FALL
2024

DATA INTENSIVE COMPUTING [CSE 587 C]

PROJECT REPORT

PROFESSOR: CHEN
XU

ADNAN SHAHID SADAR (50592332)
BRUNDA VENKATESH (50610164)
MOHAMMED ABDUL AFTAB MUDDASIR (50604245)
RACHANA RAMESH (50596083)

INDEX

HIGHLIGHTS

Project Overview:

Title: ***Customer Segmentation and Purchase Behaviour Prediction for Retail Businesses.***

Problem Statement: Forecasting purchase behaviour at a segment level using demographic and transactional data.

Objective: Develop predictive models to provide insights for marketing and inventory optimization, focusing on customer segmentation.

Key Questions and Hypotheses:

Demographic Influence:

- Younger customers prefer fashion and technology.
- Urban shoppers tend to buy higher-priced products.

Revenue Insights:

- Categories like clothing and technology contribute the most revenue.
- Seasonal trends affect quarterly revenues.

Customer Behaviour Analysis:

- Middle-aged customers (31–50 years) shop the most and prefer family-oriented items.
- Gender influences preferences (e.g., women lean towards clothing, men towards technology).

Methodology:

1. **Data Preparation**:

- Cleaning and preprocessing of data (missing values, outliers, normalization).
- Dimensionality reduction using PCA.

2. **Exploratory Data Analysis (EDA)**:

- Found key trends in demographic influences and revenue patterns.

3. **Algorithms and Models**:

- **K-Means Clustering**: Identified customer segments by age and purchase behaviour.
- **Decision Trees and Gradient Boosting**: Predicted customer behaviour and purchase patterns.
- **Logistic Regression and Bayesian Networks**: Explored relationships between demographics and high-value purchases.
- **KNN and ARIMA**: Modelled seasonal revenue trends.

Key Findings:

1. **Customer Segmentation**:

2. **Revenue Trends**:

- Clothing and technology products drive sales.
- Revenue peaks in specific seasons, highlighting inventory and marketing opportunities.

3. **Demographics and Behaviour**:

- Gender and age significantly influence purchasing preferences.
- Urban shoppers show distinct behaviour, favouring higher-priced items.

4. **Algorithm Performance**:

- Gradient Boosting and Decision Trees showed high accuracy for segmenting customers.
- ARIMA struggled with revenue forecasting, suggesting potential areas for improvement.

Application and Deployment:

Developed a user-friendly web application for insights:

- **Tools**: Python, MySQL, Amazon RDS, Streamlit.
- Hosted publicly for real-time predictions and visualization.
- Application link: https://customer-insight-engine.streamlit.app/

Conclusion and Recommendations:

- **Actionable Insights**:
    - Use customer segmentation for targeted marketing.
    - Align inventory strategies with demographic and seasonal trends.
- **Future Work**:
    - Explore deeper customer personalization using additional features.
    - Enhance predictive accuracy by integrating more advanced models.

PROJECT REPORT

Project Title:

Customer Segmentation and Purchase Behaviour Prediction for Retail Businesses

Problem Statement:

Purchase Behaviour Prediction at Segment-Level: Using Demographic and Transactional Data to Identify Trends in Retail Customer Purchases Across Different **Categories.**

Objective:

Develop a predictive model that forecasts purchase behaviour not for individual customers, but for similar customer segments (e.g., age group, gender, location).

Phase 1 Summary:

During this stage, the project examined customer buying behaviours to identify trends at the segment level. Data cleaning and preprocessing included handling missing values, addressing outliers, and feature engineering for Total_Spent and age groups.

EDA supported the confirmation of the following key hypotheses:
Demographic Influences: Age and gender have strong influences on category preferences-for instance, younger customers would tend more towards technology, females more to clothing and beauty.
Regional Variations: High-value categories enjoyed better sales in urban areas, thus demanding location-based strategies.
Revenue Insights: Clothing and technology drove the highest revenue; however, quarterly revenue trends brought out seasonal opportunities for improvement.

