DATATHON 4: Analysis Supermarket Sales

James Ding

Jody Nguyen

Naz Shalamo

Matthew Haddad

Javier Becaria

June 2, 2024

1 Introduction

In the retail industry, it is important to know about the desires of clients and their behaviors in order to optimize product offers and improve their satisfaction. This research aims at exploring complex interrelations between volumes of product sales, unit prices, customer ratings and purchase patterns across various types of products and payment methods. Average quantities sold, customer rating, unit prices are some of the features we are going to target with measures like K-means Clustering, Principal Component Clustering and exploratory data analysis (EDA).

Moreover, customer segmentation examines how demographic characteristics and consumption behavior form distinct groupings that can be used to infer consumer satisfaction levels. The main objective here is to study payment methods as related to quantity sold as well as unit prices so that general trends in transaction preferences can be established. In addition, we will identify factors influencing customer ratings by looking for reasons causing higher or lower shopping experiences.

This approach enables pinpointing areas for improving products and services and pricing strategies that work best in customer care service.

2 Data collection and cleaning process.

The Supermarket Sales database, available on Kaggle and published on May 27, 2019 [1], is a well-prepared dataset ideal for analytical projects. It offers a clean and ready-to-use collection of data, with no calculated data added and no data transformation required, enabling users to focus directly on their analyses without the need for extensive preprocessing. However,

there is some uncertainty regarding the dataset's origin, as it is unclear whether the data represents actual sales records or if it was synthetically generated for the purpose of data science practice. This ambiguity does not detract from its utility for educational and analytical purposes.

3 Questions and Objectives

3.1 Question 1 (Jody): How does the balance between average quantity sold and customer rating vary across different product lines? What do the different clusters reveal about customer satisfaction across various product quantities?

3.1.1 Objective 1

The objective is to determine the correlation between average quantity sold and customer ratings across diverse product lines & segment customer data based on the quantity of products purchased and the ratings provided. It aims to identify specific patterns that suggest underlying consumer behaviors and investigate which product quantities are consistently associated with higher or lower customer ratings. Finally, we wish to use insights from the analysis to monitor and adjust market conditions to adapt with customer needs.

3.1.2 Data visualization process and techniques

In order to achieve impact and clarity, there are several crucial measures that must be taken in order to create an effective chart. Preparing and cleansing the data is the first stage.

For the EDA analysis, we used Power BI to plot the pie chart and the scatter plot. Variable Quantity and Rating were placed in X axis and Y axis with Product Line were placed in Legend. To improve visualization, adjust visual settings such as Category Label, Legend, Axis's Titles.

For K-means and PCA analysis, we used Python to examine our targeted objectives. Necessary libraries should be imported prior to our analysis such as pandas, sklearn, altair, seaborn, plotly, and mathplotlib.

3.1.3 Discussion of visualization findings

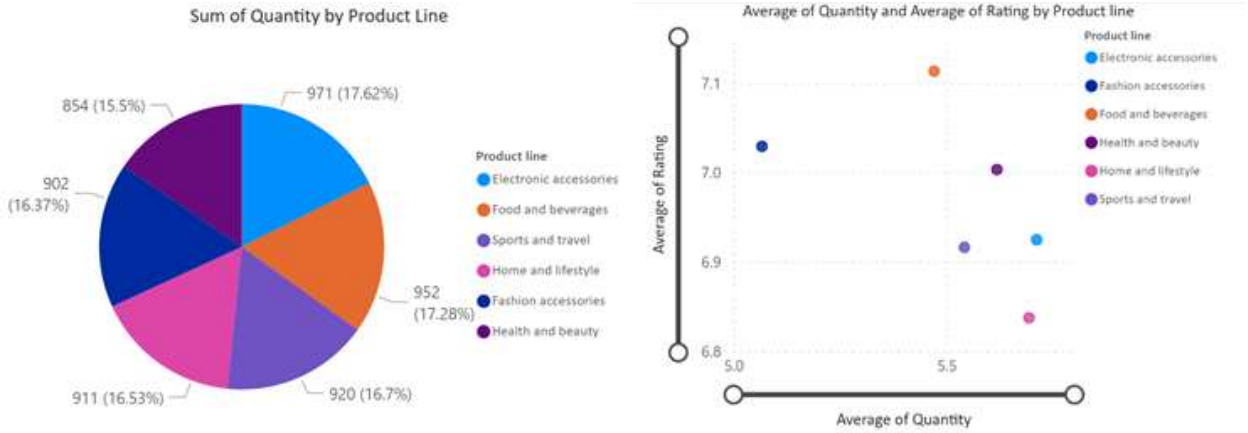


Figure 1.1. Pie Chart and Average Rating vs. Quantity by Product Line

The pie chart provides a clear distribution of product quantities sold across various product lines. The quantities sold are relatively evenly distributed among the product lines.

The second chart provides a visual comparison of consumer behavior across different types of products.

Food and Beverages: Although this category shows the highest average rating, it shows a lower average quantity. This could imply that while the products are highly rated, they are purchased in moderate quantities per transaction.

Fashion Accessories: This product line has a high average rating (around 7.0) and a relatively lowest average quantity, suggesting that these items are not often bought in large quantity due to higher unit price but overall got good customer satisfaction.

Health and Beauty, Electronic Accessories and Sports and Travel: They have similar ratings and quantities, indicating balanced customer satisfaction and purchasing behavior.

Home and Lifestyle: This product line has the lowest average rating (just below 6.9) but with a high average quantity, suggesting either its design or specific quality issues affecting customer satisfaction.

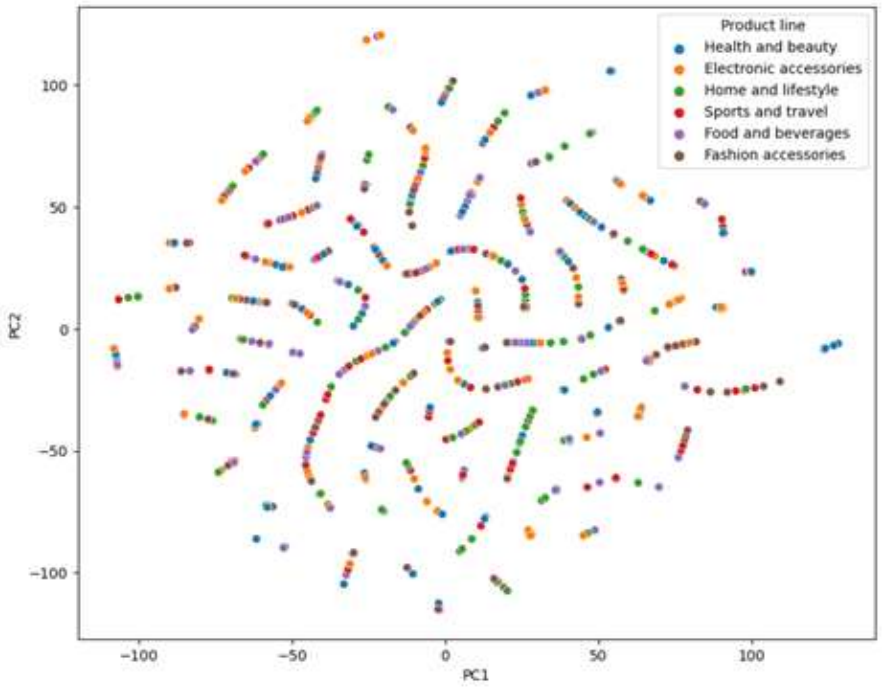


Figure 1.2. Principal Component Cluster Evaluation

The t-SNE plot shows a non-linear distribution of data points. The Health & Beauty category appears to cluster around the center. The clustering isn't very tight, suggesting moderate variability within this product line. Food and Beverage products form a few, tighter clusters in

the center regions of the plot. This product line shows less mixing with others, which might suggest specialized products that are distinct from other categories in terms of customer rating and quantity sold. Points representing Home and Lifestyle & Electronic Accessories and Sport & Travel are dispersed throughout the plot, with no distinct clustering, implying significant diversity within this product line in term of customer rating and sold quantity.

