Machine Learning

Customer Segmentation

Group 10

Meet The Team

Henry
20222002

Abdul
20231738

Yan
20222004

Executive Summary

As artificial intelligence continues to expand across industries, so does the demand for its integration into enterprise solutions. Businesses that fail to leverage AI for critical tasks such as optimizing marketing campaigns and understanding customer behavior risk falling behind in an increasingly competitive landscape. This report directly addresses that challenge by demonstrating how advanced analytics can support smarter, data-driven business decisions.

This report presents a comprehensive customer segmentation analysis aimed at identifying distinct customer groups within the provided retail store dataset. To gain insights into the diversity of customer behaviors and characteristics, the data set provided contains a plethora of varying demographic information, spending patterns, and purchase history. Our methodology involved in-depth exploratory data analysis, the application of advanced clustering techniques, and the discovery of association rules within each segment.

Exploratory Data Analysis

This section details the initial investigation into our customer datasets. Through comprehensive data profiling, descriptive statistics, and targeted visualizations, we aimed to understand data characteristics, identify missing values and outliers, and uncover preliminary patterns that would guide our feature engineering and clustering strategies.

Data Sources

Customer Info

Columns: **25**
Rows: **34,060**
Contains:

- **Customer ID**
- **Location Data**
- **Kids Home**
- **Birth Dates**
- **Lifetime Spend Per**
- **Product**
- **Loyalty Card Info**
- **Typical Hours**
- **First Transaction Year**
- **% Products Bought On Promotion**

Customer Basket

Columns: **3**
Rows: **100,000**
Contains:

- **Customer ID**
- **Invoice ID**
- **Items Purchased**

The datasets provided contained a lot of meaningful transactional and personal data that will help us in our analysis and clustering later. However, we did still discover some outliers and columns that needed to be altered to maximize our results, next we will outline the preprocessing steps we took to ready our data for the next steps.

Data Cleaning And Preprocessing

The changes we made to the data are as follows:

Augmentation

Customer Birth Date -> **Customer Age**

Kids Home, Teens Home -> **Total Children Home, No Children**

Typical Hours -> **Morning Shopper**
Afternoon Shopper
Evening Shopper

First Transaction Year -> **Customer Duration**

Degree In Name -> **Ordinal Degree Level**

List Of Goods -> All Purchased Items

Loyalty Card -> Dummy 1/0 Variable

Cleaning

Columns we dropped or didn't include due to Irrelevancy or being redundant after additions:

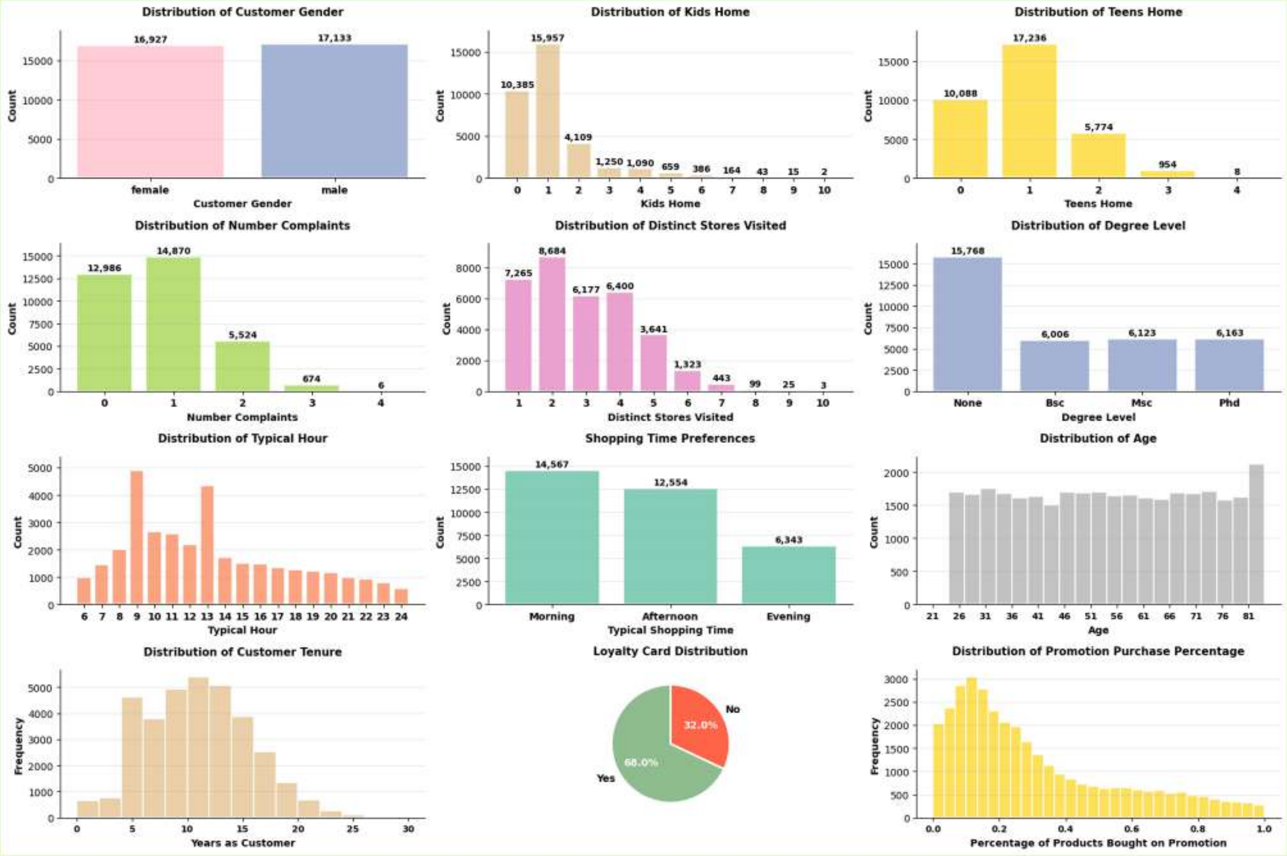
- **Customer Birthdate**
- **Unnamed: 0**
- **Customer Name**
- **Customer For**
  Customer Gender
- **Customer Location**

We also filled our missing values via KNN imputation and fixed some columns as follows:

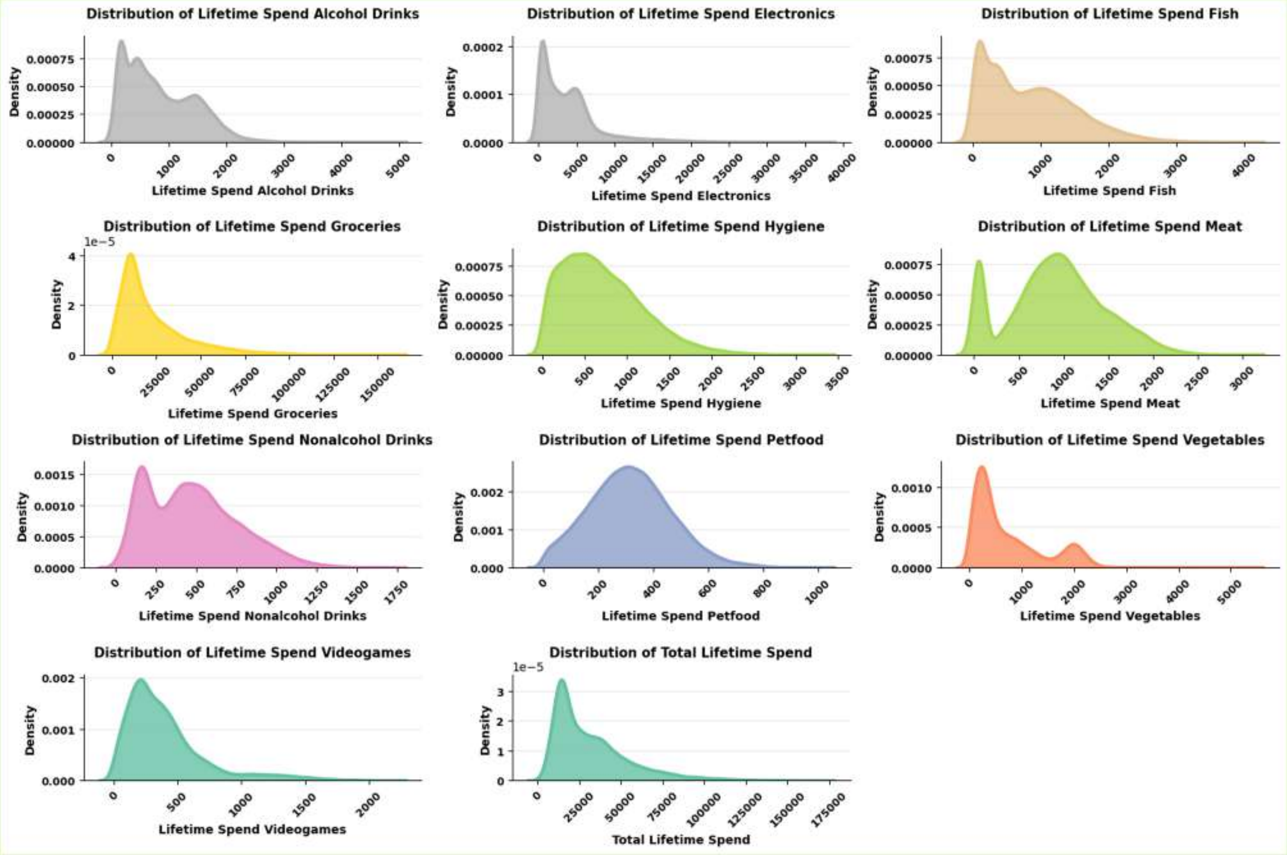
- **Taking the absolute value of the negative entries**
- **Ensuring that percentage values are < 100%**
- **Used Robust Scaling**

