Segmentation & Profiling Project

Daniel Jackson

2024-05-02

Table of Contents

1. Introduction

This segmentation and profiling section of the project will be a continuation of our data due diligence project. As we did in the data due diligence project, we will be acting as a marketing analytics manager for a telecommunications company. After cleaning the customer data set and creating a subset of the original data, we are now tasked with developing a customer segmentation that can support effective, and economically sound, customer retention efforts.

To accomplish this goal, we will be using two of the three segmentation techniques: rules-based, supervised, or unsupervised. For this project, we will be focusing on rules-based and unsupervised segmentation. We will provide insight as to why we chose those two segmentation methods later.

In this project, we will provide a summary of our key findings and recommendations that we will make as marketing analytics manager. We will highlight notable insights while providing proper conclusions and recommendations based on our findings. Then, we will run through some technical explanations describing the segmentation methods that we chose and highlight detailed findings. Tableau was used to create the visualizations that you will see in this project.

2. Summary of Findings

Using the cleaned-up subset that we created from our data due diligence project, we wanted to turn our focus to segmentation and profiling of our customers. Our goal was to develop a proper segmentation to support effective customer retention efforts. We were tasked with selecting two segmentation methods to properly profile our customer base. We wanted to provide two different approaches, so we decided to choose a rules-based segmentation and an unsupervised segmentation. Each segmentation was selected to allow us the opportunity to compare different segmentation results and to see which one provided the most insight.

2.1 Rules-Based Segmentation Findings

The rules-based approach allowed some control of how we wanted to group our customers. We decided to classify customers based on their total monetary value to the telecommunications company. We classified each customer as having Low value, Medium value, or High value. We will dive more into the technicalities of these classifications in the technical section. What we were shocked to find was that most of the customer base had Medium to High value. Overall, there were not many customers that were of Low value. Of the 4347 customers in our data set, only 174 were classified as Low value. 1957 customers were classified as Medium value and 2216 were classified as High value. We were not expecting to see so many Low value customers, which is a good thing! That means that most of the customers are of good value to the telecommunications company.

For the rules-based segmentation, we segmented the customer base by their total monetary value classification and each of the other qualitative variables in our data set. We first segmented them by region, then by profession, then by home ownership and finally by car brand. This allowed us to see how the data was segmented in a few different lights. Looking at the rules-based segmentations, we were under the impression that the data was going to be best segmented by region. When the data was segmented by the customer's value and the region, it did provide us with a very balanced segmentation. In Figure 1 below, you will see the total number of customers per segmentation by region and customer value:

Segmentation Count	**Customer Value**	**Region**
40	Low	Northeast
33	Low	Southeast
33	Low	West
28	Low	Southwest
40	Low	Midwest
385	Medium	Northeast
389	Medium	Southeast
396	Medium	West
387	Medium	Southwest
400	Medium	Midwest
444	High	Northeast
482	High	Southeast
443	High	West
417	High	Southwest
430	High	Midwest

Figure 1

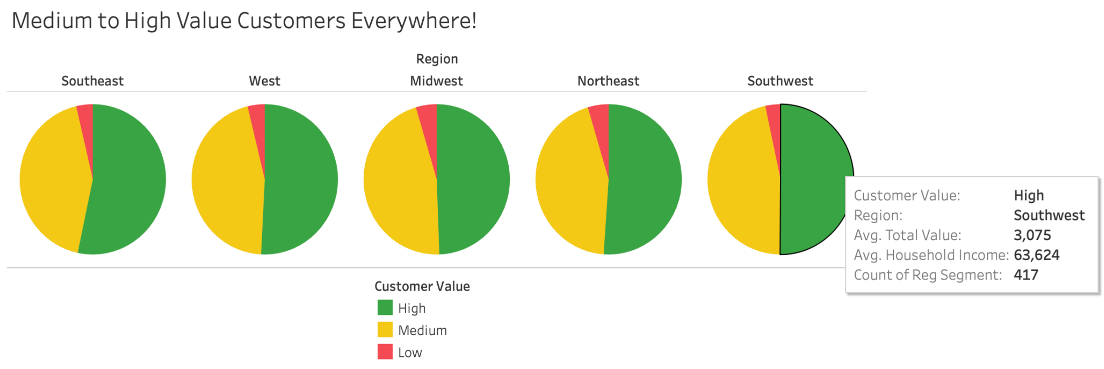

Figure 2

As you can see in *Figure 1* and *Figure 2* above, the Low customer segmentations were of small count, the medium was in the high 300s mark and the High customer counts were all higher than 400. This segmentation represented the data well as we saw that we did not have too many Low-valued customers. This meant that a majority of our customer base falls into the Medium to High value buckets. This can be seen in the table above. What we also noticed is that that higher the customer value per region, the higher the household income was in that grouping.

To see the customer base segmented in a different light, we segmented the customers by their job professions to see how valuable customers were based on their professions. As we saw when the customer base was segmented by region, we were able to see a pattern in the counts: the lower valued customers were in the smaller segments count-wise, and the higher valued customers had higher counts in their respective segments. We did see a little bit of a different story with the profession segmentation.

When we segmented the customer base by the profession, we were able to see a very different segmentation. Below is the graph representing the customers segmented by their value and their profession.