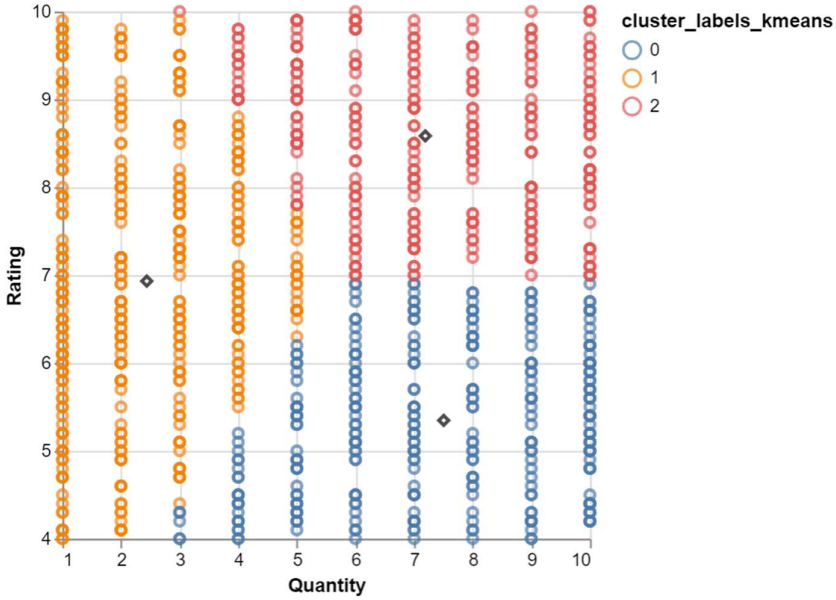


Figure 1.3. K-means Cluster Evaluation

All clusters are spread across all levels of quantity from 1 to 10, which implies that the quantity purchased does not strongly influence the rating. In other words, buying more or less of a product doesn't consistently affect how it's rated, as high, medium, and low ratings are seen at almost every quantity level. Products in Cluster 0 might require attention to understand why they are rated poorly and if there's a quality or expectation mismatch. Products in Cluster 1 are likely to meet or exceed customer expectations.

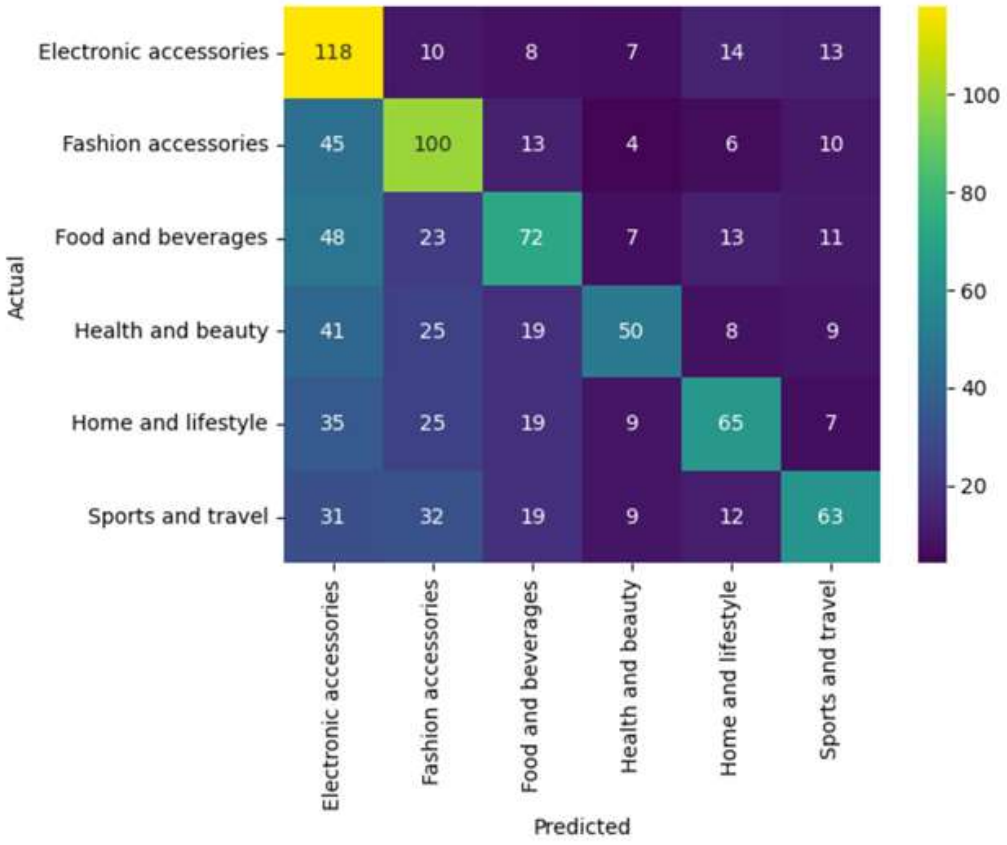


Figure 1.4. Confusion Matrix

The model performs well in correctly predicting transactions for "Electronic Accessories" and "Fashion Accessories", indicating a high number of true positives. There are noticeable confusions between some categories, suggesting that the features used by the model may overlap significantly between these categories

3.1.4 Practical recommendations

Based on the findings, with a relatively balanced distribution of sales across product lines, it is recommended to consider investing in targeted product lines that show slightly higher sales volumes and profits. The K-means clusters analysis does not provide too many insights due to no difference between purchased quantity and product ratings. However, further K-means

cluster analysis with another variable such as unit price and product ratings are strongly recommended for better pricing strategies. The off-diagonal numbers in confusion matrix, particularly where misclassifications are high, provide opportunities for further refinement of the model.

3.1.5 Limitations

The interpretation of clusters and model results can be subjective, especially without extensive domain knowledge, leading to potentially biased decisions. The model's effectiveness is heavily dependent on the variables used. If important variables that influence customer behavior or product classification are missing or incorrectly recorded, the model's predictive accuracy and the utility of the clustering might be misinterpreted. This dataset has an issue where gross income column is not reported correctly. It is reported the same as 5% of Unit price which is equal to Tax amount. Hence, while the gross income is one of the important variables for our analysis, we could not make use of it.

3.2 Question 2 (James): What patterns can be identified in customer ratings across different branches and product lines that can provide actionable insights for improving customer satisfaction?

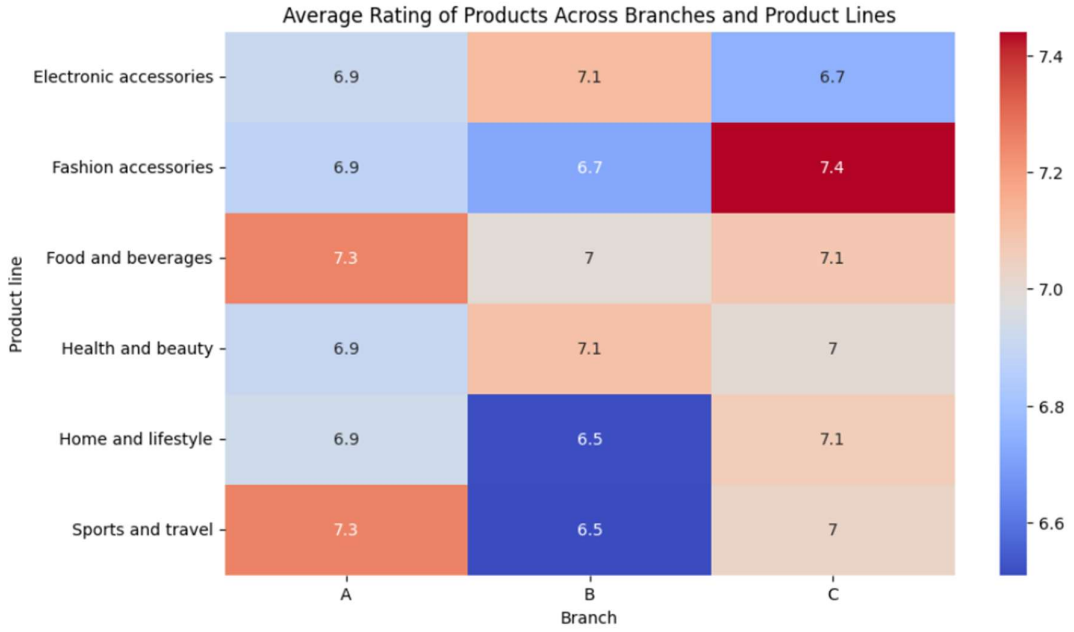
3.2.1 Objective 2

To identify patterns in customer ratings across different branches and product lines to provide actionable insights for improving customer satisfaction.

