



Capstone Project: E-Commerce Website Customers Segmentation

Problem Statement

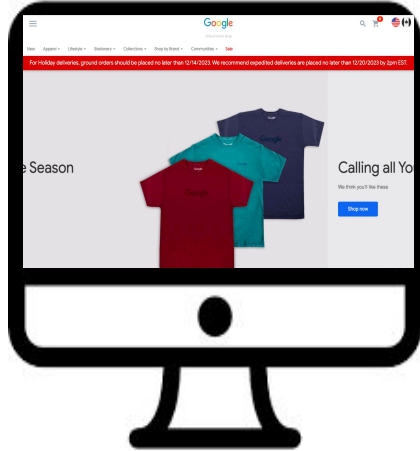
The marketing team of the Google Merchandise website wants to identify different segments from its pool of customer data, understand each segment's behavior, determine the segment that marketing team should invest its marketing expenditure to boost their spending and propose marketing tactics for the chosen segment

Process:

1. Identify different segments of customers
2. Understand each customer segment behavior
3. Choose a customer segment to invest marketing expenditure
4. Determine marketing tactics for the chosen segment

Website Overview

shop.googlemerchandisestore.com



Data is retrieved from Kaggle

Product Groups (Brand)

Google

Android

Youtube

No Brand

Product Categories

Apparel

Bag

Drinkware

Head Gear

Lifestyle

Google Nest

Google Waze

Products

Men's T-Shirt

Women's T-Shirt

Baby Short Sleeve

Tote Bag

Pouch

Tumblr

Mug

Hat

Cap

Kickball

Yoga mat

Thermostat

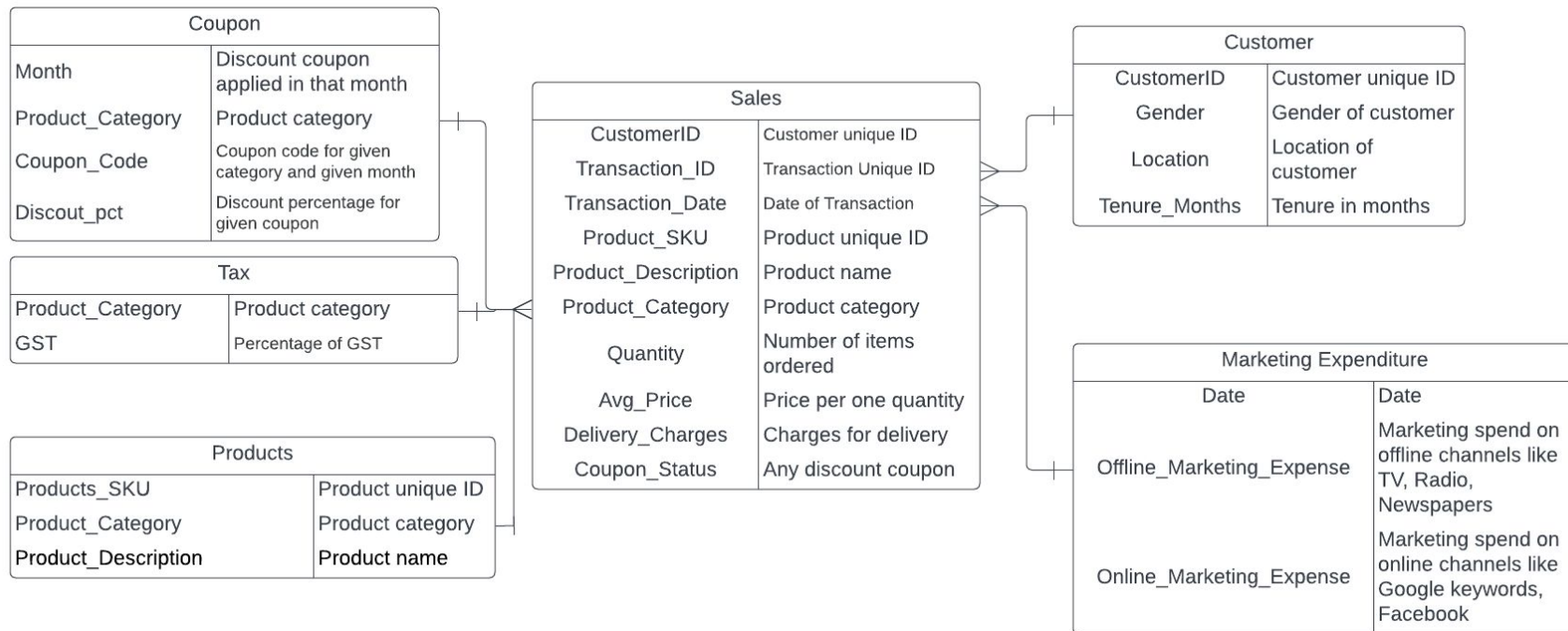
Camera

Sock

Tee

Data Preview

Data Range: 2019-01-01 to 2019-12-31 (364 days)



Data Preparation: Customer Analysis and Product Analysis

Customer Centric

Customer_Analysis	
CustomerID	Customer unique ID
Total_Spending	Total spending of customer
Delivery_Charges	Total delivery charges of customer
Num_Transactions	Total number of transactions
Total_Quantity	Total quantities
Location	Location of customer
Gender	Gender of customer
Tenure_Months	Tenure in months
Used_Coupon_Count	Total number of coupons used
Not_Use_Coupon_Count	Total number of coupons not used
Clicked_Coupon_Count	Total number of coupon clicked
Revenue	Total revenue

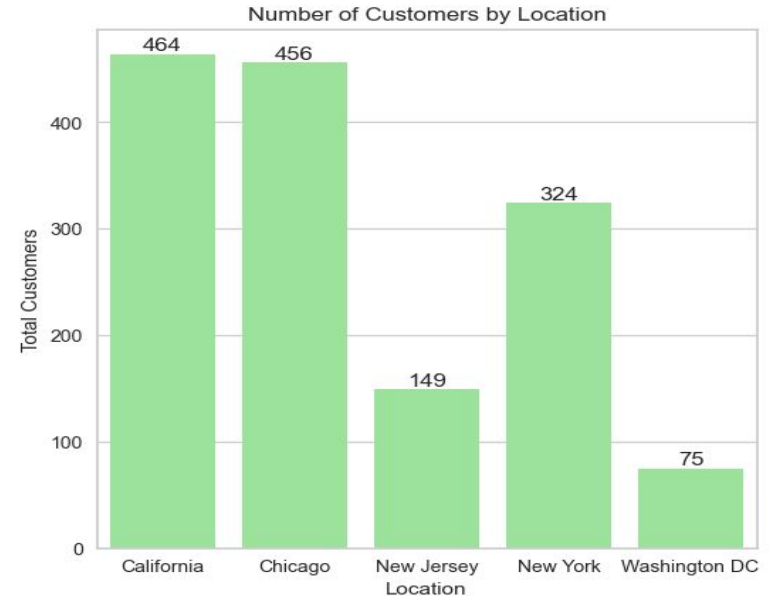
Product_Analysis	
CustomerID	Customer unique ID
Purchased_Products	List of total products purchased
Products_SKUS	Product unique ID
Product_Categories	Product category
Product_Group	Product Group



