Marketing Final Project Report

Clustering Mall Customers

Ishaan Buch
Saiansh Raizada
Tyler Cushing

I. Background and Problem

A United States supermarket has been facing a problem regarding their annual revenue. The supermarket has been in operation since 2008 and was flourishing until 2016 where problems began to arise. Since 2016, their annual revenues have declined by 10% every year. In general supermarkets have very low and incredibly thin profit margins, on average between 1% and 3%. So, not selling the amount of products that you intended to can cause a very fast decline in revenue. We suspect the problem is not due to a single supermarket, but rather across the supermarkets as a whole. Supermarkets sell a ton of items and that is where they make a profit. There is not a high markup in prices in supermarkets, so to be losing 10% of their revenue annually the problem is at hand across the board.

The United States Supermarket data was gathered from Kaggle.com and is used for marketing analysis purposes.

Research Objectives:

1. Cluster customers to analyze how their annual income effects their spending patterns
2. Analyze the different age categories that shop at the supermarket
3. How much do customers particularly spend when they shop at the supermarket
4. What is the gender breakdown of customers that shop at the supermarket

Justificant for the research:

1. Why do we think it is important?

With annual revenues declining yearly by 10%, the supermarket will not be able to stay afloat if the revenue problem is not resolved. It is important for the supermarket to know who they are selling to and what their target audience is. If they are improperly utilizing their promotions and products it could lead to a decline in revenue.

2. What types of data and quantitative analyses will help us address the problem?

The types of data that will help our quantitative analysis:
Genre of individual: Male or Female
Age of Individual
A Spending Score of the Individual (on a scale of 1-100)
Annual Income of Individual (in thousands of dollars)

3. What types of business strategies and/or regulatory policies will our research be able to recommend

Our research will be able to recommend different promotions and marketing techniques that the supermarket can focus on so that they can maximize their target audience. Knowing the type of people that shop at the store can help us focus on certain aspects of the business and narrow down where exactly the lost profit is coming from.

Pricing strategies is a second strategy we intend to address via our research. Analyzing customer purchasing patterns and behavior can help identify optimal pricing strategies for different products.

Lastly, customer segmentation can be used to help supermarkets with their product placement and selection throughout the store.

