spending
CustSpend <- df
CustSpend$tSpent <- rowSums(CustSpend[c("MntWines", "MntFruits", "MntMeatProducts",
"MntFishProducts", "MntSweetProducts", "MntGoldProds")])
CustSpend$tSpent

#plot customer spending by place of purchase
storepurchasePlot <- ggplot(data=CustSpend, aes(x=tSpent, y=NumStorePurchases)) +
geom_point() + scale_y_continuous(name="Purchased from store", labels = scales::comma)+
scale_x_continuous(name="Spending")
storepurchasePlot

webpurchasePlot <- ggplot(data=CustSpend, aes(x=tSpent, y=NumWebPurchases)) +
geom_point() + scale_y_continuous(name="Purchased from web", labels = scales::comma) +
scale_x_continuous(name="Spending")
webpurchasePlot

catalogpurchasePlot <- ggplot(data=CustSpend, aes(x=tSpent, y=NumCatalogPurchases)) +
geom_point() + scale_y_continuous(name="Purchased from catalog", labels = scales::comma) +
scale_x_continuous(name="Spending")
catalogpurchasePlot

#calc average of number of purchases by place
mean(CustSpend$NumWebPurchases)
mean(CustSpend$NumCatalogPurchases)
mean(CustSpend$NumStorePurchases)

#averages: Web 4.084821 | Catalog 2.662054 | Store 5.790179

str(CustSpend)
   #plot data shows different resutls, need to view data in different plot
cor(CustSpend$tSpent, df$NumWebPurchases)
cor(CustSpend$tSpent, df$NumCatalogPurchases)
cor(CustSpend$tSpent, df$NumStorePurchases)
```

Bar charts

```
#create bar charts for datra visualization of customer spending
p<-ggplot(data=CustSpend, aes(x=NumWebPurchases, y=tSpent)) +
  geom_bar(stat="identity")
p
```

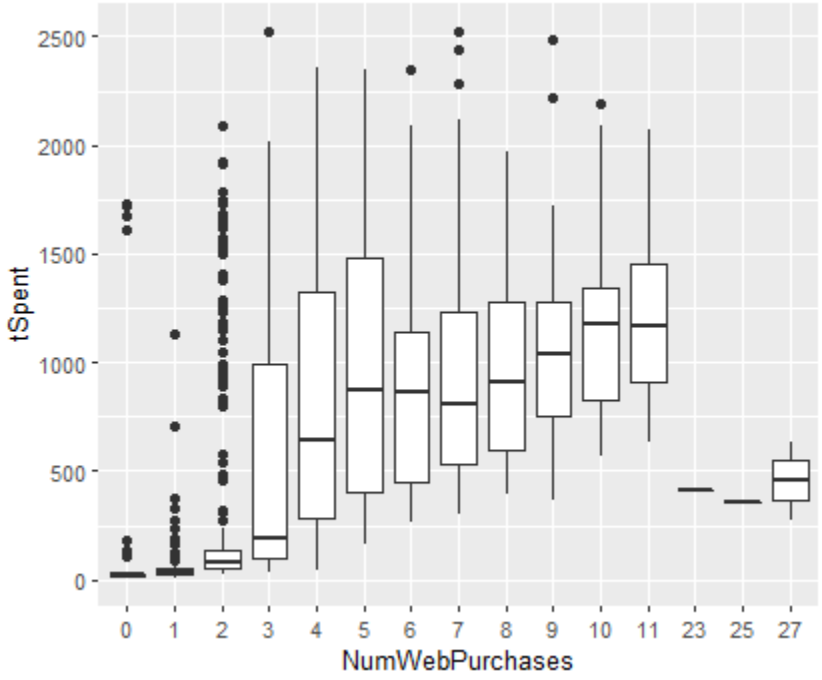
```
q<-ggplot(data=CustSpend, aes(x=NumStorePurchases, y=tSpent)) +
  geom_bar(stat="identity")
q

r<-ggplot(data=CustSpend, aes(x=NumCatalogPurchases, y=tSpent)) +
  geom_bar(stat="identity")
r

