Black Friday Project

David Contento

December 4, 2018

In this report I will be trying to construct a model that accurately predicts the amount (in dollars) of purchases a customer will make based on their individual characteristics. I will begin the process by first looking at the structure of the data and performing any cleaning and pre-processing. The data used in this project was gathered from Kaggle, and the purpose of this project was to submit the model as part of a contest.

First I check the structure of our data and determine whether or not there are any NAs in the data.

```r
str(data)
```

```
## 'data.frame':    537577 obs. of  12 variables:
##  $ User_ID                   : int  1000001 1000001 1000001 1000001 1000002 1000003 1000004 1000004 1
##  $ Product_ID                : Factor w/ 3623 levels "P00000142","P00000242",..: 671 2375 851 827 273
##  $ Gender                    : Factor w/ 2 levels "F","M": 1 1 1 1 2 2 2 2 2 2 ...
##  $ Age                       : Factor w/ 7 levels "0-17","18-25",..: 1 1 1 1 7 3 5 5 5 3 ...
##  $ Occupation                : int  10 10 10 10 16 15 7 7 7 20 ...
##  $ City_Category             : Factor w/ 3 levels "A","B","C": 1 1 1 1 3 1 2 2 2 1 ...
##  $ Stay_In_Current_City_Years: Factor w/ 5 levels "0","1","2","3",..: 3 3 3 3 5 4 3 3 3 2 ...
##  $ Marital_Status            : int  0 0 0 0 0 0 1 1 1 1 ...
##  $ Product_Category_1        : int  3 1 12 12 8 1 1 1 1 8 ...
##  $ Product_Category_2        : int  NA 6 NA 14 NA 2 8 15 16 NA ...
##  $ Product_Category_3        : int  NA 14 NA NA NA NA 17 NA NA NA ...
##  $ Purchase                  : int  8370 15200 1422 1057 7969 15227 19215 15854 15686 7871 ...
```

```r
#The data is clearly cross-sectional
#checking which variables have NA's
colSums(is.na(data))
```

```
##                    User_ID                 Product_ID
##                          0                          0
##                     Gender                        Age
##                          0                          0
##                 Occupation              City_Category
##                          0                          0
## Stay_In_Current_City_Years             Marital_Status
##                          0                          0
##         Product_Category_1         Product_Category_2
##                          0                     166986
##         Product_Category_3                   Purchase
##                     373299                          0
```

Since there are NAs in Product_Category_2 and Product_Category_3 i combined those into a dummy variable called "multi". Multi takes on a value of 1 if the product belongs to multiple categories and a value of 0 if the product belongs to only one category. Using this method i do not need to include all three dummy variables.

```r
#combining product category 2 and 3 variable into dummy:
#with 1 if in more than one product category and 0 if not
data$multi = 1
data[is.na(data$Product_Category_2), "multi"] = 0
```

```r
#removing product category 2 and 3 variables
data <- data[, -c(10:11)]
str(data)
```

```
## 'data.frame':    537577 obs. of  11 variables:
##  $ User_ID                   : int  1000001 1000001 1000001 1000001 1000002 1000003 1000004 1000004 1
##  $ Product_ID                : Factor w/ 3623 levels "P00000142","P00000242",..: 671 2375 851 827 273
##  $ Gender                    : Factor w/ 2 levels "F","M": 1 1 1 1 2 2 2 2 2 2 ...
##  $ Age                       : Factor w/ 7 levels "0-17","18-25",..: 1 1 1 1 7 3 5 5 5 3 ...
##  $ Occupation                : int  10 10 10 10 16 15 7 7 7 20 ...
##  $ City_Category             : Factor w/ 3 levels "A","B","C": 1 1 1 1 3 1 2 2 2 1 ...
##  $ Stay_In_Current_City_Years: Factor w/ 5 levels "0","1","2","3",..: 3 3 3 3 5 4 3 3 3 2 ...
##  $ Marital_Status            : int  0 0 0 0 0 0 1 1 1 1 ...
##  $ Product_Category_1        : int  3 1 12 12 8 1 1 1 1 8 ...
##  $ Purchase                  : int  8370 15200 1422 1057 7969 15227 19215 15854 15686 7871 ...
##  $ multi                     : num  0 1 0 1 0 1 1 1 1 0 ...
```

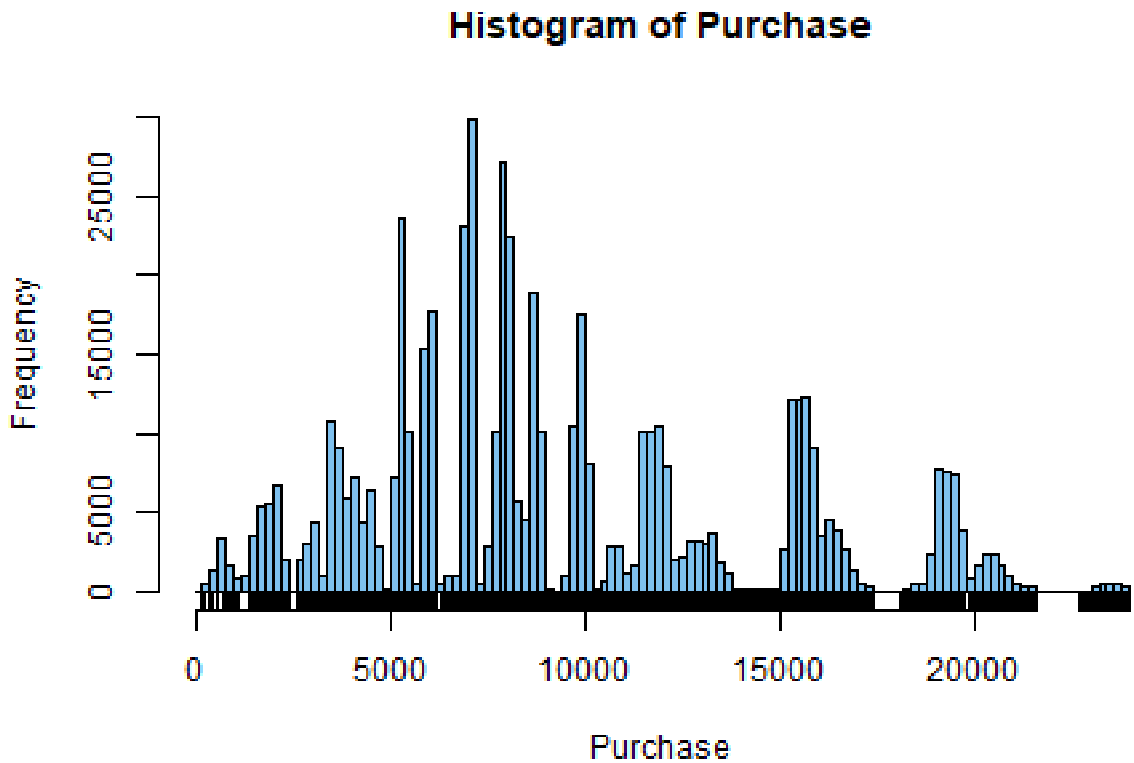
```r
colSums(is.na(data))
```

```
##                    User_ID                 Product_ID
##                          0                          0
##                     Gender                        Age
##                          0                          0
##                 Occupation              City_Category
##                          0                          0
## Stay_In_Current_City_Years             Marital_Status
##                          0                          0
##         Product_Category_1                   Purchase
##                          0                          0
##                      multi
##                          0
```

Now I will move on to the descriptive analysis. Notice that most of our variables are categorical and therefore I can only make histograms for Purchase and Occupation. It is important to note that purchases (total dollar amount of a persons transaction) is the dependent variable for this report.

I now conduct some data analysis on the variables to see if they are skewed, have outliers, or need to be transformed. I look at the Quantile-Quantile plots for Purchase in order to determine its normality.

```r
#Histogram of Purchases
par(mfrow=c(1,1))
hist(Purchase, breaks = "FD", col = "skyblue2")
rug(Purchase)
```

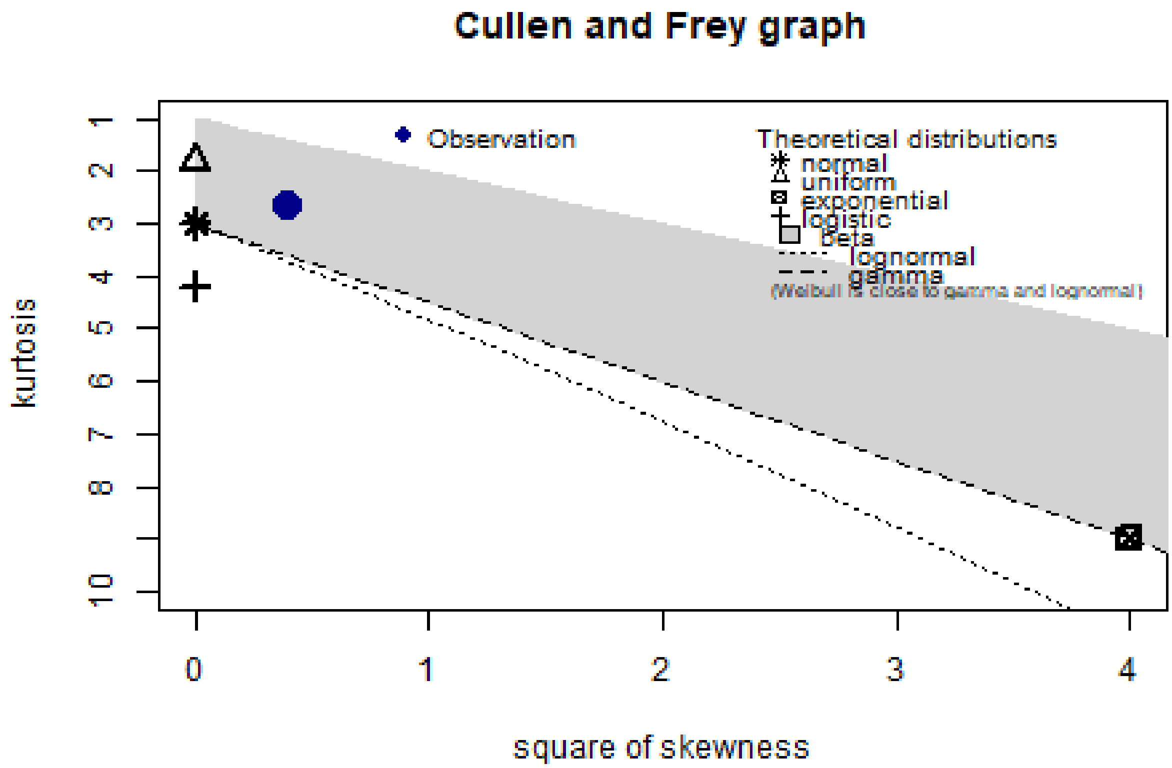



```r
S(Purchase)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##     185    5866    8062    9334   12073   23961
```

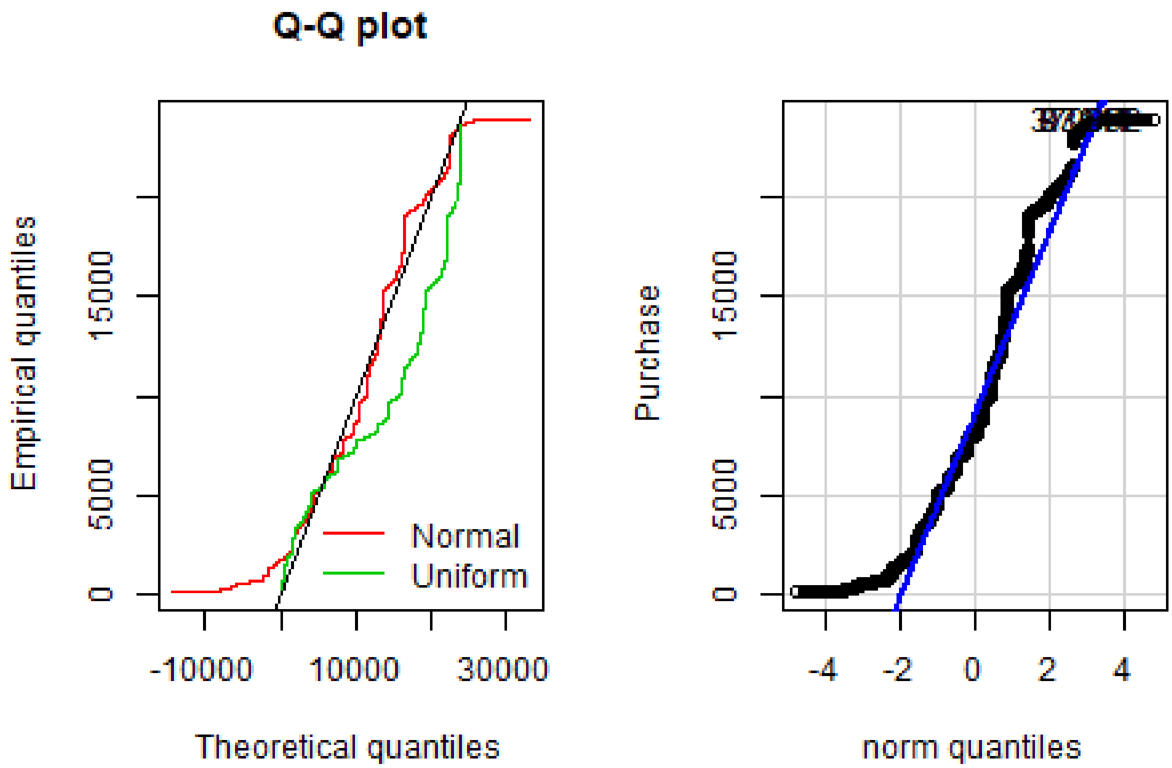
```r
par(mfrow=c(1,1))

#Cullen-Frey graph of Purchases
descdist(Purchase)
```




```
## summary statistics
## ------
## min:  185   max:  23961
## median:  8062
## mean:  9333.86
## estimated sd:  4981.022
## estimated skewness:  0.6242797
## estimated kurtosis:  2.656879
```

```
#Fitting Distributions to Purchases
fit.norm = fitdist(as.numeric(Purchase), "norm")
fit.unif = fitdist(as.numeric(Purchase),"unif")
plot.legend = c("Normal", "Uniform")
par(mfrow=c(1,2))
qqcomp(list(fit.norm, fit.unif), legendtext = plot.legend)
qqPlot(~ Purchase, data = data, id = list(n=3))
```

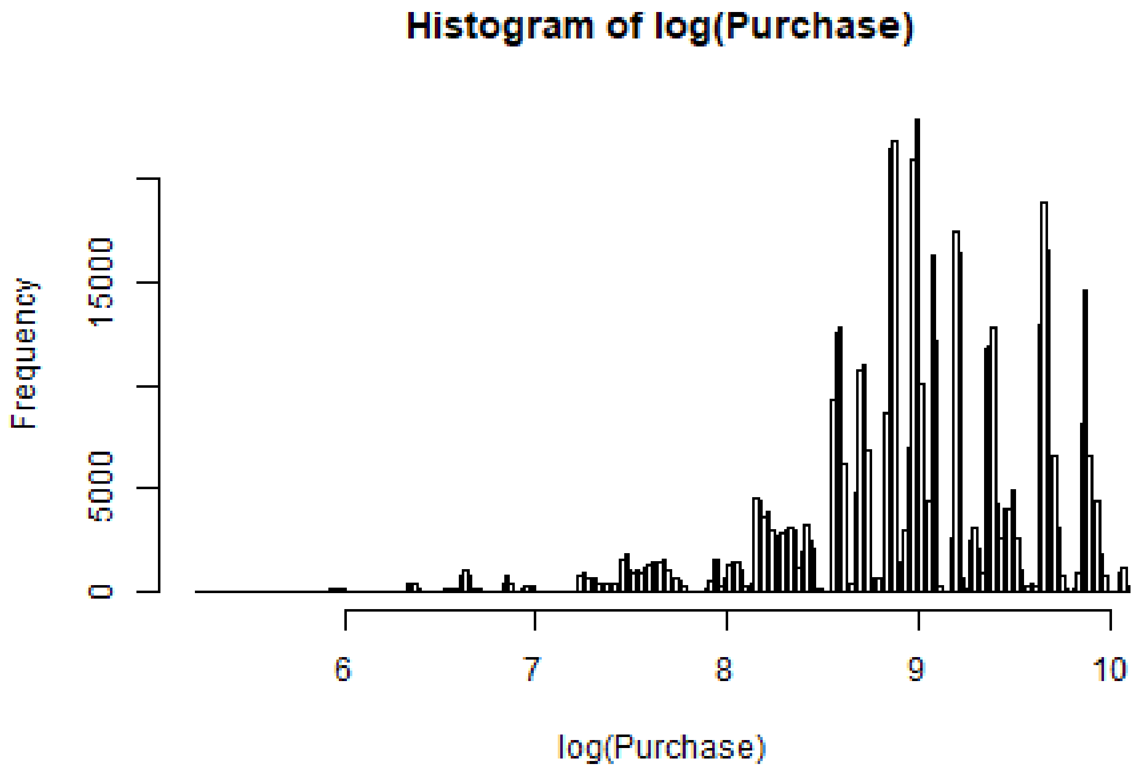



```
## [1]  87441  93017 370892
```

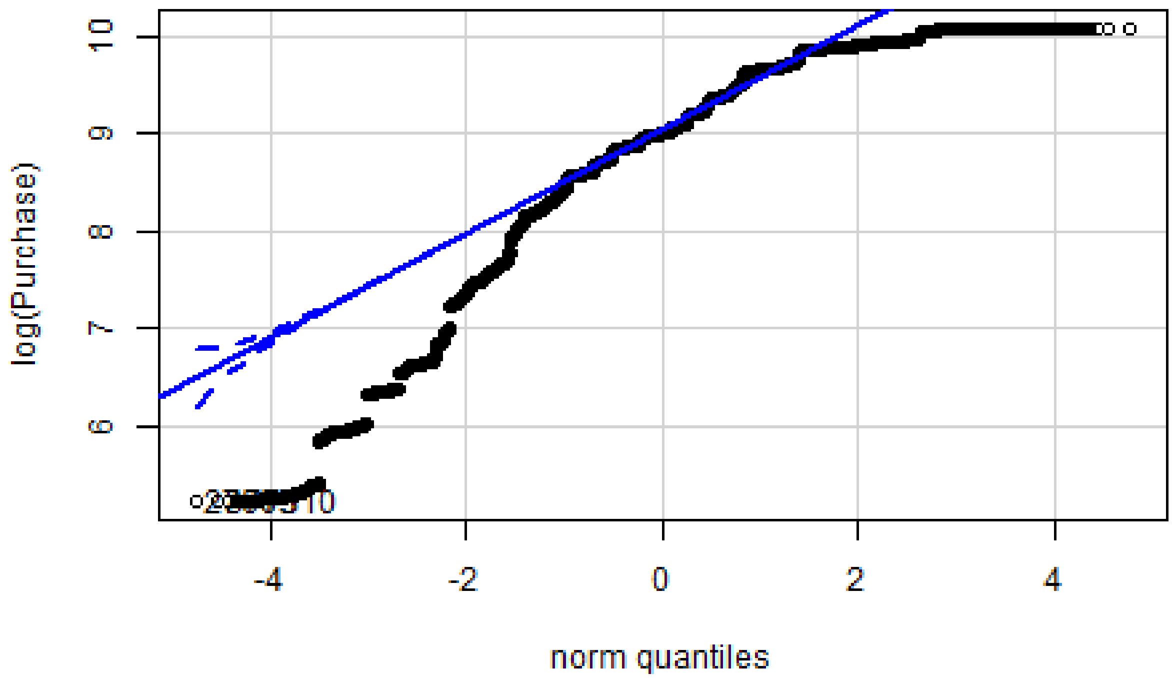
```r
#I fit a gamma distribution to the inc data
```

