Springboard Data Science Career Track
Capstone Project 3
‘Customer Segmentation and Churn Analysis’
By: Nicholas Dean
September 2021

# 1. Introduction

**Defining the problem**

Customer churn is a significant problem for any business. According to HubSpot: "*Customer churn is the percentage of customers that stopped using your company's product or service during a certain time frame*". Estimates vary, but depending on the industry, it can cost as much as 10x more to acquire a new customer than the cost of retaining customers who might churn. Additionally, Client Success estimates that existing customers will purchase additional products at a conversion rate of 65%, while acquiring brand new customers has about a 13% conversion rate - cultivating existing customers is clearly more effective. The odds are significantly in favor of retaining existing customers rather than finding new ones. Customer segmentation is a strong strategy for product marketing and sales teams to describe broad categories of their customers, and when paired with churn analysis can be used to derive targeted and actionable recommendations for improving customer retention.

The purpose of this project is to use clustering models to establish customer segments within a customer base, and to pair this with a broader churn analysis to identify factors which impact churn. Ultimately this will be used to make clear recommendations to stakeholders (leaders in marketing, sales, and operations teams) that can be used to further study churn among individual customer segments or which can be immediately implemented or tested to reduce churn.

**Summary of results.**

Using clustering algorithms three distinct customer segments were uncovered and defined:

1. The first is primarily made up of households with single, middle-aged, low volume customers.
2. Another contains large households which are price-conscious, but high-valued.
3. The final segment consists of mature households with no children, who shop less frequently, but are still high-valued.

These customer segments were created using a subset of the customer data for which I also had demographic information, which could be used to better understand spending patterns within each group.

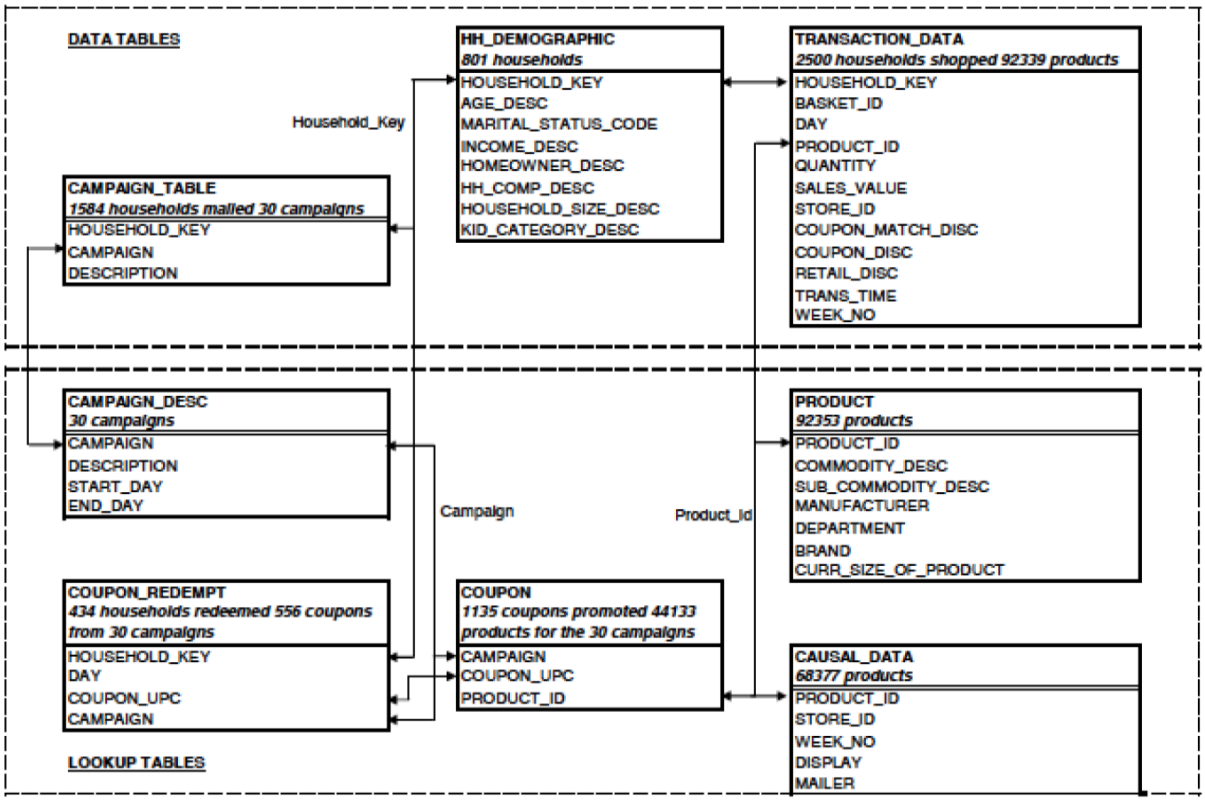
Churn was ultimately defined as households that have not purchased from the grocery store in the past 30 days. Using this definition, 18.16% of households ('households' is used interchangeably with 'customers' throughout this report) are currently considered churned. The model analysis revealed that the average shopping lag and frequency were by far the most important factors in predicting customer churn, closely followed by the average number of items

purchased and the percentage of purchases containing an item that was advertised in a mailer.

## 2. Approach

**Data acquisition & wrangling.**

The data for this project was sourced from a Dunnhumby case study titled "The Complete Journey" and was created as a study for 711 days of marketing mailing campaigns on customer purchasing behavior with a grocery store chain. While the data was initially found on the Kaggle competition website, I downloaded the data directly from Dunnhumby with the documentation and relational diagram. The data was represented as multiple tables, and loaded into a SQLite3 database according to the diagram provided by Dunnhumby below.




Using these tables I first needed to determine how I could assemble a row-wise representation of a single household, since the purpose of my project was not to assess the efficacy of different marketing campaigns, but rather to examine and describe the households themselves. This approach allowed me to identify what kind of information could be aggregated on each household and passed to machine learning algorithms for classification purposes.

Ultimately I extracted 56 different aggregate measures of customer behavior, across three distinct categories:

- Coupon and Campaign Redemption information (Percent of TypeA campaigns sent, Count of redeemed coupons, etc)
- Product information (Count of unique products purchased, percent of products purchased from a private brand, etc)
- Transaction information (Average sales value, max coupon discount, etc)

Once this information was queried, I combined all of the query results into a single data frame with a row for each household, and a column for each of the queries. With this done, I was able to start to conduct some basic data wrangling and better understand the data set I've just created.

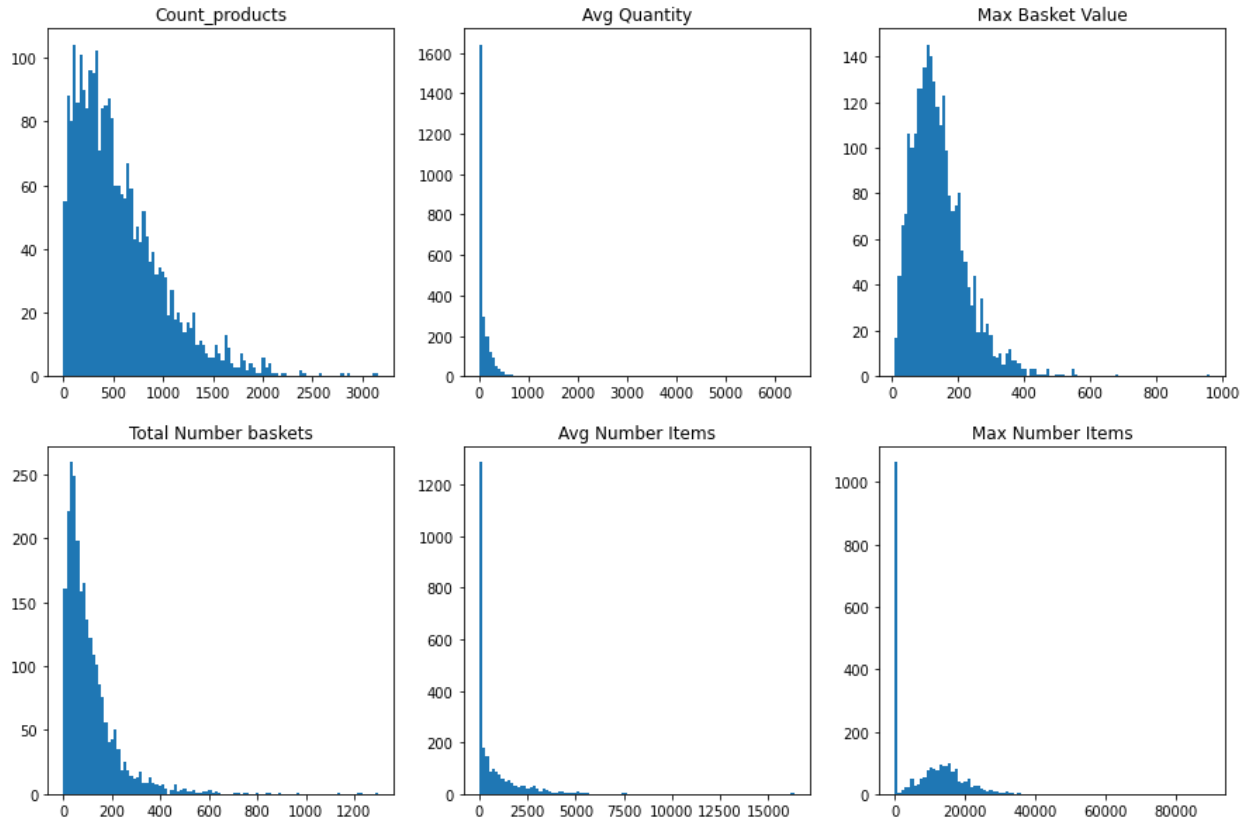
In the course of creating the queries and needing to use some as subqueries for additional information, I imputed missing values with zero, since a household with no coupons sent to them would have an coupon redemption rate of 0%, but the way I was calculating this and other features resulted in a value of NA. This proved to be an effective strategy although it did create some issues with categorical information that went unnoticed at first.

