Sprint #2: Superstore Sales: A Customer Segment Analysis

Dave Delva (217624347), Deenu Yadav (307381), Fanny Guevara (307501), Kaustubh Mulay (307493)

Executive Summary

For this project, we will use SuperStore sales data and transform the data into value based insights, in order to increase revenues, offer new products and services, better balance supply and demand, optimize marketing offerings and increase customer engagement.

Our value proposition will be delivered via a cloud based dashboard and recommender system that will facilitate the management of items that are key to achieving the financial goals of the organization. The organization will have a tool that will provide visibility into sales, inventories, and profit to maximize sales and minimize the overstocks that drain revenue and profit.

Project proposal

Utilizing existing data objects in the customer data file, we will explore the customer and category segments necessary for the analysis.

It is important for all organizations to evaluate the consequence of the current COVID-19 pandemic on the economy and its business. As it relates to the Superstore, the impacts of this pandemic will affect all lines of business especially supply chain. The extent of the impacts might not be fully known until further analysis is dedicated to this one issue. Given the time restriction for this project, the effects of COVID-19 are out of scope.

Methodology

We will present our preliminary analysis in the following sections (1) Data Manipulation, (2) Descriptive Analysis and (3) Data Modeling.

3.1 Data Manipulation

In consideration of the ethical machine learning framework and data privacy, we have removed the customer name column from our dataset to protect the customer's identity.

We start our data cleaning by converting the date variables from a factor data type to a date_time data type. We further analyze our data for missing or special values, outliers and obvious inconsistencies. As it relates to orders and states, a single outlier was identified. The state of Wyoming had a single order in the four year time frame of the dataset. As we are unable to verify the reason, we have decided to remove the entry. Further, we identified that the Row IDs and Order IDs do not have a unique relationship. Therefore, further analysis will be conducted by the aggregation of "Order ID" in order to eliminate double counting.

Aside from these 3 identifications, the data is said to be clean, consistent and normally distributed.

3.2 Descriptive Analysis

Based on our data objects, we have identified our main target variables and have separated them into 2 categories, independent and dependent variables:

- **Independent Variables**
  - Customer ID
  - Category
  - Quantity
  - Region
  - Segment
- **Dependent variable**
  - Sales
  - Profit
  - Discount
  - States
  - Sub-Category

Our multivariate analysis of Profit, Sales and Discount demonstrates an inverted relationship between Discount and Profit; as products are discounted beyond 50%, the profit drops to points of interest such as below $4,000 ***(Figure 1D)***. Further, products are typically discounted up to 20% with items being discounted above 40% seen as outliers, which are worth further exploration ***(Figure 1C)***.

Moreover, we look at Profit and Discount by Customer Segment to get an idea on how these customer segments impact sales. We can see that the Consumer segment tends to buy more products in the 80% discount range ***(Figure 2a)***. When evaluating the impact of Discount and Profit by Product Category, we can see that Profit is significantly impacted by the Office Supplies and Technology Categories as they tend to Discount items in the range of 70% - 80% ***(Figure 2b)***.

We can see that most of the sales (32%) occurred in the Western region; followed by the Eastern region (20%), Central region (23.5%), and the remaining (16.4%) in the Southern region. Further, 20% of all total sales happened in the State of California, 11% in New York, 9% in Texas, and the remaining sales divided equally between the remaining 49 states. Additionally, 51.6% of the sales are attributed to the Consumer segment, 30.2% of sales were made by the Corporate segment and 18.1% for the Home Office segment ***(Figure 3 and Figure 4)***.

3.3 Data Modelling

Using the descriptive analysis performed above, we have identified 3 segments: (1) Orders by Category, (2) Orders by Customer Type, (3) Orders by Region.

We will explore Hierarchical and K-Medoids Clustering in attempts to derive additional effective dimensions for deciding the segment type/clustering groups in our dataset.

