A Customer Segmentation Analysis on Consumer Purchasing Behavior

DSP 562 — Nellie Dawson

Video Presentation: Dawson Presentation (Youtube)

ABSTRACT

Understanding consumer purchasing behavior can help companies make decisions that benefit both themselves and the individuals to whom they serve. Through the use of clustering methods and visualizations, I conduct customer segmentation analyses to explore trends and purchasing behavior using both K-Means and Agglomerative clustering, then incorporate dimension reduction via t-SNE on both clustering techniques. I create a variety of visualizations that compare numerical and categorical features from the data using the binary Gender segmentation and the segmentation groups formed by dimension reduction for both two and three cluster groups. Through my exploration of the cluster groupings, I found that Subscription Status and Gender had some influence on making initial segmentation groups for this dataset followed by Purchase Amount and Age. Understanding this hierarchy of influential factors can assist organizations in making decisions.

BACKGROUND

Economists have been analyzing consumer behavior for hundreds of years, but with the introduction of technologies and an ever growing amount of data, we are able to understand so much more about consumers than ever before. Analyzing consumer behavior is a way for companies to determine what consumers are looking for and/or how the company can increase their profits. Marketing teams use a technique called customer segmentation to explore patterns in similarly characterized groupings. This method of grouping customers allows companies to understand how individuals are similar and how these groups tend to purchase their products. Some segmentation methods utilize clear variables such as demographic or geographic information, others are more complex qualities that indicate more about the behaviors of the individuals. I have selected a dataset that has quite a few numerical and categorical variables in order to do a comprehensive analysis of these variables and their impact on customer segmentation and the trends in purchasing behavior. Through this analysis, I compare trends within standard demographic groups to the trends within the clusters.

In 1954, George Stigler of Columbia University explored the history of these processes in his paper, *The Early History of Empirical Studies of Consumer Behavior.* Through both a theoretical and empirical lens, he explores two points of focus, the income theory and the demand theory. Economists hypothesized long ago that the income of individuals has an effect on their spending habits. Stigler

notes that some of the earliest work in attempting to understand the habits of individuals sparked early implementation of minimum wage laws. The use of statistical analysis was employed to quantify the distinction in spending habits between different levels of income groupings. On the other hand is the demand theory - the idea that there is an inverse relationship between supply and demand. Understanding what consumers are purchasing and at what price amongst other variables can be crucial to making decisions that benefit both a company and the consumers.

In the analysis, *Gender Difference and Consumer Behavior of Millenials,* Kraljević and Filipovic recognise the uniqueness of the millennial generation and their impact on both the economy and rising trends. Using the foundations of purchase behavior in terms of loyalty, price sensitivity and whether the purchases were made in store or online, they attempt to explore potential distinctions between gender groups. They found that millennial women have a large sensitivity to price, and they are much more likely to enroll in loyalty programs or use loyalty rewards. Although the true motivation is unclear, this study indicates that millennial men tend to shop online more. These researchers recommend that we focus marketing efforts on how to best suit these individuals.

Students from the Shandong University of Finance and Economics explored the impact that the COVID-19 pandemic has had on consumer behavior in *The Impact of Consumer Purchase Behavior Changes on the Business Model Design of Consumer Services Companies Over the Course of COVID-19.* They wanted to explore the idea of 'panic buying' and did so by analyzing the changes in consumer purchases (CPC) on the object, motive, place, timeframe and method of the purchase at each segmentation of gender, age, income, and education. They found that the pandemic had a significant influence on purchasing behavior of the individuals studied, and this may be due to the idea that we may potentially make riskier or more impulsive purchases due to the mentality that such intense circumstances may bring.

DATA DESCRIPTION

The dataset used in this analysis was shared on Kaggle via user Sourav Banerjee who made use of the features of ChatGPT to build a synthetically created dataset that emulates accurate and reflective experiences of a shopper. The intention of this dataset is to give insight into model building but due to

the synthetic nature of the dataset, not all variables may " faithfully replicate" how they present in the real world, so any conclusions should not be used as generalizations to make any actual decisions.

This dataset contains 3,900 observations on 18 unique variables. Each customer observation includes some demographic information, customer behavior information, and information about the item purchased from this hypothetical company. The variables in this dataset are

Customer ID, Age, Gender, Item Purchased, Category, Purchase Amount (USD), Size, Location, Color, Season, Review Rating, Subscription Status, Shipping Type, Discount Applied, Promo Code Used, Previous Purchases, Payment Method, Frequency of Purchases

The variables Age, Purchase Amount, Review Rating, and Previous Purchases are the Numerical Variables in this dataset. All other variables are qualitative/categorical variables. This customer data uses a **Customer ID** to anonymously identify each customer. Although this is synthetically generated, this anonymization of observations is good practice for data privacy. **Age** is a numeric variable ranging from 18 to 70, **Gender** responses in this dataset are binary Male/Female elements, and **Location** indicates the state in the United States where the customer is hypothetically from. The **Category** variable gives a title to the general type of item under one of the four umbrellas Clothing, Footwear, Accessories, and Outerwear, while the **Item Purchased** variable provides what the item is more specifically. The **Purchase Amount** in US Dollars range from twenty dollars to one hundred dollars, and indicates the price paid for the item and not the originally advertised price. This dataset includes **Size** options ranging from Small to Extra Large. There are columns indicating both the **Season** and **Color** of the item. This dataset assumes each individual provides a **Review Rating** on a scale of zero to five. Three variables have a binary Yes/No response for whether the individual has a **Subscription Status,** if there was a **Discount Applied**, and if there was a **Promo Code Used.** The number of **Previous Purchases** ranges from 1 previous purchase to 50 previous purchases, while the **Frequency of Purchases** gives qualitative information about how often this individual purchases item(s) from this company, and both are indicators of customer loyalty. Lastly, the **Payment Method** and **Shipping Type** are qualitative variables whose values indicate the mode of payment (Cash, Card or Digital Transfer) and any **Shipping Type** selected or if it was an in store pick up.