The findings set up a very strong foundation for predictive modelling at the segment level and provide actionable insights for inventory optimization and marketing strategies.

Questions and Hypothesis:

Q1. How do age and gender affect customer preferences for different product categories?
Hypotheses: Younger customers may prefer fashion items, while older customers may lean toward household products. Gender might also influence the likelihood of purchasing specific categories.

Q2. Do customers in different locations (shopping malls) show distinct purchasing behaviours?
Hypothesis: Customers from malls located in urban areas may buy higher-priced products compared to those in suburban or rural areas.

Q3. Which product categories are generating the highest sales, indicating a need for potential adjustments in stocking strategies?
Hypotheses: Understanding which product categories are generating the highest sales allows businesses to pinpoint their top-performing items. By knowing which categories are selling well, businesses can adjust their stocking strategies to ensure that high-demand items are adequately stocked. This helps prevent stockouts, reducing lost sales opportunities, and overstock situations that lead to increased holding costs.

Q4. How do the revenue trends across different quarters in 2021 and 2022 compare, and how can quarterly performance of past years be helpful to plan for the future of business?

Hypothesis: Understanding the revenue performance across different quarters provides valuable insights into seasonal trends and purchasing behaviours. Businesses can use this information to identify peak periods for sales and plan inventory and marketing strategies accordingly.

Q5. How do demographic factors (like gender and age) influence purchasing behaviour in different categories?
Hypotheses: Female customers are more likely to purchase clothing and beauty products compared to male customers

Q6. What is the relationship between payment method and purchasing volume in different shopping malls?
Hypotheses: Customers who use credit cards or debit cards tend to purchase a higher quantity of items compared to those using cash.

Q7. How does age affect the likelihood of purchasing high-value products in the dataset?
Hypothesis: Older customers are more likely to purchase high-value products due to higher disposable income and established financial stability compared to younger individuals.

Q8. Does gender influence the likelihood of making high-value purchases?
Hypothesis: Female customers are more likely to make high-value purchases compared to male customers, possibly due to preferences in certain product categories (e.g., clothing or shoes) that tend to have higher price points.

Phase 2 Summary:

Our investigation uses advanced data analytic tools to better understand client behaviour and segmentation. We used **K-Means Clustering** to identify client categories based on demographics and product preferences, and then used the Elbow Method to determine the best number of clusters. Statistical tests, such as **Chi-Square** and ANOVA, were used to investigate the impact of age, gender, and geography on purchase behaviours, and no significant relationships were found in the dataset. Additionally, PCA was used to minimize data dimensionality and visualize purchasing trends across locales, which showed little change between them. These findings imply that demographic and regional factors may not have a significant impact on client choices, highlighting the significance of investigating other variables, such as spending habits, for more tailored marketing techniques. This comprehensive method offers excellent advice for optimizing inventory management and customer engagement.

Our team investigated different algorithms to test hypotheses about consumer behaviour and revenue patterns. We used **Decision Tree** and Gradient Boosting algorithms to segment customers and stock products. These algorithms were highly interpretable and accurate, including tuning methods such as Gini Index and max_depth for Decision Trees and hyperparameter optimization for Gradient Boosting. To investigate seasonal revenue trends, we **used K-NN Regression** and ARIMA, with an emphasis on quarterly revenue patterns. Tuning parameters such as the number of neighbours in K-NN and the best p, d, and q values in ARIMA resulted in optimal model performance. While Gradient Boosting and Decision Trees had great predictive ability, ARIMA failed with early revenue estimates, indicating areas for development. These models offered insights on stocking tactics and seasonality trends for enhanced business planning.

