Customer Profiling Analysis

Lucas Lafferty and Greyson Brower

12/4/2021

Contents

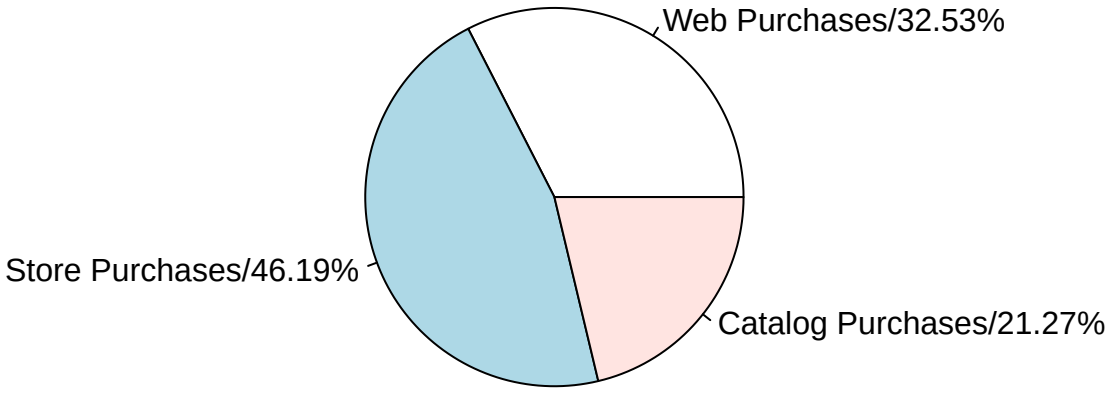
Intro to Our Regressions

We used a dataset that had 2240 observations across 29 different variables, with each variable pertaining to a customer's demographic or shopping habits. The demographic variables included income, education level, marital status, number of kids at home, number of teenagers at home, their age, and when they became a customer. The variables pertaining to shopping habits included whether or not they have complained, how recent their last purchase was, the amount they spent on wine, meat, fish, fruits, sweets, and gold, which promotional campaigns they accepted offers in, the number of purchases they made with a discount, the number of purchases made in store, on the website, or from a catalog, and the number of times they visited our website in a given month. By using this dataset, we were able to create regressions that would give us information pertaining to what demographic a user was a part of and how they purchase certain products.

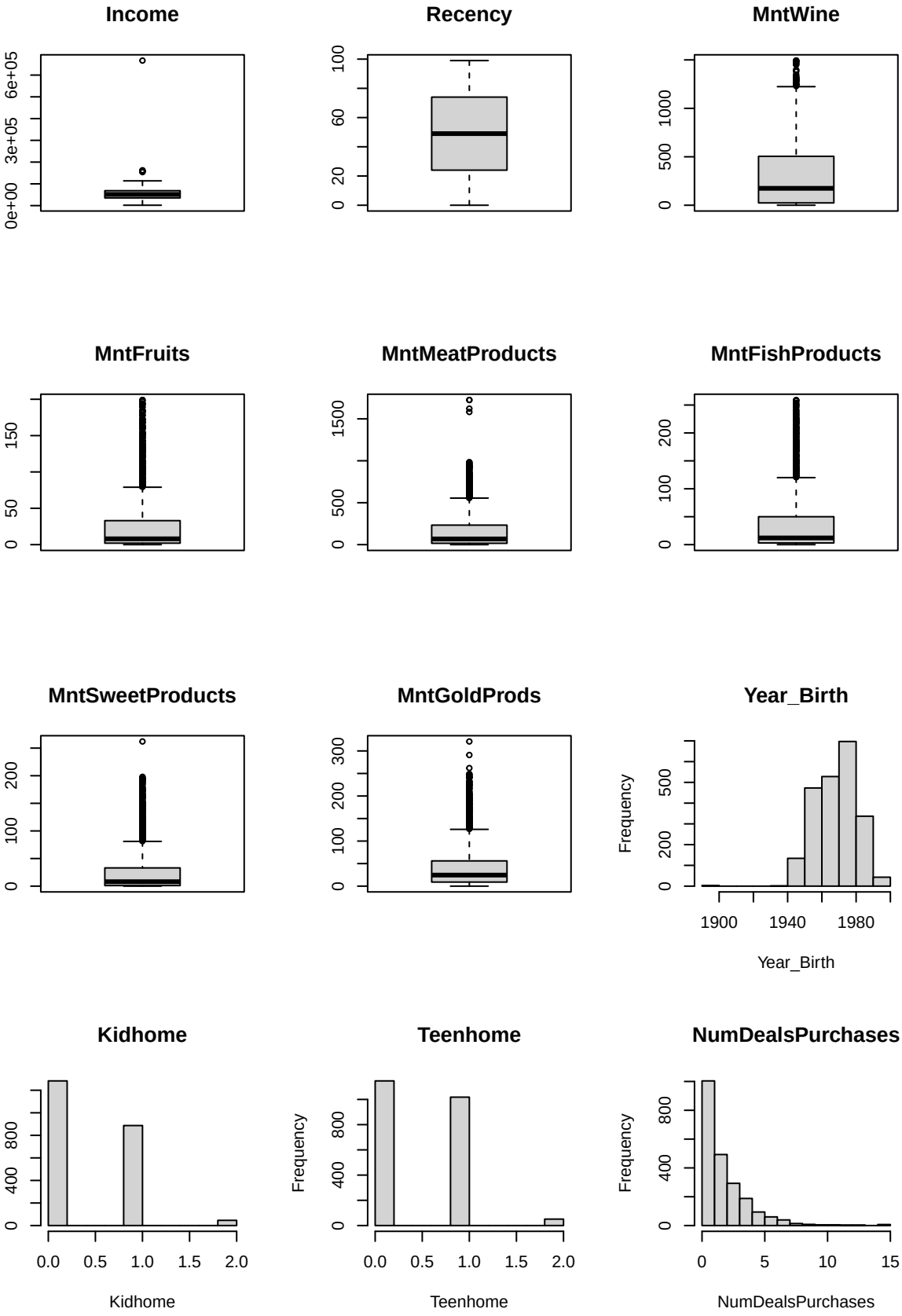
Before beginning any models, we had to explore and clean the data set. There were a few observations that were either data entry errors or purposefully misleading information, which were originally skewing our results.