Model 1: Hierarchical Clustering by Profit Segmentation

The first model we set out to use was the hierarchical clustering model. To compute hierarchical clustering, first we compute distances. To compute the distance measure we used Euclidean distance. After having the distance object defined, we computed hierarchical clusters using 'ward.D2' method. Ward's method is the classic method to compute such clusters as it starts with individual data points and merges other data points to form a cluster on iterations; this approach of Ward helps minimize total within the sum of squares of the clustering.

Figure 5a and 5b in the appendix show the cluster model and dendrogram of the generated clusters and cluster groups.

Model 2: K-Medoids Clustering

K-Medoids Clustering is an alternative approach to k-means clustering for identifying groups in the dataset. It does not require one to pre-specify the number of clusters to be generated as the calculation is done by the Silhouette width method ***(Figure 6)***. We first measured the Gower distance and then the Silhouette width; which indicated the number of optimal clusters to be 7.

We attempted the clustering using the K-Medoid method ***(Figure 7a)***. The clusters formed by the algorithm were not ideal clusters as they were spread throughout the plain and not uniformed. Therefore, we modified the input to see the output of 3 clusters ***(Figure 7b)***.

After modifying the cluster output we found similar results to the 7 cluster output. Consequently we would be considering this model to not be suitable for our data set and opt for a different segmentation method.

Modeling and Evaluation

Associative Rule Mining - Apriori Algorithm

Apriori algorithm is part of associative rule mining techniques and often it is used in market basket analysis to determine the frequency of patterns and relevant associations in a dataset. It predicts the next relevant item in the itemset based on associations and correlations between the itemsets. The Associative rule mining technique is mostly used in business decisions according to customer purchases. In order to determine correlations and associative items, the algorithm uses measures such as support, confidence and lift.

Support: The support of an itemset X, $supp(X)$ is the proportion of transaction in the database in which the item X appears. It signifies the popularity of an itemset.

$supp(X)$=Number of transaction in which$X$appears / Total number of transactions.

Confidence: Confidence of a rule is defined as follows:

$conf(X \rightarrow Y) = supp(X \cup Y) supp(X)$
It signifies the likelihood of item Y being purchased when item X is purchased.

Lift: The lift of a rule is defined as:

$lift(X \rightarrow Y) = supp(X \cup Y) supp(X) * supp(Y)$

This signifies the likelihood of the itemset Y being purchased when item X is purchased while taking into account the popularity of Y.

For this project, we have implemented apriori in two steps, one for frequent itemset generation with support (0.025) and confidence (50%), and second step to generate the candidate rules from each frequent itemset. The candidate rule provides the lhs (selected dataset item) and rhs (recommended corresponding dataset item list) components to prompt customers with a recommended list of items on the selected one.

Model Implementation and Evaluation

We have developed the Dashboard serving two functions:

1. A descriptive dashboard to simulate the performance of various segments
2. A recommender system to show the consumer preferences and purchasing behaviours

The dashboard is hosted on a secured Shiny cloud platform to simplify integration, improve agility, and ease the burden of implementation for Superstore IT resources. (**Figure 8a & 8b**)

Examples of tangible insights delivered by the dashboard:

1. Sales are up on average nationwide across all 3 of the product categories. How can the profitability of Furniture be improved?

2. Days to ship was slightly up in 3 out of the 4 regions. Standard Class shipping appears to be the culprit. What is the reason?

3. Profit ratio was down in the East driven by a -134% profit ratio in Machines. A root cause analysis is required.

Insights and summarizing results

Based on the data exploration and descriptive analysis, we were able to identify various segments and key variables that gave insight into how the organization is performing in each segment. The data provided had a great number of categorical variables. The inferences of modeling is derived from associative type algorithms which on one hand can predict a consumer behavior in buying the products together and on the other hand, allow buyers and planners to understand which items are purchased together to ensure they are also bought and stocked together.

Furthermore, it is key to understand that the 3 customer segment categories (Home Office, Corporate, Consumer) have very unique and distinct buying patterns and purchasing needs. For example, there is great variability in the frequency and quantity of items purchased. Some organizations have scheduled office supply purchasing windows while others are less structured and infrequent. We noticed that the minimum support and confidence in the dataset to determine the association and correlation relationships were impacted by unique and distinct buying patterns and purchasing needs of the customer segments.

