Data Due Diligence

Nelson Tran

Master of Data Science, Merrimack College

DSE5004 Visual Data Exploration

Dr. Michael Dupin

April 13th, 2024

Summary

Upon being given a dataset that contained a sample of 5000 customers. I realized that there is little to no customer analytics. I am a strong believer in evidence-based management, so I conducted a comprehensive assessment of the customer base which includes a general overview of demographic and financial-related variables. The demographic-related variables include, "GeoRegion (Var 1)," "AgeGroups (Var 2)," "OtherPets (Var 3)," "RegionVotingStatus (Var 4)," "TownSizeGroup (Var 5)," and "TVWatchersPC (Var 10)." The financial-related variables include, "TotalDebt (Var 6)," "AvgCreditItem (Var 7)," "CartoIncome (Var 8)," and "MonthlyPhoneBill (Var 9)." The variables obtained from the data set are created to understand the customer base more and variables that I thought would be interesting to have more insight into.

Created Variables/Features

Demographic-Related Variables

When looking at the data set, I noticed that the region and town size were using factors instead of the names of the regions or town size. I converted both variables because I thought that would be more information to look at the actual town size and region that a customer is from instead of both variables just having a number in place of where they are from.

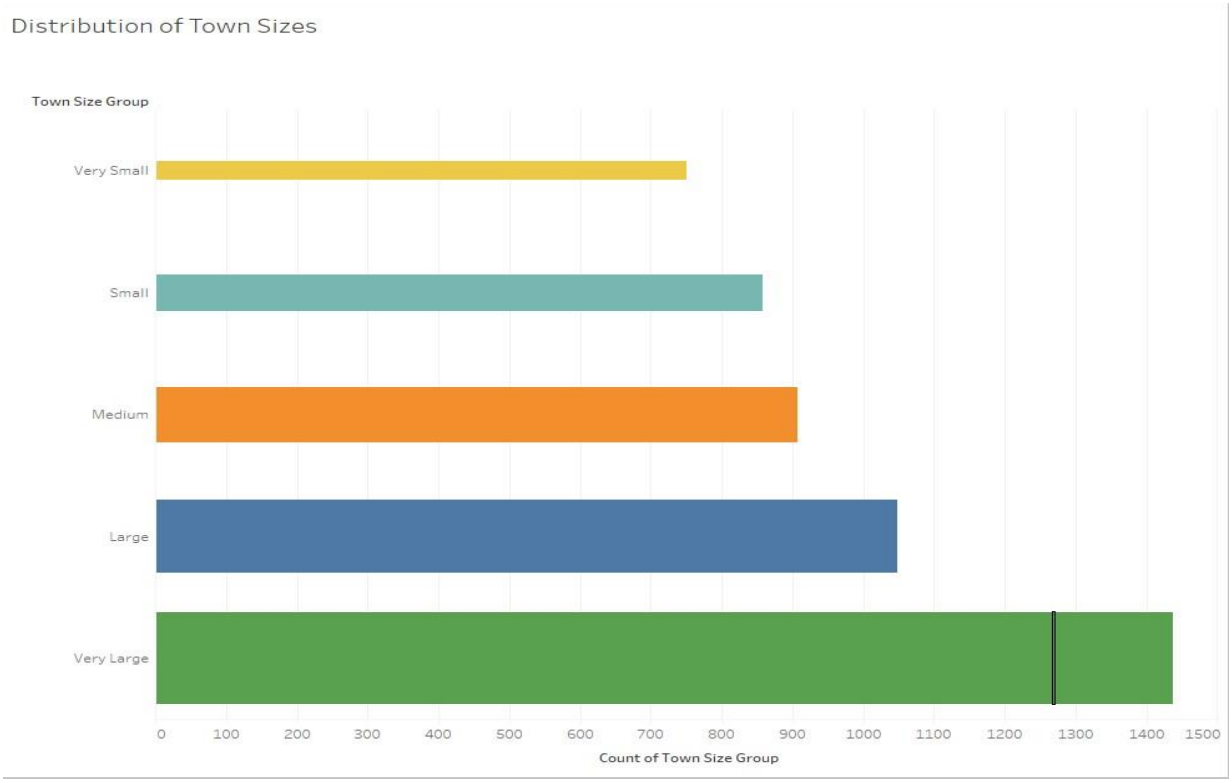

Figure 1: Distribution of Town Sizes

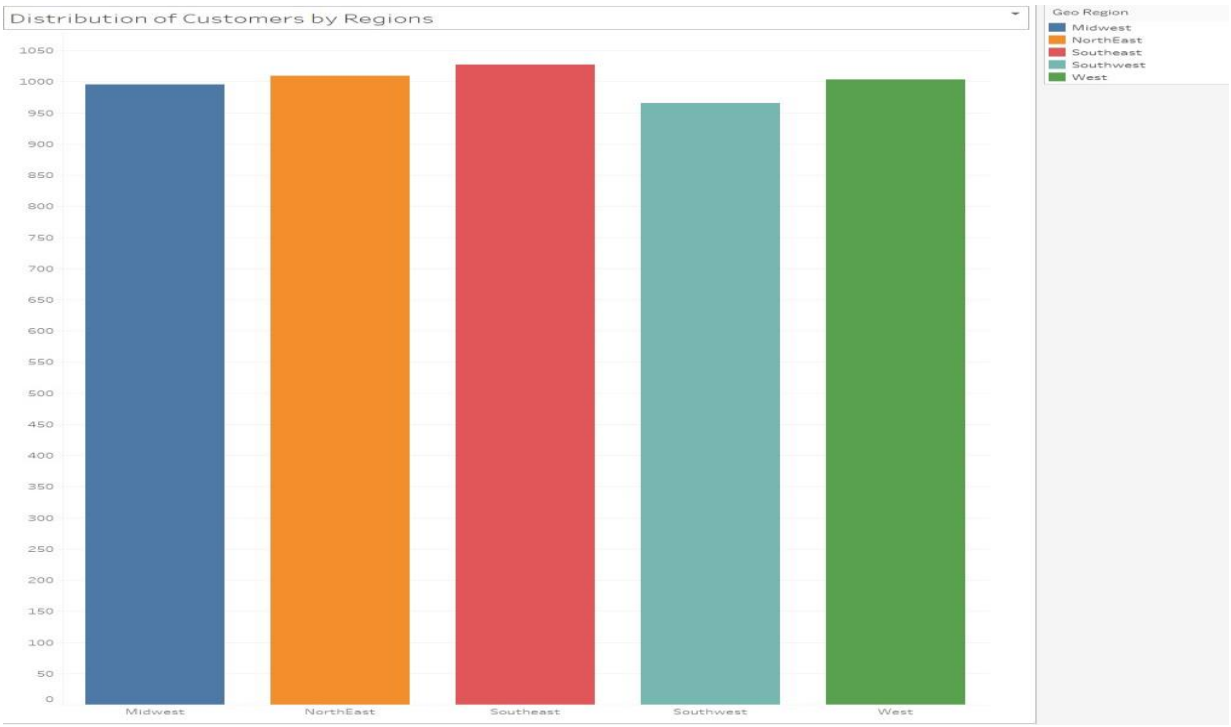

Figure 2: Distribution of Customers by Region

Analyzing both graphs shows that there is nothing special about the data. I wanted to see if there were some discrepancies where customers in the northeast would be more likely to use the company or if there would be more customers in a smaller town. Both graphs are normal and what you would expect to see if you were looking at a sample of the customers.

Var 4 is a variable that I derived from the one I made from Var 1 or "GeoRegion." Var 4 combines a customer's voting status and Var 1. Learning about the various voting statuses for different regions in the U.S. could give some insight into the customers in the dataset.