Purchase looks fairly normal from the QQ-Plots but I log the data just to make sure there isn't a better transformation.

```r
#Histogram of Log(Purchases)
hist(log(Purchase), breaks = "FD")
```



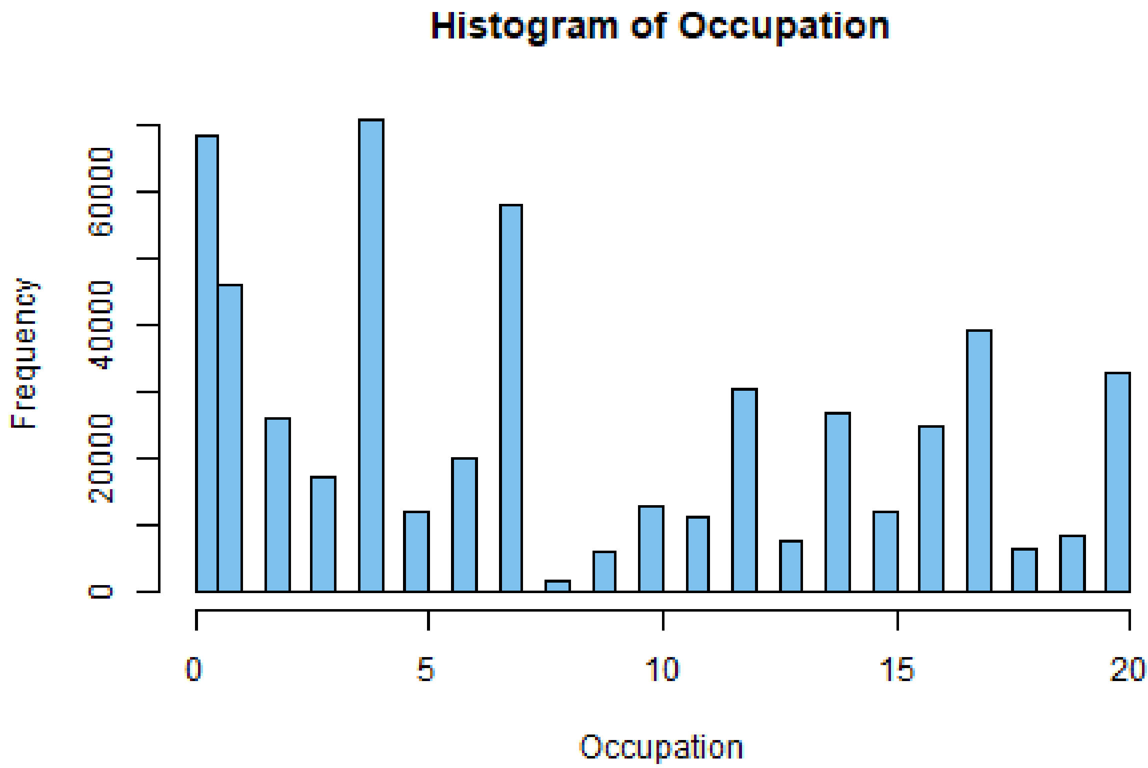

```
qqPlot(log(Purchase))
```




```
## [1]  27603 377310
```

Based on the QQ plot I can see that Purchase's should not be transformed as it makes the data less normal. Next, I looked at the histogram for Occupation.

```r
#Histogram of occupation
par(mfrow=c(1,1))
hist(Occupation, breaks = "FD", col = "skyblue2")
```

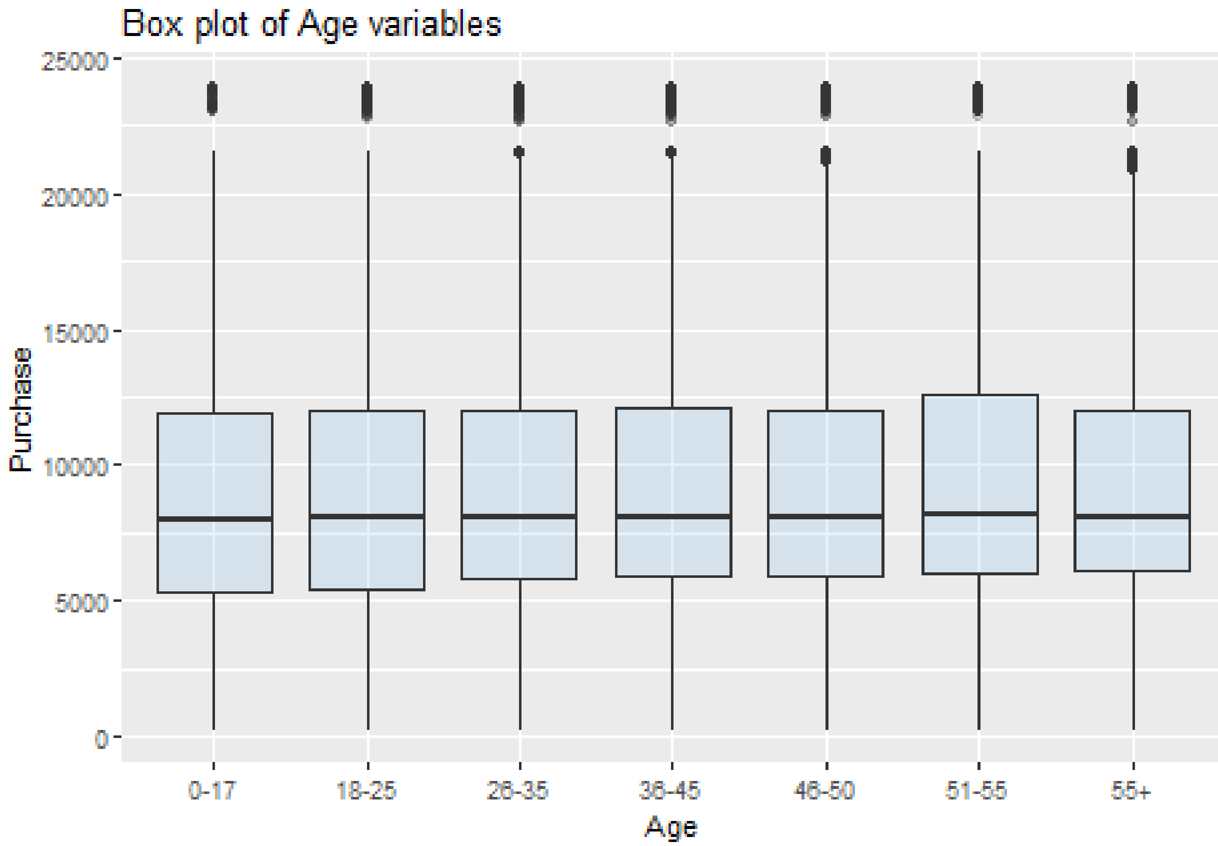


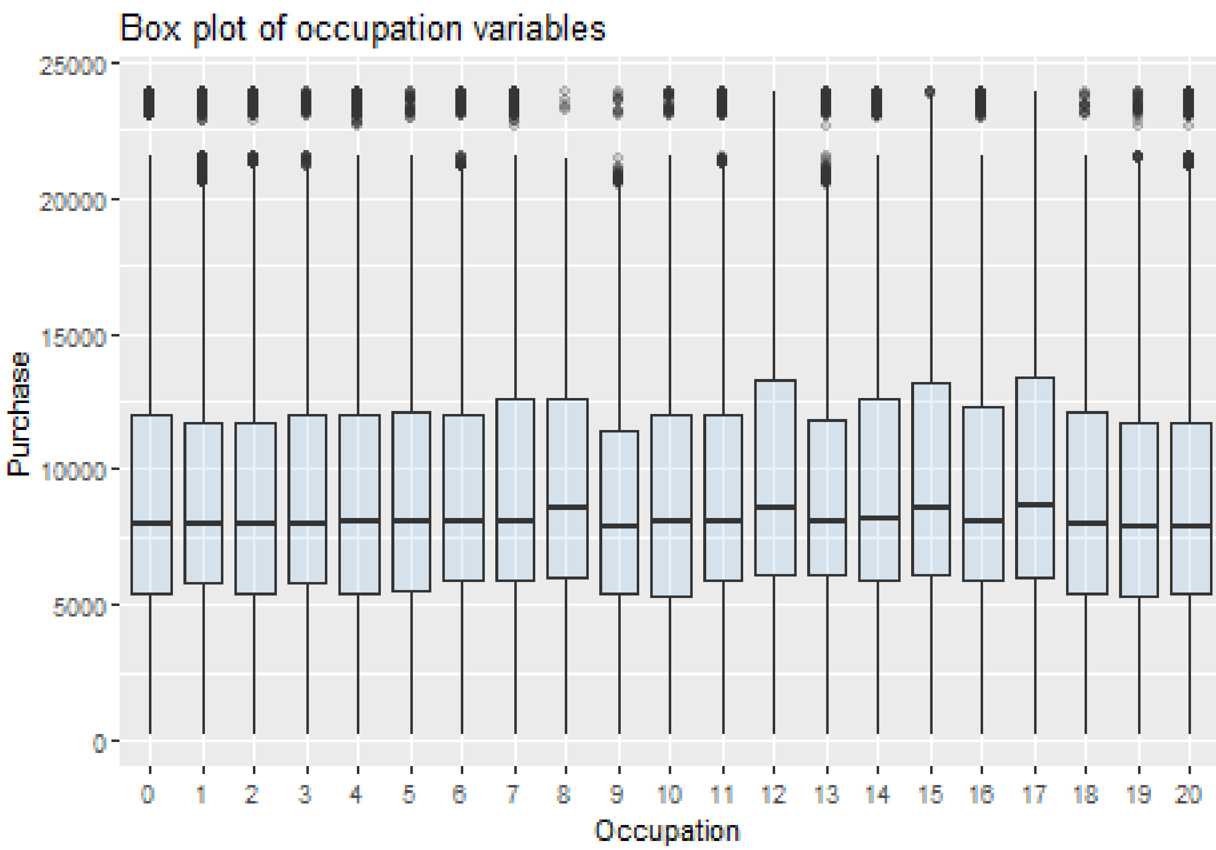

```
S(Occupation)
```

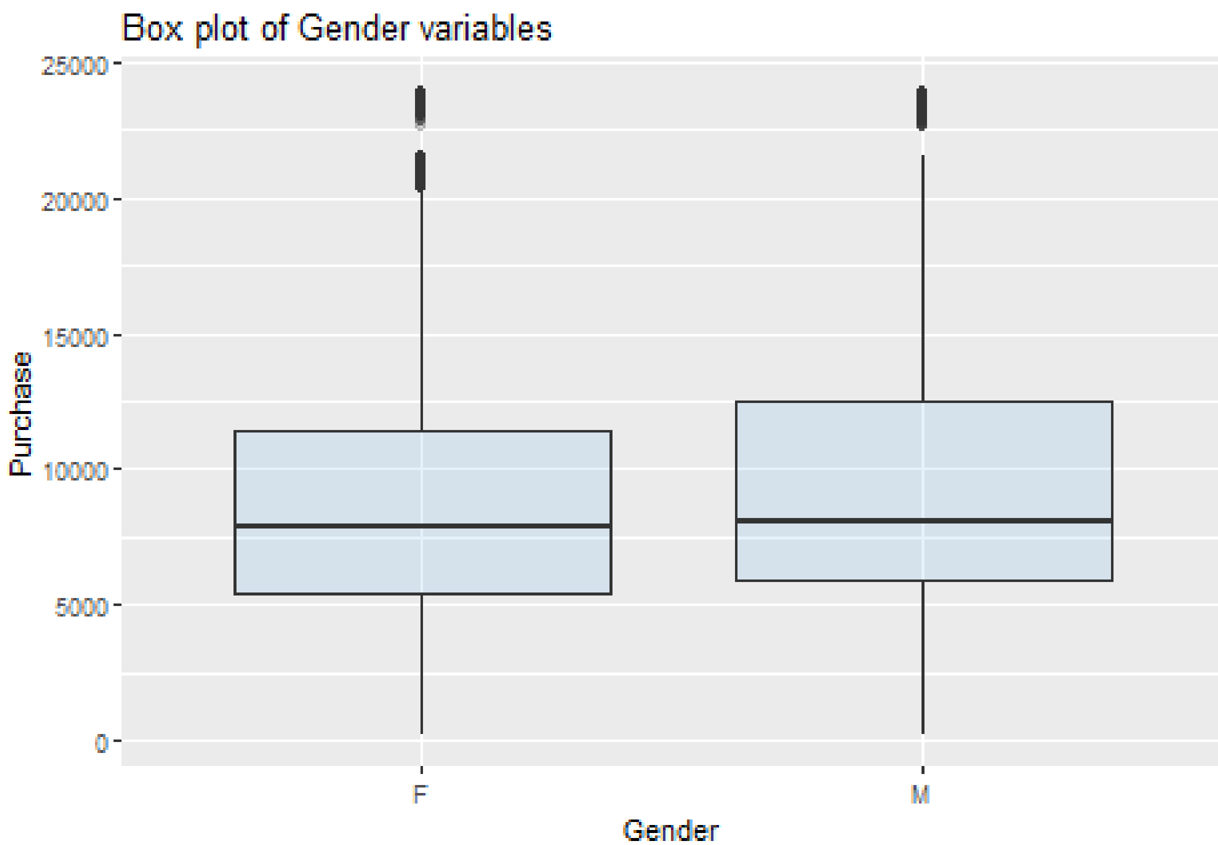
```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   0.000   2.000   7.000   8.083  14.000  20.000
```

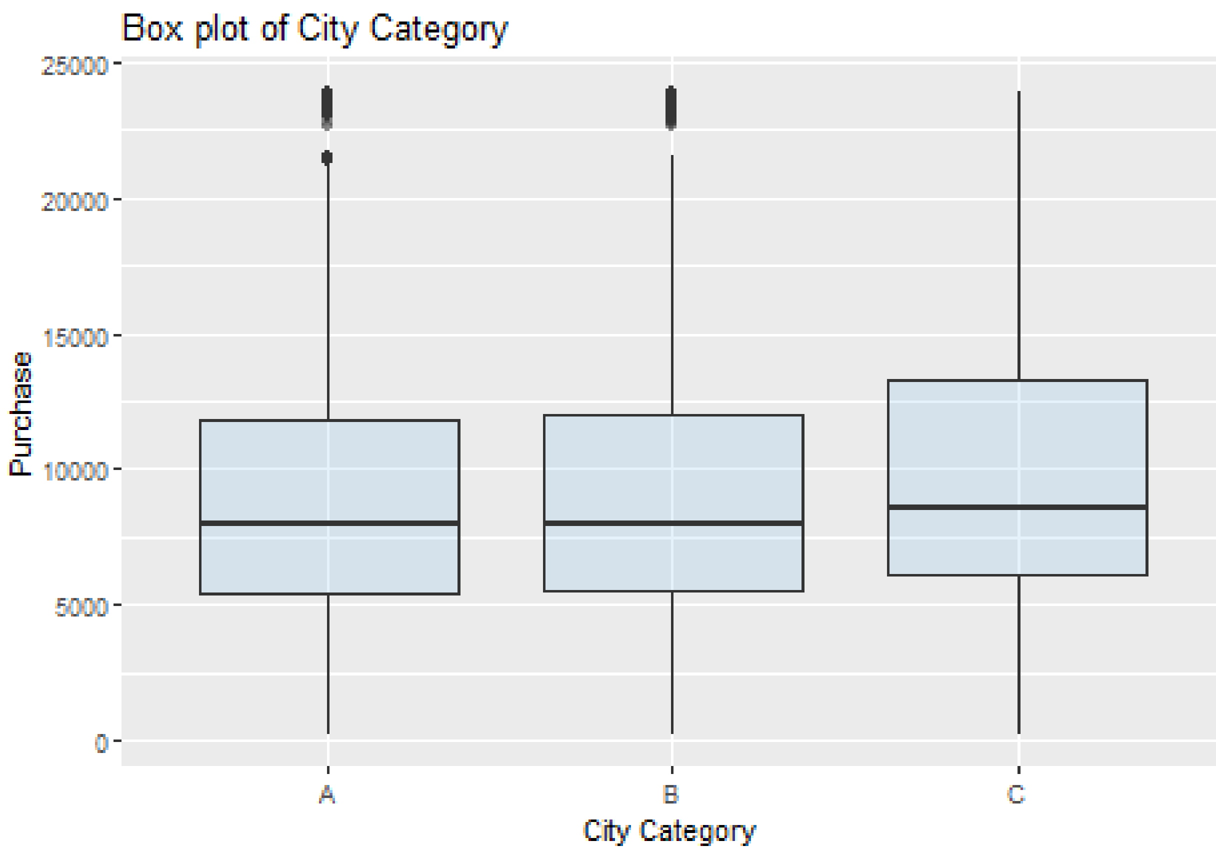
This plot illustrates how many observed individuals fall in specific Occupation categories. I can see from the histogram that the data is not skewed, does not contain any outliers, and does not need to be transformed. I confirmed these results using the Box-Cox transformation test.