As noted, this data is synthetically generated. So, while there are no missing values or outliers to clean, the issues reside in the quality of the creation of the dataset and its accuracy of the distributions to the real world. Some ways this dataset may not faithfully represent the real word include the assumption that all individuals gave a review score, and that each observation implies that each customer purchased only one item. Also, the variable Previous Purchases has a range from 1 to 50, but it could have included 0 previous purchases which would indicate a brand new consumer. Aside from the issues regarding the selection of these variable ranges, the accuracy of the variable distributions is something to consider. This distribution of most variables is relatively uniform and may hinder the generalizability of the analysis, but does not invalidate the processes or methods used. As seen in **Figure 1,** the numeric variables of Purchase Amount, Age, Previous Purchases, and Review Rating all seem to have a somewhat uniform distribution when looking at the entire dataset. The variable Review Rating has some unique peaks throughout its uniformity because it seems that more people give whole number rating scores than decimal values. In addition to the variable Previous Purchases not including 0 in the synthetic creation of the data, there seems to be a slight peak on the lower extreme, where a larger number of individuals have only 1 previous purchase, the minimum value for this variable in this synthetic dataset.

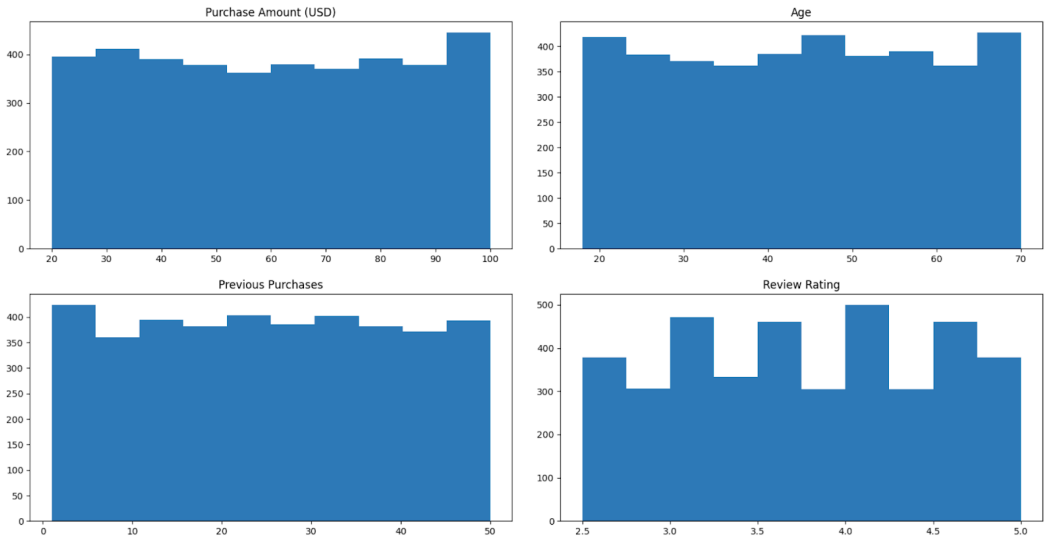

Figure 1: Distribution of Numeric Variables

Some categorical variables also have a relatively uniform distribution, while others have a bit of variation. As seen in **Figure 2**, the variables Size and Category seem to have the most variation, where it is clear most individuals purchase Size Medium and the Clothing Category compared to the other selections available. These particular variable distributions do seem to accurately reflect true purchasing behavior. Of the binary response variables, Gender and Subscription Status had the most variation, while Discount Applied and Promo Code Used are a bit more close in count as seen in **Figure 3**. Variables such as Item Purchased, Color, and Location have the most variety in choices, but similarly had relatively uniform distributions.

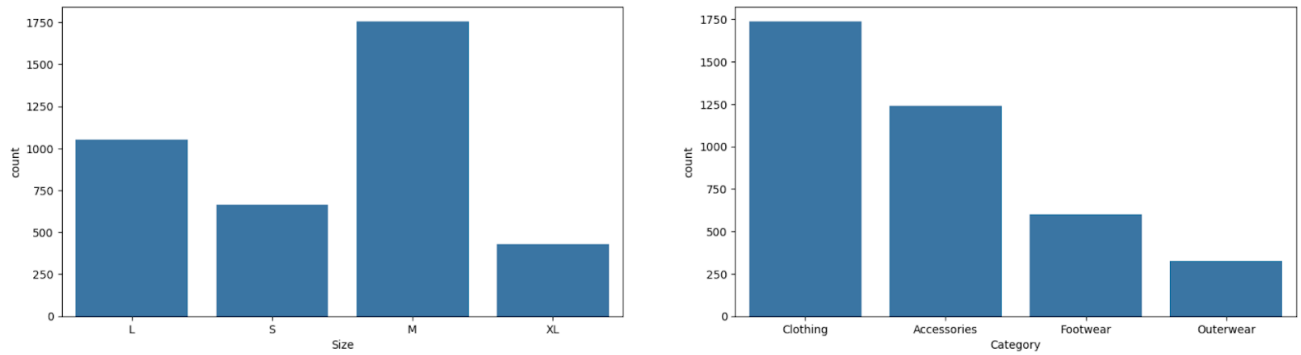

Figure 2: Distribution of Select Categorical Variables

Figure 3: Distribution of Binary Categorical Variables

METHODS

My main goal for this analysis is in the exploration and identification of patterns and trends within the customer segmentation groupings. I am aiming to gain a better understanding of how these purchasing behaviors and trends might present themselves through the data, so my project explores