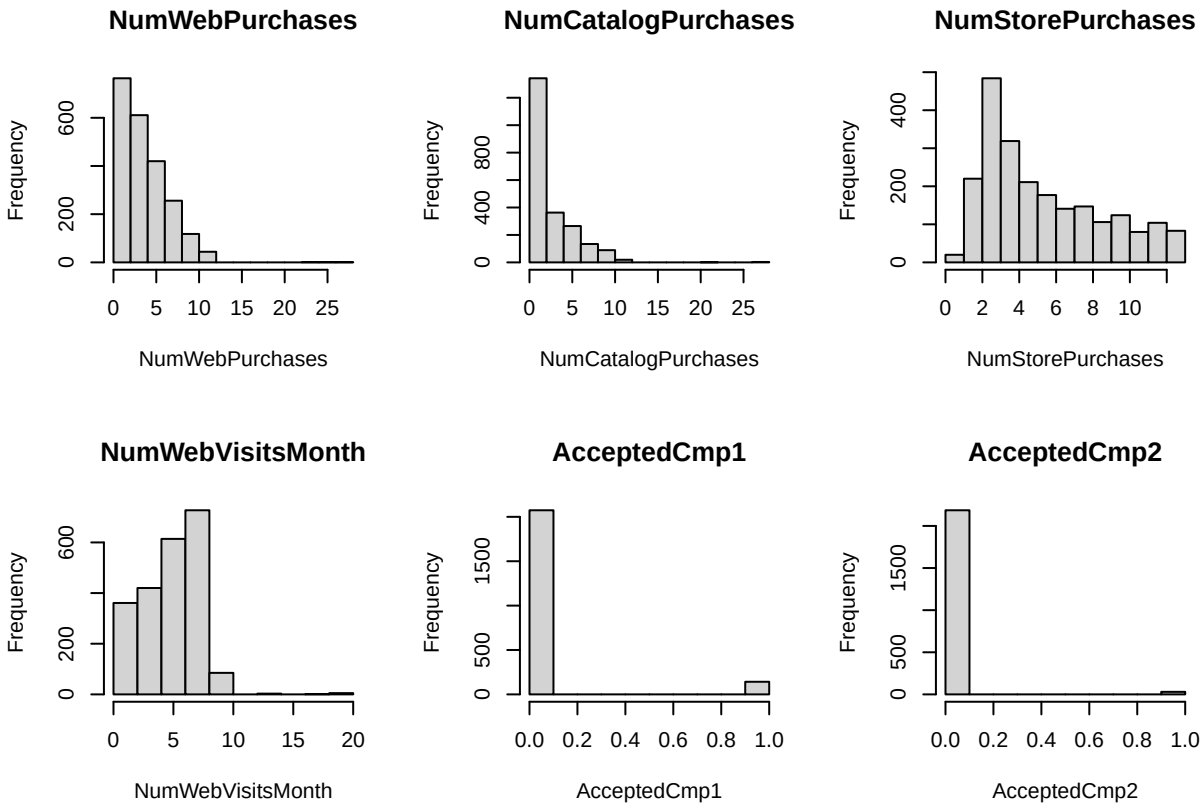
Descriptive Analytics

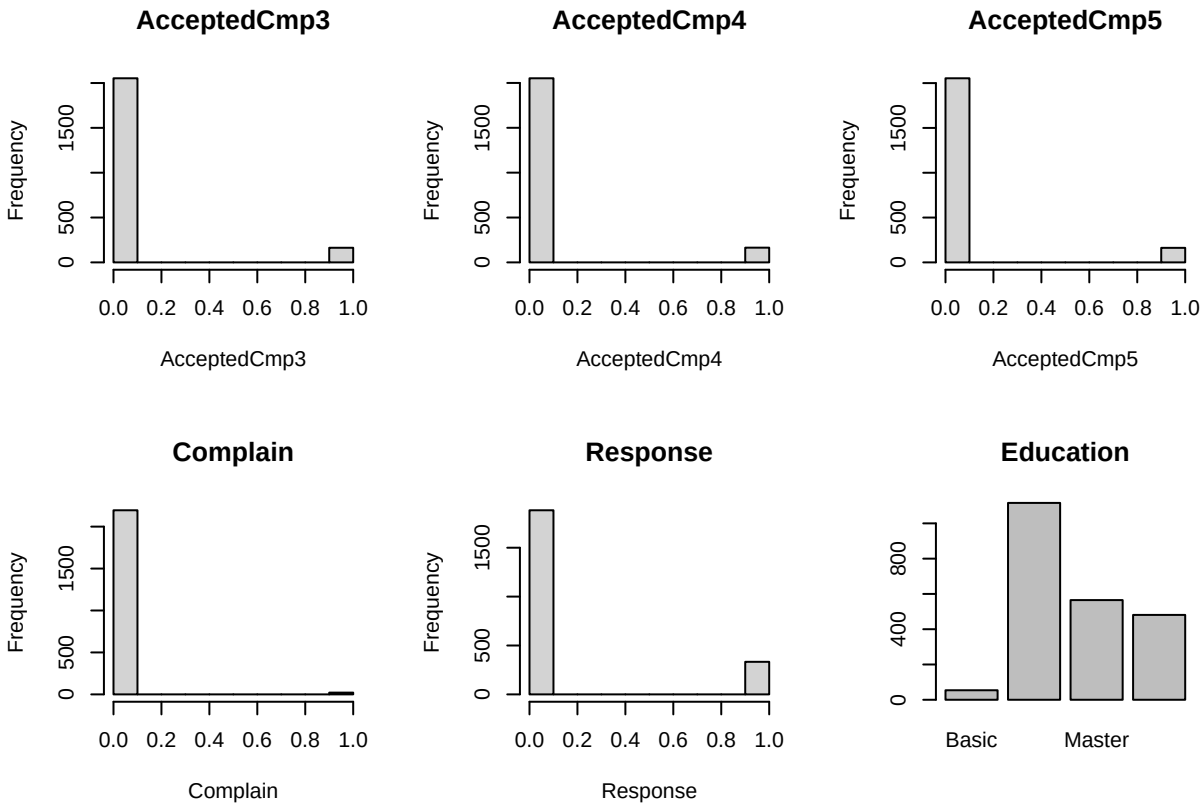
Creating Pie Chart to see Which Method of Purchase Would be Best to Predict

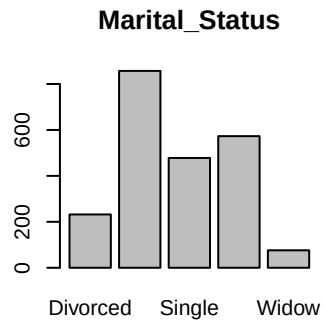


Boxplots, Histograms and Barcharts for Analysis of Data Distributions




NumWebPurchases
Frequency
0
200
600
0 5 10 15 20 25
NumWebPurchases
NumCatalogPurchases
Frequency
0
400
800
0 5 10 15 20 25
NumCatalogPurchases
NumStorePurchases
Frequency
0
200
400
0 2 4 6 8 10
NumStorePurchases
NumWebVisitsMonth
Frequency
0
200
600
0 5 10 15 20
NumWebVisitsMonth
AcceptedCmp1
Frequency
0
500
1500
0.0 0.2 0.4 0.6 0.8 1.0
AcceptedCmp1
AcceptedCmp2
Frequency
0
500
1500
0.0 0.2 0.4 0.6 0.8 1.0
AcceptedCmp2


AcceptedCmp3
Frequency
0
500
1500
0.0
0.2
0.4
0.6
0.8
1.0
AcceptedCmp3
AcceptedCmp4
Frequency
0
500
1500
0.0
0.2
0.4
0.6
0.8
1.0
AcceptedCmp4
AcceptedCmp5
Frequency
0
500
1500
0.0
0.2
0.4
0.6
0.8
1.0
AcceptedCmp5
Complain
Frequency
0
500
1500
0.0
0.2
0.4
0.6
0.8
1.0
Complain
Response
Frequency
0
500
1500
0.0
0.2
0.4
0.6
0.8
1.0
Response
Education
0
400
800
Basic
Master



Cleaning and Manipulating Data

The first observation we removed claimed to have an income of $666,666, more than $500,000 more than the next highest. There were other observations which claimed to have a date of birth of 1900 or earlier, which we also decided to remove. Finally, many data points contained derivative information in-place of marital status and education level, such as "Alone" instead of "Single" for marital status and "2n Cycle" instead of "Masters" for education level. We adjusted those values to their proper category and continued.

Correlation Matrix to check for any highly correlated Predictors

```
##                   MntWines MntFishProducts MntFruits MntGoldProds
## MntWines         1.0000000       0.3969151 0.3858442    0.3914611
## MntFishProducts  0.3969151       1.0000000 0.5930383    0.4262987
## MntFruits        0.3858442       0.5930383 1.0000000    0.3934594
## MntGoldProds     0.3914611       0.4262987 0.3934594    1.0000000
## MntSweetProducts 0.3895829       0.5834842 0.5714738    0.3567538
## MntMeatProducts  0.5680808       0.5729861 0.5467400    0.3575558
##                  MntSweetProducts MntMeatProducts
## MntWines                0.3895829       0.5680808
## MntFishProducts         0.5834842       0.5729861
## MntFruits               0.5714738       0.5467400
## MntGoldProds            0.3567538       0.3575558
## MntSweetProducts        1.0000000       0.5346240
## MntMeatProducts         0.5346240       1.0000000
```

```
##                       NumStorePurchases NumWebPurchases NumCatalogPurchases
## NumStorePurchases             1.0000000       0.5157556           0.5178870
## NumWebPurchases               0.5157556       1.0000000           0.3865388
## NumCatalogPurchases           0.5178870       0.3865388           1.0000000
```

After reviewing each of these predictors correlations, we are unable to see any highly correlated values that would lead to further action on our models.

First Question: Can we create a profile of a Customer that will Shop in-store as to Advertise the Items they purchase and increase those sales?

