**Executive Summary**

Acme Inc has been struggling to effectively target their marketings, leading to wasted resources and low conversion rates. To tackle this problem, we:

* deep-dived into data given to get familiar with it and find insights.
* tested different models and did market research to find the best model to segment the targets.
* tested different models including:
* fine-tuned the model, which is GMM model, to produce the best result.
* developed prototype recommender system based the segmentation.

**Business Statement**

Acme inc. is a clothing company. It has been struggling to create effective customer segments to execute their marketing efforts effectively. It suffered wasted resource and low conversion rates. Its marketing team’s research estimated that effective personalization can boost conversion by 15%. The data science team is leading the charge of developing such strategy and methodology to create the system.

**Data Overview and EDA**

**Data Glossary**

There are 3900 entries and 18 columns.

Customer ID:A unique identifier assigned to each individual customer, facilitating tracking and analysis of their shopping behavior over time.  
Age: The age of the customer, providing demographic information for segmentation and targeted marketing strategies.  
Gender: The gender identification of the customer, a key demographic variable influencing product preferences and purchasing patterns.  
Item Purchased: The specific product or item selected by the customer during the transaction.  
Category: The broad classification or group to which the purchased item belongs (e.g., clothing, electronics, groceries).  
Purchase Amount (USD): The monetary value of the transaction, denoted in United States Dollars (USD), indicates the cost of the purchased item(s).  
Location: The geographical location where the purchase was made, offering insights into regional preferences and market trends.  
Size: The size specification (if applicable) of the purchased item, relevant for apparel, footwear, and certain consumer goods.  
Color: The color variant or choice associated with the purchased item, influencing customer preferences and product availability.  
Season: The seasonal relevance of the purchased item (e.g., spring, summer, fall, winter), impacting inventory management and marketing strategies.  
Review Rating: A numerical or qualitative assessment provided by the customer regarding their satisfaction with the purchased item.  
Subscription Status: Indicates whether the customer has opted for a subscription service, offering insights into their level of loyalty and potential for recurring revenue.  
Shipping Type: Specifies the method used to deliver the purchased item (e.g., standard shipping, express delivery), influencing delivery times and costs.  
Discount Applied: Indicates if any promotional discounts were applied to the purchase, shedding light on price sensitivity and promotion effectiveness.  
Promo Code Used: Notes whether a promotional code or coupon was utilized during the transaction, aiding in the evaluation of marketing campaign success.  
Previous Purchases: Days passed from the last purchase  
Payment Method: Specifies the mode of payment employed by the customer (e.g., credit card, cash), offering insights into preferred payment options.  
Frequency of Purchases: Indicates how frequently a customer purchases items.

**Numerical Variables Analysis**

|  | **count** | **mean** | **std** | **min** | **25%** | **50%** | **75%** | **max** |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **Customer ID** | 3900 | 1950.50 | 1125.98 | 1.0 | 975.75 | 1950.5 | 2925.25 | 3900.0 |
| **Age** | 3900 | 44.07 | 15.21 | 18.0 | 31.00 | 44.0 | 57.00 | 70.0 |
| **Purchase Amount (USD)** | 3900 | 59.76 | 23.69 | 20.0 | 39.00 | 60.0 | 81.00 | 100.0 |
| **Review Rating** | 3900 | 3.75 | 0.72 | 2.5 | 3.10 | 3.7 | 4.40 | 5.0 |
| **Previous Purchases** | 3900 | 25.35 | 14.45 | 1.0 | 13.00 | 25.0 | 38.00 | 50.0 |

we can see that the distribution of the numerical variables. Age, Previous Purchases, and Purchase Amount (USD) seem to be evenly distributed. A graph with different colored squares

AI-generated content may be incorrect.

Figure The distribution of Age, Purchase Amount(USD), Previous Purchases

**Categorical Variables**

* Gender: Disproportionately male. It’s surprising because top item purchased are associated women.
* Category: Most popular category is clothing.
* Size: Size ‘M’ is most popular.
* Season: Evenly distributed.
* Subscription Status: Most customers don’t have subscriptions.
* Discount Applied and Promo Code Used: They are identical. We don’t need one or another.
* Payment Method: Doesn’t seem to be any particular pattern.
* Frequency of Purchases: Every 3 Months seems to be most frequent. But this is misleading because Every 3 Months and Quarterly are essentially the same. Same goes to Bi-weekly and Fortnightly. With that consideration, Quarterly is the most frequent value followed by Bi-weekly.

