Nick Ilvento
DSE 5004

Data Due Diligence Project

Id:

- CustomerID (Identifier)

Geographic:

- Region (Categorical)
- TownSize (Ordinal)

Demographic:

- Gender (Categorical)
- Age, EducationYears, HouseholdSize (Numeric)
- ActiveLifstyle, MaritalStatus (Categorical, Nominal)

Employment:

- JobCategory (Categorical)
- UnionMember, Retired (Nominal)
- EmploymentLength (Numeric)

Financial:

- HHIncome (Monetary), DebtToIncomeRatio, CreditDebt, OtherDebt (Numeric)
- LoanDefault (Categorical, Nominal)
- CreditCard (Categorical)
- CardTenure, CardItemsMonthly, CardSpendMonth (Numeric)

Housing and Transportation:

- HomeOwner (Categorical, Binomial)
- CarsOwned, CarValue (Numeric)
- CarOwnership, CarBrand (Categorical)
- CommuteTime (Numeric)

Political:

- PoliticalPartyMem, Votes (Nominal)

Telecommunications:

- PhoneCoTenure (Numeric)
- VoiceLastMonth, VoiceOverTenure, EquipmentLastMonth, EquipmentOverTenure, DataLastMonth, DataOverTenure (Numeric, Monetary)
- EquipmentRental, CallingCard (Nominal)
- WirelessData, Multiline, VM, Pager, CallerID, CallWait, CallForward, ThreeWayCalling, EBilling (Nominal)
- Internet(Ordinal)

Entertainment:

- OwnsPC, OwnsMobileDevice, OwnsGameSystem, OwnsFax, NewsSubscriber, (Nominal)
- TVWatchingHours (Numerical)

Pets:

- NumberPets, NumberCats, NumberDogs, NumberBirds, (Numeric)

Nick Ilvento
DSE 5004

Data Cleaning:

The first thing I did was to clean up CarOwnership and CarBrand. Personally, having a numeric value of -1 to represent n/a values is very confusing to ugly to me. This numeric value can also mess with statistical analysis and create weird correlation effects. To clean these up, I created new columns with a simple Excel formula. The new columns replace the -1 value with "None" for CarOwnership and "no car" for CarBrand. Now if we create visualization using CarOwnership it will display observations as having "None" cars instead of -1. This will make the visualization appear much more professional.

I also noticed that Car Value was not as clean as it should be. I went ahead and cleaned it up by changing values that were $1,000 to 0. I did this because in each case where the car value was $1,000. That household owned 0 cars. This must have been a glitch in the data entry system or somewhere else along the line. Either way, in order to create accurate metrics and plots, this data must be cleaned.

Next, I enacted a similar process when cleaning the monetary variables related to equipment and data. Similar to the last cleaning step, I believed the n/a value of "$-" to be ugly and hindering of potential visualization. I thought it would look best blank, so I created new columns that deleted the "$-" from those 4 columns: EquipmentLastMonth, EquipmentOverTenure, DataLastMonth, and DataOverTenure.

It appeared to me that a column could be added for Pets that are not dogs, cats, or birds. Many times an observation will contain a high number of Pets, but a low number of dogs, cats and birds. This disparity typically means there is something hidden or missing in the dataset. To combat this, I created a column called OtherPets. This column takes the sum of all pets and subtracts it by the count of dogs, cats, and birds. I then noticed that even in OtherPets we are still seeing high numbers so I assume these are most likely fish or insects being counted individually.