Creating a Multivariable Logistic Regression

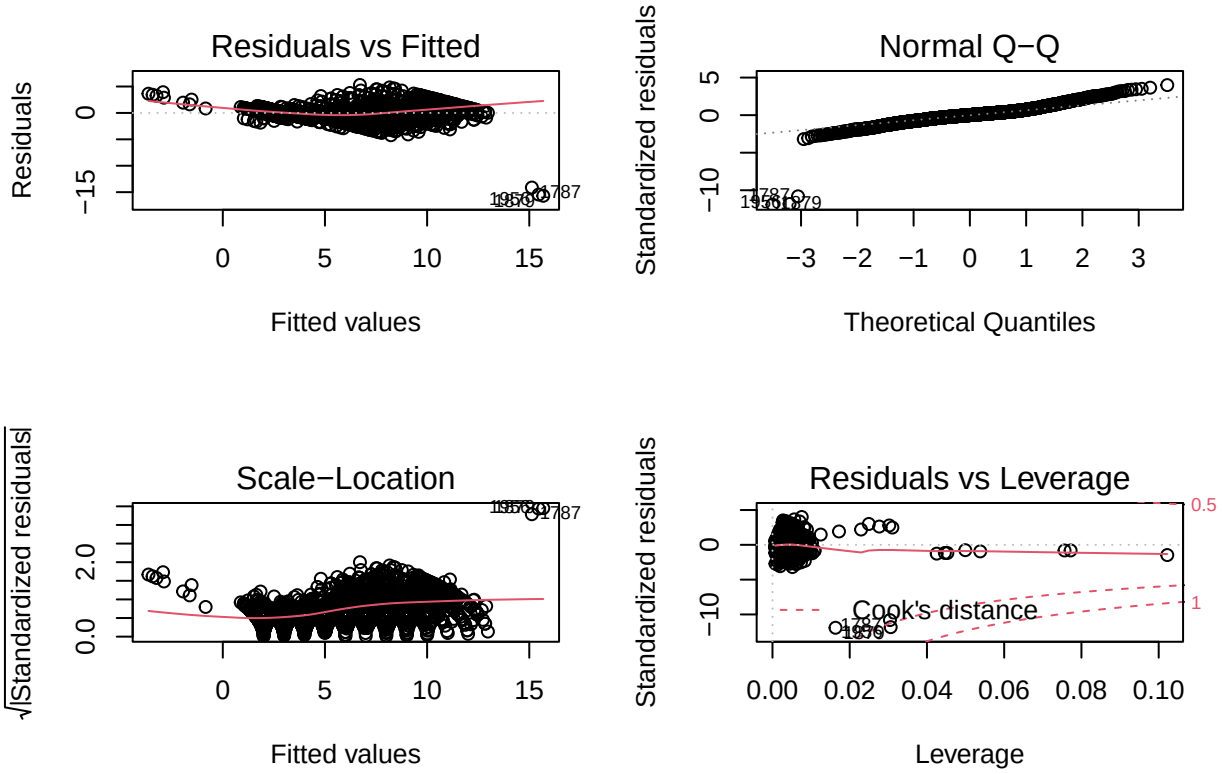
```
summary(nsp_step)
##
## Call:
## lm(formula = final_data$NumStorePurchases ~ final_data$Year_Birth +
##     final_data$NumCatalogPurchases + final_data$Income + final_data$Dt_Customer +
##     final_data$Response + final_data$NumTotalPurchases + final_data$NumWebVisitsMonth,
##     data = final_data)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -15.6853  -0.6137   0.0108   0.5674   5.2879
##
## Coefficients:
##                                  Estimate Std. Error t value Pr(>|t|)
## (Intercept)                    -3.197e+00  5.526e+00  -0.579 0.562973
## final_data$Year_Birth           7.612e-03  2.471e-03   3.080 0.002092 **
## final_data$NumCatalogPurchases -5.522e-01  1.699e-02 -32.492  < 2e-16 ***
## final_data$Income              -4.761e-06  2.386e-06  -1.995 0.046154 *
## final_data$Dt_Customer         -5.875e-04  1.556e-04  -3.776 0.000164 ***
## final_data$Response            -4.677e-01  8.311e-02  -5.627 2.06e-08 ***
## final_data$NumTotalPurchases    5.349e-01  7.374e-03  72.535  < 2e-16 ***
## final_data$NumWebVisitsMonth   -2.957e-01  1.693e-02 -17.468  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.327 on 2204 degrees of freedom
## Multiple R-squared:  0.8338, Adjusted R-squared:  0.8333
## F-statistic:  1580 on 7 and 2204 DF,  p-value: < 2.2e-16
```

We began by creating a full model, on which we will perform a stepwise function to determine which predictors are the most significant and then continue fitting the model.

Summarizing the Multivariable Linear Regression Model

R-squared value of 0.833, with all predictors being significant and the model being deemed extremely significant with an F-statistic of 1580.

Plotting the Multivariable Linear Regression Model's Statistics:



Residuals vs Fitted: The plot has a normal scatter, with a small cluster existing underneath the rest of the points.

Normal Q-Q: A majority of the points are within margin to the line defining normal distribution, except for a few at the bottom. You can note that they are the same points as the ones found in the Residuals vs Fitted Model.

Scale-Location: A majority of the points are close to the line, with a few (also a part of the original cluster found in Residuals vs Fitted) leveraging the model's line and range up.

Residuals vs Leverage: This final plot also shows the same 3 points from the original 3 model fit statistic plots being deemed outliers, as they are far from the normal cluster of points and pull the line for cooks distance down.

Adjustments Made From Plot Statistics:

We found from plotting the summary statistics of this model that there were 3 observations that could be labeled as outliers, as they had appeared on every one of the charts away from the normal spreads and fits that R had deemed to be the best. These 3 observations were: 1787, 1956 and 1879.

```
##
## Call:
## lm(formula = nsp_data$NumStorePurchases ~ nsp_data$Year_Birth +
##     nsp_data$NumCatalogPurchases + nsp_data$Income + nsp_data$Dt_Customer +
##     nsp_data$Response + nsp_data$NumTotalPurchases + nsp_data$NumWebVisitsMonth,
##     data = nsp_data)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -4.3341 -0.6261  0.0030  0.5747  5.3194
##
## Coefficients:
##                                Estimate Std. Error t value Pr(>|t|)
## (Intercept)                  -1.665e+00  4.990e+00  -0.334 0.738634
## nsp_data$Year_Birth           6.483e-03  2.232e-03   2.904 0.003715 **
## nsp_data$NumCatalogPurchases -6.018e-01  1.551e-02 -38.813  < 2e-16 ***
## nsp_data$Income              -1.151e-05  2.202e-06  -5.226 1.89e-07 ***
## nsp_data$Dt_Customer         -5.195e-04  1.406e-04  -3.696 0.000224 ***
## nsp_data$Response            -4.259e-01  7.507e-02  -5.674 1.58e-08 ***
## nsp_data$NumTotalPurchases    5.639e-01  6.799e-03  82.946  < 2e-16 ***
## nsp_data$NumWebVisitsMonth   -3.439e-01  1.549e-02 -22.193  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.199 on 2201 degrees of freedom
## Multiple R-squared:  0.8642, Adjusted R-squared:  0.8637
## F-statistic:  2000 on 7 and 2201 DF,  p-value: < 2.2e-16
```

Summarizing the Fitted Multivariable Linear Regression Model

R-squared value of 0.8637, which increased from 0.833, maintained the 7 significant predictors, and increased the models F-statistic from 1580 to 2000.

Answering the First Question:

After modeling and fitting this model there is useful information to take from this multivariable regression. In reference to those who make purchases in store (our dependent variable), we are able to infer that a younger population, with a lower income, a lower amount of visits to this company's website and a high total number of purchases will be likely to make the most in-store purchases.

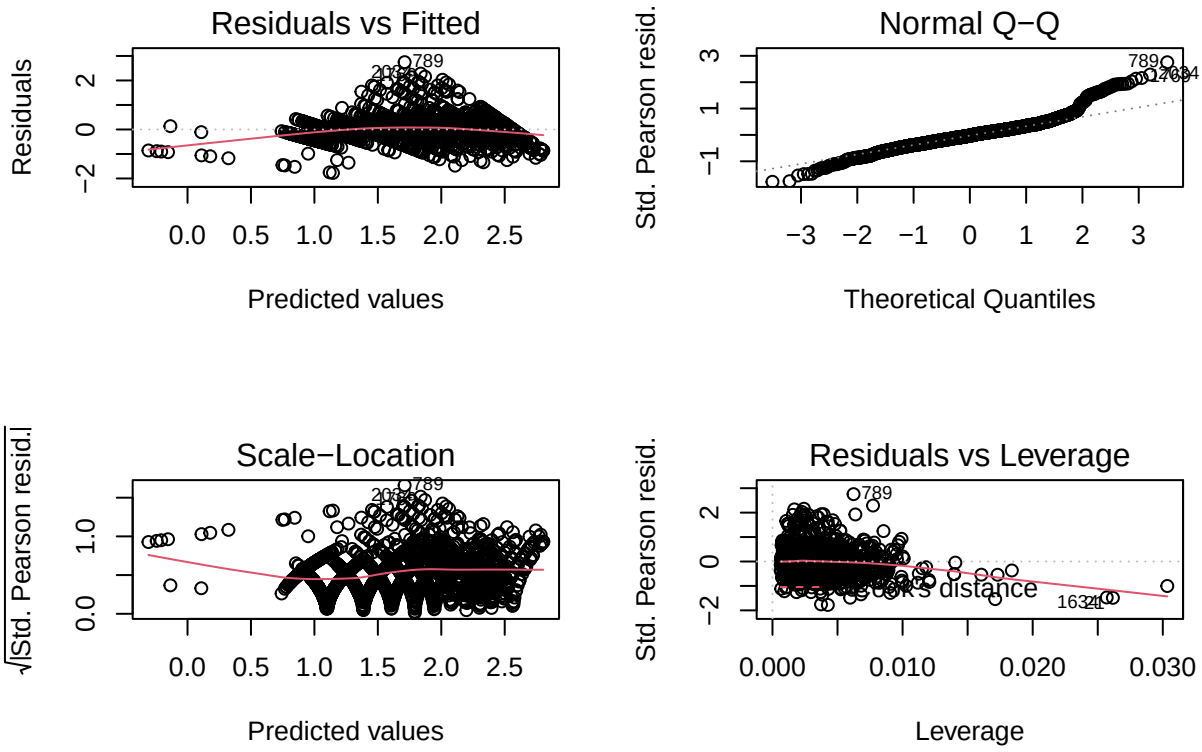
Testing a Poisson Regression to Answer the First Question

