**Segmentation & Profiling Project**

Cameron Gallien

DSA5400: Visual Data Exploration

Merrimack University

Prof. Christopher Healey

**Executive Summary**

Background

Upon my arrival at this company, I noticed a severe lack of customer analytics that have been done on our current customer base. Without knowledge of the types of customers we have, it is extremely difficult to analyze potential strategies to increase profits within the company. For that reason, I have decided to learn more about our customer base in order to increase these profits. One of the easiest ways to do that is through customer retention.

The purpose of this report is to identify patterns in our customer base that will help us figure out which types of customers are quick to leave the company, and which are likely to remain with the company for longer periods of time. To do this we will identify different segments of our current customer base in order to drive future decisions on customer retention, and develop potential marketing strategies to retain customers as long as possible.

Process

Two different segmentation methods were utilized to try to categorize our current customer base. The first technique that was used is unsupervised segmentation. By using unsupervised segmentation multiple variables could be plugged into a machine learning program which could find different patterns and clusters of customers that might not immediately be visible. I felt as though even if the method did not produce the most actionable segments, I would be able to gain some valuable knowledge about the customer base that could be used to narrow down the segments in my next approach.

This next approach utilized is rule-based segmentation. This was chosen as it allowed me to take some of the insights gleaned from the unsupervised segmentation and narrow down the segments to find a more specific and effective grouping of customers. With these segments being more precisely tuned to specific variables it would be easier to take actionable measures in regards to customer retention. In this method I used a mixture of demographic and behavioral variables to create the clusters, which seemed most likely to produce actionable marketing insights for the company.

Results

The unsupervised segmentation approach showed an optimal segmentation of about 4-5 clusters. The method provided 4 segments, with the most obvious trend being age. Those segments with the highest average age had the longest average tenure with the company compared to other segments. Aside from the insights regarding age, it was tough to find many other actionable trends.

I took the information from the unsupervised approach and expanded upon it with the rules-based segmentation approach. When analyzing the data more thoroughly, I noticed a positive correlation between customers who subscribed to the news, and average tenure with the company. By combining the variables of customer age, and whether or not a customer subscribed to the news, I was able to create 6 clusters of customers with significant differences in average tenure. The cluster that contained the oldest customers who subscribed to the news showed a significantly higher average tenure with the company. This segment was therefore nicknamed “Aged & Informed.”

After noticing the trend regarding news subscriptions, I revised the unsupervised approach to add news subscriptions as an additional variable. This created a much bigger difference in average tenure between the segments, and backed up what was shown in the rules-based approach regarding whether or not the customer subscribed to the news.

Recommendations

While both approaches ended up with similar clusters, I believe the rule-based approach provided the strongest results. This is because the Aged & Informed segment from this method had a higher average company tenure than the highest segment from the unsupervised approach. Also, by explicitly separating customers based on these rules, it is also easier to take meaningful steps and specifically target these groups with potential marketing campaigns.

Even though the unsupervised approach had very similar findings, the end result of these segments did not show as big of a difference as the segments from the rules-based approach. Clusters 1 & 2, and clusters 3 & 4 showed very few actionable differences between them, which left much more insight available in the rules-based approach.

Because of these trends, I believe it is important to focus marketing efforts on news publications, as those who read the news tend to be the longest tenured customers. I recommend two different marketing strategies targeting publications with an older audience, as well as those with a younger audience. Advertising in publications with older audiences seems obvious, as the customers in this range tend to stay with the company longer and will drive up the average tenure. I also believe that we should target publications with younger audiences as well. These customers still have higher tenures than their non-subscribing counterparts, and since they should live longer than the older customer base their potential customer tenure is even higher. A loyalty program of some sort, perhaps one where your monthly bill price fluctuates less based on your tenure with the company, might be a good idea to showcase in these publications to try to keep these customers onboard with the company for as long as possible.