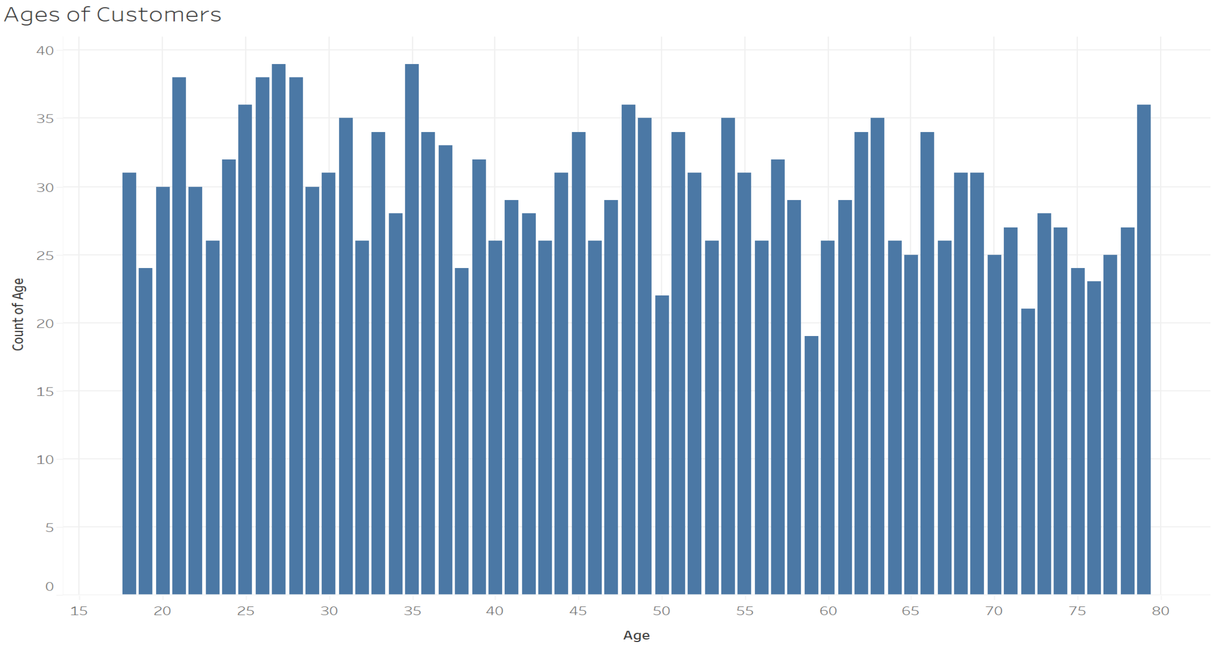
I noticed there was a slight spelling glitch within the Credit Card column. I cleaned up the column so that the values spell "Other" instead of "Othe." This will help visualizations look much cleaner and will go a long way when presenting to stakeholders.

Personally, I didn't like the way CreditDebt and OtherDebt were formatted. They were expressed in decimal format as terms of 100,000 so I created new columns that multiplied their values back by 100,000. This makes the variables much easier and quicker to read. It would also create a much simpler graph, as the exact monetary figure would be displayed on a visualization, rather than a decimal number. Decimal numbers remind me of ratios and dollars expressed as millions.