3.2.2 Data visualization process and techniques

Plotting the Heatmap: The heatmap is plotted using seaborn's heatmap function. The annot=True argument adds annotations (the average rating values) to each cell in the heatmap. The cmap='coolwarm' argument specifies the colour map for the heatmap.

3.2.3 Discussion of visualization findings



Branch Performance: The heatmap reveals that certain branches consistently receive higher ratings for specific product lines. For example, Branch A might excel in Electronics, while Branch B could have mixed performance.

Product Line Performance: The heatmap also shows how different product lines are rated across branches. For instance, Home and Lifestyle products might have higher ratings in Branch C compared to others.

Targeted Improvements: Branches with lower ratings for certain product lines can investigate the reasons behind these ratings and take corrective actions, such as improving product quality or customer service.

3.2.4 Practical recommendations

- Implement branch-specific strategies to share best practices and address areas needing improvement.
- Enhance product lines with lower ratings through quality reviews and customer

feedback.

- Provide targeted training for staff in branches with lower ratings to improve customer service and product knowledge.
- Establish a robust customer feedback mechanism to continuously monitor and address customer satisfaction issues.
- Use insights from high-rating branches and product lines for targeted marketing campaigns and promotions.

3.2.5 Limitations

Data Scope

Temporal Factors

Subjectivity in Ratings

Missing Data

External Influences

Data Quality

3.2.6 Conclusions

Our findings revealed distinct patterns in customer satisfaction:

- Certain branches, such as Branch A, consistently received higher ratings for specific product lines like Electronics, indicating strong performance in these areas.
- Variability in ratings across branches and product lines highlighted areas for potential improvement, such as the need for better customer service or product quality enhancements in specific branches or product lines.

The heatmap visualization effectively demonstrated how average ratings varied, providing a clear visual representation of customer satisfaction trends. Additionally, clustering techniques helped group branches and product lines with similar characteristics, guiding targeted improvement strategies.

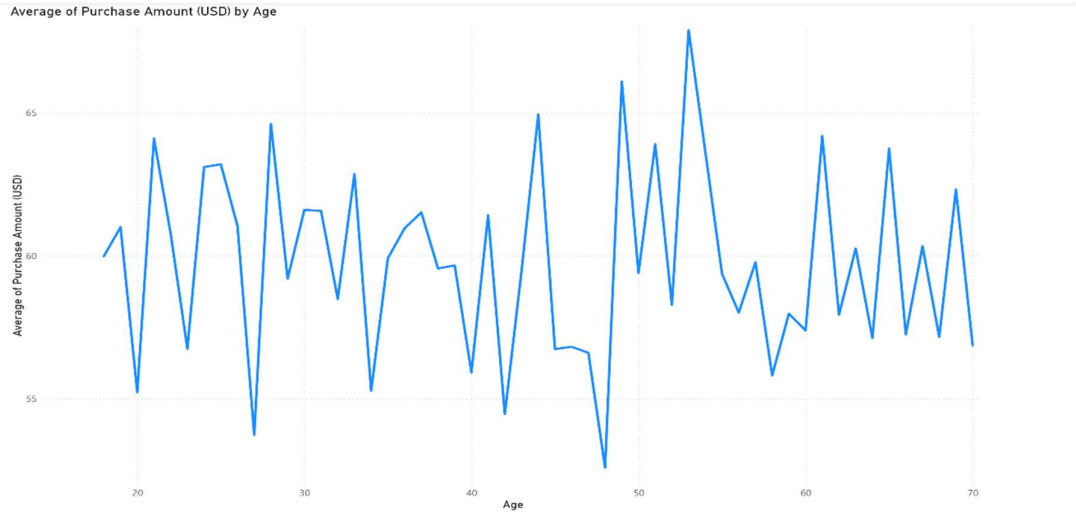
3.3 Question 3 (Naz): How does age influence the purchase amounts in customer shopping trends?

3.3.1 Objective

To see how customer demographics and purchasing behaviors segment into distinct clusters, and what insights we can derive from these segments

3.3.2 Data visualization process and techniques

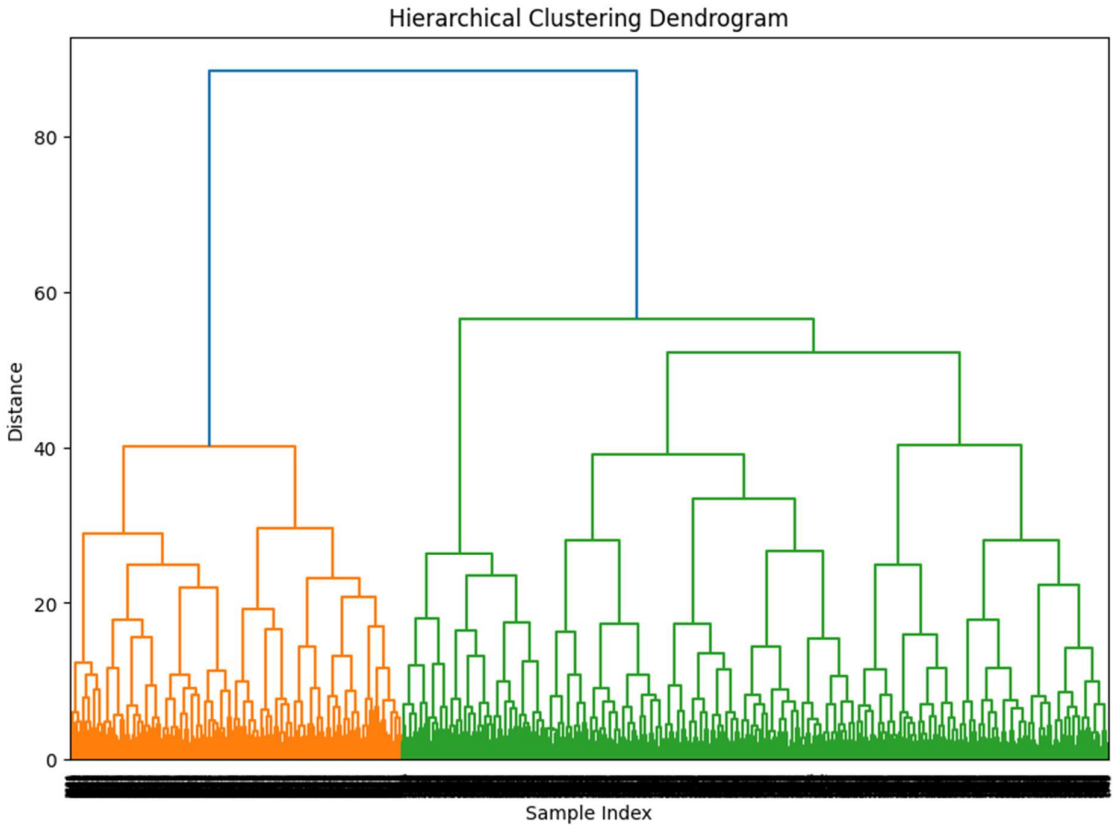
Our data visualization journey began with an exploratory data analysis (EDA) in Power BI. We imported the shopping trends dataset and created initial visualizations to understand the distribution of key variables, such as age and purchase amount. Using Power BI's intuitive interface, we generated line charts to observe average purchase amounts across different age groups, providing insights into age-specific spending behaviors. This initial EDA helped us identify patterns and trends, laying the groundwork for more sophisticated analyses.