Exploratory Data Analysis

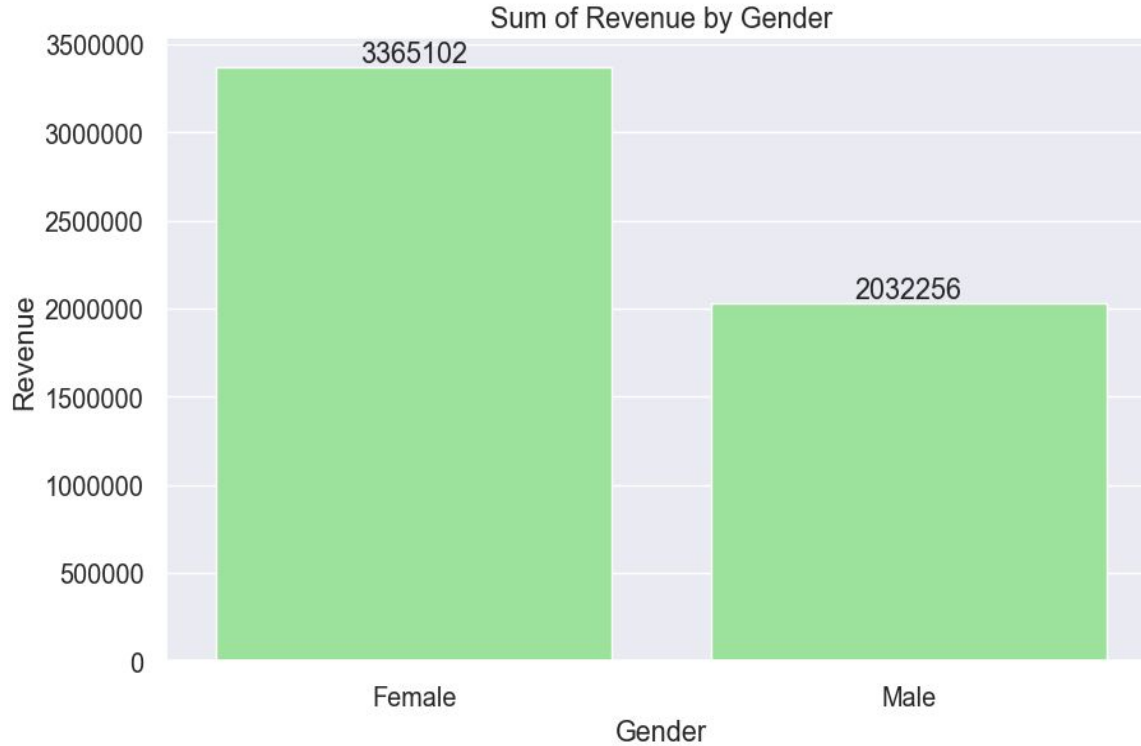


Revenue and Number of Customers per Location



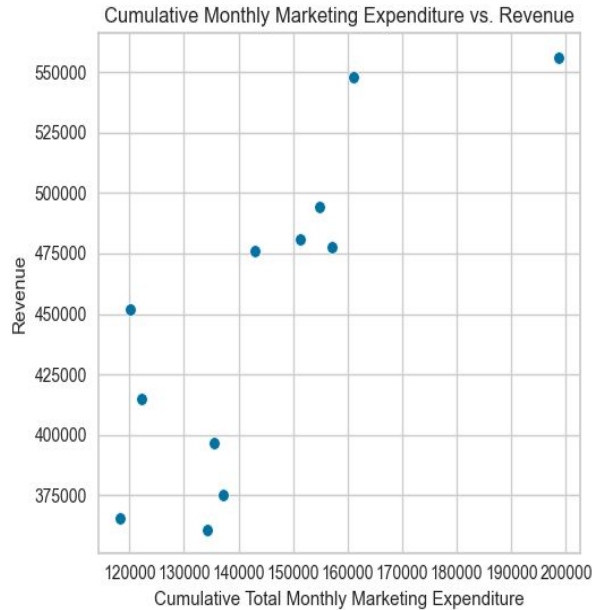
- Chicago, California, and New York are top 3 locations with highest revenue and number of customers
- New Jersey and Washington DC acquire lowest revenue and number of customers