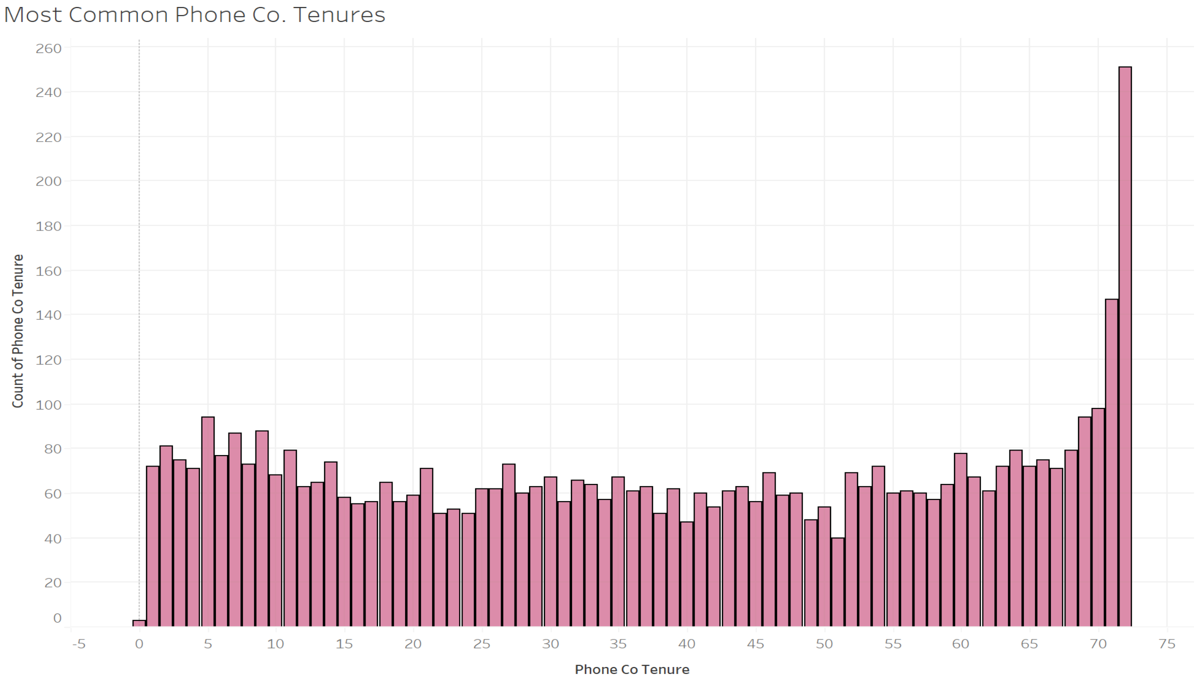
Looking at the CommuteTime column, it's hard to determine what is considered a long commute and what is considered a short one. In order to solve this problem, I created a column with standardized CommuteTime Values. This was done by subtracting from the average and dividing by the standard deviation. By putting this column in terms of Z-Scores, we can much more easily determine what type of commute each person has.

Nick Ilvento
DSE 5004

One -Variable Visualizations:

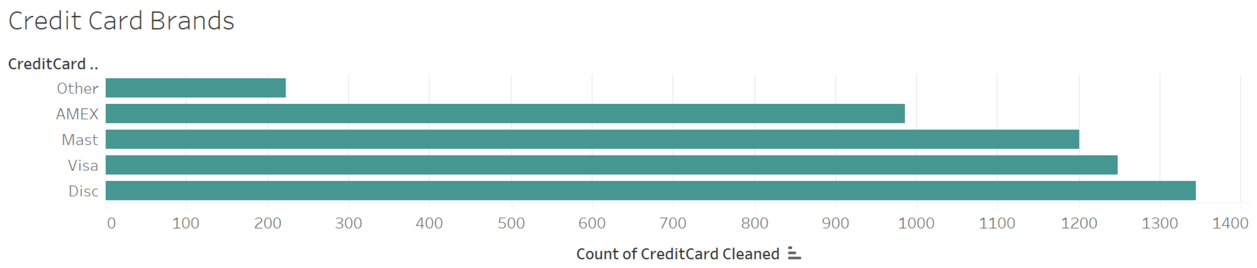


I first wanted to get a feel for the age of our customers. It appears they are distributed fairly-randomly as no patterns are seen in the histogram.

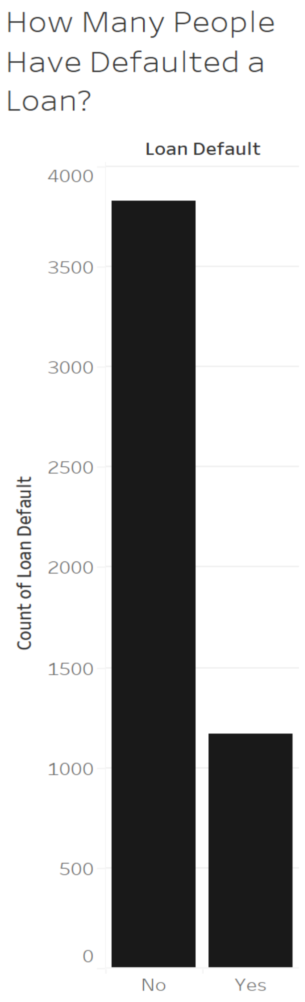


The plot of count of Phone Co Tenure for Phone Co Tenure. The data is filtered on Equipment Over Tenure, which excludes Null.