#comapring different place of purchas to spending

s <- ggplot(data=CustSpend, aes(x=tSpent, y=NumWebPurchases, fill=NumStorePurchases, )) +
  geom_bar(stat="identity", position=position_dodge())
s
```

Boxplots

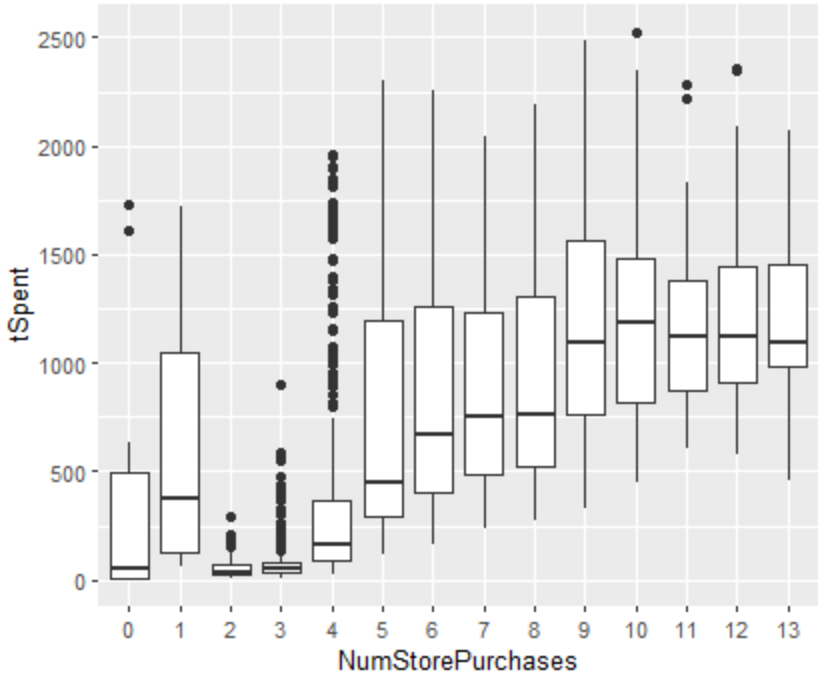
```
#Originally used hist in naming for histogram but wanted to use boxplots instead.
#create dataframe capturing csutomer spend

histdf <- data.frame()
histdf <- CustSpend

histdf$NumWebPurchases <- as.factor(histdf$NumWebPurchases)
histdf$NumStorePurchases <- as.factor(histdf$NumStorePurchases)
histdf$NumCatalogPurchases <- as.factor(histdf$NumCatalogPurchases)

#assign variable for ggplot to recal later
w <- ggplot(histdf, aes(x=NumWebPurchases, y=tSpent)) + geom_boxplot()
w + stat_summary(fun=mean, geom = "point", shape=7, size =4)
w

t <- ggplot(histdf, aes(x=NumStorePurchases, y=tSpent)) + geom_boxplot()
t + stat_summary(fun=mean, geom = "point", shape=7, size =4)
t

c <- ggplot(histdf, aes(x=NumCatalogPurchases, y=tSpent)) + geom_boxplot()
c + stat_summary(fun=mean, geom = "point", shape=7, size =4)
c