We also used a poisson regression to attempt to predict the number of in-store purchases.

##

```
## Call:
## glm(formula = NumStorePurchases ~ NumCatalogPurchases + Dt_Customer +
##     Response + NumTotalPurchases + NumWebVisitsMonth, family = poisson(link = log),
##     data = nsp_data)
##
## Deviance Residuals:
##      Min        1Q    Median        3Q       Max
## -2.50841  -0.27406  -0.02458   0.21099   2.37485
##
## Coefficients:
##                       Estimate Std. Error z value Pr(>|z|)
## (Intercept)          3.663e+00  7.979e-01   4.591 4.42e-06 ***
## NumCatalogPurchases -1.022e-01  5.373e-03 -19.026  < 2e-16 ***
## Dt_Customer         -1.578e-04  4.921e-05  -3.207  0.00134 **
## Response            -5.187e-02  2.532e-02  -2.049  0.04046 *
## NumTotalPurchases    8.986e-02  1.953e-03  46.006  < 2e-16 ***
## NumWebVisitsMonth   -7.062e-02  4.863e-03 -14.523  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for poisson family taken to be 1)
##
##     Null deviance: 3895.11  on 2208  degrees of freedom
## Residual deviance:  469.09  on 2203  degrees of freedom
## AIC: 8144.7
##
## Number of Fisher Scoring iterations: 4
```

Checking Plot Statistics of the Poission Model



There was not over-dispersion, so a quasi-poisson model was not needed. We can see that the poisson model is a good fit due to the high p-value when running the goodness-of-fit test

From reviewing the final summary of the model, you can see that younger individuals with a lower income, less purchases from the catalog, more total purchases less web visits per month and the longer they are a customer they will have a higher number of purchases from the store.

Second Question: Can we create a profile of a Customer that will complain?

We then created a logistic regression for the complaint data in an attempt to profile the users that were most likely to complain in our dataset.

```
##
## Call:
## glm(formula = final_data$Complain ~ Education + Dt_Customer +
##     MntGoldProds + AcceptedCmp4, family = binomial, data = final_data)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -0.2895  -0.1614  -0.1186  -0.0664   3.2941
##
## Coefficients:
```

```
##                       Estimate Std. Error z value Pr(>|z|)
## (Intercept)          1.546e+01  2.387e+03   0.006   0.9948
## EducationGraduation  1.660e+01  2.387e+03   0.007   0.9945
## EducationMaster      1.624e+01  2.387e+03   0.007   0.9946
## EducationPhD         1.475e+01  2.387e+03   0.006   0.9951
## Dt_Customer         -2.264e-03  1.184e-03  -1.912   0.0558 .
## MntGoldProds        -1.249e-02  7.030e-03  -1.776   0.0757 .
## AcceptedCmp4        -1.569e+01  1.344e+03  -0.012   0.9907
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 228.06  on 2211  degrees of freedom
## Residual deviance: 211.59  on 2205  degrees of freedom
## AIC: 225.59
##
## Number of Fisher Scoring iterations: 19
```

Confusion Matrix to Check Our Logistic Model's Predictions

```
##          response
## predicted    0    1
##         0 2192   20
```

From the Confusion Matrix's results, we can determine that there was not enough data to build a good predictive model. Of the 2240 observations in our dataset, only 20 complained, which caused the model to predict that nobody would complain and maintain an extremely high level of accuracy, 99.11%. As you can see in our confusion matrix, our model did not even attempt to make a prediction that the customer would complain.

Third Question: Can we create a customer profile that will assist us in predicting who will purchase the most amount of Wine?

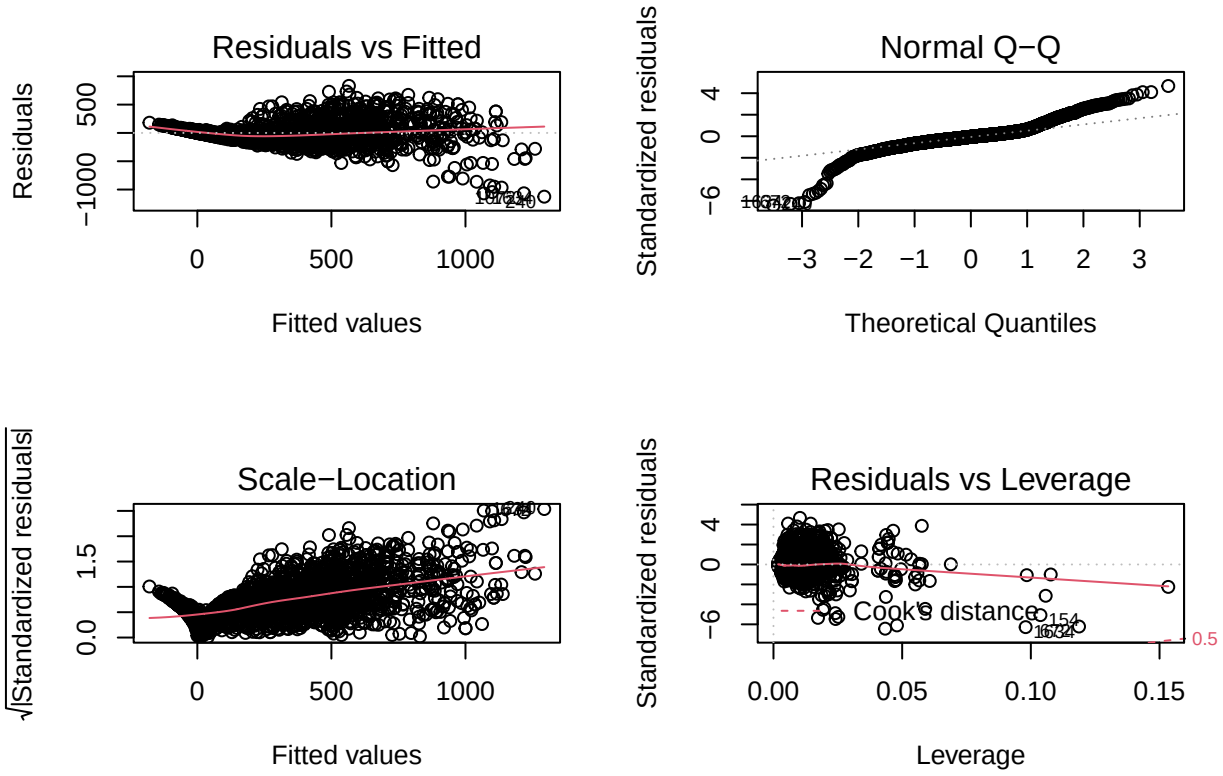
We began answering this question by creating a full model, on which a stepwise selection would be run to find the most significant predictors.

```
##
## Call:
## lm(formula = (wine_data$MntWines) ~ Year_Birth + Education +
##     Income + Kidhome + Dt_Customer + MntMeatProducts + MntSweetProducts +
##     MntGoldProds + NumDealsPurchases + NumWebPurchases + NumCatalogPurchases +
##     NumStorePurchases + NumWebVisitsMonth + AcceptedCmp3 + AcceptedCmp4 +
##     AcceptedCmp5 + AcceptedCmp1 + AcceptedCmp2, data = wine_data)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1125.81   -81.56    -9.79    58.28   830.65
##
## Coefficients:
##                       Estimate Std. Error t value Pr(>|t|)
## (Intercept)          2.522e+03  7.721e+02   3.266  0.00111 **
```

```
## Year_Birth          -6.650e-01  3.488e-01  -1.907  0.05669 .
## EducationGraduation -2.589e+01  2.590e+01  -1.000  0.31766
## EducationMaster     -2.153e+00  2.645e+01  -0.081  0.93514
## EducationPhD         5.628e+01  2.709e+01   2.077  0.03789 *
## Income               3.795e-03  3.524e-04  10.768  < 2e-16 ***
## Kidhome             -3.980e+01  9.639e+00  -4.129 3.78e-05 ***
## Dt_Customer         -9.577e-02  2.141e-02  -4.473 8.12e-06 ***
## MntMeatProducts      1.126e-01  2.864e-02   3.930 8.76e-05 ***
## MntSweetProducts    -4.684e-01  1.202e-01  -3.898 1.00e-04 ***
## MntGoldProds         1.656e-01  9.009e-02   1.838  0.06624 .
## NumDealsPurchases   -5.303e+00  2.387e+00  -2.222  0.02641 *
## NumWebPurchases      1.754e+01  1.944e+00   9.025  < 2e-16 ***
## NumCatalogPurchases  2.108e+01  2.249e+00   9.374  < 2e-16 ***
## NumStorePurchases    2.641e+01  1.722e+00  15.339  < 2e-16 ***
## NumWebVisitsMonth    2.189e+01  2.645e+00   8.278  < 2e-16 ***
## AcceptedCmp3         4.365e+01  1.532e+01   2.848  0.00444 **
## AcceptedCmp4         1.747e+02  1.661e+01  10.516  < 2e-16 ***
## AcceptedCmp5         2.384e+02  1.786e+01  13.344  < 2e-16 ***
## AcceptedCmp1         4.897e+01  1.788e+01   2.739  0.00622 **
## AcceptedCmp2         1.014e+02  3.526e+01   2.874  0.00409 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 178.9 on 2191 degrees of freedom
## Multiple R-squared:  0.7214, Adjusted R-squared:  0.7188
## F-statistic: 283.6 on 20 and 2191 DF,  p-value: < 2.2e-16
```