A chart of different colored squares

AI-generated content may be incorrect.

Figure Distribution of categorical variables

* Item Puchased: Blouse, Jewerly, and Pants are the most popular items.
* Location: Montana is the most frequent value. It’s surprising because Montana is a relatively small state. The data might have been sampled with strata.

A graph of different colored bars

AI-generated content may be incorrect.

Figure Distribution of Item Purchased, and Location

**Feature Engineering**

1. Frequency of Purchase: This indicates how frequently a customer shop on the website. We turned this into numerical scale per year. For instance Weekly is 52 as there are 52 week in a year. In a same manner, Monthly would be 12.
2. Loyalty\_Score: A combination of metrics. Frequency + Recency + Subscription\_Status \* 10.

**Models Tested**

**Model Selected**

A graph of different colored bars

AI-generated content may be incorrect.

Figure Silhouette Score Results

Based on the silhouette scores, we concluded that the clusters that are created with GMM clustering presents the best result. The followings are the characteristics of each cluster.

A diagram of a group of boxes

AI-generated content may be incorrect.

A chart of a cluster of data

AI-generated content may be incorrect.

A diagram of a distribution across clusters

AI-generated content may be incorrect.

A chart of different colored squares

AI-generated content may be incorrect.

**Cluster Summary**

**Cluster 0: Premium High-Margin Shoppers**

Average Spending: $82 per transaction (82% of maximum possible spend)  
Purchase pattern: Buy expensive items approximately once every 25-26 days  
Engagement: Exceptionally high (4.3/5.0) - actively interacting with brand touchpoints  
Business value: Likely responsible for 30-40% of total profit despite lower purchase frequency  
Real-world example: Luxury shoppers who carefully select high-end items rather than making frequent purchases

**Cluster 1: Steady Mid-Tier Customers**

Average Spending: $57 per transaction

Purchase pattern: Shop approximately every 2-3 weeks (15 times per period)

Recency: Last purchased about 30 days ago

Engagement: Good but not exceptional (3.7/5.0)

Real-world example: Regular customers who reliably purchase mid-priced items on a predictable schedule

**Cluster 2: Enthusiast Power Users**

Average Spending: $68 per transaction  
Purchase pattern: Most frequent shoppers at 38 purchases (approaching weekly visits)  
Recency: Longest time since last purchase (39 days) - potential risk of churn  
Engagement: Very high (4.2/5.0) - brand enthusiasts  
Real-world example: Passionate hobbyists or product category enthusiasts who frequently buy related items

**Cluster 3: Recent New/Reactivated Customers**

Average Spending: $51 per transaction  
Purchase pattern: Infrequent (8-9 purchases total)  
Recency: Most recent shoppers (purchased within the last 13 days)  
Engagement: Moderate (3.4/5.0)  
Real-world example: Newly acquired customers or recently reactivated dormant customers who haven't established a regular pattern yet

**Cluster 4: Budget-Conscious Frequent Shoppers**

Average Spending: $40 per transaction (lowest spending group)  
Purchase pattern: Very frequent (35 purchases) - about every 10 days  
Recency: Moderately recent (29 days since last purchase)  
Engagement: Lowest of all groups (3.2/5.0)  
Real-world example: Price-sensitive frequent shoppers who make smaller purchases, likely deal-hunters or necessity-based shoppers

**Conclusion / Recommendation**

From the exercise we explored different ways to segment our customer. With the model we created, we ended up with 5 distinctly different segments. Premium High-Margin Shoppers, Steady Mid-Tier Customers, Enthusiast Power Users, Recent New/Reactivated Customers, and Budget-Conscious Frequent Shoppers. Based on this information, we were able to build a prototype recommender system that recommends items based on what other people purchase in the cluster that they are assigned to. The data provided seems to be a snapshot data. With historic data, we can enrich our recommender system even further and personalized.

A screenshot of a product recommendation service

AI-generated content may be incorrect.

Figure Recommender System Prototype