Data Transformation and Simplification Project

Introduction

In this project, I stepped into the role of a new marketing analytics manager for a telecommunications company to work on making sense of the data set representing 5,000 customers. In this role, I am helping the company begin a new program to provide services to customers with financial need to help advance their own or their children’s education. While transforming and simplifying the data set, I kept this goal in mind while prioritizing fields that would be particularly helpful in the analysis; these fields include salary, education, debt, marital status, and household size. Additionally, I considered the types and number of plans, services, devices, and subscriptions each customer had to help determine what they may not currently have access to and what they may be struggling to afford.

1. Setting the Stage

1.1 Hypothetical Scenario

As a new marketing analytics manager for a telecommunications company, I am looking to help the company begin a new program for customers in financial need. This program will provide customers with services, plans, and/or devices that will advance their own or their dependents' education. The analysis will include identifying customers in need based on their household income, debt, and level of education. I will need to consider the services, plans, and/or devices the customer already has as well as those they may need for educational purposes.

1.2 Guiding Questions

As I worked through making sense of the data, I kept a running list of guiding questions important to the analysis, shown below.

- What qualifications must customers meet to be eligible for the program?
- What services, plans, and/or devices should the company provide as part of this program?
- Who is eligible for this program?
- What are the demographics of those who are eligible for this program?
- What services, plans, and/or devices do customers already have?
- What services are accessed the least by customers who have financial need?
- What services do customers with dependents utilize the most?

2 Transforming the Data

2.1 Examining Data Values and Data Types

After performing an initial clean step in my data flow, I began examining the data values and data types. My initial observations included:

- Region, Town size, and Home ownership: these three fields were numerical, with numbers corresponding to different categories; in my opinion, a more intuitive representation of the values would be to change it to categorical data (string values) and eliminate the use of values associated with different categories
- Equipment monthly usage, Equipment lifetime usage, Monthly data usage, Lifetime data: these fields were registering as string values because there were dollar signs in front of each value; since these are not monetary values, the dollar signs could be removed and the field could be converted from string values to numerical values
- Home internet plan: values included yes, no, and values associated with plan types; this could be simplified to a simple yes or no for my analysis purposes
- Null values occurred in the following fields: town size, gender, job category, household size, number of pets, number of cats, number of dogs, number of birds, home ownership, voice service tenure usage
- Vehicle ownership, vehicle brand, vehicle value: included negative values in the 497 rows with zero vehicles owned
- Credit card type: typo of "Othe" should be "Other"

Knowing that some of these fields would not have a substantial impact on my analysis, I proceeded by transforming the fields I anticipated would be most likely to be useful.

2.1.1 Region column

In the data set provided, the Region column had numerical values corresponding to the region where the customer resides:

- 1 = Northeast
- 2 = Southeast
- 3 = Midwest
- 4 = Southwest
- 5 = West Coast

To transform the data, I changed the Region column to string data values and replaced the values (1, 2, 3, 4, and 5) with the corresponding categories (Northeast, Southeast, Midwest, Southwest, and West Coast). I made this transformation to make it easier to filter and aggregate the data by region. Creating these categories allows analysis based on customers' geographical location and will be more intuitive to use with the regional names than a key with values. This is seen in the examples of the simple bar charts shown in Figure 1 and Figure 2. I can conclude from the bar charts that there are roughly the same number of customers from each region, with the most customers living on the West Coast and the fewest customers living in the Southwest. I suspect that this may also correspond with the population distribution across the United States, since the West Coast (particularly California) is heavily populated.

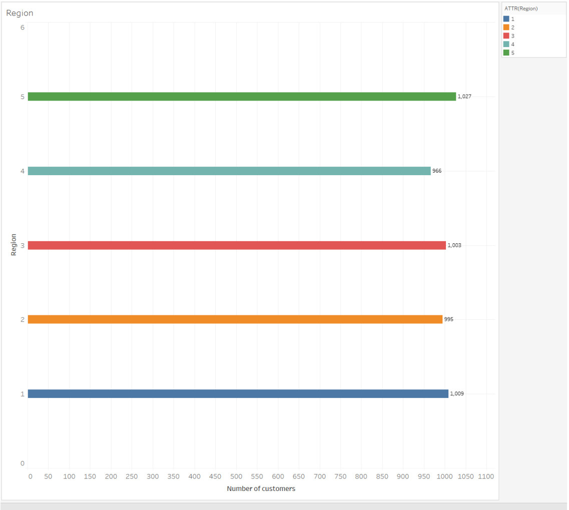


Figure 1 *This bar graph shows the results of using the codified values for the Region column; the results would be similar for the codified values used in the Town size column.*

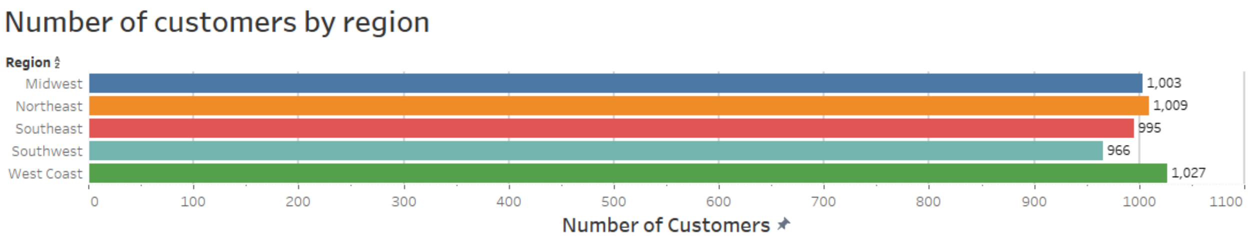


Figure 2 *This bar graph shows the results of using the transformed Region column.*

I can conclude from the bar charts that there are roughly the same number of customers from each region, with the most customers living on the West Coast and the fewest customers living in the Southwest. I suspect that this may also correspond with the population distribution across the United States, since the West Coast (particularly California) is heavily populated.

2.1.2 Town size column

In the data set provided, the Town size column also had numerical values corresponding to the size of the town where the customer resides:

- 1 = Small
- 2 = Medium
- 3 = Large
- 4 = Extra-Large
- 5 = Mega-City

To transform the data, I changed the Town size column to string data values and replaced the values (1, 2, 3, 4, and 5) with the corresponding categories (Small, Medium, Large, Extra-Large, Mega-City). I made this transformation to make it easier to filter and aggregate the data by Town size. Creating these categories allows analysis based on customers' Town size and will be more intuitive to use than a key with values. This allowed me to create the bar graph in Figure 3. As I proceed with my analysis, I could use these categories to filter by Town size to make comparisons with more ease, as shown in Figure 4.