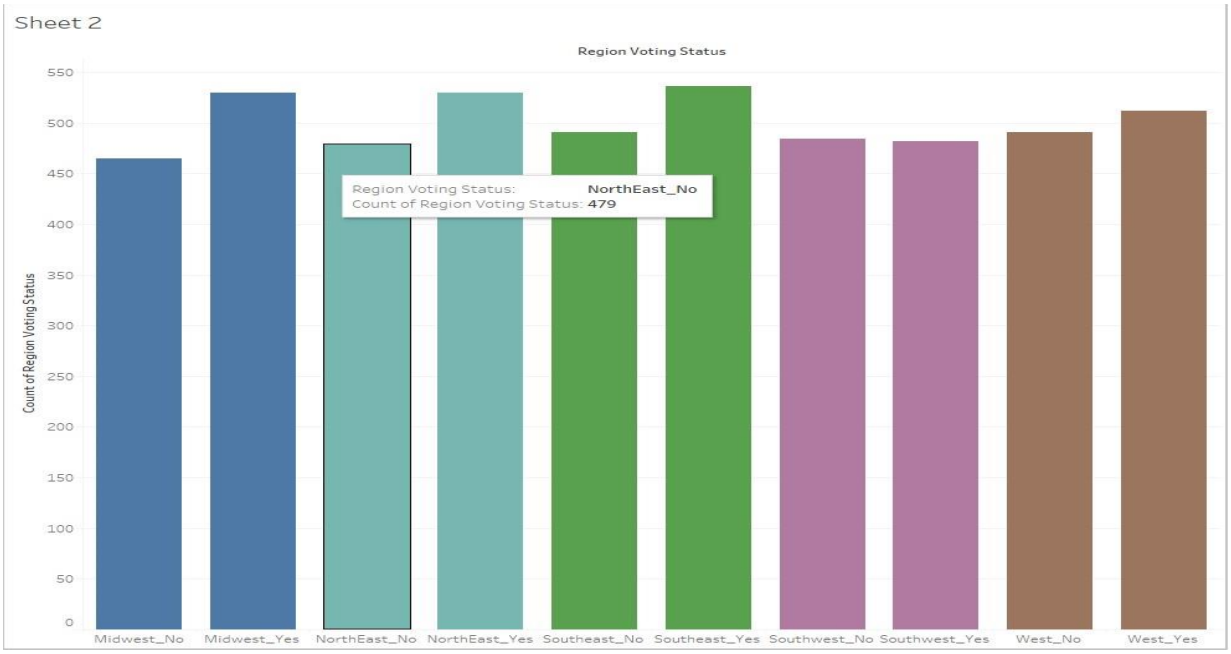

Figure 3: Region Voting Status

I figured that most people in the U.S. would be inclined to vote and there would be fewer people that didn't vote for several reasons. There seems to be the notion that customers tend to vote than not vote. I found that the southeast region looks to be a 50-50 split for this status which differs from the rest of the regions.

Sometimes age can play a big role in which types of customers a company would get but, a few years difference does not matter too much. I wanted to split this variable into 3 groups, young adults, middle age, and older adults. The young adult group would be from ages 18-37, middle age is 37-55, and older adults are 55 and above.

