**Data Due Diligence Project**

Cameron Gallien

DSA5400: Visual Data Exploration

Merrimack University

Prof. Christopher Healey

For this project my main goal was to analyze exactly what drives sales for this particular telecommunications company, and see if that would provide any insights for potential marketing strategies going forward.

I started by grouping the variables provided into the following groups:

|  |  |  |  |
| --- | --- | --- | --- |
| Demographics | Financials | Products Owned | TeleComm Specific |
| CustomerID | HHIncome | OwnsPC | PhoneCoTenure |
| Region | DebtToIncomeRatio | OwnsMobileDevice | VoiceLastMonth |
| TownSize | CreditDebt | OwnsGameSystem | VoiceOverTenure |
| Gender | OtherDebt | OwnsFax | EquipmentRental |
| Age | LoanDefault | HomeOwner | EquipmentLastMonth |
| EducationYears | CarValue | CarsOwned | EquipmentOverTenure |
| JobCategory | CardItemsMonthly | CarOwnership | CallingCard |
| UnionMember | CardSpendMonth | CarBrand | WirelessData |
| EmploymentLength |  | NewsSubscriber | DataLastMonth |
| Retired |  | CreditCard | DataOverTenure |
| MaritalStatus |  | CardTenure | Multiline |
| HouseholdSize |  |  | VM |
| NumberPets |  |  | Pager |
| NumberCats |  |  | Internet |
| NumberDogs |  |  | CallerID |
| NumberBirds |  |  | CallWait |
| CommuteTime |  |  | CallForward |
| PoliticalPartyMem |  |  | ThreeWayCalling |
| Votes |  |  | EBilling |
| Active Lifestyle |  |  |  |
| TVWatchingHours |  |  |  |

My reasoning for grouping them as such was to see if any of the specific groupings provided more information on what drives spending than the others. For example, is there a specific type of person who is more likely to have a higher monthly bill than others? Or is this determined more by a person’s financial situation? I also wanted to find out if certain products that use a lot of data would drive up the price someone pays for their telecommunication services. The last category contains all the information specific to their contracts with the telecommunications company, such as the services that they sign up for, and how much they spend.

Once the categories were created, I narrowed them down to specific variables of interest that I felt would give me the most information necessary:

|  |  |  |  |
| --- | --- | --- | --- |
| Demographics | Financials | Products Owned | TeleComm Specific |
| Region | HHIncome | OwnsPC | PhoneCoTenure |
| TownSize | DebtToIncomeRatio | OwnsMobileDevice | VoiceLastMonth |
| Gender | CardSpendMonth | OwnsGameSystem | VoiceOverTenure |
| Age | CarValue | OwnsFax | EquipmentRental |
| EducationYears |  | HomeOwner | EquipmentLastMonth |
| JobCategory |  | NewsSubscriber | EquipmentOverTenure |
| Retired |  |  | DataLastMonth |
| MaritalStatus |  |  | DataOverTenure |
| HouseholdSize |  |  |  |
| CommuteTime |  |  |  |
| PoliticalPartyMem |  |  |  |
| Votes |  |  |  |
| Active Lifestyle |  |  |  |
| TVWatchingHours |  |  |  |

The variables that I took out were done so because they either did not provide much valuable information (Customer ID’s, Car Brand, Employment Length), contained too many errors to include (Internet), or in many cases the information was just better encapsulated in a different metric (Card Items Monthly, Credit Debt, Other Debt, Equipment Rental). I also took out many of the services such as Caller ID, Call Forwarding, and Voicemail because without a way to tell their specific value they did not give much information as to what a customer was spending. I believe these metrics were better summarized in the variables regarding how much the customer spent with the company.

My next step was to create additional columns in the data frame. I began by cleaning the data, which primarily consisted of turning all the variables related to money from characters to numerical values. This was necessary in order to do any calculations with the data because in their original state they were strings of characters that mathematical operations could not be done on. During this process, many values that did not have an amount listed (DataLastMonth, EquipmentOverTenure, etc.) were changed to NA’s. These values coincided with customers who did not subscribe to the service, so the NA’s were replaced by zeroes as someone who does not subscribe would obviously not be paying for those services. The variable CarValue also included many values of -$1,000 whenever a customer did not own a car, so I replaced all of those values with 0.

Next I created a column titled AllServices that provided a value of True if the customer was paying for Voice, Equipment Rental, and Data, and a False value if they were not paying for all three services. This was created to see if there were any correlations between the amount of services a person signed up for and their total amount spent.

I created columns for the total amount a customer spends each month across all services (AllSvcTotal), and then divided that by how many months they’ve been a customer to get their monthly spending (AllSvcMonth). I did this because I felt it was a more accurate indicator of the value of each customer than the total amount that they’ve spent, considering those values could be skewed by how long they’ve been a customer. I also created a column titled IncomePerMember, which divided a customer’s household income by the household size to see if that provided any better information than the total Household Income.