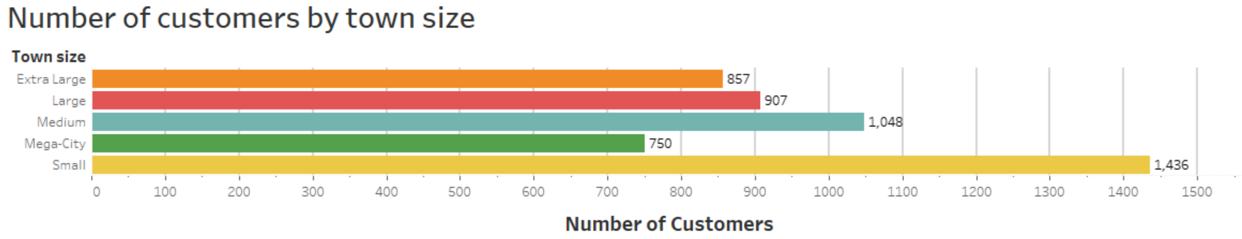


***Figure 3** This bar graph shows the results of using the transformed Town size column.*

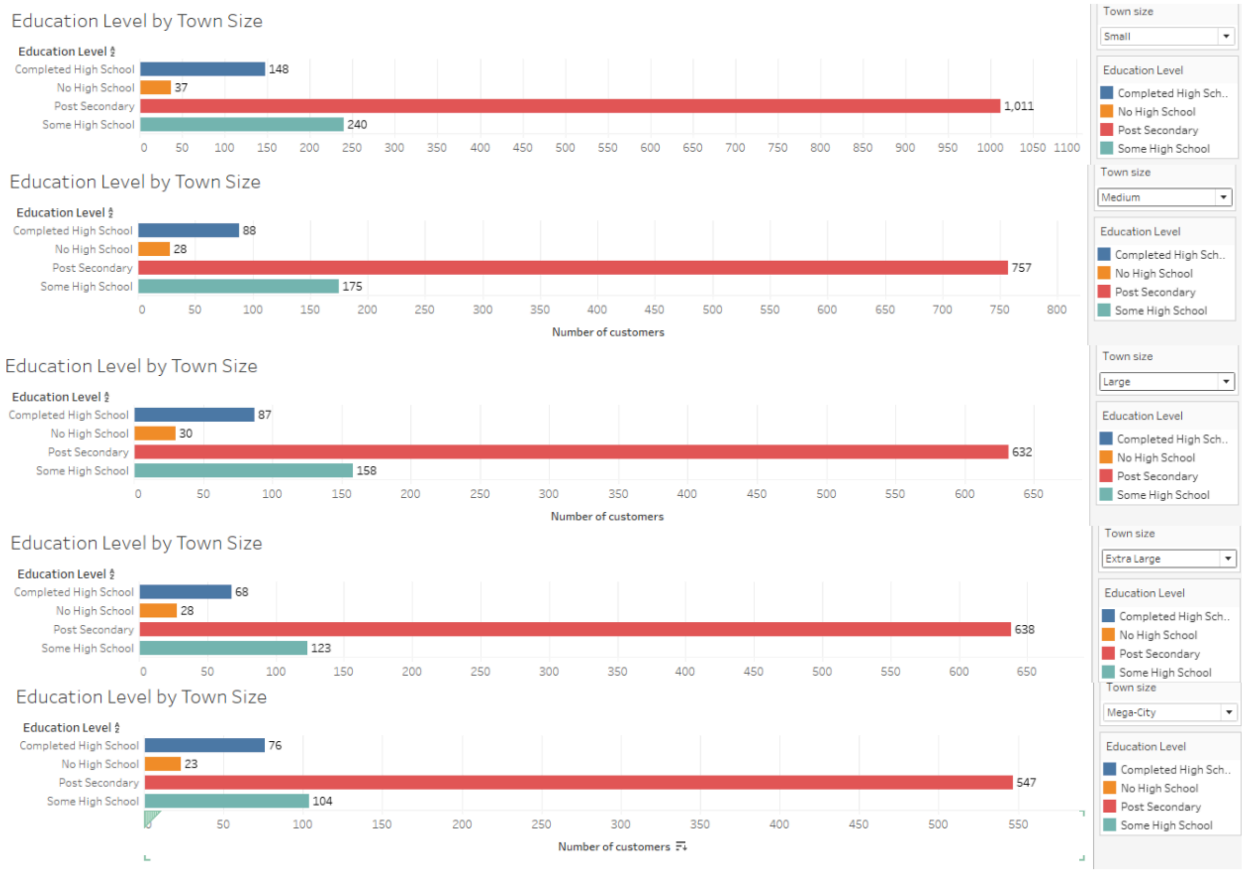


***Figure 4** This shows the breakdown of Education Level filtered by Town size.*

Based on the figures above, it is clear that regardless of Town size, most customers have at least some Post Secondary education. However, Large towns have the highest ratio of No High School to Post Secondary and also have a higher proportion of customers falling in the Some High School education level. This is significant because it signals that more customers from Large towns may benefit from the program being launched to help advance customer's education level.

2.1.3 Home Ownership column

In the data set provided, the Home Ownership column also had numerical values corresponding to whether or not the customer owns a home:

- 1 = Customer Owns a Home
- 0 = Customer does not own a home

To transform the data, I changed the Home Ownership size column to string data values and replaced the values (1, 0) with “Yes” to signify home ownership or “No” to signify not owning a home. Similar to the [2.1.1 Region column] and [2.1.2 Town Size column], I made this transformation to make it easier to filter and aggregate the data by Home Ownership is necessary during the analysis.

2.1.4 Credit Card column

In the Credit Card column, I edited the “Othe” category to “Other” in order to fix the typo.

2.1.5 Monthly card spend column

Upon examination of the null values in the Monthly card spend column, I discovered that customers with null values have 0 monthly card items. This indicates that they did not spend anything on the card. Therefore, I created a calculated field, changing the null values to 0 and replacing the original Monthly card spend column with this calculated field. This will allow the option of analyzing monthly card spending without excluding or eliminating the customers who are not spending anything on their credit cards.

2.1.6 Voice service usage and Voice service tenure usage columns

As with the Monthly card spend column, I believed that voice service tenure usage had null values that logically could be replaced by zeros. However, when I later examined that data and noticed this was not logical based on the Voice service usage values, I concluded that this had been a mistake. However, I ended up removing both of these fields as I was concerned with the services customers had, and not the amount of usage of these services, in my analysis.

2.1.7 Equipment monthly usage, Equipment lifetime usage, Monthly data usage, and Lifetime data usage columns

For the Equipment monthly usage, Equipment lifetime usage, Monthly data usage, and Lifetime data usage columns, I created calculated fields to remove the $ symbol and changed null values to 0. I then changed the fields from string values to number (decimal) values. As with 2.1.6 Voice Service usage and Voice service tenure usage columns, I ended up removing both of these fields as I was concerned with the services customers had, and not the amount of usage of these services, in my analysis.

2.1.8 Removal of irrelevant data

With the guiding questions I had created while considering the scenario and my examination of the fields in the data set, I carefully selected the following columns to remove: Union member, Number of pets, Number of cats, Number of dogs, Number of birds, Vehicles owned, Vehicle ownership type, Vehicle brand, Political affiliation, Voting history, Active fitness, Credit card type, Commute time, Card tenure, Monthly card items, Monthly card spend, and Loan default. I also considered eliminating retired customers, but after some reflection and examination of the bar graph in Figure 5, I determined that there were a number of retired customers who had not completed high school. I did not want the program to discriminate against customers based on age. Furthermore, most retired people are on fixed incomes and could benefit from this program.

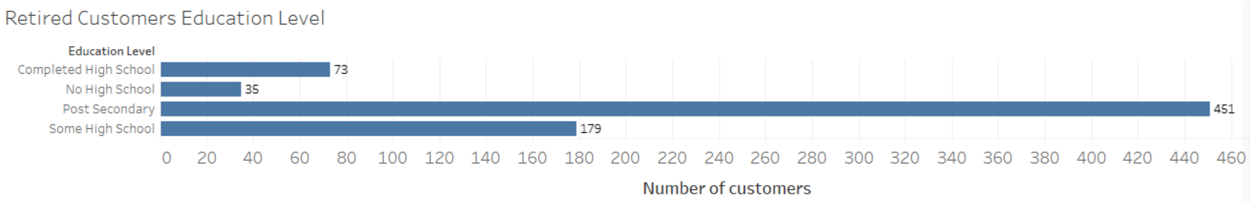


Figure 5 *This bar graph shows the Education Level of retired customers.*

3 Demographic information

3.1 Education level

One of the data fields of particular interest to me in this scenario is Education years. When I began creating visualizations, such as the one shown in Figure 6, I noticed that it was difficult to interpret information that may be helpful. While it is clear in Figure 6 that customers with more years of education have a higher annual household income, many of the values for each number of years below $150,000 are clustered and overlap.