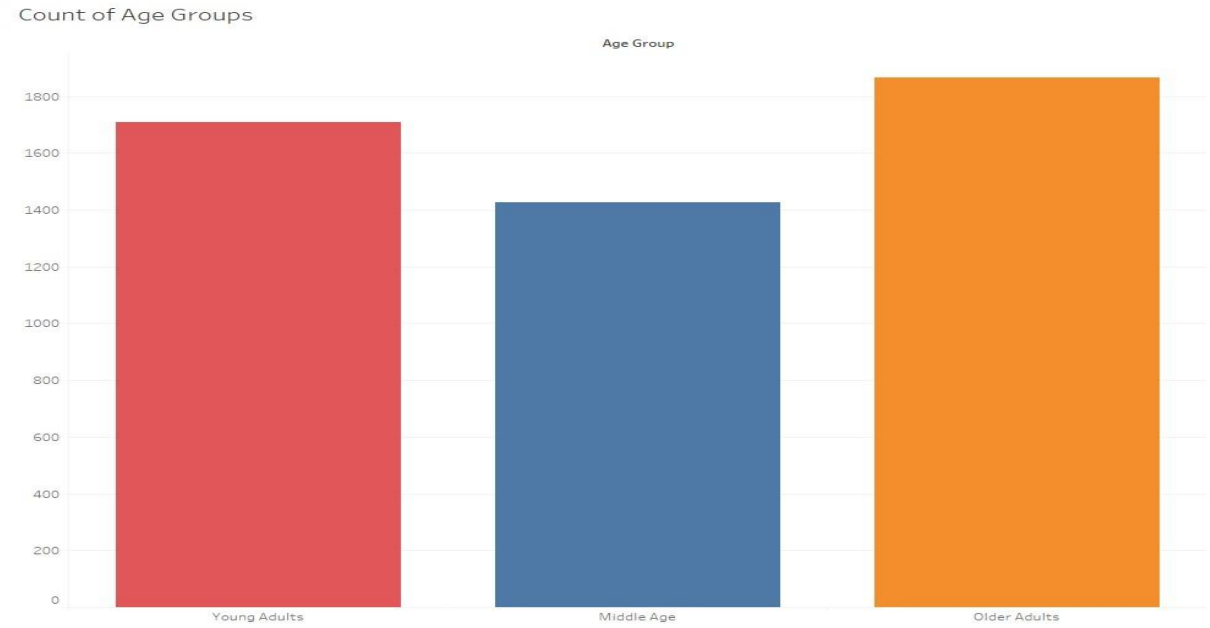

Figure 4: Age groups, young adults, middle age, and older adults respectively.

I thought it was surprising that there were more young adults than there would be middle-aged customers, but I believe that could be a factor in where I set the ages of the three groups. I felt that these numbers for each group were a good fit but, I wanted to point out that this result of the variable and its outcomes could be due to my input.

I observed in the dataset that the number of total pets did not always add up to the number of cats, dogs, and birds a customer had. I created this variable to try and figure out which other pets a customer had, though this is generalized so we would not know which type of other pets a customer had. I wanted to see if there were any other preferences in pet choice that weren't dogs, cats, and birds.

The last demographic variable that I created is Var 10 which comes from how much TV a customer watched if they had a PC or not. I wanted to see the impact of owning a PC, especially when patrons could stream many TV shows, sports, etc. on the PC instead.