II. **Data Summary and Exploratory Data Analysis**

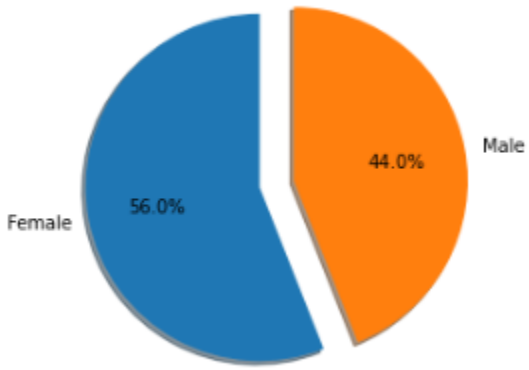

Figure 1: Male vs Female Count

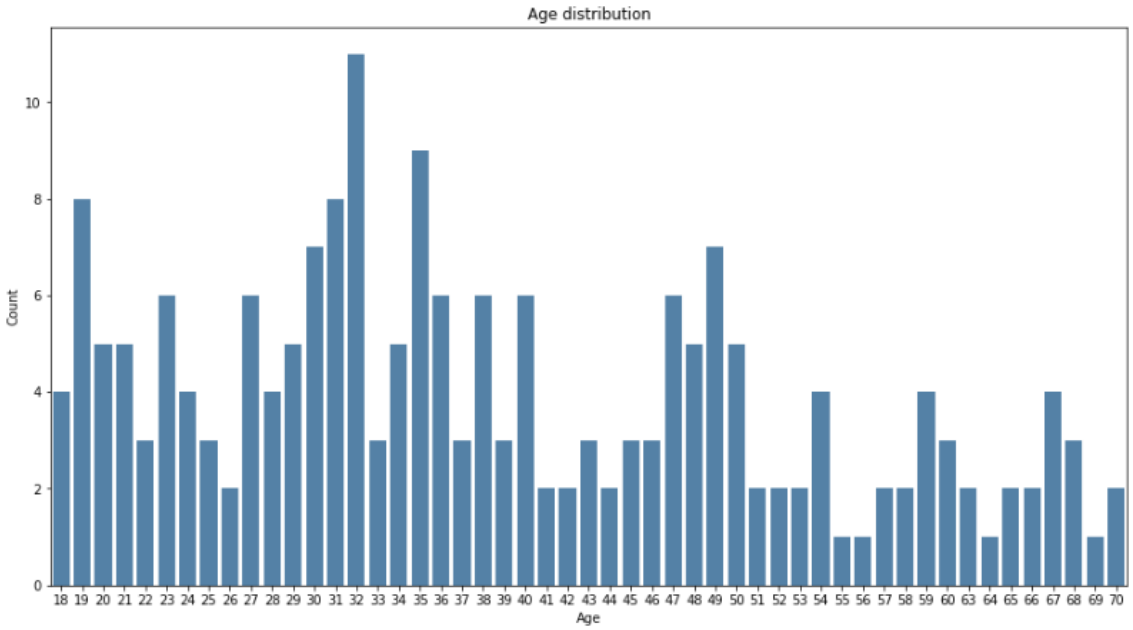

Figure 2: Age Distribution

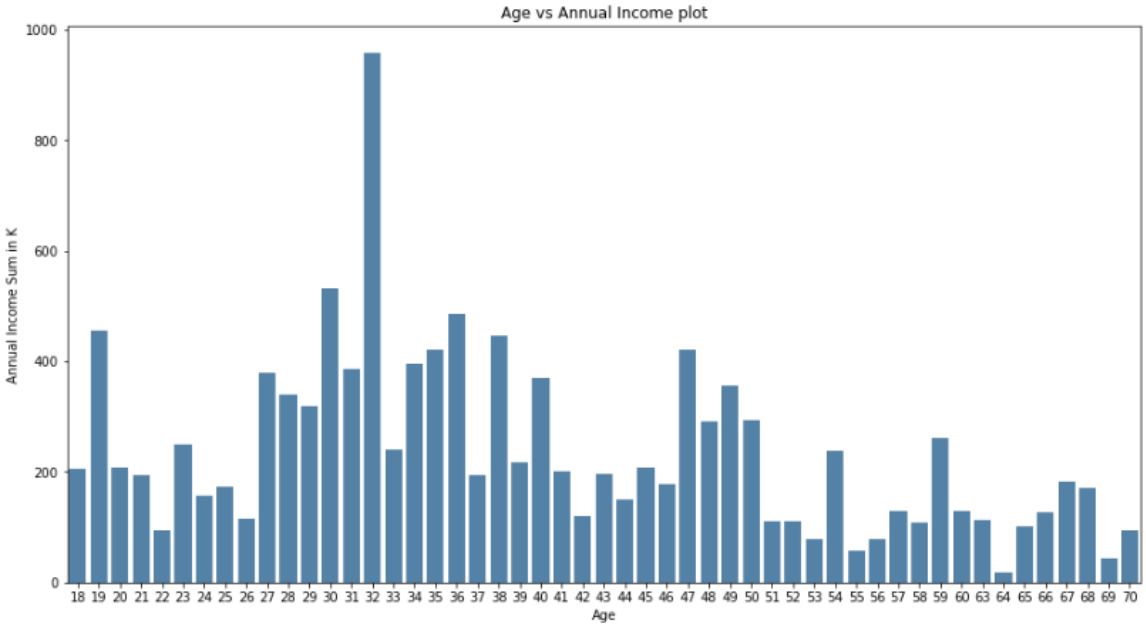

Figure 3: Age vs Annual Income

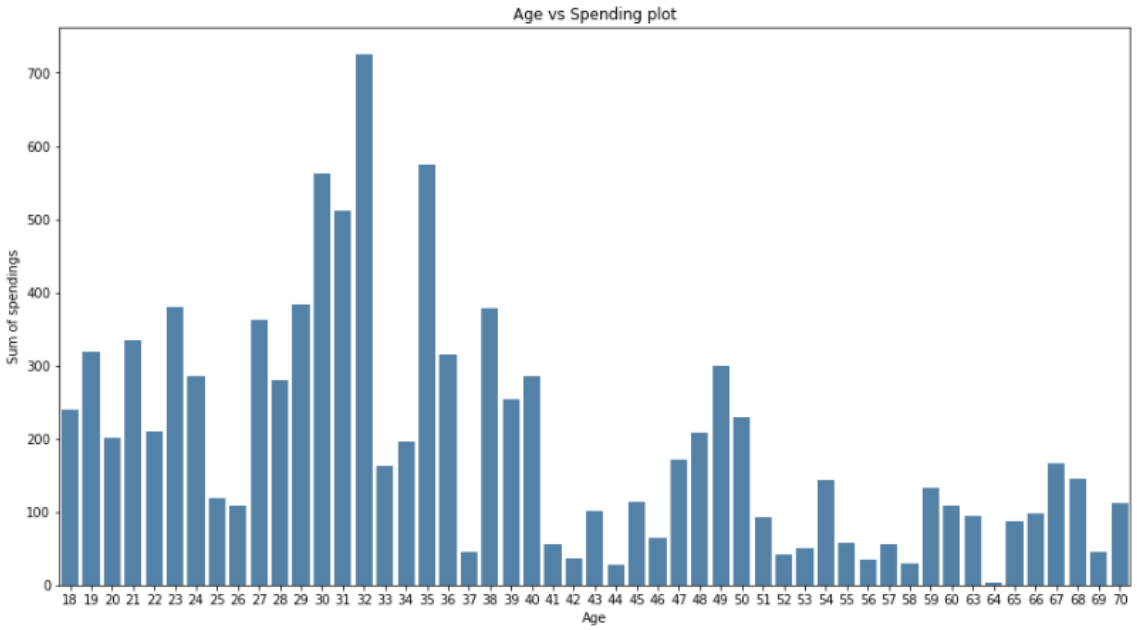

Figure 4: Age vs Spending

Summary Statistics of Customers			
	Age	Annual Income(K)	Spending Score (1-100)
Count	200.00	200.00	200.00
Mean	38.85	60.56	50.20
Std	13.97	26.26	25.82
Min	18.00	15.00	1.00
25%	28.75	41.50	34.75
50%	36.00	61.50	50.00
75%	49.00	78.00	73.00
Max	70.00	137.00	99.00