One of the first problems I encountered with the data was a large number of outliers across multiple features (unusually large values). The first and most notable of these was with the 'count_products' feature which reflects the total number of *distinct* products purchased by a household over the course of the study (711 days, just under two years). Some households had purchased over 3,000 products in that time period, which would mean consuming an average of 4.5 *unique* products every day and never repeating a purchase. On its face this isn't THAT large, but is a remarkable amount of consumption, particularly if a household is also presumably eating out once a week or purchasing items repeatedly.

In addition to the 'count_products' feature there was also a large number of households who visited the store over 400 times, and more than 25% of households purchased more than 14,000 items at a single time. While these are unlikely numbers for a normal household, this isn't all that ridiculous for a corporate client. Catering services need to purchase food from somewhere, and so too do corporations who supply in-office food. Given that I didn't have access to actual stakeholders to better understand these outliers I tried to isolate those households which were consistent outliers.

In order to do so, I identified six features with distributions which were significantly skewed to the right by outliers (*average quantity of an item purchased, total count of distinct products, maximum transaction value, total number of purchases, average number of items, and max number of items*). From there I filtered out the households which were in the top 25% of each feature. I then compared those households which appeared in the top quartile of each of those features with those households in the top quartile of each of the other features. While there are 685 households in each quartile of the data set, there were only 85 households which

appeared as outliers in each of those six features. Below are the distributions of those features:




Those 85 households (referred to hereafter as the 'high-roller' group) certainly skewed the data. This group purchased an average of 1387 unique products in each basket over a 2 year period - a normal person could maybe purchase 1387 items in that time, let alone unique ones. While I could clearly drop these households, I don't have a consulting business partner to provide more context on the makeup of this grocery chain's customers and therefore can't rule out that these numbers are in fact correct, and simply reflect a different kind of customer than the name 'household implies. I ultimately decide to keep this information in the data set, and make recommendations to consult with marketing and sales teams to identify these types of customers who make inordinately large purchases.

The other major issue that appeared during data wrangling is that of the 2,500 households for which I have data, 1,699 of them were missing demographic information. No information on marital status, income, age group, etc, was provided for these households. This makes the vast majority of my data unsuitable for customer segmentation, where customer groups are usually defined by a mixture of measurable behavior and also their demographic characteristics. While I could simply use the entire data set for segmentation, invariably the missing demographic data would create false segments. I could drop all demographic data, however that prevents any

conclusions around segment behavior from being drawn which can be extrapolated to new customers for which demographic information CAN be gleaned, but transaction behavior can only be discovered over time.

There exists the option of imputing all the missing information for the demographic features of each household. The best option would be to impute the missing values to match the distribution of the households for which I was provided demographic data. However, in doing so I would need to have more information. It is somewhat unreliable to impute 1,699 values based on the distribution of 801. Rather than reduce the validity of any findings I chose to use the entire data set and drop demographic features when examining churn analysis, and drop all households for which there is no customer demographic information and retain demographic features to use with clustering.

**Storytelling & inferential statistics.**

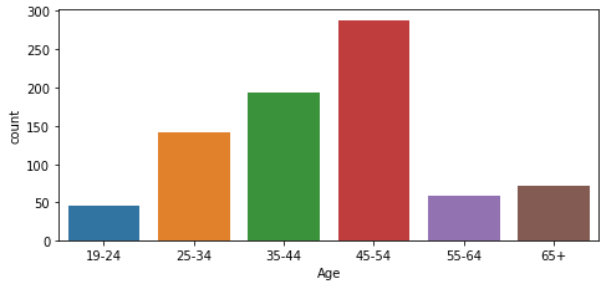
After determining the approach I would take for both the clustering and churn analysis portions of this project I started to examine several questions I had regarding the data that could help to inform my approach to both churn and customer segmentation. These questions can be found below All details can be found in my exploratory data analysis notebook[1], and in this report I comment on a subset of what is implemented there.

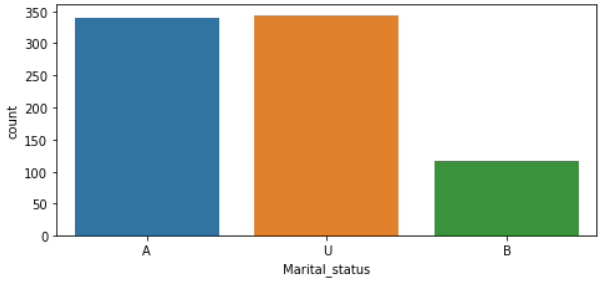
- Number of households in each of the demographic categories (Age, Marital Status, Income, Homeowner status, Household Comp/Size, Kids)
- Create a correlation heatmap of non-categorical features - which exhibit strong positive or negative correlation? (nothing significant was found and this was eliminated from the final report)
- How does the income range affect monetary value to the store, and the average sales value of each basket?
- How does having children change spending habits? More frequency and lower sales value?
- What does the relationship between % of coupons redeemed and average sales value and monetary value look like?
- What does the percent of TypeA campaigns look like in relation to the % of coupons redeemed?
- How do private brand shoppers spend money differently? Higher or lower average sales value/frequency?
- How does the average shopping lag relate to the average basket value?
- How does % of purchased on display in store relate to % of purchases on display in the mailers?
- Is there a significant difference between the % of purchases on display vs % of baskets with an item on display?

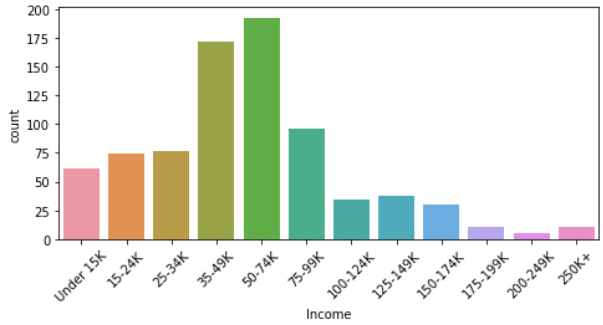
The first question regarding clustering is really one of understanding what demographic patterns or trends exist within this sample of households, and which of these features interact with transactional or campaign response features. At a glance, it's obvious that the customer

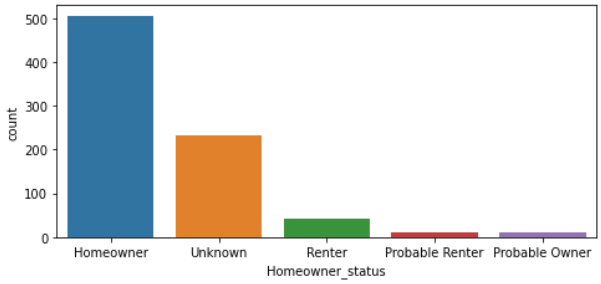
[1] Customer Segmentation and Churn: Exploratory Data Analysis

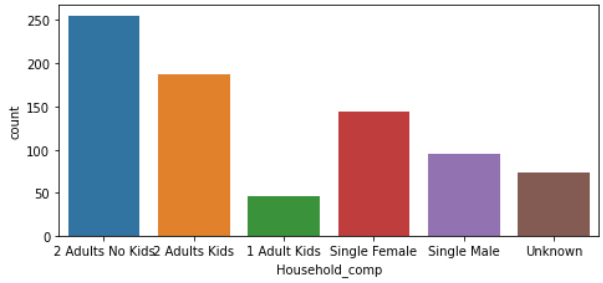
base that was studied is largely between the ages of 35 and 55, earns a low-middle class income, and owns their own home. However, there is a lot of information that is not available; over 1600 households were missing any demographic information, while unknown is still a large category for much of the demographic data I did have access to. With this data 'unknown' is the largest or second largest group of households for 'marriage status', 'homeowner status', and 'child-raising status'. You can see the details below, in the context of marital status 'A' indicates married, 'B' indicates single, and 'U' indicates unknown.