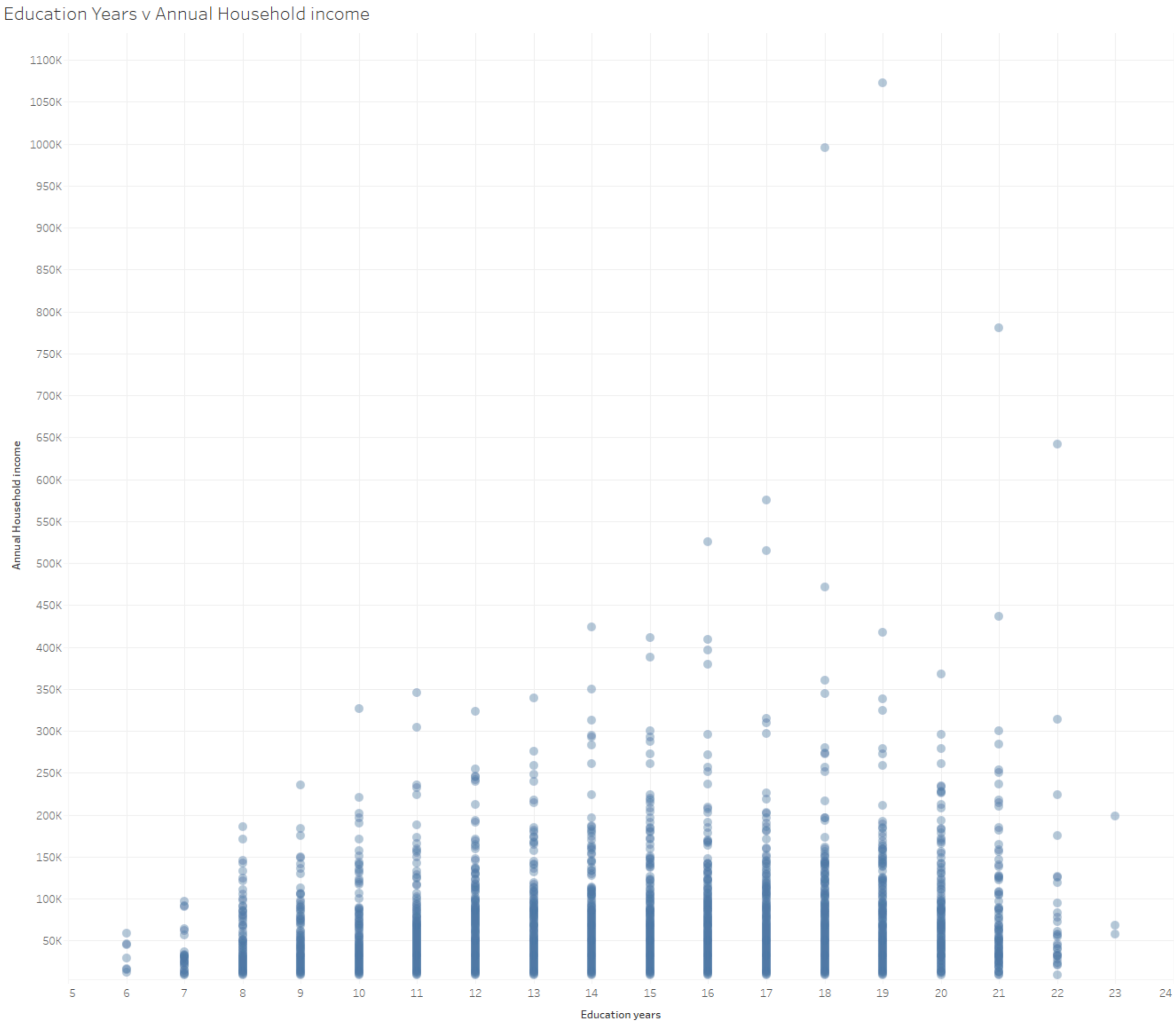


Figure 6 *The scatterplot in this visualization shows the relationship between number of years of education and annual household income.*

I considered how I might be able to better understand the relationships between a customer's education and the data values relating to finances (i.e., annual household income, debt to income ratio). I examined the distribution of education years in a histogram (shown in Figure 7) and also considered how to logically group the number of Education years. I felt as though I had two choices at this point - to arbitrarily choose ranges to break down Education years by or to break down Education years into logical categories. I opted for the latter because it felt as though it would help identify customers in more need of further education - particularly those who had not completed high school. While those who had studied beyond high school would constitute a greater number, their additional years of education put them in a lesser position to benefit from the company's new program.

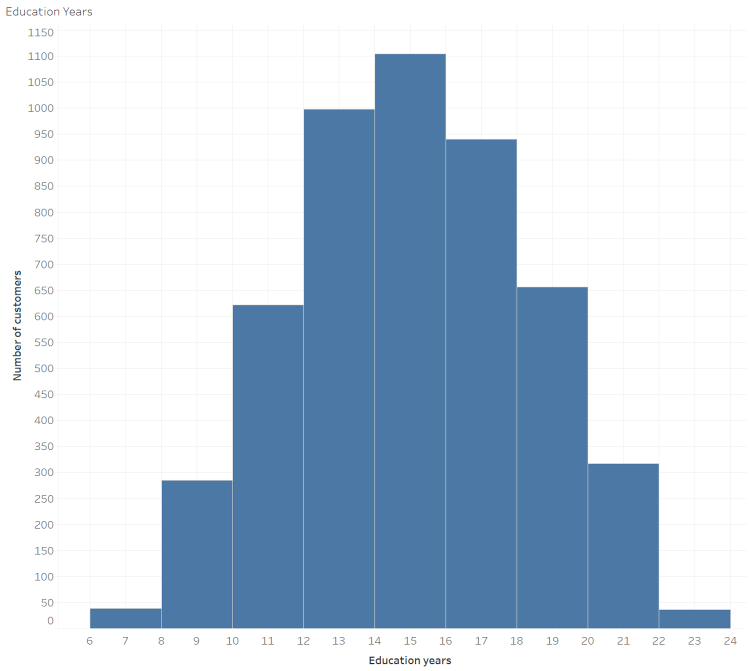


Figure 7 *This histogram shows the distribution of customers across Education years with a bin size of 2 years.*

Having made my decision, I created a calculated field using the formula:

```
IF ([Education years]<=8) THEN "No High School"
ELSEIF (9<=[Education years] AND [Education years] <= 11) THEN "Some High School"
ELSEIF ([Education years] = 12) THEN "Completed High School"
ELSEIF ([Education years] >= 13) THEN "Post Secondary"
ELSE "Other"
END
```

I included ELSE “Other” to ensure that no values fell in that category, indicating that I accidentally left data values out. This ended up being a good strategy as I was able to figure out that I had at first excluded [Education years] = 13; I fixed this oversight by changing [Education years] > 13 to [Education years] >=13.

Having created the Education level field, I could now aggregate by Education level when analyzing financial data, as shown by the box plots in Figure 8. These box plots clearly visualize the discrepancies between Annual Household Income and Education level; while the bottom 50% of incomes are roughly the same regardless of Education level, the top 50% vary greatly, with Post Secondary having the most customers making six figures.