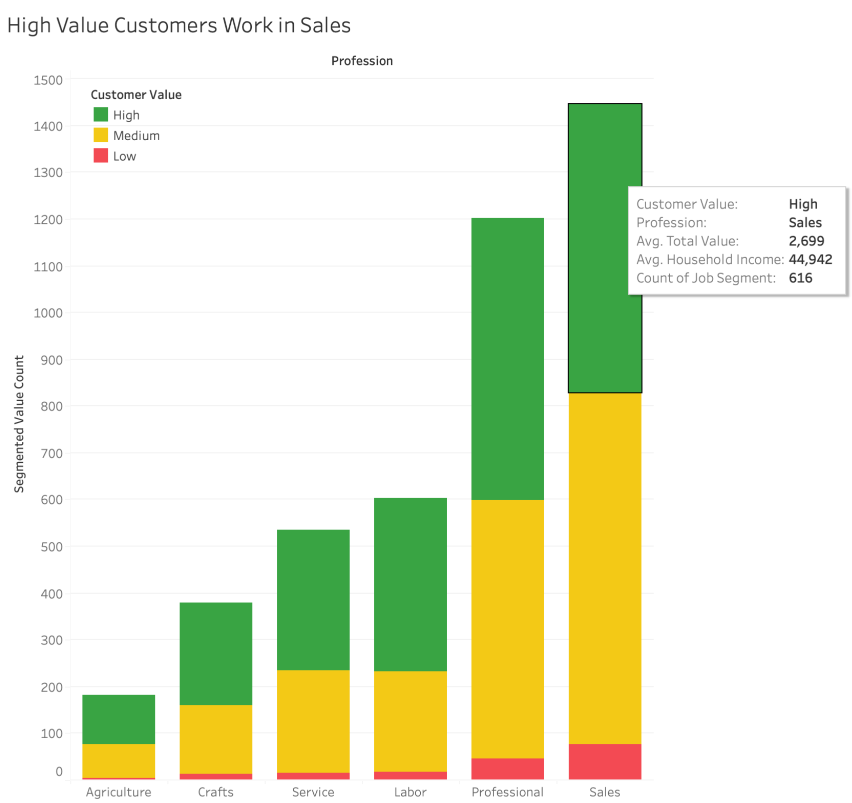

Figure 3

As you can see in graph above, a majority of the High and Medium value customers work as a Professional or work in Sales. This gives us some insight on who we may want to target customers when trying to capture new business.

Based on the two segmentations that we focused on above, we decided to choose the rules-based segmentation based on the customer's profession. This tells a better story of our customer base. Our customers that work in Agriculture account for a very small percentage of our customer base, while those in Sales and Professionals account for a major portion of our customer base. Plus, we can see just how valuable each customer is based on what they do for a profession. The goal behind this segmentation was to group our customer base and create an effective way to retain customers. Based on our findings, we want to ensure that we are retaining those customers that work in Sales and work as Professionals.

We also created some segmentations based on if the customer is a homeowner or what car brand the customer drives. For the purposes of what we selected for our preferred rules-based segmentation, we did not include those visualizations here. If you would like to check out those visuals, you can find them in the appendix under *Figure 10* and *Figure 10.1*.

2.2 Unsupervised Segmentation Findings

In the rules-based segmentation approach, it allowed us some control over how we wanted to segment the customer base. This focused on our qualitative variables more than our quantitative variables. To contrast that approach, we wanted to perform an unsupervised segmentation method to group our customers. This focused more on the quantitative variables in our data set. We used a K-means clustering approach to do our segmentations. We chose a value of six clusters to segment our data. We will get more into why we chose this number in the technical section.

This unsupervised segmentation technique created six groups. Like our rules-based segmentation, we care about a customer's total value to the company. Therefore, we graphed the customers total value against three of the other quantitative variables: total debt, household income, and car values. We used our created clusters in each graph to see where they were properly clustered the best, and to see which customers had the highest value.

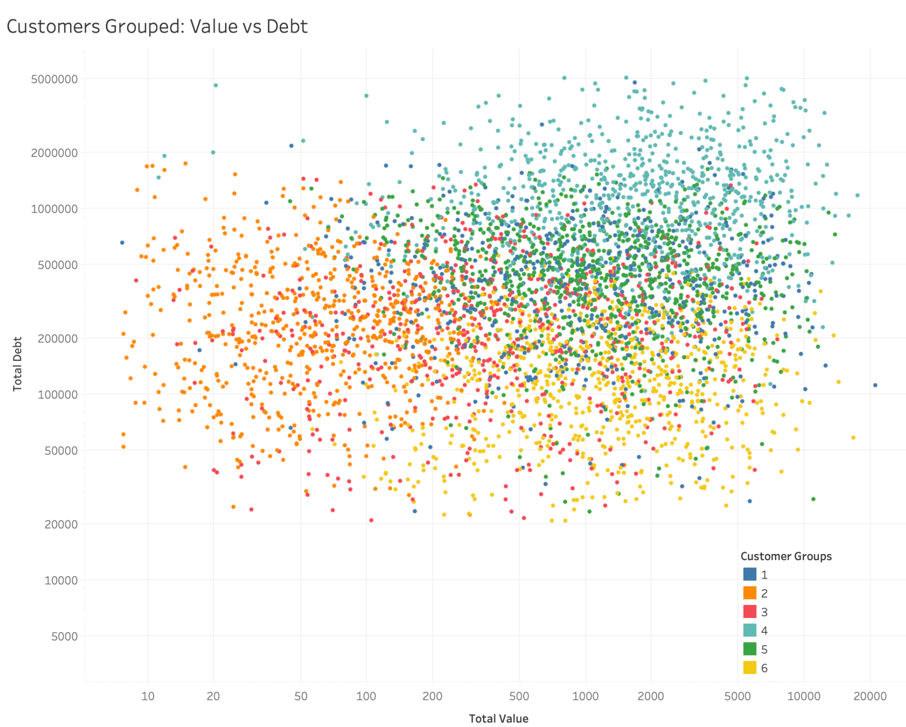

Figure 4

In *Figure 4* above, we grouped our clusters of customers in a visual with total value and total debt, we see that our customer group 4 shows that customers with high value, also have high debt. You can see it more clear below in *Figure 4.1* when just group number 4 is highlighted.

