Module 08: Segmentation & Profiling Project

Nischal Pant

Introduction:

In this project, we are tasked with identifying, segments of the dataset Customer.csv, which we initially cleaned and analyzed. We will be using two different methods of segmentation in this project. Rules-Based Segmentation and Supervised Segmentation. Our main goal in the project is to identify segments of the population that we can use to effectively use towards marketing, company expansion, and customer retention. We will also be breaking down our project using the STAR approach (Situation, Task, Action, and Result) to further explain our findings, provide insights, and make data-driven decisions.

Situation:

We have a very large dataset of customers of our mega telecommunication company. They are very diverse and are very insightful.

Task:

We are tasked with finding high-value customers, and we must create segments to determine which group of the population we want to target based on variables of the Customer.csv data set. We want high-value customers.

Action:

We will use two segmentation methods, to determine the groups we would like to target. For that, we must create a set of rules by which we will abide, and analyze our data (rules-based) and then further analyze those segments with a supervised learning method (logistic regression) to see if our variables can be used to predict our dependent variable.

New variable creation:

We created new variables to make our decision tree, segmentation, and regression models possible. Household Income was separated into medium, high, and low based on thresholds, 35000 and below was low, 75000 and below was medium and above that was high, and above 100000 was Very high.

CustomerValueCatagory was also created to as well by adding dataovertenure, voicelastmonth and equipment lastmonth variables. Then using the mean, the mode, and the high values we created thresholds to determine if they were high or low-value customers. We can see that the mean was 64.10, the median was 49.75 and the mode was 15. We determined that any customer over 100 was considered high value and anything lower was low.

We also created a Debttoincomeratio category which we did the same, if the customer was under 4, the customer was low and if it was over 4 and less than 6 it was medium, and over 6 was high.

DataOverTenureCatagory for low, medium and high was also created using the same methodology, which mean and under being low, mode and below being medium and higher would be considered high. The distribution of the histogram for DataOverTenure was normal so this would be the best course of action.

Results:

We will further discuss the results of our findings and then make data-driven decisions.

Segment creation and profiling:

1. Men

2. Women with Low Debt-to-Income Ratio, Multiline Usage, and Equipment Rental

3. Women with Low Debt-to-Income Ratio, Multiline Usage, Equipment Rental, and High Data Over Tenure

4. Women with High Debt-to-Income Ratio, Multiline Usage, Equipment Rental, and High-value customer

These segments were created using rule-based, and logistic regression models.

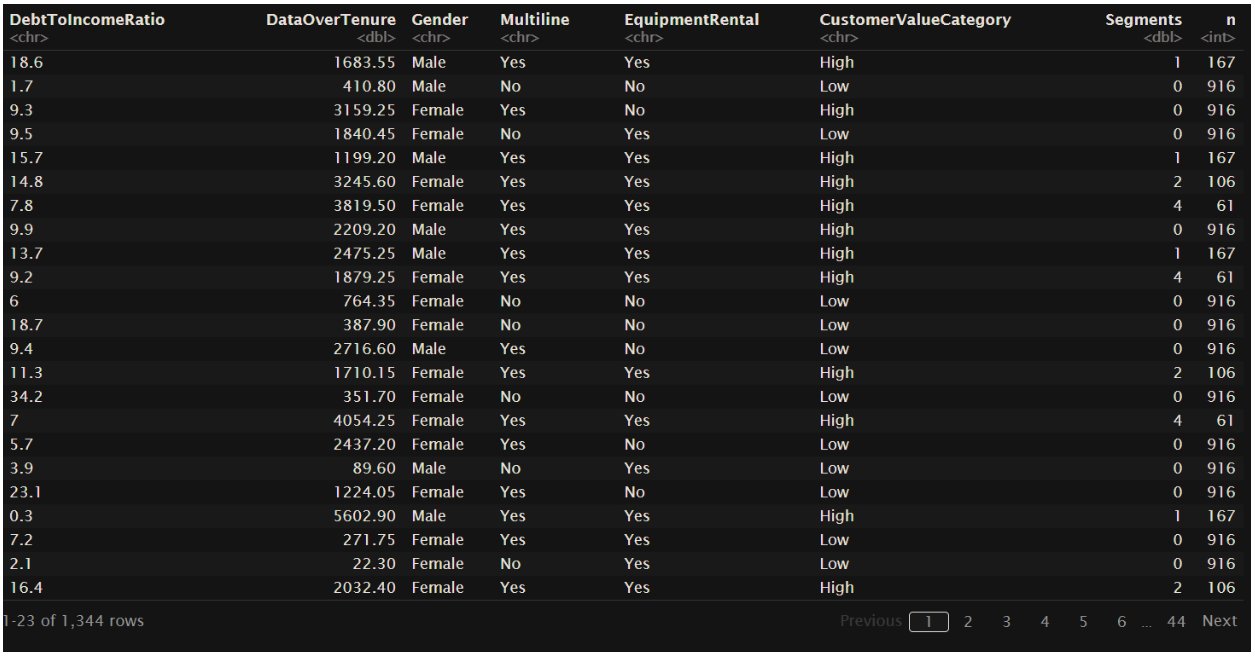
Rule Based:

The majority of the segments are centered around women. We want to first create a diversion for the most simple binary variable we can find and we have gender for that. If we target the entire variable we cannot create much segregation and the segmentation could be harder to understand. This is simply a rule and should not be considered anything else.