patterns through the use of customer segmentation and visualizations to attempt to determine influential factors in making these groupings. Because of the exploratory nature of my analysis, I have chosen to use unsupervised methods of clustering which means I do not have a ground truth to compare the clusters against. The use of unsupervised clustering methods does not require training and test splits or cross validation. The preprocessing that this data requires includes scaling of the numeric features and encoding the categorical variables using dummy variables. Scaling is required because clustering methods can be extremely sensitive to the distances between points and certain variables could prove to be more influential in the clustering methods than they truly are. The true number of identifiable clusters is not determined ahead of time in this unsupervised analysis, so the accuracy of these both K-Means and Agglomerative is assessed using the silhouette score. Silhouette scores use the shape and density of the individual clusters to determine accuracy because there is not a true variable to confirm against. I explore the clustering accuracy of both KMeans and Agglomerative clustering methods with and without t-SNE dimension reduction.

I investigate the cluster algorithms of KMeans and AgglomerativeClustering ability to build a variety of different numbers of clusters. Therefore, I used mostly default values for the remaining parameters of K-Means, and only compared accuracy between different linkage modes for AgglomerativeClustering. After these initial attempts at clustering, I implement dimension reduction through t-SNE to visualize the cluster groupings and explore its effect on the silhouette score.

The use of visualizations is key in identifying patterns or trends in purchasing behavior. Once clusters are identified via these various methods, I reassign the cluster labels to the original dataset, and create visualizations to explore trends in purchasing behavior. To explore trends in purchasing behavior, I build sets of visualizations. Each visualization is intended to compare three different segmentation groupings. These three segmentation groupings I compare are the standard demographic group of Gender and the groups determined by the dimension-reduced K-Means clustering methods with both two and three clusters. Based on these demographic and cluster segmentation groups, I compare purchasing behavior on some influential numeric and categorical features. Some of the distributions I explore are Purchase Amount (USD), Age, and Subscription Status. Through my exploration of these visualizations, I identify features that have the most impact on segmentation.

RESULTS

For these unsupervised clustering methods of KMeans and AgglomerativeClustering, I used all variables aside from the Customer ID to attempt to build clusters and segment individuals based on their purchasing behavior. Both clustering methods use the euclidean distances so ensuring these variables are within the same ranges can limit bias of an unnecessarily influential variable from the dataset. Therefore, numeric values were scaled and categorical variables were encoded with dummy variables. As my goal was to tune the number of clusters, I used the default parameters for K-Means. This includes using the k-means++ initialization at 1 iteration, and the maximum iterations is set to 300. For Agglomerative Clustering, I did tune the linkage parameter and found that average was slightly more successful than ward in some cases, but the cluster distributions were questionable, so I continued with the default linkage parameter of ward for AgglomerativeClusering.

Unfortunately, due to the high dimensionality and generally uniform distribution of the data across many variables, there were not incredibly successful results in the standard clustering methods. Prior to dimension reduction, K-Means clustering resulted in silhouette scores of 0.098, 0.069, and 0.058 for two, three, and four clusters respectively, and Agglomerative Clustering resulted in silhouette scores of 0.098, 0.054, and 0.038 for two, three, and four clusters respectively. These scores are incredibly low, but I did not expect great results due to the uniformity and synthetic nature of this dataset. To continue with my analysis, I looked to dimension reduction techniques to potentially improve these scores.

Implementing the dimension reduction techniques of t-SNE did result in better silhouette scores as seen in **Figures 4 and 5**, but it was interesting to note that both K-Means and Agglomerative Clustering resulted in the same silhouette scores for when two clusters were made both before and after dimension reduction. It was clear from the visualizations that K-Means and Agglomerative Clustering both identify seemingly identical clusters with two cluster groupings, where silhouette scores for K-Means and Agglomerative were both 0.569 for two clusters. When looking at distributions of observations in two clusters, both the initial and dimension reduced clustering methods saw similar if not the same exact distribution. This consistency across all models may indicate that there may be some subset of the data that all models recognize as a unique cluster.