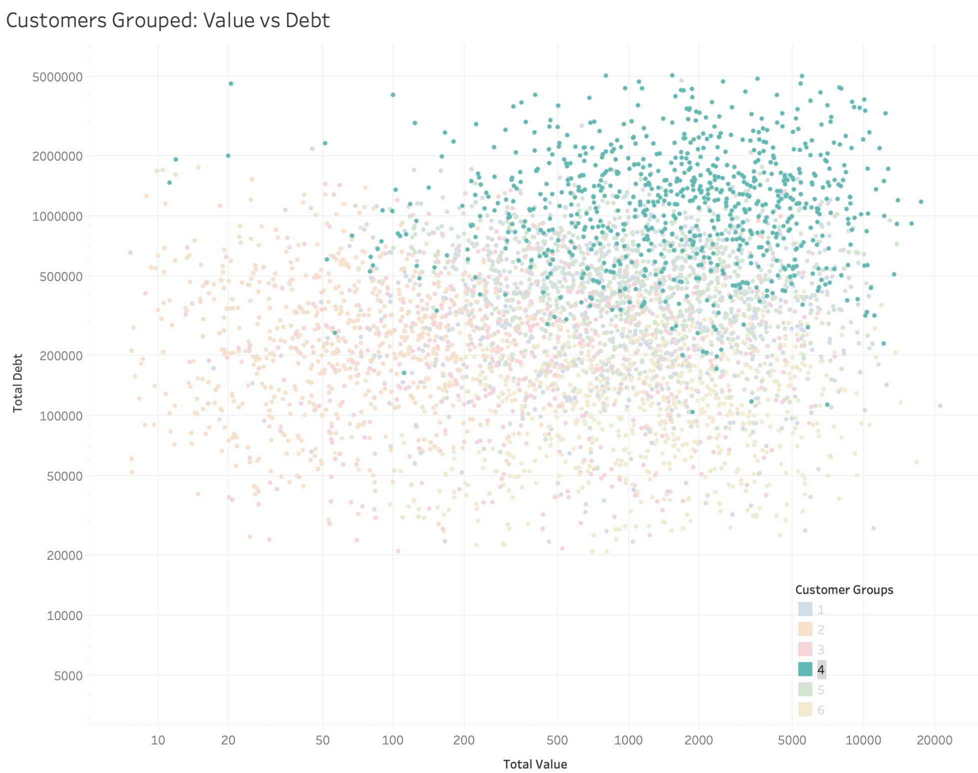

Figure 4.1

Let us look at our customer groupings in *Figure 5* when plotting total value and household income.

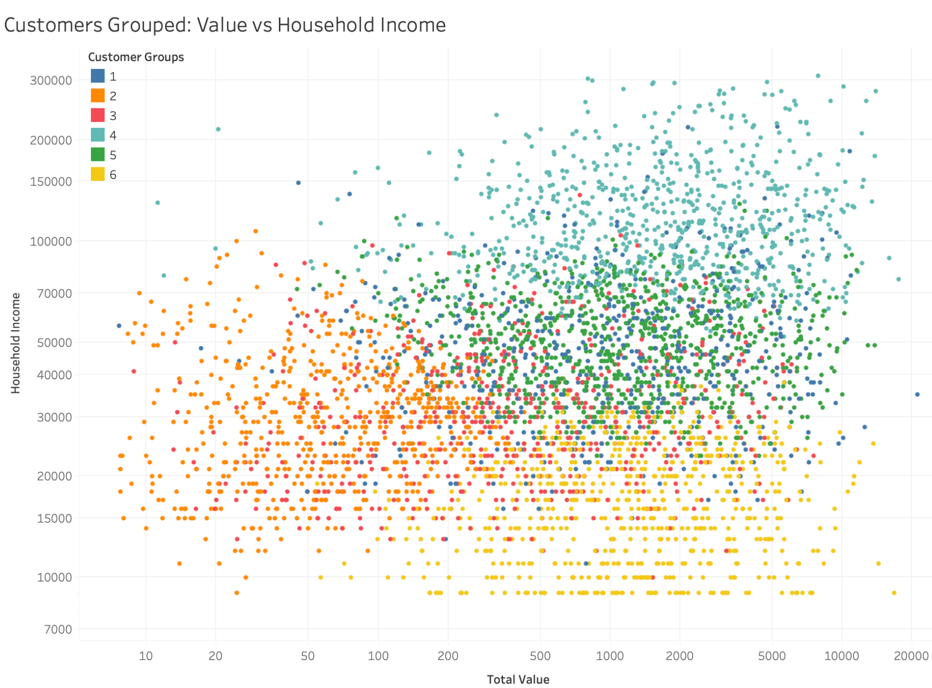

Figure 5

We can see that group 4 is clustered with high customer value. In this visual, we see that those with high household income, also have high customer value. Another interesting cluster has appeared in this visual. Customer group 6 seems to have low household income by medium to high value. This could provide some insight to our customer retention efforts, as we could focus our customer services on those customer groups to ensure that we do not lose them. Below you will see two graphs: one with customer group 4 highlighted (*Figure 5.1*) and one with customer group 6 highlighted (*Figure 5.2*).

Customer Group 4 Highlighted:

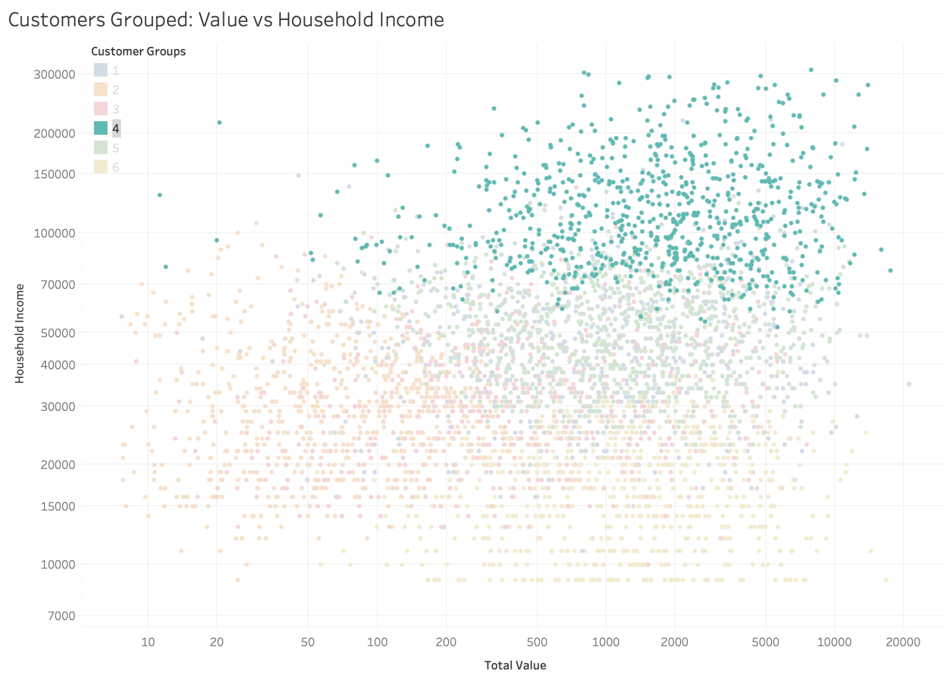

Figure 5.1

Customer Group 6 Highlighted:

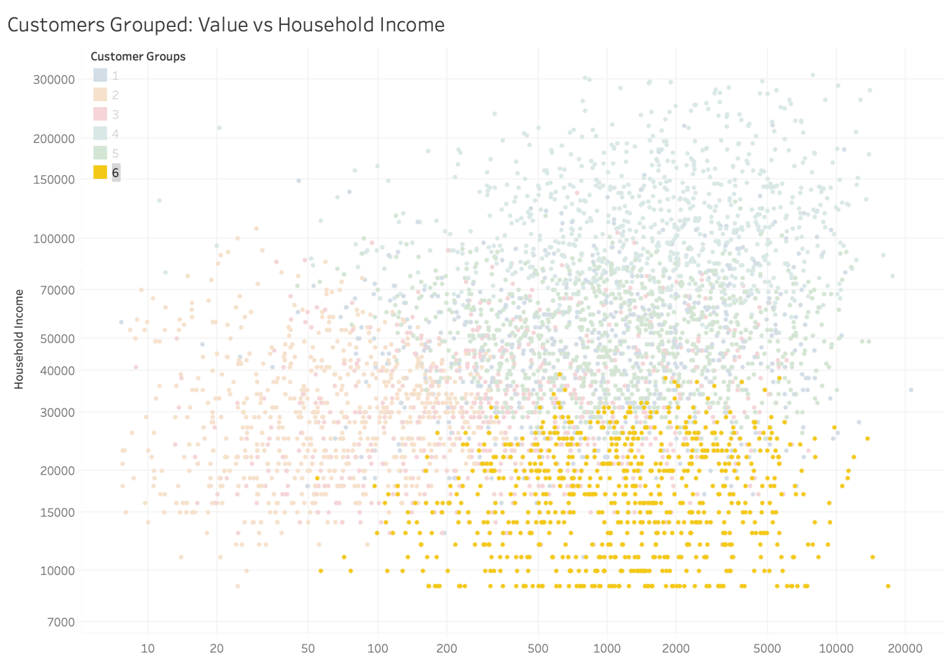

Figure 5.2

Now, let us look at our customer groups when you plot total value and car value together to see if that provides us any insight. This can be seen below in *Figure 6*.

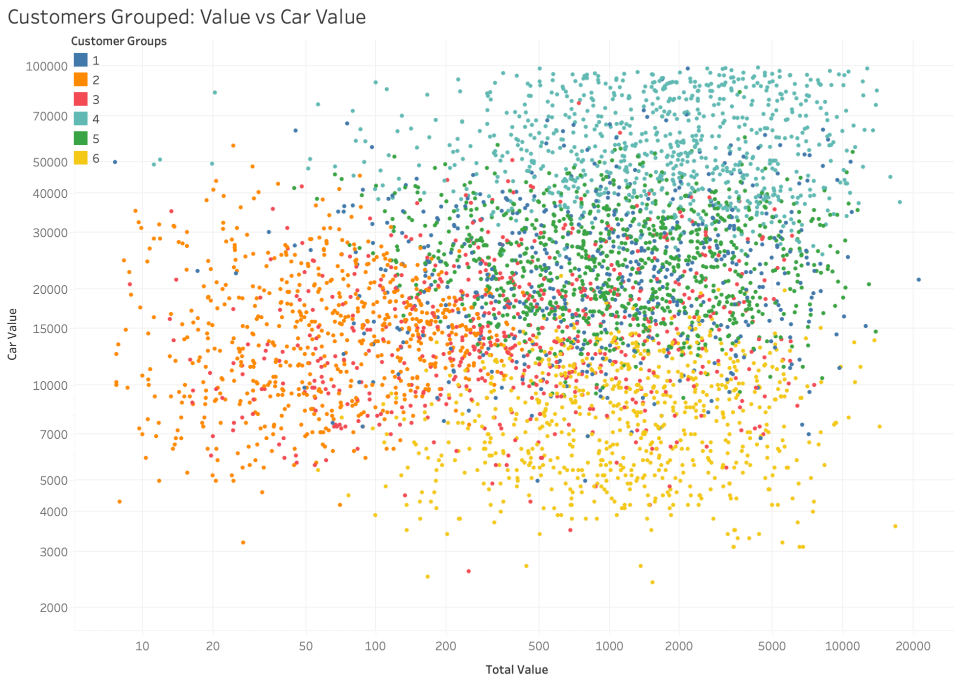

Figure 6

Looking at that 4th customer grouping again, we see that both the total value and car value are high. Based on the three main visuals that we plotted, our customers grouped into that 4th group spend the most as they have high debt, high household income, and high values for their car, plus they are high value customers to the telecommunications company. These customers should be treated with high priority as we do not want to lose customers that can afford to spend money. This can be seen below in *Figure 6.1*.