 #view plots
w
t
c
```

Creating a boxplot for each place of purchase by spending helped map the number of purchases by place and reflect spending as the number of items per place increases.

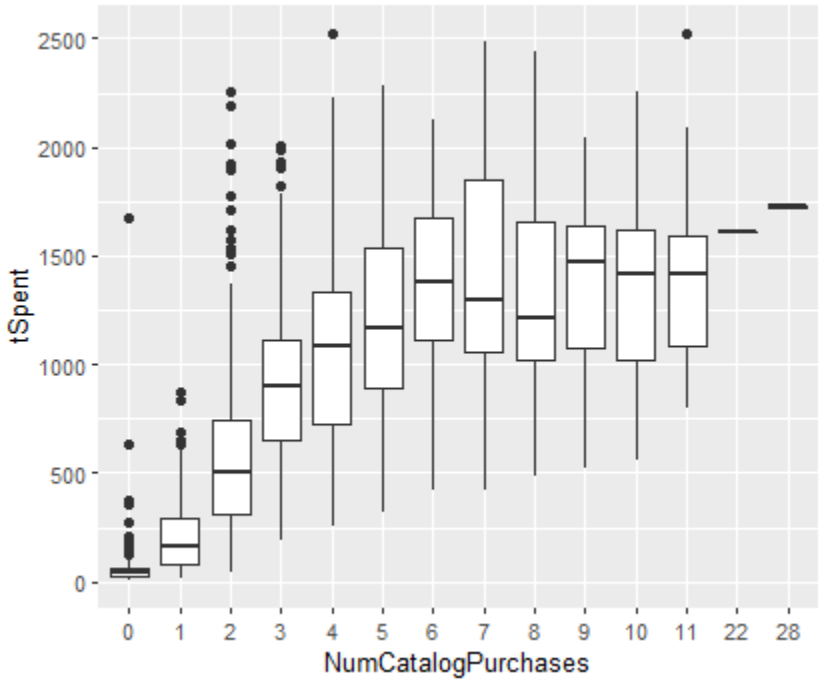
Web purchases: As the number of purchases increase there seems to be a drop in spending and a higher concentration data as the count of purchases increases pasts 8, meaning as customer purchase more items through the website spending drops. Therefore, it could be argued that customers spend less overall on web purchases as they order more items.



Store purchases: As the number of purchases increases there seems to be an increase in spending and a higher concentration of the data, meaning as customer purchase more items through the website spending increases with a max at 9 items and drops off slightly. We may be able to assume that customers spend more overall on store purchases as they order more items, and more overall than web purchases for 6 purchases but lower for smaller orders with less purchases.



Catalog purchases: Lastly, the number catalog purchases increases with spending under 6 purchases and spending beings to drop and plateau with more variation from the mean up to a certain point, 12 items, with yet more items up to 22 and 28.



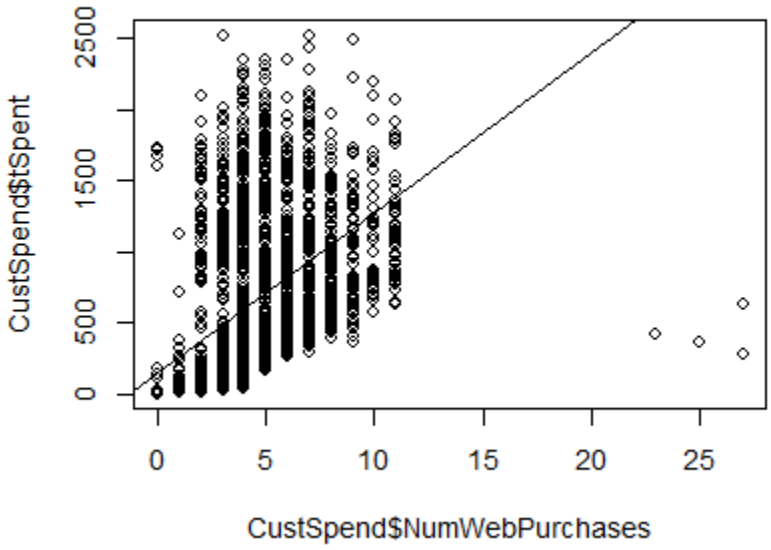
With this in mind we wanted to test each place of purchase in a linear model.

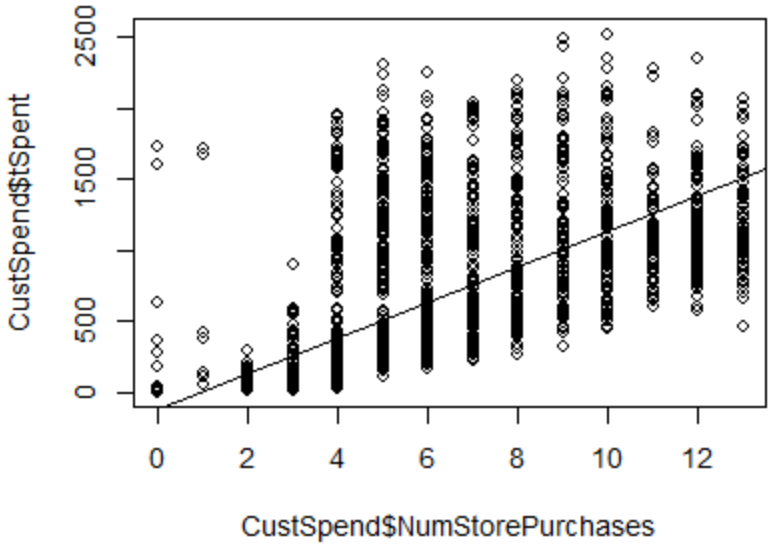
#Creates a linear model

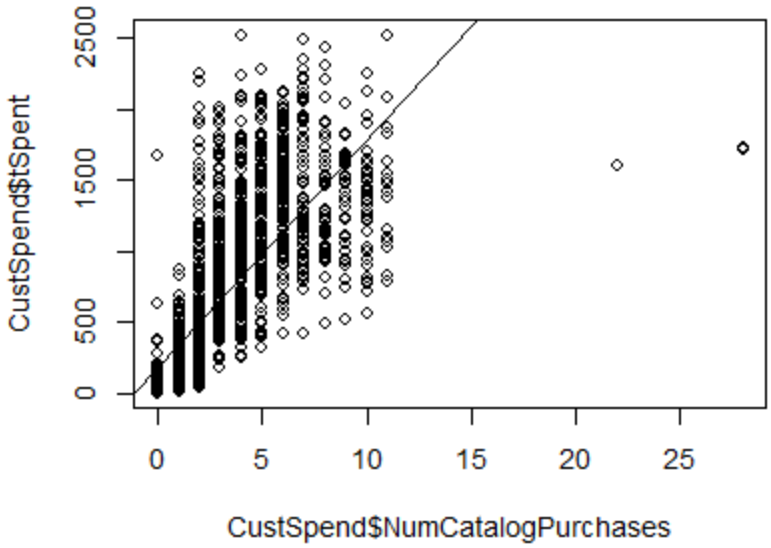
```
weblm <- lm(tSpent ~ NumWebPurchases,data = CustSpend)