I then wanted to get a feel for how long our customers have been with us. It's clear that there are more customers who have been with us for 71 or 72 months compared to any other amount. Ages are normally distributed, so that cannot be the cause. I wonder if we started collecting data 72 months ago and members were all placed under that time frame. Very Interesting.

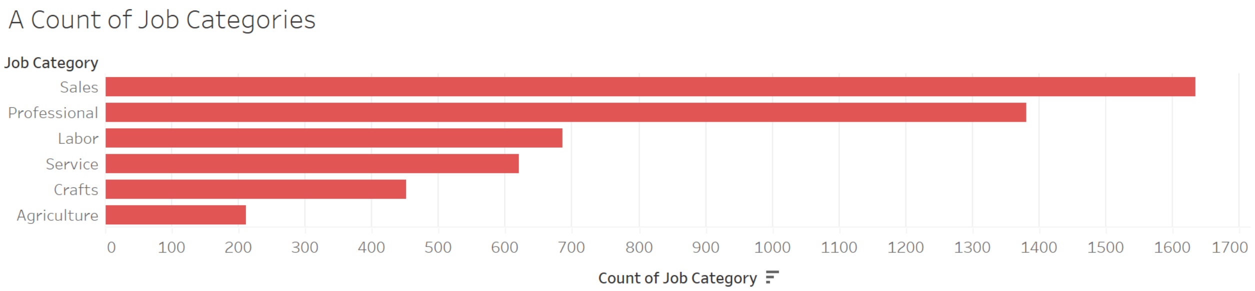


The count of credit card brands could be useful to our Finance or advertising department. Our previous data cleaning of the 'Other' Category makes this chart look much cleaner than it would.



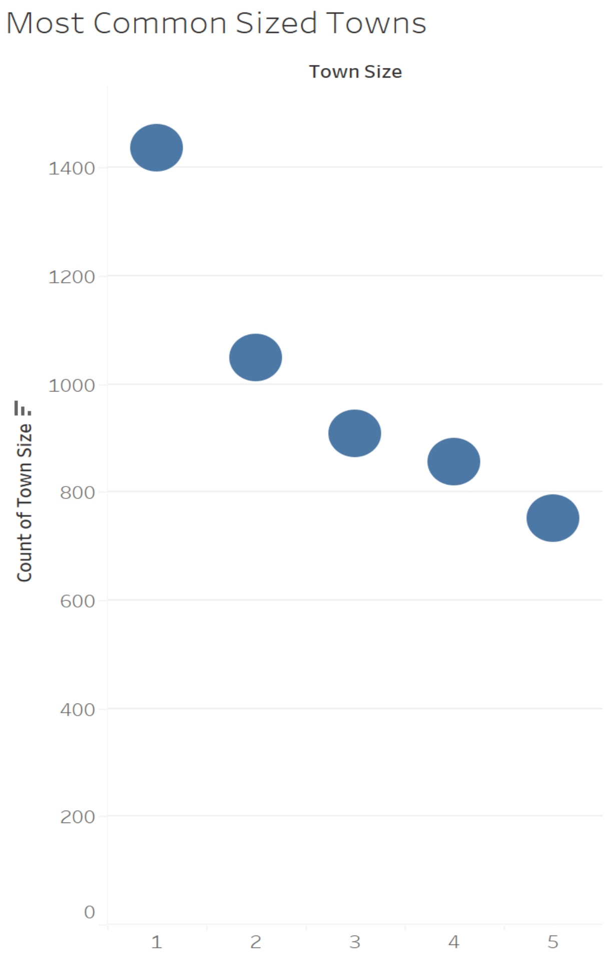
-It's good to see that the vast majority of our Customers have not Defaulted on a Loan. This means for the most part our Customers are financially smart and secure.