Visualizations

General Data Visualizations

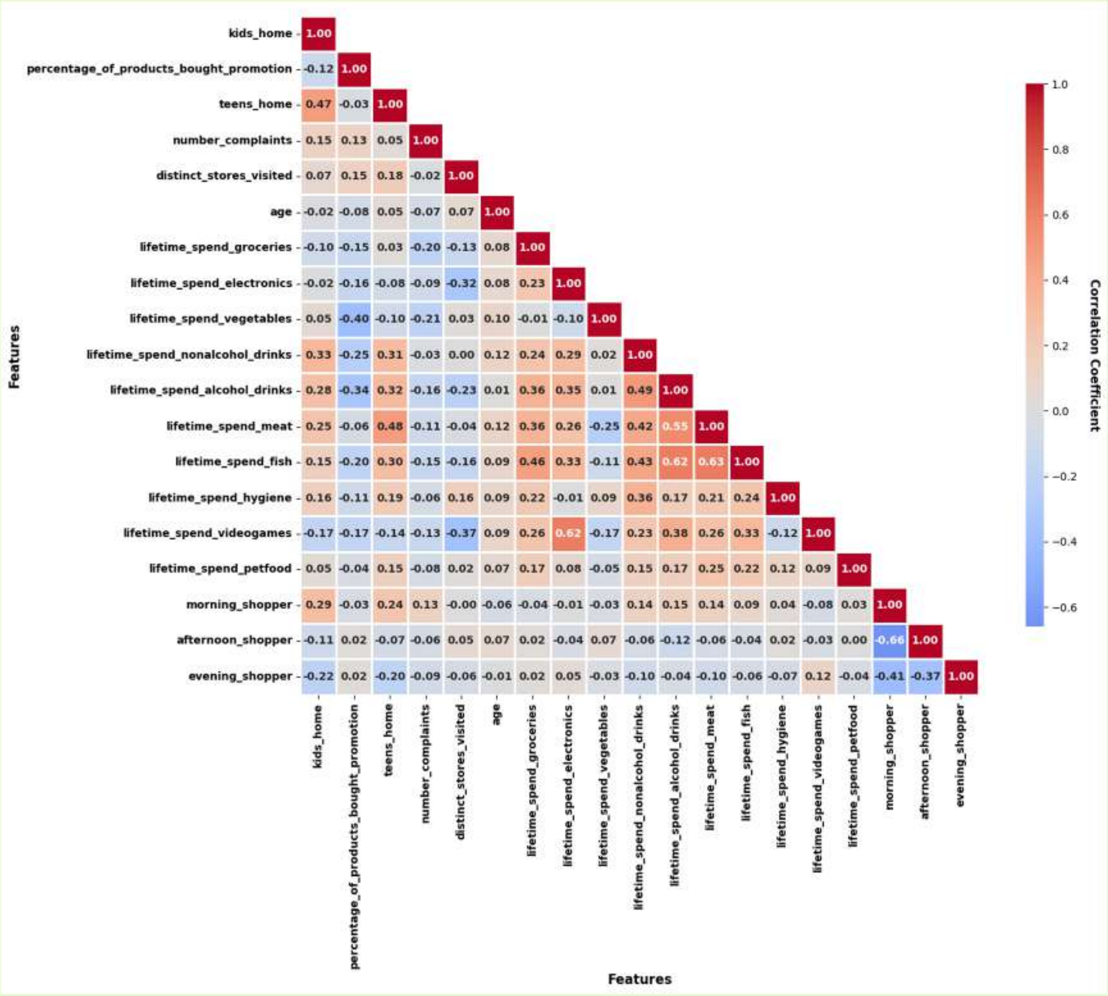


Customer Spending Distributions

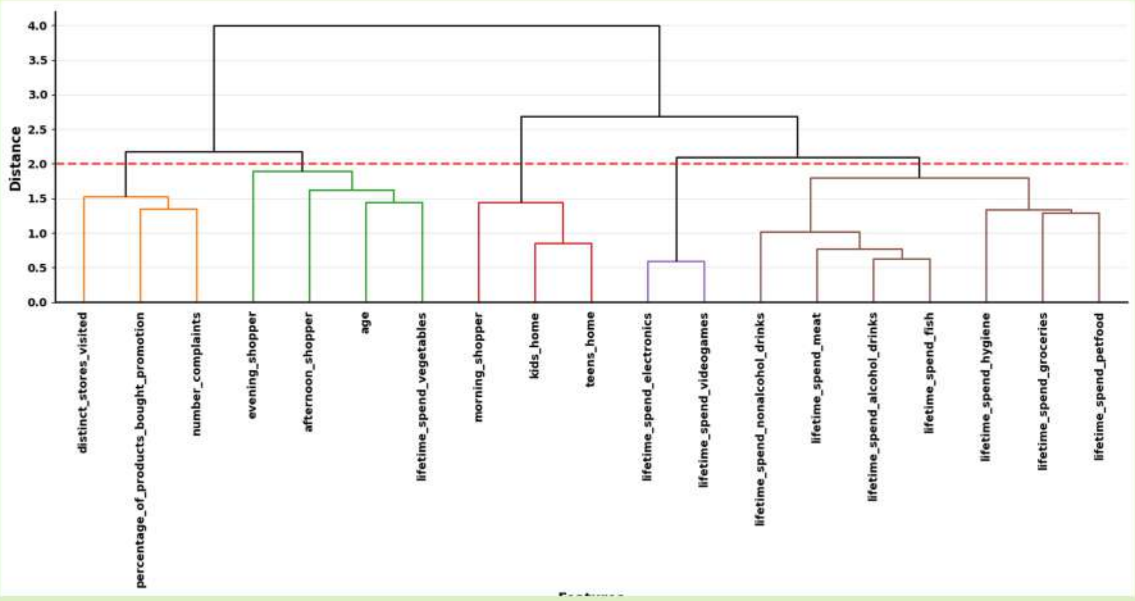


Visualizations

Feature Correlation Heatmap

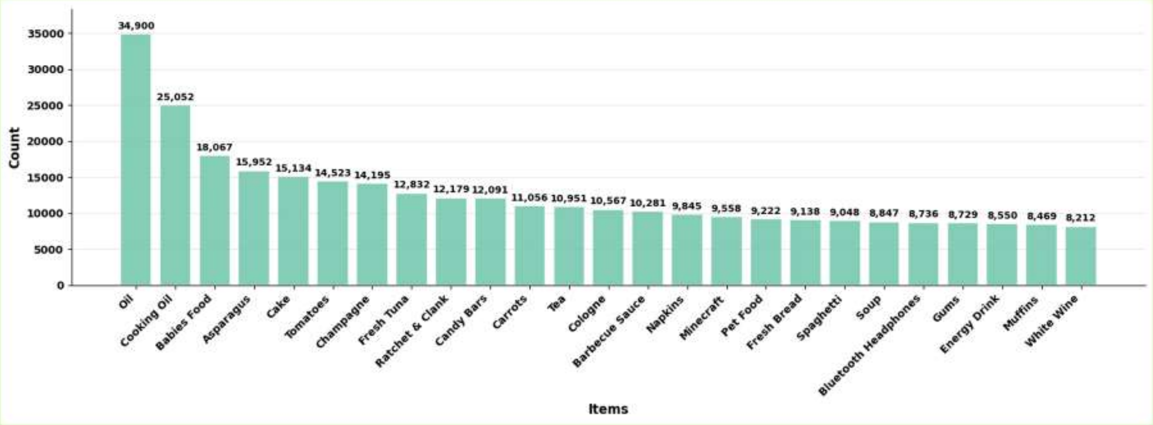


Dendrogram Of Numeric Features

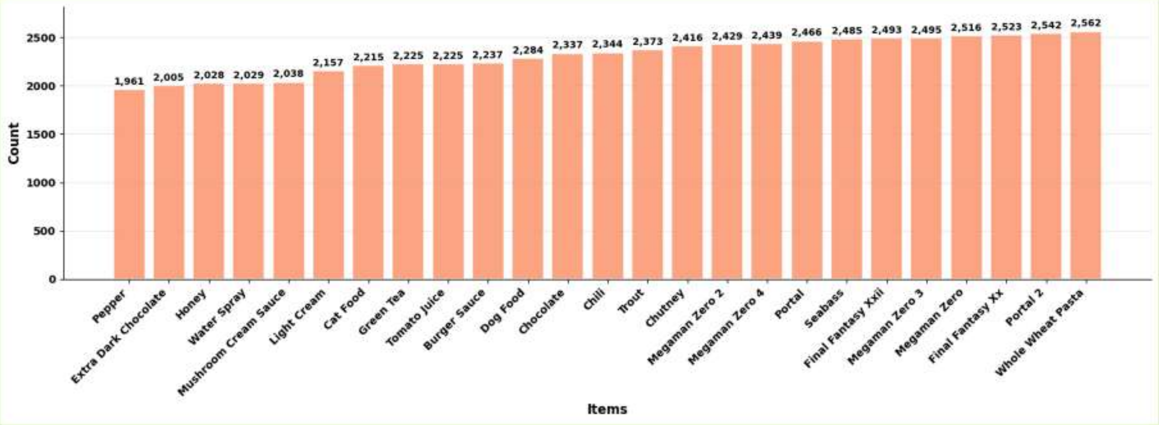


Visualizations

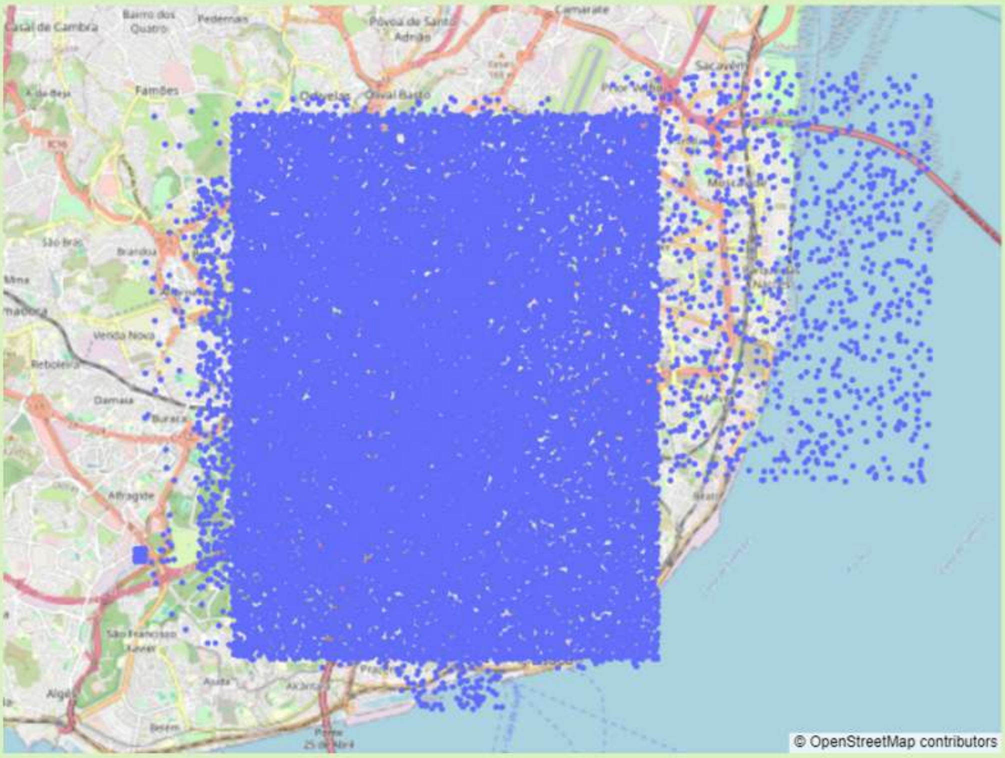
Most Purchsed Items



Least Purchsed Items



Map Plot



Visualizations

interpretations

General Plot

Gender Is Equally Balanced
More People With Children Than Without
More People With Teens Vs Kids
Similar Number Of People With Degree Vs Not
Peaks In Hours 9:00 And 13:00
Evening Shopping A Lot Less Frequent
Vast Majority Of Customers Have Been Customers For 5+ Years
Equally Balanced Age

Geographical Visualization

We see that there are 2 square plots, one around the university which indicates to a student cluster, and one on the left around the makro store, indicating another cluster filled with potential business owners, which is backed up by the spending habits of the people in these area, which is why we extracted them as a cluster.

Dendrogram Plot

distinct_stores_visited, promotion_percentage, number_complaints cluster together.

evening/afternoon_shopper, age, spend_vegetables cluster together.

morning_shopper, kids_home, teens_home are closely related.