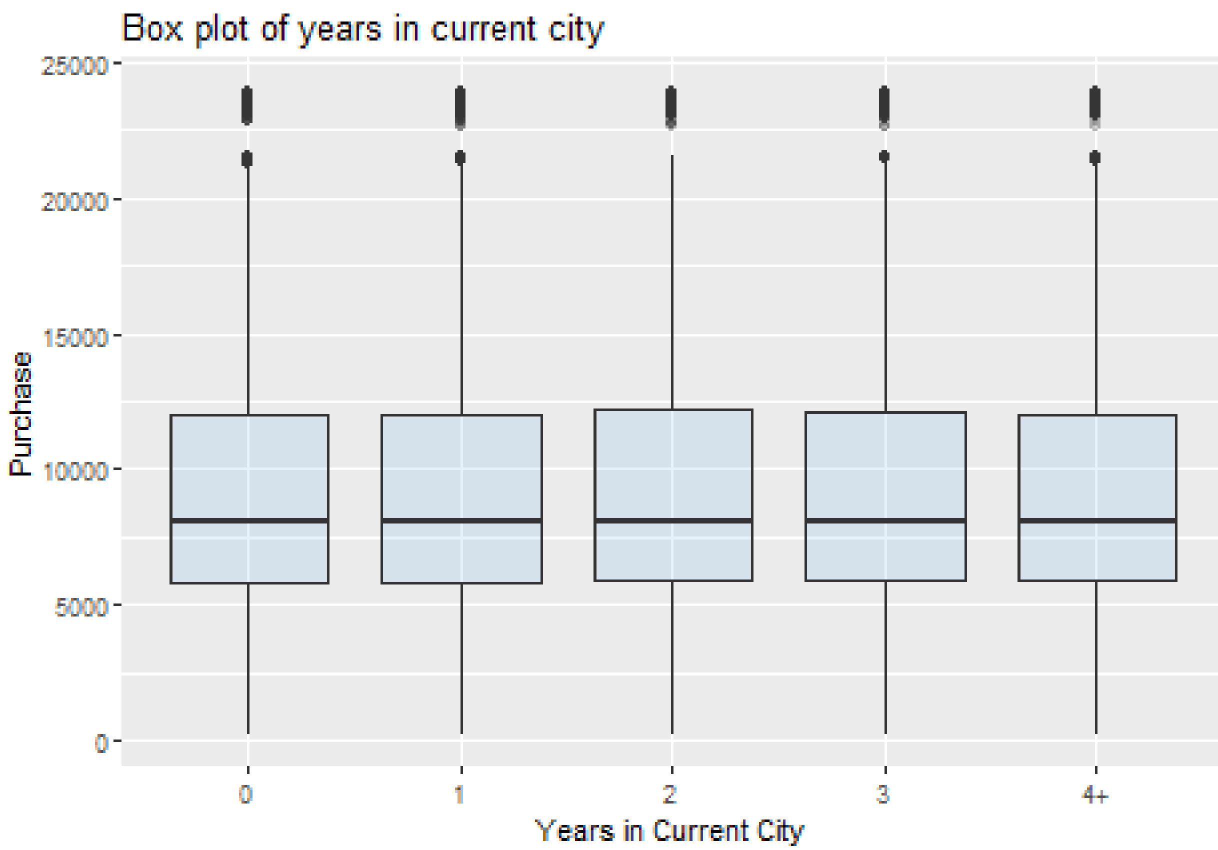
Next, I looked at the box plots to see what the spread of purchase amount looks like for each of the variables.

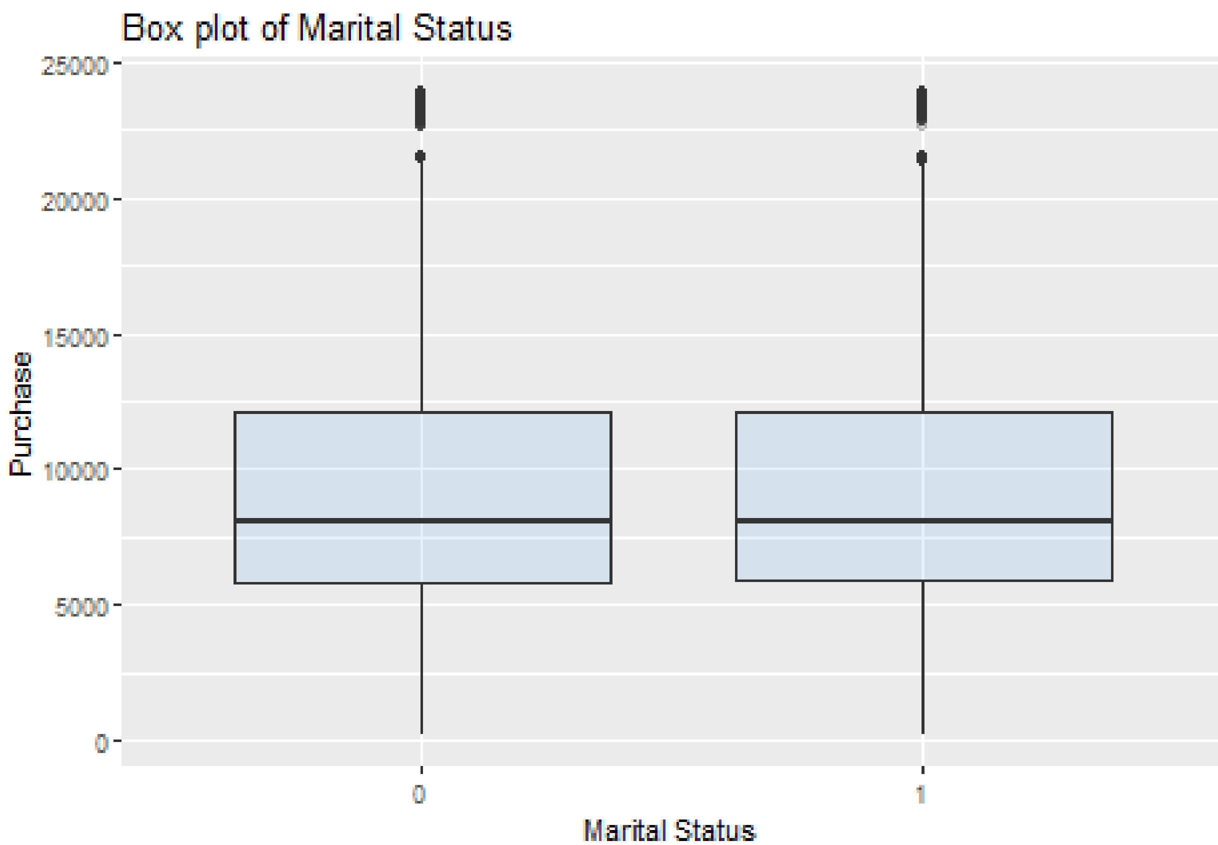

Box plot of Age variables
25000
20000
15000
10000
5000
0
Purchase
0-17
18-25
26-35
36-45
46-50
51-55
55+
Age

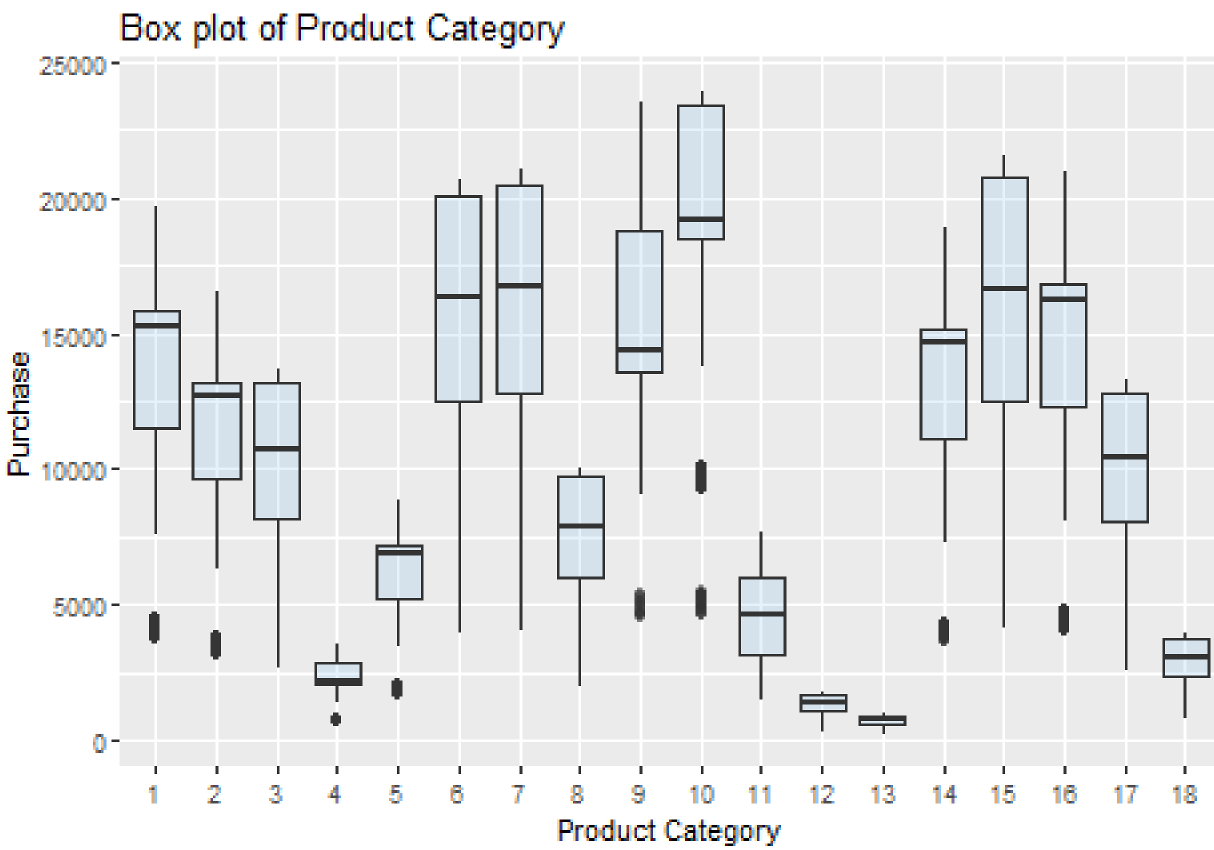

Box plot of occupation variables
Purchase
25000
20000
15000
10000
5000
0
0 1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20
Occupation

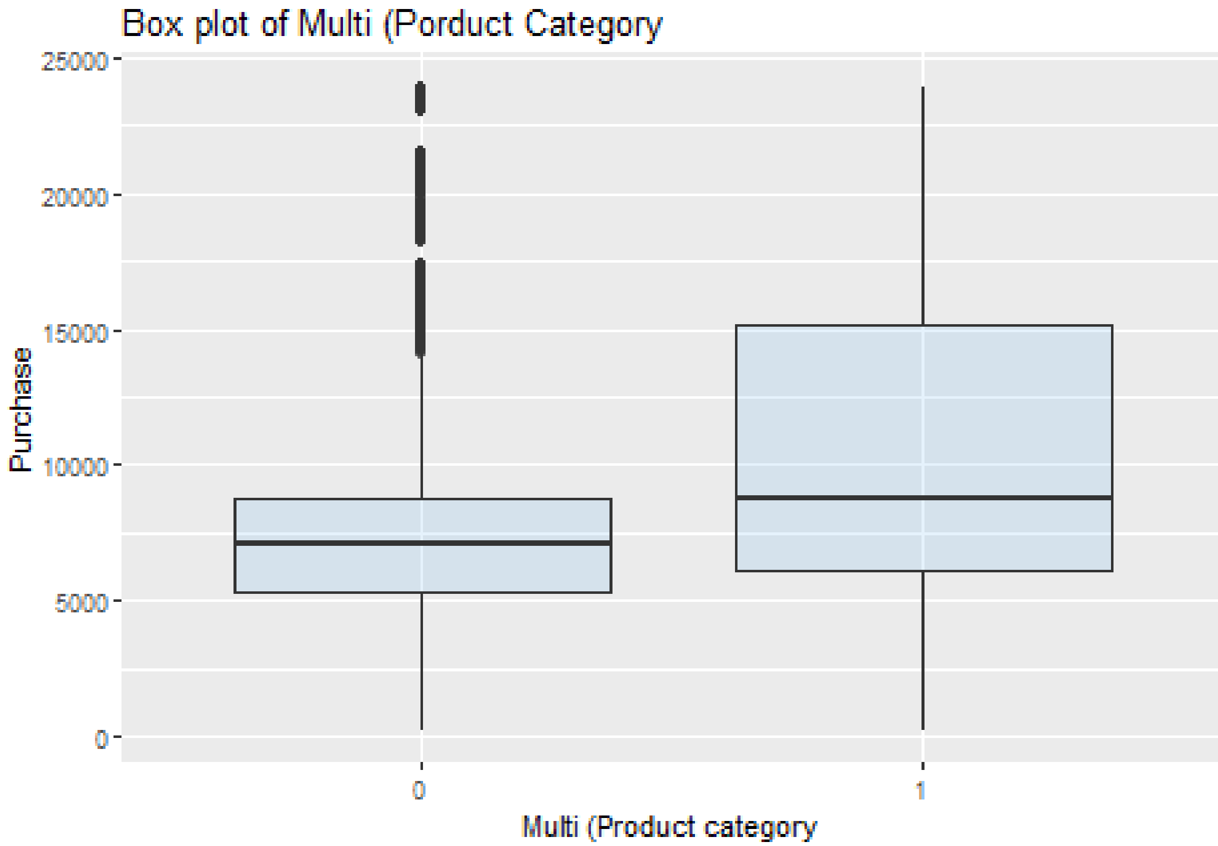

Box plot of Gender variables
25000
20000
15000
10000
5000
0
Purchase
F
M
Gender


Box plot of City Category
25000
20000
15000
10000
5000
0
Purchase
A
B
C
City Category


Box plot of years in current city
Purchase
25000
20000
15000
10000
5000
0
0
1
2
3
4+
Years in Current City


Box plot of Marital Status
25000
20000
15000
10000
5000
0
Purchase
0
1
Marital Status


Box plot of Product Category
25000
20000
15000
10000
5000
0
Purchase
1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
16
17
18
Product Category



I make the following observations based on the graphs above:

1. The mean purchase amount is fairly evenly distributed across the specific category levels within each variable except for Product Category and multi.
2. I noticed that there is a larger spread in purchase amount when the product belongs to only one category.
3. Different product categories have different purchase amount means with different spreads.