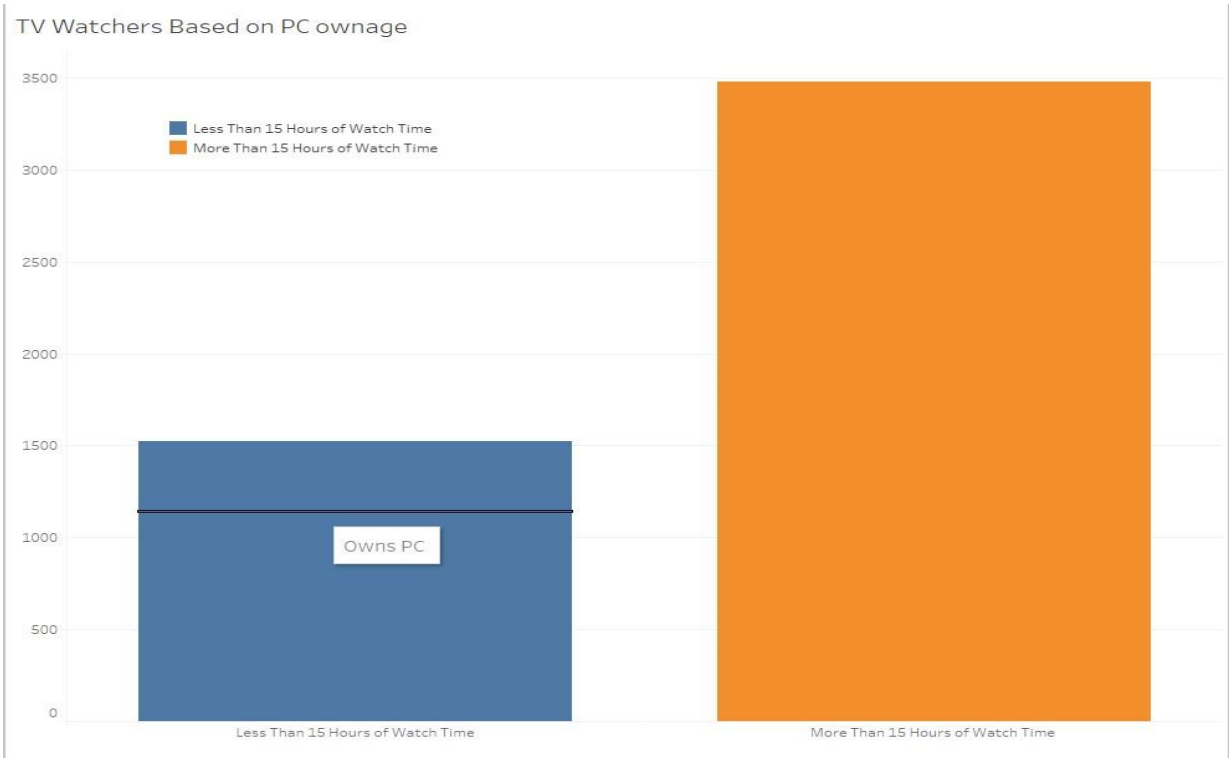

Figure 5: TV watch time based on PC ownage.

I wanted to leave a bit of room when it came to watching TV, so I used a 15-hour cutoff for watch time to consider errors with streaming or if a program is only available on cable. From Figure 5, we can infer that there is indeed a significance in watch time if a customer owned a PC or not.

Financial-Related Variables

Variables 6, 7. 8, and 9 are all financial-related variables. Var 6 is about the total debt of a customer which includes both credit card debt and any other debts a customer may have. I do understand that there are many kinds of debts a customer could have but I wanted to see the total for each customer. Var 7 is the average cost of an item charged on a credit card. This is done by dividing the amount of monthly spending on a credit card and the number of items purchased. This was done to see if a customer is a big spender or if they like to buy many items at low prices which could give insight into their shopping habits. Var 8 is the car value to income ratio. This is the value of a car divided by their household income. This is no indicator of what kind of car they drive so this could be information on the type of car that they drive since there is typically a price difference between economic and luxury car brands. Var 9 is the monthly phone bill. Typically, a customer would have both voice and data, so this is created by adding the price of other data and voice over their tenure divided by the length of their tenure. I don't think there is anything particularly interesting with this variable but could show how much of a customer's income goes into a phone bill.