Spending Plot

Most spending categories are right-skewed, with many low spenders and a few high spenders.
Groceries and Total Lifetime Spend have extreme outliers, suggesting a small group spends significantly more.
Petfood shows a more normal distribution, indicating consistent spending among pet owners.
Some categories like Meat and Vegetables have bimodal patterns, hinting at distinct consumer groups.

Correlation Heatmap

Kids/Teens Home: Strongly correlated
- Meat & Fish Spend: Highly correlated
- Meat/Fish & Groceries Spend: Strong positive correlations.
- Nonalcohol & Alcohol Drinks Spend: Positively correlated

Age & Electronics/Videogames Spend: Weak negative link (younger spend more).
- Complaints & Distinct Stores: Slight positive correlation (0.18).

Item Frequencies

Top 3 Items Purchased:
- Oil
- Cooking Oil
- Babies Food

3 Least Purchased Items
- Pepper

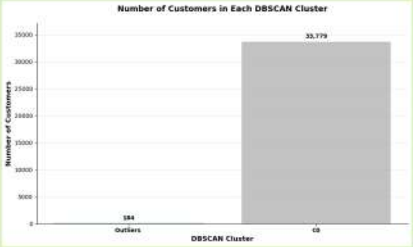
Extra Dark Chocolate
- Honey

Modelling

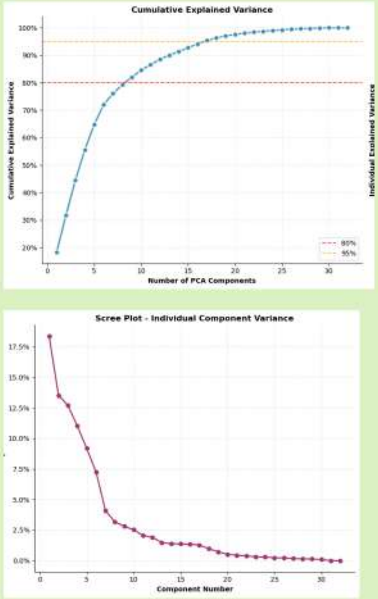
Customer Segmentation

Ran a dbscan grid serach to identify optimal hyperparameters to locate and remove outliers and removed 184

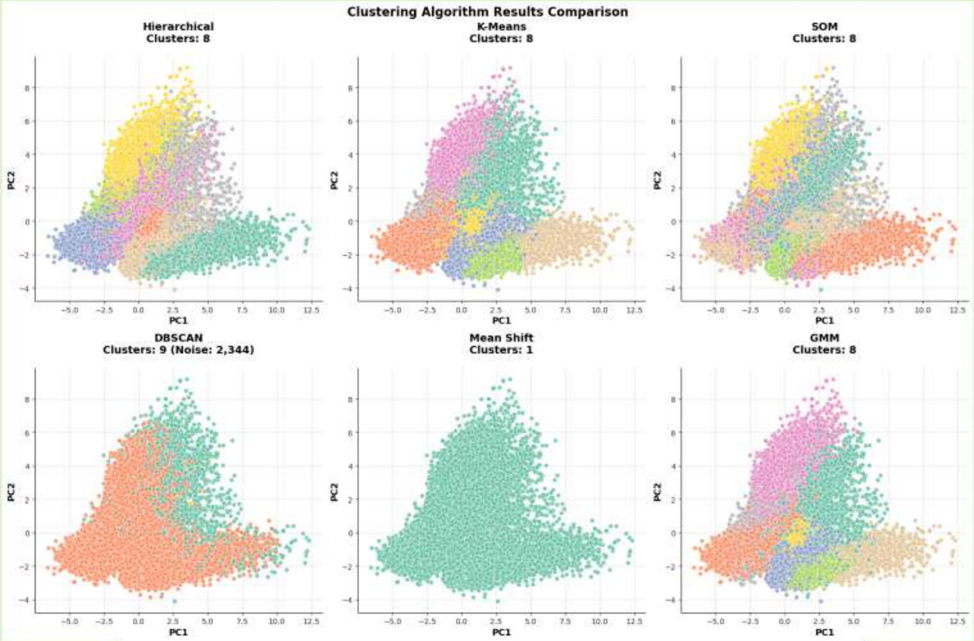
- Used pca as a dimentially reduction tool, because high correlation between the original varliables

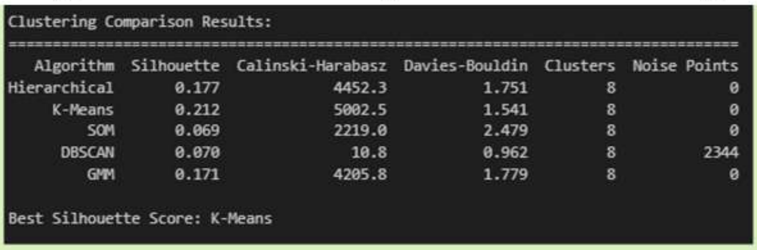


Selecting Number of PCA Componants



Model Selection





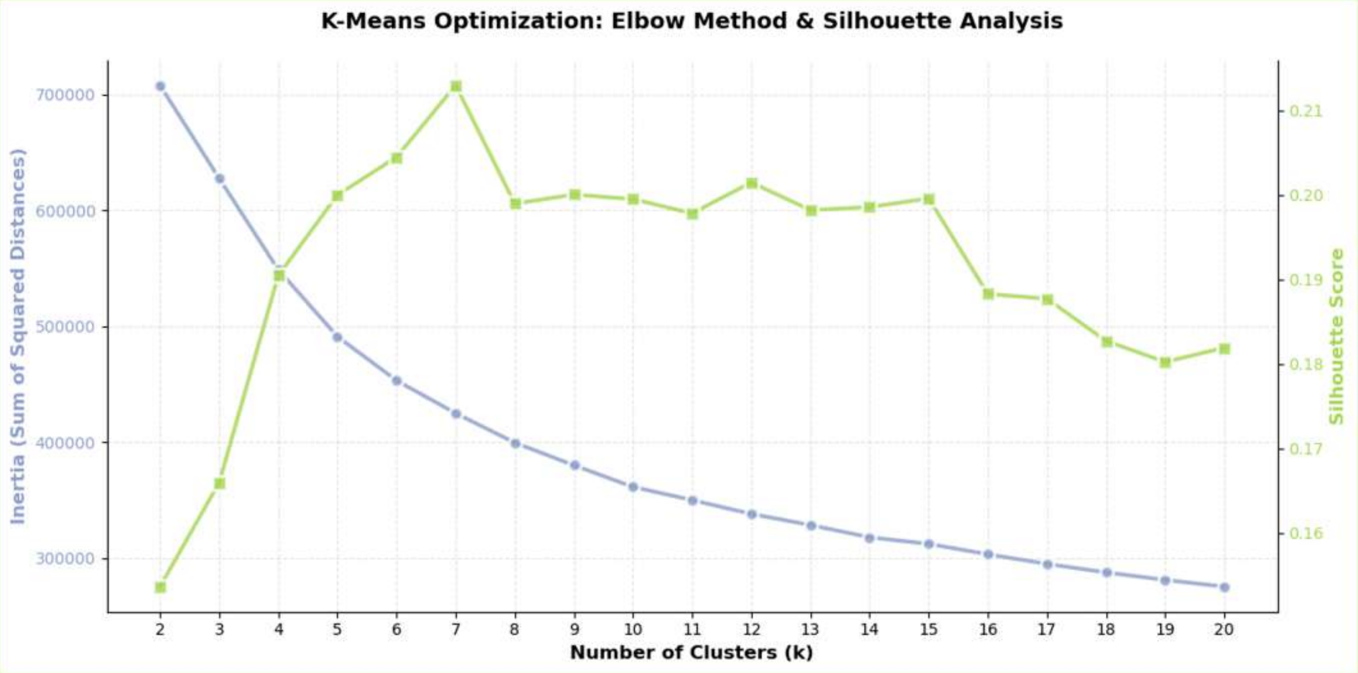
```
Clustering Comparison Results:
================================================================================
   Algorithm  Silhouette  Calinski-Harabasz  Davies-Bouldin  Clusters  Noise Points
Hierarchical       0.177             4452.3           1.751         8             0
     K-Means       0.212             5002.5           1.541         8             0
         SOM       0.069             2219.0           2.479         8             0
      DBSCAN       0.070               10.8           0.962         8          2344
         GMM       0.171             4205.8           1.779         8             0

Best Silhouette Score: K-Means
```