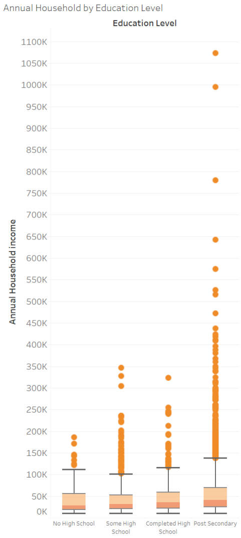


Figure 8 *This visualization has four box plots showing the Annual Household income by Education level.*

3.2 Age Groups

When I first started exploring the variables of interest to me, I created the scatter plot shown in Figure 9 to see the relationship between Annual Household income and Percent Non-credit Debt. Considering that with age, people may need to take out car loans or mortgages (which would then likely be paid off as they reach a more advanced age), I wanted to see the breakdown based on age. As can be seen in Figure 9, the large amount of data makes it difficult to see much, particularly near the vertical axis. Having successfully broken down Education years into groups, I decided to go forward with a similar strategy for age.

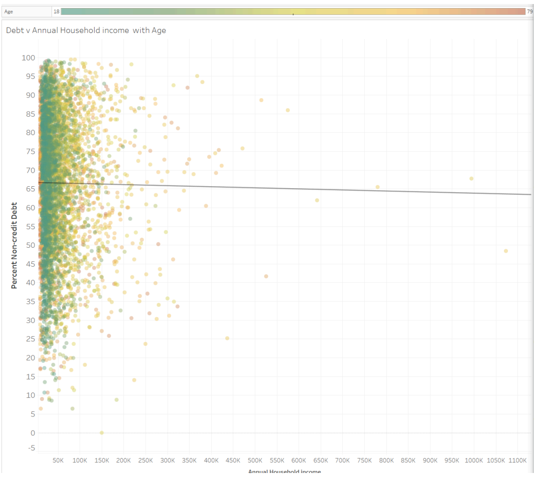


Figure 9 *In this scatter plot, I could see that there may be a weak negative correlation between Annual Household income and Percent Non-credit debt; furthermore, there seems to be more younger customers with a lower Annual Household income but it is hard to decipher with so many data values.*

To break down the ages into groups, I once again researched logical age groups and examined the distribution of ages in a histogram (shown in Figure 10). After experimenting with the bin width I determined that the data was roughly uniform, with fewer customers at the extremes (under 20 and over 70). Considering my research and the distribution of the data, I chose how I wanted to break down the ages: 18-24, 25-34, 35-44, 45-54, 55-64, 65-74, over 75.

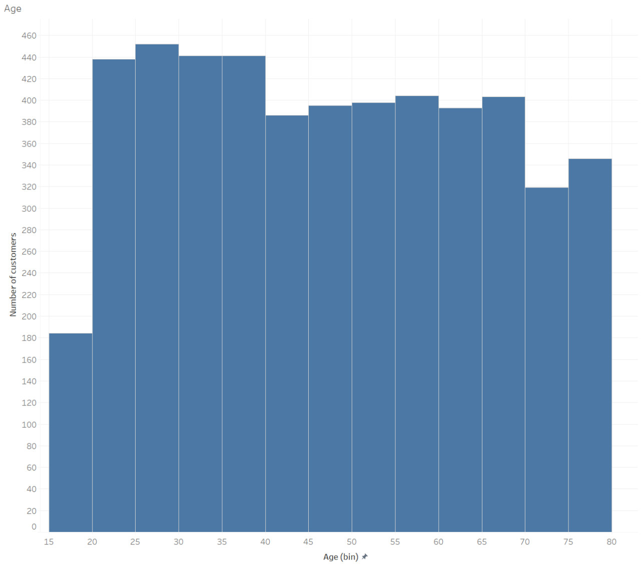


Figure 10 *This histogram shows the distribution of customers' ages with a bin size of 5 years.*

With the age groups determined, I created a calculated field using the formula:

```
IF ([Age]>=18 AND [Age]<=24) THEN "18 to 24"
ELSEIF (25<=[Age] AND [Age] <= 34) THEN "25 to 34"
ELSEIF (35<=[Age] AND [Age] <= 44) THEN "35 to 44"
ELSEIF (45<=[Age] AND [Age] <= 54) THEN "45 to 54"
ELSEIF (55<=[Age] AND [Age] <= 64) THEN "55 to 64"
ELSEIF (65<=[Age] AND [Age] <= 74) THEN "65 to 74"
ELSE "Over 75"
END
```

To ensure no values were missing or double counted, I added the totals from each age group to verify there were 5,000 data values. I also checked the catch all "Over 75" category, making sure the only data values included in that category were indeed 75 and above.

With the categories for both Education level and Age group, I was able to create a series of scatter plots (shown in Figure 11) to see income and debt for different ages and levels of education. This visualization clearly shows where there are differences between age groups for each level of education - particularly in the age groups that are 45 and above. For example, when filtering by Annual household income, I discovered that the only customers with incomes above $500,000 had Post Secondary education and are between the ages of 45 and 74.

With the Education years and Age now grouped in the Education Level and Age Group fields, I added a clean step and removed the two former columns from the data set.

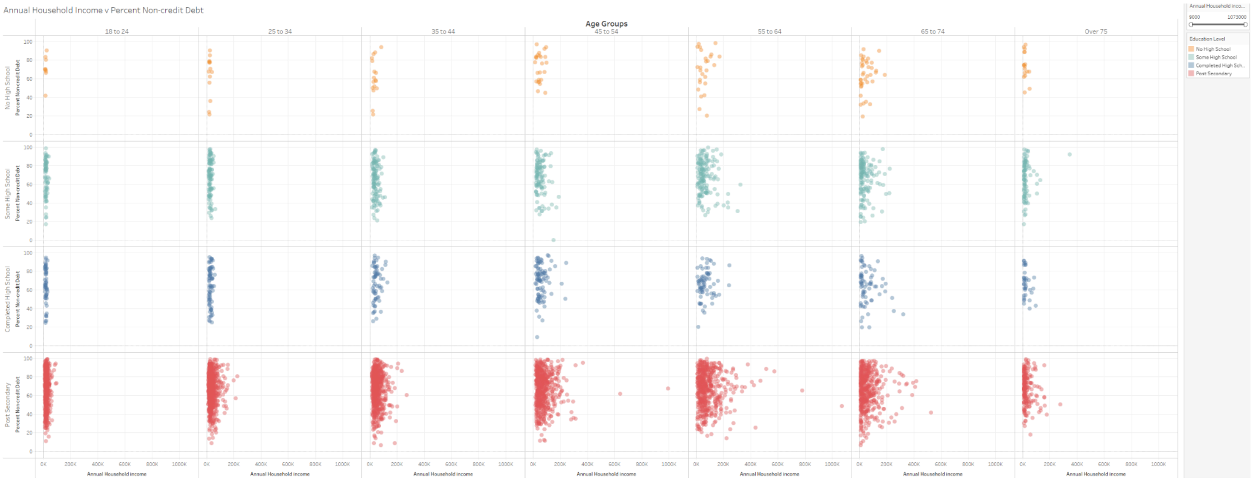


***Figure 11** This visualization shows a series of scatterplots for each combination of education level and age group, comparing annual household income and percent non-credit debt.*

4 Financial information

4.1 Percent Non-credit Debt Column

The program proposed is intended for customers in need to access services, plans, and devices to advance their education or their dependents' education. To decide how best to summarize the financial situation, I began by creating and examining the relationship between Credit Debt and Other Debt, shown in Figure 12. This graph showed a weak positive correlation between the variables and that most customers fell in the bottom left corner of the graph.