Two-Column Graphical Analysis

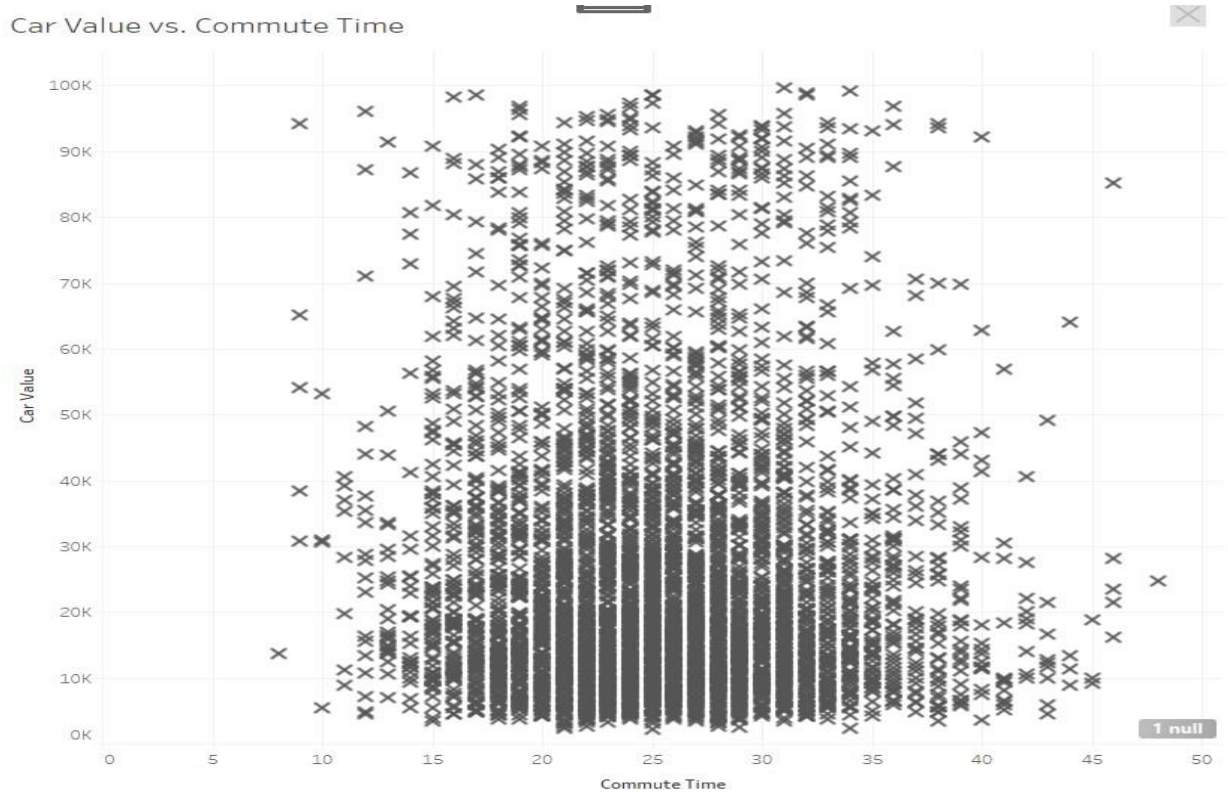

Figure 6: Car Value vs. Commute Time

Commuting time could play a big role in which type of car a customer could have. You would want to have a car with higher miles per gallon when you are commuting long distances or with long commute times. There don't seem to be any discrepancies in the graph but there is a higher density toward lesser value vehicles as not all customers can have a high household income.

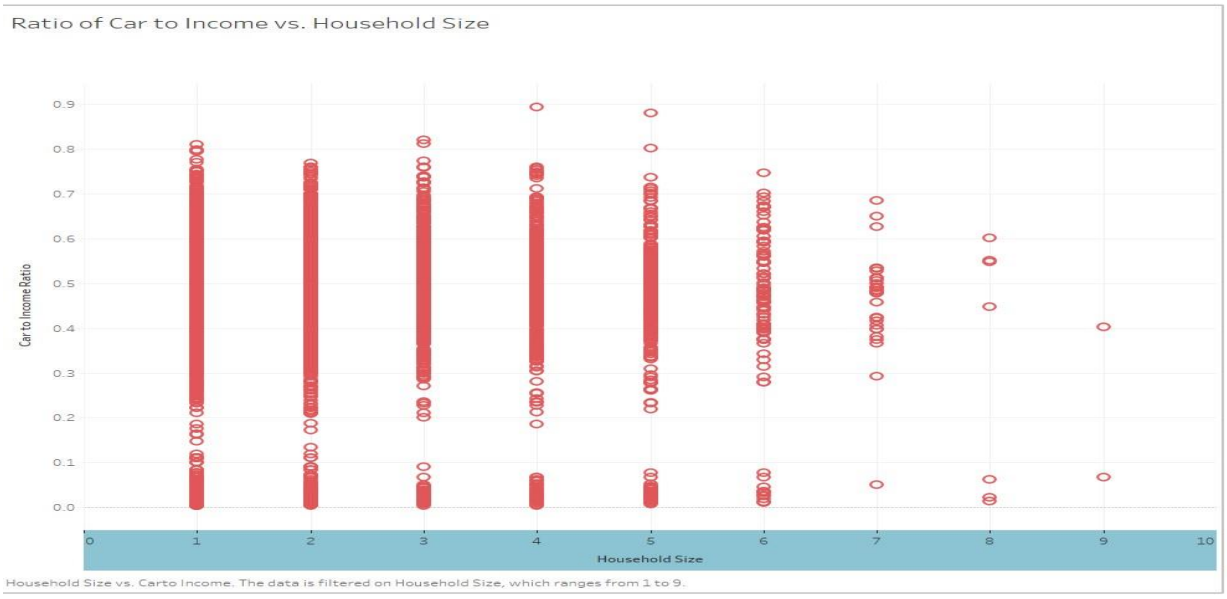

Figure 7: Ratio of Car to Income vs. Household size

The bigger a household size the more an income must be allocated toward other necessities. A car in some cases is not a necessity so it's expected that households with fewer people will allocate their money towards materialistic things like a car. Of course, we must consider a person's priorities but on a surface level, the graph is what I thought it to be.