Gender vs. Revenue

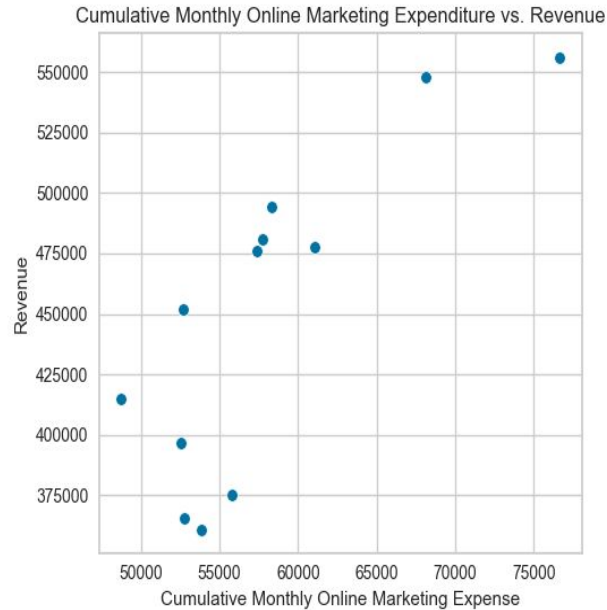


- Female spends higher than male

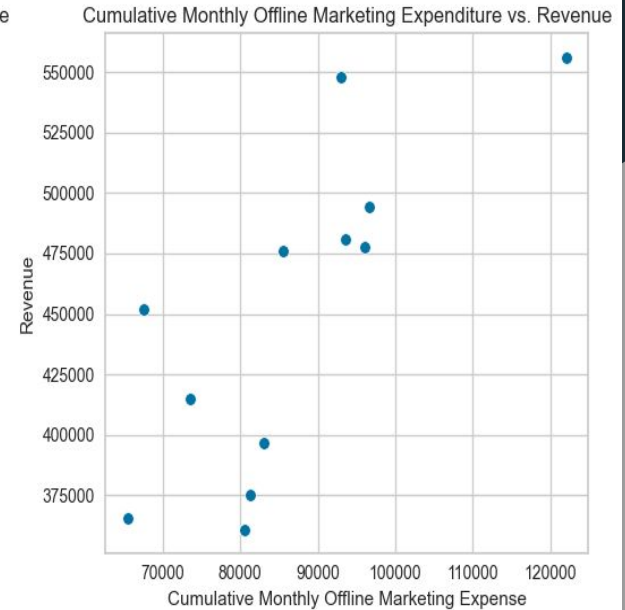
Marketing Expenditure vs. Revenue



Correlation: 0.796



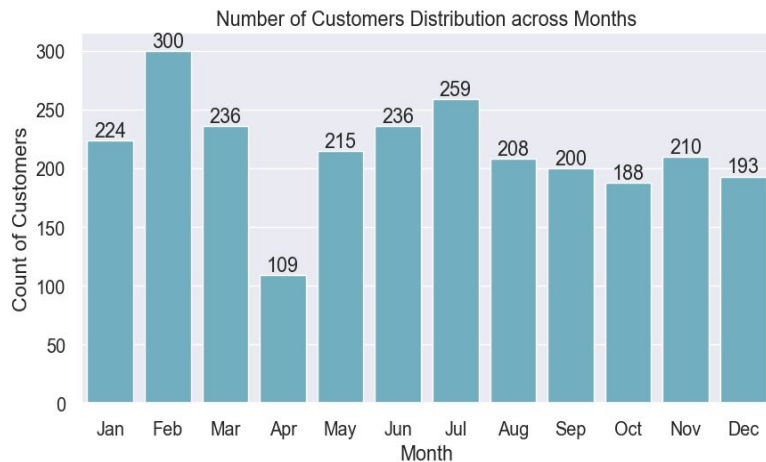
Correlation: 0.811



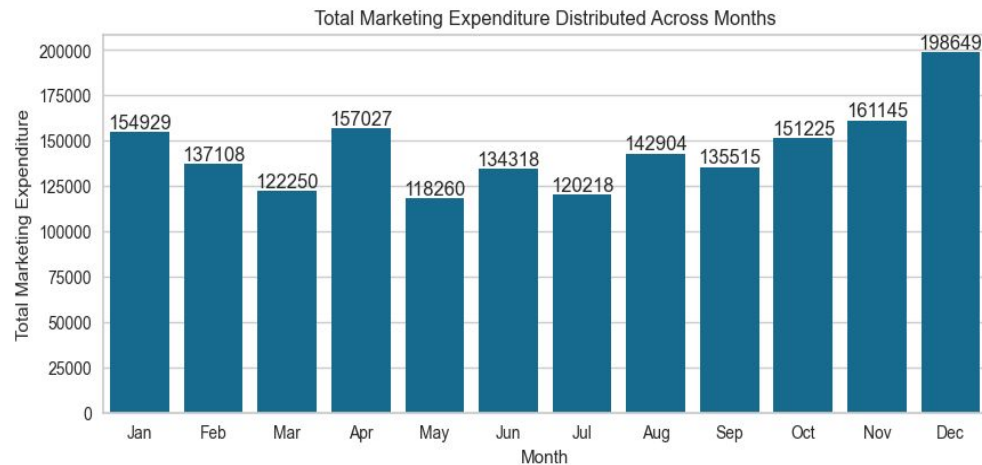
Correlation: 0.757

- High marketing expenditure both online and offline reflects high revenue obtained

Marketing Expenditure vs. Customers Distribution



- Top 2 months with highest customers: February and July
- April has lowest customers



- Top 2 months with highest marketing expenditure: November and December


Total Coupon Usage vs. Revenue



Correlation: 0.175

High coupons used
among customers did not
generate high revenue

Data Analysis Conclusion

- Allocated higher marketing expenditure to target low revenue and customers factors:
 - Location: New Jersey and Washington DC
 - Gender: Male
 - Months: April, November, and December Increase purchasing
- High investment in marketing yields higher revenue
- Releasing more coupons do not generate higher revenue