We ran multiple clustering models and evaluated them based on Silhouette, Calinski-Harabasz and Davies-Bouldin metrics, and based on the results we decided on using a K-Means model going forward

Modelling

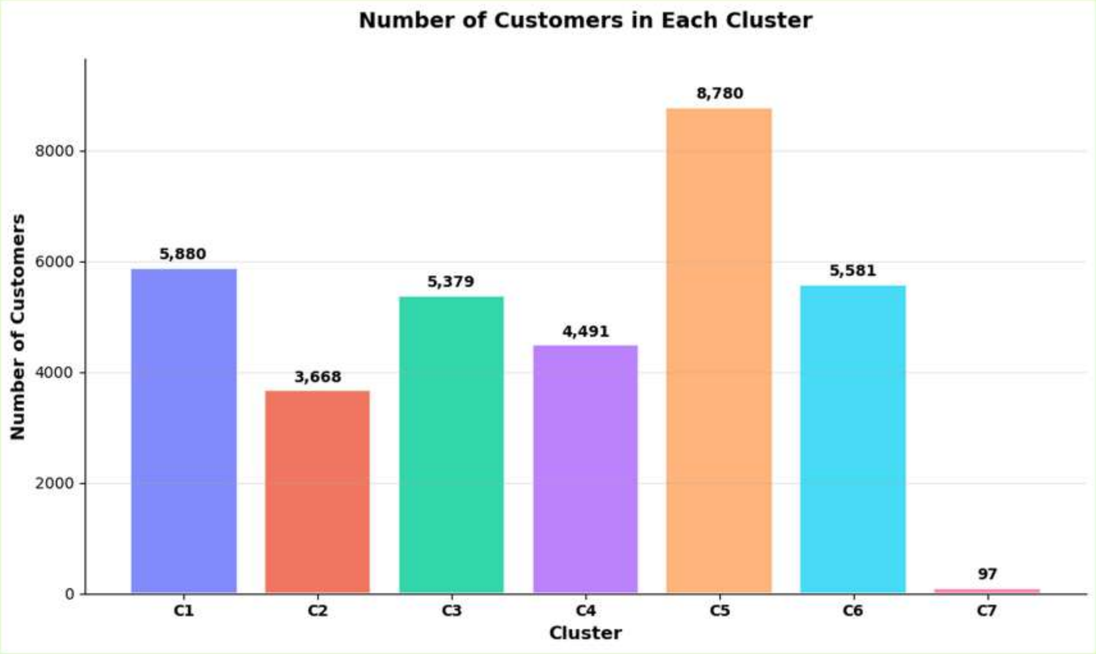
K-Means

Cluster Selection

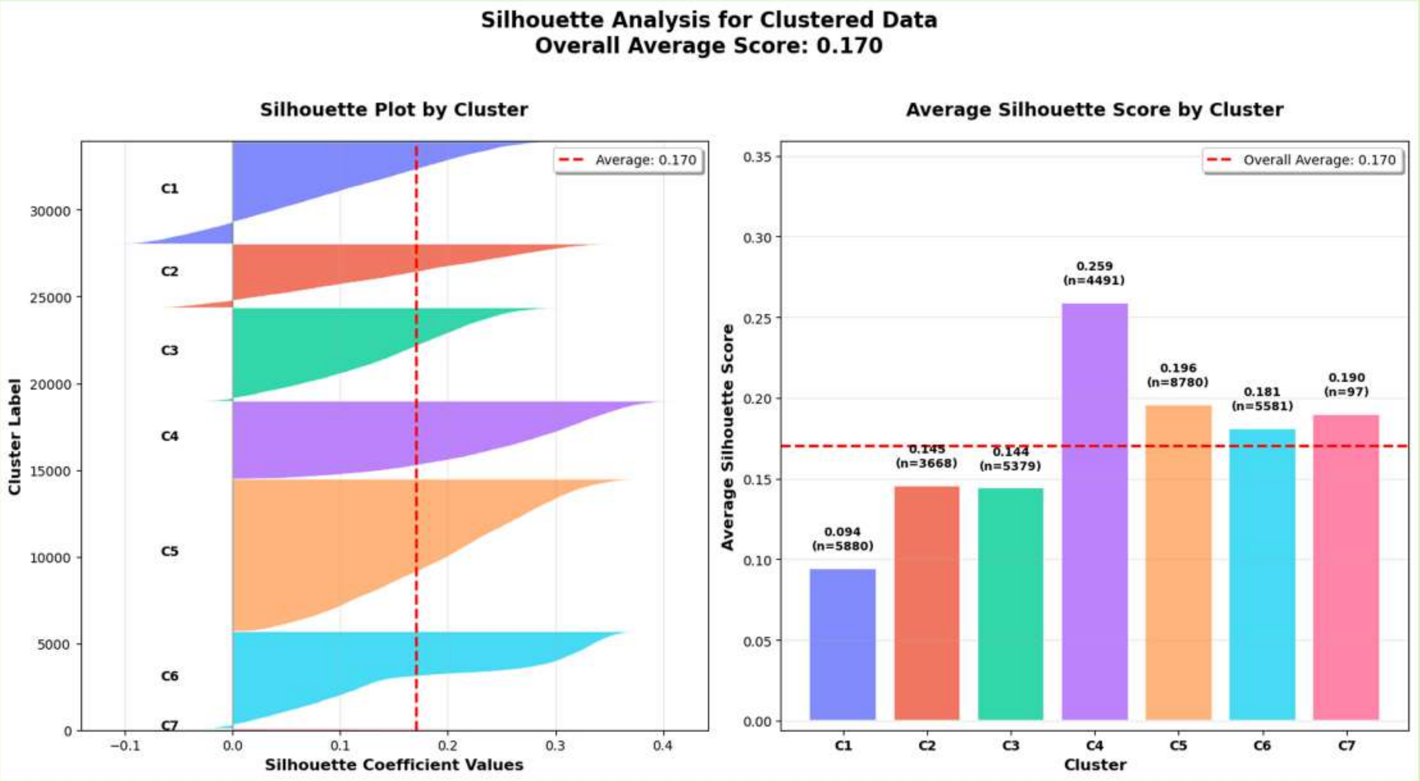


This plot displays K-Means optimization via the Elbow Method (Inertia, blue) and Silhouette Analysis (green). While both suggest good clustering efficiency around k=6, we initially chose this value. However, by including the pre-separated "Makro clients" as a distinct group, our final segmentation yielded seven total customer clusters.