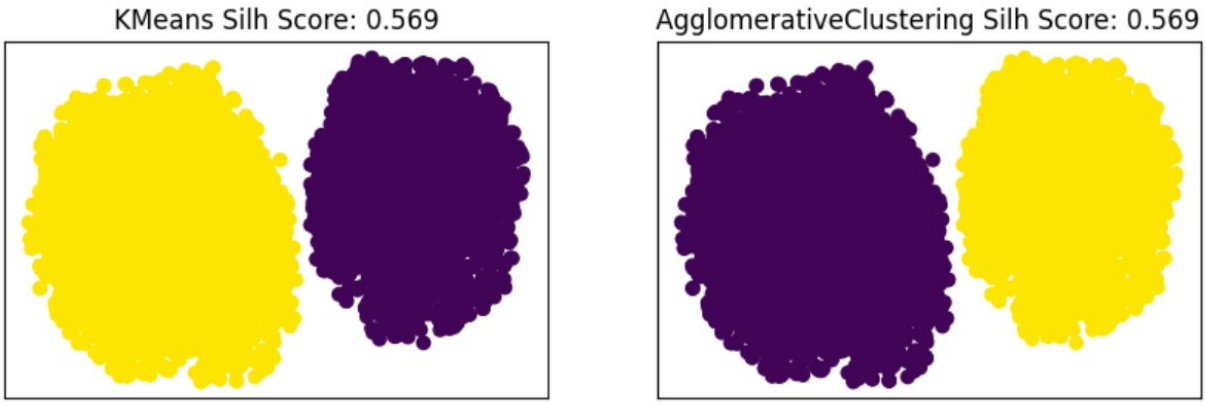

Figure 4: Visualization of Two Clusters with t-SNE Dimension Reduction

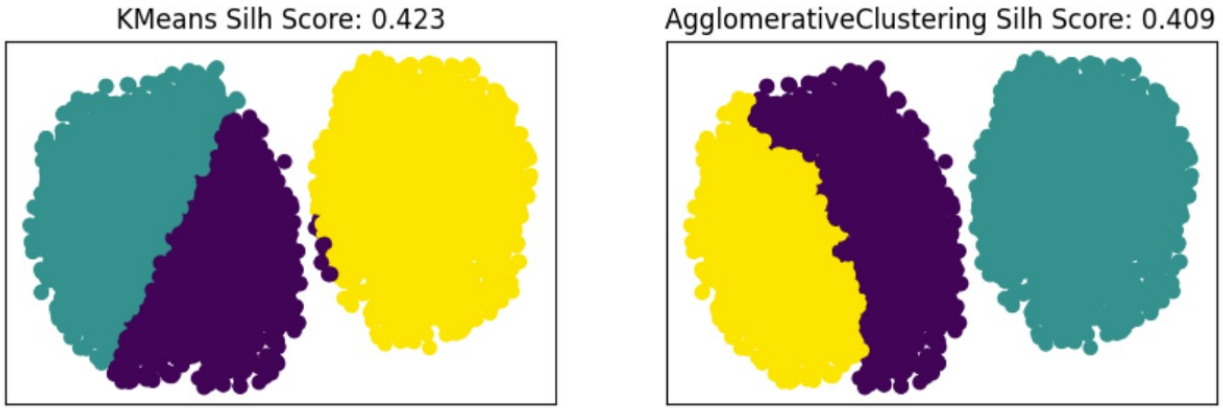

Figure 5: Visualization of Three Clusters with t-SNE Dimension Reduction

Similar to the results prior to dimension reduction, we see slightly different silhouette scores when the number of clusters is increased to three. K-Means performs slightly better than Agglomerative with a silhouette score of 0.423 compared to Agglomerative's silhouette score of 0.409. While these silhouette scores are not exceptional, they are a drastic improvement from the standard clustering methods prior to dimension reduction. There is some difference in distribution of the observations into three clusters when looking at the clusters formed before and after the dimension reduction was implemented. Because these silhouette scores are greater when t-SNE dimension reduction techniques are applied, I continue my analysis by using the two clusters and three clusters formed by dimension reduction techniques for my visualizations.

The first visualization created aims to explore the distribution of gender within the original dataset and within the cluster groupings. The visualizations in **Figures 6 and 7** clearly show that one cluster has all males, while the other cluster grouping has a mixture of males and females. In exploring these splits, it seems that all individuals with a Subscription Status were males, so all methods saw this Male with Subscription as a pretty clear cluster.

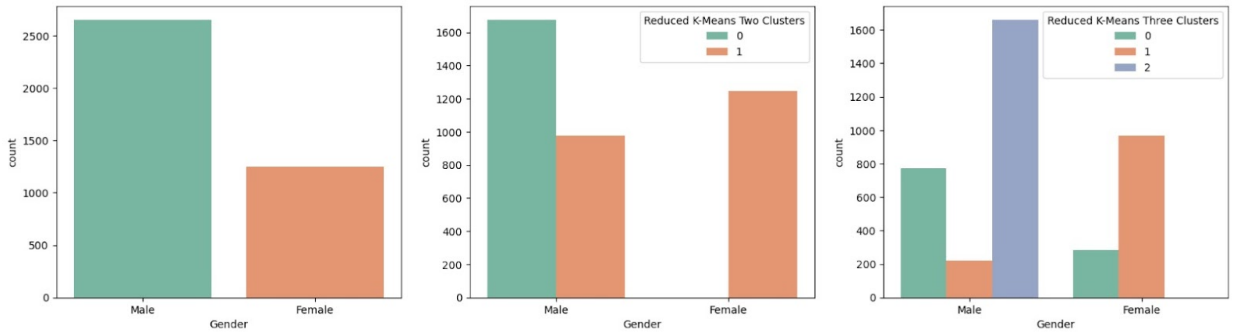

Figure 6: Distribution of Gender across Segmentation Groups

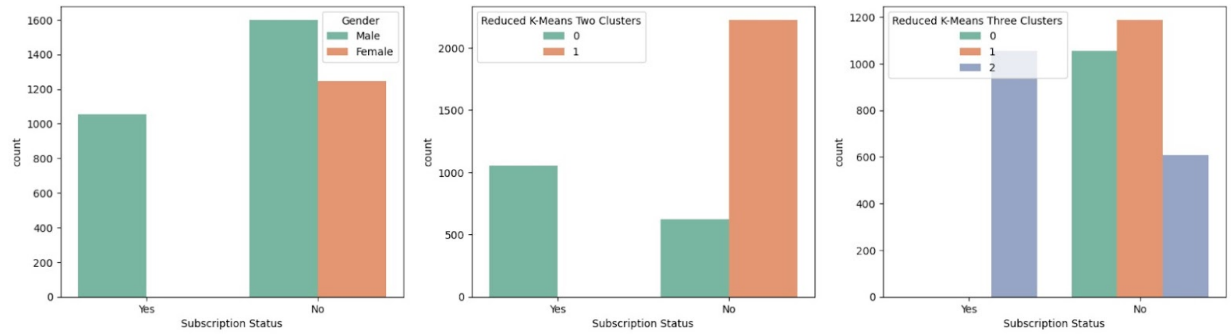

Figure 7: Distribution of Subscription Status across Segmentation Groups

Next, I then explored the distribution of Purchase Amount and Age of the cluster groupings. There does not seem to be any trends in Purchase Amount when looking strictly at the demographic segmentation groups of Gender or even the two clusters. However, in the distribution plots in **Figures 8 and 9** that splits the data between the three cluster groups, there seems to be some interesting trends. Cluster 2 includes all of the Males with a subscription status, so when we look at the remaining two clusters for trends, we see some clear patterns in both Purchase Amount and Age. Cluster 0 seems to skew right in both Purchase Amount and Age, while Cluster 1 seems to skew left in both these