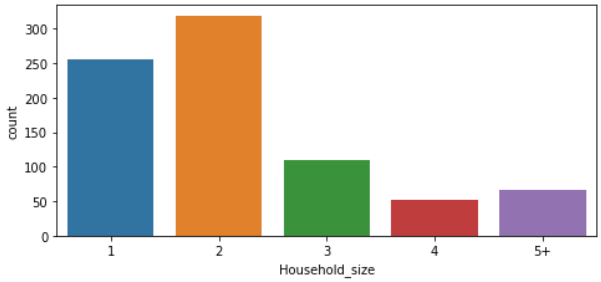


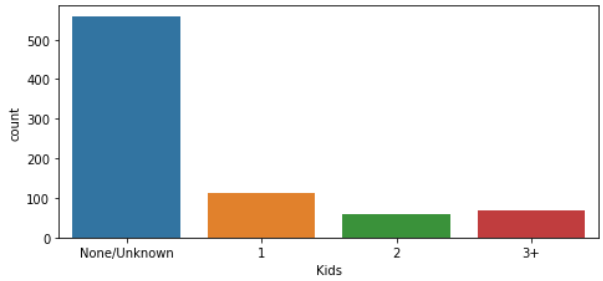



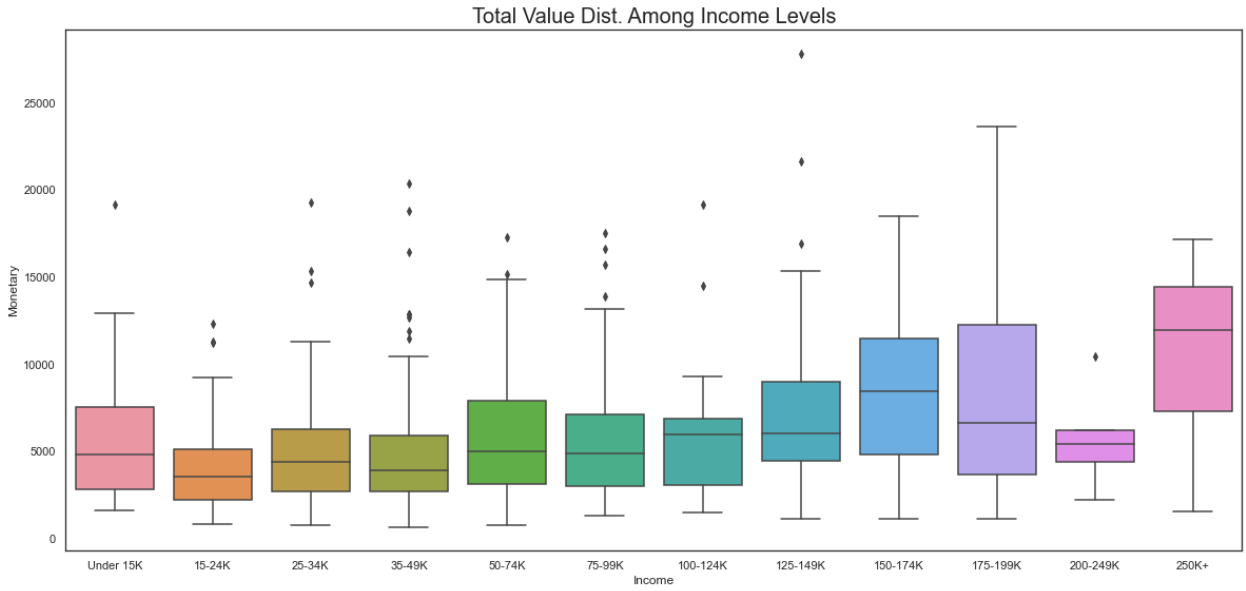



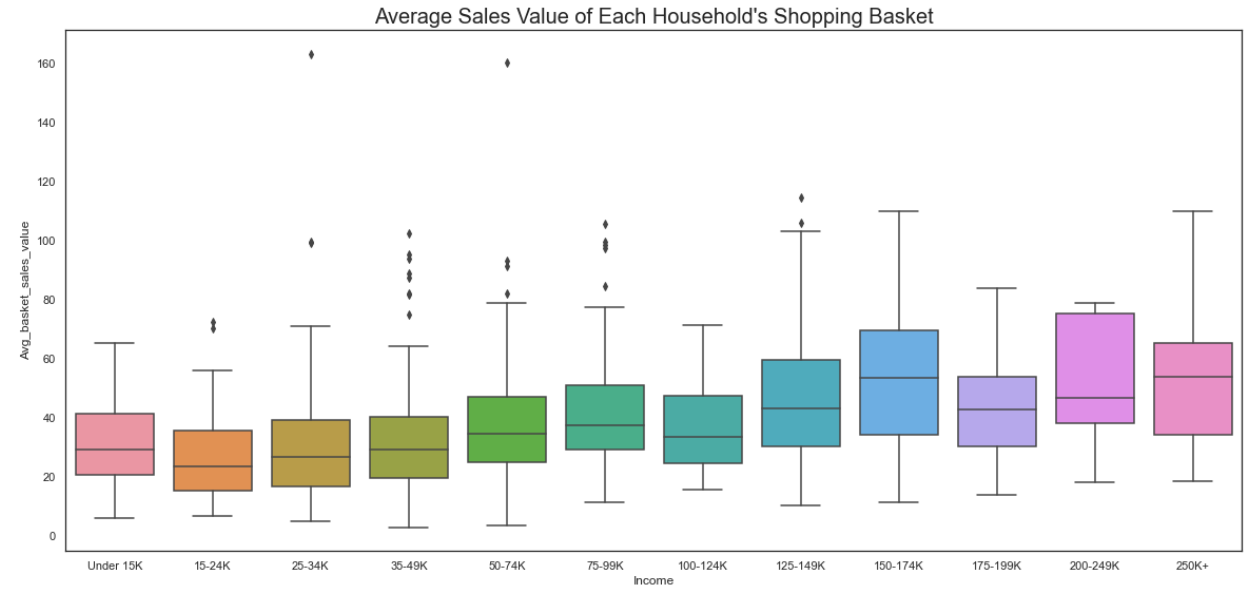




First and most obviously, I wanted to explore how income level affected the overall value of a household to the grocery chain. Given that both grocery stores and income brackets tend to be geographically clustered this data set offers a unique view into how different income levels

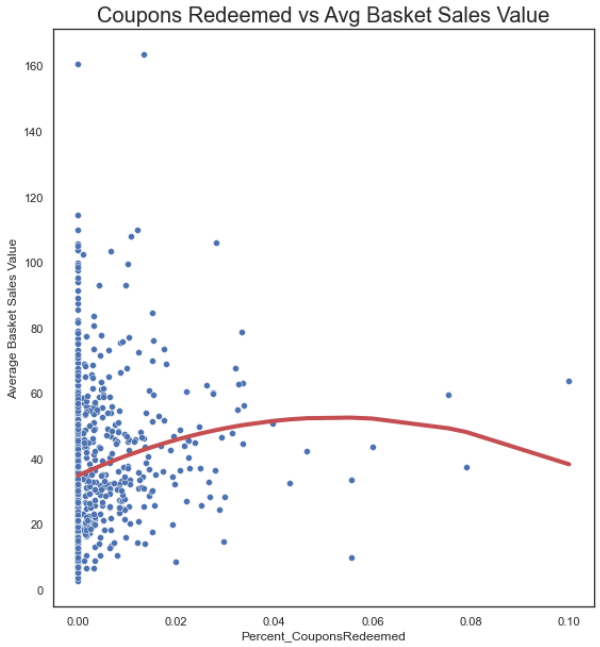
spend money at a grocer. Below I've examined the relationship of both average and total sales value of a household as it relates to income brackets:

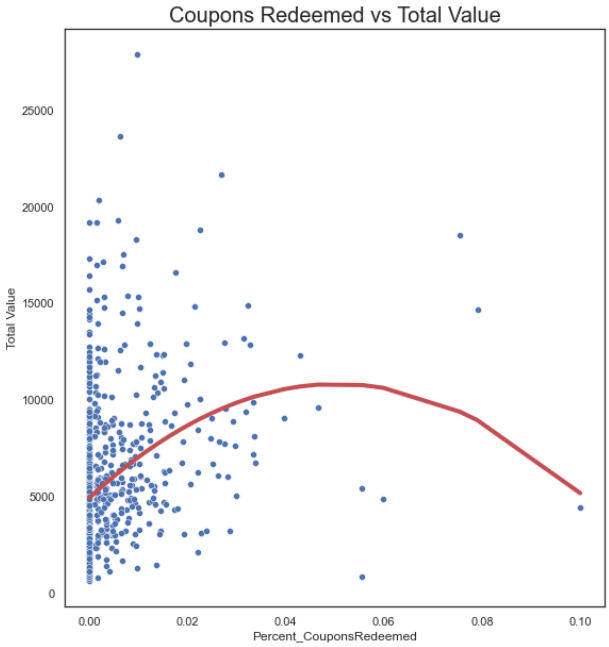



It is interesting to note that the total value of customers seems to rise steadily with income, however the increase in total value with the highest income level is much less pronounced than the increase with the average sales value. This could be the result of higher income levels having more disposable income to dine out, and thus preparing food at home less and shopping less overall. It should also be noted that there is a significant drop in total value from households earning 175-199K to 200-249K, however the general upward trend recovers with the next income group.