Conclusion

The interactive dashboard and recommender system being delivered by this project will facilitate the management of items (such as, but not limited to inventory, sales performance - current and historical, customer preferences) that are key to achieving the value proposition set out to be delivered in this project.

The platform is intended to be utilized throughout the organization in an effort to encourage cross functional collaboration to develop concise, strategic plans and goals based on sound metrics and employ marketing tactics based on said metrics. The dashboard will facilitate the analysis of current and past performance in an easy-to-consume format yielding a better understanding of trends and market segments that will feed into an increase in profit through better planning and buying.

In closing, the key benefits of the tools could be summarized into the following points:

- Improved Accuracy
    - Inventory
    - Shipping
    - Buying
- Increased Profits
- Increased Return On Investment (ROI)
- Reduced Markdowns
- Reduced Profit Loss

Sprint #1: Superstore Sales: A Customer Segment Analysis

Appendix

Figure 1: Multivariate analysis for Profit, Sales and Discount

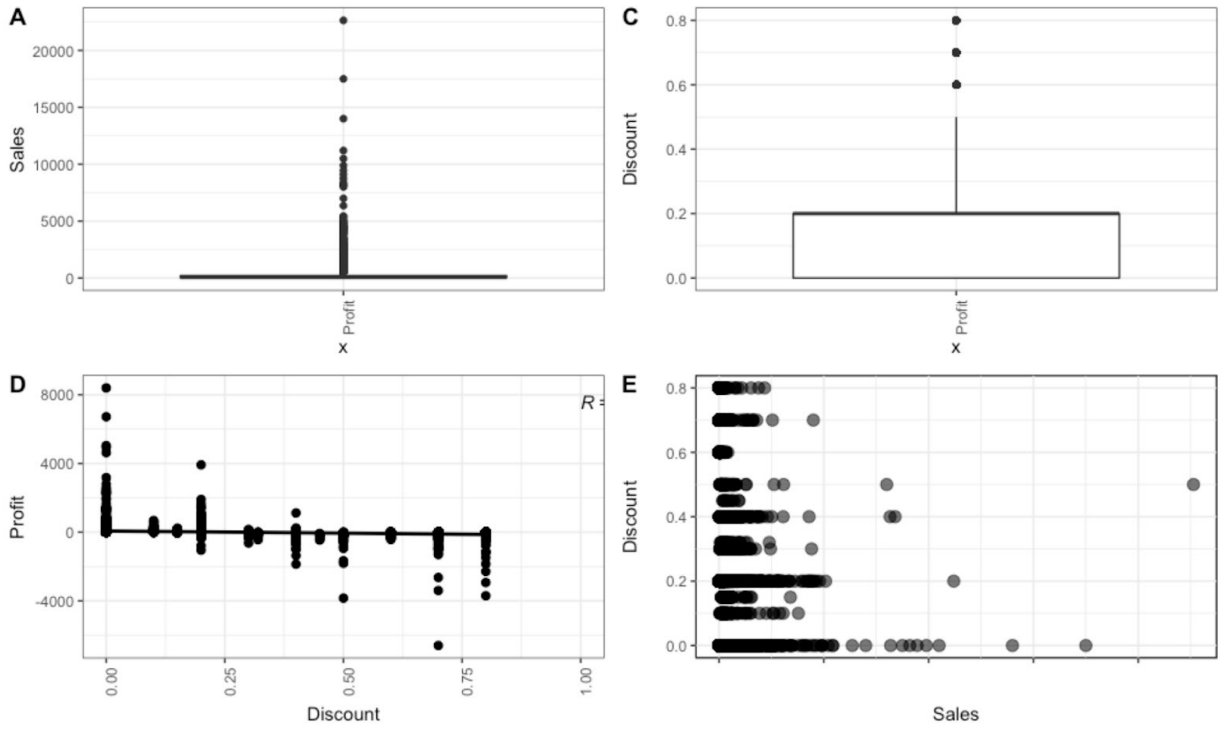


Figure 2a: Univariate Analysis of Profits for Customer Segments

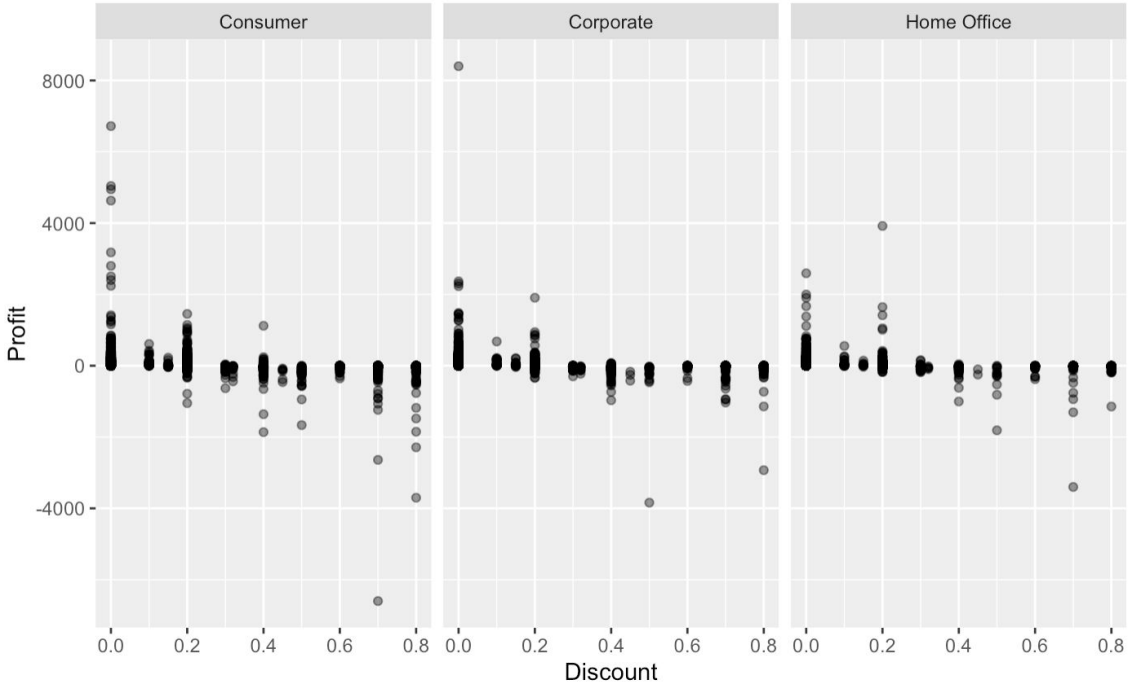


Figure 2b: Univariate Analysis of Profits for Category

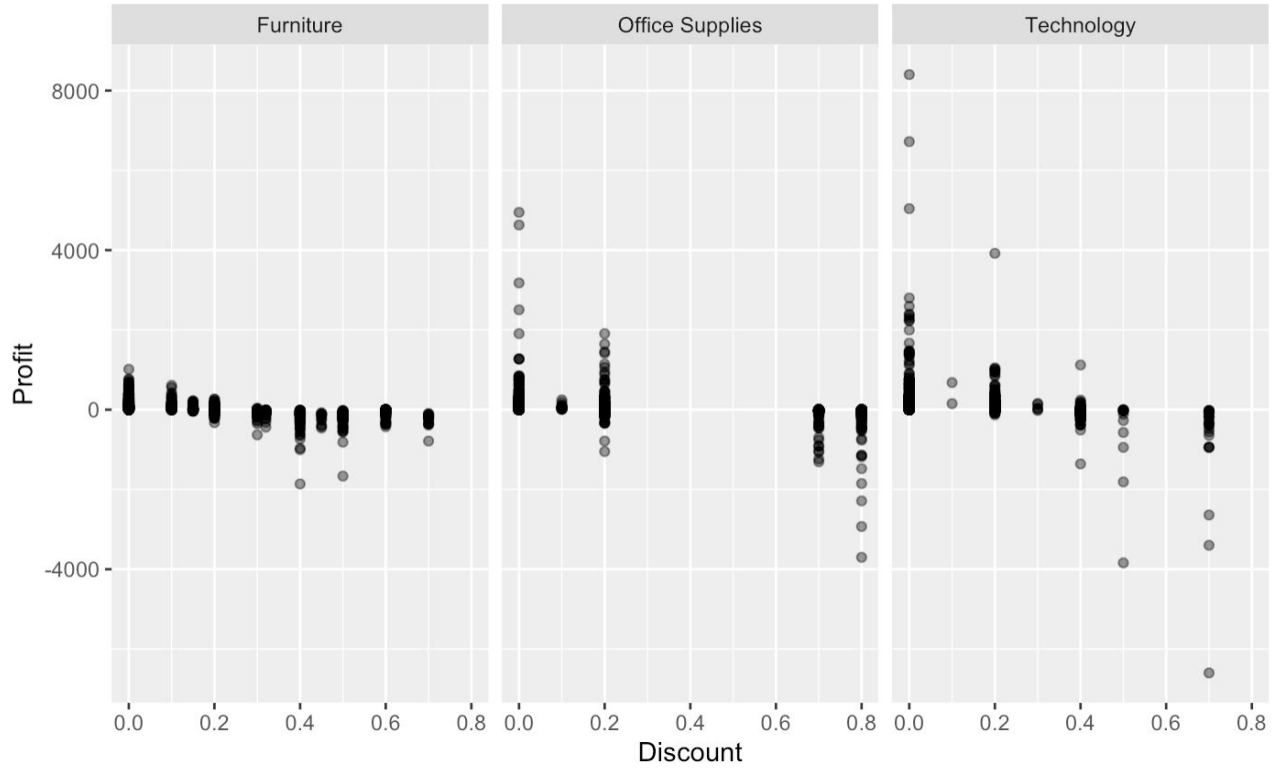
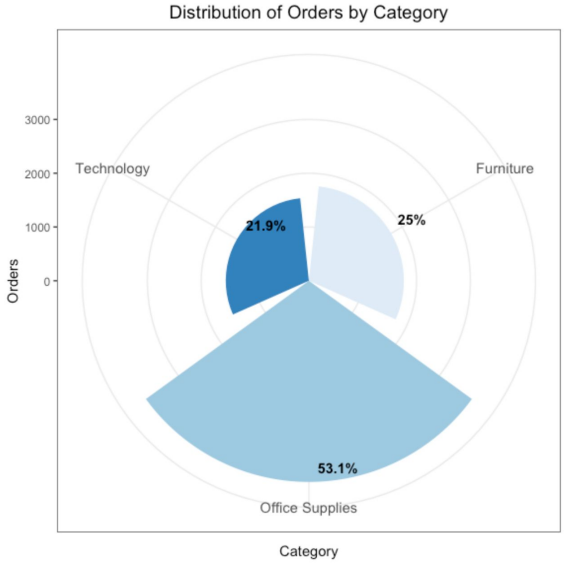
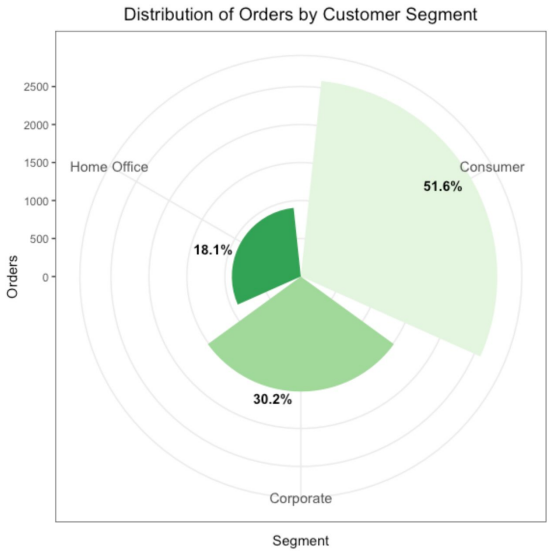