distribution plots. Despite these values having no correlation when looking at the entire dataset, Cluster 0 has a high density of individuals purchasing lower priced items with a simultaneously high density of individuals on the lower end of the age spectrum. Conversely, Cluster 1 holds a high density of individuals purchasing higher priced items as well as a higher density of individuals on the higher end of the age range here. With the current cluster groupings formed, other numerical variables of Previous Purchases and Review Rating seem to have no distinguishable distribution trends. Some categorical variables of Category and Size also don't have any notable differences amongst the determined cluster groupings.

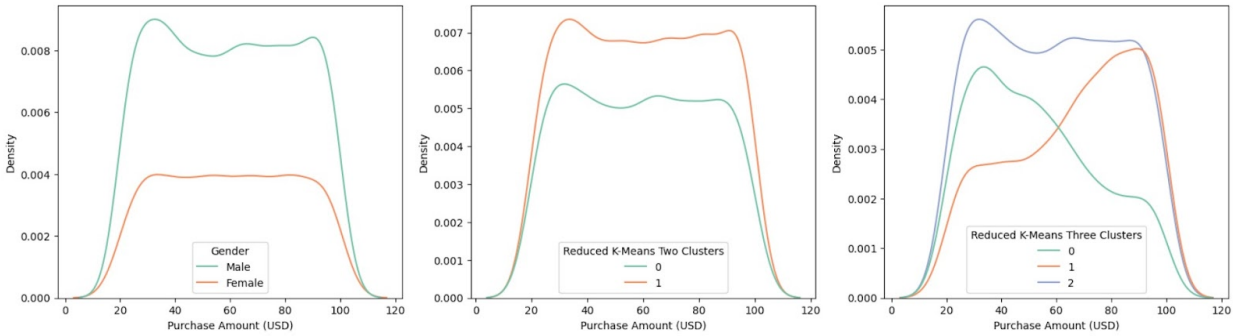

Figure 8: Distribution of Purchase Amount across Segmentation Groups

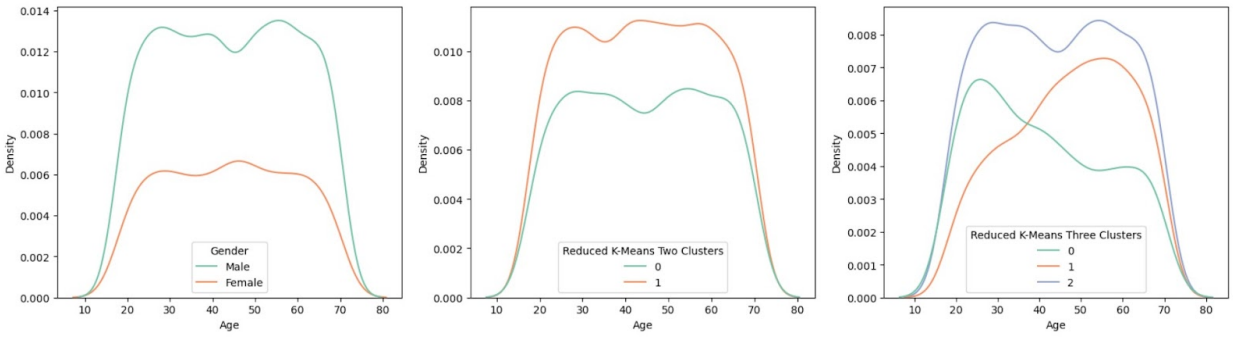

Figure 9: Distribution of Age across Segmentation Groups

CONCLUSIONS

Due to high dimensionality and uniformity of the data, standard clustering techniques without dimension reduction were incredibly unsuccessful. However, despite the uniformity of the data, I was delighted to get weak to moderate success after dimension reduction was implemented.

While I was limited in my scope of this project, it would be very interesting to continue to explore the cluster analysis further as we increase the number of clusters and see what features are impactful in building these cluster groups. Although at this time, Previous Purchases, Review Rating, Category, and Size seem to have no distinguishable distribution trends, there is more potential to expand this customer segmentation analysis. Just as the trends weren't visible until the third cluster was made, there may be more trends to recognize if we were to increase the number of clusters. Additionally, I would be interested in exploring and comparing different demographic groups to these and future clusters. Demographic groups available within this dataset include location and age. While both of these variables have an extensive range, it could be possible to group them into age brackets or location regions for future analysis and comparison.

Although this project currently lacks potential for generalizability, with the knowledge I gain from these insights, I can potentially apply these methods to a dataset that is created using actual data and could then use that more accurate information to make suggestions to a potential company. It would be interesting to see if the data would still recognize Subscription Status as an influential feature if it were a mixture of Males and Females as opposed to all Males. Now that I have explored customer segmentation and developed some visualizations with these techniques, I would be interested in seeing what happens when using data that is not so uniform or more faithfully represents true shopper experiences.