Feature Engineering

RFM Analysis: What is it and Why it is needed

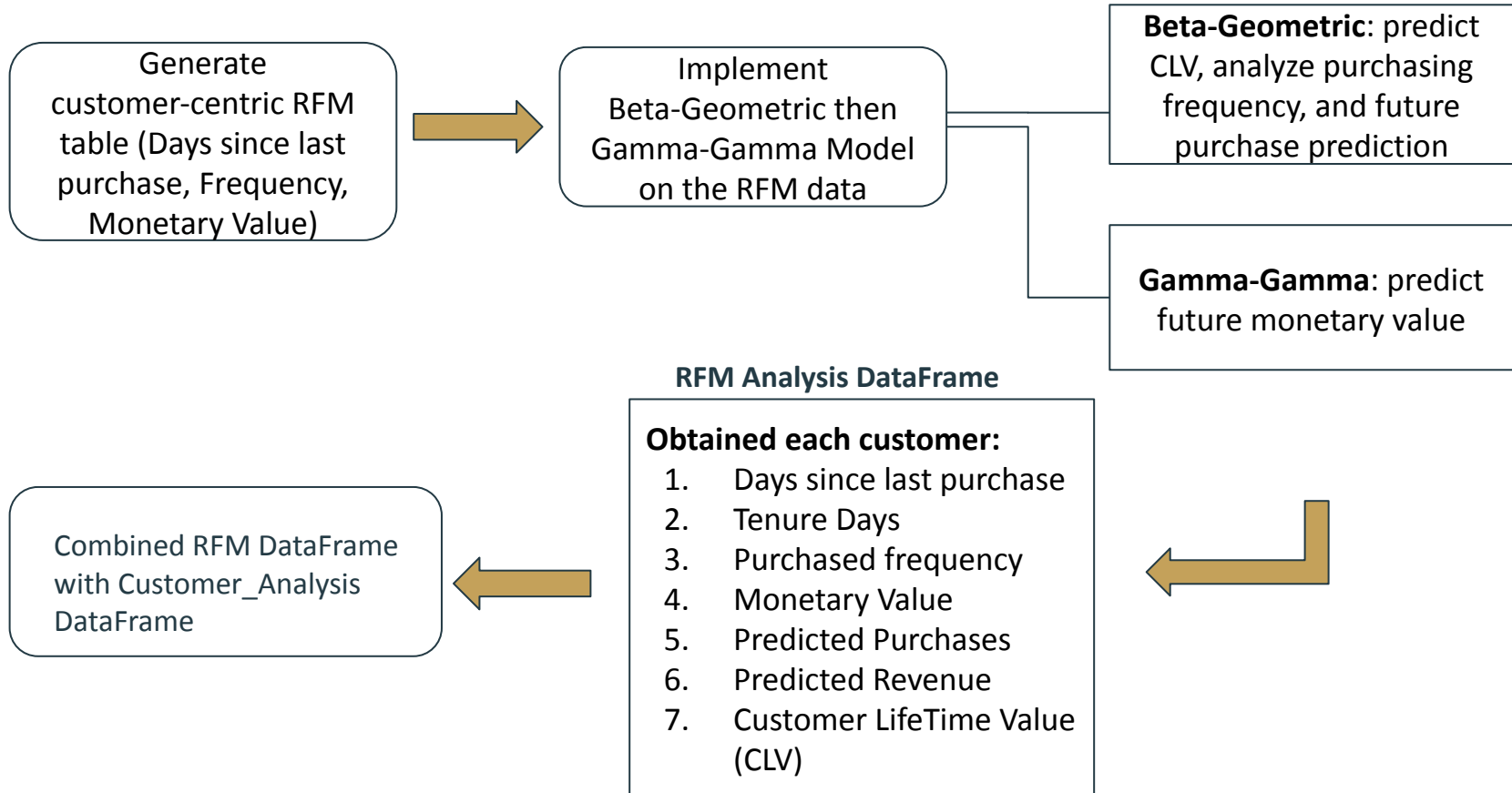
Definition of RFM analysis

- **Recency** (Days Since Last Purchase): The number of days between the customers last purchase and their first purchase. How many days does it take for customers to make the next purchase
- **Frequency**: How many days or how often during the observed period does a customer made purchases
 - Calculated based on the number of days customers made purchase
- **Monetary Value**: Total revenue gained from the customer during their tenure

Why RFM analysis is needed

- Obtained additional relevant features such as frequency and Days Since Last Purchase that can be included in the modeled data.
- Better understanding of customer purchasing behavior and future purchasing behavior, can lead to better segmentation of customers

Feature Engineering using RFM Analysis Procedure



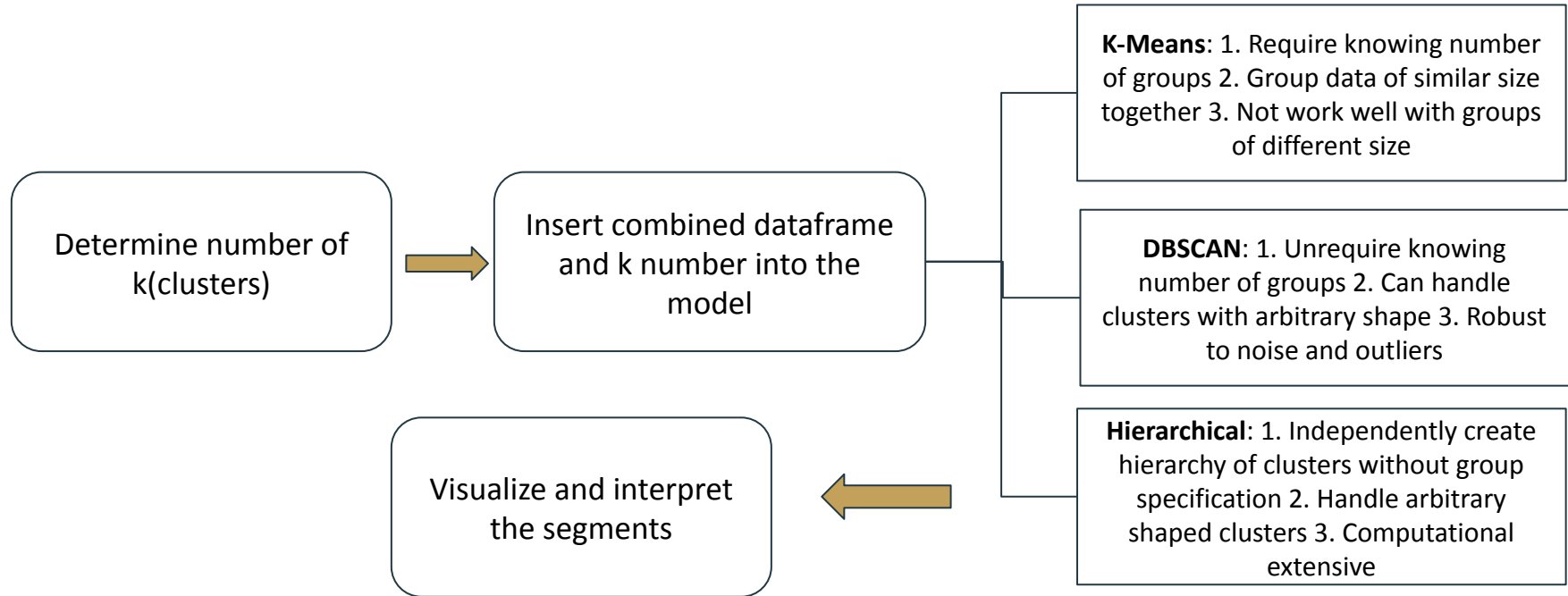
RFM Analysis

RFM Analysis		
Feature	Definition	Calculation
CustomerID	Unique customer ID	Last Purchase Date - Preceding Purchase Date
Days_Since_Last_Purchase	Number of days since customer last purchase	
Tenure_Days	Number of days customer have been with the store	
Purchased_Frequency	Number days customer made purchase	
Monetary_Value	Revenue gained from customer during their tenure	
Predicted_Purchases	Expected number of future purchases of the customer	recency (Days_Since_Last_Purchase) and frequency (historical transaction frequency) to estimate future purchase behavior.
Predicted_Revenue	Expected future revenue the customer is expected to generate	Predicted Purchases x Predicted Average Transaction Value
Customer LifeTime Value (CLV)	Total value a customer is expected to generate over their entire relationship with the store	Expected Number of Transactions x Predicted Average Transaction Value