Figure 3: Distribution of Orders by Category, Region, and Customer Segments





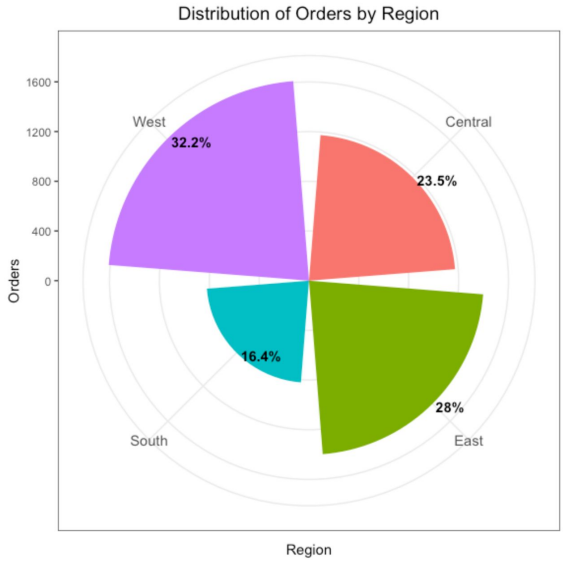


Figure 4: Additional Descriptive Analysis

- Sales can be further broken down by product categories, which show that Office Supply leads the sales at 53.1%, while Furniture and Technology items share the rest of the sales at 25% and 21.9%, respectively.
- There was an increase in online orders annually from 2014 to 2017, while the average profit and average sales fluctuated. Office supplies dominated by 60% of the consumer orders placed online between that same time.
- Analysis of the subcategories of products shows that there is more customer demand for Home Office products such as binders, papers, phones, storage, chairs and accessories. A trend we expect to increase in the short term due to the shift towards working from home because of the Covid-19 self isolation requirements made by various governments across the globe.

Figure 5a: Hierarchical Cluster in Circular Dendrogram Tree for Profit Segmentation

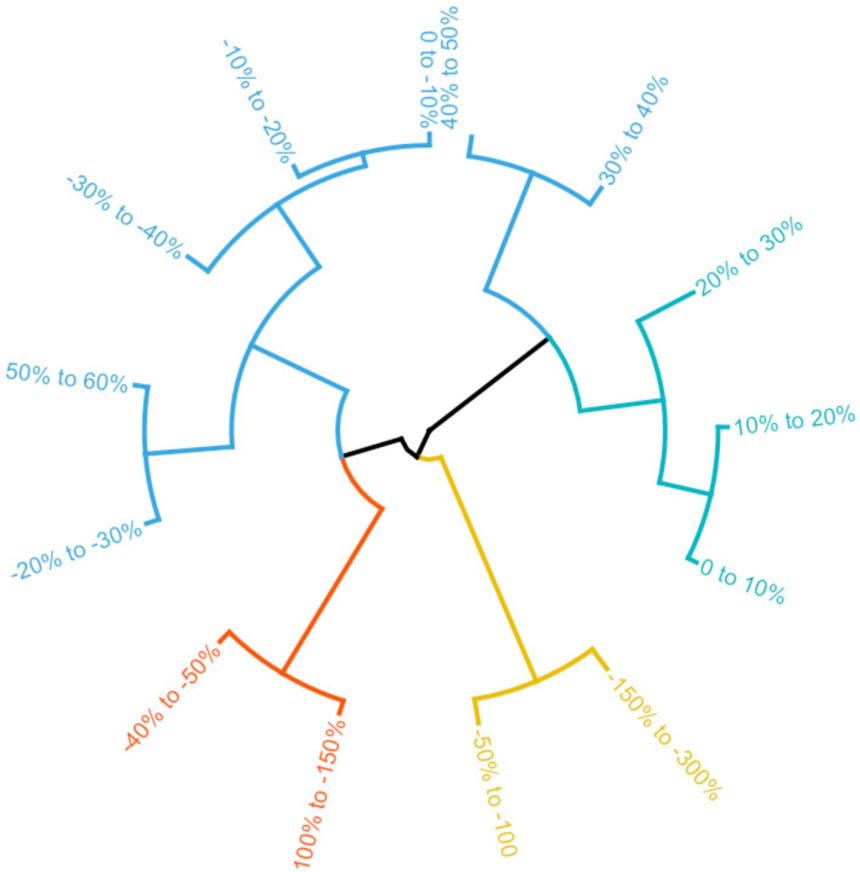


Figure 5b: Hierarchical Cluster Analysis in Dendrogram Tree for Profit Segmentation