Cluster Results



Cluster Results

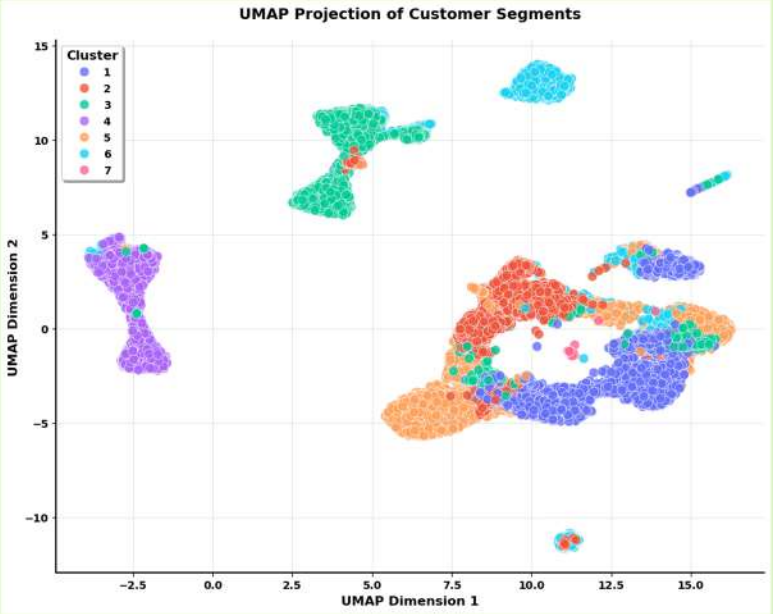


These visuals give a clear overview of our final customer segments.

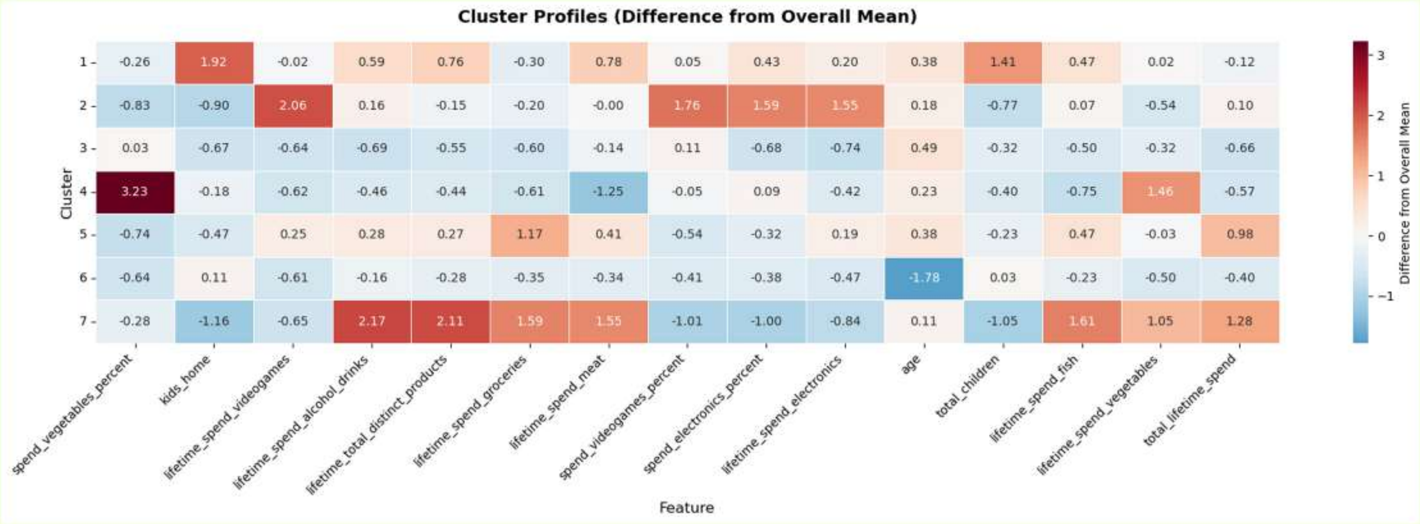
The bar chart showing the number of customers in each cluster highlights the range in group sizes – with Cluster C5 being the largest (8,780 customers), followed by C1, C6, and C3. Clusters C2 and C4 are mid-sized, while C7 is much smaller, with just 97 customers (Makro Cluster). This spread suggests that our customer base is diverse in both behavior and size.

The silhouette analysis offers insight into the quality of these clusters. With an overall average silhouette score of 0.170, some clusters—like C4, C5, C6, and C7—stand out as more well-defined, while others show more overlap. This helps us assess how distinct and meaningful each segment is, guiding how we tailor our strategies for each group.

The UMAP projection reveals well-defined and distinct customer segments, with clear separation between the majority of clusters. While there are a few outliers scattered throughout the plot, they do not significantly impact the overall structure or integrity of the clustering. The compactness and separation of the main groups suggest strong underlying patterns in the data.



Cluster Classification



This heatmap is integral in defining each customer segment. It visualizes how each cluster's average for a given feature deviates from the overall average customer. Red indicates a higher-than-average value, while blue indicates a lower-than-average value, enabling us to pinpoint distinctive characteristics.

By analyzing these significant positive and negative deviations for each cluster, we were able to assign descriptive names that encapsulate their core behaviors and demographics:

Cluster 1: "Family Shoppers" shows higher kids_home and total_children, alongside increased spend_fish and spend_alcohol_drinks percentages.

Cluster 2: "Tech & Gaming Enthusiasts" exhibits notably higher lifetime_spend_videogames, spend_videogames_percent, lifetime_spend_electronics, and spend_electronics_percent, with fewer children.

Cluster 3: "Balanced Budget Shoppers" appears more balanced across these specific features, with no extreme deviations, suggesting a more moderate profile across major spending categories.

Cluster 4: "Healthy Shoppers" is characterized by a significantly higher spend_vegetables_percent and lifetime_spend_vegetables, and lower spending on meat and alcohol.

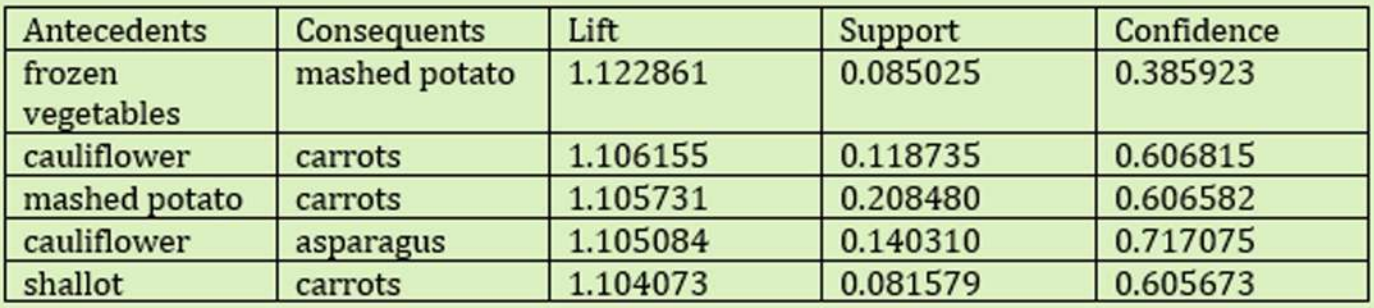
Cluster 5: "Big Spenders" stands out with higher lifetime_total_distinct_products, spend_groceries, and overall total_lifetime_spend.