SOURCES

Dataset: https://www.kaggle.com/datasets/iamsouravbanerjee/customer-shopping-trends-dataset

Stigler, George J. "The Early History of Empirical Studies of Consumer Behavior." *Journal of Political Economy*, vol. 62, no. 2, Apr. 1954, pp. 95–113, https://doi.org/10.1086/257495.

Kraljević, Radojka, and Zrinka Filipović. "Gender Differences and Consumer Behavior of Millennials." *Acta Economica et Turistica*, vol. 3, no. 1, 27 June 2017, pp. 5–13, content.sciendo.com/view/journals/aet/3/1/article-p5.xml, https://doi.org/10.1515/aet-2017-0002.

Tao, Hu, et al. "The Impact of Consumer Purchase Behavior Changes on the Business Model Design of Consumer Services Companies over the Course of COVID-19." *Frontiers in Psychology*, vol. 13, 2022. *Frontiers*, www.frontiersin.org/articles/10.3389/fpsyg.2022.818845/full, https://doi.org/10.3389/fpsyg.2022.818845.

APPENDIX:

Video Presentation: [Video Presentation - Youtube]

Python Code: DAWSON_Final Project Code.pdf

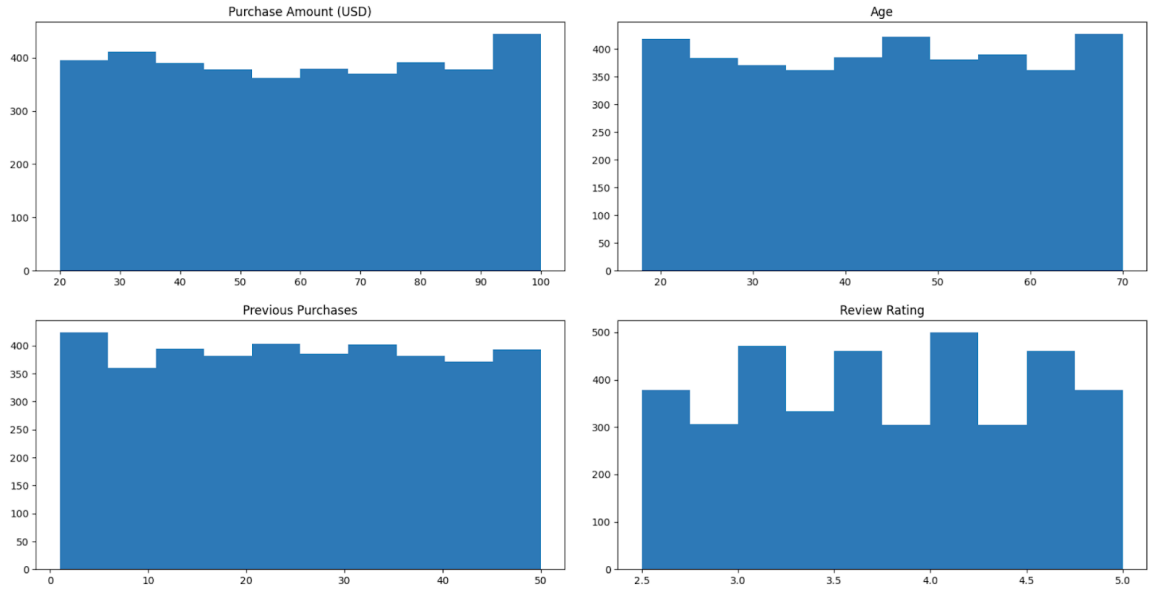

Figure 1: Distribution of Numeric Variables

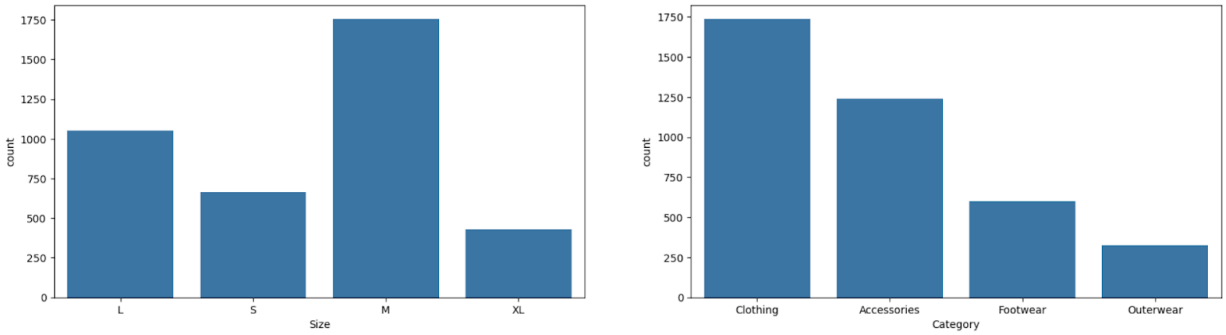

Figure 2: Distribution of Select Categorical Variables

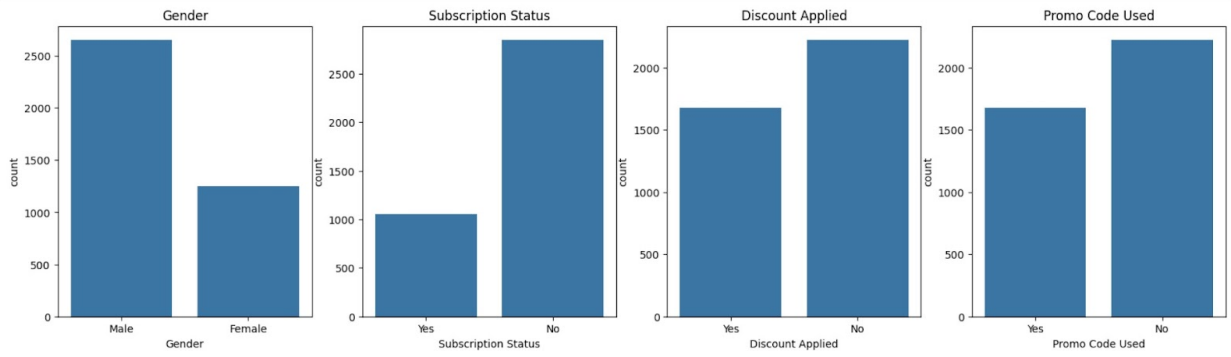

Figure 3: Distribution of Binary Categorical Variables

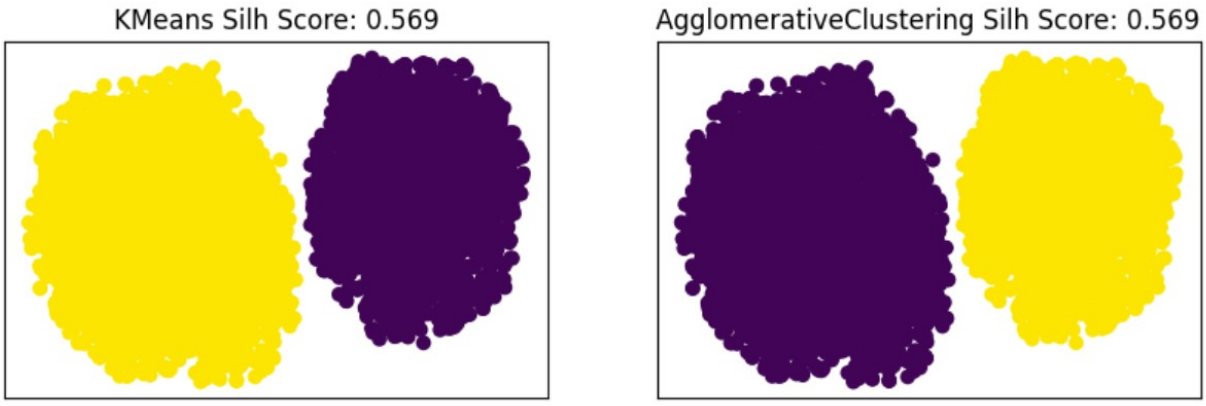

Figure 4: Visualization of Two Clusters with t-SNE Dimension Reduction

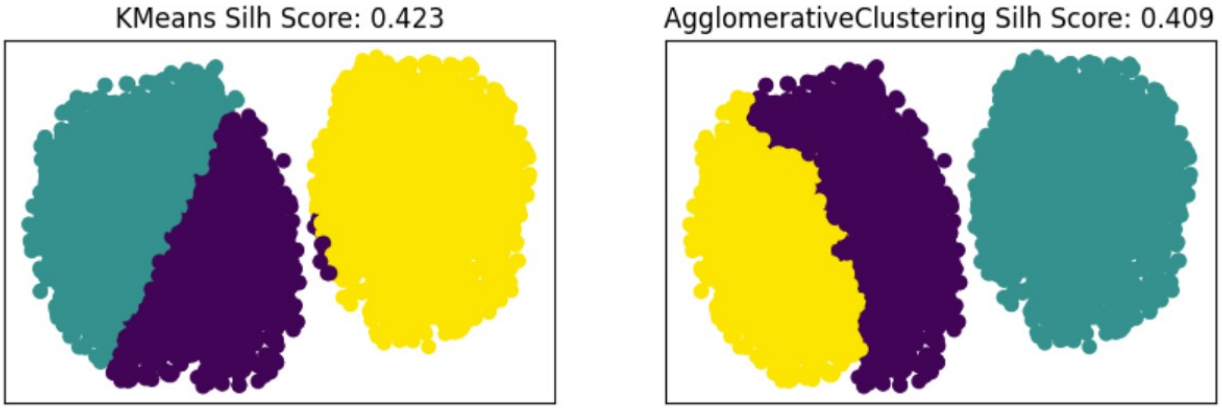

Figure 5: Visualization of Three Clusters with t-SNE Dimension Reduction

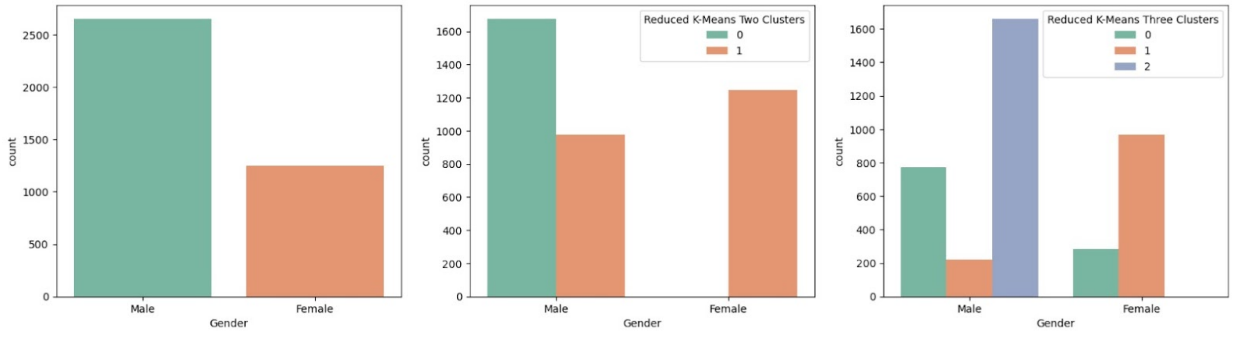

Figure 6: Distribution of Gender across Segmentation Groups

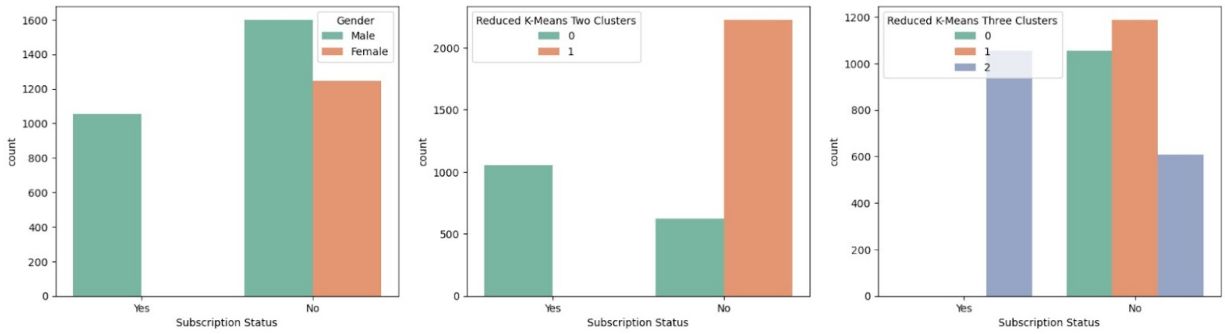

Figure 7: Distribution of Subscription Status across Segmentation Groups

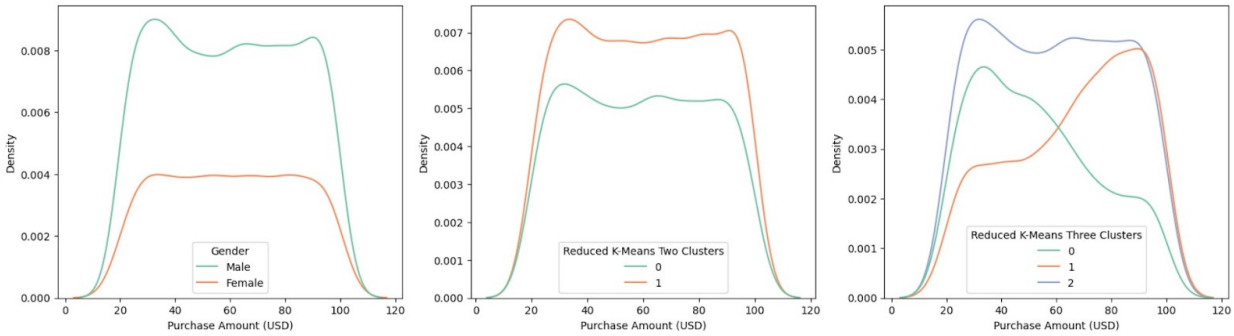

Figure 8: Distribution of Purchase Amount across Segmentation Groups