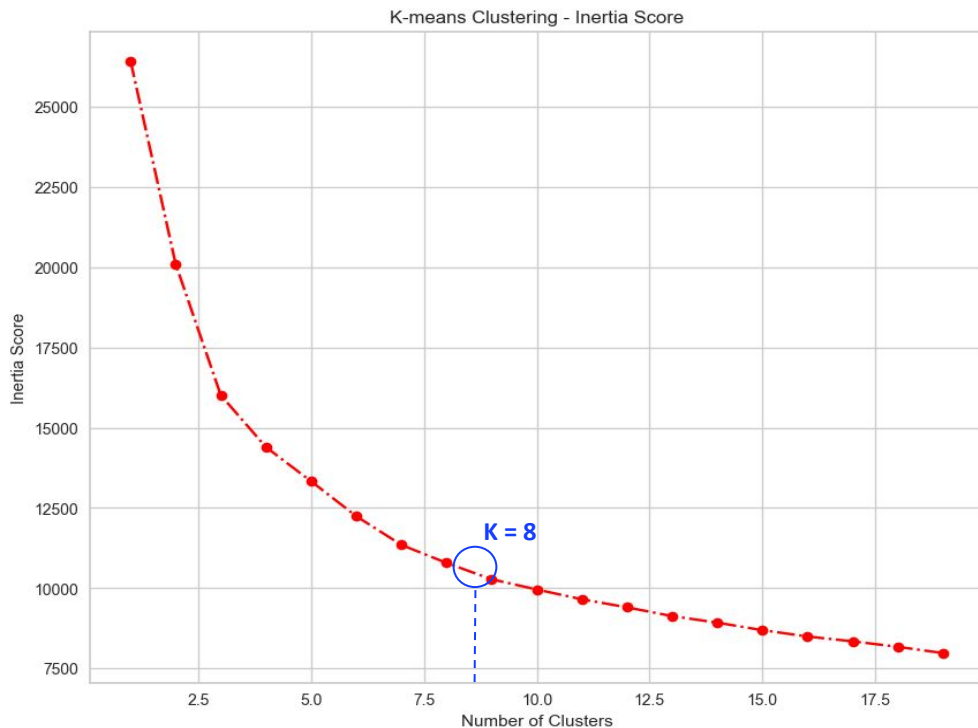


Modeling: Unsupervised Learning Clustering

Unsupervised Clustering Execution Process

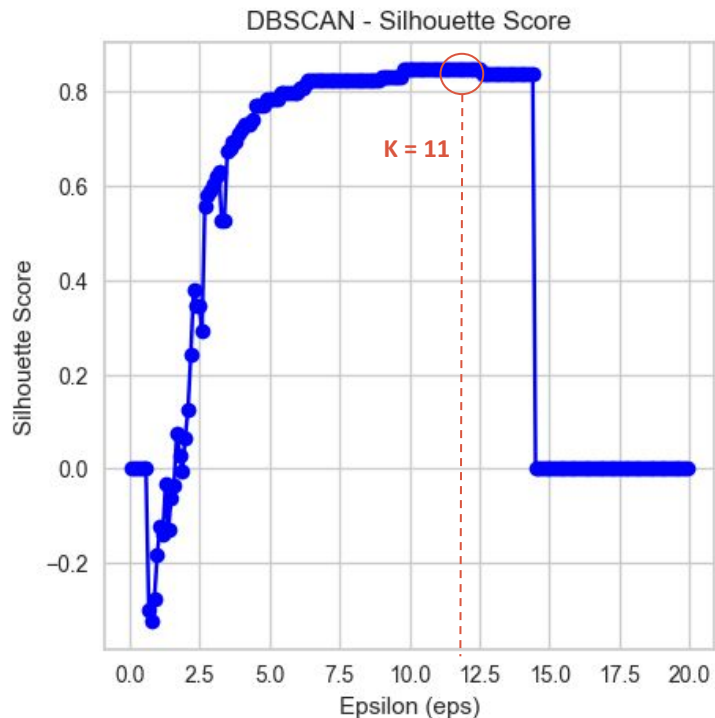


K-Means Clustering: Number of Clusters Evaluation



- Inertia Score Evaluation (Elbow method):
 - How compact together the data points are within each cluster
 - Elbow method: choose K when the decrease of Inertia score flattens out
 - Optimal number of clusters = 8

DBSCAN Clustering: Number of Clusters Evaluation



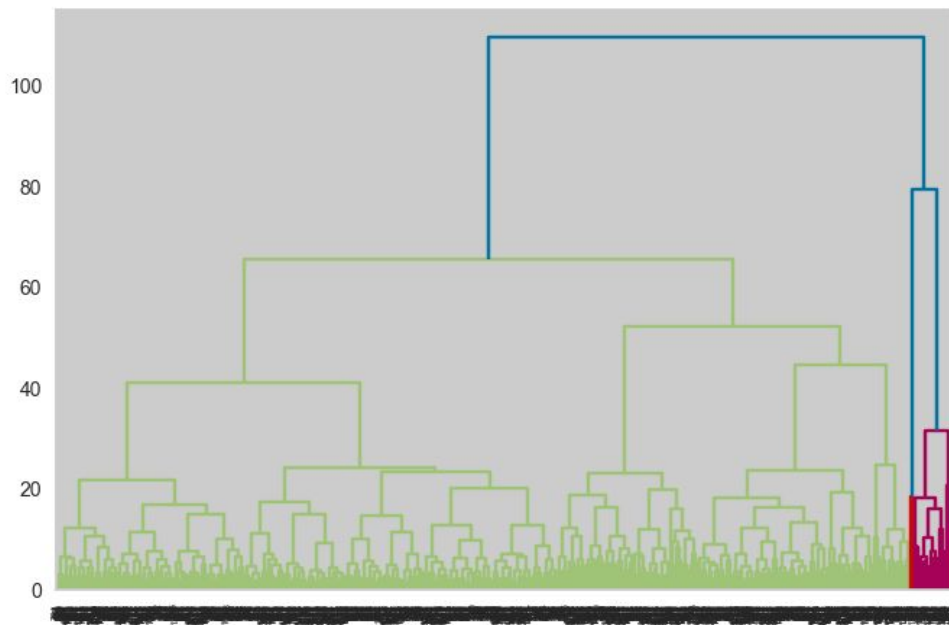
Insert into the model



```
array([-1, 0])
```

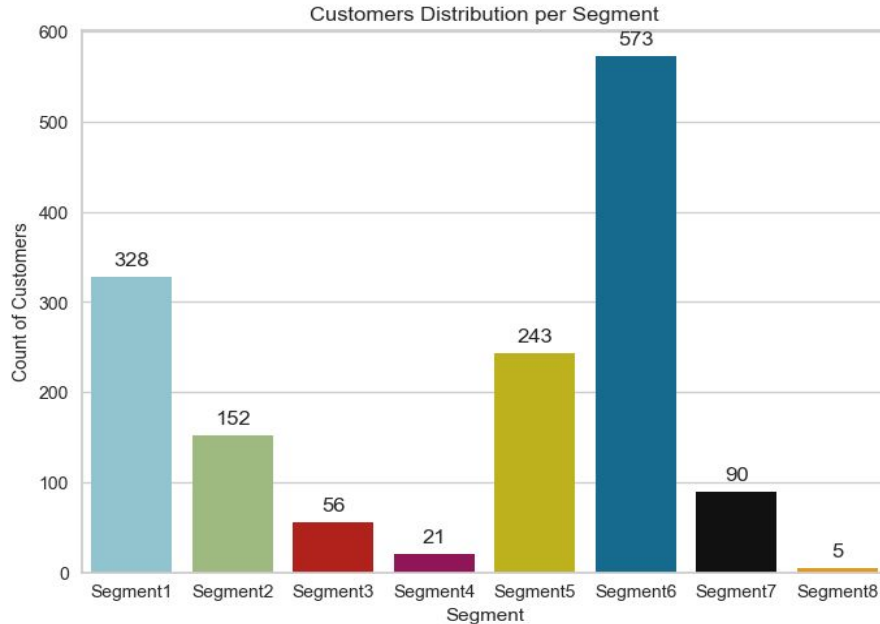
- Silhouette Score Evaluation:
 - How similar an object is to its own cluster (cohesion) compared to other clusters (separation)
 - Choose the epsilon with highest k
 - Output: 2 outcome labels(groups) with negative value -> clustering failed