Lastly, I created a column titled TotalTech, which added up how many additional products a customer had (PC, Mobile Device, Game System, & Fax). This variable could provide insights on whether owning these additional products caused customers to spend more on their telecommunications bills in exchange for the services required to optimize those devices.

The next and final step was analysis. I started by looking at how the customers were distributed by region (a). This histogram showed that the customer data was pretty evenly distributed by region of the country. This led me to a box plot of spending per region (b), which showed that the amount being spent monthly has very little variation across the country.

Next I wanted to see if there were any patterns between a person’s financial situation, and how much they spend monthly on telecommunications services. This led me to create a scatterplot of household income compared to monthly charges (c). As you can see by the scatterplot, there does not seem to be a causal relationship between the among someone makes and their monthly telecom bill. Those who spend the most per month on services are actually on the lower end of the spectrum. This led to my next scatterplot, which plots a customer’s Debt To Income ratio against their average monthly bill (d). This plot does show that as a person’s debt ratio grows to very high levels, the amount spent on their monthly bill goes down gradually. This offers insights that weren’t as evident in the previous scatterplot.

One very interesting finding that I stumbled upon was the impact of a person’s education level on their monthly bill (e). While this scatterplot has points that are very condensed due to the amount of data, it clearly shows that the highest spenders on a monthly basis are those with more years of education. This could point marketing teams towards higher education institutes, as those customers may be more likely than the others to spend more with the company.

I also sought to see if there was any correlation between a customer’s job category and monthly spending (f). Aside from some basic outliers there is very little variation between job types when it comes to monthly spending.

The next box plot shows the relationship between whether or not someone subscribes to all services available, and what their bill equals (g). This plot shows that when someone is subscribed to all services (Data, Equipment, and Voice) their monthly spending is much higher. This is pretty obvious to some degree, however I was surprised to see that it was as high as it was. There is a substantial jump in a customer’s monthly bill if they subscribe to all three services from the company. This could potentially lead to the company lowering the cost if a customer bundles all of the services together, which could in turn allow more customers to add additional services. It also shows how important it is for the company to get existing customers to add additional products.

The last variable I chose to analyze was the total pieces of additional technology that a customer has on top of their telecom products. These were PCs, mobile devices, fax machines, and gaming consoles. When graphed against the average spent per month on their bill, it is impossible not to see a trend (h). The more devices that a person owns, the more they tend to spend per month on their telecommunication services. With this in mind, I was curious as to how many customers had additional pieces of technology (i). Surprisingly, there are very similar amounts of customers with 0-3 additional pieces of technology, with approximately half of that amount having 4 devices. An interesting marketing strategy that could perhaps be incorporated is giving away some of these devices for those who sign up for their service, as the additional money spent by these customers would make up for the initial cost of the device.

The biggest takeaway from the analysis is that the amount someone spends with the company has a lot less to do with a customer’s finances, and much more to do with the services they use. Those who have many devices which utilize telecommunication services tend to spend the most out of any other variable that I could find. There are also additional insights to be found in the amount of services that a customer is signed up for, as being subscribed to all three services has a major impact in their monthly bill. Lastly, I was surprised to see very little correlation when it came to the region of the country and a customer's job category. Neither variable showed much change when it comes to their monthly spending. While I did not expect there to be a massive amount of difference between these categories, I was surprised to see just how little variation there was across these variables.

(a)

Chart, bar chart

Description automatically generated

(b)

Chart, box and whisker chart

Description automatically generated

(c)

Chart, scatter chart

Description automatically generated

(d)

Chart, scatter chart

Description automatically generated

(e)

Chart, bar chart

Description automatically generated

(f)

Chart

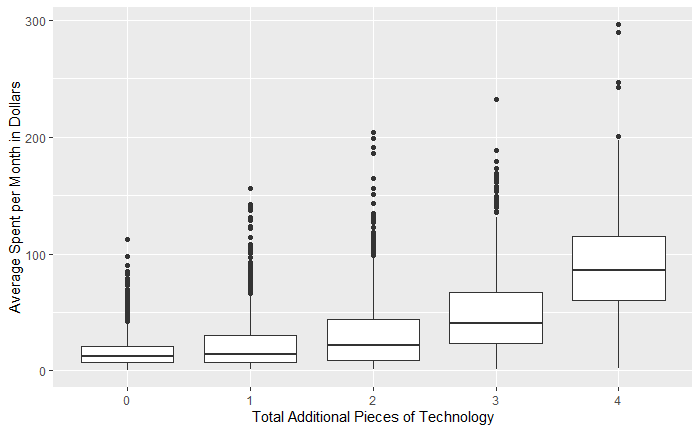
Description automatically generated

(g)

Chart, box and whisker chart

Description automatically generated

(h)



(i)

Chart, bar chart

Description automatically generated