Summarizing the Original Model to Predict Who Will Purchase Wine:

Adjusted R-squared of 0.7188, with 18/20 predictors included in this model having a significance at a minimum of 90% confidence. The models F-statistic was 283.6, showing we have a highly signifcant model.



After plotting the model's plot statistics, we can see serious heteroscedasticity in the Scale-Location graph. We can also see the Residuals vs. Fitted graph does not have much of an even spread on the left tail. The Normal Q-Q has extremely frayed edges from the line of best fit. Finally, we can see in the Residuals vs Leverage plot has a few points negativly leveraging the model.

We decided from this information, we needed a transformation.

Prior to transforming our data, we decided to check for the amount of 0s contained in this variable, to confirm whether or not we were going to remove data points in our set.

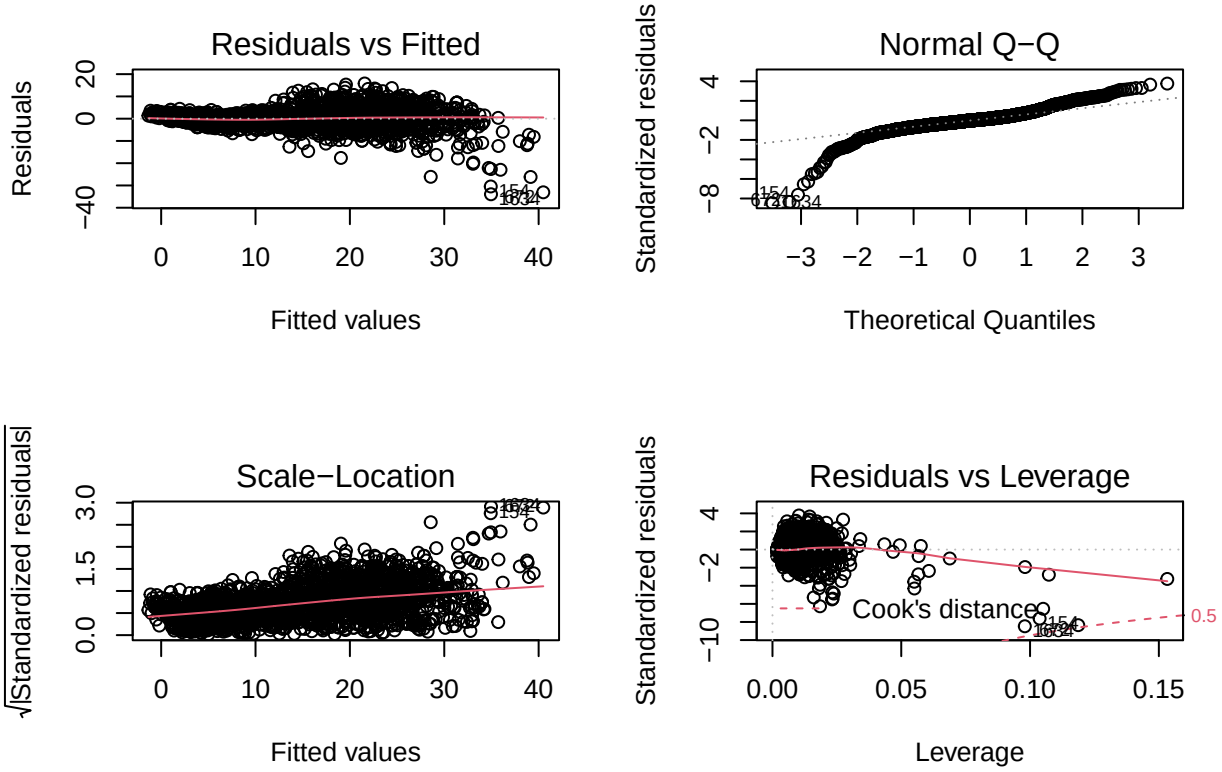
There are only 7 observations in MntWines that contain a 0 , so we did not feel the need to remove any of the observations. Transforming the data would solve the issues of potentially getting a negative number predicted.

```
##
## Call:
## lm(formula = (wine_data$MntWines)^(1/2) ~ Year_Birth + Education +
##     Income + Kidhome + Dt_Customer + MntMeatProducts + MntSweetProducts +
```