Hierarchical Clustering: Number of Clusters Evaluation



- Hierarchical Clustering:
 - “Ward” method of linkage minimizes the variance when merging clusters and is known for producing compact and spherical clusters.
 - Outcome: 3 clusters

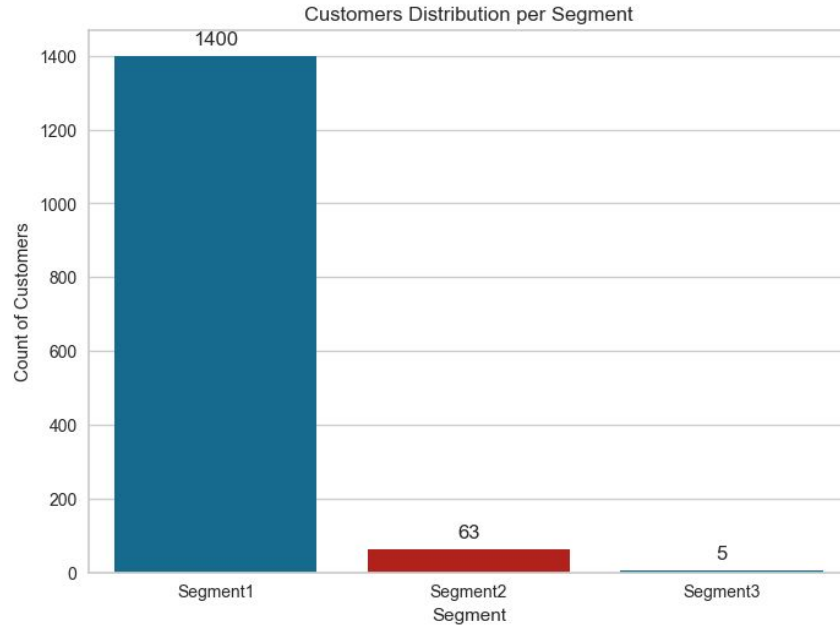
K-Means: Label Outcome Evaluation



Data Distribution: Proportionate data distribution

- More customers distributed in Segment1, Segment2, Segment5, and Segment6
- Targeting these segments maximize potential return of marketing investment
 - Higher market reach and higher revenue capturing

Hierarchical: Label Outcome Evaluation



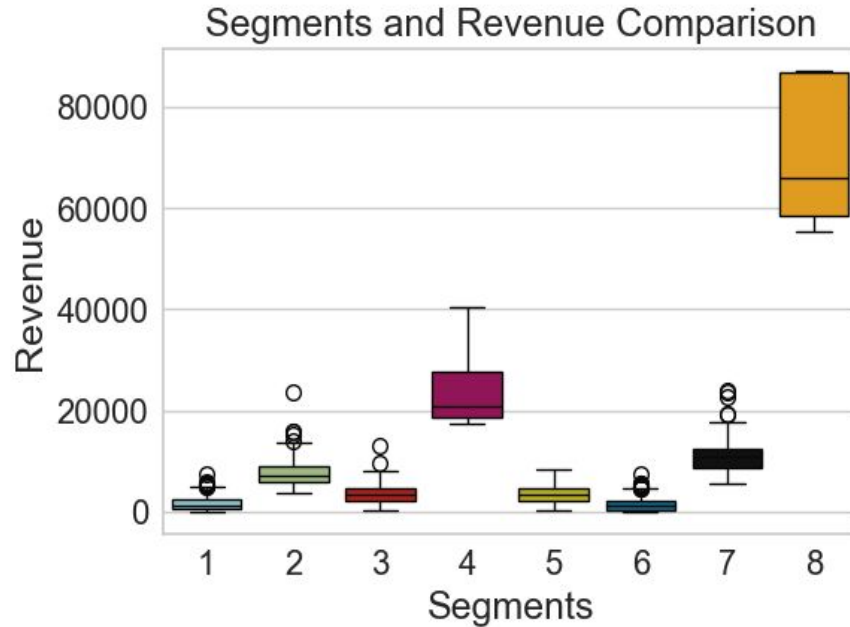
- Most customers are distributed in Segment1
- Imbalanced Data

Data Distribution: Proportionate data distribution



Chosen Clustering Method: K-Means

Segments and Revenue Comparison

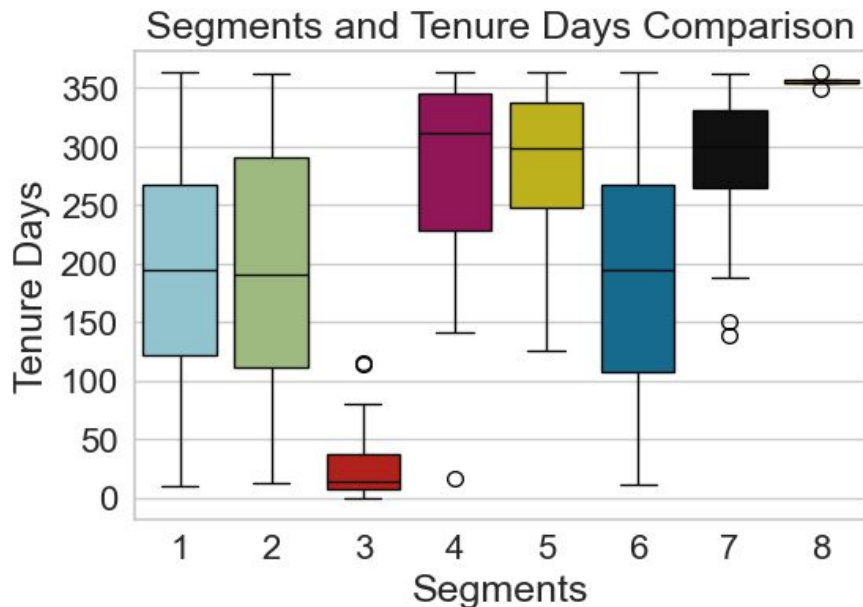


Mean Revenue by Segment

Segment	Revenue
Segment6	1382
Segment1	1573
Segment5	3475
Segment3	3549
Segment2	7773
Segment7	11211
Segment4	23863
Segment8	70779

Segment 6 and
Segment 1 generate
lowest revenue.