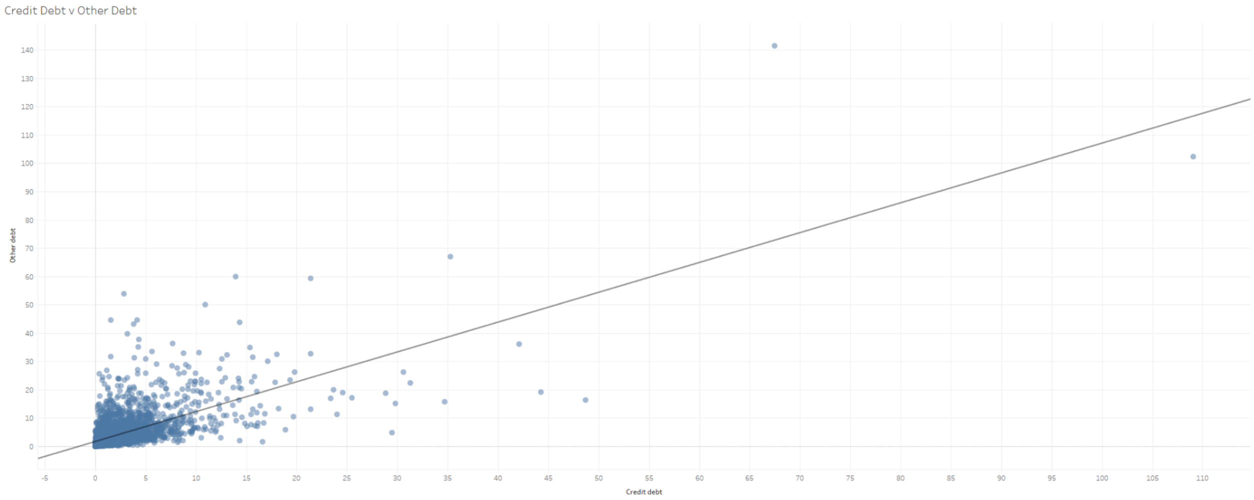


***Figure 12** This scatter plot shows a weak positive relationship between credit debt and other debt.*

Knowing that I wanted a value that could compare a customer's credit debt to their other debt, I decided to create a calculated field that I named Percent Non-Credit Debt. The values in this column represent the percentage of debt that falls in the "other" category (in other words, Non-credit debt). I calculated the field using the formula:

[Other debt]/([Credit debt]+[Other debt])*100

When examining the new field, I discovered that there was a single null value. Upon investigating why it was null, I discovered that the debt was $0 in that case so I changed the null value to 0. Once this field was created, I added a clean step and removed credit debt and other debt fields since the new field presented the same information in a different form (as a percentage). With this new field, I could clearly see that most customers' debt was mostly non-credit debt (over 50%), as seen in Figure 13.

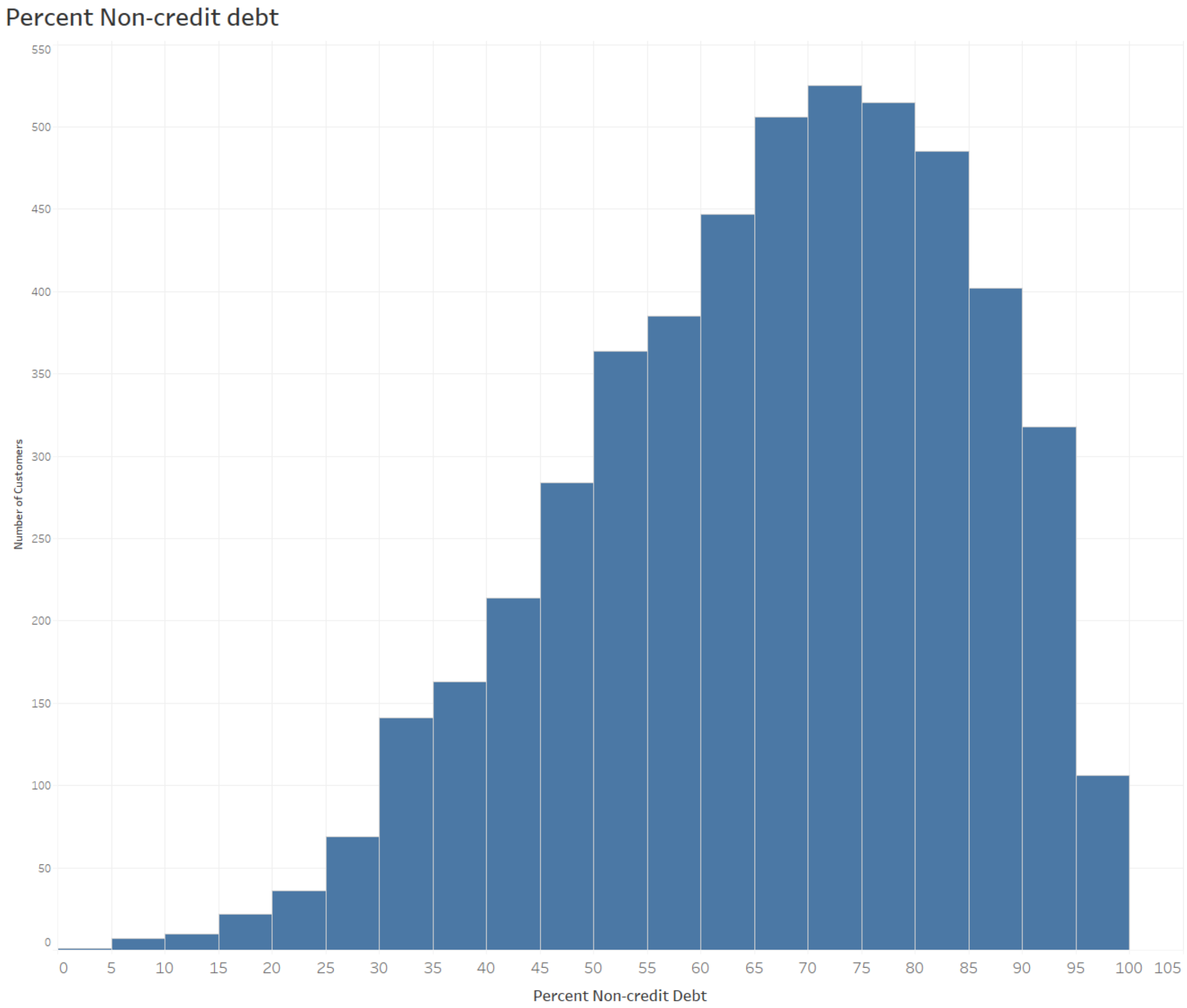


Figure 13 *This histogram shows the distribution of Percent Non-credit debt.*

5 Services, Plans, Devices, and Subscriptions

5.1 Plans

Once the qualifying customers are selected, I know I will want to be able to look at the type(s) of plans each customer had as well as the number of plans, etc. This will allow me to color code and filter data to see which customers are in need of which services, plans, devices, and subscriptions. Initially, when beginning to consolidate and rethink the data values around the different services, plans, devices, and subscriptions offered, I began to create a field for Services and Plans. My first attempt at a formula was as follows:

```
IF [Voice service usage]>0 AND [International plan] = "No" AND [Wireless data plan] = "No" THEN "Voice service"
ELSEIF [Voice service usage]>0 AND [International plan] = "Yes" AND [Wireless data plan] = "No" THEN "Voice service, International plan"
ELSEIF [Voice service usage]>0 AND [International plan] = "No" AND [Wireless data plan] = "Yes" THEN "Voice service, Wireless data plan"
ELSEIF [Voice service usage]>0 AND [International plan] = "Yes" AND [Wireless data plan] = "Yes" THEN "Voice service, International plan, Wireless data plan"
ELSEIF [Voice service usage]=0 AND [International plan] = "Yes" AND [Wireless data plan] = "Yes" THEN "International plan, Wireless data plan"
ELSEIF [Voice service usage]=0 AND [International plan] = "Yes" AND [Wireless data plan] = "No" THEN "International plan"
ELSEIF [Voice service usage]=0 AND [International plan] = "No" AND [Wireless data plan] = "Yes" THEN "Wireless data plan"
ELSEIF [Voice service usage]=0 AND [International plan] = "No" AND [Wireless data plan] = "No" THEN "No plans or services"
END
```