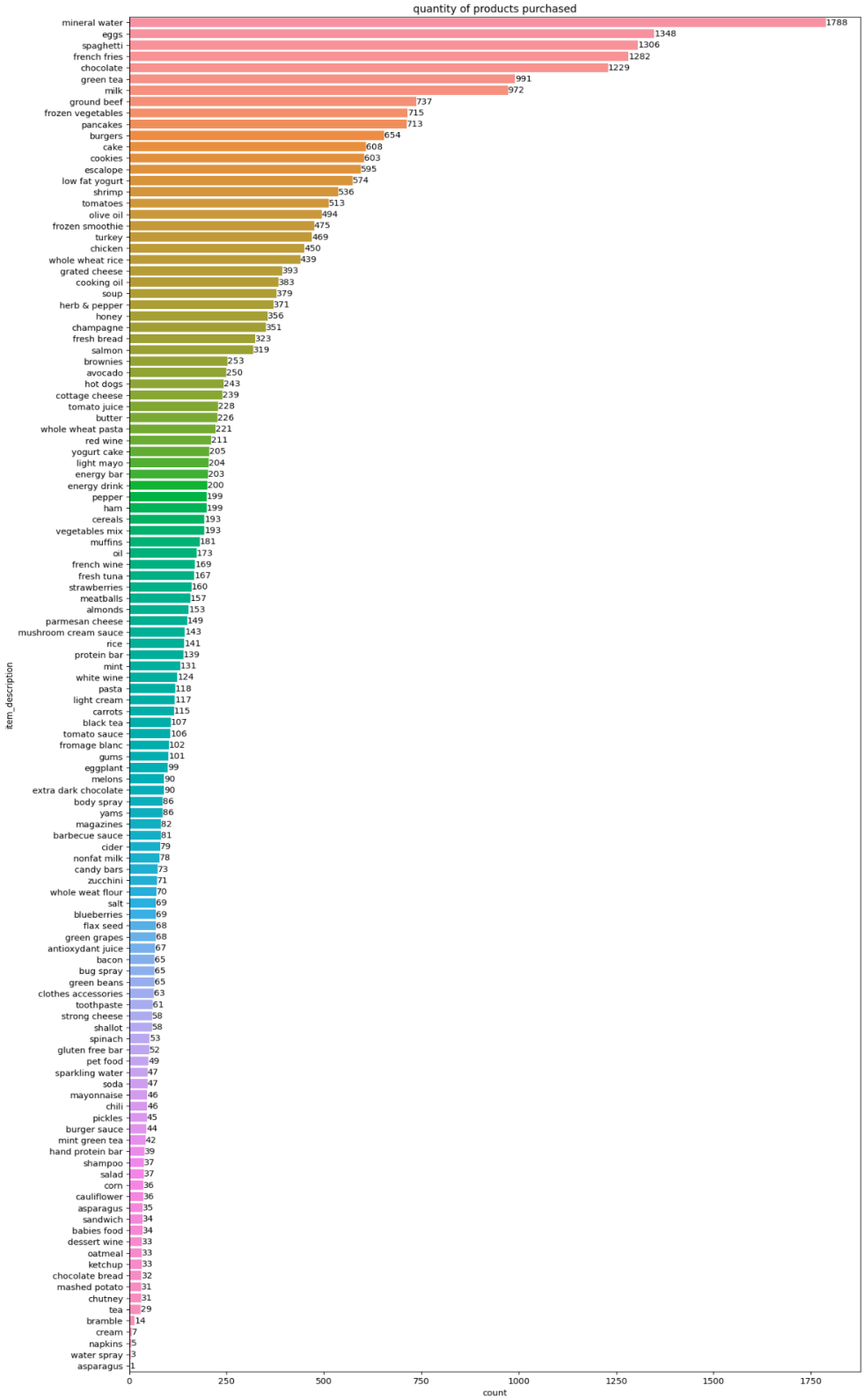

Figure 5: Frequency Chart

Notable observations:

- On average, a consumer buys 4 products.
- The highest number of products purchased in a single instance was 20.
- The top 10 most purchased products are as follows:
    - Mineral water
    - Eggs
    - Spaghetti
    - French fries
    - Chocolate
    - Green tea
    - Milk
    - Ground beef
    - Frozen vegetables
    - Pancakes
        - It is worth noting that most items here are groceries