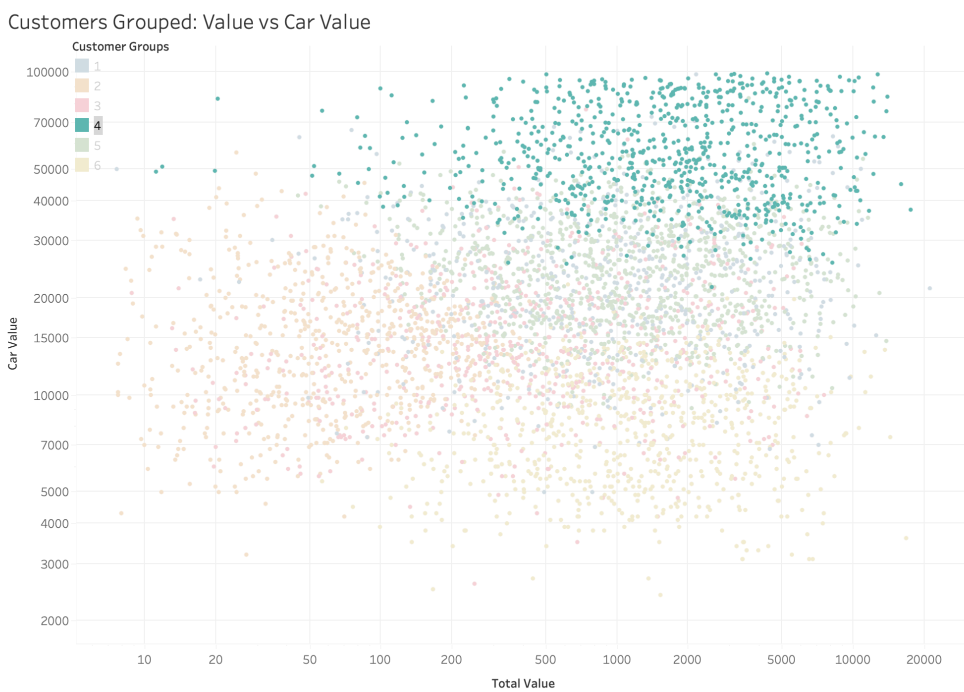

Figure 6.1

Based on what we saw in our unsupervised segmentation, our most valuable customers were grouped into the 4th customer grouping. What the unsupervised segmentation did is group customers based on their quantitative variables and clustered them based on similar tendencies. Using some different visuals, we were able to see how these customers were grouped together when comparing two different quantitative variables.

3. Choice of Segmentation

Looking at each segmentation method, they both offer their respective pros and cons. The rules-based segmentation approach that we chose allowed us to set the parameters for what we consider a customer to be worth and how we should classify them. This customization allowed a lot of control that would set the narrative. However, this method will disregard whatever we want it to disregard based on how we set the rules. The rules-based segmentation visuals that we created were more focused on the qualitative variables in our data set such as region and job profession.

The unsupervised clustering that we did on the data took in all the quantitative variables, such as household income, total value and total debt and grouped the customers. This gave a different perspective on how the customers can be grouped without the control we had in the rules-based segmentation. This method disregards how we want the customers to be grouped. The only control that we had was deciding how many groups that we wanted to see the data divided into.

We chose these two segmentation methods over supervised segmentation. We wanted one segmentation focused on our qualitative variables in the data set and we wanted another segmentation focused on the quantitative variables. We were able to do this with both the rules-based and unsupervised segmentations respectfully. Because we wanted the data to speak for itself, we chose unsupervised over supervised. With unsupervised segmentation, you can see how the different clusters group together by their natural tendencies, rather than selecting an individual response variable to segment around. We did that in our rules-based segmentation with our customer value variable and wanted to offer as different of a segmentation as possible. Thus, we went with unsupervised over supervised.

Our original goal was to group customers in a manner to see how valuable they are to the telecommunication's company. Truthfully, both methods are a great choice as they can be used together to create a marketing plan to ensure that the company retains their high valued customers. If we had to choose just one segmentation method, we would recommend the rules-based method. As you see above in that section, we were able to segment the customer base as Low, Medium, and High valued customers. When we plotted those groupings by job profession, it was clear that the most valuable customers were both in Sales and worked as Professionals. This leads to a simple conclusion and allows the telecommunications company to act on those findings to ensure that they retain those customers that work in Sales and as Professionals. The company can also focus on how they can generate more customers for the future within those respective professional fields.

4. Technical Section

In this technical section, we will walk through our rules-based segmentation and our unsupervised segmentation methods. We will give some behind-the-curtains insight on how we created our segmentations that led us to our findings summarized above. We included our Rmarkdown code in the appendix for those that want to see the coding logic behind each segmentation.

4.1 Rules-Based Methods

We will start with our rules-based segmentation approach. We used RStudio for our segmentation coding. For seeing what customers are worth to the telecommunications company, we looked at the total value variable that we created in our data due diligence project. This variable represents the total value that a customer is worth to the company over their entire time as a customer. In the data due diligence project, we did perform a log transformation on the total value variable to normalize the distribution of the variable for modeling purposes. We wanted to create a rule based qualitative variable from the total value variable and classify each customer as having Low value, Medium value or High value with the telecommunications company. We called this variable "customer_value". If the customer's total value was less than 1/3 of the max value, we classified them as Low. If the customer's total value was between 1/3 and 2/3 of the max value, we classified them as Medium. If the customer's total value is greater than 2/3 of the max value, we classified them as High. This was the rule that we created for the rules-based segmentation.

Once we created our rule for our rules-bases segmentation, we then used RStudio to segment our data set by the four other qualitative variables (region, profession, homeownership, and car brand) in the data to see how the customers were grouped. We used Tableau to visualize each segmentation to see which qualitative variable, when combined with our customer_value variable, best represented the customers demographics. In doing so, we decided that the segmentation with the customer's profession best represented the value of our customers and allowed us to gain insight on which customers need to be retained moving forward.

In *Figure 7* below, you will see our customer value rules-based segmentation represented by the average income for each profession. This visualization is a good way to see that our high value customers, no matter the profession, have higher household incomes. And the lower valued customers have the lower household incomes.