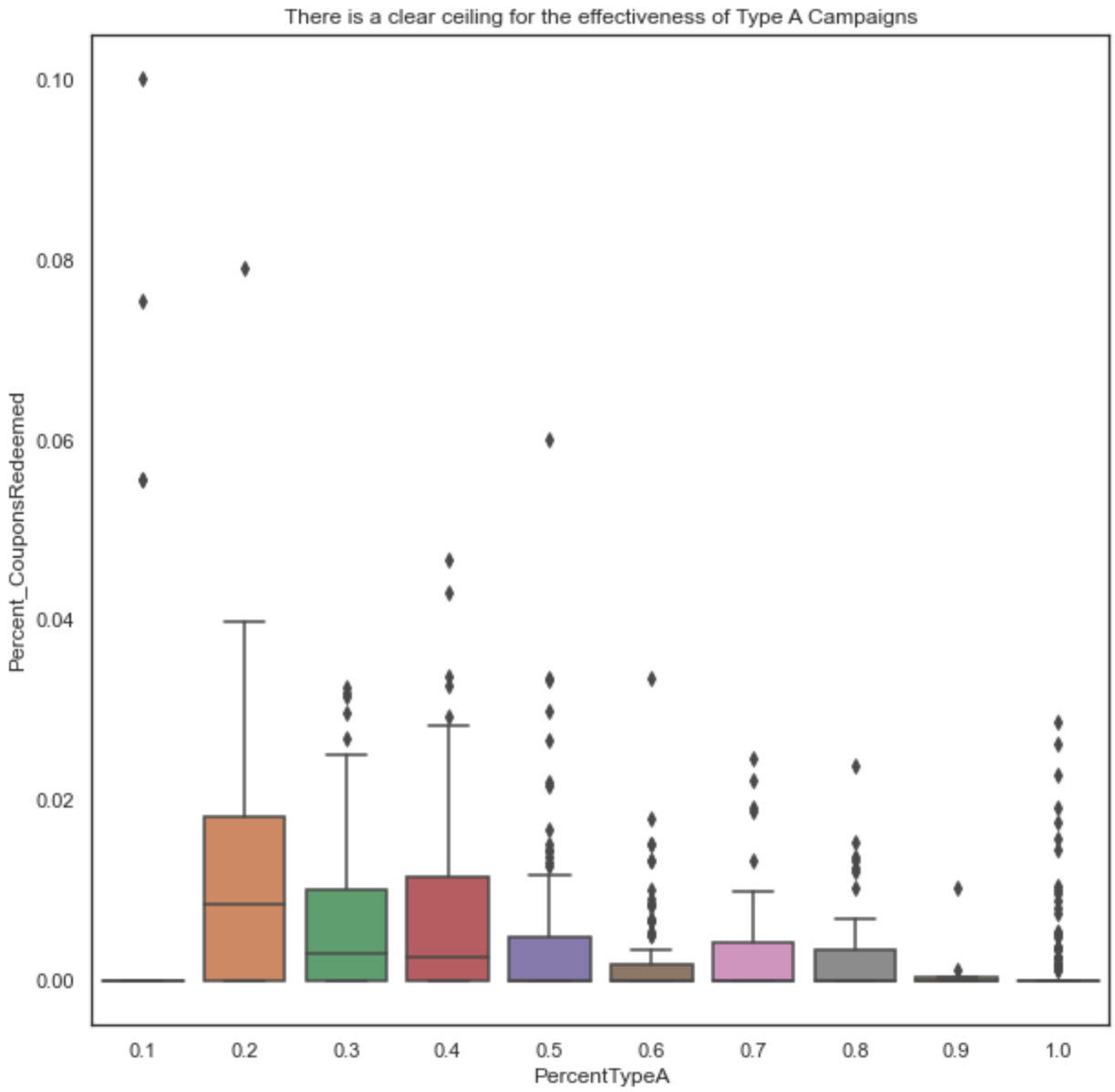
This trend is similar to how child-raising status affects a household's value to the grocery chain; as the number of children in a home increases, so does the value of that household increase. This is rather intuitive given that with more people to feed a household will need to purchase a higher volume of goods. The first non-intuitive insight came after examining household value with regards to coupon redemption rates.




Surprisingly the average basket value and the total value of a household are not negatively correlated with the percent of coupons a household redeems. In fact, the value seems to climb up to a certain point at around 5 percent coupon redemption, at which point it begins to fall as households redeem a higher percentage of coupons. While this trend is more pronounced with a household's total value, it is consistent for the average sales value as well. This would indicate that driving campaigns that have coupon redemption rates of around 5% are best for driving grocery store value.

This then warranted the question - which campaigns sent to households are more effective? This study sent out three types of campaigns. Type A campaigns are targeted advertising - sending the household a tailored set of 16 coupons based on their previous purchasing behavior. Type B and C campaigns are 'shotgun' advertising - sending the entire book of available coupons to a household. I previously decided to examine if campaigns were redeemed at different rates depending on what portion of a household's advertising was targeted vs shotgun, however another valuable question to study further would be which campaigns have a coupon redemption rate near 5%.

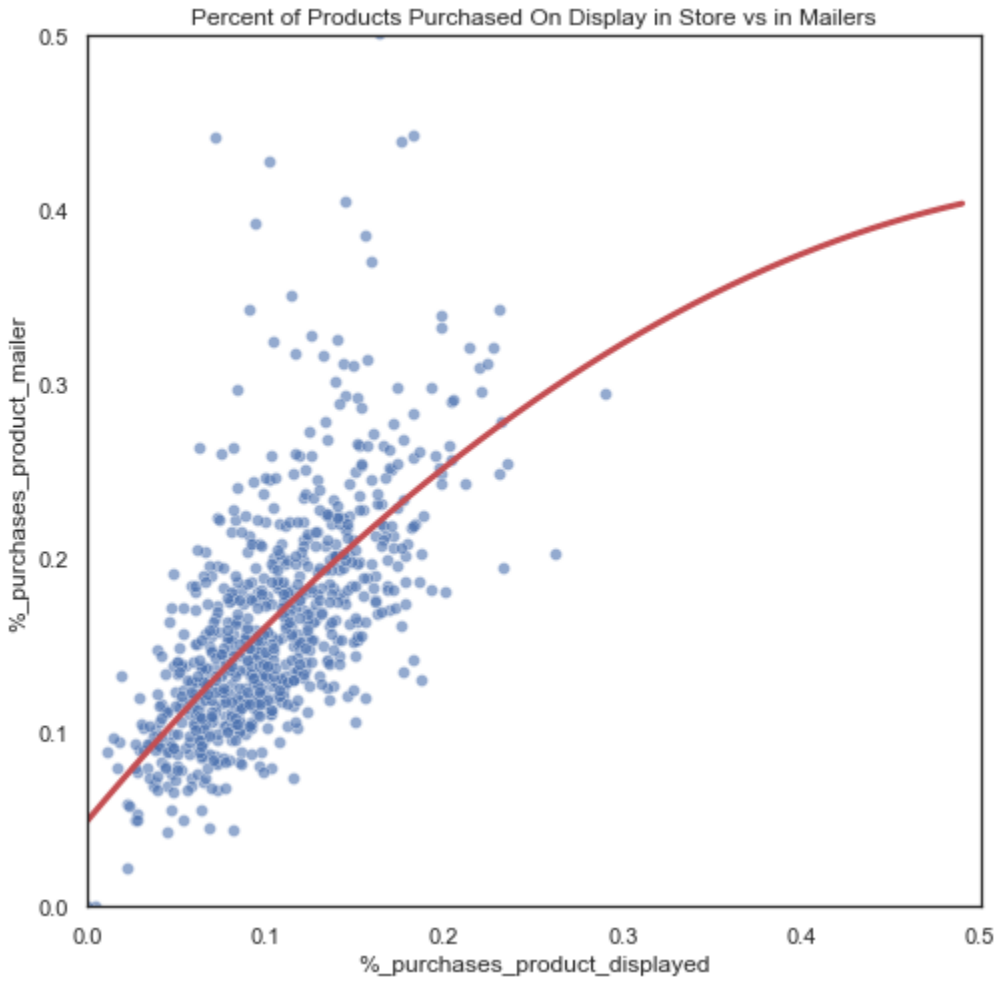



While for the majority of households the percent of campaigns sent to them which are Type A is not impactful on the percent of coupons sent which they redeem; it should be noted that no household redeemed more than 4 percent of their coupons unless Type A campaigns were less than 50% of the campaigns sent to them. This could indicate that the 'targeted' mail advertising campaigns which are Type A are effective, but only up to a certain point. It would be interesting to experiment with these campaigns to see if tailoring coupons in targeted advertising based on what we predict someone will want to try is more effective; however this is outside the scope of this project.

One hypothesis that I was forced to reject was that households who purchased from private brands more frequently would spend more on average. This was posited based purely on conjecture and the relative price difference between many national brands (eg: pepperidge

farm cookies) and their private counterparts (in this case a smaller local bakery). The correlation coefficient between the percentage of households' products purchased from private brands and the total value or frequency with which households' purchase from this grocery chain was small, and thus there is no significant relationship.