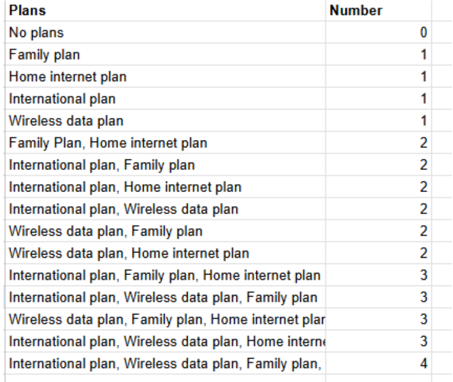
As I worked on making sure I included all service and plan types, I realized that my formula was quickly becoming complicated, so I decided to create two separate fields. When I first revised the field to just include Plans, I used this formula:

```
IF [International plan] = "No" AND [Wireless data plan] = "No" THEN "No plans"
ELSEIF [International plan] = "Yes" AND [Wireless data plan] = "No" THEN "International plan"
ELSEIF [International plan] = "No" AND [Wireless data plan] = "Yes" THEN "Wireless data plan"
ELSEIF [International plan] = "Yes" AND [Wireless data plan] = "Yes" THEN "International plan, Wireless data plan"
END
```

I then realized that there was also a Family plan and Home internet plan. In order to include them, I would need to account for the 16 combinations that could happen. Before rewriting the formula, I also had to transform the Home internet plan field so the data was Boolean and not strings. To do so, I created a calculated field to change Other values (e.g., “4”) to “Yes”. I then proceeded to write what would be the final formula for the Plans field:

```
IF [International plan] = "No" AND [Wireless data plan] = "No" AND [Family plan] = "No" AND [Home internet plan] = "No" THEN "No plans"
ELSEIF [International plan] = "Yes" AND [Wireless data plan] = "No" AND [Family plan] = "No" AND [Home internet plan] = "No" THEN "International plan"
ELSEIF [International plan] = "No" AND [Wireless data plan] = "Yes" AND [Family plan] = "No" AND [Home internet plan] = "No" THEN "Wireless data plan"
ELSEIF [International plan] = "Yes" AND [Wireless data plan] = "Yes" AND [Family plan] = "No" AND [Home internet plan] = "No" THEN "International plan, Wireless data plan"
ELSEIF [International plan] = "Yes" AND [Wireless data plan] = "Yes" AND [Family plan] = "Yes" AND [Home internet plan] = "No" THEN "International plan, Wireless data plan, Family plan"
ELSEIF [International plan] = "No" AND [Wireless data plan] = "Yes" AND [Family plan] = "Yes" AND [Home internet plan] = "No" THEN "Wireless data plan, Family plan"
ELSEIF [International plan] = "Yes" AND [Wireless data plan] = "No" AND [Family plan] = "Yes" AND [Home internet plan] = "No" THEN "International plan, Family plan"
ELSEIF [International plan] = "No" AND [Wireless data plan] = "No" AND [Family plan] = "Yes" AND [Home internet plan] = "No" THEN "Family plan"
ELSEIF [International plan] = "No" AND [Wireless data plan] = "No" AND [Family plan] = "Yes" AND [Home internet plan] = "Yes" THEN "Family plan, Home internet plan"
ELSEIF [International plan] = "No" AND [Wireless data plan] = "Yes" AND [Family plan] = "No" AND [Home internet plan] = "Yes" THEN "Wireless data plan, Home internet plan"
ELSEIF [International plan] = "Yes" AND [Wireless data plan] = "No" AND [Family plan] = "No" AND [Home internet plan] = "Yes" THEN "International plan, Home internet plan"
ELSEIF [International plan] = "No" AND [Wireless data plan] = "No" AND [Family plan] = "No" AND [Home internet plan] = "Yes" THEN "Home internet plan"
ELSEIF [International plan] = "Yes" AND [Wireless data plan] = "Yes" AND [Family plan] = "No" AND [Home internet plan] = "Yes" THEN "International plan, Wireless data plan, Home internet plan"
ELSEIF [International plan] = "Yes" AND [Wireless data plan] = "No" AND [Family plan] = "Yes" AND [Home internet plan] = "Yes" THEN "International plan, Family plan, Home internet plan"
ELSEIF [International plan] = "No" AND [Wireless data plan] = "Yes" AND [Family plan] = "Yes" AND [Home internet plan] = "Yes" THEN "Wireless data plan, Family plan, Home internet plan"
ELSEIF [International plan] = "Yes" AND [Wireless data plan] = "Yes" AND [Family plan] = "Yes" AND [Home internet plan] = "Yes"
THEN "International plan, Wireless data plan, Family plan, Home internet plan"
END
```

Now that I had the plans for each customer consolidated, I researched how to create a field that would provide a count of the number of plans. My first step was to create the mapping table shown in Figure 14. This table summarizes the combinations that a customer could have as well as the number of plans associated with each combination.



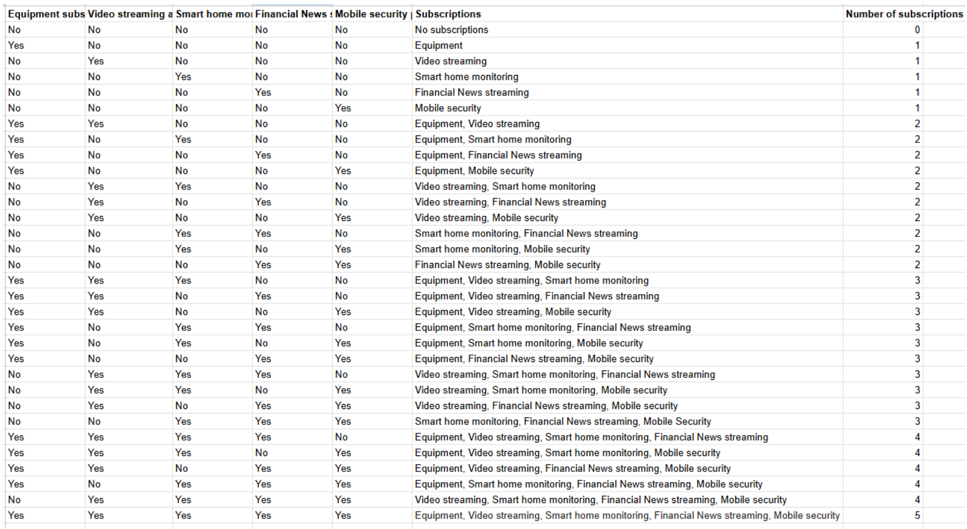
Plans	Number
No plans	0
Family plan	1
Home internet plan	1
International plan	1
Wireless data plan	1
Family Plan, Home internet plan	2
International plan, Family plan	2
International plan, Home internet plan	2
International plan, Wireless data plan	2
Wireless data plan, Family plan	2
Wireless data plan, Home internet plan	2
International plan, Family plan, Home internet plan	3
International plan, Wireless data plan, Family plan	3
Wireless data plan, Family plan, Home internet plar	3
International plan, Wireless data plan, Home interne	3
International plan, Wireless data plan, Family plan,	4

Figure 14 *This is the mapping table used to create the calculated field, Number of plans.*

I exported the table as a csv file and then connected it with the Customer data flow. I created a clean step in the data flow then joined the mapping table with it. I removed the duplicate field, Plans, after ensuring no data was lost. At first, I was concerned that I had expected 6 values in plans to correspond with 2 plans until I realized that two of the combinations resulting in 2 plans contained zero customers. When I later returned to the data flow, I realized that I had lost rows in this join. To prevent this omission, I went back and made it a left join to ensure the rows from the original data table were preserved.