We wanted to use different variables that would show promising usage. Any good telecommunication company's KPI will have to consider its services' usage. To target customers who are already heavily using the services and Yes to multiline, yes to equipment rental, and high data over tenure seemed like the best variable to use for the segmentation.

High-value customers are almost a redundant usage as it takes into account dataovertenure and equimentrentallastmonth (which is not part of the segments we have for this project). Still, we can see that the n for this segment (4) was significantly lower which could be used for further feature selection to improve our segments.

- Women with Low Debt-to-Income Ratio, Multiline Usage, and Equipment Rental n = 167



DebtToIncomeRatio	DataOverTenure	Gender	Multiline	EquipmentRental	CustomerValueCategory	Segments	n
<chr>	<dbl>	<chr>	<chr>	<chr>	<chr>	<dbl>	<int>
18.6	1683.55	Male	Yes	Yes	High	1	167
1.7	410.80	Male	No	No	Low	0	916
9.3	3159.25	Female	Yes	No	High	0	916
9.5	1840.45	Female	No	Yes	Low	0	916
15.7	1199.20	Male	Yes	Yes	High	1	167
14.8	3245.60	Female	Yes	Yes	High	2	106
7.8	3819.50	Female	Yes	Yes	High	4	61
9.9	2209.20	Male	Yes	Yes	High	0	916
13.7	2475.25	Male	Yes	Yes	High	1	167
9.2	1879.25	Female	Yes	Yes	High	4	61
6	764.35	Female	No	No	Low	0	916
18.7	387.90	Female	No	No	Low	0	916
9.4	2716.60	Male	Yes	No	Low	0	916
11.3	1710.15	Female	Yes	Yes	High	2	106
34.2	351.70	Female	No	No	Low	0	916
7	4054.25	Female	Yes	Yes	High	4	61
5.7	2437.20	Female	Yes	No	Low	0	916
3.9	89.60	Male	No	Yes	Low	0	916
23.1	1224.05	Female	Yes	No	Low	0	916
0.3	5602.90	Male	Yes	Yes	High	1	167
7.2	271.75	Female	Yes	Yes	Low	0	916
2.1	22.30	Female	No	Yes	Low	0	916
16.4	2032.40	Female	Yes	Yes	High	2	106

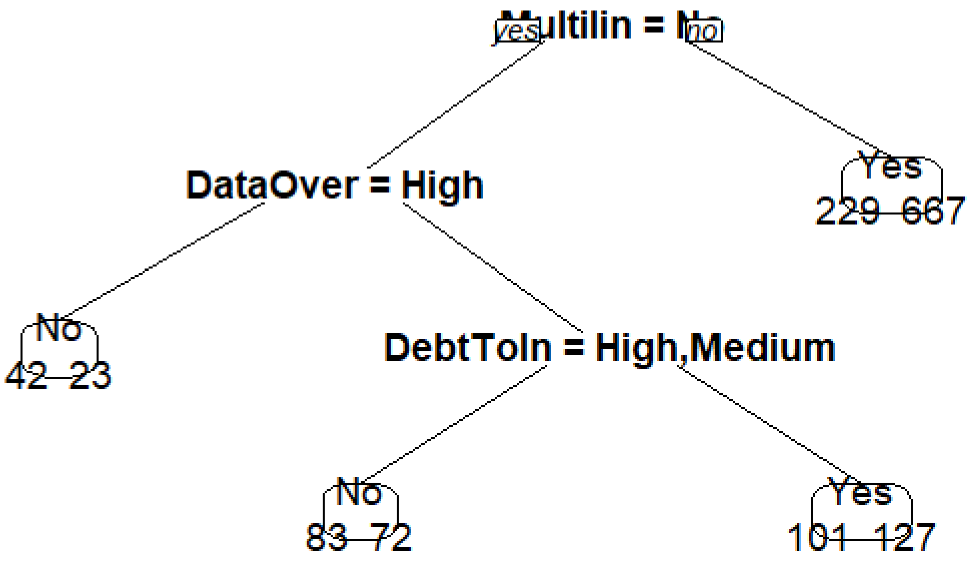
1-23 of 1,344 rows

Logistic Regression model:

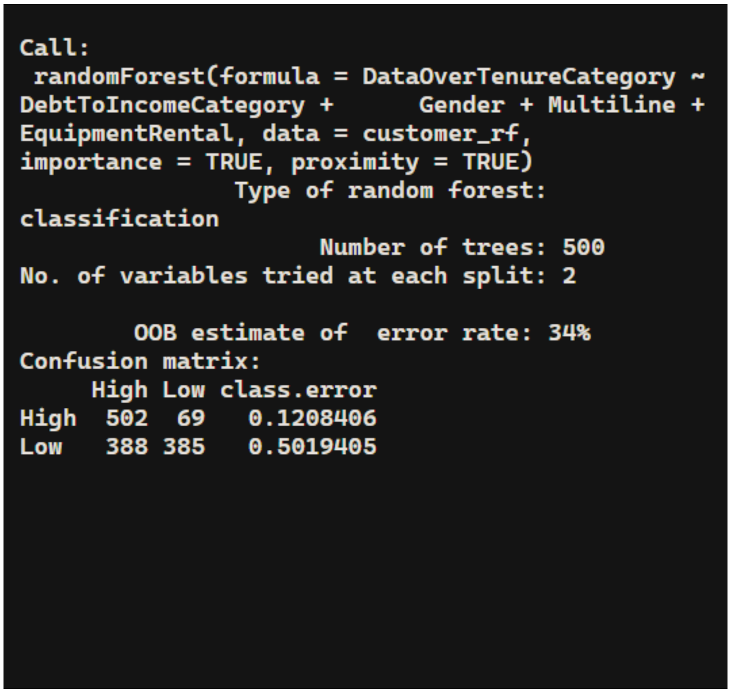
This segment had the largest output of 167 individuals. When we compare our rule-based segmentation with our regression models we can see many similarities in the output. For the regression model, we categorized the CustomerValue category by adding values of VoiceLastMonth, EquipmentLastMonth, and DataLastMonth value. We created a threshold for high, medium, and low based on the mean, mode, and high values in the entires. Using the customer value category (as our dependent variable) we were able to find the significance of the other variables and determined that DataOverTenureCategoryLow, MultilineYes, and EquipmentRentalYes had the most significant correlation which we indeed had in our segment.

Both models have similar outputs but there is a lot of things that are unconsidered and not addressed in the project. These must be described in a later section on things that need to be improved.

Summary of findings and explanation:

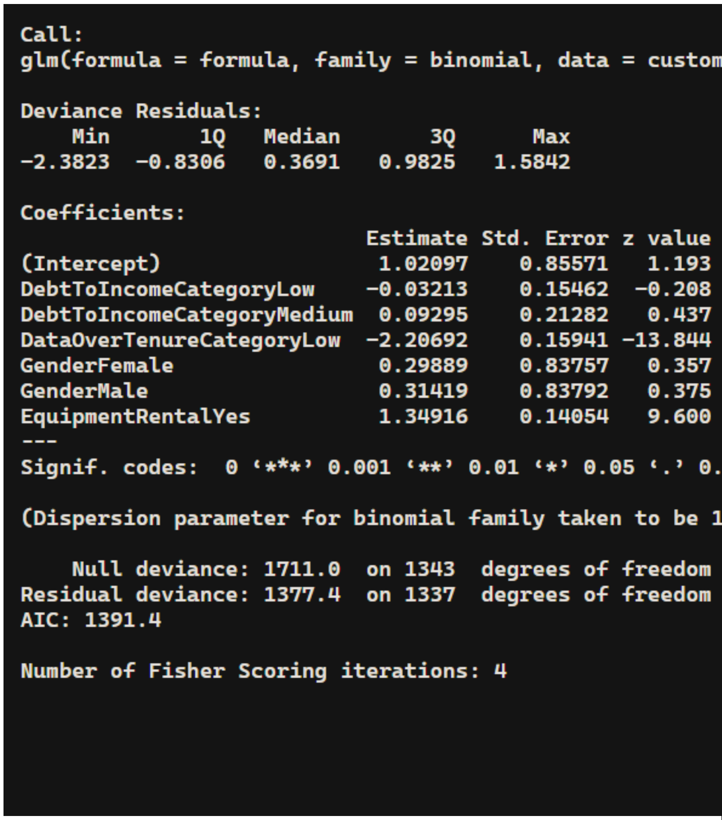


The figure above shows a tree model used for understanding our variables. Multiline (whether they had multiple lines or not), Debt to income category (low debt to income ratio was broken into high, low, and medium), and Data Over tenure (how much data they used over time, high and low). As you can see each node separated a yes and no result. People with multiple lines, high data-over tenure, and medium/high debt-to-income ratio yielded n = 100, and 127 respectively, and low debt-to-ratio income of n = 83 and 72 respectively. These are the target segments we would potentially like to use.



```
Call:
 randomForest(formula = DataOverTenureCategory ~
DebtToIncomeCategory +      Gender + Multiline +
EquipmentRental, data = customer_rf,
importance = TRUE, proximity = TRUE)
               Type of random forest:
classification
                     Number of trees: 500
No. of variables tried at each split: 2

        OOB estimate of  error rate: 34%
Confusion matrix:
     High Low class.error
High  502  69   0.1208406
Low   388 385   0.5019405
```