Now that I have looked at some of the univariate and bivariate characteristics, I create a regression model with all of the additive terms.

```r
#Model 1
mod.1 <- lm(Purchase ~ Gender + Age + Occupation +  City_Category+Stay_In_Current_City_Years +
              Marital_Status+Product_Category_1 + multi, data)
S(mod.1)
```

```
## Call: lm(formula = Purchase ~ Gender + Age + Occupation + City_Category +
##          Stay_In_Current_City_Years + Marital_Status + Product_Category_1 + multi,
##          data = data)
##
## Coefficients:
##                                Estimate Std. Error  t value Pr(>|t|)
## (Intercept)                   9132.1573    48.2767  189.163  < 2e-16 ***
## GenderM                        523.1621    14.9960   34.887  < 2e-16 ***
## Age18-25                       349.5520    41.7249    8.378  < 2e-16 ***
## Age26-35                       543.4284    40.5208   13.411  < 2e-16 ***
## Age36-45                       633.9828    41.6525   15.221  < 2e-16 ***
```

```
## Age46-50                        606.6089    45.7437   13.261  < 2e-16 ***
## Age51-55                        934.5792    46.7410   19.995  < 2e-16 ***
## Age55+                          739.0741    51.2880   14.410  < 2e-16 ***
## Occupation                        6.1314     0.9947    6.164 7.10e-10 ***
## City_CategoryB                  162.8656    15.8793   10.256  < 2e-16 ***
## City_CategoryC                  717.7012    17.1714   41.796  < 2e-16 ***
## Stay_In_Current_City_Years1      28.5154    20.5123    1.390  0.16448
## Stay_In_Current_City_Years2      62.9861    22.8902    2.752  0.00593 **
## Stay_In_Current_City_Years3      27.0766    23.2618    1.164  0.24443
## Stay_In_Current_City_Years4+     44.4407    23.8489    1.863  0.06240 .
## Marital_Status                  -53.9095    13.8472   -3.893 9.89e-05 ***
## Product_Category_1             -355.3684     1.8840 -188.628  < 2e-16 ***
## multi                          1139.8707    15.2280   74.854  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard deviation: 4685 on 537559 degrees of freedom
## Multiple R-squared: 0.1153
## F-statistic:  4120 on 17 and 537559 DF,  p-value: < 2.2e-16
##      AIC      BIC
## 10612982 10613194
```

Since Stay_In_Current_City was not significant (except for Stay_In_Current_City_2) I used a Chow-Test to determine if the estimated coefficient of any of the "Stay in Current City" variables are equal to zero.

```
hyp <- c("Stay_In_Current_City_Years1 = 0", "Stay_In_Current_City_Years2 = 0",
         "Stay_In_Current_City_Years3 = 0", "Stay_In_Current_City_Years4+ = 0")
linearHypothesis(mod.1, hyp)
```

```
## Linear hypothesis test
##
## Hypothesis:
## Stay_In_Current_City_Years1 = 0
## Stay_In_Current_City_Years2 = 0
## Stay_In_Current_City_Years3 = 0
## Stay_In_Current_City_Years4  +  = 0
##
## Model 1: restricted model
## Model 2: Purchase ~ Gender + Age + Occupation + City_Category + Stay_In_Current_City_Years +
##     Marital_Status + Product_Category_1 + multi
##
##   Res.Df      RSS Df Sum of Sq      F  Pr(>F)
## 1 537563 1.18e+13
## 2 537559 1.18e+13  4 186223076 2.1209 0.07539 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Because the p-value is greater than 0.05, there is insufficient evidence to justify keeping "Stay in Current City" in the model. I removed this in the following model.

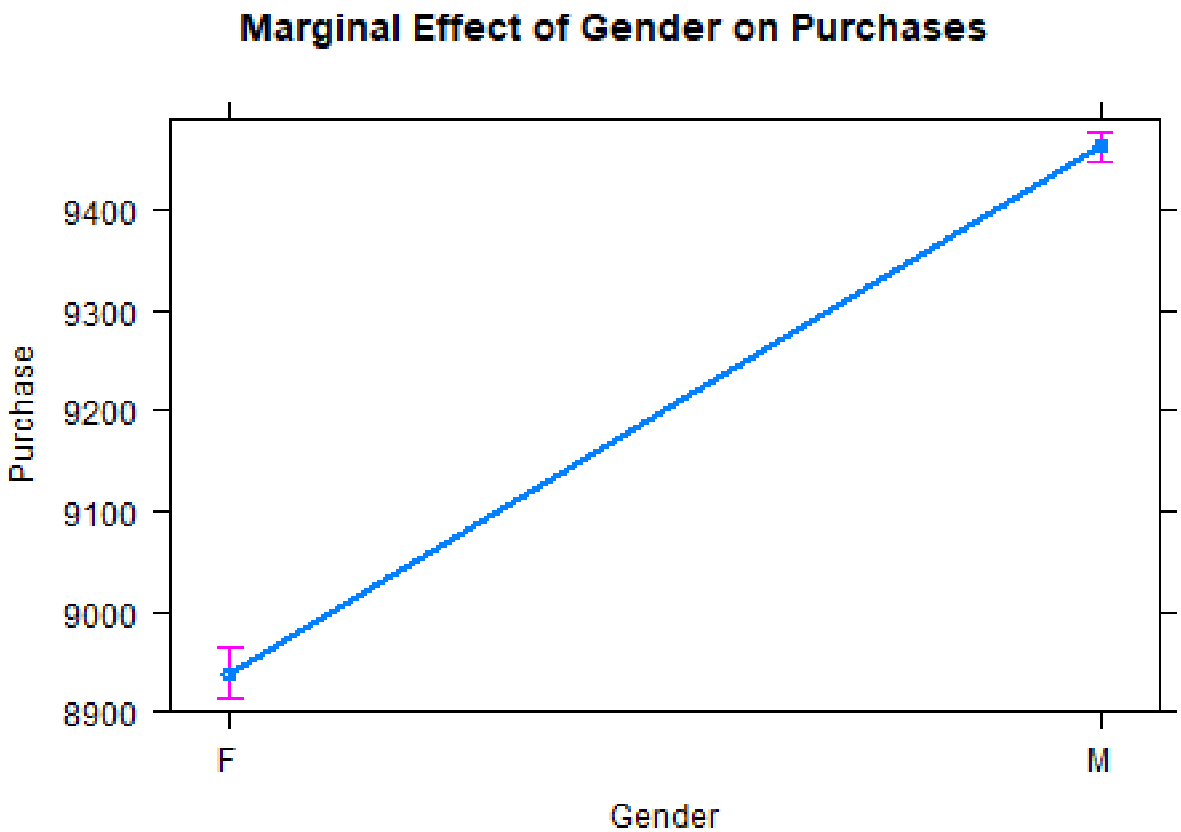
```
#Model 2 (stay in city removed)
mod.2 <- lm(Purchase ~ Gender + Age + Occupation + City_Category +
              Marital_Status+Product_Category_1 + multi, data)
S(mod.2)
```

```
## Call: lm(formula = Purchase ~ Gender + Age + Occupation + City_Category +
```

```
##          Marital_Status + Product_Category_1 + multi, data = data)
##
## Coefficients:
##                    Estimate Std. Error  t value Pr(>|t|)
## (Intercept)        9164.473     45.747  200.330  < 2e-16 ***
## GenderM             523.371     14.983   34.932  < 2e-16 ***
## Age18-25            348.786     41.709    8.362  < 2e-16 ***
## Age26-35            543.423     40.500   13.418  < 2e-16 ***
## Age36-45            633.682     41.641   15.218  < 2e-16 ***
## Age46-50            605.493     45.716   13.245  < 2e-16 ***
## Age51-55            933.942     46.696   20.001  < 2e-16 ***
## Age55+              738.935     51.273   14.412  < 2e-16 ***
## Occupation            6.151      0.994    6.188 6.11e-10 ***
## City_CategoryB      163.903     15.859   10.335  < 2e-16 ***
## City_CategoryC      719.382     17.154   41.937  < 2e-16 ***
## Marital_Status      -53.926     13.846   -3.895 9.83e-05 ***
## Product_Category_1 -355.386      1.884 -188.655  < 2e-16 ***
## multi              1140.011     15.228   74.863  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard deviation: 4685 on 537563 degrees of freedom
## Multiple R-squared: 0.1153
## F-statistic:  5387 on 13 and 537563 DF,  p-value: < 2.2e-16
##      AIC      BIC
## 10612982 10613150
```

Now that all the variables are statistically significant, I looked at the effects plots. Our R^2 remains the same.

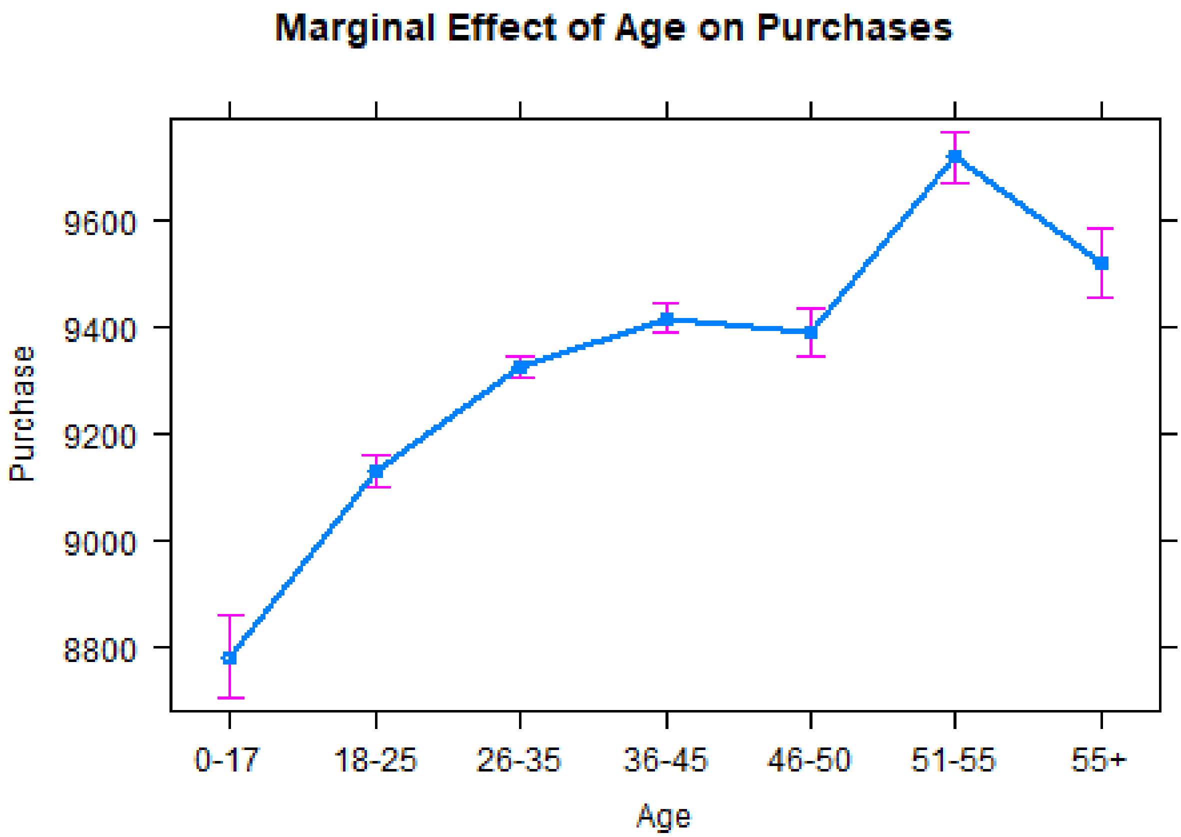
```r
plot(effect(mod = mod.2, "Gender"), main="Marginal Effect of Gender on Purchases")
```



From the Gender effect plot, I can determine that males spent more on Black Friday than females. The spread on Purchases for males is also smaller than the spread on Purchases for females.

Intuitively, this may be a result of males buying more expensive/big-ticket items on Black Friday than females.

```
plot(effect(mod = mod.2, "Age"), main="Marginal Effect of Age on Purchases")
```

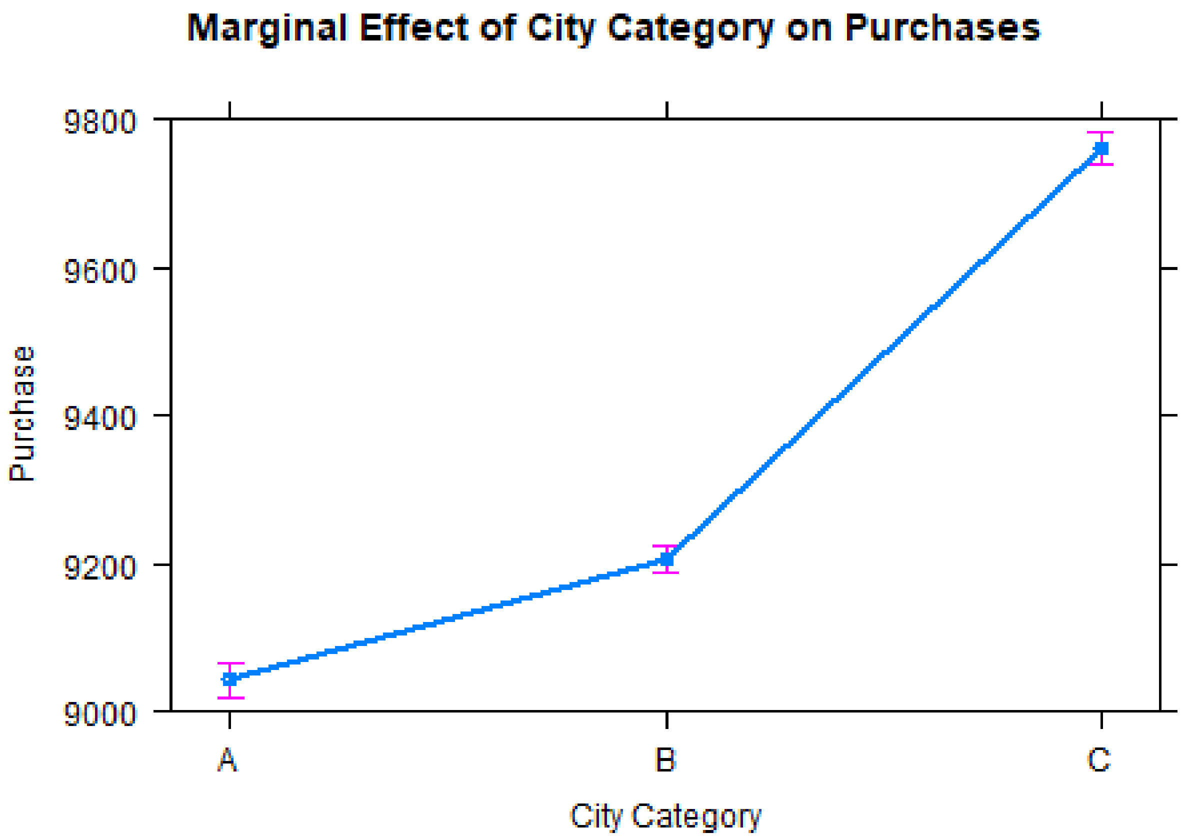


Here I see that in general, purchases go up as age increases, with more variability amongst the lower and higher age ranges.

However, I do see that purchases plateau between the age groups of 36-45 and 46-50.

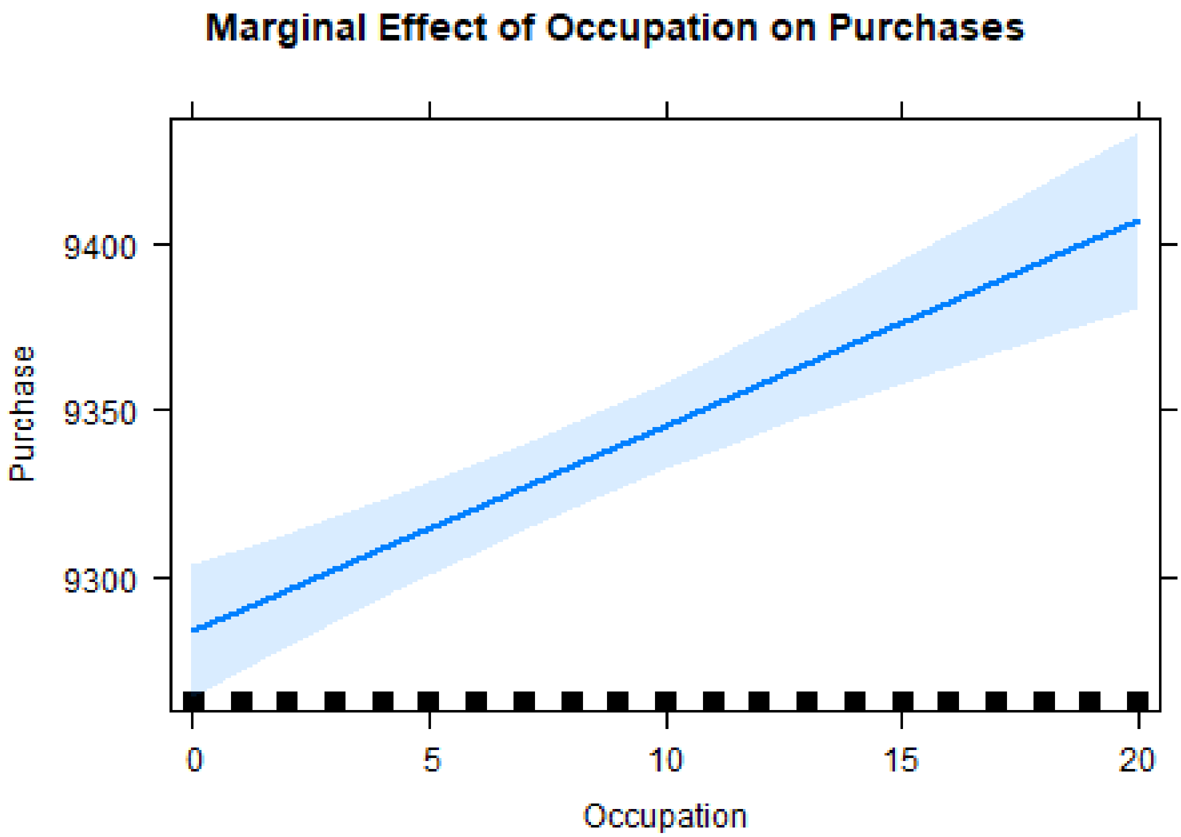
Intuitively, I believe the variability is due to younger age groups because they could be spending either their money or their parents' money. For the older age groups the variability could be due to retirees in this group with lower disposable income.