5.2 Subscriptions

Having seen the easier method of using a mapping table, I chose to use this technique to try to create both the Subscriptions and Number of Subscriptions fields. I once again created a mapping table in Google sheets (shown in Figure 15) and downloaded it as a csv file that I could connect to the customer data.



Equipment subs	Video streaming a	Smart home mo	Financial News	Mobile security	Subscriptions	Number of subscriptions
No	No	No	No	No	No subscriptions	0
Yes	No	No	No	No	Equipment	1
No	Yes	No	No	No	Video streaming	1
No	No	Yes	No	No	Smart home monitoring	1
No	No	No	Yes	No	Financial News streaming	1
No	No	No	No	Yes	Mobile security	1
Yes	Yes	No	No	No	Equipment, Video streaming	2
Yes	No	Yes	No	No	Equipment, Smart home monitoring	2
Yes	No	No	Yes	No	Equipment, Financial News streaming	2
Yes	No	No	No	Yes	Equipment, Mobile security	2
No	Yes	Yes	No	No	Video streaming, Smart home monitoring	2
No	Yes	No	Yes	No	Video streaming, Financial News streaming	2
No	Yes	No	No	Yes	Video streaming, Mobile security	2
No	No	Yes	Yes	No	Smart home monitoring, Financial News streaming	2
No	No	Yes	No	Yes	Smart home monitoring, Mobile security	2
No	No	No	Yes	Yes	Financial News streaming, Mobile security	2
Yes	Yes	Yes	No	No	Equipment, Video streaming, Smart home monitoring	3
Yes	Yes	No	Yes	No	Equipment, Video streaming, Financial News streaming	3
Yes	Yes	No	No	Yes	Equipment, Video streaming, Mobile security	3
Yes	No	Yes	Yes	No	Equipment, Smart home monitoring, Financial News streaming	3
Yes	No	Yes	No	Yes	Equipment, Smart home monitoring, Mobile security	3
Yes	No	No	Yes	Yes	Equipment, Financial News streaming, Mobile security	3
No	Yes	Yes	Yes	No	Video streaming, Smart home monitoring, Financial News streaming	3
No	Yes	Yes	No	Yes	Video streaming, Smart home monitoring, Mobile security	3
No	Yes	No	Yes	Yes	Video streaming, Financial News streaming, Mobile security	3
No	No	Yes	Yes	Yes	Smart home monitoring, Financial News streaming, Mobile Security	3
Yes	Yes	Yes	Yes	No	Equipment, Video streaming, Smart home monitoring, Financial News streaming	4
Yes	Yes	Yes	No	Yes	Equipment, Video streaming, Smart home monitoring, Mobile security	4
Yes	Yes	No	Yes	Yes	Equipment, Video streaming, Financial News streaming, Mobile security	4
Yes	No	Yes	Yes	Yes	Equipment, Smart home monitoring, Financial News streaming, Mobile security	4
No	Yes	Yes	Yes	Yes	Video streaming, Smart home monitoring, Financial News streaming, Mobile security	4
Yes	Yes	Yes	Yes	Yes	Equipment, Video streaming, Smart home monitoring, Financial News streaming, Mobile security	5

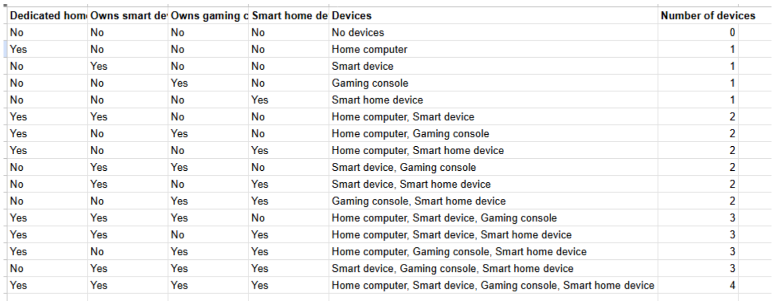
Figure 15 *This is the mapping table used to create the calculated fields for Subscriptions and Number of Subscriptions.*

I then joined the Customer data table from my most recent clean step with the Subscription mapping table using an inner join. After the join, I checked to make sure that there were still 5,000 rows and also checked that the totals for 0, 1, 2, 3, 4, and 5 subscriptions were correct. During this process, I discovered that customers were being counted more than once. When I looked at the settings, there was only a join clause for Financial News streaming subscriptions so I added join clauses for the rest of the subscription types. Once I made this correction, the values were as expected.

Once the new fields (Subscriptions and Number of subscriptions) were as I wanted them to be, I added a clean step and removed the following columns: Equipment subscription, Equipment monthly usage, Equipment lifetime usage, Video streaming add-on, Smart home monitoring, Mobile security package, and Financial News streaming subscription.

5.3 Devices

With my confidence growing in my ability to use mapping tables, I continued with this technique to create two more fields: Devices and Number of devices. I created a mapping table (see Figure 16), downloaded it as a csv, and connected the mapping table to the Customer data.



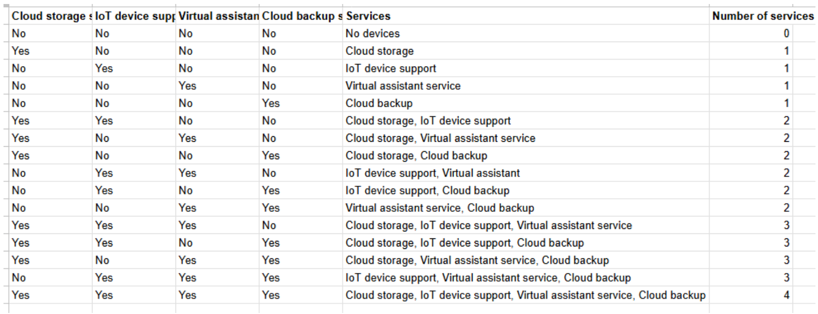
Dedicated hom	Owns smart de	Owns gaming c	Smart home de	Devices	Number of devices
No	No	No	No	No devices	0
Yes	No	No	No	Home computer	1
No	Yes	No	No	Smart device	1
No	No	Yes	No	Gaming console	1
No	No	No	Yes	Smart home device	1
Yes	Yes	No	No	Home computer, Smart device	2
Yes	No	Yes	No	Home computer, Gaming console	2
Yes	No	No	Yes	Home computer, Smart home device	2
No	Yes	Yes	No	Smart device, Gaming console	2
No	Yes	No	Yes	Smart device, Smart home device	2
No	No	Yes	Yes	Gaming console, Smart home device	2
Yes	Yes	Yes	No	Home computer, Smart device, Gaming console	3
Yes	Yes	No	Yes	Home computer, Smart device, Smart home device	3
Yes	No	Yes	Yes	Home computer, Gaming console, Smart home device	3
No	Yes	Yes	Yes	Smart device, Gaming console, Smart home device	3
Yes	Yes	Yes	Yes	Home computer, Smart device, Gaming console, Smart home device	4

Figure 16 *This is the mapping table used to create the calculated fields for Devices and Number of devices.*

Having learned from the mishap when I joined the Subscription mapping table with the Customer data, I made sure to set a join clause for all the columns involving devices. I performed the join between the Devices mapping table and the most updated cleaned Customer data. I checked that values made sense and that there were no duplicates or missing values before adding a clean step and removing the following columns: Smart home devices, Owns smart device, Owns gaming console, and Dedicated home computer.