The figure on the left was a random forest analysis with DataOverTenureCatagory (with values low, medium, and high) that used a dependent variable to see if we could use machine learning by constructing many decision trees to create connections between the variables. The output shows 500 trees and 2 splits at each node but the estimated error rate was 34%, thus causing us to stray away from this analysis. Further feature selection was needed for this analysis.

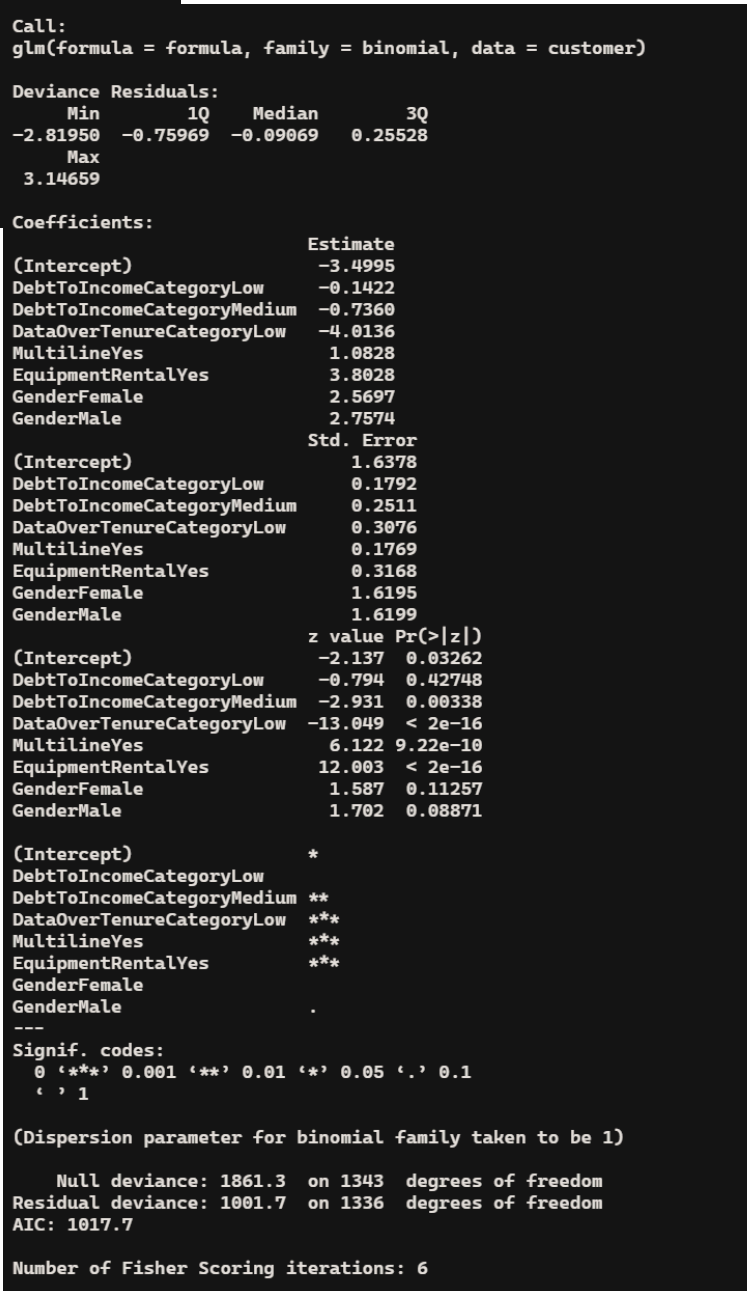


```
Call:
glm(formula = formula, family = binomial, data = customer)

Deviance Residuals: 
    Min       1Q   Median       3Q      Max  
-2.3823  -0.8306   0.3691   0.9825   1.5842  

Coefficients:
                           Estimate Std. Error z value Pr(>|z|)    
(Intercept)                 1.02097    0.85571   1.193    0.233    
DebtToIncomeCategoryLow    -0.03213    0.15462  -0.208    0.835    
DebtToIncomeCategoryMedium  0.09295    0.21282   0.437    0.662    
DataOverTenureCategoryLow  -2.20692    0.15941 -13.844   <2e-16 ***
GenderFemale                0.29889    0.83757   0.357    0.721    
GenderMale                  0.31419    0.83792   0.375    0.708    
EquipmentRentalYes          1.34916    0.14054   9.600   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 1711.0  on 1343  degrees of freedom
Residual deviance: 1377.4  on 1337  degrees of freedom
AIC: 1391.4

Number of Fisher Scoring iterations: 4
```

The figure on the left used a logistic regression model to analyze our variables. We categorized the CustomerValue category by adding values of VoiceLastMonth, EquipmentLastMonth, and DataLastMonth value and created a threshold for high, medium, and low based on the mean, mode, and high values in the entires. Using the customer value category we were able to find the significance of the other variables and determined that dataovertenurecatagorylow, and equipmentrentalyes correlated.