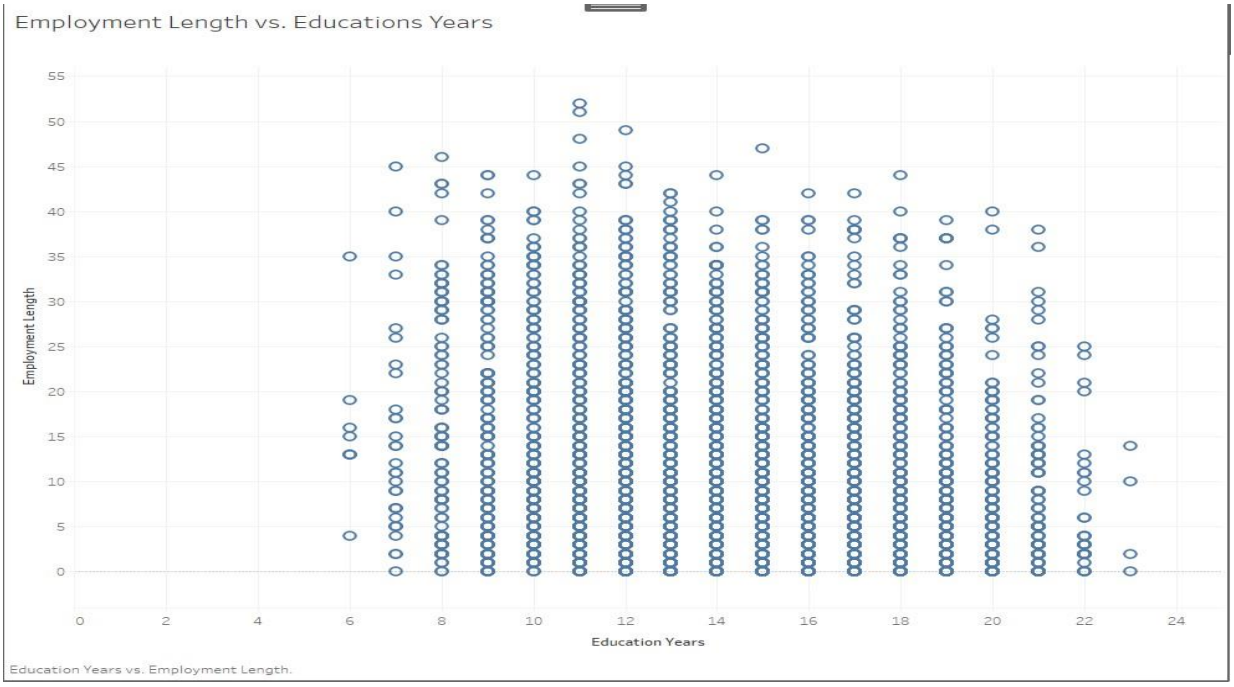

Figure 8: Employment Length vs. Education Years.

I thought there could be some interesting takeaways when comparing a customer's employment length with their education years but there doesn't seem to be a pattern. I still thought that these could be two variables that could be meaningful to at least see in a plot.