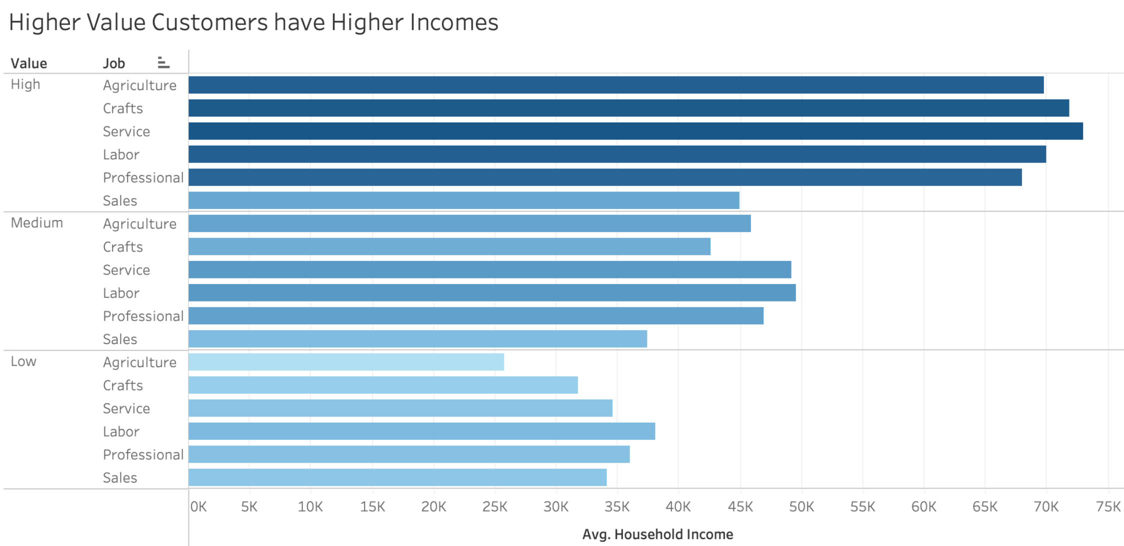

Figure 7

4.2 Unsupervised Segmentation

For our second segmentation approach, we did so with an unsupervised segmentation. We already created a rules-based segmentation based on a customer's total value being classified as Low, Medium, or High. We wanted to then focus on the quantitative variables in our data set. We used a K-means clustering technique on the quantitative variables to segment our data. The quantitative variables that we used were number of pets, household size, car value, household income, total debt, and total value. We compared segment solutions from K = 1 to K = 12 and chose the best K value to represent our data. In *Figure 8* below, you will see the plot of the total within-cluster sum of squares from the K = 1 to the K = 12 range.

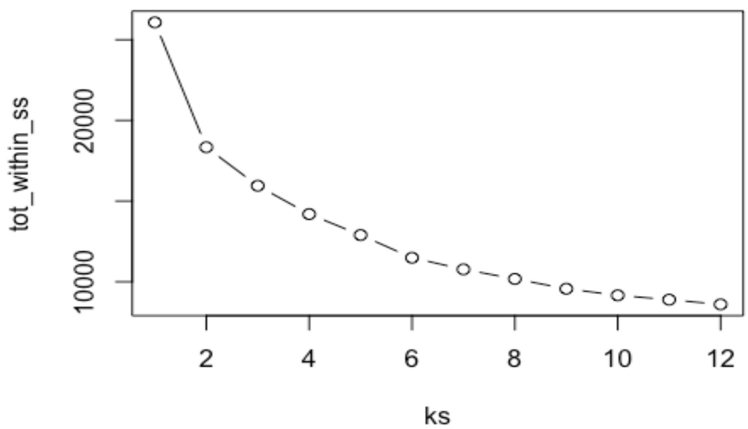

Figure 8

We can see that as the number of K's increase, the total within-cluster sum of squares decreases. We see the biggest drop off in the sum of squares occurs from K =1 to K = 2. After K = 2, the drop in the sum of squares is less dramatic. We ended up choosing K = 6 in our clustering technique as our optimal number of clusters. Using the elbow method when analyzing the plot, we can see that the last, somewhat dramatic, drop off in the sum of squares is at K = 6. The drop off after K = 6 tends to even out as K increases.

Using K = 6 in our clustering method, we created six groupings of our customers based on our quantitative variables. We then used Tableau to compare different variables against our total customer value variable and let our clusters represent the data points. This allowed us to see how our customers were grouped together when we compared different quantitative variables.

In *Figure 9* below, you will see a line graph showing the average household income and the average total value for each of the 6 customer groups. As you can see, customer group 4 has the highest average customer value, and the highest average income. This group make the most, and they enjoy spending that money with the telecommunications company.

Figure 9

5. Conclusion

After comparing our two segmentation methods, we chose our rules-based segmentation method when we grouped our customers by Low, Medium, and High value by their job profession. This allowed us to see where our most valuable customers were in our data set. We saw that those that worked in Sales and as Professionals were our most valuable customers. Using this information, we recommended that the telecommunications implement a marketing program to ensure that those customers are retained. These customers are the most valuable and need to be kept moving forward to ensure that they do not take their business elsewhere. It also showed us that customers that worked in Agriculture and Crafts had a lower count of high valued customers. We see that those working in Service and Labor are in that middle ground between Sales/Professionals and Agriculture/Crafts. Assuming that there is enough marketing funding after we ensure our most valuable customer base is retained, we want to make another recommendation to capture new business. The company could devise a marketing program geared towards those customers working in Agriculture, Crafts, Service and Labor to generate more Medium value customers to High value customers.