#Prints the model results
weblm
summary(weblm)

storelm <- lm(tSpent ~ NumStorePurchases,data = CustSpend)
storelm
summary(storelm)

cataloglm <- lm(tSpent ~ NumCatalogPurchases,data = CustSpend)
cataloglm
summary(cataloglm)
```

Linear model summary:

- The adjusted R-squared values for web purchases was 0.26 with a p-value of < 2.2e-16
- The store purchases returned an adjusted R-squared value of 0.45 and a p-value of 2.2e-16 as well.
- The catalog purchases reported an adjusted R-squared value of 0.60 and a p-value of 2.2e-16.







In these charts we can also see in ab-line which shows the line of best fit for the data.
In this case, catalog purchases have a positive slop but the values drop off after 10 purchases. Similarly, web purchases drop off at 10 purchases while store purchases continue even after 12 purchases.

Summary assessment and actionable steps

Summary Assessment

Our findings show that our best customers are between the ages of 40 and 70, and have an income of $70,000-85,000. Top customers can be identified by those with upper end incomes who more frequently buy wine and meat. The SVM prediction model determining these variables has a 94% accuracy and low p-value, so will serve as a good prediction for identifying others who will be top customers. A linear regression model also supports these findings.

Through multiple scatter plot analyses, we determined that there is no strong correlation between how a customer shops (in-store, website, catalog) and whether or not they will take advantage of promotions. The correlation between in-store purchases and promotions is a weak positive correlation with a correlation value of 0.069. The correlation between website purchases and promotions is a weak positive correlation with a correlation value of 0.234. And the correlation between catalog purchases and promotions is a weak negative correlation with a correlation value of –0.009.

Multiple linear regression analysis shows that we can predict with 46% variance whether or not a customer will accept the first run of a promotion. The variables that predict this behavior are: year of birth, education, how many teens are in the home, date the customer joined our program, fish products purchased, sweets purchased, web purchases, and catalog purchases.

We used linear modeling and boxplot analysis to determine that customers are likely to spend more in-store than through other modes of purchase. Customers spend less overall during web purchases as they purchase more items. During in-store transactions, customers spend more as they buy more items as compared to web purchases, and overall they are likely to spend more if they are in the store. For catalog purchases, customers spend more on items if they purchase under 6 items and begin to spend less on items and levels off as they purchase more items.

Actionable steps

Since we now know who the highest spending customers are, those with higher incomes who purchase wine and meat, we don't need to focus promotions on these customers since they are already buying at a higher rate. Instead, we can focus promotions on customers with lower incomes who make purchases such as sweets and vegetables. We do however, want to make sure that we keep wine and meats in stock in order to keep our top customers coming to the store.

Since we know that there isn't a strong correlation between promotions and how customers make purchases, we do not need to spend time focusing promotions on people by mode of purchase, We do know, however, that people with a low income who have teenagers and purchase fish respond well to promotions, so we should continue to focus promotions on that demographic.

We also need to focus on getting people into the store. Those who come in to the store purchase more items and spend more on those items. In the future, we should find ways to use promotions that incentivize in-store purchases. This could take the form of extra savings when something is purchased in-store.