III. **Data Analyses and Key Findings**

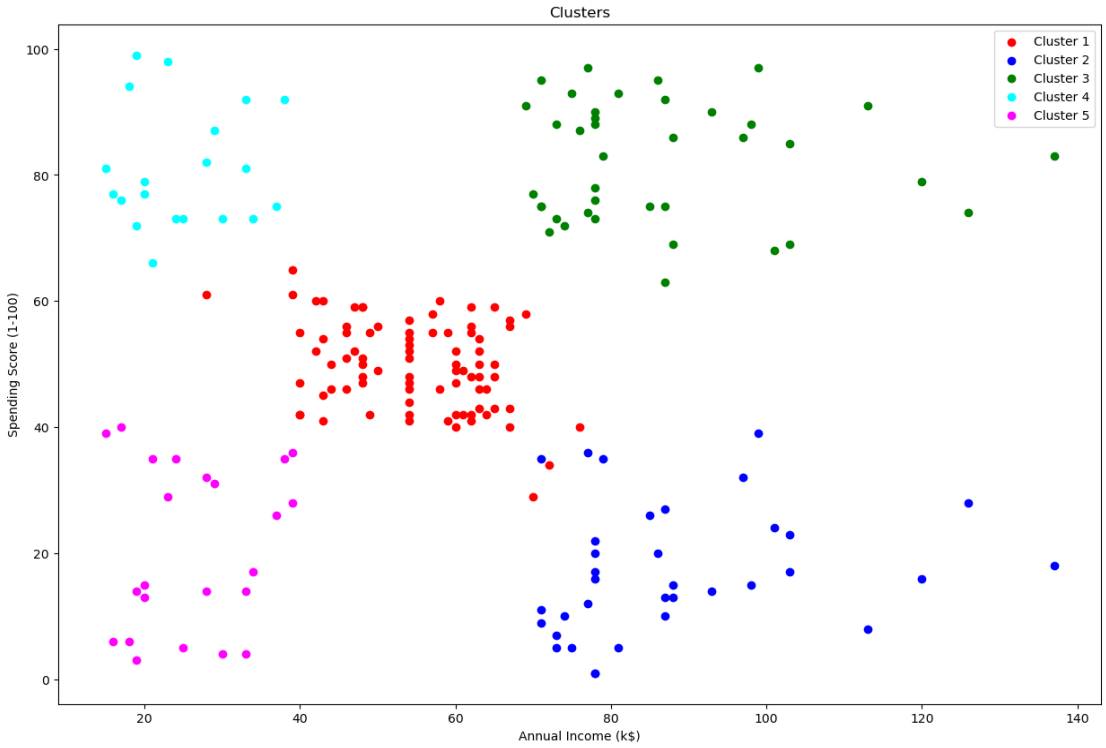

Figure 6: Cluster Analysis

- Cluster 5 (Magenta) : People with low annual income and who spend lower in the mall. These are customers still less targeted.
- Cluster 2 (Blue) : People with high annual income and who still don't spend much money in the mall. These customers can be targeted so that they can be tempted to spend more in the mall. These are the customers with 3rd priority to be targeted (1 - Cluster 5, 2 - Cluster 2).
- Cluster 4 (Cyan) : People with low annual income and who spend a lot in the mall. These Customers are less targeted by the mall (if the mall wishes to so that the people with low income don't need to spend too much in the mall - ethics).
- Cluster 3 (Green) : People with higher annual income and who spend more in the mall. These customers are likely to buy more in the mall so they are the priority targets for the mall.
- Cluster 1 (Red) : People with average annual income and spend averagely in the mall.

Market Basket Analysis:

First, the necessary packages are imported: pandas, numpy, missingno, matplotlib.pyplot, seaborn, plotly.express, plotly.graph_objects, make_subplots from plotly.subplots, StandardScaler and MinMaxScaler from sklearn.preprocessing, Counter from collections, LabelEncoder from sklearn.preprocessing, apriori and association_rules from mlxtend.frequent_patterns, KMeans and DBSCAN from sklearn.cluster, silhouette_score, calinski_harabasz_score, davies_bouldin_score from sklearn.metrics, and PCA from sklearn.decomposition.

Then, a CSV file is read into a Pandas DataFrame using pd.read_csv. The DataFrame contains transaction data for a market basket analysis. Next, the transaction data is processed to create a new DataFrame with each row representing a single item in a transaction. This is done by iterating over each transaction and creating a new row for each item in the transaction. The new DataFrame has two columns: “ID_client” and “item_description”. The next step is to analyze the data and generate some visualizations. First, a bar plot is created using seaborn to show the frequency of each item purchased. The top 10 most purchased items are identified and displayed in the plot. The top 10 least purchased items are also identified. Next, association rules are used to identify the most commonly purchased pairs of items. Apriori algorithm is used for generating frequent itemsets and association_rules is used for generating association rules. The minimum support is set to 1% and the minimum confidence is set to 40%. The results are displayed in a DataFrame.