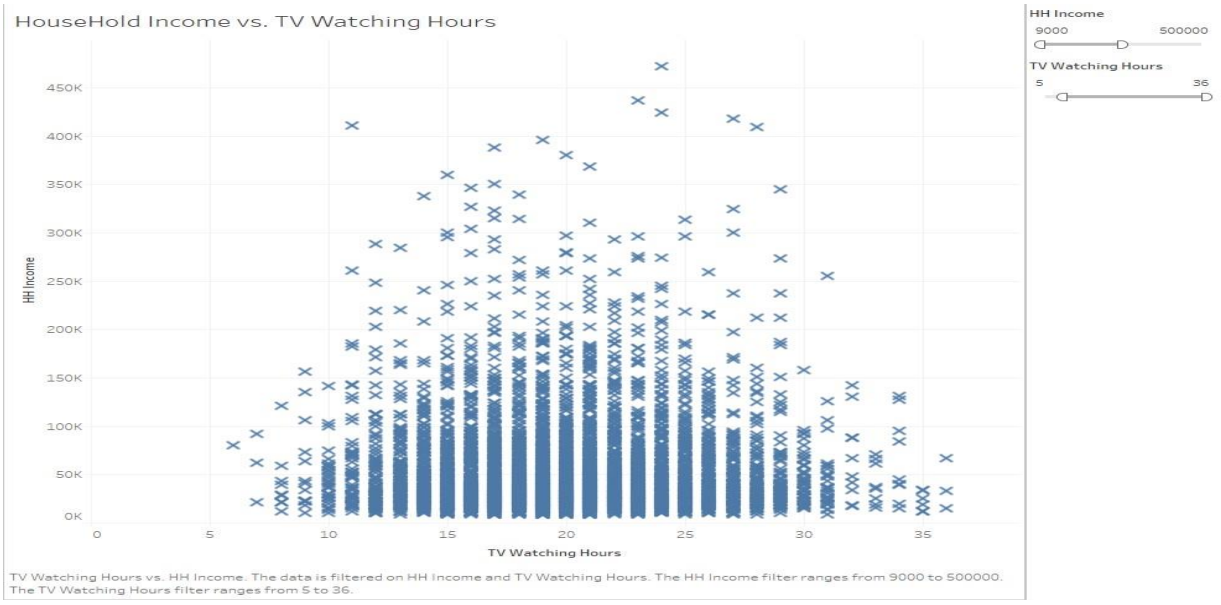

Figure 9: Household Income vs. TV Watching Hours

I wanted to see if there was any correlation between how much a household makes to their TV-watching time. I think if there was one thing you can infer from this graph is that income on average does not factor into how much TV a customer is watching.

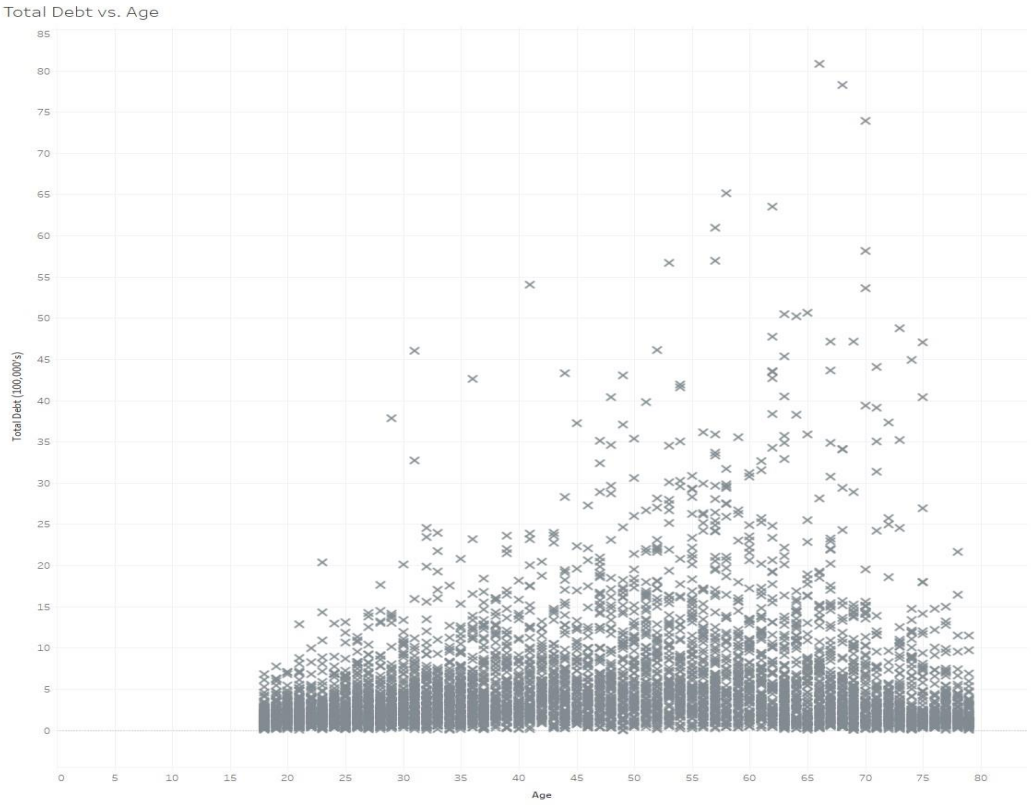

Figure 10: Total Debt vs. Age

This graph is interesting because I would think that many younger adults would have more debt compared to their older counterparts. This is a bit of a surprise to me as the graph shows that older folks tend to have higher amounts of debt. I thought that young adults would have more debt because of purchases or schooling but this doesn't appear to be the case.