6. Appendix

Visualizations

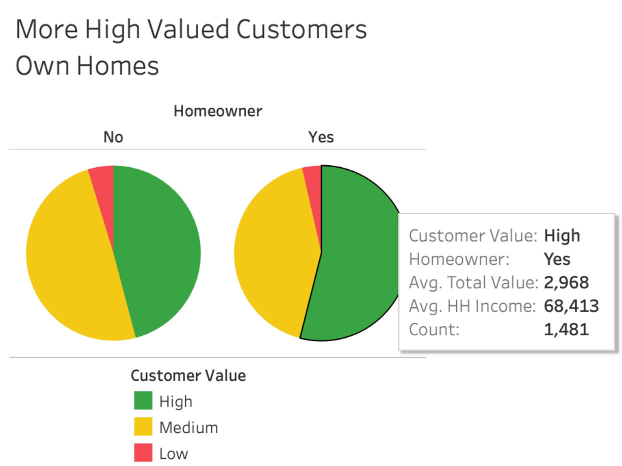

Figure 10

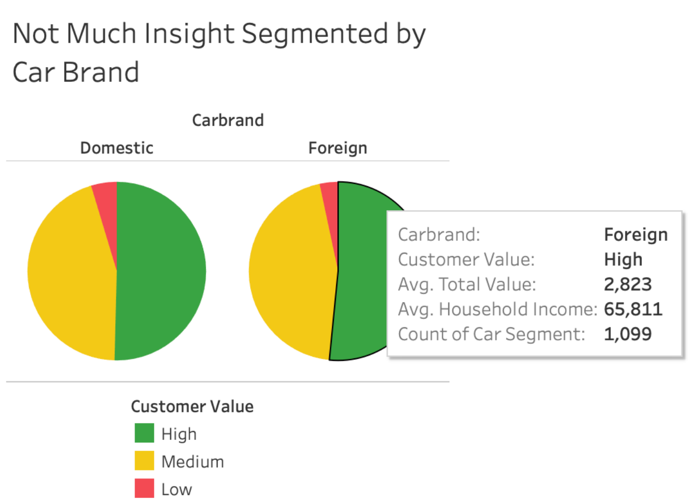

Figure 10.1

Code

```
# Packages used
library(dplyr)
library(tidyr)

# Read in data
sub_df = read.csv("combined_customer_data.csv")

# Check dimension of data frame
dim(sub_df)
# 4347 observations
# 11 columns

# Let us remove X column
sub_df = sub_df[, -which(names(sub_df) == "X")]

dim(sub_df)
# 4347 observations
# 10 columns

str(sub_df)
# Let us change numberpets, household size to numeric vectors and change
homeowner to Yes or No (1 for yes, 0 for no)
sub_df = sub_df %>%
  mutate(
    numberpets = as.numeric(numberpets),
    householdsize = as.numeric(householdsize),
    homeowner = ifelse(homeowner == 1, "Yes", "No")
  )

str(sub_df)

# Check min, max and average
min(sub_df$log_tot_value)
# 2.03 min value
min_val = min(sub_df$log_tot_value)

max(sub_df$log_tot_value)
# 9.95 max value
max_val = max(sub_df$log_tot_value)

mean(sub_df$log_tot_value)
# 6.46 average value
mean_val = mean(sub_df$log_tot_value)

# 1/3 threshold
one_third_val = ((1/3) * max_val)

# 2/3 threshold
```

```
two_third_val = (2 * one_third_val)


# Let us create customer_value qualitative variable using our breaks that we created using the cut function
sub_df$customer_value = cut(sub_df$log_tot_value, breaks=c(0, one_third_val, two_third_val, max_val), labels=c("Low", "Medium", "High"))
unique(sub_df$customer_value)
# Check if any NA values in customer_value variable
any(is.na(sub_df$customer_value))
# FALSE

table(sub_df$customer_value)

# Now let us combine this rule with a few variables

# Region
unique(sub_df$region)
sub_df = sub_df %>% mutate(reg_segment = case_when(
      customer_value == "Low" & region == "northeast" ~ 1,
      customer_value == "Medium" & region == "northeast" ~ 2,
      customer_value == "High" & region == "northeast" ~ 3,
      customer_value == "Low" & region == "southeast" ~ 4,
      customer_value == "Medium" & region == "southeast" ~ 5,
      customer_value == "High" & region == "southeast" ~ 6,
      customer_value == "Low" & region == "west" ~ 7,
      customer_value == "Medium" & region == "west" ~ 8,
      customer_value == "High" & region == "west" ~ 9,
      customer_value == "Low" & region == "southwest" ~ 10,
      customer_value == "Medium" & region == "southwest" ~ 11,
      customer_value == "High" & region == "southwest" ~ 12,
      customer_value == "Low" & region == "midwest" ~ 13,
      customer_value == "Medium" & region == "midwest" ~ 14,
      customer_value == "High" & region == "midwest" ~ 15
    ))

# 1 = low and northeast
# 2 = medium and northeast
# 3 = high and northeast
# 4 = low and southeast
# 5 = medium and southeast
# 6 = high and southeast
# 7 = low and west
# 8 = medium and west
# 9 = high and west
# 10 = low and southwest
# 11 = medium and southwest
# 12 = high and southwest
# 13 = low  and midwest
# 14 = medium and midwest
```

```
# 15 = high and midwest

# Check for any NA values in segment variable
any(is.na(sub_df$reg_segment))
# FALSE

# Summarize results of each segment
sub_df %>%
  add_count(reg_segment) %>%
  group_by(reg_segment,n) %>%
  summarise_all("mean")

# Remove qualitative variables
sub_df %>%
  select(-c("region","homeowner","jobcategory", "carbrand",
"customer_value")) %>%
  add_count(reg_segment) %>%
  group_by(reg_segment,n) %>%
  summarise_all("mean")

# Homeowner
unique(sub_df$homeowner)
sub_df = sub_df %>% mutate(home_segment = case_when(
      customer_value == "Low" & homeowner == "Yes" ~ 1,
      customer_value == "Medium" & homeowner == "Yes" ~ 2,
      customer_value == "High" & homeowner == "Yes" ~ 3,
      customer_value == "Low" & homeowner == "No" ~ 4,
      customer_value == "Medium" & homeowner == "No" ~ 5,
      customer_value == "High" & homeowner == "No" ~ 6
    ))

# 1 = low and homeowner
# 2 = medium and homeowner
# 3 = high and homeowner
# 4 = low and non-homeowner
# 5 = medium and non-homeowner
# 6 = high and non-homeowner

# Summarize results of each segment
sub_df %>%
  add_count(home_segment) %>%
  group_by(home_segment,n) %>%
  summarise_all("mean")

# Remove qualitative variables
sub_df %>%
  select(-c("region","homeowner","jobcategory", "carbrand",
"customer_value")) %>%
  add_count(home_segment) %>%
```

```r
  group_by(home_segment,n) %>%
  summarise_all("mean")

# Job Category
unique(sub_df$jobcategory)
sub_df = sub_df %>% mutate(job_segment = case_when(
      customer_value == "Low" & jobcategory == "Professional" ~ 1,
      customer_value == "Medium" & jobcategory == "Professional" ~ 2,
      customer_value == "High" & jobcategory == "Professional" ~ 3,
      customer_value == "Low" & jobcategory == "Sales" ~ 4,
      customer_value == "Medium" & jobcategory == "Sales" ~ 5,
      customer_value == "High" & jobcategory == "Sales" ~ 6,
      customer_value == "Low" & jobcategory == "Labor" ~ 7,
      customer_value == "Medium" & jobcategory == "Labor" ~ 8,
      customer_value == "High" & jobcategory == "Labor" ~ 9,
      customer_value == "Low" & jobcategory == "Agriculture" ~ 10,
      customer_value == "Medium" & jobcategory == "Agriculture" ~ 11,
      customer_value == "High" & jobcategory == "Agriculture" ~ 12,
      customer_value == "Low" & jobcategory == "Service" ~ 13,
      customer_value == "Medium" & jobcategory == "Service" ~ 14,
      customer_value == "High" & jobcategory == "Service" ~ 15,
      customer_value == "Low" & jobcategory == "Crafts" ~ 16,
      customer_value == "Medium" & jobcategory == "Crafts" ~ 17,
      customer_value == "High" & jobcategory == "Crafts" ~ 18
    ))

# 1 = low and professional
# 2 = medium and professional
# 3 = high and professional
# 4 = low and sales
# 5 = medium and sales
# 6 = high and sales
# 7 = low and labor
# 8 = medium and labor
# 9 = high and labor
# 10 = low and agriculture
# 11 = medium and agriculture
# 12 = high and agriculture
# 13 = low  and service
# 14 = medium and service
# 15 = high and service
# 16 = low and crafts
# 17 = medium and crafts
# 18 = high and crafts

# Summarize results of each segment
sub_df %>%
  add_count(job_segment) %>%
  group_by(job_segment,n) %>%
  summarise_all("mean")
```

```
# Remove qualitative variables
sub_df %>%
  select(-c("region","homeowner","jobcategory", "carbrand",
"customer_value")) %>%
  add_count(job_segment) %>%
  group_by(job_segment,n) %>%
  summarise_all("mean")

# Car Brand
unique(sub_df$carbrand)
sub_df = sub_df %>% mutate(car_segment = case_when(
      customer_value == "Low" & carbrand == "Domestic" ~ 1,
      customer_value == "Medium" & carbrand == "Domestic" ~ 2,
      customer_value == "High" & carbrand == "Domestic" ~ 3,
      customer_value == "Low" & carbrand == "Foreign" ~ 4,
      customer_value == "Medium" & carbrand == "Foreign" ~ 5,
      customer_value == "High" & carbrand == "Foreign" ~ 6
    ))

# 1 = low and domestic
# 2 = medium and domestic
# 3 = high and domestic
# 4 = low and foreign
# 5 = medium and foreign
# 6 = medium and foreign

# Summarize results of each segment
sub_df %>%
  add_count(car_segment) %>%
  group_by(car_segment,n) %>%
  summarise_all("mean")

# Remove qualitative variables
sub_df %>%
  select(-c("region","homeowner","jobcategory", "carbrand",
"customer_value")) %>%
  add_count(car_segment) %>%
  group_by(car_segment,n) %>%
  summarise_all("mean")
```