Segments and Tenure Days Comparison

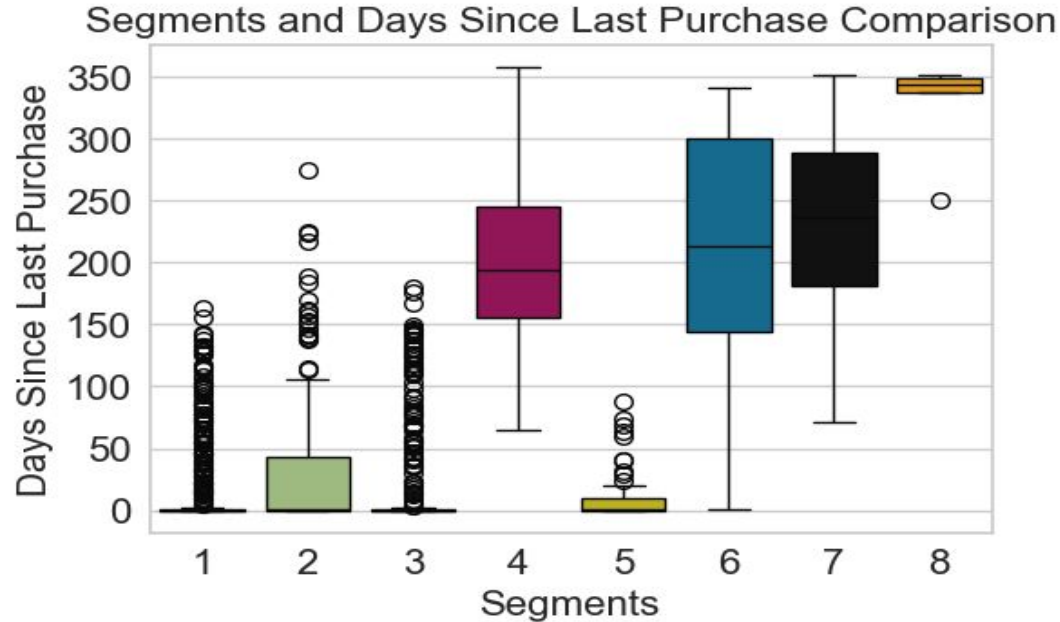


Mean Tenure Days by Segment

Segment	Tenure_Days
Segment3	26
Segment6	188
Segment1	192
Segment2	198
Segment4	280
Segment5	287
Segment7	293
Segment8	355

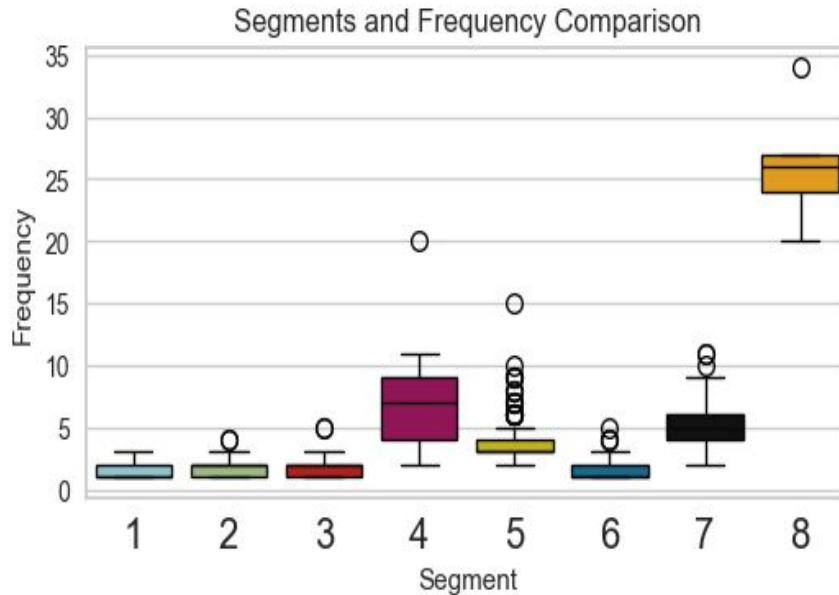
Segment1, 3, and 6 tenure days (in business with the store) are the lowest.

Segments and Days Since Last Purchase Comparison



- Days since last purchase = 0 in Segment 1, 2, 3 and 5 represent customers that purchase only once
- Segment 6 has highest average duration between 2 purchases

Segments and Frequency Comparison



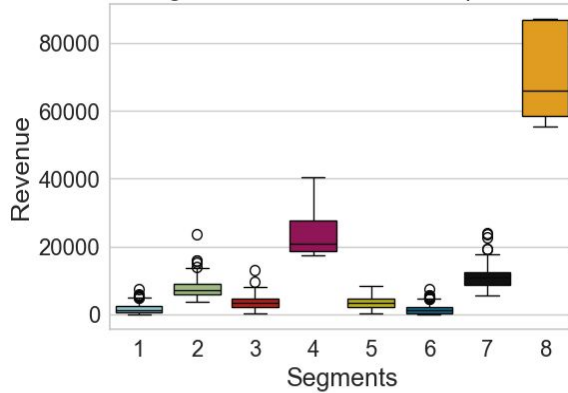
Mean Frequency by Segment

	Segment	frequency
5	Segment6	1.301920
0	Segment1	1.365854
1	Segment2	1.750000
2	Segment3	1.857143
4	Segment5	3.736626
6	Segment7	5.055556
3	Segment4	7.142857
7	Segment8	26.200000

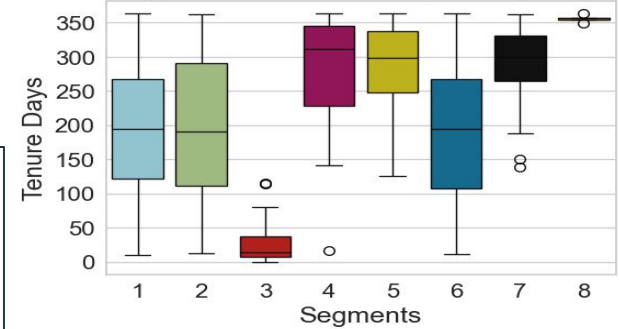
Segment 1 and Segment 6 are the lowest frequently purchase groups

