Our first step was to read in our data set into RStudio. We see that our data set contains 5000 observations and 60 variables. To make the analysis easier for coding purposes, we changed all the variables to lower case. We then looked to see how many columns in the data set had NA values. We saw that there was a variable named “x” that had NA values for every observation. We went ahead and just removed that entire column.

The first variable that we chose to include in our subset was number of pets, which tells us how many pets each customer has. The number of pets variable had 6 NA values. The number of cats, number of dogs and number of birds variables had 7, 8 and 34 NA values respectively. We assumed that the NA values in these columns equaled 0. After mutating all the NA values to 0, we just added the number of cats, the number of dogs and the number of birds variables together in the number of pets column. This gave us a more accurate idea to how many total animals each customer owns.

The second variable that chose for our subset was household size, which tells us how many people live in the house. We saw that household size had 8 NA values. To fill those NA values, we found the average household size based on if customers are married or unmarried. We found that the average household size for married couples is about 3 while the average household size for unmarried couples is about 1. We then imputed those into the NA values based the customer’s marital status.

The third variable that we chose for our subset was homeowner. The homeowner variable is a binary column that says either 1 for homeowner or 0 for non-homeowner. Since this is a binary column, we decided to replace the NA values for the mode for the column. The mode was 1 meaning that more customers having own a home compared to those who do not.

The fourth variable that we chose for our subset was job category, which tells us what each customer does for work. When looking at the job category variable, we found multiple empty strings. We confirmed this by checking the unique values for job category. Those unique values were: Professional, Sales, Labor, Agriculture, Service, Crafts, and an empty string. After mutating those empty strings to NA, we counted that there were 15 NA values in the column. We then looked at the highest count of job category values for those that are in the union and for those that are not in the union. For customers in the union, the “Professional” job category was the highest count. For non-union customers, the “Sales” job category had the highest count. Therefore, we imputed “Professional” for all NA values for union customers, and we imputed “Sales” for all non-union customers.

The fifth variable that we chose for our subset was region. The region variable had 5 unique values: 1, 2, 3, 4 and 5. Each number represented a region that the customer lives in. For visual purposes, we wanted to convert those numbers to their respective regions. Using the data dictionary that we were provided, we were able to see what region each number represented:

* 1: Northeast
* 2: Midwest
* 3: West
* 4: Southwest
* 5: Southeast

We then converted each number to the actual name of the region.

The sixth variable that we chose for our subset was household income, which tells us what the household income is for each customer. We found that the household income variable was a character vector, meaning each observation was represented by a string of characters. Since we wanted this variable to be numeric, we removed all non-numeric characters and then converted the column into a numeric column. We then looked at the distribution of the household income. The distribution was right skewed. Therefore, we tried a log transformation on the data, which made the distribution more normal. We then created a new variable to take the log of household income. We looked at a boxplot of the transformed variable to see if there were any outliers. We then removed those outliers from the data set as a whole.

The seventh variable that we chose for our subset was one that we created from multiple variables in the data set. We wanted to look at the total value of each customer over their tenure with the company. We looked at three variables: equipment over tenure, voice over tenure and data over tenure. These three variables gave us the dollar amount that each customer paid over their lifetime with the company. After converting the three variables to numeric variables, we added them together to find a total dollar value for each customer. We checked the distribution of the total value for each customer and saw that it was heavily skewed right. We then used a log transformation on the variable. This made the distribution slightly skewed left. However, this left skew was much less skewed than the right skewedness in the non-transformed total value distribution. After we did the log transformation, we then analyzed a boxplot and removed any outliers present.

The eight variable that we chose for our subset was one that we also created from other variables in the data set. We wanted to look at the total debt for each customer by adding the credit debt and other debt variables together. Using the data dictionary, we saw that both columns were expressed in $100,000 USD. Therefore, we multiplied each variable by 100000 and added them together to create a total debt variable. We saw that the distribution was skewed heavily to the right. We removed any customers that had zero debt (which was only one) and did a log transformation. This made our distribution much more normal. We then removed any outliers present.

The ninth variable that we chose for our subset was car brand, which tells us if a customer’s vehicle is either domestic or foreign. Along with the domestic and foreign classifications in the variable, there was also a “-1” classification. This was for any customer that did not have a vehicle. We converted any “-1” classification to be represented by “None”.

The tenth and final variable that we chose for our subset was car value, which tells us the value of each customer’s vehicle. For those that do not own a vehicle, there was a “$(1,000.00)” string. We then mutated the column to change the “$(1,000.00)” string to “0” if the customer’s car brand was equal to “None”. We then removed non-numeric characters and converted the car value variable to a numeric column. We then looked at the distribution of the variable and found that it is right skewed. Since we wanted to try a log transformation, we decided to remove any customer that does not have a vehicle and only focus on those that do have one. After doing so, we found that a log transformation did make the distribution much more normal. We then removed any outliers present.

Our cleaned-up subset of the original data has 4347 observations and 10 variables. Below you will find a table that shows each of the 10 variables that we have created and the characteristics of each variable. The first 4 variables are the qualitative variables, and the last 6 variables are the quantitative ones.

|  |  |  |
| --- | --- | --- |
| **Variable** | **Data Type** | **Variable Type** |
| Homeowner | Qualitative | Binary |
| Job Category | Qualitative | Nominal |
| Region | Qualitative | Nominal |
| Car Brand | Qualitative | Nominal |
| Number of Pets | Quantitative | Discrete |
| Household Size | Quantitative | Discrete |
| Log Household Income | Quantitative | Continuous |
| Log Total Value | Quantitative | Continuous |
| Log Total Debt | Quantitative | Continuous |
| Log Car Value | Quantitative | Continuous |

Now we will look at some visualizations using Tableau.