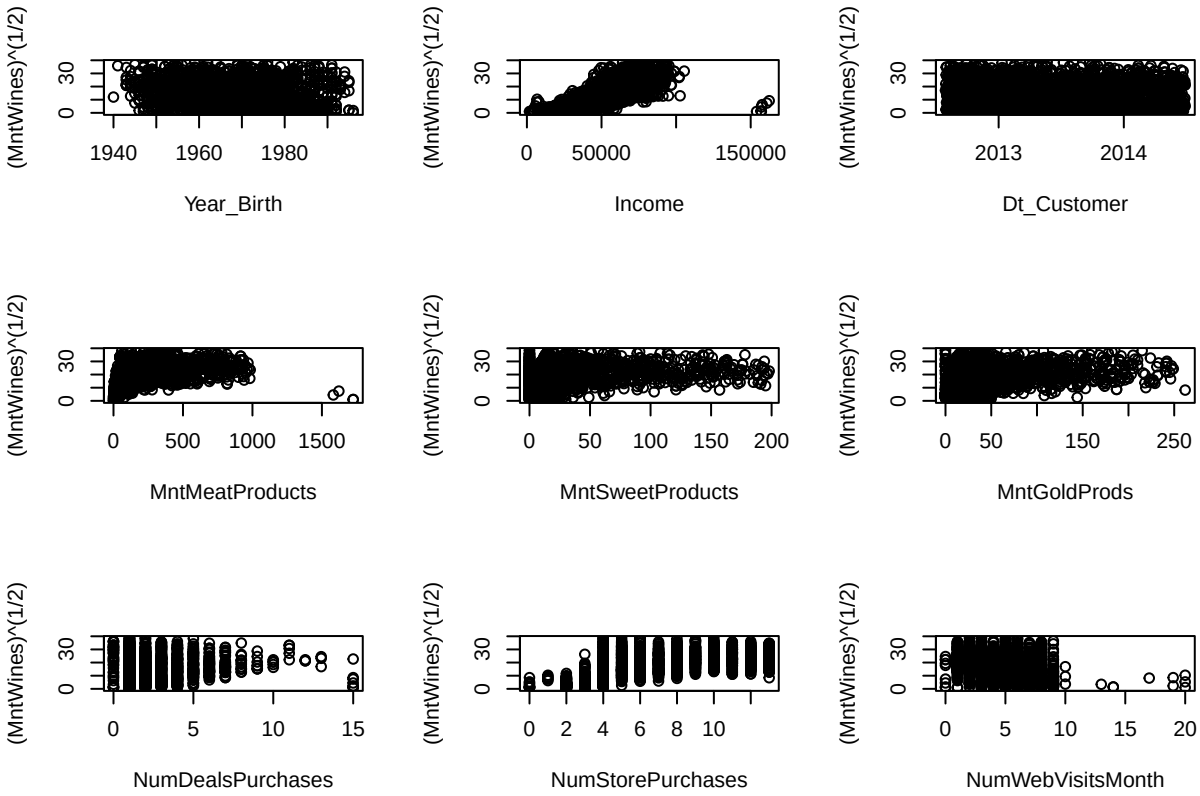
```
##     MntGoldProds + NumDealsPurchases + NumWebPurchases + NumCatalogPurchases +
##     NumStorePurchases + NumWebVisitsMonth + AcceptedCmp3 + AcceptedCmp4 +
##     AcceptedCmp5 + AcceptedCmp1, data = wine_data)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -33.953  -1.859  -0.094   1.710  15.800
##
## Coefficients:
##                       Estimate Std. Error t value Pr(>|t|)
## (Intercept)          7.826e+01  1.821e+01   4.298 1.80e-05 ***
## Year_Birth          -2.474e-02  8.226e-03  -3.008 0.002657 **
## EducationGraduation  1.708e+00  6.108e-01   2.797 0.005203 **
## EducationMaster      2.118e+00  6.239e-01   3.395 0.000699 ***
## EducationPhD         3.724e+00  6.389e-01   5.829 6.39e-09 ***
## Income               1.427e-04  8.310e-06  17.178  < 2e-16 ***
## Kidhome             -1.821e+00  2.273e-01  -8.011 1.83e-15 ***
## Dt_Customer         -2.327e-03  5.050e-04  -4.609 4.28e-06 ***
## MntMeatProducts      2.223e-03  6.743e-04   3.297 0.000993 ***
## MntSweetProducts    -1.146e-02  2.833e-03  -4.047 5.36e-05 ***
## MntGoldProds         8.147e-03  2.124e-03   3.835 0.000129 ***
## NumDealsPurchases    1.558e-01  5.627e-02   2.768 0.005688 **
## NumWebPurchases      6.942e-01  4.574e-02  15.179  < 2e-16 ***
## NumCatalogPurchases  5.007e-01  5.297e-02   9.453  < 2e-16 ***
## NumStorePurchases    8.389e-01  4.055e-02  20.690  < 2e-16 ***
## NumWebVisitsMonth    4.475e-01  6.223e-02   7.191 8.80e-13 ***
## AcceptedCmp3         1.268e+00  3.606e-01   3.515 0.000449 ***
## AcceptedCmp4         4.173e+00  3.825e-01  10.909  < 2e-16 ***
## AcceptedCmp5         4.149e+00  4.186e-01   9.912  < 2e-16 ***
## AcceptedCmp1         9.690e-01  4.208e-01   2.303 0.021391 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 4.218 on 2192 degrees of freedom
## Multiple R-squared:  0.8229, Adjusted R-squared:  0.8214
## F-statistic: 536.1 on 19 and 2192 DF,  p-value: < 2.2e-16
```

Summarizing the Transformed Model to Predict Who Will Purchase Wine:

Adjusted R-squared of 0.8214 from 0.7188, which is a significant increase. All of the predictors increased in significance, raising the minimum level of confidence across all to 95%. The amount of predictors also decreased, dropping to 19 from 20, reducing the complexity of our model. The F-statistic also rose from 283.6 to 536.1, showing the extreme increase in significance of our model.



You can see from the transformation, the model's plot statistiics look much better. Besides having some fringe points on the Normal Q-Q and the Residuals vs Leverage plots, almost all of the heteroscedasticity was removed when looking at the Residusals vs Fitted and Scale-Location plots.



When plotting the scatterplots of the transformed data in correlation to each of the predictor variables, you can see that the square root transformation had some affect on the dispersion of our data. With the exception of DT_Customer, Year_Birth and NumWebVisitsMonth, the other variables seem to rise properly in correlation to the transformation.

Although we do lose some of the interpretability from the transformation, the inferential value gained in the model was well worth it.

```
##
## Call:
## lm(formula = (wine_data$MntWines)^(1/2) ~ Year_Birth + Education +
##     Income + Kidhome + Dt_Customer + MntMeatProducts + MntSweetProducts +
##     MntGoldProds + NumDealsPurchases + NumWebPurchases + NumCatalogPurchases +
##     NumStorePurchases + NumWebVisitsMonth + AcceptedCmp3 + AcceptedCmp4 +
##     AcceptedCmp5 + AcceptedCmp1, data = wine_data)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -33.953  -1.859  -0.094   1.710  15.800
##
```

```
## Coefficients:
##                       Estimate Std. Error t value Pr(>|t|)
## (Intercept)          7.826e+01  1.821e+01   4.298 1.80e-05 ***
## Year_Birth          -2.474e-02  8.226e-03  -3.008 0.002657 **
## EducationGraduation  1.708e+00  6.108e-01   2.797 0.005203 **
## EducationMaster      2.118e+00  6.239e-01   3.395 0.000699 ***
## EducationPhD         3.724e+00  6.389e-01   5.829 6.39e-09 ***
## Income               1.427e-04  8.310e-06  17.178  < 2e-16 ***
## Kidhome             -1.821e+00  2.273e-01  -8.011 1.83e-15 ***
## Dt_Customer         -2.327e-03  5.050e-04  -4.609 4.28e-06 ***
## MntMeatProducts      2.223e-03  6.743e-04   3.297 0.000993 ***
## MntSweetProducts    -1.146e-02  2.833e-03  -4.047 5.36e-05 ***
## MntGoldProds         8.147e-03  2.124e-03   3.835 0.000129 ***
## NumDealsPurchases    1.558e-01  5.627e-02   2.768 0.005688 **
## NumWebPurchases      6.942e-01  4.574e-02  15.179  < 2e-16 ***
## NumCatalogPurchases  5.007e-01  5.297e-02   9.453  < 2e-16 ***
## NumStorePurchases    8.389e-01  4.055e-02  20.690  < 2e-16 ***
## NumWebVisitsMonth    4.475e-01  6.223e-02   7.191 8.80e-13 ***
## AcceptedCmp3         1.268e+00  3.606e-01   3.515 0.000449 ***
## AcceptedCmp4         4.173e+00  3.825e-01  10.909  < 2e-16 ***
## AcceptedCmp5         4.149e+00  4.186e-01   9.912  < 2e-16 ***
## AcceptedCmp1         9.690e-01  4.208e-01   2.303 0.021391 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 4.218 on 2192 degrees of freedom
## Multiple R-squared:  0.8229, Adjusted R-squared:  0.8214
## F-statistic: 536.1 on 19 and 2192 DF,  p-value: < 2.2e-16
```

Answering the Third Question of Customer Profiling those who Purchase the Most Wine:

We can see that older customers with no kids at home, higher education levels, and higher incomes tend to purchase more wine than other customers.

Fourth Question: Can we create a customer profile that will assist us in predicting who will purchase the most amount of Meat?