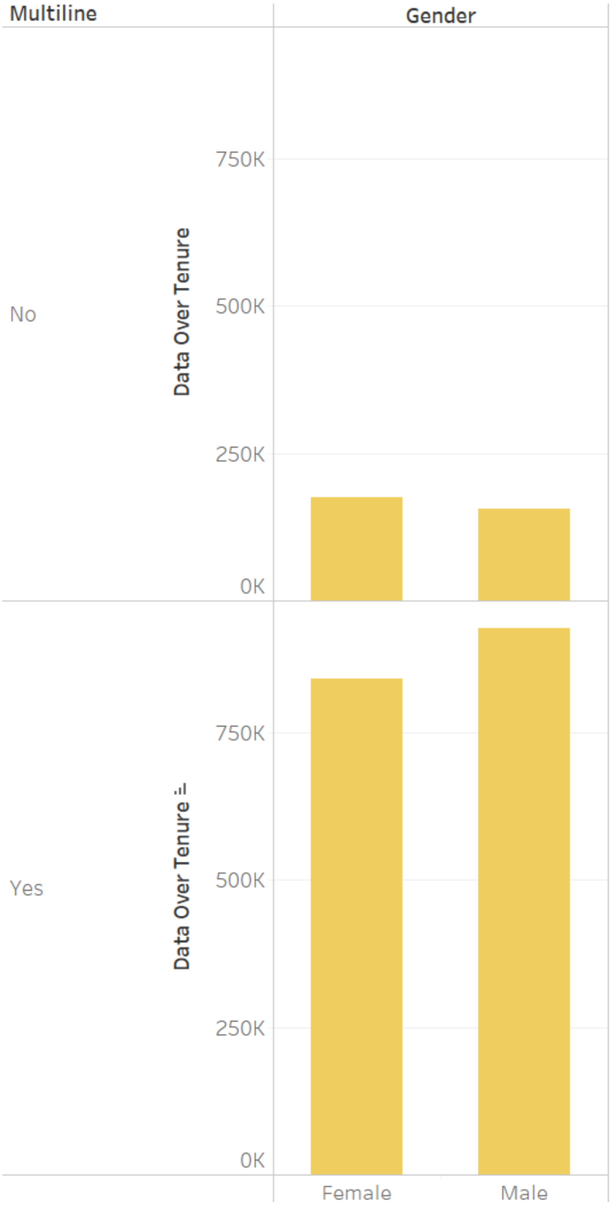
We in fact used this variable in our rule-based segmentation. The majority of our high-value data is correlated to these variables (taking into account the errors)

The figure on the right is another regression model that we executed which showed that, DataOverTenureCategoryLow, MultilineYes, and EquipmentRentalYes had the most significant correlation.