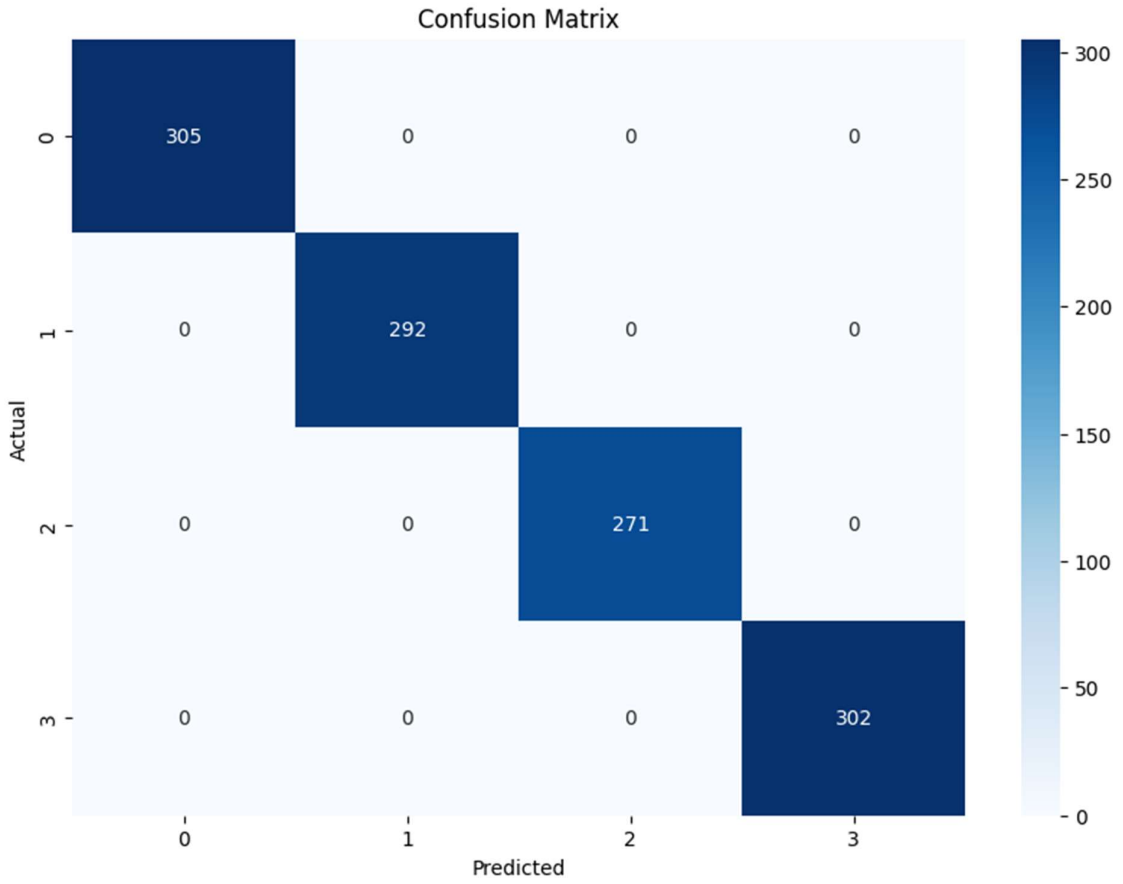
Following the initial EDA in Power BI, we transitioned to Google Colab to apply advanced dimensionality reduction and clustering techniques. We used Principal Component Analysis (PCA) to reduce the dataset's dimensionality, allowing us to project data points into a new space defined by the first two principal components. This step was crucial in revealing underlying structures within the data. The scatter plot generated from the PCA results

showcased the distribution of customers based on purchasing behaviors and product preferences, colored by product characteristics.

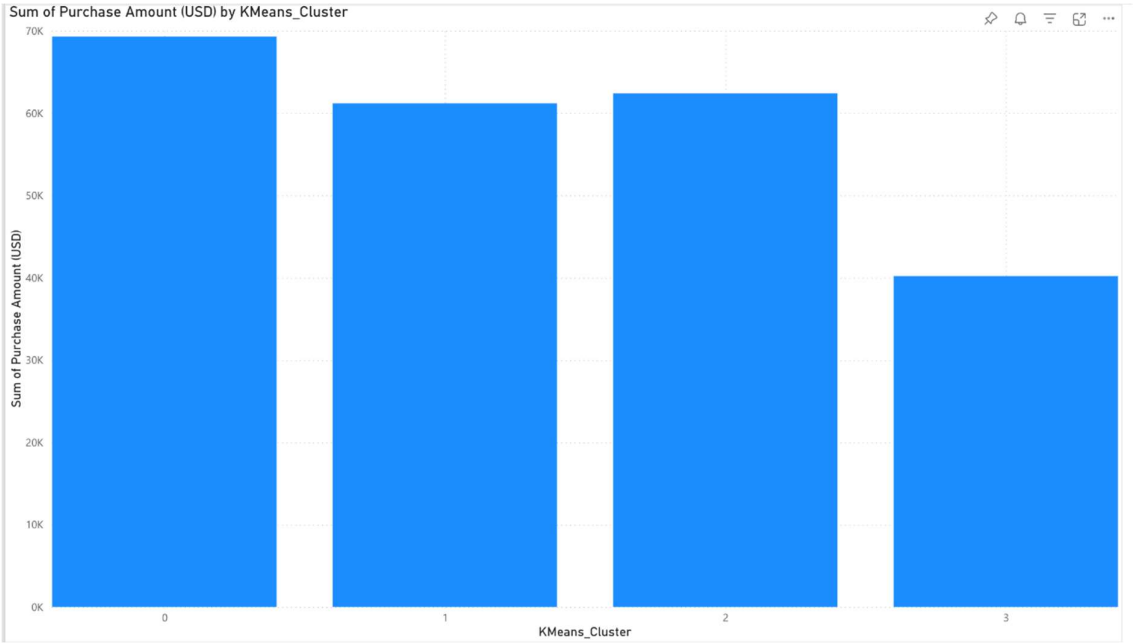
To delve deeper into customer segmentation, we performed hierarchical clustering, visualizing the results with a dendrogram. This dendrogram illustrated the hierarchical relationships between data points, helping us identify natural groupings within the customer base. We then employed k-means clustering, using the elbow method to determine the optimal number of clusters. The resulting scatter plot, colored by k-means clusters, provided a clear visualization of distinct customer segments. These clustering techniques allowed us to identify and analyze groups of customers with similar characteristics, providing a robust basis for targeted marketing strategies.



To further enhance our analysis, we built a Random Forest classifier to predict cluster membership based on customer demographics and purchasing behaviors. This model was trained and validated in Google Colab, with the confusion matrix demonstrating perfect accuracy in predicting cluster labels. The classification model confirmed the effectiveness of the chosen features in segmenting the customer base. By leveraging these insights, we developed a comprehensive understanding of distinct customer segments, enabling data-driven decisions for marketing and customer engagement strategies.



By combining the strengths of Power BI for initial EDA and Google Colab for advanced analytical techniques, we achieved a thorough and multifaceted analysis of the shopping trends dataset. This process allowed us to derive actionable insights and identify key customer segments with precision.



3.3.3 Discussion of visualization findings

The visualizations and analyses conducted reveal significant insights into customer purchasing behaviors and demographic trends. The initial exploration in Power BI highlighted how average purchase amounts vary across different age groups, indicating that age-specific marketing strategies could be highly effective. The line chart, for instance, showed peaks and troughs in spending, suggesting that certain age groups might be more inclined to spend more at different stages of their lives. This foundational understanding paved the way for deeper analyses.

Moving into more advanced techniques, the Principal Component Analysis (PCA) and clustering methodologies provided a clearer segmentation of the customer base. The PCA scatter plot illustrated distinct groupings based on purchasing behaviors and product preferences, which were further refined using hierarchical and k-means clustering. The dendrogram from hierarchical clustering revealed natural clusters, while the k-means

clustering identified four optimal segments, as indicated by the elbow method. These clusters, when visualized, showed distinct spending patterns and demographic characteristics. For example, Clusters 0 and 1 were identified as high-value segments, contributing significantly to the overall revenue, while Cluster 3 represented a lower-spending group. Understanding these clusters allows for targeted marketing efforts, such as personalized promotions for high-value segments and engagement strategies to boost spending in lower-value segments. The Random Forest classification model's perfect accuracy further validated these clusters, confirming that customer demographics and purchasing behaviors are strong predictors of cluster membership.

3.3.4 Practical recommendations

Based on the insights gained from our visualizations and analyses, several practical recommendations can be implemented to enhance marketing and customer engagement strategies. Firstly, focus on high-value segments identified in Clusters 0 and 1 by offering personalized promotions, loyalty programs, and exclusive deals to maximize their lifetime value and retention. For Cluster 3, which represents lower-spending customers, consider implementing targeted campaigns to increase engagement and spending, such as

introducing special discounts, product bundles, or tailored offers that meet their specific preferences. Additionally, leverage the detailed characteristics of each cluster to optimize product recommendations and inventory management, ensuring that the most popular items among high-value segments are readily available.

3.3.5 Limitations

While the analysis provided valuable insights, several limitations should be noted. The dataset used may not represent the entire customer base, potentially leading to biased results. Additionally, the clustering techniques assumed that the chosen features were the most relevant, potentially overlooking other influential factors. The perfect accuracy of the classification model, while promising, raises concerns about overfitting, indicating that the model may perform less effectively on unseen data. Future analyses should incorporate a broader range of features and datasets and apply cross-validation techniques to ensure the robustness and generalizability of the findings.