```
#Marginal effect plot of City Category
plot(effect(mod = mod.2, "City_Category"), main="Marginal Effect of City Category on Purchases", xlab="(
```



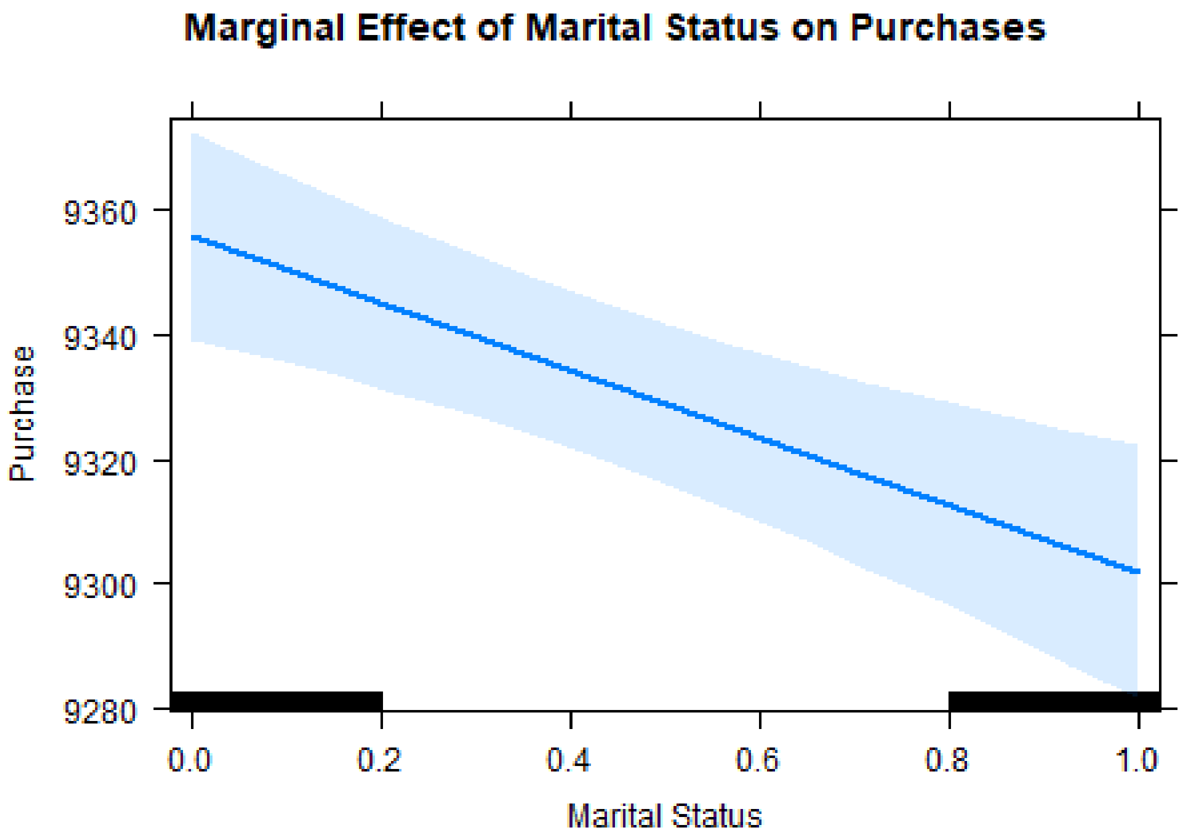
In City Category, I can see that the overall dollar value of purchases made in City C were much higher than B and C. This could mean that items are more expensive in city C or there is more variety (people shop more).

```
#Marginal effect plot of Occupation
plot(effect(mod = mod.2, "Occupation"), main="Marginal Effect of Occupation on Purchases", xlab="Occupa
```



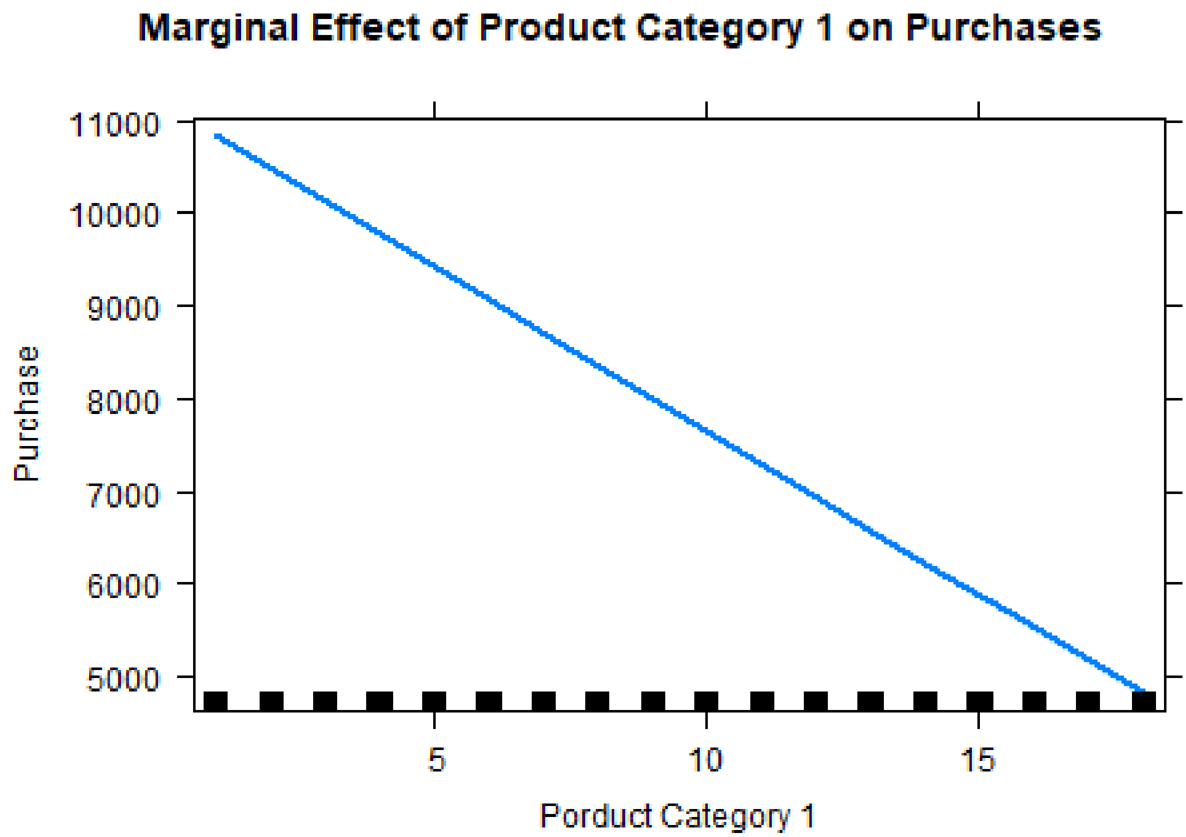
The Occupation effect plot shows us that as the Occupation category increases, then purchases increase. This could indicate that the larger the occupation category, the higher the income. If this were the case, intuitively, it makes sense that there is a little more variation at the lower and higher occupation category values. This is because at lower levels of income, you are likely to have a different spending behavior than other people in the same income bracket.

```
#Marginal effect plot of Marital Status
plot(effect(mod = mod.2, "Marital_Status"), main="Marginal Effect of Marital Status on Purchases", xlab=
```



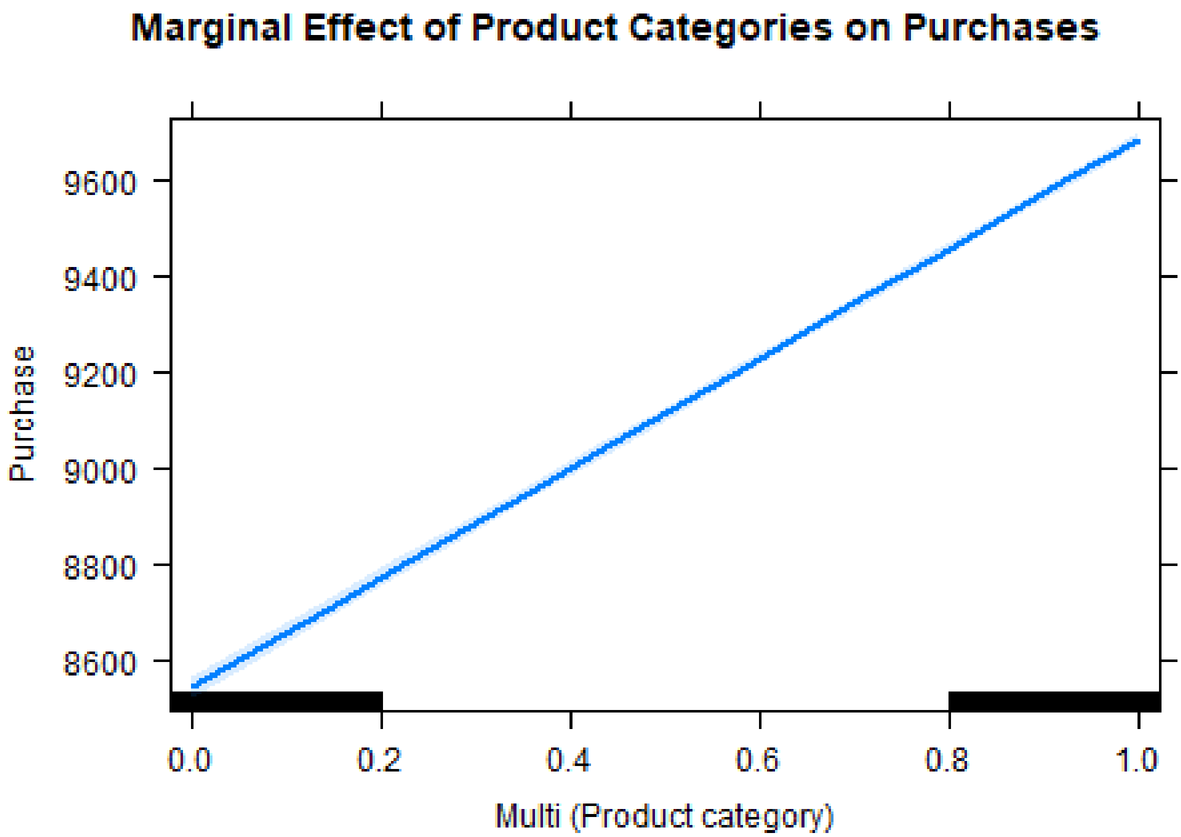
From this effects plot, it seems that purchases are lower for married individuals versus single individuals.

```
#Marginal effect plot of product category 1
plot(effect(mod = mod.2, "Product_Category_1"), main="Marginal Effect of Product Category 1 on Purchases
```



For Product Category_1, I see that the higher the number of the category, the lower the purchase dollar amount. This could mean that higher category numbers are items that are cheaper or that less people buy them.

```
#Marginal effect plot of product category
plot(effect(mod = mod.2, "multi"), main="Marginal Effect of Product Categories on Purchases", xlab="Mul
```



This plot shows that if a product belongs to more than one category then, the dollar value of purchases increase. This could indicated that the items hold more value if they belong to multiple categories or they are items that are purchased more.

From the effects plots, I noticed that Gender and multi variables looked almost identical and wanted to test if there was any degree of collinearity between the two variables. In order to do this, I looked at their correlation.

```r
#Correlation of product categories and gender
cor(multi, (as.numeric(Gender))-1)
```

```
## [1] 0.01197696
```

Because the result is so low I can safely assume that our multi and Gender variables are significantly different from each other.

Now that I have the effects plots, I use the Ramsey RESET test in order to determine whether or not our current model is mis-specified.

```r
#Ramsey RESET test
resettest(mod.2, power=2, type="regressor")
```

```
##
##  RESET test
##
## data:  mod.2
## RESET = 24835, df1 = 4, df2 = 537560, p-value < 2.2e-16
```

Now I decided to look at interaction terms and possible variable transformations because I found from the Ramsey Reset Test that our model is mis-specified with just the additive terms.

I first try adding the interaction term of gender and age because spending at the different age groups might be different depending on gender.

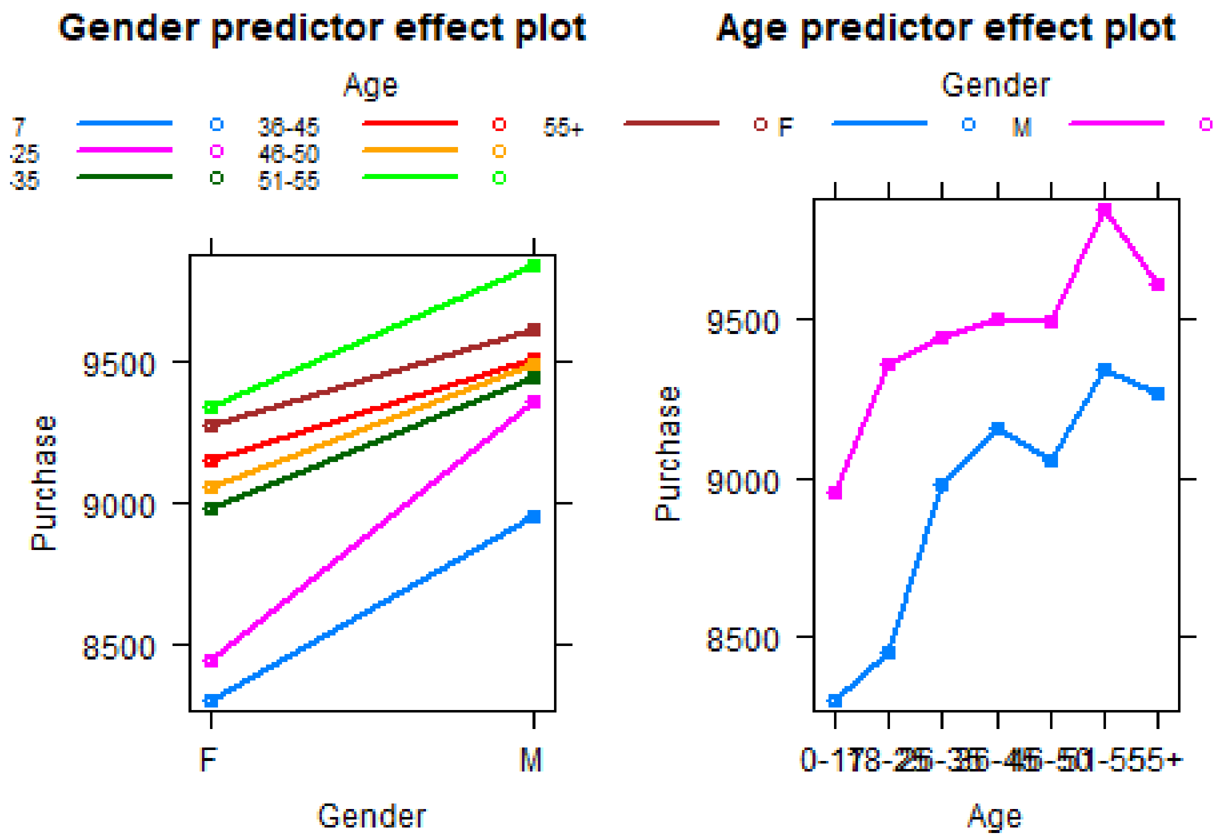
```
#Model 3 (Adding gender:age interaction term) BENCHMARK
mod.3 <- lm(Purchase ~ Gender + Age + Occupation + Marital_Status + Product_Category_1 +
              City_Category + multi + Gender:Age, data)
S(mod.3)
```

```
## Call: lm(formula = Purchase ~ Gender + Age + Occupation + Marital_Status +
##          Product_Category_1 + City_Category + multi + Gender:Age, data = data)
##
## Coefficients:
##                     Estimate Std. Error  t value Pr(>|t|)
## (Intercept)         9073.021     70.096  129.438  < 2e-16 ***
## GenderM              655.554     81.785    8.016 1.10e-15 ***
## Age18-25             148.948     73.252    2.033 0.042015 *
## Age26-35             675.167     70.115    9.629  < 2e-16 ***
## Age36-45             849.590     72.747   11.679  < 2e-16 ***
## Age46-50             754.962     79.139    9.540  < 2e-16 ***
## Age51-55            1036.082     82.424   12.570  < 2e-16 ***
## Age55+               965.924     94.756   10.194  < 2e-16 ***
## Occupation             6.464      0.995    6.496 8.23e-11 ***
## Marital_Status       -50.459     13.854   -3.642 0.000270 ***
## Product_Category_1  -355.260      1.884 -188.605  < 2e-16 ***
## City_CategoryB       166.108     15.876   10.463  < 2e-16 ***
## City_CategoryC       717.925     17.172   41.807  < 2e-16 ***
## multi               1140.442     15.227   74.898  < 2e-16 ***
## GenderM:Age18-25     248.626     88.888    2.797 0.005157 **
## GenderM:Age26-35    -191.445     85.238   -2.246 0.024704 *
## GenderM:Age36-45    -304.618     88.249   -3.452 0.000557 ***
## GenderM:Age46-50    -223.032     95.419   -2.337 0.019419 *
## GenderM:Age51-55    -155.584     98.767   -1.575 0.115195
## GenderM:Age55+      -317.237    111.948   -2.834 0.004600 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard deviation: 4685 on 537557 degrees of freedom
## Multiple R-squared: 0.1155
## F-statistic:  3695 on 19 and 537557 DF,  p-value: < 2.2e-16
##      AIC      BIC
## 10612829 10613064
```

After adding in the interaction term of Gender and Age, I see that the Adjusted R-Squared improves and the variable estimates do not vary by much. Additionally, the estimates remain statistically significant. The interaction terms of Gender on the different age groups have varying degrees of statistical significance. Going forward I use this model as a benchmark for comparison.

Next I show the predictor effects plots of the interaction between Gender and Age on Purchase.