Now that we have created 4 different rule-based segments, let us use Tableu to visualize each to help us select best rules-bases segment to represent our customers.

```
sub_df %>%
  write.csv("segmentation_df.csv")

# Let us re-read data in
sub_df = read.csv("combined_customer_data.csv")

sub_df = sub_df[, -which(names(sub_df) == "X")]
```

```r
sub_df = sub_df %>%
  mutate(
    numberpets = as.numeric(numberpets),
    householdsize = as.numeric(householdsize),
    homeowner = ifelse(homeowner == 1, "Yes", "No")
  )

# Now, let us create subset of data that only has numeric variables
num_df = sub_df[sapply(sub_df, is.numeric)]

# Let us scale each variable before we perform K-means clustering
# Assuming num_df contains only numeric variables
scaled_df = as.data.frame(scale(num_df))

# Now, let us take a look and see what K should equal for our clustering
segmention
ks = 1:12
tot_within_ss = sapply(ks, function(k) {
    set.seed(1223)
    cl = kmeans(scaled_df, k)
    cl$tot.withinss
})
plot(ks, tot_within_ss, type = "b")

# Based on plot, we see biggest drop off from K = 1 to K = 2.
# We see that once K = 6, the drop in tot_within_ss value tends to drop
relatively constantly as the value of K increases. Therefore, let us use K =
6 for our segmentation.

set.seed(1223)
num_clusters = 6
kclust = kmeans(scaled_df, centers = num_clusters, nstart = 10)

#add segments to original dataset
sub_df$kmeans_six = as.factor(kclust$cluster)

sub_df %>%
  select(-c("region","homeowner", "jobcategory", "carbrand")) %>%
  add_count(kmeans_six) %>%
  group_by(kmeans_six,n) %>%
  summarise_all("mean")

# Add K = 5 and K = 7 segments to data set so we can compare them visually
# K = 5
set.seed(1223)
num_clusters = 5
kclust = kmeans(scaled_df, centers = num_clusters, nstart = 10)

#add segments to original dataset
```

```r
sub_df$kmeans_five = as.factor(kclust$cluster)

sub_df %>%
  select(-c("region","homeowner", "jobcategory", "carbrand")) %>%
  add_count(kmeans_five) %>%
  group_by(kmeans_five,n) %>%
  summarise_all("mean")

# K = 7
set.seed(1223)
num_clusters = 7
kclust = kmeans(scaled_df, centers = num_clusters, nstart = 10)

#add segments to original dataset
sub_df$kmeans_seven = as.factor(kclust$cluster)

sub_df %>%
  select(-c("region","homeowner", "jobcategory", "carbrand")) %>%
  add_count(kmeans_seven) %>%
  group_by(kmeans_seven,n) %>%
  summarise_all("mean")

# Export sub_df to CSV file again for unsupervised segmentation
sub_df %>%
  write.csv("unsup_segmentation_df.csv")
```