3.3.6 Conclusion

In conclusion, the data visualization and analysis journey effectively segmented the customer base into distinct clusters based on demographics and purchasing behaviors. By utilizing Power BI for initial exploration and Google Colab for advanced clustering and classification techniques, we identified key customer segments with distinct characteristics and spending patterns. These insights provide a strong foundation for targeted marketing strategies and personalized customer engagement initiatives. While there are some limitations to consider, the findings offer valuable guidance for optimizing marketing efforts.

3.4 Question 4 (Matthew): How does the unit price affect different product line ratings?

3.4.1 Objective:

Use clustering and classification learning to determine patterns and predictions between the

unit price of goods and their respective ratings.

3.4.2 Data visualization process and techniques

To begin the query, Power BI was used to perform EDA on the Supermarket_sales data set with a focus on the variables *Unit price, Ratings,* and *Product line.* Four distinct charts were made in Power BI:

1. Pie Chart: This chart helps visualize the sum of pricing of goods between each product line to determine if product lines have biased pricing towards certain lines.

2. Box plot: To identify outlying points of unit price in each given rating, this helped to show that no matter the rating, the price of the unit has roughly the same mean and interquartile range.

3. Hexbin plot: Due to the dense random scatter plot that rating vs unit price shows, a contour plot, Hexbin plot, was shown to average out the density in intervals to make visual inference easier for the reader.

4. Correlation Matrix Plot: This was used to see how correlated the two variables are and to set a baseline for how accurate our future finds should be.

5. Bar Graphs: To visualize the variance of each PCA and their cumulative summation.

6. Scatter Plots: To visualize the distribution of product lines when grouped by PCA. Furthermore, to show the clustering of k-means analysis.

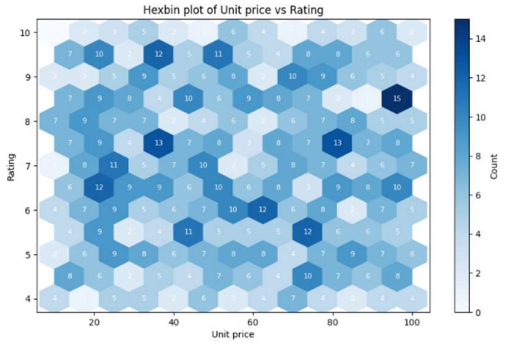
7. Line Chart: To illustrate the elbow graph to deduce a k for the k-means analysis.

3.4.3 Discussion of visualization findings

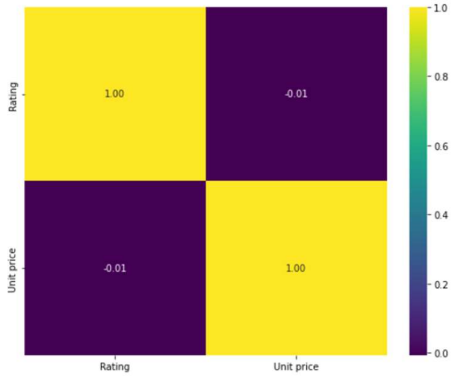
Exploratory Data Analysis

The pie chart shows that between all product lines, the unit pricing totals roughly the same between all categories. Fashion accessories are the most expensive at 10.17K and Health and beauty is the least expensive at 8.34K. This of course could not only be attributed to the pricing of the products, but the total amount of each product bought. The boxplot shows that there are no statistical outliers when compared to the interquartile ranges of each rating range. Interestingly there were no ratings below 4. The means of each rating range fall within a small margin as well as the Q1 and Q3 ranges.