```
#plotting the predictor effects of gender and age on purchase
plot(predictorEffects(mod.3, ~ Gender:Age), lines=list(multiline=TRUE))
```



I can see from the plots that males Purchases are higher at every age group. The difference in Purchases between males and females is large at the age group of "0-17" and decreases as the age group increases. So female spending increases faster than male spending as the age group increases and begins to converge with male spending. These two plots are showing us the same information.

Next I try adding the interaction term of gender and occupation because the spending might be different in the occupation categories depending on the gender.

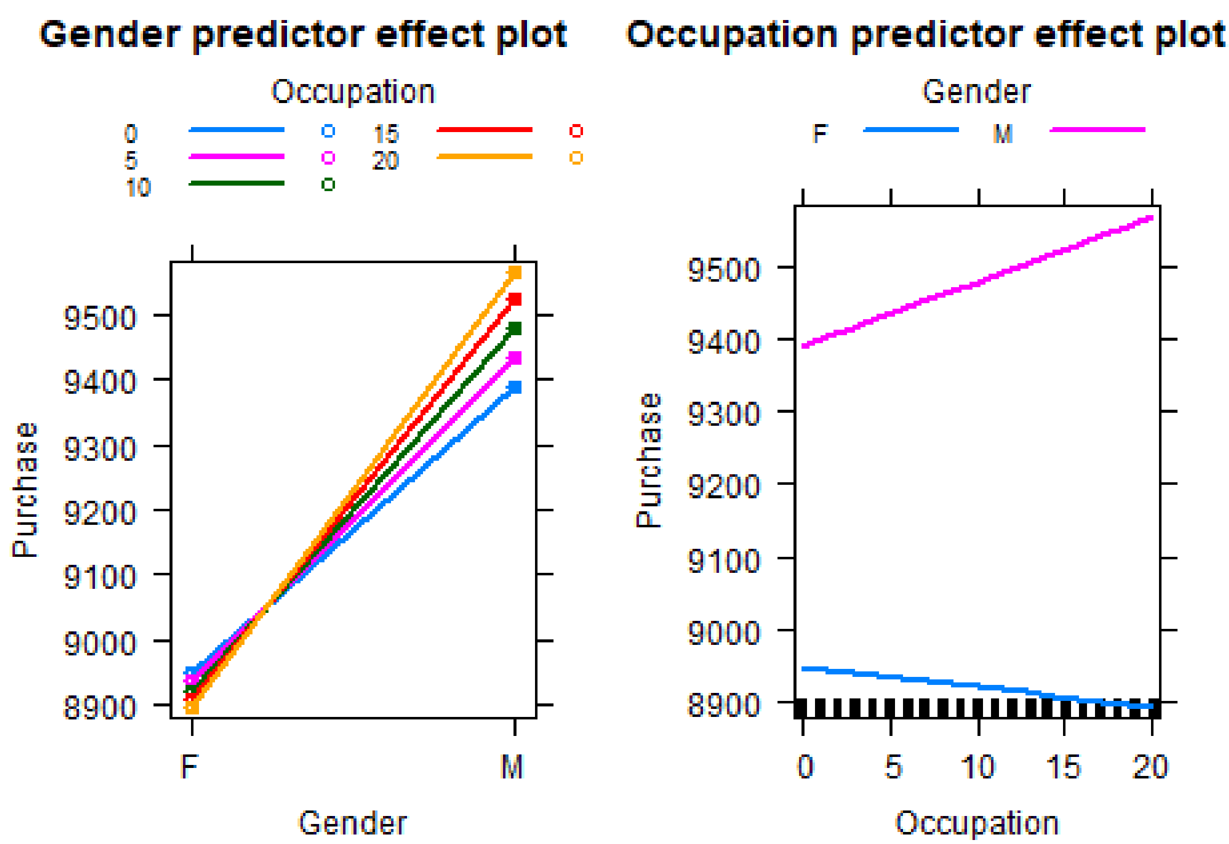
```
#Model 4 (Adding gender:occupation interaction term)
mod.4 <- lm(Purchase ~ Gender + Age + Occupation + Marital_Status + Product_Category_1 +
              City_Category + multi + Gender:Occupation, data)
S(mod.4)
```

```
## Call: lm(formula = Purchase ~ Gender + Age + Occupation + Marital_Status +
##           Product_Category_1 + City_Category + multi + Gender:Occupation, data =
##           data)
##
## Coefficients:
##                      Estimate Std. Error  t value Pr(>|t|)
## (Intercept)          9227.009     47.473  194.364  < 2e-16 ***
## GenderM               440.335     22.546   19.531  < 2e-16 ***
## Age18-25              345.953     41.712    8.294  < 2e-16 ***
## Age26-35              538.604     40.511   13.295  < 2e-16 ***
## Age36-45              629.584     41.649   15.117  < 2e-16 ***
## Age46-50              596.101     45.755   13.028  < 2e-16 ***
## Age51-55              925.705     46.725   19.812  < 2e-16 ***
## Age55+                730.781     51.299   14.246  < 2e-16 ***
```

```
## Occupation           -2.805      2.071   -1.354 0.175631
## Marital_Status      -50.301     13.865   -3.628 0.000286 ***
## Product_Category_1 -355.423      1.884 -188.677  < 2e-16 ***
## City_CategoryB      166.449     15.867   10.490  < 2e-16 ***
## City_CategoryC      721.500     17.159   42.048  < 2e-16 ***
## multi              1139.643     15.228   74.840  < 2e-16 ***
## GenderM:Occupation   11.617      2.357    4.929 8.28e-07 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard deviation: 4685 on 537562 degrees of freedom
## Multiple R-squared: 0.1153
## F-statistic:  5004 on 14 and 537562 DF,  p-value: < 2.2e-16
##      AIC      BIC
## 10612960 10613139
```

The adjusted R-Squared does not change when I add this interaction term compared to the model with just interaction terms. Additionally, Occupation becomes insignificant, while the interaction of Occupation and Gender is significant.

```
#Marginal effect of Gender:occupation interaction (with occupation variable)
plot(predictorEffects(mod.4, ~ Gender:Occupation), lines=list(multiline=TRUE))
```



I can see from the plots that males spend more at every Occupation category. Females spend the least in occupation category 20 while males spend the most in occupation category 20. This is true for every occupation level.

I now try adding the gender occupation interaction term, but take out the occupation variable because it was insignificant in the previous model with this interaction term.

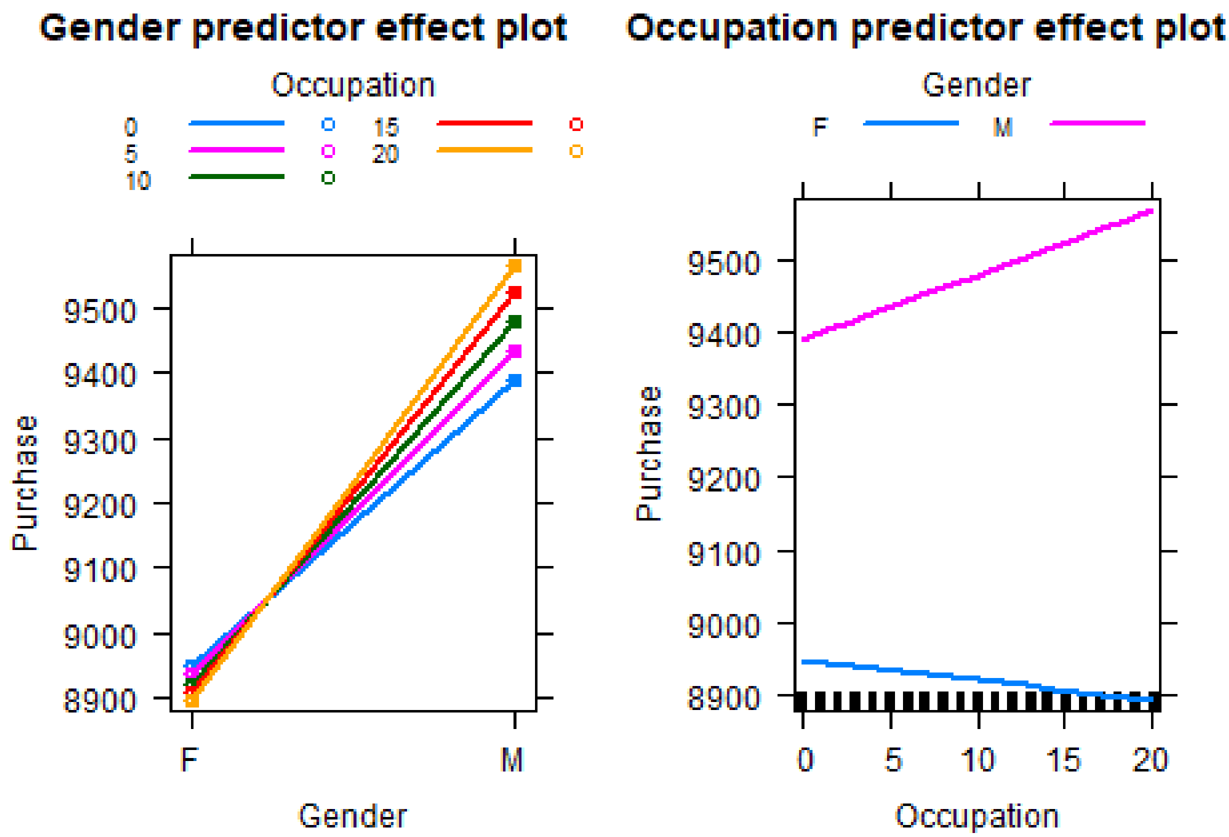
```r
#Model 5 (adding gender:occupation interaction term, but without occupation term)
mod.5 <- lm(Purchase ~ Gender + Age + Marital_Status + Product_Category_1 +
              City_Category + multi + Gender:Occupation, data)
S(mod.5)
```

```
## Call: lm(formula = Purchase ~ Gender + Age + Marital_Status +
##          Product_Category_1 + City_Category + multi + Gender:Occupation, data =
##          data)
##
## Coefficients:
##                    Estimate Std. Error  t value Pr(>|t|)
## (Intercept)        9227.009     47.473  194.364  < 2e-16 ***
## GenderM             440.335     22.546   19.531  < 2e-16 ***
## Age18-25            345.953     41.712    8.294  < 2e-16 ***
## Age26-35            538.604     40.511   13.295  < 2e-16 ***
## Age36-45            629.584     41.649   15.117  < 2e-16 ***
## Age46-50            596.101     45.755   13.028  < 2e-16 ***
## Age51-55            925.705     46.725   19.812  < 2e-16 ***
## Age55+              730.781     51.299   14.246  < 2e-16 ***
## Marital_Status      -50.301     13.865   -3.628 0.000286 ***
## Product_Category_1 -355.423      1.884 -188.677  < 2e-16 ***
## City_CategoryB      166.449     15.867   10.490  < 2e-16 ***
## City_CategoryC      721.500     17.159   42.048  < 2e-16 ***
## multi              1139.643     15.228   74.840  < 2e-16 ***
## GenderF:Occupation   -2.805      2.071   -1.354 0.175631
## GenderM:Occupation    8.812      1.131    7.790 6.71e-15 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard deviation: 4685 on 537562 degrees of freedom
## Multiple R-squared: 0.1153
## F-statistic:  5004 on 14 and 537562 DF,  p-value: < 2.2e-16
##      AIC      BIC
## 10612960 10613139
```

The Adjusted R-Squared does not change after I take out the occupation category and statistical significance do not change for the variables. However, the interaction of females with occupation becomes insignificant.

Next I show the predictor effects plots of the gender occupation interaction term in this model.

```r
#Marginal effect of Gender:occupation interaction (without occupation variable)
plot(predictorEffects(mod.5, ~ Gender:Occupation), lines=list(multiline=TRUE))
```



The predictor effects of spending based on gender and occupation does not change from the previous model.

After observing in the effects plots that the marginal effect of occupation on purchase is linear and strictly increasing, I think that the occupation category is separated based on purchasing power because the effect plot shows that as the occupation category increases the spending increases. Due to this, I think adding a quadratic term to occupation will improve the model as it will capture the decreasing marginal returns in spending to an increase in the occupation category.

```
#Model 7 (with quadratic occupation)
mod.6 <- lm(Purchase ~ Gender + Age + Occupation + I(Occupation^2) + City_Category +
              Marital_Status + Product_Category_1 + multi, data)
S(mod.6)
```

```
## Call: lm(formula = Purchase ~ Gender + Age + Occupation + I(Occupation^2)
##          + City_Category + Marital_Status + Product_Category_1 + multi, data =
##          data)
##
## Coefficients:
##                      Estimate Std. Error  t value Pr(>|t|)
## (Intercept)         8977.8279    47.3348  189.666  < 2e-16 ***
## GenderM              515.6934    14.9878   34.407  < 2e-16 ***
## Age18-25             411.0227    41.8974    9.810  < 2e-16 ***
## Age26-35             617.5623    40.7796   15.144  < 2e-16 ***
## Age36-45             698.8090    41.8492   16.698  < 2e-16 ***
## Age46-50             674.4405    45.9280   14.685  < 2e-16 ***
## Age51-55            1001.2764    46.8925   21.353  < 2e-16 ***
## Age55+               795.0995    51.3931   15.471  < 2e-16 ***
```

```
## Occupation           61.4957     3.7502   16.398  < 2e-16 ***
## I(Occupation^2)       -2.9421     0.1922  -15.305  < 2e-16 ***
## City_CategoryB       155.3145    15.8653    9.790  < 2e-16 ***
## City_CategoryC       705.3858    17.1746   41.072  < 2e-16 ***
## Marital_Status       -48.6963    13.8471   -3.517 0.000437 ***
## Product_Category_1 -354.9086     1.8836 -188.417  < 2e-16 ***
## multi              1138.9584    15.2248   74.809  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard deviation: 4684 on 537562 degrees of freedom
## Multiple R-squared: 0.1156
## F-statistic:  5021 on 14 and 537562 DF,  p-value: < 2.2e-16
##      AIC      BIC
## 10612750 10612929
```

From the output I can see that the Adjusted R-Squared of 0.1156 has improved and all of the variables are statistically significant. I can see that the estimate of the quadratic occupation term is negative, which indicates that there is decreasing marginal returns to spending as the occupation category increases as I expected.

Next I try adding in the gender and age occupation interaction term with the quadratic occupation term in the model.

```
#model with quadratic occupation and gender:age
mod.7 <- lm(Purchase ~ Gender + Age + Occupation + I(Occupation^2) + Gender:Age + City_Category +
              Marital_Status + Product_Category_1 + multi, data)
S(mod.7)
```

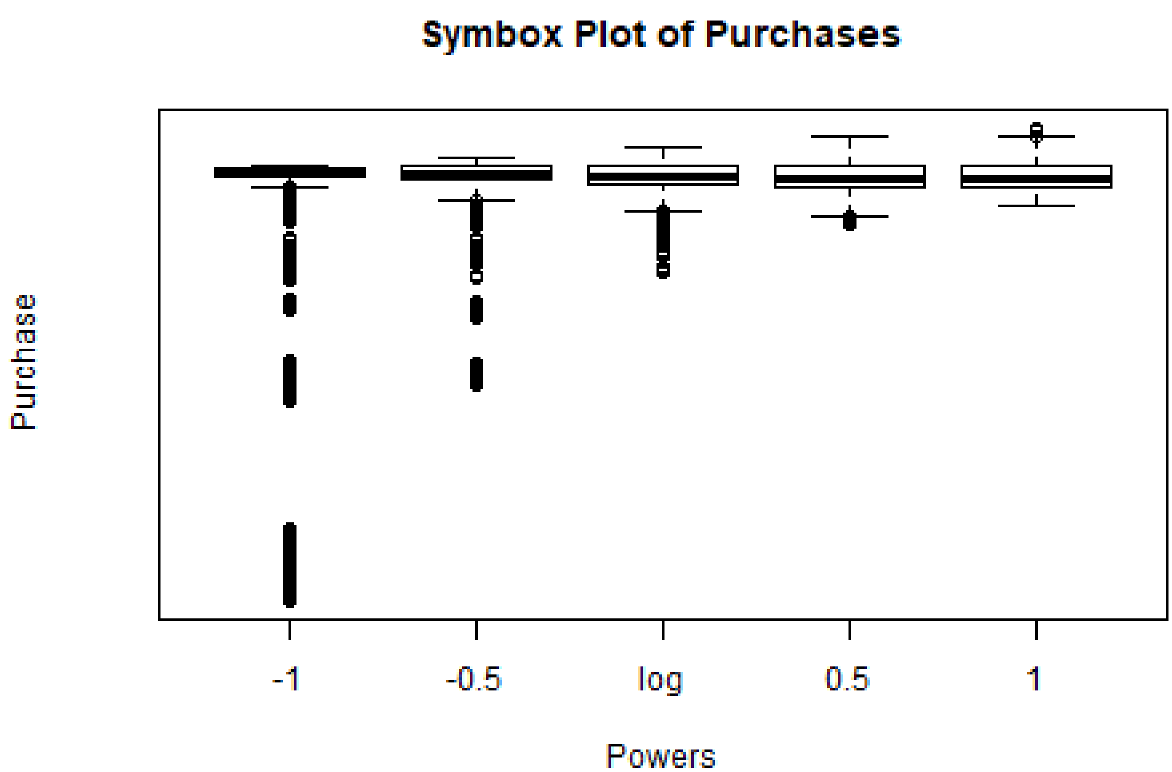
```
## Call: lm(formula = Purchase ~ Gender + Age + Occupation + I(Occupation^2)
##         + Gender:Age + City_Category + Marital_Status + Product_Category_1 +
##         multi, data = data)
##
## Coefficients:
##                     Estimate Std. Error  t value Pr(>|t|)
## (Intercept)        8875.2704    71.2743  124.523  < 2e-16 ***
## GenderM             666.1997    81.7707    8.147 3.73e-16 ***
## Age18-25            227.4473    73.4173    3.098 0.001948 **
## Age26-35            757.6103    70.3087   10.775  < 2e-16 ***
## Age36-45            927.8264    72.9123   12.725  < 2e-16 ***
## Age46-50            849.1295    79.3639   10.699  < 2e-16 ***
## Age51-55           1112.2854    82.5587   13.473  < 2e-16 ***
## Age55+             1024.1485    94.8131   10.802  < 2e-16 ***
## Occupation           61.5485     3.7526   16.402  < 2e-16 ***
## I(Occupation^2)      -2.9279     0.1923  -15.224  < 2e-16 ***
## City_CategoryB      157.2922    15.8831    9.903  < 2e-16 ***
## City_CategoryC      703.7400    17.1938   40.930  < 2e-16 ***
## Marital_Status      -45.3900    13.8553   -3.276 0.001053 **
## Product_Category_1 -354.7798     1.8835 -188.364  < 2e-16 ***
## multi              1139.4141    15.2235   74.846  < 2e-16 ***
## GenderM:Age18-25    224.4765    88.8835    2.526 0.011553 *
## GenderM:Age26-35   -205.2082    85.2244   -2.408 0.016047 *
## GenderM:Age36-45   -324.5758    88.2401   -3.678 0.000235 ***
## GenderM:Age46-50   -260.0536    95.4294   -2.725 0.006429 **
## GenderM:Age51-55   -169.7914    98.7500   -1.719 0.085541 .
## GenderM:Age55+     -322.6017   111.9247   -2.882 0.003948 **
```