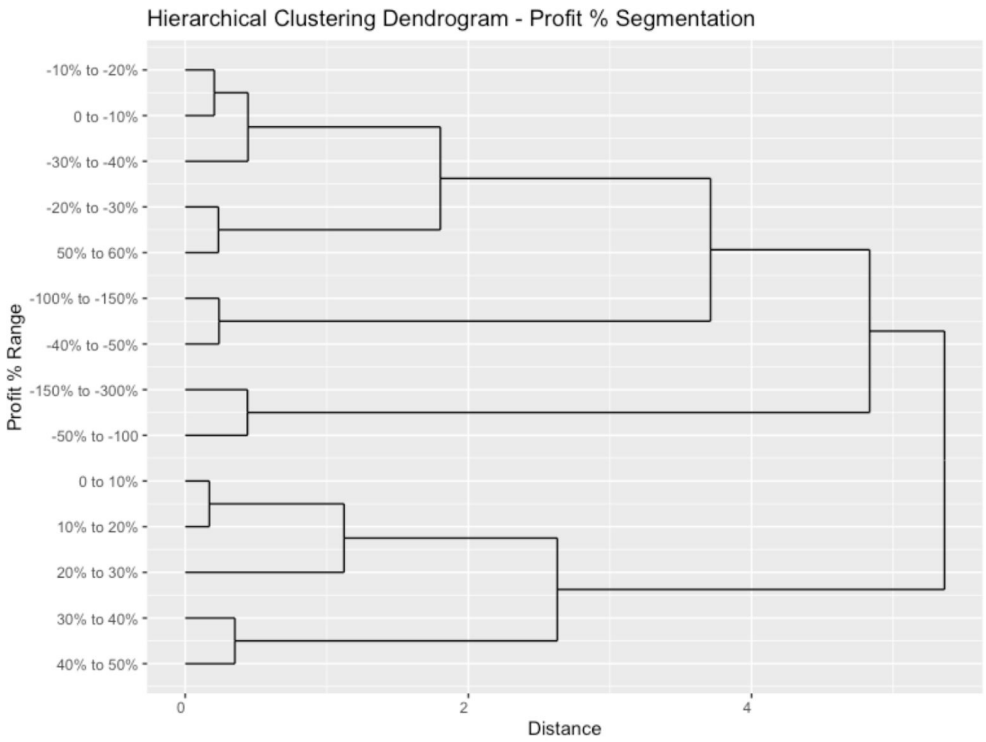


Figure 6: Visualization of Silhouette Width indicating the optimal number of clusters