Rating and Unit price

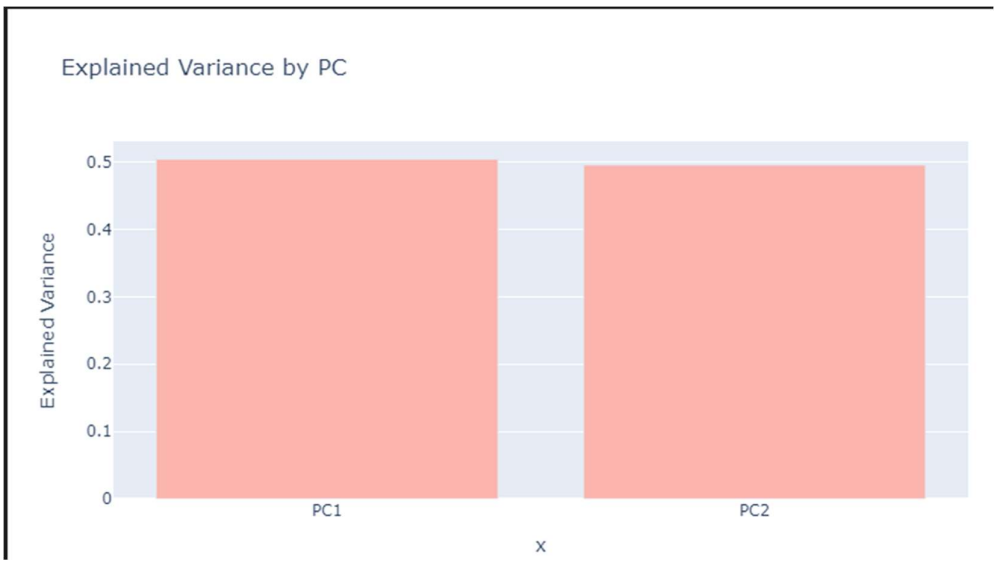


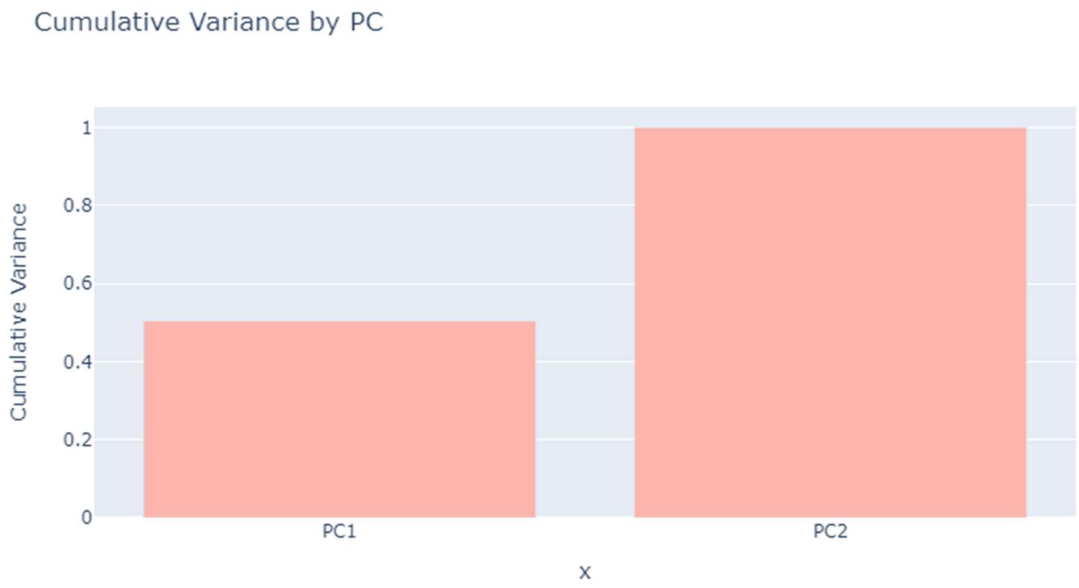
Rating and Unit price



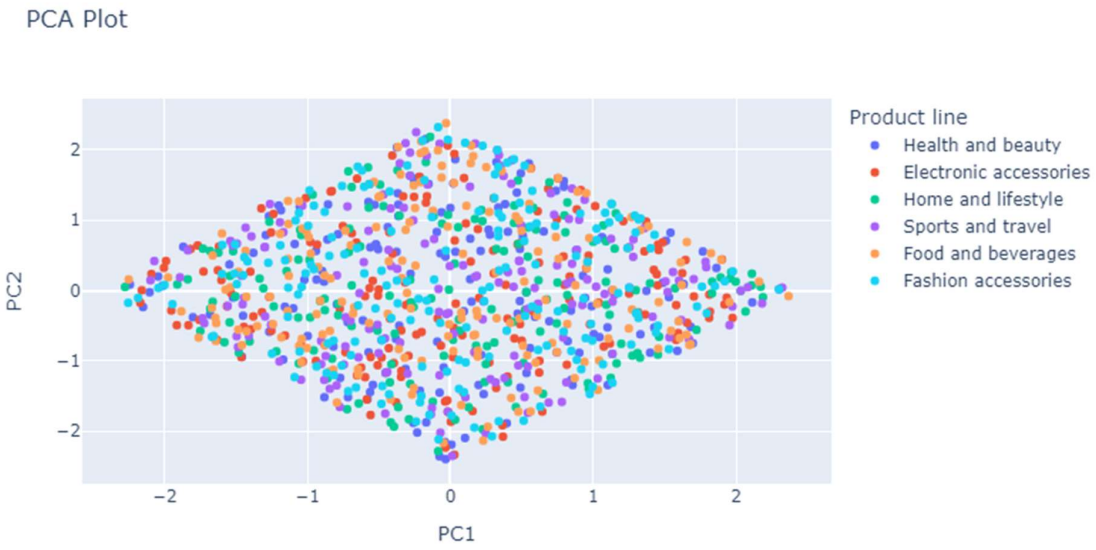
The hexbin shows that the scattering of ratings versus unit price is mostly random with low and high ratings being found in all areas of the graph. There is, however, a trend that to rating lower than 7 has reached more than 13 counts, with only one hex hitting the highest 15 count at rating 8.5. These 15 counts are found in the second highest unit pricing range and could maybe indicate that the higher the price the better rating a product could receive. Looking at the correlation matrix we see that Rating a Unit price and almost 0 correlation, we should expect random plots in our upcoming analysis.

Principle Component Analysis



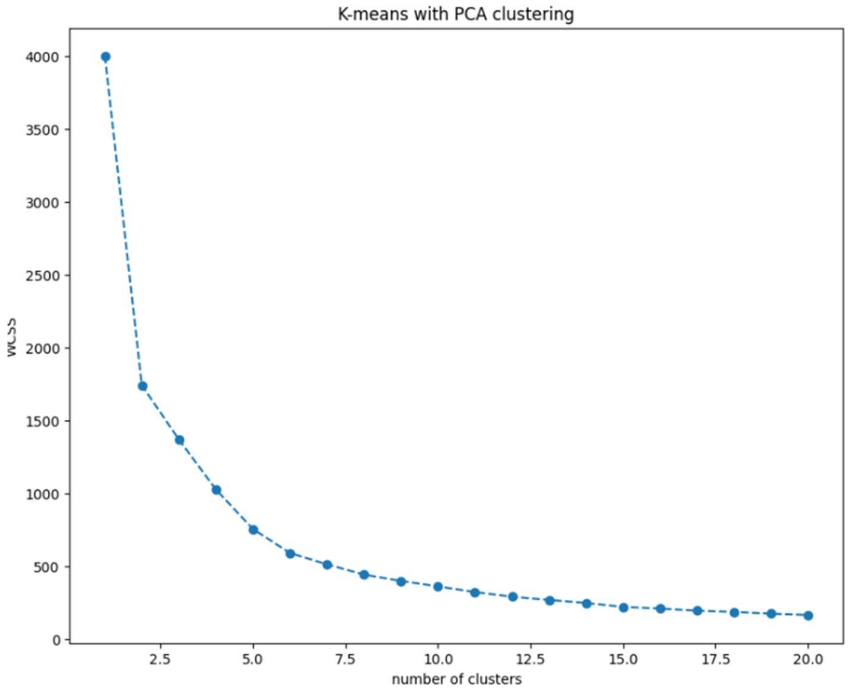


A PCA was done on the variables Ratings and Unit price. PC1 and PC2 share almost the same variance indicating that they share the same importance when modelling. This can be further illustrated in the cumulative variance plot showing that with just PC1 and PC2 100% of the variance is accounted for.

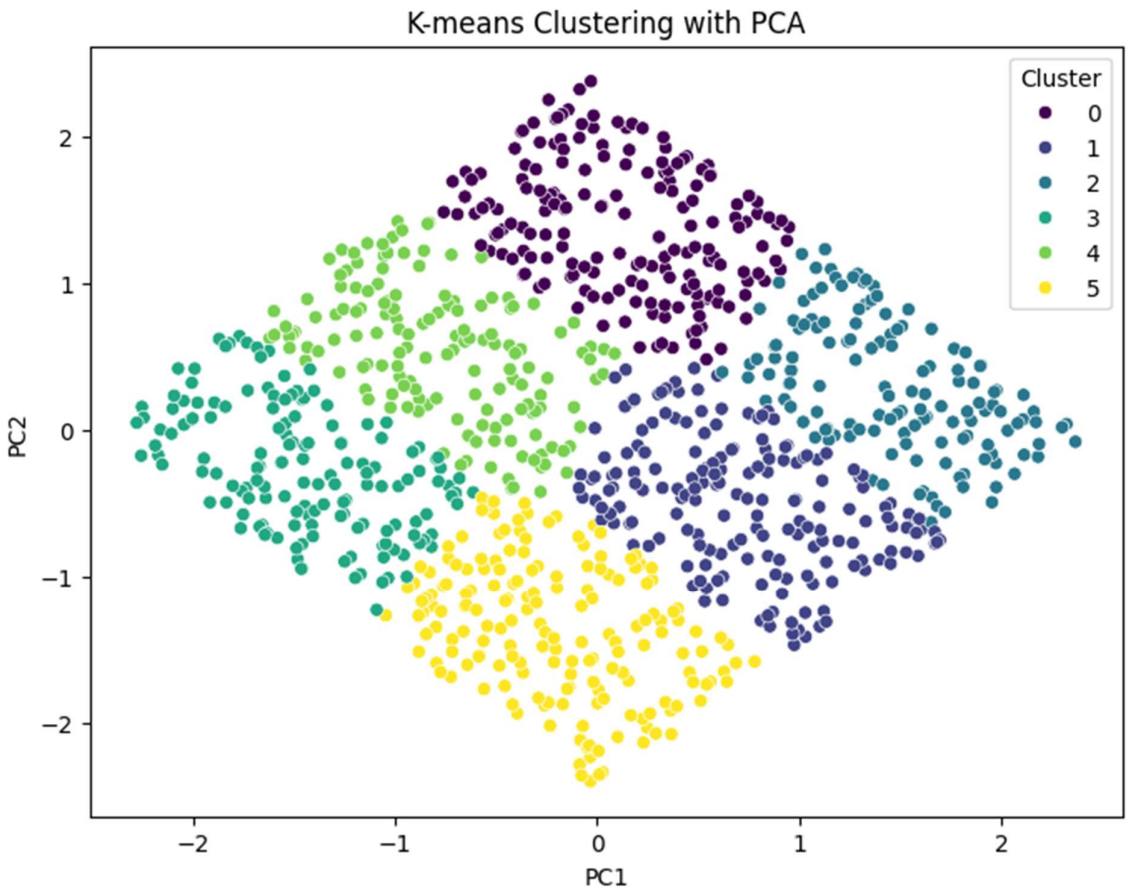


When we plot the PC1 vs PC2 and look at the product lines we see a very even and random spread of each product line. This is corroborated by the correlation plot done in EDA.