```
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard deviation: 4684 on 537556 degrees of freedom
## Multiple R-squared: 0.1159
## F-statistic:  3524 on 20 and 537556 DF,  p-value: < 2.2e-16
##      AIC      BIC
## 10612600 10612846
```

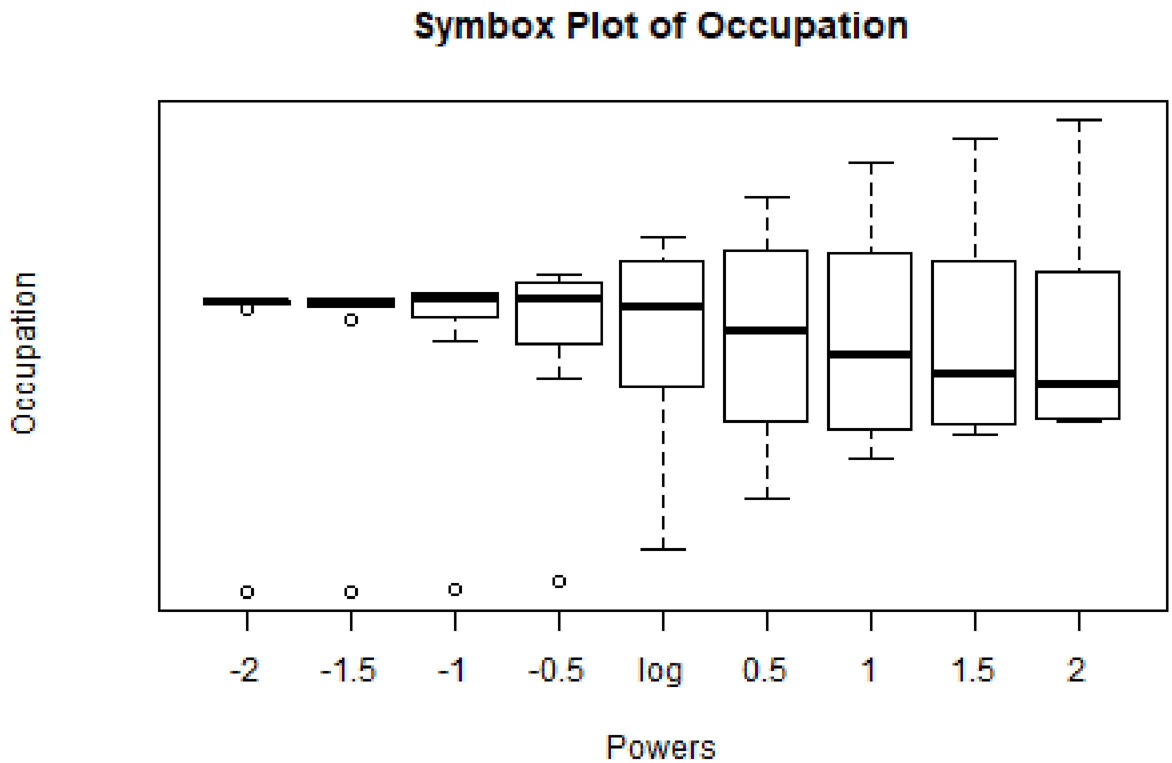
The Adjusted R-Squared increases after I add the interaction term. The variables significance do not change. However, the different interactions between age and gender have varying degrees of significance.

```r
#Box-Cox transformation test
powerTransform((Occupation+1)~1,data, family = "bcPower")
```

```
## Estimated transformation parameter
##        Y1
## 0.3784342
```




```
## Warning in symbox.default(Occupation, powers = c(-2, -1.5, -1, -0.5, 0, :
## start set to 0.2
```



The output above from the power transform function suggests that I should perform a square root transformation to the occupation variable (I rounded the transformation parameter to the nearest tenth). I can confirm the output of the power transform function by looking at the symbox transformation plot of the occupational variable. The symbox plot also suggests that I should perform a square root transformation.

```
#Model 8 (with sqrt(occupation) and gender:age terms)
mod.8 <- lm(Purchase ~ Gender + Age + I(sqrt(Occupation)) + Gender:Age +
              Marital_Status + Product_Category_1 + City_Category, data)
S(mod.8)
```

```
## Call: lm(formula = Purchase ~ Gender + Age + I(sqrt(Occupation)) +
##          Gender:Age + Marital_Status + Product_Category_1 + City_Category, data =
##          data)
##
## Coefficients:
##                        Estimate Std. Error  t value Pr(>|t|)
## (Intercept)           10128.975     69.439  145.868  < 2e-16 ***
## GenderM                 607.469     82.204    7.390 1.47e-13 ***
## Age18-25                131.060     73.627    1.780  0.07507 .
## Age26-35                654.290     70.489    9.282  < 2e-16 ***
## Age36-45                846.198     73.126   11.572  < 2e-16 ***
## Age46-50                747.610     79.563    9.396  < 2e-16 ***
## Age51-55               1014.595     82.855   12.245  < 2e-16 ***
## Age55+                  957.554     95.249   10.053  < 2e-16 ***
## I(sqrt(Occupation))      43.621      4.673    9.334  < 2e-16 ***
## Marital_Status          -56.063     13.926   -4.026 5.68e-05 ***
## Product_Category_1     -414.372      1.719 -241.096  < 2e-16 ***
```

```
## City_CategoryB        175.324     15.958   10.987  < 2e-16 ***
## City_CategoryC        748.749     17.256   43.391  < 2e-16 ***
## GenderM:Age18-25      278.463     89.347    3.117  0.00183 **
## GenderM:Age26-35     -149.657     85.675   -1.747  0.08067 .
## GenderM:Age36-45     -277.930     88.704   -3.133  0.00173 **
## GenderM:Age46-50     -181.617     95.916   -1.894  0.05829 .
## GenderM:Age51-55      -96.017     99.269   -0.967  0.33343
## GenderM:Age55+       -294.042    112.523   -2.613  0.00897 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard deviation: 4709 on 537558 degrees of freedom
## Multiple R-squared: 0.1064
## F-statistic:  3555 on 18 and 537558 DF,  p-value: < 2.2e-16
##      AIC      BIC
## 10618368 10618592
```

After adding the square root occupation term I can see that the significance for the Age18-25, GenderM:Age26-35, GenderM:Age46-50, and GenderM:Age51-55 variables dropped, when compared to model 7. the GenderM:Age51-55 variable became completely insignificant, while the other terms dropped to a significance level of 10%. The R_squared for model 8 is also lower than 7, which suggests that the model is worse at explaining the variation in the data. In order to confirm which model is better I looked at the BIC and AIC for each regression.

```r
#comparing AIC and BIC of Model 7 and 8
AIC(mod.8, mod.7)
```

```
##       df      AIC
## mod.8 20 10618368
## mod.7 22 10612600
```

```r
BIC(mod.8, mod.7)
```

```
##       df      BIC
## mod.8 20 10618592
## mod.7 22 10612846
```

From the output above I can see that model 7 has a lower AIC and BIC compared to model 8, which includes the square root occupation term. Since AIC and BIC are used for model comparison I can conclude that model 7 fits and explains the data better than model 8.

```r
#Variance Inflation Factor for Model 7
vif(mod.7)
```

```
##            GenderM           Age18-25           Age26-35
##            30.3870            19.6340            29.0600
##           Age36-45           Age46-50           Age51-55
##            20.8430            11.7260            10.8710
##             Age55+         Occupation    I(Occupation^2)
##             8.2332            14.6890            14.6230
##     City_CategoryB     City_CategoryC     Marital_Status
##             1.5073             1.5486             1.1370
## Product_Category_1              multi   GenderM:Age18-25
##             1.2230             1.2162            22.8720
##   GenderM:Age26-35   GenderM:Age36-45   GenderM:Age46-50
##            37.9090            24.4390            12.3730
##   GenderM:Age51-55     GenderM:Age55+
```

```
##           11.7930            8.8515
```

Evaluating the variance inflating factor suggests there is a high degree of colinearity between most of our variables. However, this is to be expected. Since our variables are almost entirely categorical, there is not a wide range of variation and therefore little opportunity for the values of each variable to not be colinear. While I will not use these results to change our model decision, it is worth noting for the sake of completeness.

```
#Creating Stepwise model
step.model <- stepAIC(mod.2, direction = "backward", trace = FALSE)
summary(step.model)
```

```
##
## Call:
## lm(formula = Purchase ~ Gender + Age + Occupation + City_Category +
##     Marital_Status + Product_Category_1 + multi, data = data)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -10403.1  -3152.7   -717.7   2423.4  17913.0
##
## Coefficients:
##                    Estimate Std. Error  t value Pr(>|t|)
## (Intercept)        9164.473     45.747  200.330  < 2e-16 ***
## GenderM             523.371     14.983   34.932  < 2e-16 ***
## Age18-25            348.786     41.709    8.362  < 2e-16 ***
## Age26-35            543.423     40.500   13.418  < 2e-16 ***
## Age36-45            633.682     41.641   15.218  < 2e-16 ***
## Age46-50            605.493     45.716   13.245  < 2e-16 ***
## Age51-55            933.942     46.696   20.001  < 2e-16 ***
## Age55+              738.935     51.273   14.412  < 2e-16 ***
## Occupation            6.151      0.994    6.188 6.11e-10 ***
## City_CategoryB      163.903     15.859   10.335  < 2e-16 ***
## City_CategoryC      719.382     17.154   41.937  < 2e-16 ***
## Marital_Status      -53.926     13.846   -3.895 9.83e-05 ***
## Product_Category_1 -355.386      1.884 -188.655  < 2e-16 ***
## multi              1140.011     15.228   74.863  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 4685 on 537563 degrees of freedom
## Multiple R-squared:  0.1153, Adjusted R-squared:  0.1152
## F-statistic:  5387 on 13 and 537563 DF,  p-value: < 2.2e-16
```

When I do a backwards selection on the baseline model (only additive terms), the results suggest to use all the additive terms. These results will be used to compare the results of applying the same process to our selected model.

```
#Creating Stepwise Model
step.model2 <- stepAIC(mod.7, direction = "backward", trace = FALSE)
summary(step.model2)
```

```
##
## Call:
## lm(formula = Purchase ~ Gender + Age + Occupation + I(Occupation^2) +
##     Gender:Age + City_Category + Marital_Status + Product_Category_1 +
##     multi, data = data)
##
```

```
## Residuals:
##      Min       1Q   Median       3Q      Max
## -10409.4  -3154.2   -702.8   2426.3  18160.2
##
## Coefficients:
##                      Estimate Std. Error  t value Pr(>|t|)
## (Intercept)         8875.2704    71.2743  124.523  < 2e-16 ***
## GenderM              666.1997    81.7707    8.147 3.73e-16 ***
## Age18-25             227.4473    73.4173    3.098 0.001948 **
## Age26-35             757.6103    70.3087   10.775  < 2e-16 ***
## Age36-45             927.8264    72.9123   12.725  < 2e-16 ***
## Age46-50             849.1295    79.3639   10.699  < 2e-16 ***
## Age51-55            1112.2854    82.5587   13.473  < 2e-16 ***
## Age55+              1024.1485    94.8131   10.802  < 2e-16 ***
## Occupation            61.5485     3.7526   16.402  < 2e-16 ***
## I(Occupation^2)       -2.9279     0.1923  -15.224  < 2e-16 ***
## City_CategoryB       157.2922    15.8831    9.903  < 2e-16 ***
## City_CategoryC       703.7400    17.1938   40.930  < 2e-16 ***
## Marital_Status       -45.3900    13.8553   -3.276 0.001053 **
## Product_Category_1  -354.7798     1.8835 -188.364  < 2e-16 ***
## multi               1139.4141    15.2235   74.846  < 2e-16 ***
## GenderM:Age18-25     224.4765    88.8835    2.526 0.011553 *
## GenderM:Age26-35    -205.2082    85.2244   -2.408 0.016047 *
## GenderM:Age36-45    -324.5758    88.2401   -3.678 0.000235 ***
## GenderM:Age46-50    -260.0536    95.4294   -2.725 0.006429 **
## GenderM:Age51-55    -169.7914    98.7500   -1.719 0.085541 .
## GenderM:Age55+      -322.6017   111.9247   -2.882 0.003948 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 4684 on 537556 degrees of freedom
## Multiple R-squared:  0.1159, Adjusted R-squared:  0.1159
## F-statistic:  3524 on 20 and 537556 DF,  p-value: < 2.2e-16
```

In performing the backward selection, I see that all but the interaction between gender and the 51 to 55 age range and also the 18 to 25 age range are significant. Next I will repeat the process using a stepwise selection process.

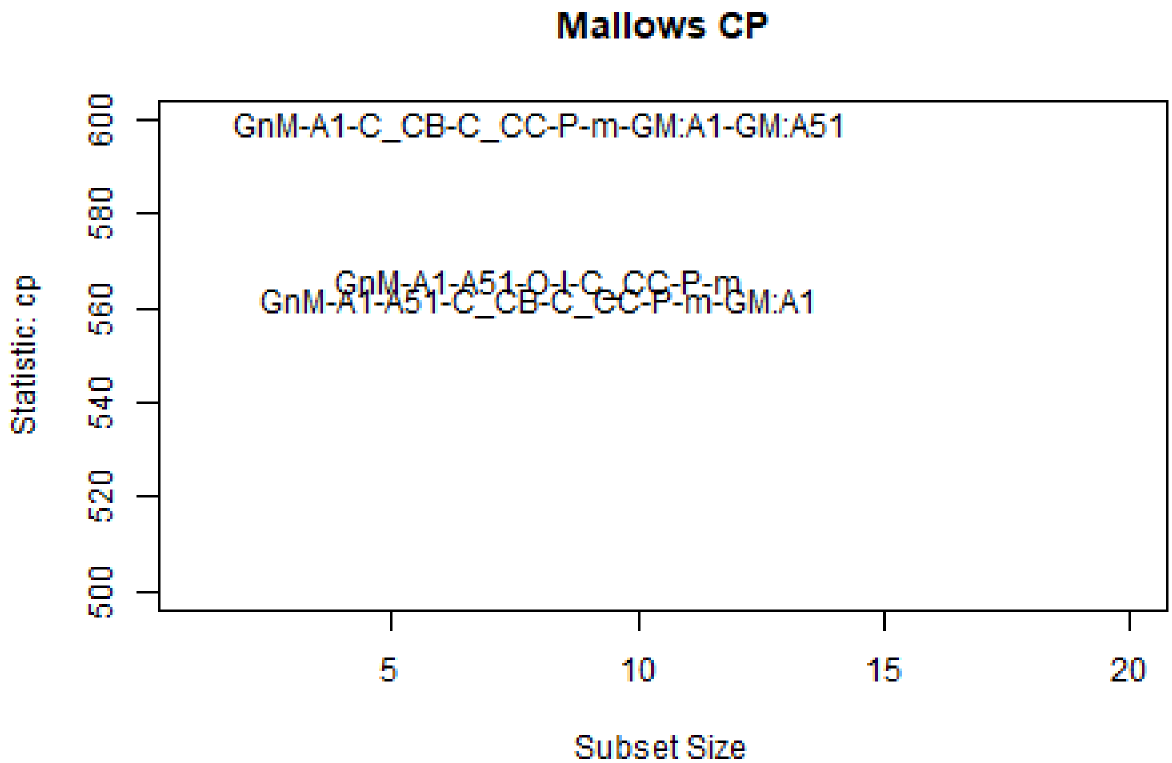
```r
#Creating Stepwise Model
step.model3 <- stepAIC(mod.7, direction = "both", trace = FALSE)
summary(step.model3)
```