Nick Ilvento
DSE 5004



Count of Job Category for each Job Category. The view is filtered on Job Category, which excludes Null.

Job Categories can be a very helpful piece of data when analyzing customers. It's clear the majority of Customers work Professional/Sales Jobs. On average, these jobs tend to be more high paying than others, especially the others listed. This is again, evidence of the financial prowess shown by most of our Customers.



Count of Town Size for each Town Size. The view is filtered on Town Size, which excludes Null.

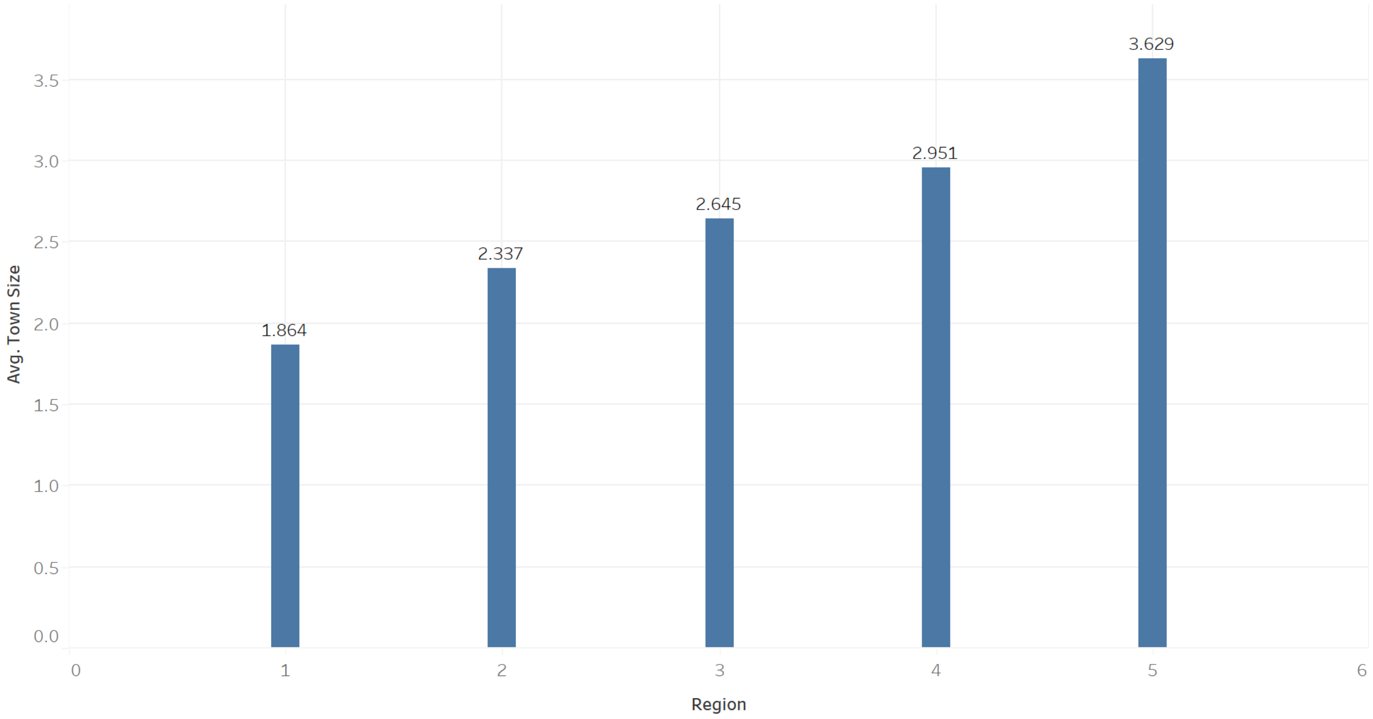
The size of our customer's towns helps us gain a grasp on our demographics as a whole. As expected, many of our customers are from a Town size of 1 or 'Very Large.' However, it's interesting that the other sizes have little variance between them. There is only a difference in about 300 customers between the rural towns and size 2 towns. Considering the population density of America, I would say we have a more respective amount of rural and suburban customers, compared to large cities.