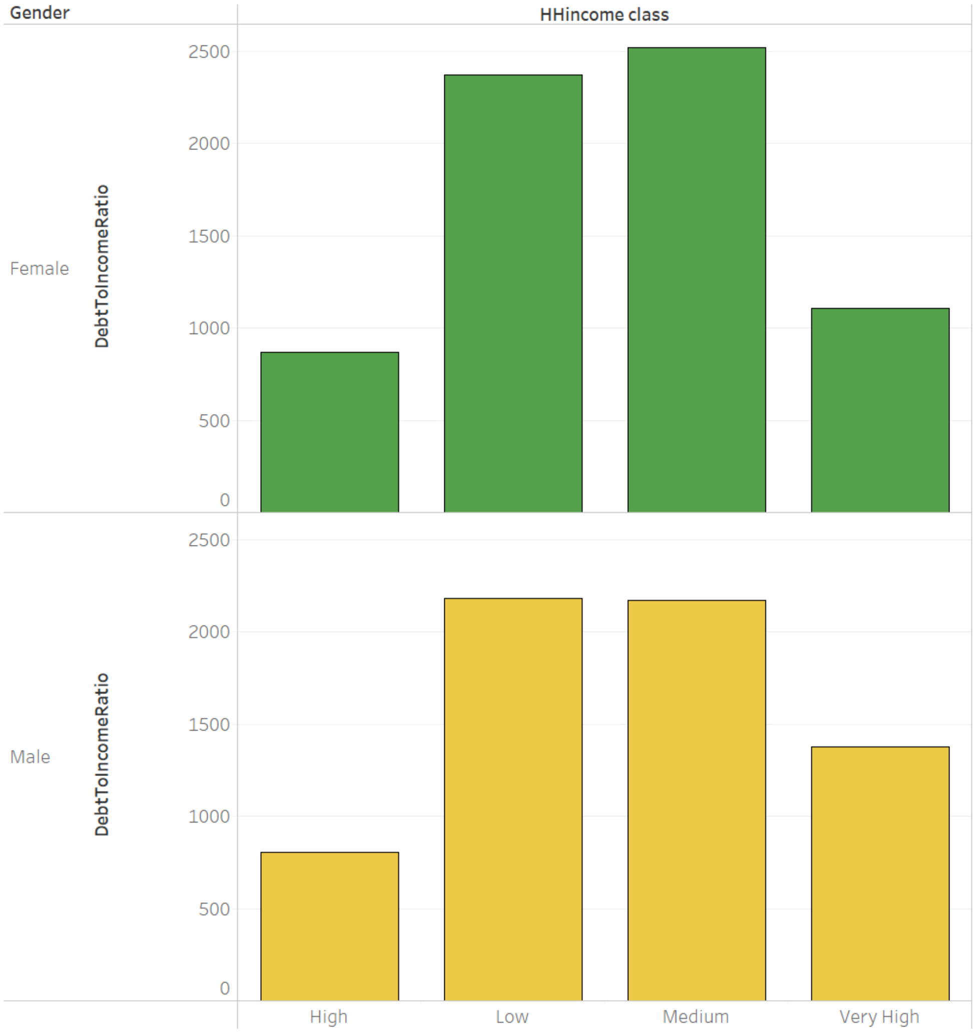
```
Call:
glm(formula = formula, family = binomial, data = customer)

Deviance Residuals: 
     Min        1Q    Median        3Q  
-2.81950  -0.75969  -0.09069   0.25528  
     Max  
 3.14659  

Coefficients:
                           Estimate
(Intercept)                 -3.4995
DebtToIncomeCategoryLow     -0.1422
DebtToIncomeCategoryMedium  -0.7360
DataOverTenureCategoryLow   -4.0136
MultilineYes                 1.0828
EquipmentRentalYes           3.8028
GenderFemale                 2.5697
GenderMale                   2.7574
                           Std. Error
(Intercept)                    1.6378
DebtToIncomeCategoryLow        0.1792
DebtToIncomeCategoryMedium     0.2511
DataOverTenureCategoryLow      0.3076
MultilineYes                   0.1769
EquipmentRentalYes             0.3168
GenderFemale                   1.6195
GenderMale                     1.6199
                           z value Pr(>|z|)
(Intercept)                 -2.137  0.03262
DebtToIncomeCategoryLow     -0.794  0.42748
DebtToIncomeCategoryMedium  -2.931  0.00338
DataOverTenureCategoryLow  -13.049  < 2e-16
MultilineYes                 6.122 9.22e-10
EquipmentRentalYes          12.003  < 2e-16
GenderFemale                 1.587  0.11257
GenderMale                   1.702  0.08871
                              
(Intercept)                *  
DebtToIncomeCategoryLow       
DebtToIncomeCategoryMedium ** 
DataOverTenureCategoryLow  ***
MultilineYes               ***
EquipmentRentalYes         ***
GenderFemale                  
GenderMale                 .  
---
Signif. codes:  
  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1
  ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 1861.3  on 1343  degrees of freedom
Residual deviance: 1001.7  on 1336  degrees of freedom
AIC: 1017.7

Number of Fisher Scoring iterations: 6
```

Visualization of Gender vs Data Over Tenure vs Multiline



This figure was created in tableau which visualized data over the tenure variable against the multiline variable, separated by gender. For customers with yes entries for multline, we had far taller bars and male customers were larger than the female.