**Technical Report**

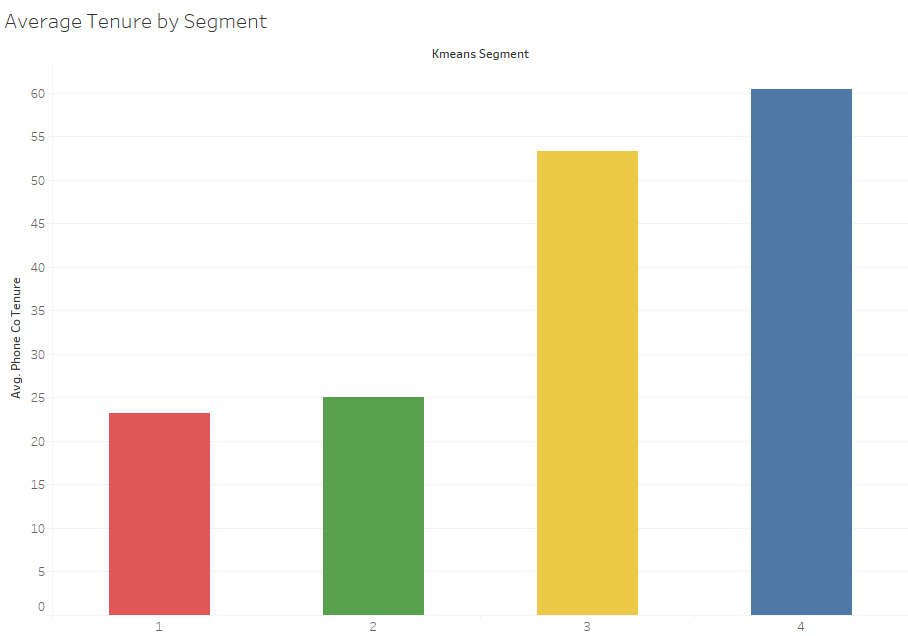
Unsupervised Segmentation

Unsupervised segmentation analysis (more specifically, K-Means clustering) consists of examining specific variables and using machine learning to group customers together based on similar characteristics within those variables. The variables that I decided to utilize were tenure, age, household income, total pieces of technology, average spent per month, and years of education. After seeing the results of the rules-based segmentation which will be touched upon later, the variable of news subscribers was also added to the analysis. The graph that was produced below showed that the optimal amount of clusters fell somewhere between 3 & 7.

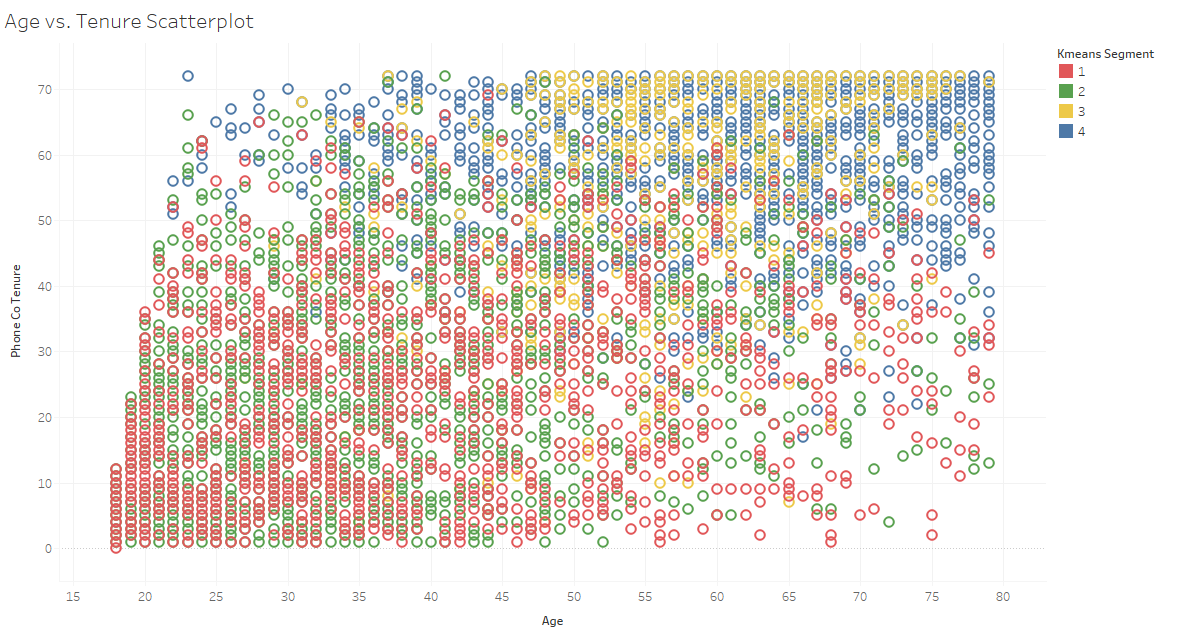
A picture containing line, diagram, text, plot

Description automatically generated

When running 7 clusters the data became too spread out to obtain any knowledgable insights from, and trends were difficult to find. This led to a 3 cluster experiment, which did show the trend of age being a factor, however one of the segments was much bigger than the others. After some tinkering, I decided upon 4 clusters as these clusters showed the biggest difference between average phone tenure.