```
##
## Call:
## lm(formula = Purchase ~ Gender + Age + Occupation + I(Occupation^2) +
##     Gender:Age + City_Category + Marital_Status + Product_Category_1 +
##     multi, data = data)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -10409.4  -3154.2   -702.8   2426.3  18160.2
##
## Coefficients:
##                      Estimate Std. Error  t value Pr(>|t|)
## (Intercept)         8875.2704    71.2743  124.523  < 2e-16 ***
```

```
## GenderM              666.1997    81.7707    8.147 3.73e-16 ***
## Age18-25             227.4473    73.4173    3.098 0.001948 **
## Age26-35             757.6103    70.3087   10.775  < 2e-16 ***
## Age36-45             927.8264    72.9123   12.725  < 2e-16 ***
## Age46-50             849.1295    79.3639   10.699  < 2e-16 ***
## Age51-55            1112.2854    82.5587   13.473  < 2e-16 ***
## Age55+              1024.1485    94.8131   10.802  < 2e-16 ***
## Occupation            61.5485     3.7526   16.402  < 2e-16 ***
## I(Occupation^2)       -2.9279     0.1923  -15.224  < 2e-16 ***
## City_CategoryB       157.2922    15.8831    9.903  < 2e-16 ***
## City_CategoryC       703.7400    17.1938   40.930  < 2e-16 ***
## Marital_Status       -45.3900    13.8553   -3.276 0.001053 **
## Product_Category_1 -354.7798     1.8835 -188.364  < 2e-16 ***
## multi               1139.4141    15.2235   74.846  < 2e-16 ***
## GenderM:Age18-25     224.4765    88.8835    2.526 0.011553 *
## GenderM:Age26-35    -205.2082    85.2244   -2.408 0.016047 *
## GenderM:Age36-45    -324.5758    88.2401   -3.678 0.000235 ***
## GenderM:Age46-50    -260.0536    95.4294   -2.725 0.006429 **
## GenderM:Age51-55    -169.7914    98.7500   -1.719 0.085541 .
## GenderM:Age55+      -322.6017   111.9247   -2.882 0.003948 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 4684 on 537556 degrees of freedom
## Multiple R-squared:  0.1159, Adjusted R-squared:  0.1159
## F-statistic:  3524 on 20 and 537556 DF,  p-value: < 2.2e-16
```

In terms of significance, the results of backward selection and stepwise selection on our model are identical. The only terms that may possibly be desirable to remove is the previously mentioned interaction term and the 18 to 25 age range. I will continue the model evaluation with a Mallow CP.

```
#Mallow CP
ss=regsubsets(Purchase ~ Gender + Age + Occupation + I(Occupation^2) + City_Category + Gender:Age +
                Marital_Status + Product_Category_1 + multi,method=c("exhaustive"),nbest=3,data=data)
subsets(ss,statistic="cp",legend=F,main="Mallows CP",col="steelblue4", ylim = c(500,600))
```




```
##                        Abbreviation
## GenderM                         GnM
## Age18-25                         A1
## Age26-35                         A2
## Age36-45                         A3
## Age46-50                         A4
## Age51-55                        A51
## Age55+                          A55
## Occupation                        O
## I(Occupation^2)                   I
## City_CategoryB                 C_CB
## City_CategoryC                 C_CC
## Marital_Status                    M
## Product_Category_1                P
## multi                             m
## GenderM:Age18-25              GM:A1
## GenderM:Age26-35              GM:A2
## GenderM:Age36-45              GM:A3
## GenderM:Age46-50              GM:A4
## GenderM:Age51-55             GM:A51
## GenderM:Age55+               GM:A55
```

```
data$A1 = 0
data[(data$Age=="18-25"), "A1"] = 1

data$A51 = 0
data[(data$Age=="51-55"), "A51"] = 1
```

```r
data$City_C = 0
data[data$City_C == "C", "City_C"] = 1

test1 = lm(Purchase ~ Gender + A1 + A51 + Occupation + I(Occupation^2) + City_C + Product_Category_1 + m
test2 = lm(Purchase ~ Gender + A1 + A51 + City_Category + Product_Category_1 + multi + Gender:A1, data)

AIC(test1, mod.7, test2)
```

```
##       df      AIC
## test1  9 10615065
## mod.7 22 10612600
## test2 10 10613140
```

```r
BIC(test1, mod.7, test2)
```

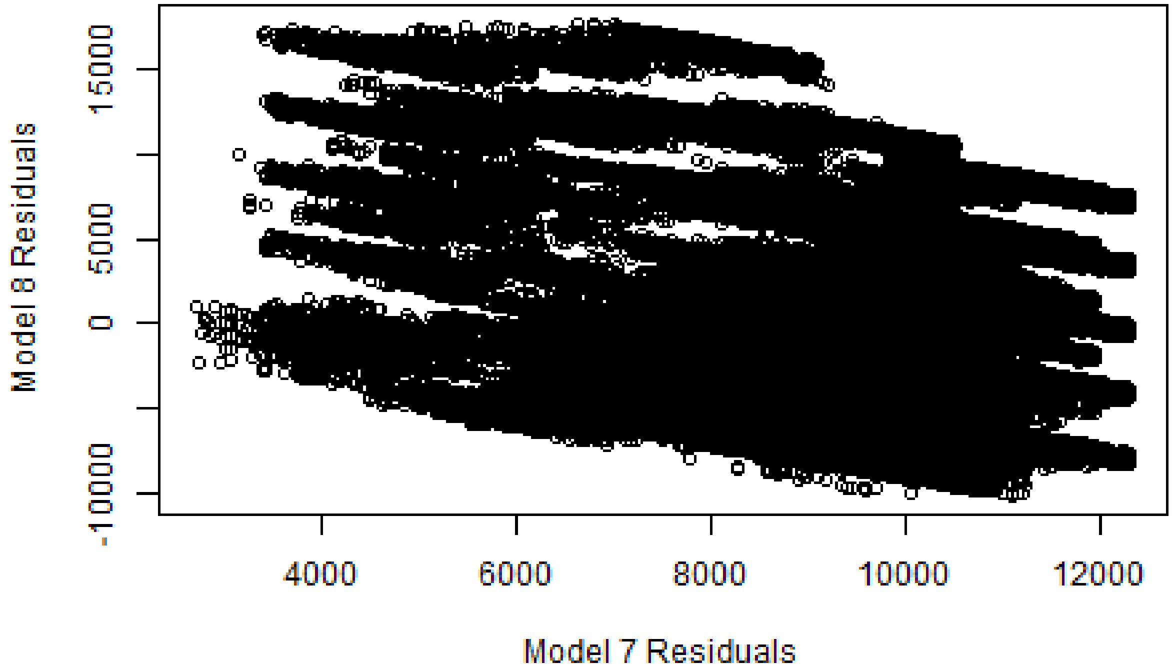
```
##       df      BIC
## test1  9 10615166
## mod.7 22 10612846
## test2 10 10613251
```

For the Mallow CP, the resulting graph has been rescaled to exclude all the results that are well above the models of interest so as to see the suggested models more clearly. Interestingly enough, when I compare the two best suggestions from Mallow C with our original model using AIC and BIC, our original model scores better. As a result, I will use the following model:

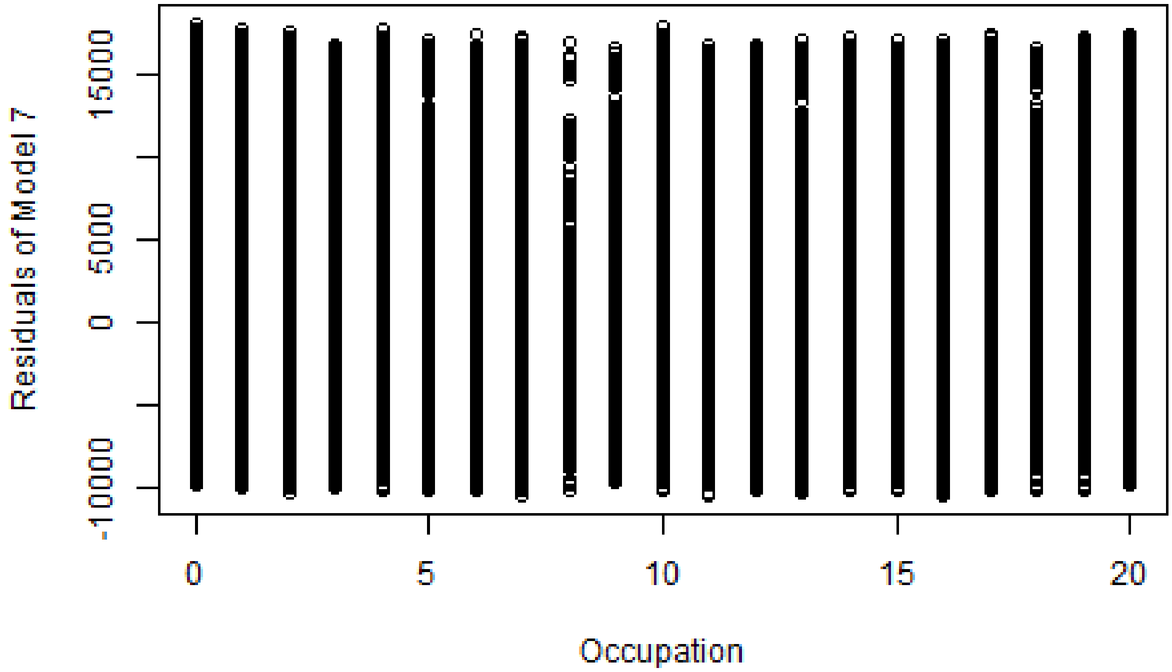
$$Purchases = \beta_o + \beta_1 Male + \beta_2 Occupation + \beta_3 Occupation^2 + \beta_4 MaritalStatus + \beta_5 Multi + \beta_6 Age + \beta_7 Male \times AGE + \beta_8 City$$

Now that I have come to our model, I will plot the residuals, plot the fitted values, and perform a 5-fold cross-validation as a robustness check.

```r
#Plot of Residuals from models
plot(mod.7$fitted.values, mod.8$residuals, ylab="Model 8 Residuals", xlab="Model 7 Residuals")
```




```r
#Plot of occupation variable and Model 7 residuals
plot(Occupation, mod.7$residuals, ylab="Residuals of Model 7")
```



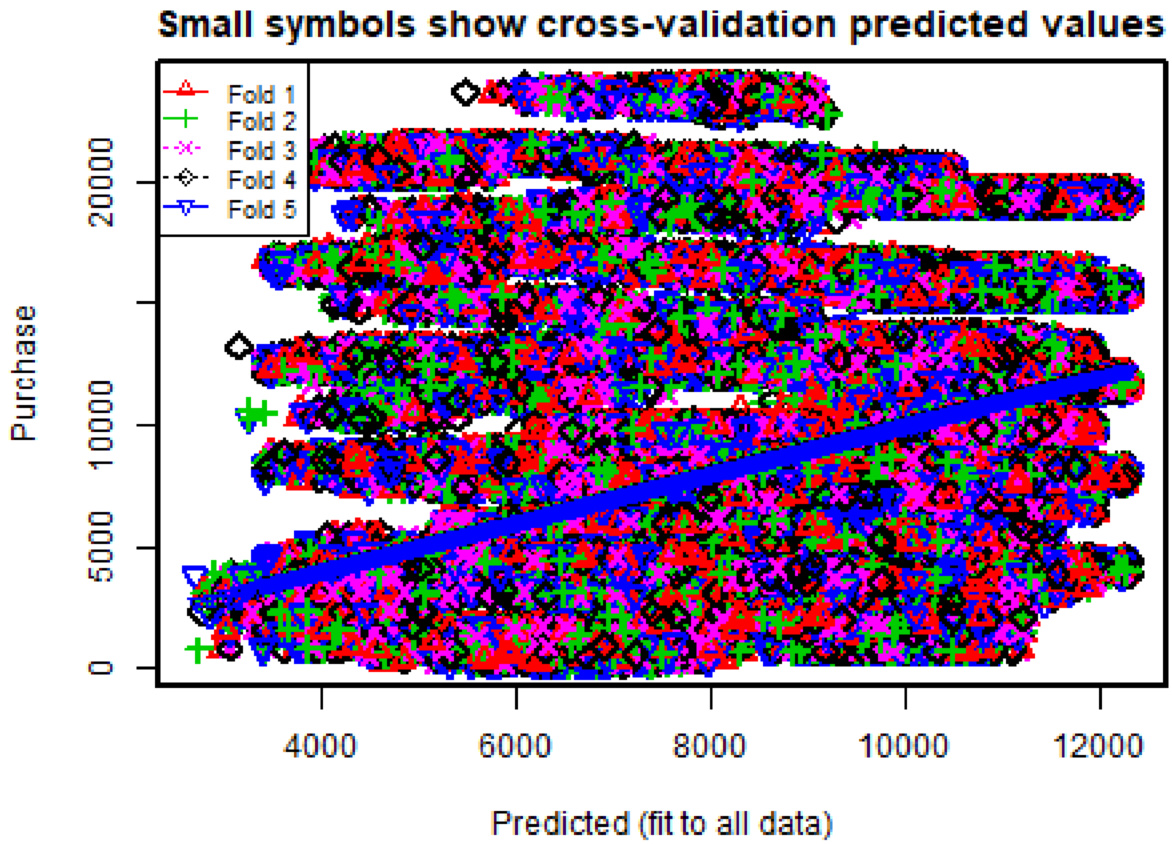

```
#Model 7 is preferred
```

The residual plot and fitted values plot do not seem to show us very much. This is likely because most of the values are categorical. I can say that the residual plot seems to suggest that the errors are random about zero, which is good. I now move on to cross-validation.

It is important to note that I did not test for heteroskedasticity since the majority of our data is categorical.

```
cv.lm(data = data, form.lm = mod.7, m = 5, plotit = TRUE, printit = FALSE)
```

```
## Warning in cv.lm(data = data, form.lm = mod.7, m = 5, plotit = TRUE, printit = FALSE):
##
##  As there is >1 explanatory variable, cross-validation
##  predicted values for a fold are not a linear function
##  of corresponding overall predicted values.  Lines that
##  are shown for the different folds are approximate
```



From the 5-fold cross validation test, I can see that there is a high level of agreement across the different folds. This indicates that our model does reasonably well at predicting purchases for out of model observations. The five-fold cross validation test seaperates the data into five segments and uses 4/5 of those segments to predict the other 1/5.

```
summary(mod.7)
```

```
##
## Call:
## lm(formula = Purchase ~ Gender + Age + Occupation + I(Occupation^2) +
##     Gender:Age + City_Category + Marital_Status + Product_Category_1 +
##     multi, data = data)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -10409.4  -3154.2   -702.8   2426.3  18160.2
##
## Coefficients:
##                     Estimate Std. Error  t value Pr(>|t|)
## (Intercept)        8875.2704    71.2743  124.523  < 2e-16 ***
## GenderM             666.1997    81.7707    8.147 3.73e-16 ***
## Age18-25            227.4473    73.4173    3.098 0.001948 **
## Age26-35            757.6103    70.3087   10.775  < 2e-16 ***
## Age36-45            927.8264    72.9123   12.725  < 2e-16 ***
## Age46-50            849.1295    79.3639   10.699  < 2e-16 ***
## Age51-55           1112.2854    82.5587   13.473  < 2e-16 ***
## Age55+             1024.1485    94.8131   10.802  < 2e-16 ***
```

```
## Occupation           61.5485     3.7526   16.402  < 2e-16 ***
## I(Occupation^2)      -2.9279     0.1923  -15.224  < 2e-16 ***
## City_CategoryB      157.2922    15.8831    9.903  < 2e-16 ***
## City_CategoryC      703.7400    17.1938   40.930  < 2e-16 ***
## Marital_Status      -45.3900    13.8553   -3.276 0.001053 **
## Product_Category_1 -354.7798     1.8835 -188.364  < 2e-16 ***
## multi              1139.4141    15.2235   74.846  < 2e-16 ***
## GenderM:Age18-25    224.4765    88.8835    2.526 0.011553 *
## GenderM:Age26-35   -205.2082    85.2244   -2.408 0.016047 *
## GenderM:Age36-45   -324.5758    88.2401   -3.678 0.000235 ***
## GenderM:Age46-50   -260.0536    95.4294   -2.725 0.006429 **
## GenderM:Age51-55   -169.7914    98.7500   -1.719 0.085541 .
## GenderM:Age55+     -322.6017   111.9247   -2.882 0.003948 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 4684 on 537556 degrees of freedom
## Multiple R-squared:  0.1159, Adjusted R-squared:  0.1159
## F-statistic:  3524 on 20 and 537556 DF,  p-value: < 2.2e-16
```

Reviewing the results of our model, I see that males have consistently higher purchases relative to females across all age groups. Also, while purchases generally increase with age, there are dips in purchases for age group 46 to 50 and group 55 and older. Also, for each increase in the occupation value I see a 25.88 reduction in purchases plus 157.75 for each additional occupation value, on average. For city categories, I see that city category B and C are associated with 165.14 and 716.71 more purchases respectively, all else equal. Being married is associated with a 47.87 reduction in purchases, all else constant. For product category, an increase of the product category number by 1 is associated with a 355.1 reduction in purchases, all else constant. A product having multiple categories, however, is associated with a 1140.35 increase in purchases, all else constant.