Much of the data in this study is focused on the efficacy of mailed advertising campaigns and coupons to households. However I am also aware from previous experience working with retail alcohol sales teams that 'end cap' space (space in a grocery store where products might be prominently displayed, typically the end of a row of shelves) is considered prime real estate for driving sales. I chose to explore this belief, and compare the efficacy of displaying products in store vs displaying them in a mailed advertisement.



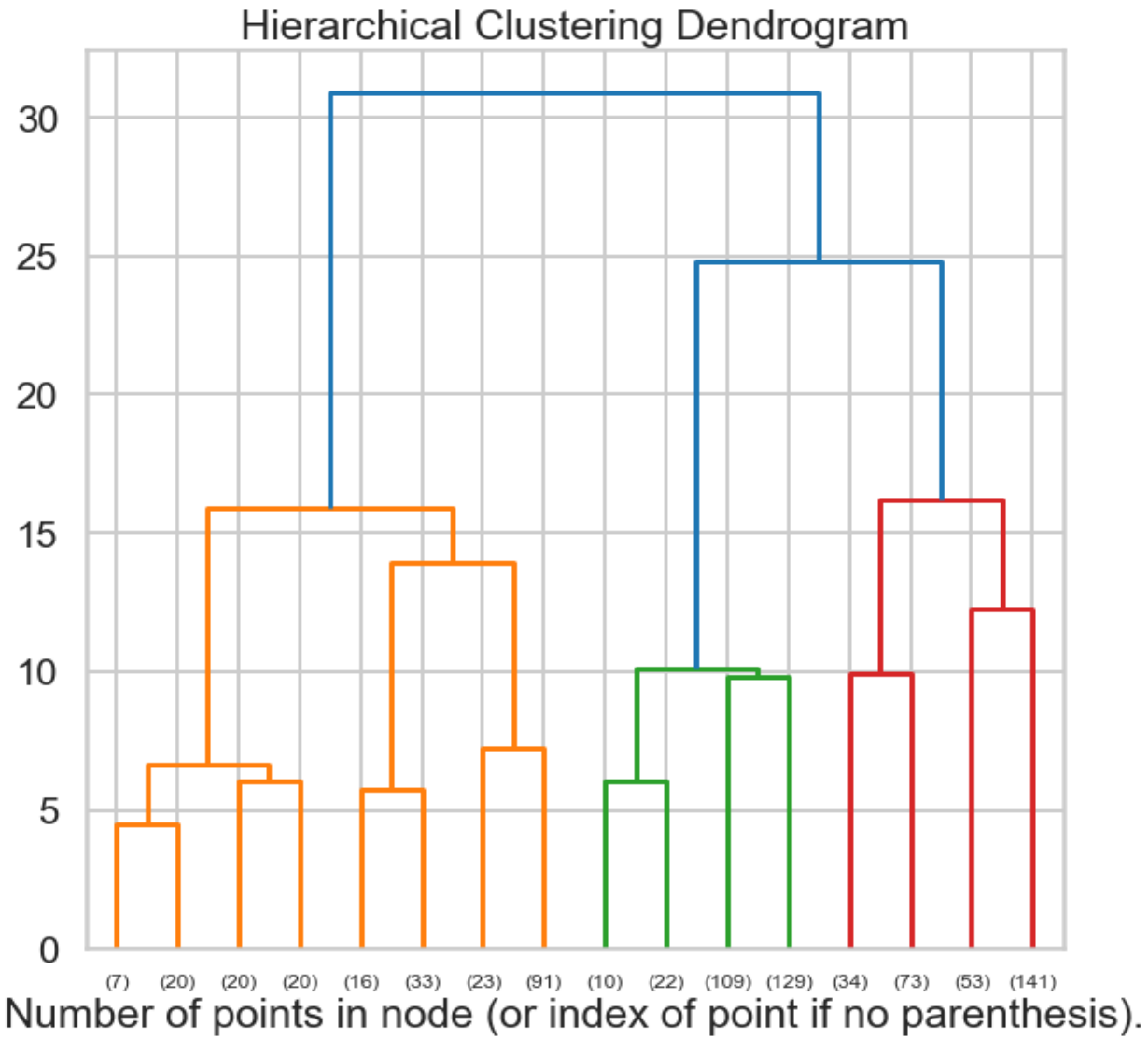

Rather surprisingly, there are more products purchased which are on display in the mailers sent to households than those which are displayed in the store. This is something that could also be examined in more detail with different segments, and should be compared to the percent of total products displayed in both the mailer and store as well. This trend is slight enough that if the ratio of products that are actually displayed in either place in a given week is

taken into account it may eliminate the trend. However, an examination in that level of detail is outside the scope of this project.

Finally, the most important finding from this exploratory analysis is that the average discount from specifically loyalty cards (rather than mailed coupons) is strongly correlated (correlation coefficient is greater than 0.75) with the average sales value of a basket. While this doesn't provide significant insight into how to approach clustering for customer segmentation, this is important to note since further exploration could examine this relationship in more detail and identify opportunities to grow revenue. If a given segment shops frequently but has low loyalty membership, loyalty programs could be offered to incentivize membership, and increased spending from those segments. This data set does **not** contain information on loyalty program membership, only loyalty program discounts. I'm able to infer membership in loyalty programs based on this, but unable to directly measure how loyalty membership impacts spending behavior. This is the final recommendation for further exploration or follow-on work before I began to craft customer segments.

**Customer segmentation.**

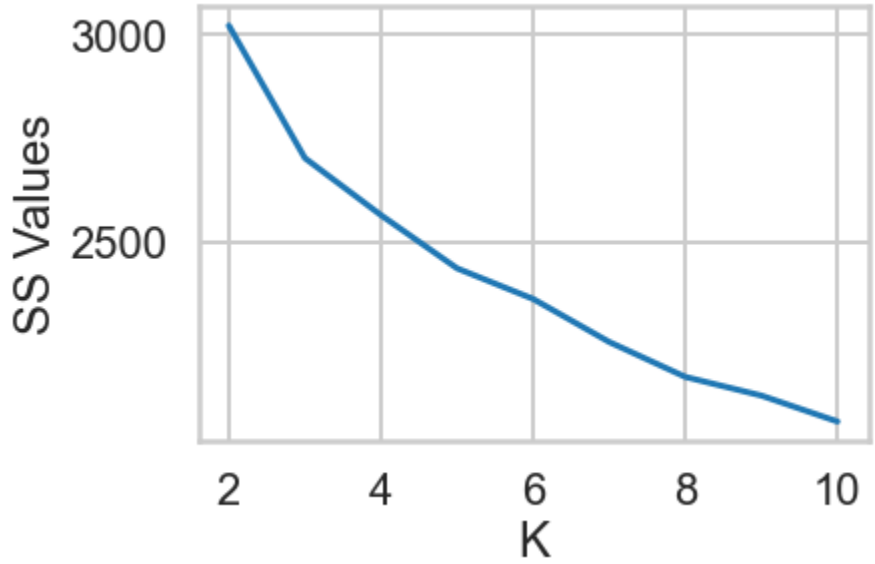
In order to approach customer segmentation in a way that eliminated as many (potentially erroneous) assumptions that I brought to the table I chose to examine the clusters created by both the k-means and agglomerative (also called hierarchical) algorithms. Using the k-means method I chose to use the elbow method to identify the ideal number for 'k', and with hierarchical clustering I simply plotted a dendrogram and examined this to identify the likely number of clusters. You can see the dendrogram on the next page:




Now, there are some instances where I could choose a large number of clusters such as 17 (the bottom level of the dendrogram). This would provide small clusters about which I could become very specific. However this is hugely impractical for customer segmentation, where the goal is to categorize customers into broad 'personas' which can be used to wield marketing and sales tools more effectively and to tailor company products for defined customer categories. Seventeen different marketing approaches is far too many to manage, however a smaller number such as three (denoted in the above plot) is much easier to both digest and understand.

In addition to selecting the proper number of clusters for agglomerative modelling, I also tried four different options for node linkage. Both single and average linkage resulted in a single cluster; while ward and complete provided similar clusters. The difference between clusters using the ward linkage method is slightly more well-defined and is what I'll use for the purpose of analysis.