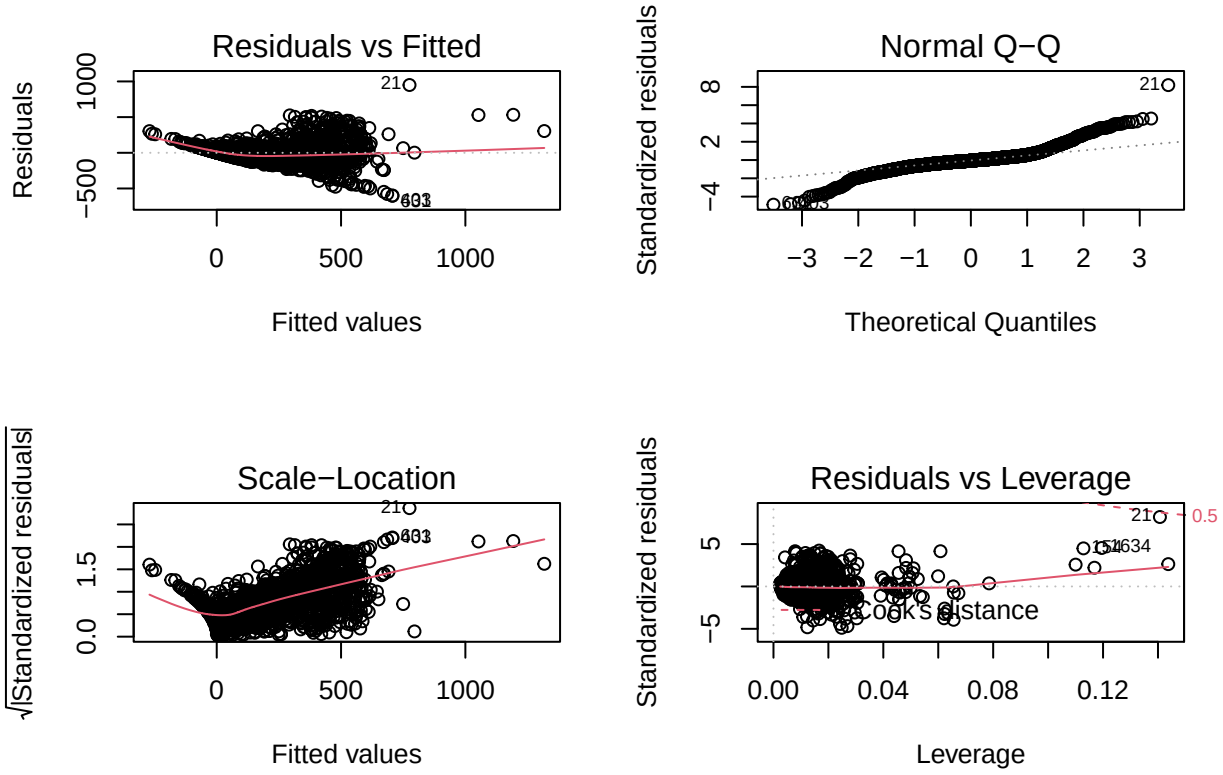
We began answering this question by creating a full model, on which a stepwise selection would be run to find the most significant predictors.

```
##
## Call:
## lm(formula = meat_data$MntMeatProducts ~ Year_Birth + Education +
##     Income + Kidhome + Teenhome + Dt_Customer + Recency + MntWines +
##     MntFruits + MntFishProducts + MntSweetProducts + MntGoldProds +
##     NumDealsPurchases + NumWebPurchases + NumCatalogPurchases +
##     NumStorePurchases + NumWebVisitsMonth + AcceptedCmp3 + AcceptedCmp4 +
##     AcceptedCmp5 + AcceptedCmp2 + Response, data = meat_data)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -602.53  -51.99   -8.59   40.28  949.63
```

```
##
## Coefficients:
##                         Estimate Std. Error t value Pr(>|t|)
## (Intercept)         -3.426e+01  5.640e+02  -0.061 0.951566
## Year_Birth           4.113e-01  2.555e-01   1.610 0.107544
## EducationGraduation  8.240e+00  1.817e+01   0.453 0.650334
## EducationMaster     -4.251e+00  1.856e+01  -0.229 0.818897
## EducationPhD        -9.004e+00  1.910e+01  -0.471 0.637355
## Income               3.287e-03  2.594e-04  12.669  < 2e-16 ***
## Kidhome             -1.477e+01  6.851e+00  -2.156 0.031212 *
## Teenhome            -7.319e+01  6.254e+00 -11.702  < 2e-16 ***
## Dt_Customer         -5.100e-02  1.523e-02  -3.348 0.000826 ***
## Recency              1.857e-01  9.497e-02   1.955 0.050652 .
## MntWines             5.503e-02  1.492e-02   3.689 0.000230 ***
## MntFruits            4.634e-01  9.327e-02   4.969 7.27e-07 ***
## MntFishProducts      3.565e-01  7.045e-02   5.059 4.56e-07 ***
## MntSweetProducts     2.346e-01  9.108e-02   2.576 0.010062 *
## MntGoldProds        -2.190e-01  6.411e-02  -3.417 0.000645 ***
## NumDealsPurchases    6.264e+00  1.794e+00   3.492 0.000489 ***
## NumWebPurchases     -3.443e+00  1.434e+00  -2.401 0.016432 *
## NumCatalogPurchases  2.494e+01  1.526e+00  16.338  < 2e-16 ***
## NumStorePurchases   -2.300e+00  1.303e+00  -1.766 0.077578 .
## NumWebVisitsMonth   -5.554e+00  1.882e+00  -2.952 0.003195 **
## AcceptedCmp3        -3.040e+01  1.100e+01  -2.763 0.005769 **
## AcceptedCmp4        -5.274e+01  1.189e+01  -4.437 9.56e-06 ***
## AcceptedCmp5         3.838e+01  1.304e+01   2.943 0.003281 **
## AcceptedCmp2        -6.956e+01  2.472e+01  -2.814 0.004944 **
## Response             3.960e+01  8.889e+00   4.455 8.80e-06 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 125 on 2184 degrees of freedom
## Multiple R-squared:  0.6927, Adjusted R-squared:  0.6893
## F-statistic: 205.1 on 24 and 2184 DF,  p-value: < 2.2e-16
```

Summarizing the Original Model to Predict Who Will Purchase Meat:

Adjusted R-squared of 0.6893, with 20/24 predictors significant at a minimum of 90% confidence. The F-statistic was 205.1, showing a highly signficant model.



After plotting the model's plot statistics, we can see some heteroscedasticity in the Scale-Location graph. We can also see the Residuals vs. Fitted graph does not have much of an even spread on the left tail. The Normal Q-Q has moderately frayed edges from the line of best fit. Finally, we can see in the Residuals vs Leverage plot has a few points negativly leveraging the model.

Prior to transforming our data, we decided to check for the amount of 0s contained in this variable, to confirm whether or not we were going to remove data points in our set.