1: Very Large
2: Large
3: Suburban
4: Small Town
5: Rural

Nick Ilvento
DSE 5004

Two Variable Visualizations:

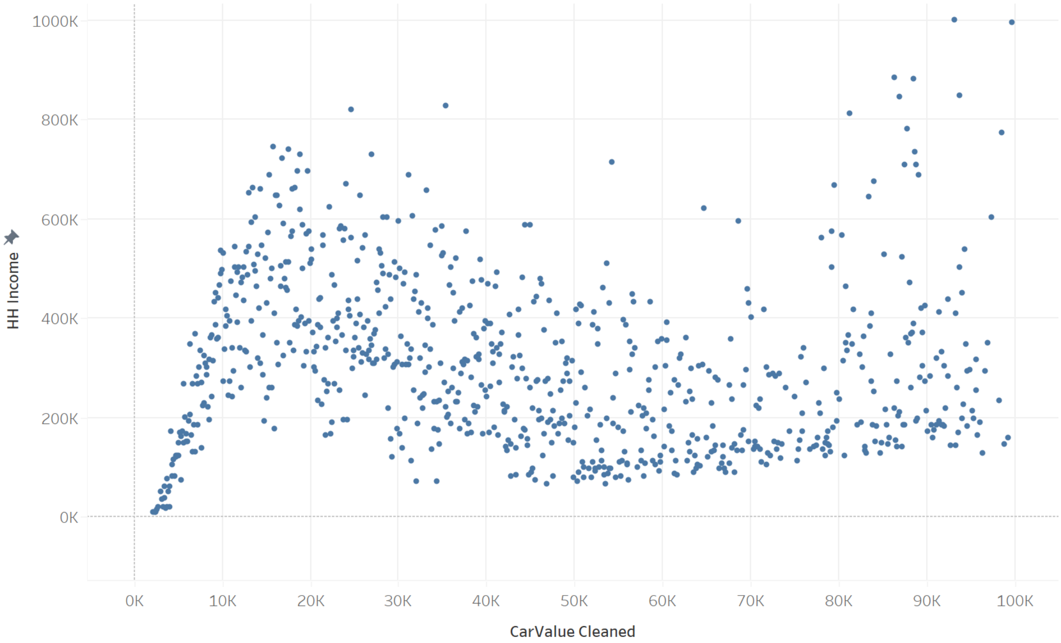
Average Town Size in each Region



1: NorthEast 2: Midwest 3: West 4: Southwest 5: Southeast

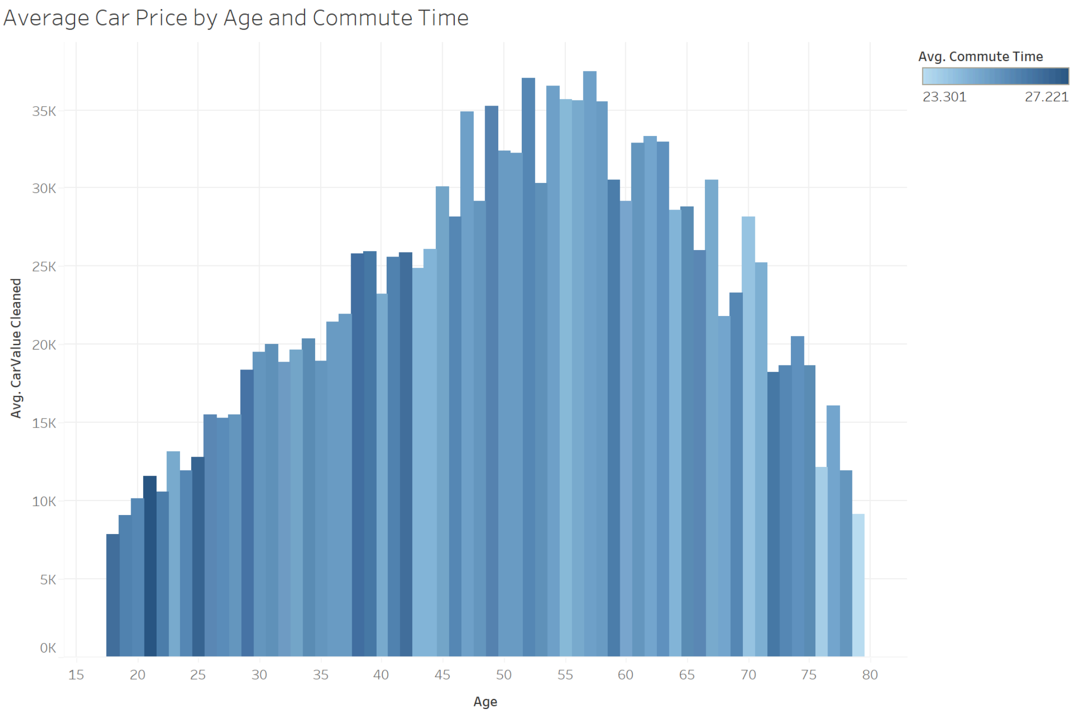
This chart displays the average town size in each region listed above. This can help us decide which demographic we may want to target for different kinds of advertising. The north-east tends to be more city based and the Southeast tends to be more on the suburban side.