K-Means Clustering Segment Evaluation

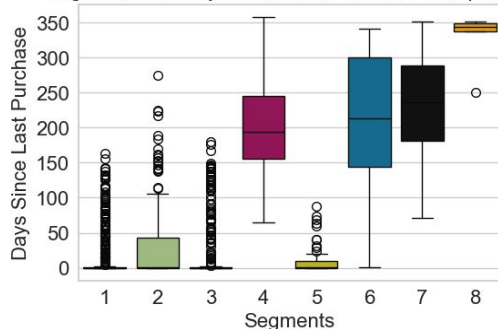
Segments and Revenue Comparison



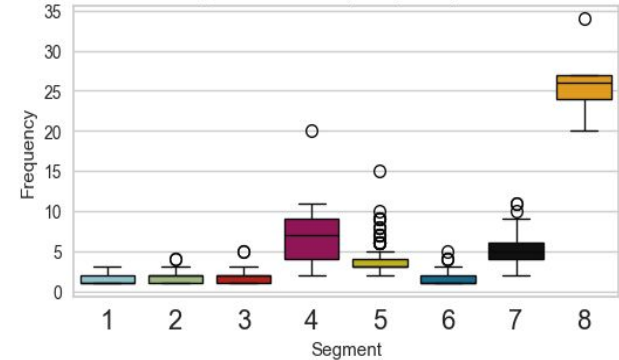
Segments and Tenure Days Comparison



Segments and Days Since Last Purchase Comparison



Segments and Frequency Comparison



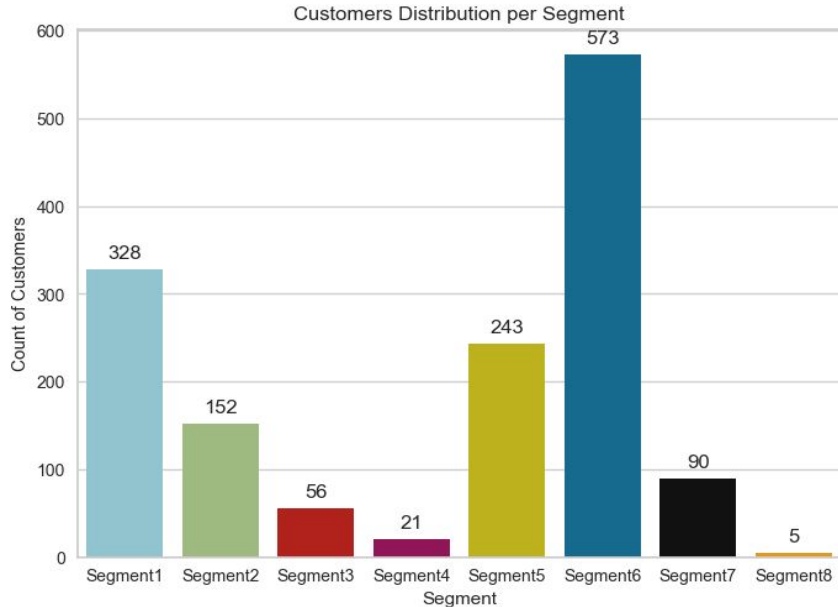
- **Focus Low Valued Segments**
Segment1 and Segment6:
 - 1. Lowest revenue
 - 2. Low tenure days
 - 3. Majority made purchase only once
 - 4. Low purchase frequency



Recommendation

Segment to Focus Marketing Expenditure

- Among 8 segments : Target Segment 1 and Segment 6



Segment 1 and Segment 6

- **Low-valued segments**
- Majority of customers resided in Chicago, California, and New York
- 60% of customers are female
- Top purchased product group is Google
- Top 5 product categories: Office, Apparel, Drinkware, Lifestyle, and Nest
- Top purchased products: Maze Pen, Google 22 oz Water Bottle, Google Sunglasses, Sport Bag, Ballpoint LED Light Pen, Google Kick Ball, Foam Can and Bottle Cooler

Marketing Strategy 1: Targeted Campaign Email Marketing

- Targeted campaign email marketing - email marketing for specific audiences
 - Specified the customers emails as Segment 1 and Segment 6
 - Email content:
 - Re-engage customers with their top product categories: Office, Drinkware, Apparel and top product group: Google
 - Promoting unpopular product categories such as Accessories, Google Waze, Gift Cards and unpopular product group: Youtube, using special promotions such as price reduction
 - Bundle deals
 - Increase email marketing sent out in November and December for Black Friday and Christmas event, and April as the lowest number of customers month
 - Allocate higher email marketing resource to specific demographics to increase awarenesses and purchases:
 - Gender: Male
 - Location: New Jersey and Washington DC

Marketing Strategy 2 : Product Bundling

Product 1	Product 2	Support	Confidence	Lift
Nest Cam Indoor Security Camera - USA	Nest Cam Outdoor Security Camera - USA	0.34	0.73	1.51
Nest Protect Smoke + CO White Battery Alarm-USA	Nest Learning Thermostat 3rd Gen-USA	0.21	0.75	1.57
Google Tee Green	Google Tee Blue	0.01	0.4	17.82

Two products bundle based on historical purchasing data of two products frequently purchased together

- Support: Frequency of purchasing the product: total purchase of A / total transaction
- Confidence: Likelihood of purchasing this item when the other item is purchased: $\text{support}(A, B) / \text{support}(A)$
- Lift: The likelihood of purchasing both product compared to purchasing individually: $\text{support}(A, B) / \text{support}(A) * \text{support}(B)$

Marketing Strategy 3 : Product Recommendation

Product_SKU	Product_Name
GGOEAFKQ020599	Purchased Android Sticker Sheet Ultra Removable
GGOEAFKQ020499	8 pc Android Sticker Sheet
GGOEGHGR019499	Google Sunglasses
GGOEGHPB071610	Google Twill Cap
GGOEGDHC018299	Google 22 oz Water Bottle
GGOEGHGH019699	Google Sunglasses
GGOEAKDH019899	Windup Android
GGOEGAAH034015	Google Men's Vintage Badge Tee Sage
GGOEGHGC019799	Google Sunglasses
GGOEGFKQ020799	Google Doodle Decal

Product_SKU	Product_Name
GGOEGFKQ020799	Purchased Google Doodle Decal
GGOEGHGC019799	Google Sunglasses
GGOEGHPB071610	Google Twill Cap
GGOEAFKQ020499	8 pc Android Sticker Sheet
GGOEAFKQ020599	Android Sticker Sheet Ultra Removable
GGOEGHGR019499	Google Sunglasses
GGOEGAAH034015	Google Men's Vintage Badge Tee Sage
GGOEGDHC018299	Google 22 oz Water Bottle
GGOENEBQ078999	Nest Cam Outdoor Security Camera - USA
GGOEGHGH019699	Google Sunglasses

Product_SKU	Product_Name
GGOEAFKQ020599	Purchased Android Sticker Sheet Ultra Removable
GGOEAFKQ020499	8 pc Android Sticker Sheet
GGOEGHGR019499	Google Sunglasses
GGOEGHPB071610	Google Twill Cap
GGOEGDHC018299	Google 22 oz Water Bottle
GGOEGHGH019699	Google Sunglasses
GGOEAKDH019899	Windup Android
GGOEGAAH034015	Google Men's Vintage Badge Tee Sage
GGOEGHGC019799	Google Sunglasses
GGOEGFKQ020799	Google Doodle Decal

Top 9 rank similar product name based on Cosine Similarity Score rank descending

Limitations

- Data from Kaggle might not be valid
- Limited relevant features of customers and limited data range: only 1 year
- Undefined company's resources to determine the potential return of marketing investment



Thank you