Cluster 6: "Students" shows a lower age and total_children (consistent with a student profile), with a slight increase in number_complaints. Confirmed by matching with our student block on the map

Cluster 7: "Business Owner Makro" is distinguished by substantially higher lifetime_total_distinct_products, high spend_groceries, spend_meat, and spend_fish, and low numbers of children and tech-related spend, reflecting their commercial buying patterns. Also matches the Block on our map

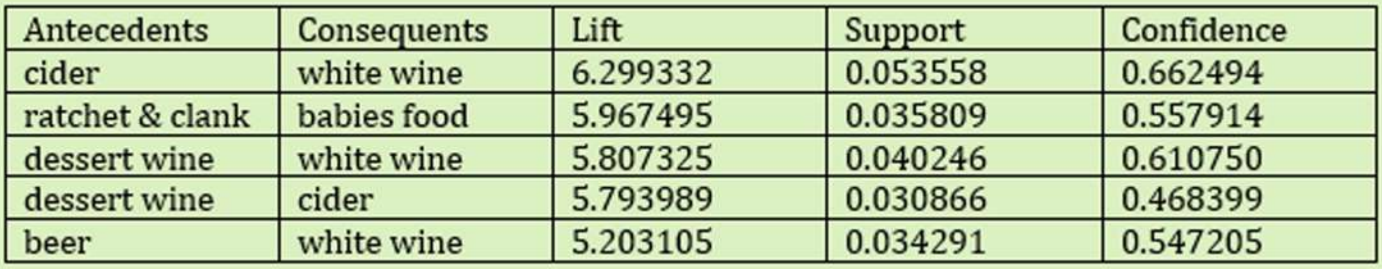
Association Rules

Healthy Shoppers



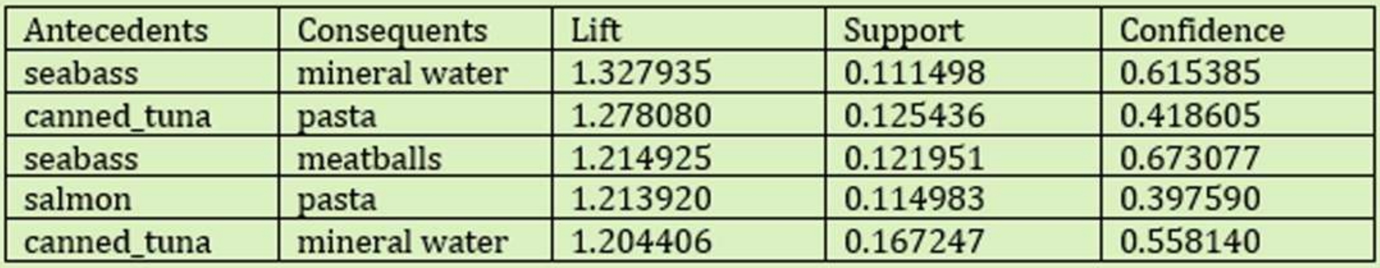
Antecedents	Consequents	Lift	Support	Confidence
frozen vegetables	mashed potato	1.122861	0.085025	0.385923
cauliflower	carrots	1.106155	0.118735	0.606815
mashed potato	carrots	1.105731	0.208480	0.606582
cauliflower	asparagus	1.105084	0.140310	0.717075
shallot	carrots	1.104073	0.081579	0.605673

Big Spenders



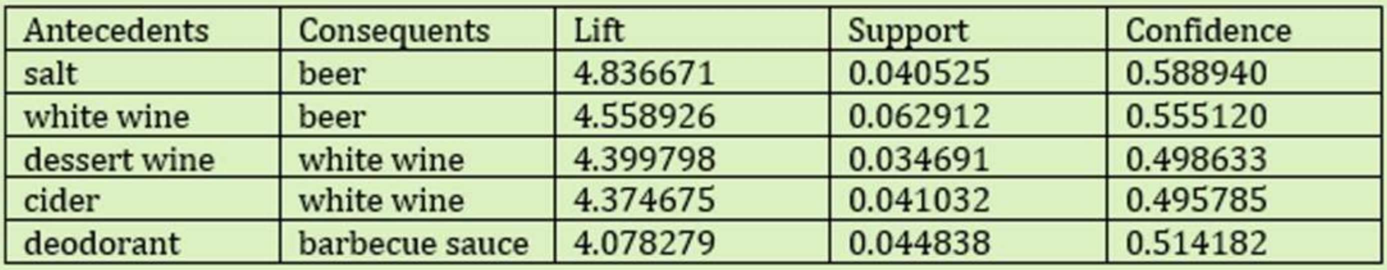
Antecedents	Consequents	Lift	Support	Confidence
cider	white wine	6.299332	0.053558	0.662494
ratchet & clank	babies food	5.967495	0.035809	0.557914
dessert wine	white wine	5.807325	0.040246	0.610750
dessert wine	cider	5.793989	0.030866	0.468399
beer	white wine	5.203105	0.034291	0.547205

Business Owners (MAKRO)



Antecedents	Consequents	Lift	Support	Confidence
seabass	mineral water	1.327935	0.111498	0.615385
canned_tuna	pasta	1.278080	0.125436	0.418605
seabass	meatballs	1.214925	0.121951	0.673077
salmon	pasta	1.213920	0.114983	0.397590
canned_tuna	mineral water	1.204406	0.167247	0.558140

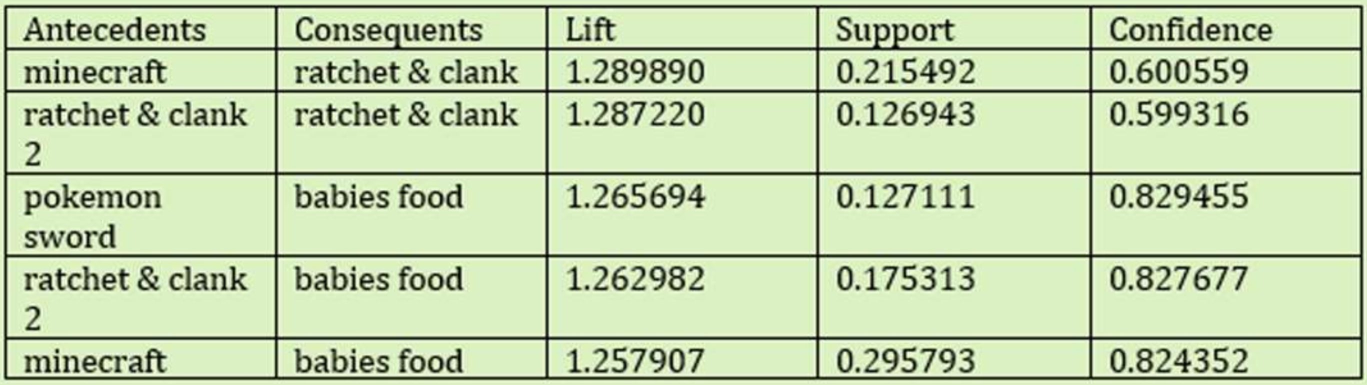
Students



Antecedents	Consequents	Lift	Support	Confidence
salt	beer	4.836671	0.040525	0.588940
white wine	beer	4.558926	0.062912	0.555120
dessert wine	white wine	4.399798	0.034691	0.498633
cider	white wine	4.374675	0.041032	0.495785
deodorant	barbecue sauce	4.078279	0.044838	0.514182