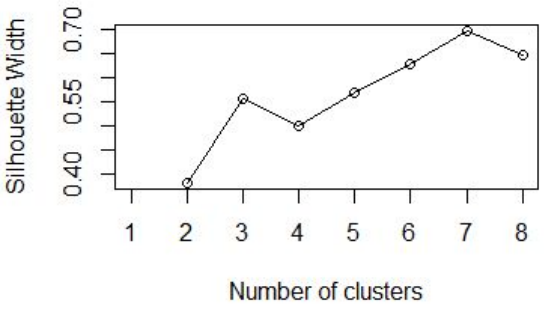


Figure 7a: Visualization of the 7 Clusters

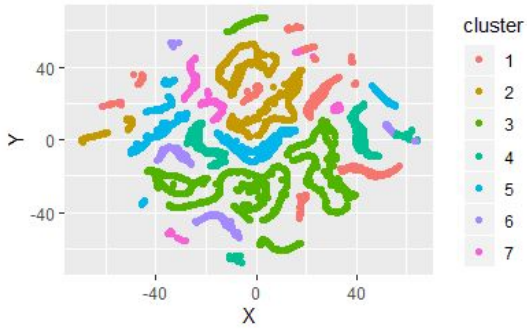
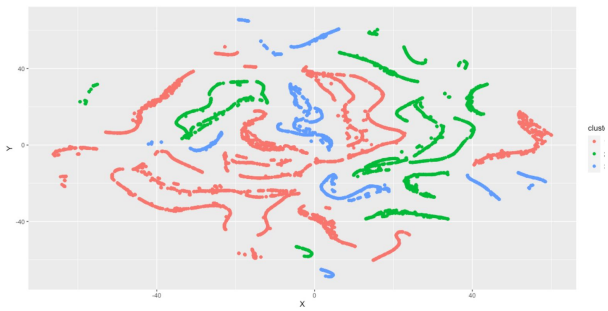


Figure 7b: Visualization of the 3 Clusters



Sprint #1: Superstore Sales: A Customer Segment Analysis

Figure 8a: Analytic Dashboard

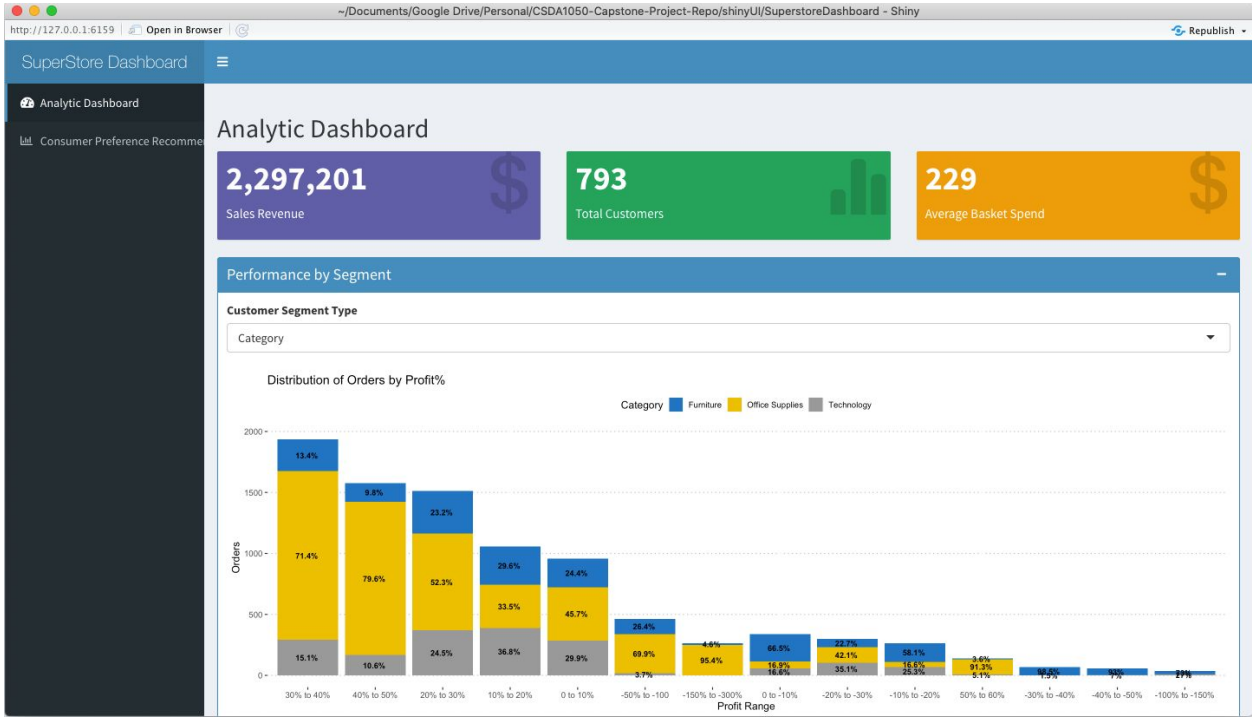


Figure 8b: Consumer Preference Recommender Dashboard