As you can see from the bar chart, segments 3 & 4 have the highest average phone tenures out of all segments, but there isn’t an overwhelming difference between them. The scatterplot below reinforces this, as the yellow and blue segments (for segment 3 & 4 respectively) are clustered very strongly in the upper part of the chart.



Still, with the provided visualizations, it is difficult to see the differences between many of the groups. Once this information was placed in a table, it was easier to see the differences between the segments.

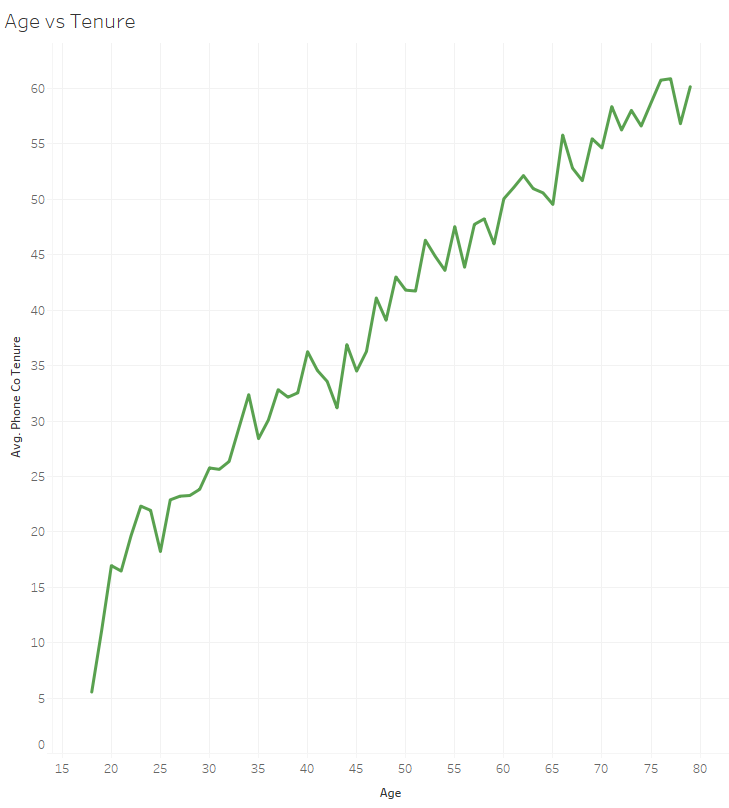
A picture containing text, font, line, screenshot

Description automatically generated

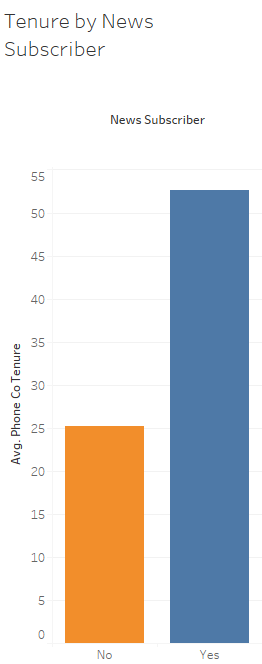
The biggest difference came down to household income, in which segment 3 has a much higher average household income than the other 3 segments, which is one valuable insight that was not as apparent in the rules-based approach. Also, you can see that the average monthly bill (Avg. All Svc Month) was extremely low for segment 1. But as you can see, the first two clusters and the second two clusters are very similar in terms of tenure with the company, which makes it difficult to truly target the customers in each specific segment.