When it comes to using the k-means method things are a bit more complicated. Unlike with agglomerative clustering the algorithm isn't mapping the entire data set into a single cluster, and identifying the multiple sub-clusters that make it up. Instead, the algorithm is randomly creating a 'k' number of centroids, and calculating the clusters that would orbit those centroids. This means that I need to test each instance of 'k' to assess which number of k is identifying the best structure in the data. You can see the plot of the sum of squares (or distance from points in a cluster to it's centroid) below. We know that as I create more clusters this number will fall, what I'm looking for is an 'elbow' in this plot which will allow me to reduce the sum of squares significantly more for that value of 'k' than for higher values:

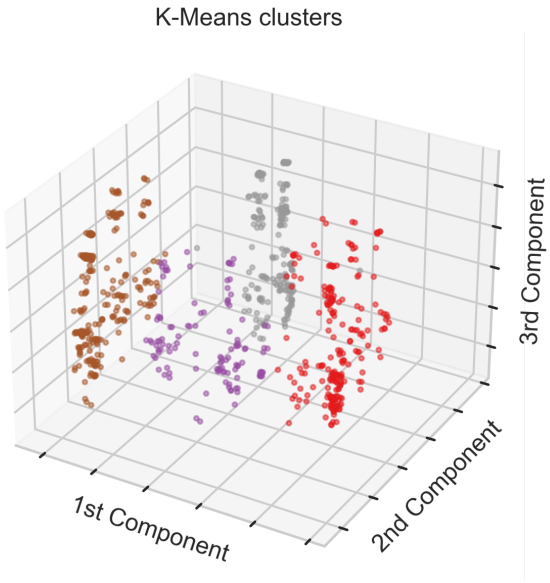


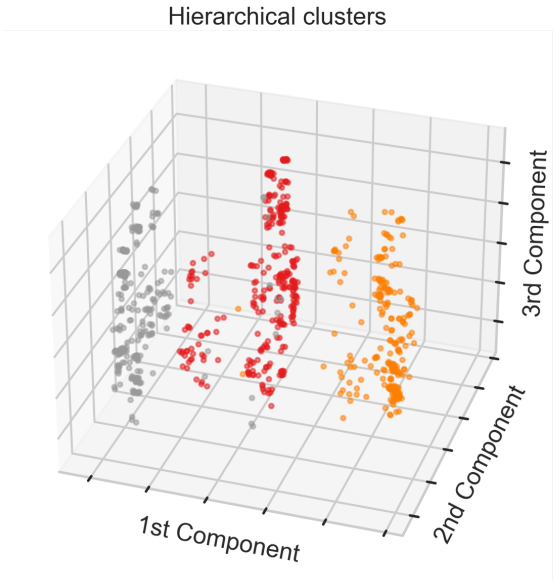

You can see here that there is a clear elbow at 'k' = 3 and 5, with another elbow at 'k' = 8. Eight clusters is far too unwieldy for business usage, so I compared graphically the structure of clusters created when the value of k varied from 3 to 6.

After iterating over various numbers of clusters, passing 6 cluster centers to the K-means algorithm provided the most distinct demographic clusters while also showing the largest degree of difference in transaction behavior. This structure is not recreated when passing 4 clusters to the K-means algorithm, but rather blends all the clusters together so that there is little difference between them. However, both clusters 0 and 3, as well as 4 and 5, showed remarkable similarity in terms of both demographic makeup and transaction behavior so I was able to combine those clusters for the purpose of customer segment analysis.

Moving forward with 4 segments created by the k-means algorithm, and 3 created by agglomerative clustering, I needed to visualize both sets of clusters in order to assess which segments would represent the data in the best way, and decide which clusters would be used for customer segmentation.

The best way to visualize these clusters when I'm working with such a large number of features is to pass the data through a Principal Component Analysis (PCA) algorithm to y reduce the dimensionality of the data to 2 or 3 dimensions where the clusters can be visualized. In doing so, I identified 3 components as the best way to represent the nuance of the data while also making the structure of these clusters much more evident. Those visualizations are found below:



Visually, you can see that the agglomerative (hierarchical) algorithm has uncovered a much more distinct structure than the k-means model. While the k-means clusters seems to have identified the 'bleeding' effect between two of the agglomerative clusters (notice the large gap in the center of the red cluster), overall the structure between the clusters is much less distinct than it is with the hierarchical approach, and these are the clusters I'll use and recommend to stakeholders for use in customer segmentation. The description of those clusters can be found below:

**A0**: (Red cluster in above plot) This is the largest cluster that the Agglomerative model produced. It's overwhelmingly made up of single-person households which are poorer on average than those belonging to another cluster. This cluster also has the largest group of households between the ages of 45 and 54 years old, almost half of the households in this cluster belong to that group. This cluster is the most recent purchasers, who typically have the smallest average sales value, smallest discounts, and the smallest basket size. "Frequent low volume customers"

**A1**: (Grey cluster in above plot)This is the smallest cluster, and also is the youngest cluster with most of its households made up of people between the age of 25 and 44. This cluster is also the only cluster to be made up of households with children, and not a single member has fewer than 3 members. Finally, this is the wealthiest cluster on average. A1 has the largest average discount (coupon, loyalty, or rebate), as well as the most coupons redeemed and the highest average sales value as well as highest total value. "High value, high discount"

**A2**: (Orange cluster in above plot)The final cluster is exclusively made up of 2-person households with no children. Generally speaking this cluster is older on average than the other two, with the largest number of households over the age of 65. Additionally, this cluster is generally poorer on average than cluster A1, however it has the largest number of households earning over 150k. This cluster is the one which has shopped the least recently, but has the second highest loyalty discount. Generally speaking it seems to follow the behavior trends of A0, but with less magnitude (frequent but low value discounts). "High interval shoppers"

**Churn analysis**

The biggest challenge when using predictive machine learning to study churn with an organization is defining what constitutes churn . My instinct was to define churn as having gone more than a month without purchasing from one of the multiple grocery store locations. Generally speaking this is how I shop - once a week to a grocery store, but there are three options in my area. Sometimes I'll go a few months without visiting one of them but will go back when I need a specific item and then re-enter them into the 'rotation'. However, this is how I shop and not necessarily indicative of how others will shop.

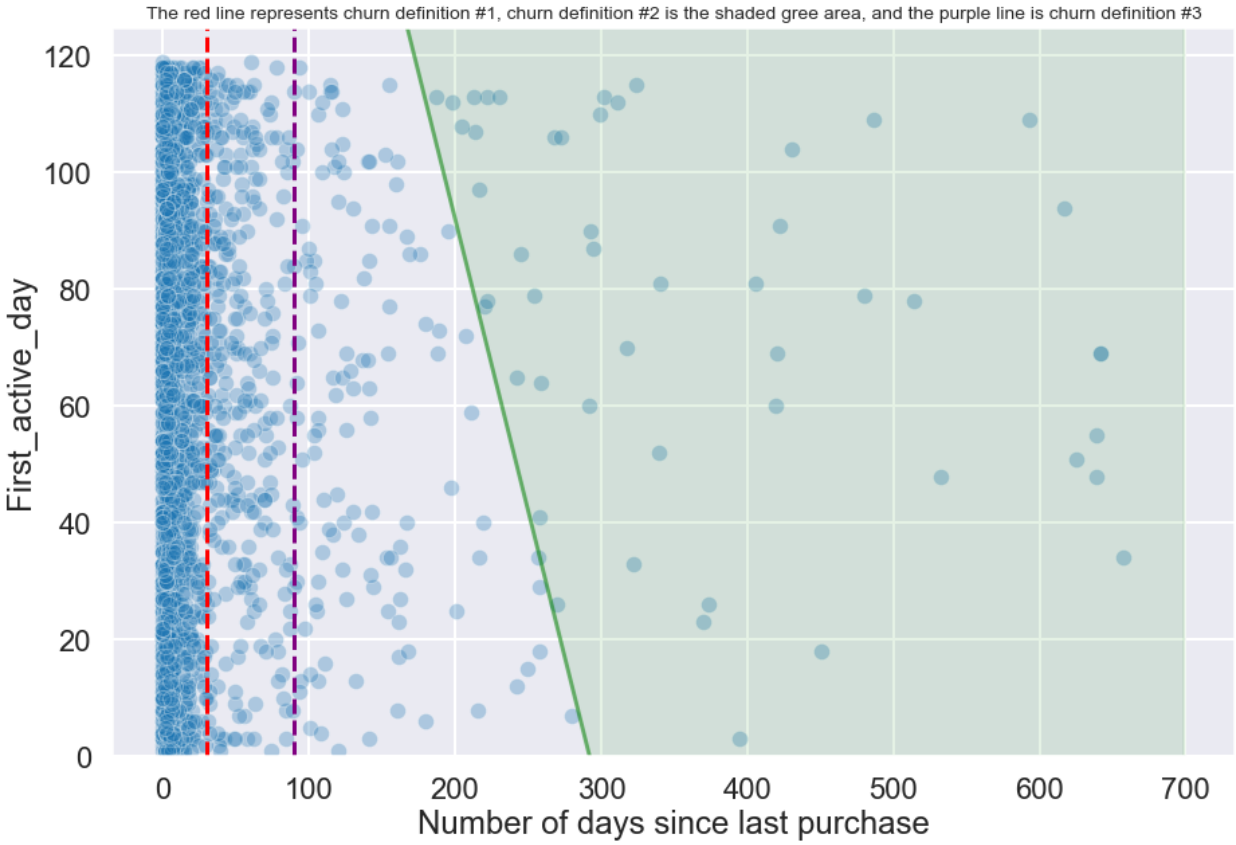
Ideally I could speak to someone at this store to assess how long produce and other essential purchases from a grocery store will last, however that's not possible so I'll take a look

and see which definition of churn will allow me to provide insights that are both sensitive to not over-defining churn, but also can provide a significant impact to revenue. The typical person will have to shop at least twice a month - fresh produce won't keep longer than 2-3 weeks.

However there are several other options which can be considered and then evaluated visually:

1. Churn is allowing 30 days (1 month) to pass without visiting one of the grocery stores (This infers that a customer has shopped 2 times at a different grocery chain without returning to the chain being examined).
2. Churn is no longer shopping from this brand of grocery chain 1 year after the first active day (This limits churn to those who are never 'resurrected').
3. Churn is allowing 90 days to pass without visiting one of the grocery stores (This infers that a customer shopped 6 times at a different chain, and would capture the behavior which I personally exhibit).
4. Some other definition of churn based on patterns noticed in the data.