K-Means

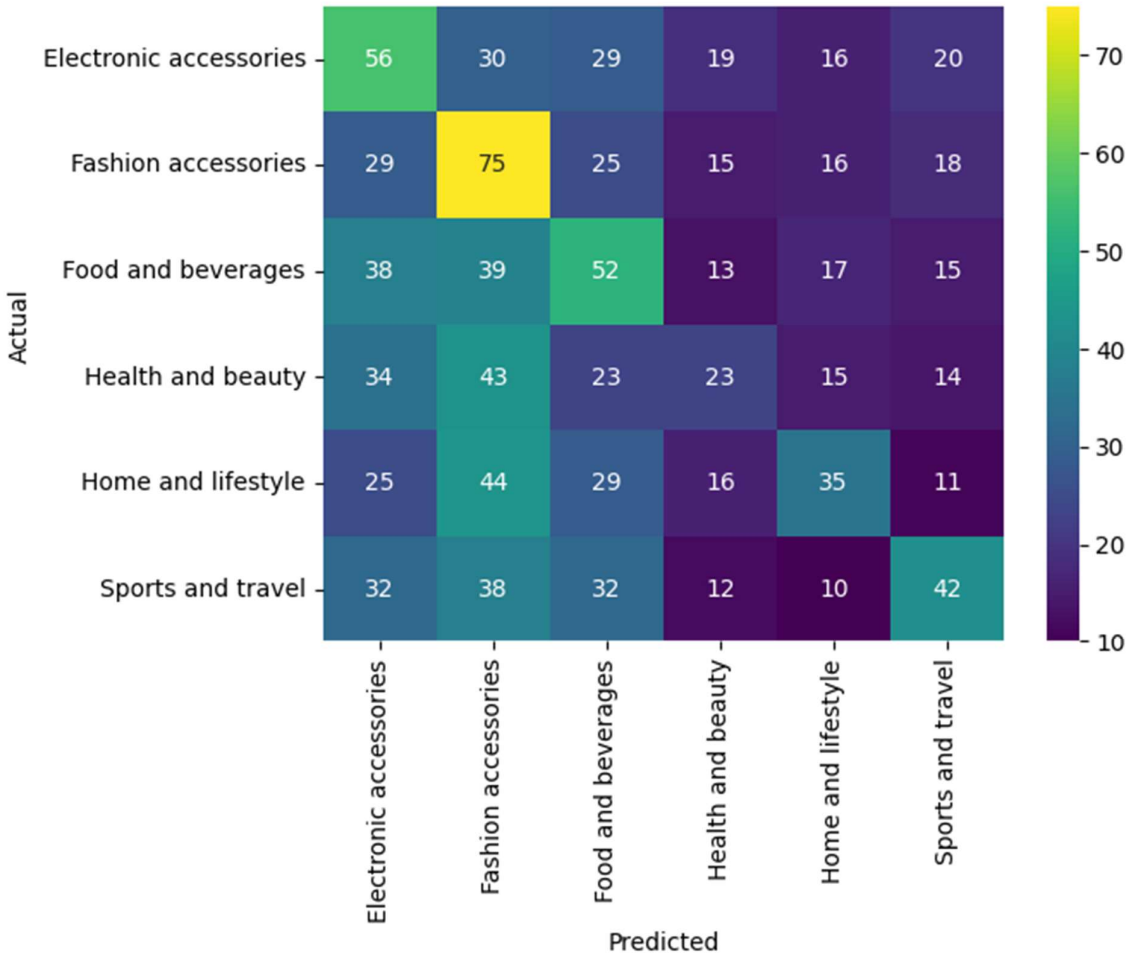


Before plotting K-Means, an elbow graph was made to determine the correct number of clusters to use. Due to the ambiguity and subjectiveness of using such a technique it would be reasonable to choose 2 or around 6 clusters. For the sake of this experiment and the number of product lines we will choose n=6.



When looking at the k-means clustering with PCA we see 6 extremely distinct clusters. This pattern could indicate that the relationship between Ratings and Unit price is not linear, allowing our correlation coefficients to be close to 0 while still maintaining this extreme clustering separation.

K-Nearest Neighbours



The confusion matrix of the k-nearest Neighbours when k=22. k = 22 was chosen by using the formula: k = sqrt(n/2) where n=1000. Here we see a multitude of incorrect predictions that yield a miscalculation rate of 0.717 or roughly 71%. This model is not accurate. Interestingly if k=3 this miscalculation rate drops steeply to roughly 51%. This could be due to more precise fitting between nodes.

3.4.4 Practical recommendations

For the Stores:

Review Pricing Strategies for Health and Beauty Products: Since Health and Beauty products have the lowest total unit pricing, the supermarket could explore strategies to boost sales in

this category. This could include promotional campaigns, bundling offers, or loyalty programs targeting these products to increase their purchase frequency.

Target Marketing Campaigns Based on Product Lines: Given the even spread of product lines across the principal components and K-means clusters, the supermarket could use targeted marketing strategies that cater to the specific preferences of different customer segments identified in the clusters.

Improve Products with Lower Ratings: Products with ratings below 7 but high unit prices should be reviewed for quality improvements. Addressing customer concerns about these products could help boost their ratings and sales.

For this Analysis:

Improve Predictive Models: The high misclassification rate in the K-nearest Neighbours model indicates room for improvement. Exploring other algorithms, hyperparameter tuning, and incorporating additional features could help build more accurate models for predicting product ratings and customer preferences.

3.4.5 Limitations

Temporal Aspects: The data appears to be from a specific time period. Customer preferences and market conditions can change over time, so the analysis might not reflect current trends.

Simplistic Model Assumptions: Using PCA reduces dimensionality but also simplifies the data, which might result in loss of important information. Similarly, the elbow method for determining the number of clusters can be subjective and may not always lead to the best number of clusters.

K-Nearest Neighbors (KNN) Performance: The KNN model shows a high misclassification rate, indicating that it might not be the best model for this dataset. Other models and techniques should be explored to improve predictive accuracy.

3.4.6 Conclusions

Exploratory data analysis revealed no significant outliers in unit price across ratings, and a weak correlation between unit price and rating. While principal component analysis (PCA) and K-means clustering found distinct groupings, the high misclassification rate in K-nearest neighbors (KNN) suggests this data may be better suited for other machine learning models. Future analysis could explore these limitations by incorporating time-based data, more complex models, and additional features to improve prediction accuracy and provide actionable insights for the supermarket.

3.5 Question 5 (Javier): Is it possible to identify patterns of payment options by analyzing quantity and unit price?

3.5.1 Objective

Evaluate clustering of Payment Options by comparing the cost of the item and the quantity sold from a line of product.

3.5.2 Data visualization process and techniques

The visualization of the data was performed in two parts. The first one was the visualization of the Quantity and Unit Price variables and its relationship with respect to the Payment options.

Power BI was used to generate a pie chart about the percentage of the quantity of products

sold by different payment options.

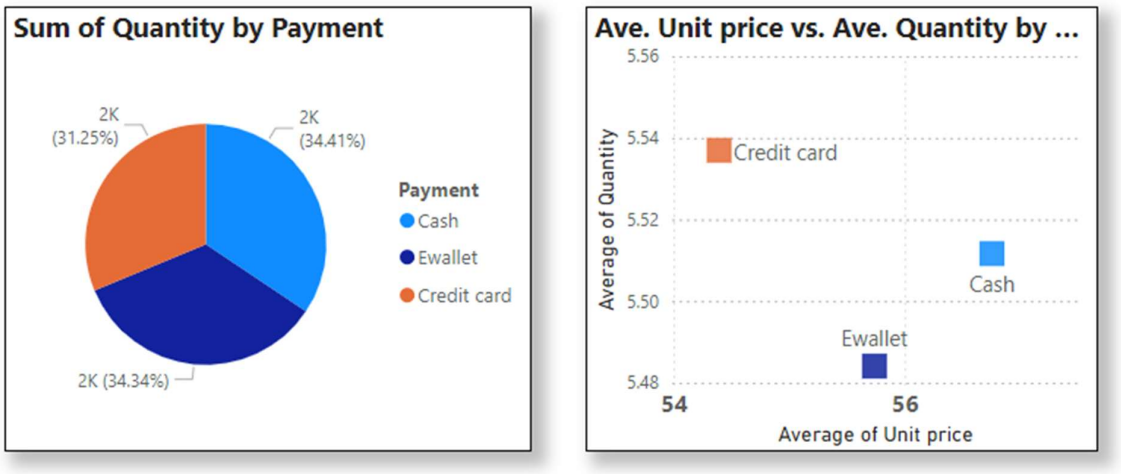


Figure 5.1. Pie Chart and Average Unit Price vs. Quantity by Payment Option

Since the Payment Option seemed to be evenly distributed among the sample, ~33% each, and the average of Unit Price vs. Quantity was close to each other, it was pertinent to evaluate the relationship between the two variables.