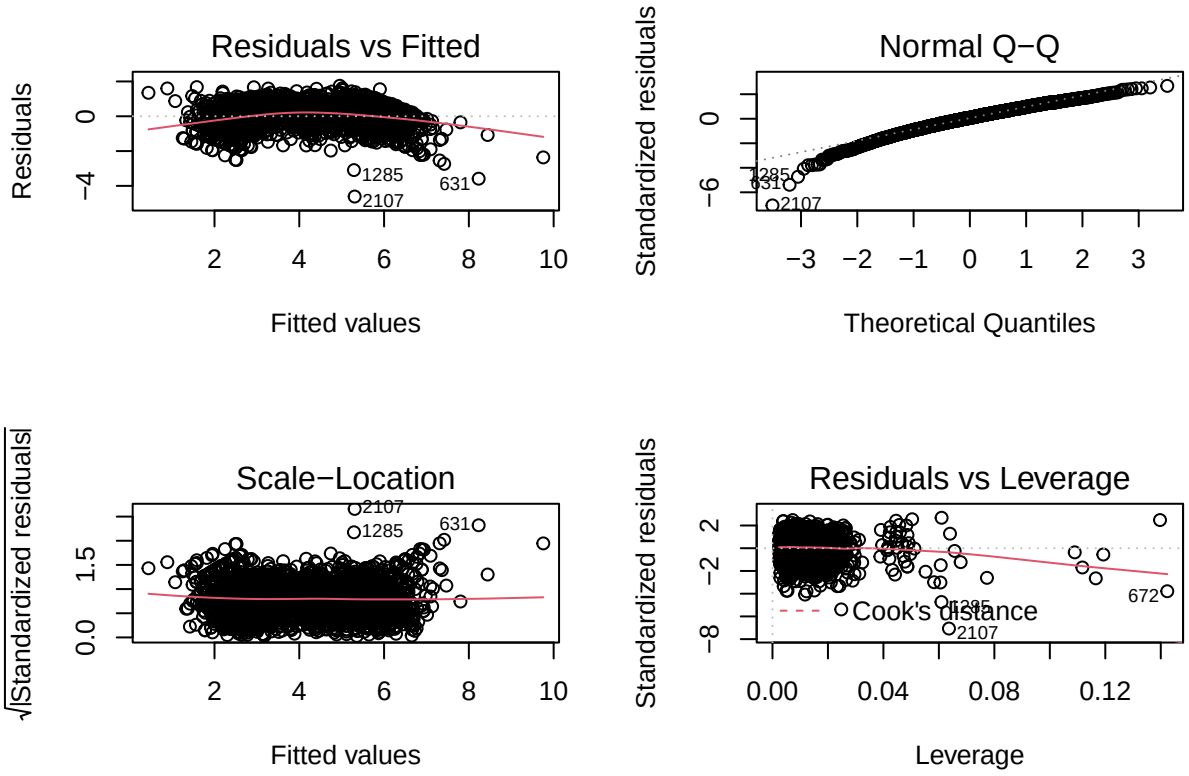
There was only one observation that had no meat purchases in our data, which allowed the transfomation to solve the issues that the observation would have caused the model.

```
##
## Call:
## lm(formula = log(meat_data$MntMeatProducts) ~ Education + Income +
##     Kidhome + Teenhome + Dt_Customer + Recency + MntWines + MntFruits +
##     MntFishProducts + MntSweetProducts + MntGoldProds + NumDealsPurchases +
##     NumWebPurchases + NumCatalogPurchases + NumStorePurchases +
##     NumWebVisitsMonth + AcceptedCmp3 + AcceptedCmp4 + AcceptedCmp1 +
##     AcceptedCmp2 + Response, data = meat_data)
```

```
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -4.6133 -0.3819  0.0513  0.4539  1.7517
##
## Coefficients:
##                       Estimate Std. Error t value Pr(>|t|)
## (Intercept)          6.611e+00  1.338e+00   4.941 8.35e-07 ***
## EducationGraduation  7.374e-01  9.794e-02   7.528 7.47e-14 ***
## EducationMaster      6.594e-01  1.000e-01   6.594 5.34e-11 ***
## EducationPhD         5.857e-01  1.027e-01   5.701 1.36e-08 ***
## Income               2.254e-05  1.396e-06  16.149  < 2e-16 ***
## Kidhome             -2.133e-01  3.658e-02  -5.831 6.35e-09 ***
## Teenhome            -4.797e-01  3.192e-02 -15.029  < 2e-16 ***
## Dt_Customer         -3.499e-04  8.262e-05  -4.235 2.38e-05 ***
## Recency              1.704e-03  5.119e-04   3.328 0.000890 ***
## MntWines             3.670e-04  7.751e-05   4.735 2.33e-06 ***
## MntFruits            2.230e-03  5.032e-04   4.432 9.78e-06 ***
## MntFishProducts      2.276e-03  3.802e-04   5.985 2.52e-09 ***
## MntSweetProducts     1.344e-03  4.900e-04   2.742 0.006156 **
## MntGoldProds         7.011e-04  3.456e-04   2.028 0.042652 *
## NumDealsPurchases    1.140e-01  9.666e-03  11.797  < 2e-16 ***
## NumWebPurchases      1.030e-01  7.719e-03  13.345  < 2e-16 ***
## NumCatalogPurchases  8.776e-02  8.223e-03  10.673  < 2e-16 ***
## NumStorePurchases    7.214e-02  6.997e-03  10.310  < 2e-16 ***
## NumWebVisitsMonth   -3.788e-02  1.012e-02  -3.744 0.000186 ***
## AcceptedCmp3        -1.391e-01  5.926e-02  -2.346 0.019041 *
## AcceptedCmp4        -3.021e-01  6.395e-02  -4.723 2.47e-06 ***
## AcceptedCmp1        -1.566e-01  6.765e-02  -2.315 0.020699 *
## AcceptedCmp2        -2.181e-01  1.332e-01  -1.638 0.101570
## Response             2.430e-01  4.780e-02   5.084 4.00e-07 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.6744 on 2185 degrees of freedom
## Multiple R-squared:  0.8288, Adjusted R-squared:  0.8269
## F-statistic: 459.8 on 23 and 2185 DF,  p-value: < 2.2e-16
```

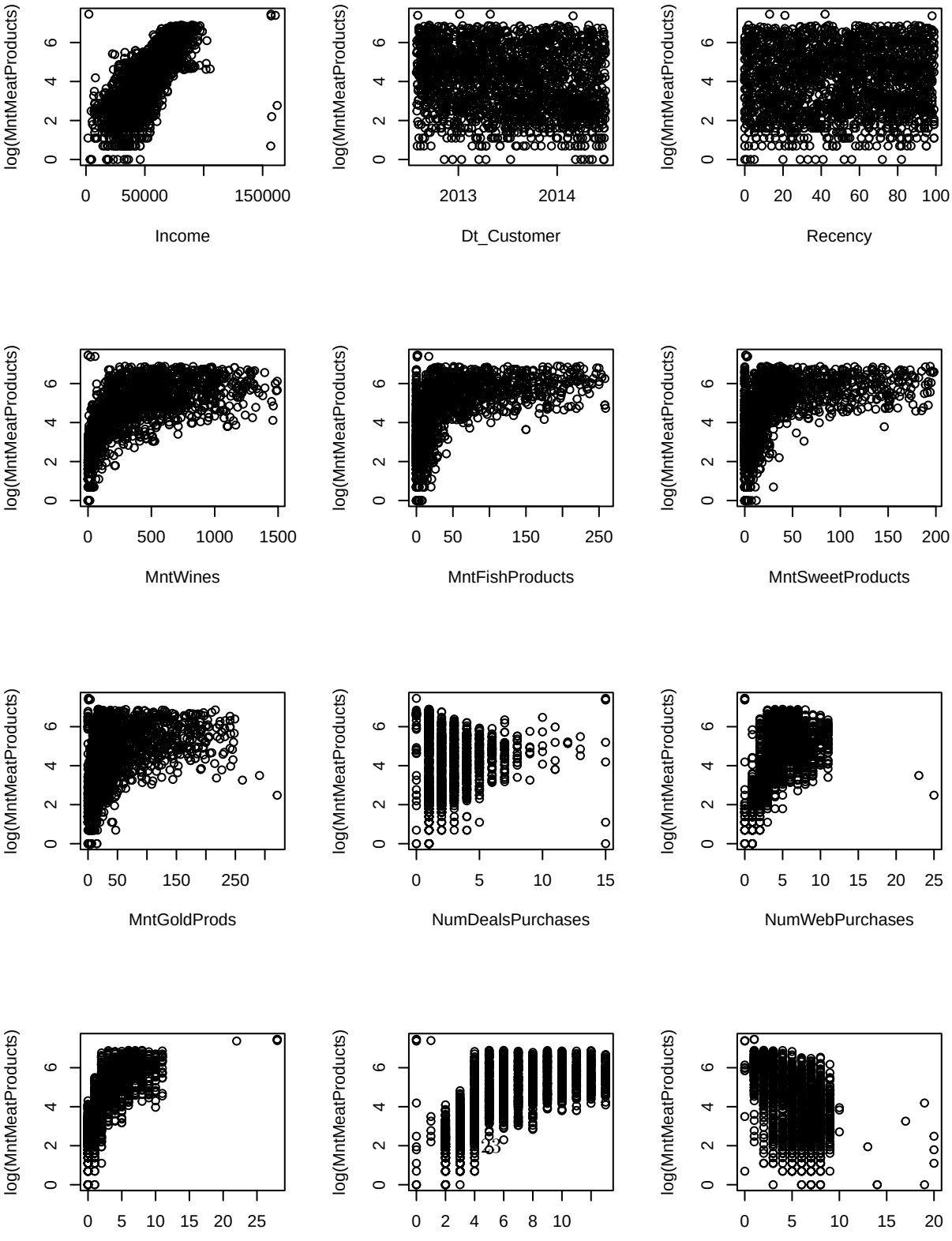
Summarizing the Transformed Model to Profile Customers who Purchase the most Meat:

Adjusted R-squared of 0.8269, which is a significant increase from 0.6893. The minimum level of significance increased as well for all signficant predictors in our model, increasing from 90% confidence to 95% confidence. The F-statistic also rose from 205.1 to 459.8, showing the increase in signficance in our model.



You can see from the transformation, the model's plot statistiics look much better. Besides having some fringe points on the Normal Q-Q and the Residuals vs Leverage plots, almost all of the heteroscedasticity was removed when looking at the Residusals vs Fitted and Scale-Location plots.

Although we do lose some of the interpretability from the transformation, the inferential value gained in the model was worth it.



When plotting the scatterplots of the transformed data in correlation to each of the predictor variables, you can see that many of the predictors rise smoothly with the logorithmic curve, with the exception of Dt_Customer, Recency and NumWebVisitsMonth (bottom right).

Answering the Fourth Question of what kind of Customers purchase the most meat:

From this model, we can see that customers with no kids or teens at home, a higher education level, higher income, and less frequent purchases tend to spend more on meat products.

Limitations

Most of the regression models we had created were without limitation. The only regression we truly had an issue with was our Logistic Regression on the binary variable Complaint. In our 2240 observation dataset, only 20 of the customers had complained. Due to the small sample size of customers that had complained versus customers that had not, we were unable to create a model that could accurately predict whether a customer will complain in the future. If we were to have more time, a possible solution would be to impute data based off of the small portion given in our dataset and create more complaints as to continue testing the model, but even then there is not much information to impute upon.