Finally, KMeans, DBSCAN and PCA algorithms are used for clustering and dimensionality reduction. The results of the clustering algorithms are evaluated using silhouette_score, calinski_harabasz_score and davies_bouldin_score. The results are visualized using plotly.

IV. **Actionable Insights, Limitations, and Future Research**

Cluster 5 (Magenta) : People with low annual income and who spend less in the mall. These customers represent a large untapped market for the mall and can be targeted with promotions and discounts to encourage them to spend more in the mall. Additionally, the mall can also target these customers with lower-priced merchandise and affordable dining options.

Cluster 2 (Blue) : People with high annual income but who still don't spend much money in the mall. These customers can be targeted with high-end merchandise and dining options, as well as exclusive events and promotions. The mall can also reach out to these customers through targeted advertising and personalized shopping experiences.

Cluster 4 (Cyan) : People with low annual income but who spend a lot in the mall. These customers are a prime target for the mall to offer them promotions and discounts to keep them coming back. The mall can also offer affordable financing options and better customer service to make their shopping experience more enjoyable.

Cluster 3 (Green) : People with higher annual income who spend more in the mall. These customers are the top priority for the mall as they are likely to continue spending money in the mall. The mall can target these customers with high-end merchandise, exclusive events, and personalized shopping experiences. Additionally, the mall can also offer loyalty programs and special rewards for these customers to further incentivize them to spend more in the mall.

Cluster 1 (Red) : People with average annual income and spend averagely in the mall. These customers are likely to continue spending at a steady pace in the mall, so the mall should focus on providing them with a consistent and enjoyable shopping experience. The mall can target these customers with regular promotions, in-store events, and a well-curated selection of merchandise to keep them coming back.

Limitations of Cluster Analysis

- Assumptions: Cluster analysis assumes that the data is normally distributed and that the clusters are spherical. If these assumptions are not met, the results of the analysis may not be accurate.
- Sensitivity to Initial Conditions: The results of cluster analysis can be sensitive to the initial conditions, such as the starting centroids or the method used to determine the number of clusters.
- Subjectivity: The interpretation of the results of cluster analysis can be subjective, as different analysts may come to different conclusions based on the same data.
- Limitations of Clustering Algorithms: Different clustering algorithms have different strengths and weaknesses, and the choice of algorithm can impact the results of the analysis.

Future Research Opportunities

- Further analysis of the customer segments: The current analysis could be expanded to include more variables such as customer demographics, customer preferences, purchase history, and shopping behavior.
- Experimental design: Conducting experiments to test the efficacy of different marketing and pricing strategies in different customer segments. This could help the supermarket understand which strategies work best for different types of customers.
- Predictive modeling: Building predictive models using machine learning algorithms to forecast future sales, customer behavior, and revenue. This could help the supermarket make informed decisions about pricing, promotions, and product placement.
- Customer satisfaction surveys: Conducting customer satisfaction surveys to understand customer needs and preferences, and to identify areas where the supermarket can improve its customer experience.
- A/B testing: Conducting A/B tests to compare the impact of different marketing strategies on customer behavior and revenue. This could help the supermarket optimize its promotions and marketing efforts.

- Natural language processing: Analyzing customer feedback and comments from social media and other sources to identify trends and patterns in customer behavior and preferences.
- Geo-spatial analysis: Mapping customer locations and behavior patterns to identify geographical areas where the supermarket is losing revenue.
- Collaborative filtering: Using collaborative filtering algorithms to recommend products and promotions to customers based on their past behavior and preferences.

V. **Appendix**

1. Jupyter Notebook “clustering-mall-customers2-hierarchical-clusterin.ipynb”
2. Jupyter Notebook “MarketBasketOptimisation.ipynb”
3. Dataset obtained from Kaggle located in folder “supermarket_marketing”
    a. “Supermarket_CustomerMembers.csv”
    b. “Market_Basket_Optimisation.csv”