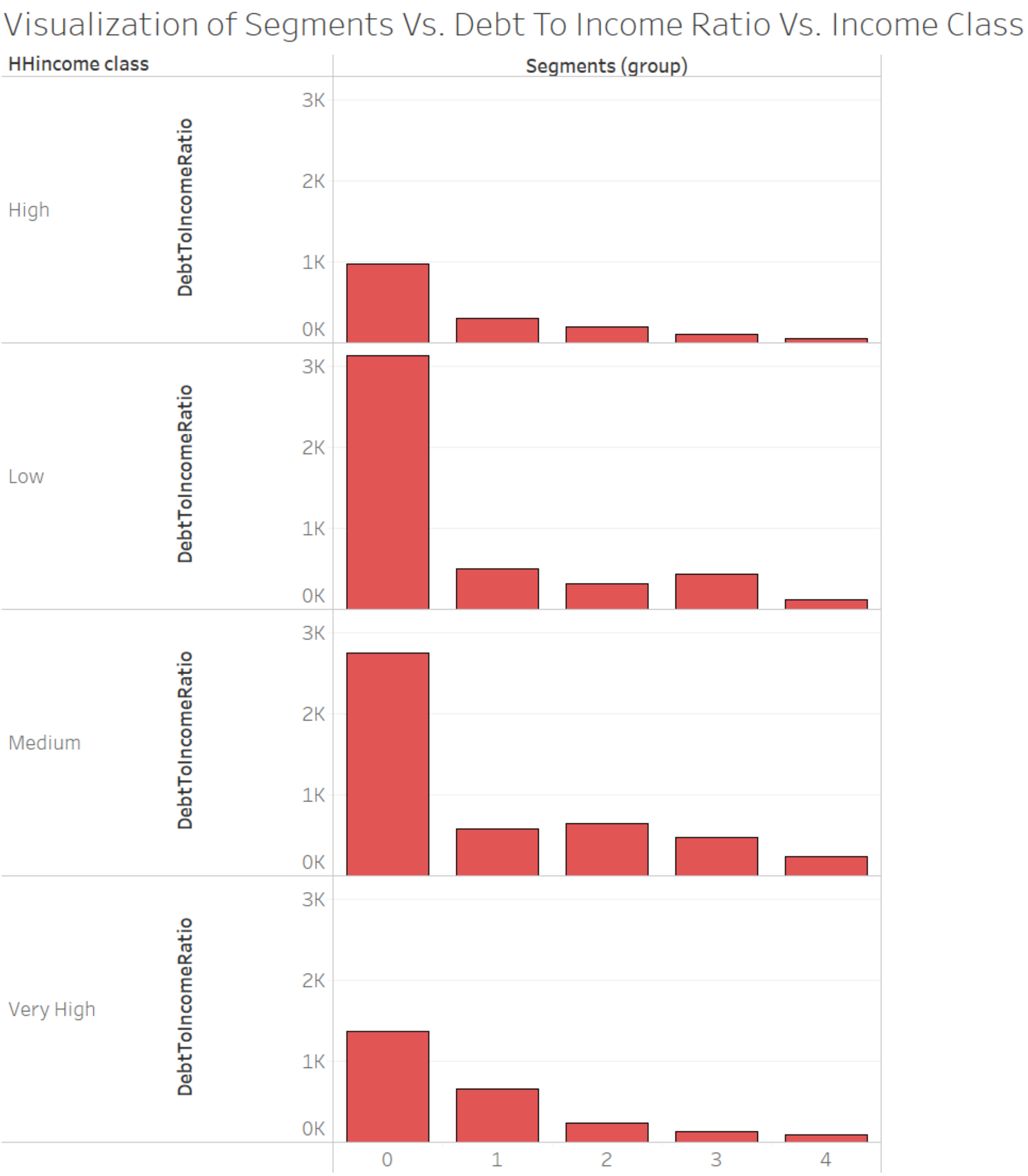
Visualization of Household income vs Sum of Debt to Income ratio seperated by Gender