Plot showing the number of days since a customer shopped relative to the day they first shopped




It would seem that most customers have shopped at this grocery store chain at least 1 year after their initial purchase from the store, perhaps this could be a good opportunity to examine churn since these customers clearly haven't been revived. However, addressing the

cause for these customer's churn won't help the bottom line of the store since it's such a small percentage of overall households.

What would be better is to focus on what will capture insights that will drive customers to purchase more frequently. In order to do so I'll look at 90-day churn since these customers could conceivably return but have also purchased from another store multiple times without returning to the grocery store being examined. In this project, I use Definition #3 from above, and at the conclusion of this study we found that there are 6.8% of households which can be considered churned.

Prior to testing out different classification models in order to study churn, I need to alter the data set that I'm going to work with to better reflect the churn problem. To this end, I used the entire data set of 2,500 households and dropped the categorical demographic features (these features only had information for 801 households). I also dropped other categorical features that I engineered such as most frequent product purchases, and most frequent store purchases. Retaining these features *could* prove useful, however it will dramatically increase the dimensionality of the data set used with the supervised learning algorithms, and I have no specific business questions that have been asked regarding those.

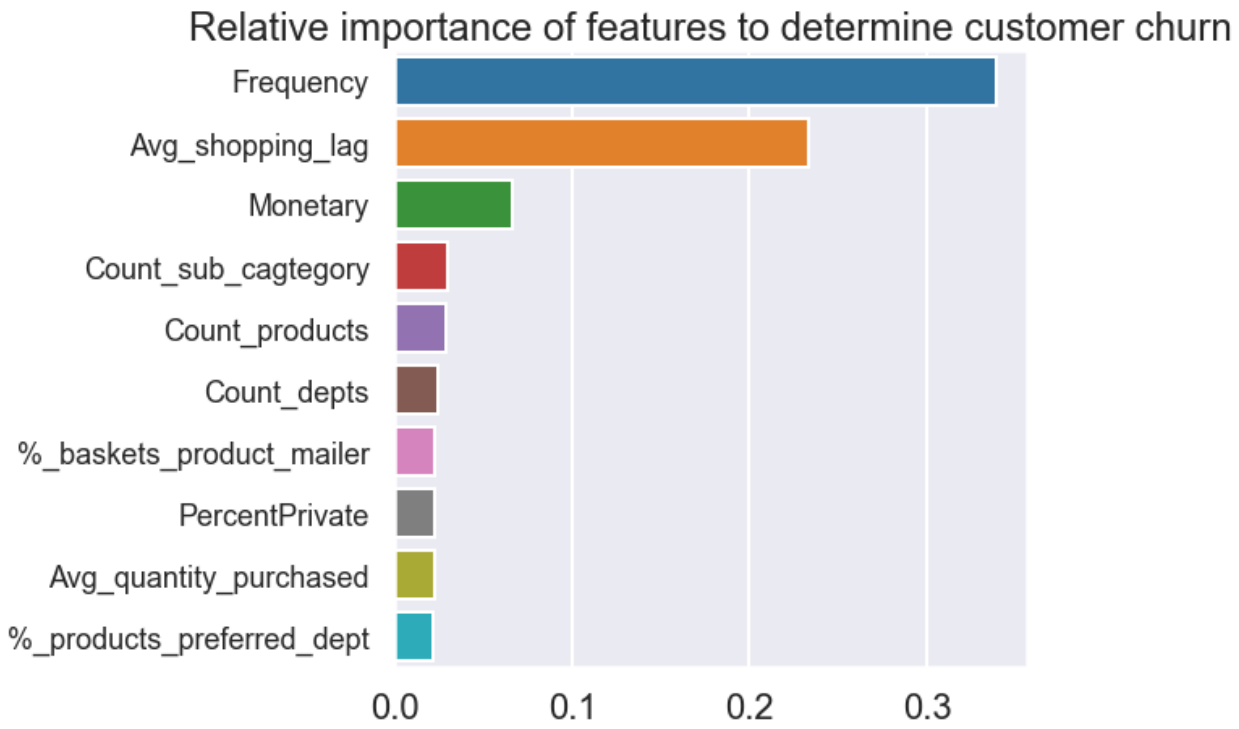
Now, before I began to select the best classification model to use to study churn, I needed to select which metric I would use to evaluate these models. I could use overall accuracy, which simply measures the % of predictions the model makes which are true. However I'm working with a heavily skewed data set - only 6.8% of all households are considered churned. Similarly, I could also measure precision, which measures the % of churn predictions which are *actually* churned.

This is closer to what I'm interested in, however there isn't a real consequence to predicting a household as churned when it is in fact retained - maybe they'll get an extra coupon. Recall is the metric I chose to work with. This measures the % of actual churned households which were predicted as churned. Using this metric to select models will mean that I'm selecting the classification model that is identifying the *most* churned households, which will give me the ability to generalize any findings on churn with more confidence. Of course, there is a relationship between precision and recall, and a high recall can come at the expense of a low precision.

After deciding to use recall as the metric to select the best performing model, I used a dummy classifier to create a baseline sense of what these metrics would look like if I used a model that had *no* predictive power. A dummy classifier simply predicts the class which the previous observation belonged to - in this instance retained - for every observation. This results in a recall score for churn of 0 since no households were predicted as churn. Another useful metric for evaluating these models is the 'macro average' which takes the accuracy score for each class and creates a composite score that accounts for the class imbalance. For a dummy classifier this macro average is 0.47 - extremely poor.

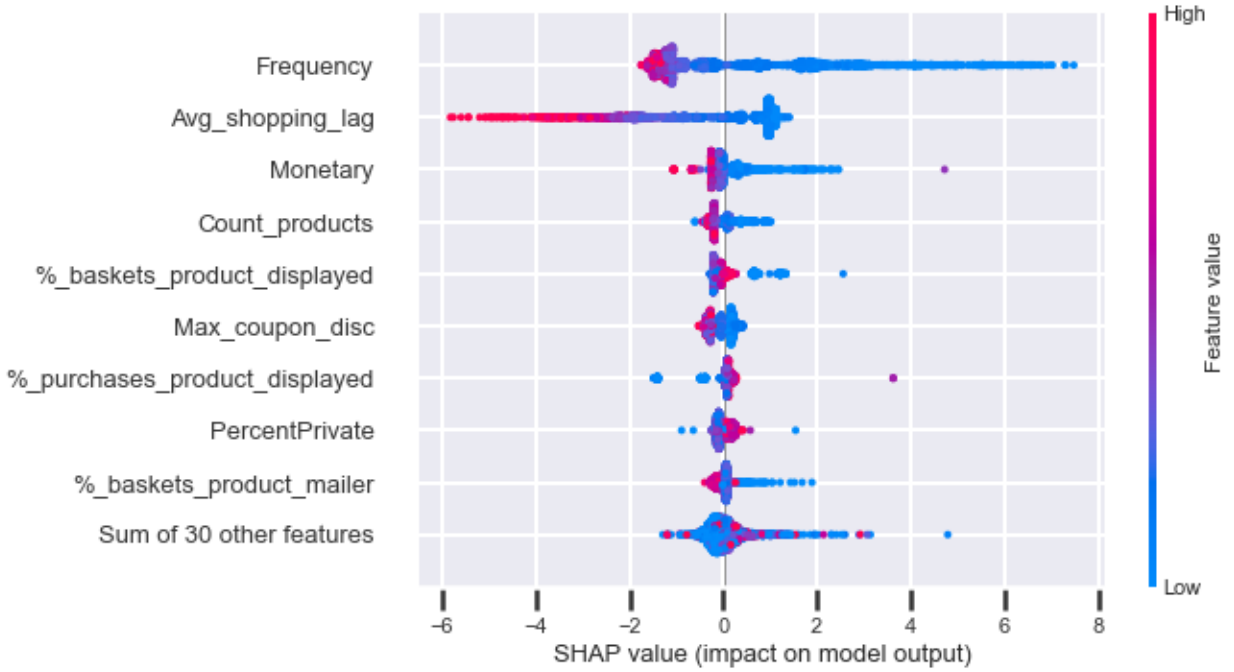
I then iterated over multiple classification models, trying a linear support-vector classifier, support-vector classifier (using the rbf kernel), k-nearest neighbors, random forest, and gradient boosting models. Only the ensemble methods and linear support-vector models predicted the churn class at all, and only the gradient boosting model provided a recall score above 0.4 and had a macro accuracy average of 0.87 - quite good.

This means that I'll use the feature importances from this model to inform recommendations for stakeholders on which factors determine a household as churned. By far, frequency (the number of shopping trips a household took over the course of the study) and average shopping lag (the average number of days between those trips) were the two most important factors for predicting churn.