5.4 Services

Once again, I decided to use the mapping table technique to create two new calculated fields - Services and Number of services. I created the mapping table shown in Figure 17 and then downloaded it as a csv to connect with the Customer data. I joined the Services mapping table with the most recent clean Customer data, ensuring that I set a join clause for each of the corresponding columns. After the join, I checked for missing values or duplicates before adding a clean step and removing the following columns: Cloud storage, IoT device support, Virtual assistant service, and Cloud backup service.

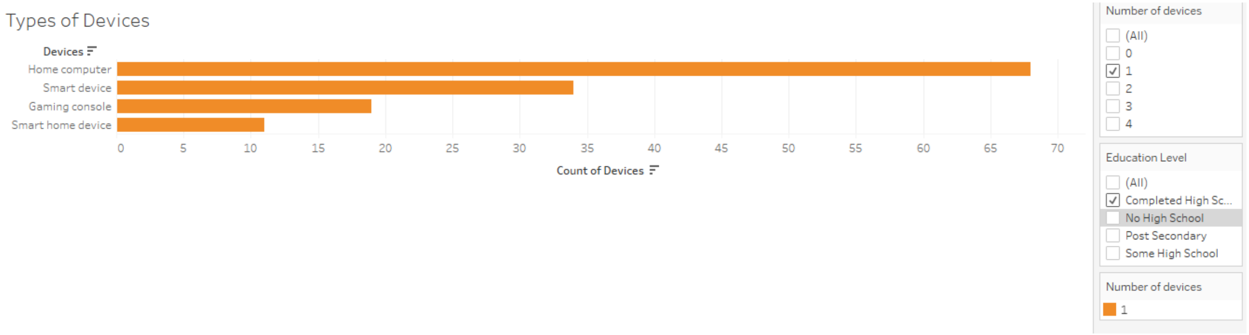


Cloud storage s	IoT device supp	Virtual assistan	Cloud backup s	Services	Number of services
No	No	No	No	No devices	0
Yes	No	No	No	Cloud storage	1
No	Yes	No	No	IoT device support	1
No	No	Yes	No	Virtual assistant service	1
No	No	No	Yes	Cloud backup	1
Yes	Yes	No	No	Cloud storage, IoT device support	2
Yes	No	Yes	No	Cloud storage, Virtual assistant service	2
Yes	No	No	Yes	Cloud storage, Cloud backup	2
No	Yes	Yes	No	IoT device support, Virtual assistant	2
No	Yes	No	Yes	IoT device support, Cloud backup	2
No	No	Yes	Yes	Virtual assistant service, Cloud backup	2
Yes	Yes	Yes	No	Cloud storage, IoT device support, Virtual assistant service	3
Yes	Yes	No	Yes	Cloud storage, IoT device support, Cloud backup	3
Yes	No	Yes	Yes	Cloud storage, Virtual assistant service, Cloud backup	3
No	Yes	Yes	Yes	IoT device support, Virtual assistant service, Cloud backup	3
Yes	Yes	Yes	Yes	Cloud storage, IoT device support, Virtual assistant service, Cloud backup	4

Figure 17 *This is the mapping table used to create the calculated fields for Services and Number of services.*

6 Next Steps and Final Thoughts

With the customer data transformed and cleaned, analysis to determine which customers will be included in the new program will be easier through the use of colors and filters. The analysis can determine combinations and ranges that customers must fulfill to qualify, using visualizations such as the ones shown in Figure 18 and Figure 19. Once these customers are selected, further analysis can be done to determine which plans, subscriptions, devices, and services the customers are in need of or may need financial assistance to continue to have access needed for furthering their education or their dependents' education.



Number of devices
□ (All)
□ 0
☑ 1
□ 2
□ 3
□ 4

Education Level
□ (All)
☑ Completed High Sc...
□ No High School
□ Post Secondary
□ Some High School

Number of devices
🟧 1

Figure 18 *This is an example of a bar graph filtered by both the number of devices and Education Level. Different relationships could be discovered by adjusting the filters.*

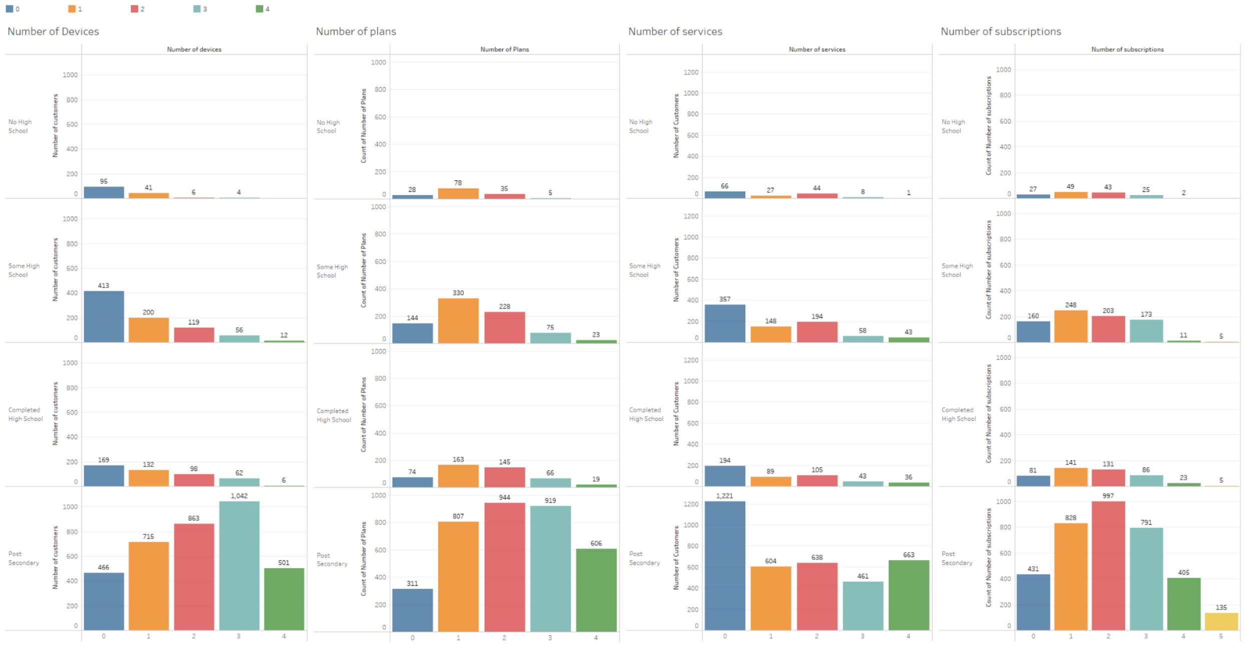


Figure 19 *This dashboard was created to show the Number of devices, plans, services, and subscriptions broken down by Education level.*

In reflecting on the process of simplifying and transforming this data set, I have realized how my approach shifted and changed throughout the project. When first looking at the data set, I was trying to find ways to simplify it and, in retrospect, wish I had used visualizations more frequently. However, I almost needed to spend some time with the data Tableau Prep before creating visualizations in Tableau Desktop. This initial simplification took place almost exclusively around examining and consolidating the fields involving plans, devices, services, and subscriptions.

Often, after adding a new clean data stp, I would open Tableau Desktop to see how the variables looked. After the initial consolidation described above, I began to look at the histograms and scatter plots included throughout this project to inform my transformation of the data. While I kept a running document of the steps I took along the way, I found that once I began working on this report I could see the true value of using the visualizations. In retrospect, I wish I had begun going back and forth between Prep and Desktop sooner but realize that this was part of the learning curve for me. In the future, after initially cleaning the day, I would spend more time looking at visualizations and using graphs to make sense of the data in addition to examining the data table.