Car Values Compared to Household Values



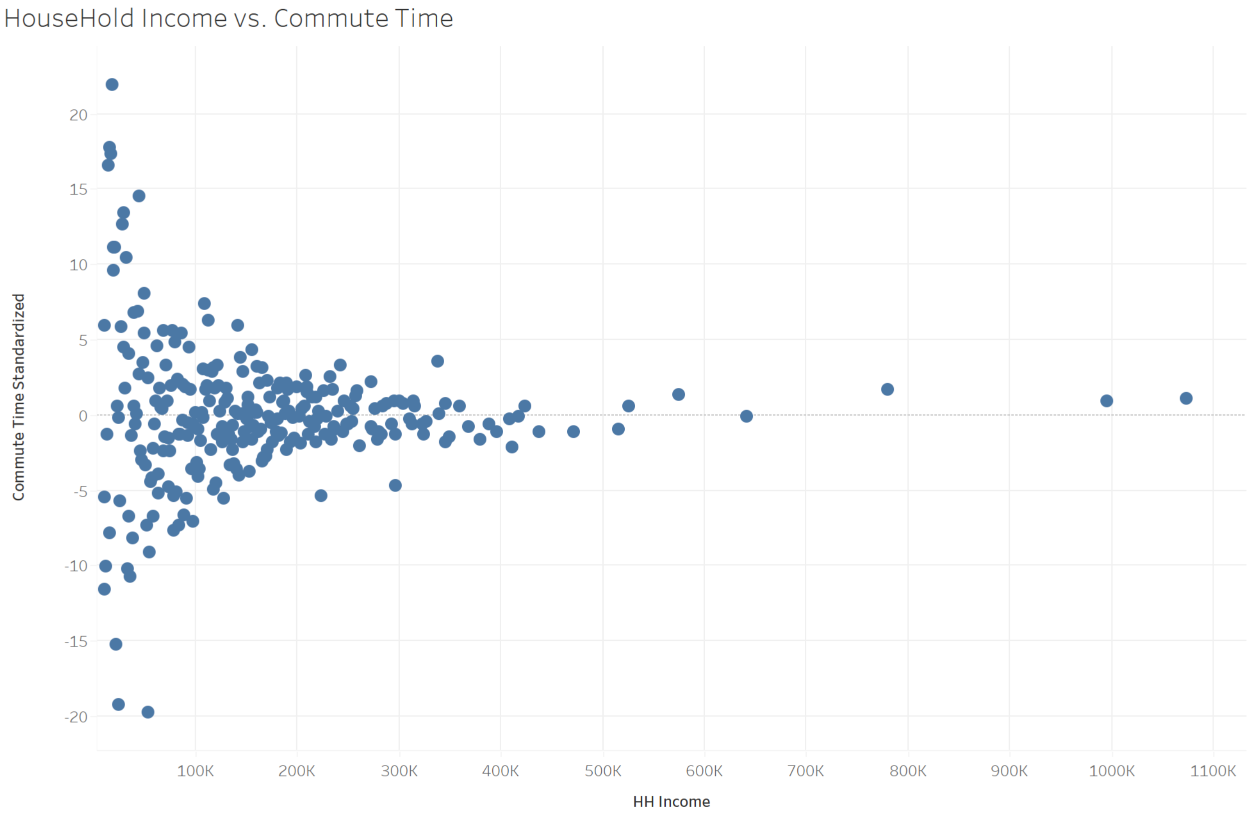
Nick Ilvento
DSE 5004

This chart above displays something very interesting about our customers. It appears that people with relatively lower car values tend to have Household Incomes above the people with higher Car Values. This is an interesting negative relationship and not something I would expect. It shows that there really is no significant relationship between car value and household value, unless you are observing the extremes.



This chart shows the relationship between Age and Car Price of our customers. This can help us to determine which Age ranges are worth our marketing efforts and give us clues to where our customers might live. It is also interesting to note that when Age and Car Price are both low, the average commute time seems to be higher. This is assumingly less time being spent on the phone, which could be a reason to slow down on the marketing to this demographic.

Nick Ilvento
DSE 5004



After standardizing commute times, we are able to look at the average commute based on Income. We can see an interesting cone-shaped pattern which tells us that there is much more variation of commute times in low income households. It is interesting that as a person's income increases their commute time converges towards the average. This may however be due to the sample size of low income households versus high ones.

Summary:

Learning about a customer basis is in my eyes one of the most important things a business can do to become successful. There is no business without its customers so learning about them is a mandatory step in order to obtain new ones and keep our old ones. There are many different kinds of people in our world, so catering to all of them is no easy task. However, if we can simplify our data and take a deep dive into the details, we can discover insights that will in turn raise profits.

Our data entry system is fairly clean and works well the majority of the time. There are not a lot of missing values and there's consistency within each column. I only noticed a few spelling glitches and formatting issues with the system, which were very easy to smoothen out in Excel. There are also many different types of data that we collect, which allows us to make our analysis and visualizations very easily.

Nick Ilvento
DSE 5004

Through visualization, I was able to pick up on various trends regarding our customers. Many of them work Professional or Sales jobs and are very scattered throughout the country. Many of them also live in large cities, but we have more who live in smaller towns relative to the population of the Nation. This means despite the Large Cities being our most popular demographic. Relatively, I assume most of our customers live in Suburban and Rural areas. More evidence of this is shown by Commute Time data. The average commute is 25 minutes long and as the income of the household increases, the average commute time converges toward this metric. We can assume from this that many people are commuting from outside the city, to the inside of the city for their professional and sales jobs. This knowledge will in turn help us determine the best ways to advertise to the majority of our customers.

A thriving business is able to gain new customers from all different areas and all different living-situations. I believe that we do this well here and that we have a high level of diversity among our clients. This allows our analytics team to take deep dives into subsets of our collected data and pull some very interesting insights. Marketing around these insights will improve our customer base and grow our company's size and profits.