Rules-Based Segmentation

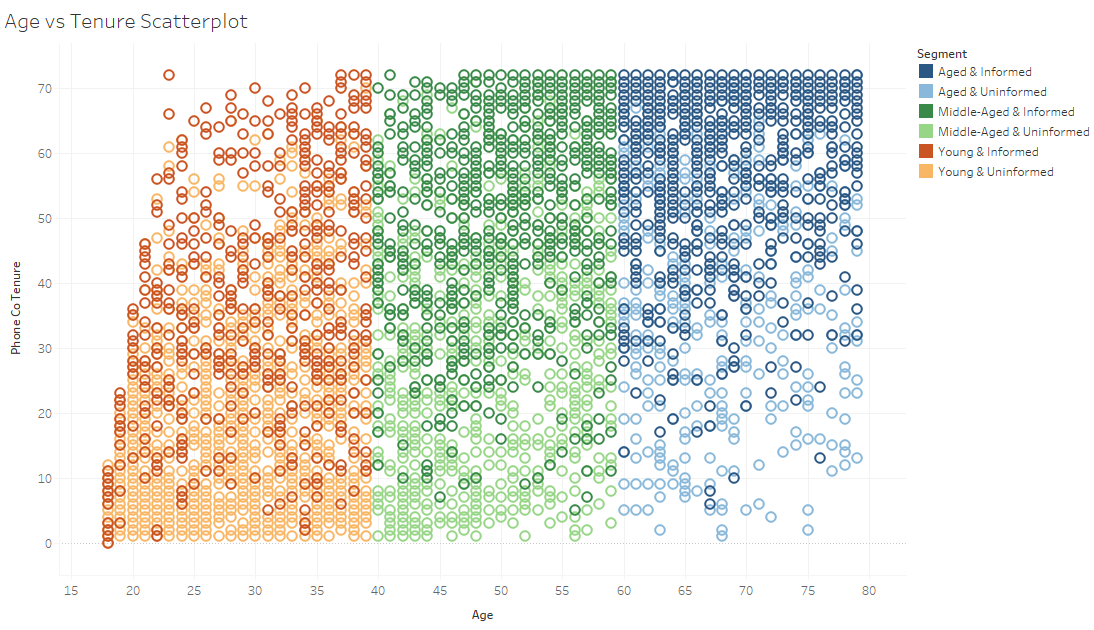
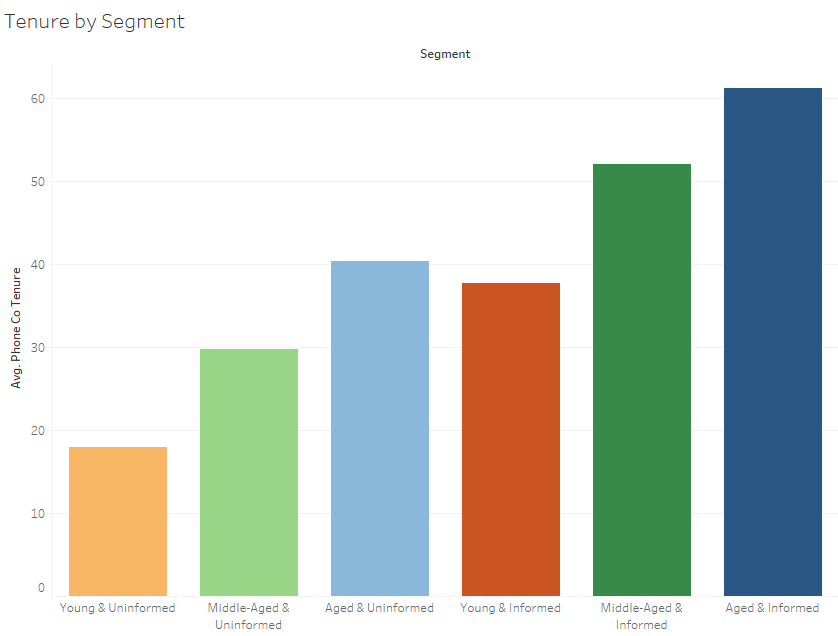
Rules-based segmentation involves selecting specific variables associated with customer values to group customers into specific clusters. Since the unsupervised approach showed a correlation between age and tenure with the company, I decided delve further into this variable.



This line chart shows the obvious correlation between the two variables, which would be enough to produce actionable insights on its own, but I wanted to see if there were other values that could further segment our data. This led me to the News Subscriber variable.



As the graph above clearly shows, the average tenure with the company more than doubles when a customer is a news subscriber. These two variables were the strongest I could find, so I decided to use Age and News Subscriber as my two variables. I separated customers into 3 age groups: Young (18-39), Middle-Aged (40-59), and Aged (60 and older). In addition, I separated these groups by whether or not they subscribed to the news to get 6 segments: Young/Middle-Aged/Aged & Uninformed, and Young/Middle-Aged/Aged & Informed.



A screenshot of a graph

Description automatically generated with low confidence

While I knew that age and news subscribers had a positive correlation to tenure, I was not expecting the results to appear so evidently. The bar chart shows a clear positive trend in tenure as the “uninformed” segments age (the left 3 bars), which continues with the “informed” segments (right 3 bars). The scatterplot reinforces this, as the dark green and dark blue segments (Middle-Aged & Informed, and Aged & Informed, respectively) are much more heavily clustered near the top of the plot. The table shows very few other trends, but one interesting one is that these same two segments have the highest average monthly bill (Avg. All Svc Month) which is even more reason why these customers should be targeted.

From this data it is easy to see why news subscribers, specifically those in the Aged & Informed segment should be targeted. This segment not only has the highest retention level, but also has the highest average monthly bill, which means that these are our most high value customers.