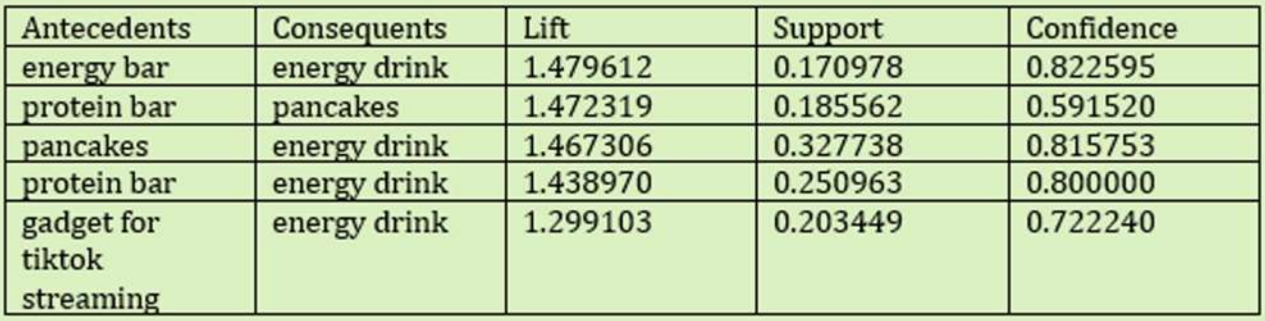
Association Rules

Family Shoppers



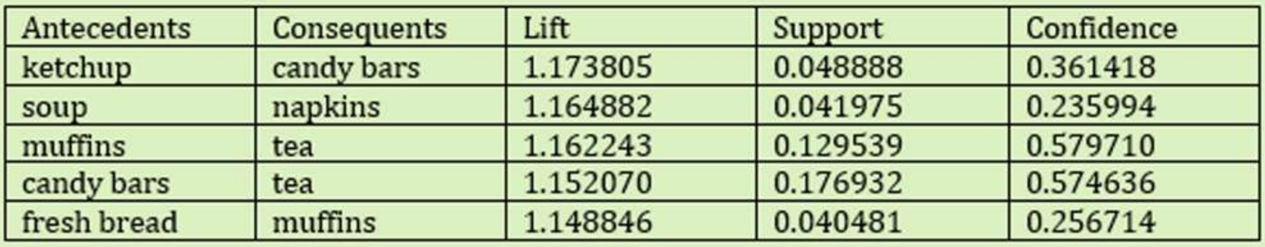
Antecedents	Consequents	Lift	Support	Confidence
minecraft	ratchet & clank	1.289890	0.215492	0.600559
ratchet & clank 2	ratchet & clank	1.287220	0.126943	0.599316
pokemon sword	babies food	1.265694	0.127111	0.829455
ratchet & clank 2	babies food	1.262982	0.175313	0.827677
minecraft	babies food	1.257907	0.295793	0.824352

Tech And Gaming Enthusiasts



Antecedents	Consequents	Lift	Support	Confidence
energy bar	energy drink	1.479612	0.170978	0.822595
protein bar	pancakes	1.472319	0.185562	0.591520
pancakes	energy drink	1.467306	0.327738	0.815753
protein bar	energy drink	1.438970	0.250963	0.800000
gadget for tiktok streaming	energy drink	1.299103	0.203449	0.722240

Balanced Budget Shoppers



Antecedents	Consequents	Lift	Support	Confidence
ketchup	candy bars	1.173805	0.048888	0.361418
soup	napkins	1.164882	0.041975	0.235994
muffins	tea	1.162243	0.129539	0.579710
candy bars	tea	1.152070	0.176932	0.574636
fresh bread	muffins	1.148846	0.040481	0.256714

Targeted Marketing

From Our Generated Clusters We Came Up With Some Marketing Strategies For The Store To Target Each Cluster Of Customers To Potentially Increase Customer Spend

Family Shoppers

Insight: Strong co-purchase of games like Minecraft, Ratchet & Clank, and essential items like babies food.

Strategies:
Bundle Deals: Offer family bundles that include a children’s game (e.g., Minecraft) + babies food at a discount.

Cross-Promotions: Target ads for baby products on gaming pages and vice versa.

Family Loyalty Program: Introduce rewards for recurring purchases in both toys/games and baby categories.

Tech & Gaming Enthusiasts

Insight: High lift for energy snacks/drinks and items like TikTok gadgets.

Strategies:

Energy + Tech Packs: Create “Gamer Fuel Kits” with energy drinks, protein bars, and tech accessories.

Content Campaign: Collaborate with streamers to promote these packs with limited-time codes.

Impulse Shelf Placement: Place energy snacks and small gadgets near checkout for quick grabs.

Balanced Budget Shoppers

Insight: Modest but practical combinations tea + muffins and candy bars

Strategies:
“Everyday Essentials” Discounts: Promote weekly deals on simple, commonly paired goods.

Combo Coupons: Offer a “Buy 2 Save More” coupon for items like tea + muffins or bread + candy.

Meal Planning Kits: Suggest low-cost breakfast or snack kits.

Healthy Shoppers

Insight: Fresh and frozen veggies often bought together, especially cauliflower, carrots, and asparagus.

Strategies:
Healthy Recipe Packs: Bundle these ingredients with a recipe card (e.g., “Roasted Veggie Medley”).

Nutrition-Focused Email Campaigns: Share cooking tips for using these veggies.

"Fresh Start" Discounts: Target discounts to customers showing interest in frozen/fresh produce.

Big Spenders

Insight: Alcohol (like wine, cider), gourmet food (dessert wine, babies food, etc.) cluster strongly.

Strategies:

Luxury Pairing Events: Host in-store or online wine & food pairing events.

Premium Product Bundles: Create “Weekend Indulgence” packs with white wine + cider + dessert items.

VIP Tier: Offer exclusive access to limited-edition wines and gourmet products.

Students

Insight: High-lift combos like salt + beer, white wine + beer.

Strategies:
- Student Survival Kits: Include beer, snacks and hygiene products.

- Weekend Combo Deals: Discount packs for party or game nights (beer + chips + salt).

- Campus Ambassadors: Use peer marketers to promote quirky bundles on social media.

Business Owner (Makro Shoppers)

Insight: Strong food service associations—seabass + mineral water, canned tuna + pasta.

Strategies:

Wholesale Meal Kits: Offer foodservice-ready meal prep kits (fish + sides).

Bulk Discount Campaigns: Provide tiered discounts for bulk purchases of frequently co-bought items.

Menu Planning Services: Partner with chefs to create meal suggestions that use popular pairings.

These shopper segments and strategies highlight the power of targeted promotions based on co-purchase behavior. By aligning marketing efforts with the unique habits and preferences of each group, businesses can drive engagement, boost sales, and build stronger customer loyalty. Personalization at this level turns raw transaction data into actionable, revenue-generating insights.

Conclusion

In this project, we performed customer segmentation using clustering techniques on retail transaction data. The main goal was to identify distinct groups of customers based on their purchasing patterns and then extract actionable insights through association rule mining.

The clustering process yielded meaningful and well-defined segments with relatively good evaluation scores, indicating a strong separation between customer types. Each cluster revealed unique behavioral traits that aligned with real-world consumer personas, such as Family Shoppers, Tech & Gaming Enthusiasts, Balanced Budget Shoppers, Healthy Shoppers, Big Spenders, Students, and Business Owners (Makro segment).

To deepen our understanding of each segment, we applied association rules within each cluster. This allowed us to uncover product pairings and preferences specific to each group. For example, Family Shoppers often purchased kids' games along with baby food, while Tech & Gaming Enthusiasts showed strong affinities for energy drinks and digital accessories. Big Spenders leaned toward premium items like wine and gourmet foods, whereas Students had quirky, budget-conscious combinations such as salt and beer.

These findings were translated into tailored marketing strategies for each segment. From bundle offers and influencer campaigns to loyalty programs and wholesale kits. Overall, the combination of clustering and association rules proved highly effective in surfacing practical, data-driven strategies to better serve different customer needs and improve targeted marketing efforts.