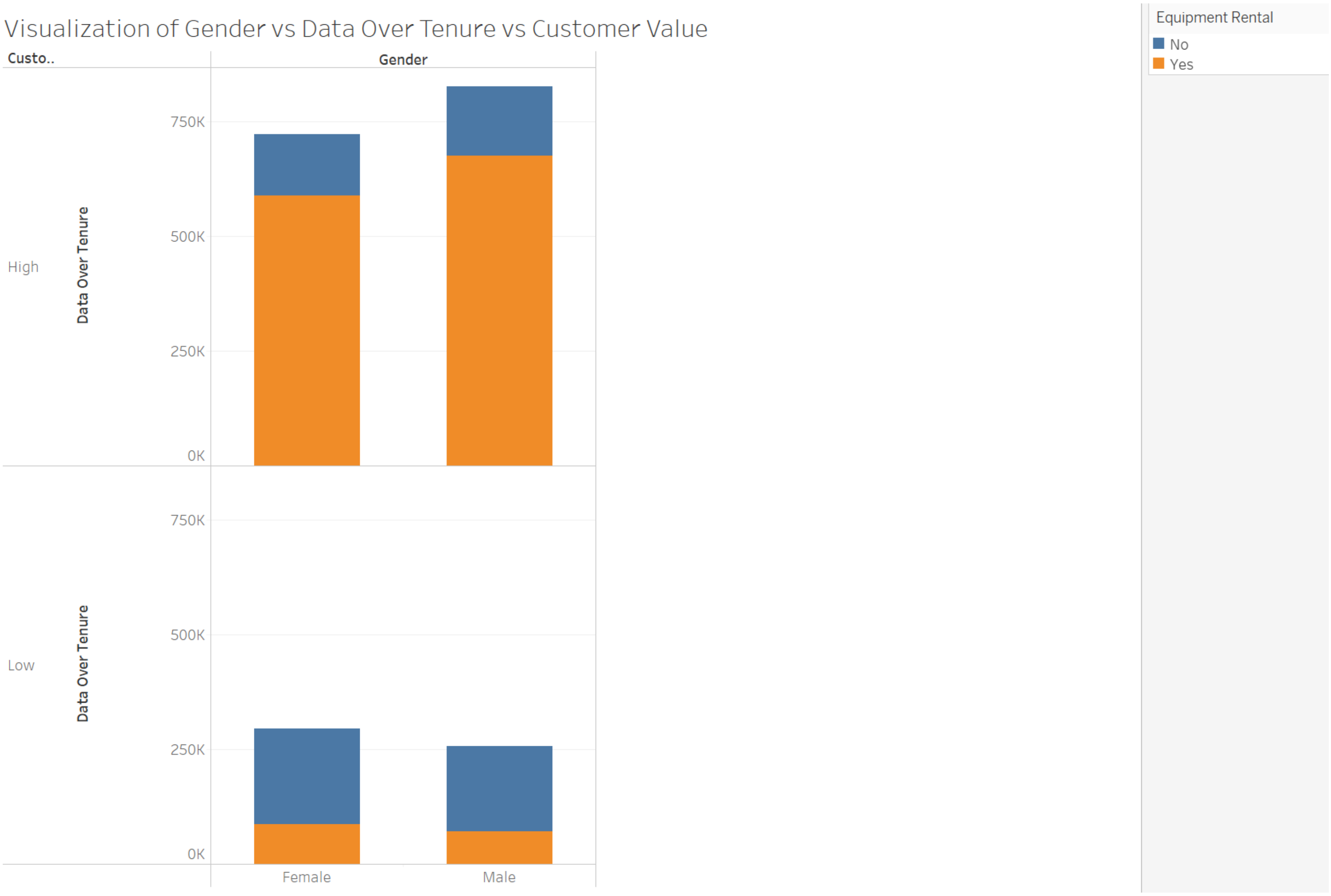
The figure above shows the Household income category, high, low, medium, and very high, against the debt-to-income ratio separated by gender. It is apparent that female entries of medium and very high household income is greater than of the men.

The figure below depicts the segments that we decided to choose to see how many n values we could collect. We created 4 segments

- 0. .Men n= 916
- 1. . Women with Low Debt-to-Income Ratio, Multiline Usage, and Equipment Rental n = 167
- 2. Women with Low Debt-to-Income Ratio, Multiline Usage, Equipment Rental, and High Data Over Tenure n = 106
- 3. . Women with High Debt-to-Income Ratio, Multiline Usage, and Equipment Rental and High-value customer n = 6



For the segments, we can see that segment 1 had the highest n = 167. Women with Low Debt-to-Income Ratio, Multiline Usage, and Equipment Rental. This was achieved by creating bins for all the segments and then plotting against the other variables.



Here is another graph showing the visualization of DataOvertenure vs. customer Value separated by gender. It can be observed that males have a higher sum of dataovertenure and have rent equipment.

Improvements:

- For the rule-based segmentation: I should have included both genders in the segment. It was a pretty naive mode to make a rule to only include females, which creates a huge gap in the market. A lot of data was not excluded when this rule was used but needs to be addressed when improving this project.
- I created only 4 segments. This is too little to find any meaningful insights based on the numbers that we got for n for each segment. If we had more segments, we could have made more graphs.
- Multi-correlation analysis should have been performed beforehand to find even better variables for the segments. We could potentially have variables that could improve our regression models. Possibly some bivariate analysis of our variables would yield a better understanding.
- More time could have been spent to create better visualizations for the project. If you could see the plots i originally made, this could have made the project slightly better.

- This project uses rule-based and regression analysis but is not performed entirely separately. Their steps were influenced by each other thus showing many similarities.

Conclusion:

Through the use of rule-based segmentation and regression model (supervised) we were able to create segments of our dataset. Our objective was to find high value customers, and see if our variables that we choose for our rule-based and regression models could indeed be used to retain customers.

Through many different analyses, we were able to categorize various factors such as Household Income, Customer Value Category, and Debt-to-Income Ratio and create meaningful segments such as Women with Low Debt-to-Income Ratio, Multiline Usage, and Equipment Rental, and Women with High Debt-to-Income Ratio, Multiline Usage, Equipment Rental, and High-value customer.

While further improvement is required we have been able to identify a few things that can be carried over. Variables mentioned in the discussions are one of them.