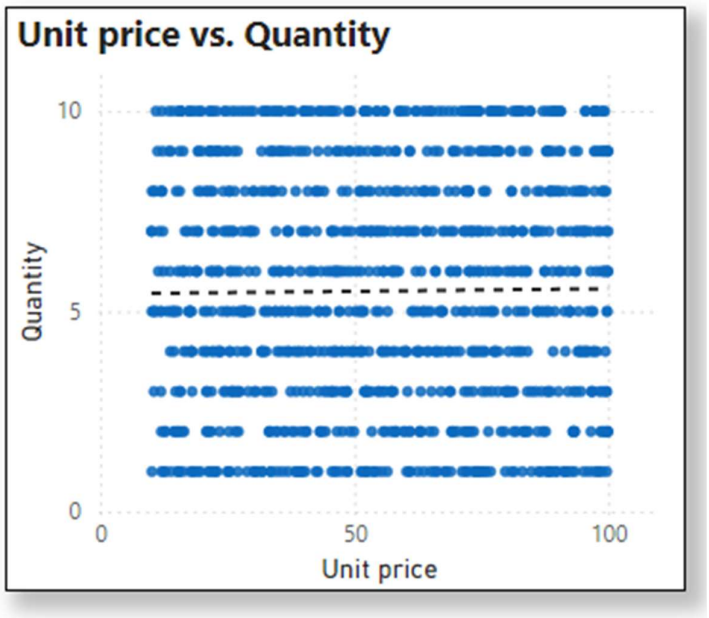


Figure 5.2. Disperse graph of Quantity vs. Unit Price

A linear trend line was plotted in Power BI to visualize any correlation between the variables. It was evident that the data did not offer any visible correlation between them.

The second set of plots were generated in Python-Colab while performing the Principal Component (PC) analysis for the data and the K-Neighbor predictive analysis.

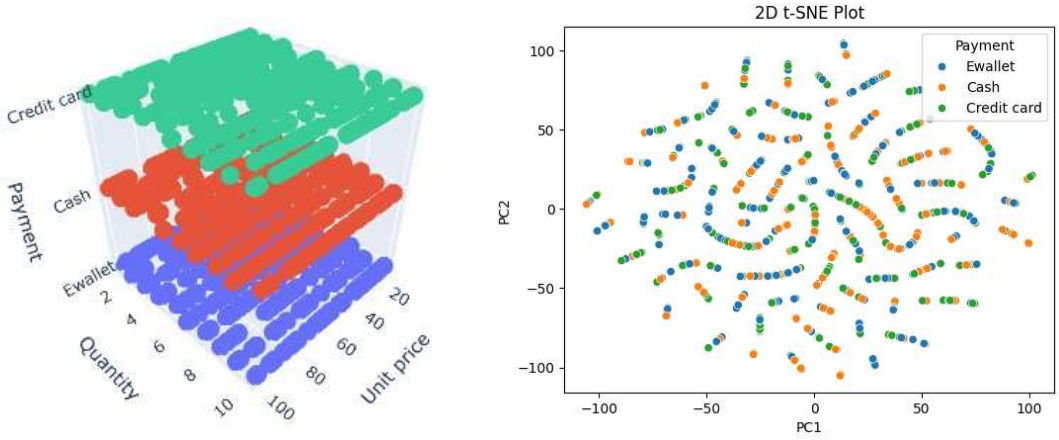


Figure 5.3. Principal Component Cluster Evaluation

Finally, a confusion table with the predictive analysis of the k-Neighbor model was used to represent the accuracy of the prediction of method of payment when knowing the quantity and the unit price of the article.

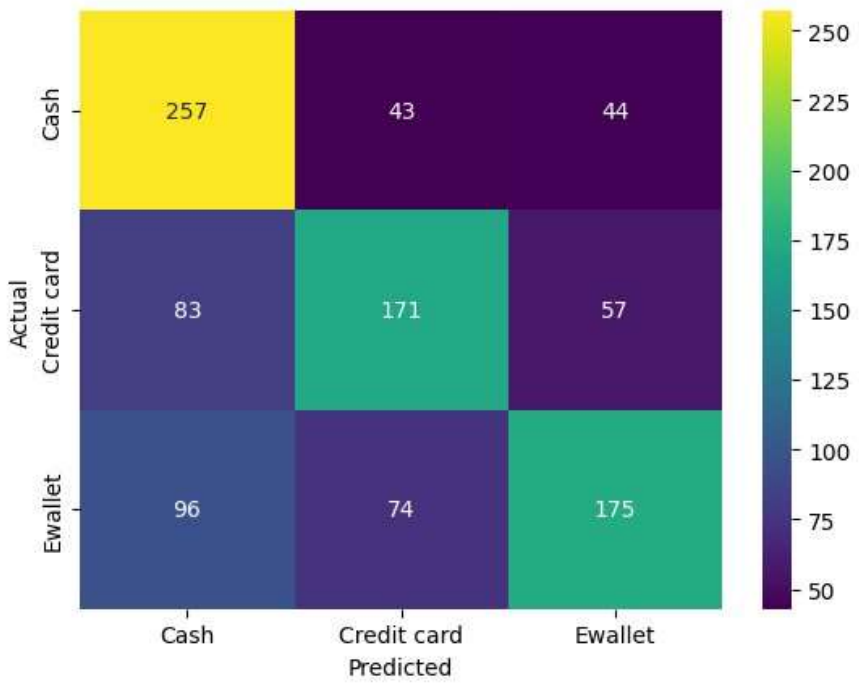


Figure 5.4. Confusion table k-Neighbor Prediction

3.5.3 Discussion of visualization findings

The variables selected for the evaluation, Unit Price, and Quantity, did not show evidence of correlation as observed in Figure 5. 1. and 5.2. The dispersion of the data and the flat regression line in the graph shows no correlation between the variables.

Once the PC analysis started, which aimed to determine clustering of the data based on these two variables, it was not surprising to continue reaching to the conclusion that the data was not correlated even at the various levels of Payment Options as observed Figure 5.3.

PC1 and PC2 did not help to separate the data even after performing data standardization as observed in Figure 5.3.

Finally, the Confusion Matrix, Figure 5.4., created between the actual and the predicted values for the K-Neighbor model helped us to predict with a 60% accuracy the Payment method

knowing the unit price and quantities.

3.5.4 Practical recommendations

It is beneficial for PC evaluations to determine if any correlations are present between the variables we want to review. PC evaluations usually reduce the dimensions trying to find clusters based on the relationships between the data. Therefore, if no correlation is observed for the study variables, it is better to capture characteristics from other variables. Also, I encountered difficulties in implementing this method with independent categorical variables. It is recommended to use non categorical variables when using the Principal Component analysis.

3.5.5 Limitations

The almost perfect distribution of the data hampered our investigation of finding relationships and prediction of the Payment Option by using the Quantity and Unit price. Based on this uncommon characteristic of the dataset, I suspect the data was created and does not correspond to actual Supermarket data. It is better to find actual data to perform these analyses.

3.5.6 Conclusion

- Total Quantities per Payment Option are very evenly distributed.
- Payment options are slightly influenced by the relationship between Unit Price and Quantity of the purchase. However, the differences are very subtle.
- There is no evident correlation between Quantity and Unit Price.

- Categories did not show clustering as all products are evenly distributed in unit price and quantity
- Principal Component (PC) analysis did not reveal any distinct clustering of the products after transforming the data and creating two projections based on PC1 and PC2.
- 60% accuracy of the Payment method prediction by using the k-mean model.

4 Conclusions

- From question (1), based on the findings, with a relatively balanced distribution of sales across product lines, it is recommended to consider investing in targeted product lines that show slightly higher sales volumes and profits.
- From question (2), Variability in ratings across branches and product lines highlighted areas for potential improvement, such as the need for better customer service or product quality enhancements in specific branches or product lines. However, Certain branches, such as Branch A, consistently received higher ratings for specific product lines like Electronics, indicating strong performance in these areas.
- From question (3), the cluster analysis for Age vs. Quantity help us to identify high-value segments that contributed to the overall revenue (Cluster 1 and 2), while Cluster 3 represented a lower-spending group. Understanding these clusters allows for targeted marketing efforts, such as personalized promotions for high-value segments and engagement strategies to boost spending in lower-value segments
- From question (4), the high misclassification rate in K-nearest neighbors (KNN) suggests this

data may be better suited for other machine learning models. Future analysis could explore these limitations by incorporating time-based data, more complex models, and additional features to improve prediction accuracy and provide actionable insights for the supermarket.

- For question (5), it was evident that no correlation existed between the quantity and the unit price in the given dataset. This observation was clear from the preliminary data analysis using Power BI and further confirmed by a deeper Principal Component Analysis (PCA) performed with Python in Google Colab. However, it was interesting to note that the k-means clustering model achieved a 60% accuracy in predicting the payment method.

5 Reference

1. Supermarket sales. (2019, May 27). Kaggle. https://www.kaggle.com/datasets/aungpyaeap/supermarket-sales