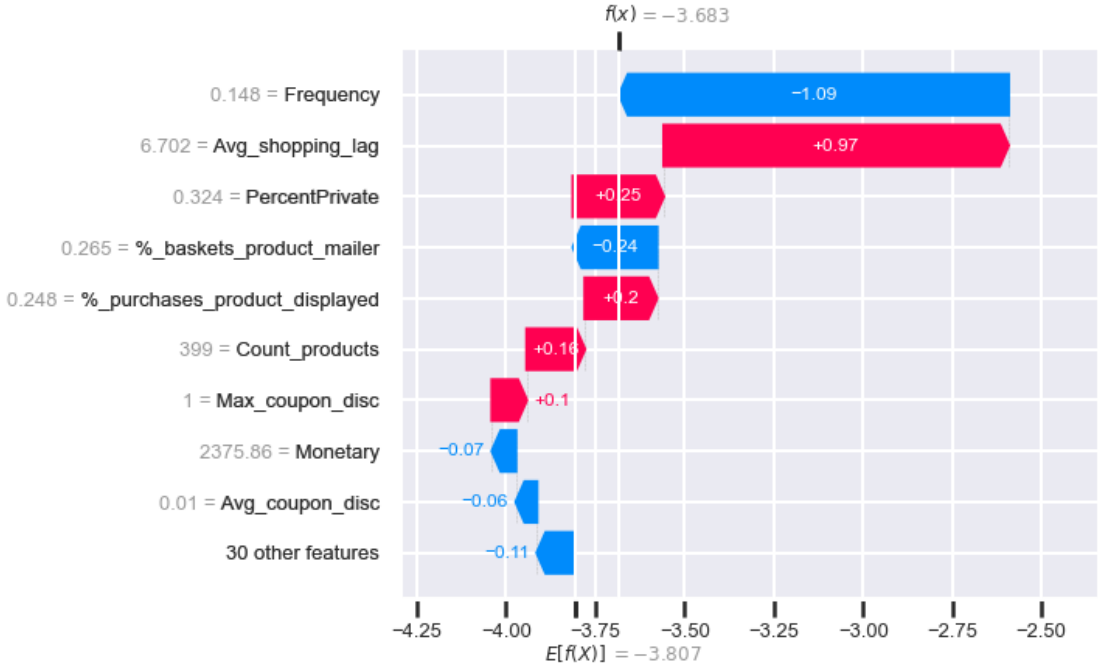

Rather interestingly the count of unique products and sub-categories and departments were also important factors to determining if a household has churned. This is related to the percentage of products purchased from a households preferred department (defined as the department a household shops from the most. Incidentally this was the 'Grocery' department for all households and could be read as the percentage of products purchased from the grocery department. Finally, the percent of transactions in which a product purchased was advertised in the mailer, and the percent of products purchased from a private brand were all indicative of churn to some degree.

These insights from the built- in sci-kit learn functionality are useful, however I decided to examine the features more closely with SHAPley values in order to better understand and quantify how these features all interacted. You can see the SHAP summary plot on the next page.




In this plot, negative SHAP values indicate influencing the model to predict a household as retained (or reducing the probability of a household being churned). Using this along with the color bar on the right we can see that lower than average frequencies strongly indicate that a household is churned, while higher than average shopping lags indicate that a household is retained. SHAP awarded features a different set of importances than the feature_importances_ method in sci-kit learn, which resulted in a different set of features being provided as most important. However we can see that having a low number of unique products purchased still influences a model to predict users as churned, and that a large number of purchases where a product was advertised in a mailer indicate a more ‘sticky’ customer.

I took this one step further and decided to examine a household which was correctly predicted as churned and to see how the SHAP values played out with this household.




Using this waterfall plot in conjunction with the summary plot we can see the individual SHAP values for each feature from among the 10 most impactful features for this household. The x-values on this plot are *log-odds*, where the probability of a household churning (a value from 0-1) has been logarithmically transformed. This means that an x-value of 0 is equal to a probability of 0.5 that a household is churned, negative values indicate probability less than 0.5, and positive represent probability greater than 0.5. Similar to the previous summary plot we can see that the average shopping lag was by far the most impactful feature that indicated this household would be 'churned'. This was closely followed by the percent of private products they purchased, the number of unique products purchased, and the percent of products purchased which were displayed. Given that the expected probability for the whole dataset is -3.807, or a 0.0217 probability that a household is churned, while and the actual output for this household is -3.683 (a probability of 0.0245 that *this* household is churned), small changes to these variables would dramatically reduce the probability of this household being churned. If the probability that a household is churned is greater than the expected probability for the entire dataset, the model predicts that household as churned.

To this end, *in this particular case*, if this household were to purchase *more* national brands, or a larger number of unique products we could easily have reduced their probability of churning. Note, that I did not make any recommendations around the percent of products purchased that are advertised. The behavior is extremely similar and the two forms of advertising seem to cancel one another out here - rather than attempt to tweak this I think it's far better to use this type of influence to alter other features for this household (perhaps advertising

more unique products or products which will necessitate a higher shopping lag such as a 24 pack of paper towels instead of a 16 pack).

While the above would be what needs to change to reduce the likelihood of churn for that particular household, what stakeholders are most interested in is assessing how to reduce churn across all households. Using the SHAP summary plot and feature importances from sci-kit learn we can tell that different measures of the count of unique products a household purchases are important. Leveraging the marketing methods stakeholders have to push customers to try new products would likely be a very effective method for reducing churn. Similarly, we can see that households with a high shopping lag (more than 5.3 days between shopping trips) are much less likely to churn. Pushing 'staple' products which will last for a longer number of days would be the best way to influence shopping lag for households in order to reduce churn as well.

## 3. Conclusions and Future Work

Using unsupervised learning methods I was able to create distinct customer segments for use, and create a classification model that presented strong recall for predicting the churn class. While the classification models for churn are a good starting point, these can be iterated on with feature engineering in order to coax better accuracy out of the models and improve upon the recommendations which can be made for stakeholders. I would recommend the following as next steps to continue to provide strong results:

- First, I would conduct an ANOVA analysis of the features in this data set in order to test for collinearity. Particularly with regards to the number of products a household has purchased, there seems to be many features which are redundant and could be eliminated. While this was acceptable for clustering, it doubtless has reduced the accuracy of predicting churn and the complexity of the model can be reduced by taking this simple step.

- After eliminating redundant features, I would approach clustering again using the same parameters with the agglomerative model and re-establish membership to the different customer segments. Quick analysis should be done here to ensure the overall structure of the clusters remains the same after eliminating redundant features.

- Finally, I would conduct churn analysis *with resampling* for each customer segment. This will allow me to both measure the minority class much more accurately, as well as provide actionable and targeted recommendations for each customer segment. Since the number of unique products seems to be heavily influential on churn, understanding how each segment responds to marketing would be valuable for influencing that purchase behavior

## 4. Recommendations to Stakeholders

As far as immediate recommendations go, I have five recommendations that can be acted upon without further study or evaluation. These recommendations are **not** based on casual relationships, but rather are intuitive actions that can be taken to alter behavior that the churn model indicates heavily influence whether or not a customer will churn. Thus these recommendations should be validated with A/B testing where possible:

- Firstly, provide more and larger coupons in the mailing advertisements. Higher maximum coupon discounts are associated with reduced churn probability, as well as a higher frequency of purchasing products that were advertised in the mailer.

- Secondly, increase the size of 'utility' items sold in the store in order to increase the average shopping lag (one of the best predictors of churn) of customers. For instance, selling a 24 ct pack of paper towels rather than a 16 ct pack. These kinds of utility items can influence households to shop with a larger period of time in between trips and are products that are simply convenient to purchase anywhere. If a customer needs toilet paper they're going to purchase it regardless of the amount you sell.

- Find ways to increase loyalty program participation, since an increased loyalty discount is heavily correlated (correlation coefficient of 0.82) with increased average sales value for a household. Strategies used by other grocery chains for this purpose include pricing techniques such as 'instant markdown' or visually contrasting normal vs loyalty prices.

- No household redeemed more than 4 percent of their coupons unless Type A campaigns were less than 50% of the campaigns sent to them. Type A campaigns are tailored based upon past purchasing behavior, and based on the churn analysis a *higher* number of unique products is associated with an increased 'stickiness' for a household. I recommend using Type A campaigns to offer coupons for *new* products based on past purchases, or to reduce the number of Type A campaigns overall.

- Lastly, study further the customers who purchased a high percentage of products on display. Were the products they purchased on display cheaper than other similar products they have purchased? Why is it that purchasing goods marketed in one way causes the model to predict a customer as churned, while purchasing goods marketed via mail as retained? Understanding how marketing efforts influence different households in different ways - particularly in the context of customer segmentation - will better enable targeted efforts to reduce churn.

## 5. Citations and consulted sources:

*Springboard Mentor:*
A.J. Sanchez

*Consulted Data Scientists:* Finn Qiao - Data Scientist @ DoorDash, Byron Becker - Software Engineer @ Amazon

HubSpot. (February 16, 2021) What Is Customer Churn? [Definition] [Blog Post]. Retrieved from: https://blog.hubspot.com/service/what-is-customer-churn.

Client Success. () The True Cost of Customer Churn: Part 1 [Blog Post]. Retrieved from: https://www.clientsuccess.com/blog/true-cost-customer-churn-part-1/.