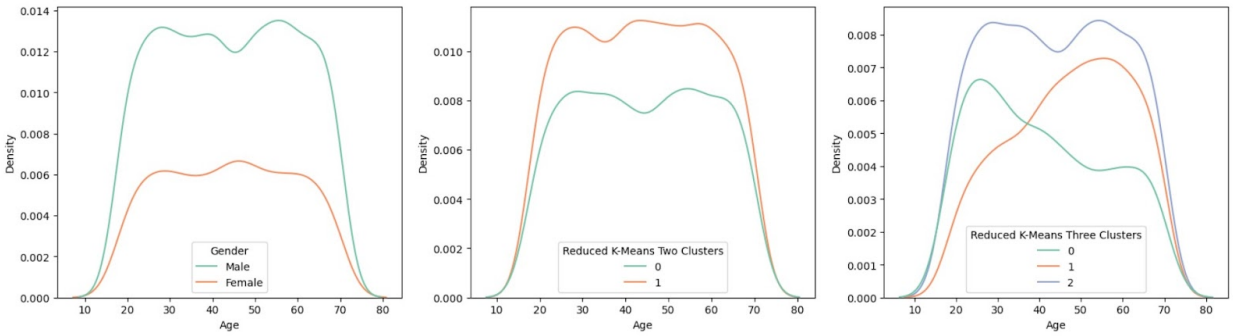

Figure 9: Distribution of Age across Segmentation Groups

Datasource: [CSV - Kaggle]

Sample:

	Age	Gender	Item Purchased	Category	Purchase Amount (USD)	Location	Size	Color	Season	Review Rating	Subscription Status	Shipping Type	Discount Applied	Promo Code Used	Previous Purchases	Payment Method	Frequency of Purchases
0	55	Male	Blouse	Clothing	53	Kentucky	L	Gray	Winter	3.1	Yes	Express	Yes	Yes	14	Venmo	Fortnightly
1	19	Male	Sweater	Clothing	64	Maine	L	Maroon	Winter	3.1	Yes	Express	Yes	Yes	2	Cash	Fortnightly
2	50	Male	Jeans	Clothing	73	Massachusetts	S	Maroon	Spring	3.1	Yes	Free Shipping	Yes	Yes	23	Credit Card	Weekly
3	21	Male	Sandals	Footwear	90	Rhode Island	M	Maroon	Spring	3.5	Yes	Next Day Air	Yes	Yes	49	PayPal	Weekly
4	45	Male	Blouse	Clothing	49	Oregon	M	Turquoise	Spring	2.7	Yes	Free Shipping	Yes	Yes	31	PayPal	Annually