Using advanced modelling tools, our team tested two crucial hypotheses about customer behaviour. To test the first hypothesis, we used Logistic Regression, modified with balanced class weights and outlier management, to determine how age effects the likelihood of high-value transactions. The model performed well with high precision and recall, as evidenced by ROC AUC scores, and provided insights into the link between age and purchasing behaviour. To test the second hypothesis, we employed a Bayesian Network trained with Maximum Likelihood Estimation (MLE) to examine gender's influence on high-value purchases. This method highlighted probabilistic relationships and indicated unique purchase patterns between genders. Visualizations like ROC curves and stacked bar charts improved the interpretability of the results, highlighting the potential for targeted marketing campaigns based on demographic information.

Phase 3:

Question 1: Who are the top customers contributing the most revenue?

Hypothesis: Customers with the highest purchase frequency and higher average transaction value are the top contributors to total revenue.

Potential: Identifying top revenue-contributing customers using RFM analysis aligns with business goals of customer retention and revenue optimization. It's a practical approach to segment high-value customers and implement targeted marketing strategies.

Popularity: The use of RFM is a standard practice in CRM, and there are many citations for it in marketing literature. For example, one study applied RFM analysis with clustering methods to create a precise customer profile.

https://iieta.org/journals/mmep/paper/10.18280/mmep.100135

Background: RFM has very strong historical roots, from its initial direct marketing origins to computationally intensive methods such as clustering; hence, it is one of the robust ways to find out patterns in customer behaviour.

https://iieta.org/journals/mmep/paper/10.18280/mmep.100135

Controversy: Only focusing on monetary value may b expressed by M_Score, as customers with small but frequent purchases could also be loyal ones. Works have debated whether the estimation of CLV is more effective in long-term customer retention strategies than border metrics.

https://iieta.org/journals/mmep/paper/10.18280/mmep.100135

Data Complexity: The question involves moderate data complexity: Extracting transaction data to get revenue and frequency metrics. Integration of other metadata, such as shopping mall, provides further detail.

Question Depth: Moderate depth: It considers explicit filtering criteria based on monetary and overall RFM scores. It is of limited hypothesis depth but is effective for finding high-value segments under defined parameters.

Creativity: The question is conventional but effective for business intelligence, we can enhance creativity by: Combining revenue data with quantitative factors (e.g., customer feedback). Using advanced visualizations like heatmaps for shopping patterns.

Question 2: How many customers fall into each RFM segment, and what is their behaviour?

Hypothesis: Customers are segmented into distinct RFM (Recency, Frequency, Monetary) categories, and each segment exhibits specific behaviours regarding purchasing patterns, engagement, and overall value to the business.

Potential: Relevance: Very important for focused marketing and resource allocation in various industries, such as retail and e-commerce. Impact: Helps optimize customer retention strategies and enhance overall profitability.

Popularity: RFM analysis is prevalent in CRM and marketing studies. Its effectiveness is documented in various applications, such as segmenting customer behaviour in retail.

https://doi.org/10.32628/IJSRST2183118

International Journal of Contemporary Economics and Administrative Sciences ISSN: 1925 – 4423 Volume :8, Issue: 1, Year:2018, pp. 1-19

Background: There is a lot of history related to RFM in marketing analytics, from direct marketing to modern data-driven methods. It leverages customer transaction data to inform business decisions.

https://doi.org/10.32628/IJSRST2183118

Controversy: Debate on Effectiveness: Some researchers argue that RFM over-simplifies customer behaviour. Critics suggest that other models, such as CLV analysis, may yield more granular insights.

https://doi.org/10.32628/IJSRST2183118

Data Complexity: Moderate Complexity: Requires structured transaction data with handling for missing or inconsistent entries. Efficient Segmentation implies data pre-processing

https://doi.org/10.32628/IJSRST2183118

Question Depth: Moderate to High Depth: This covers the customer segments categorization into the RFM segment and studying their behaviours, which do not require the formulation of many hypotheses.

https://doi.org/10.32628/IJSRST2183118

Creativity: Though common, the inclusion of interactive visualizations with filtering increases the analytical depth and user engagement.

Q3. How do age groups influence the quantity of products purchased and their preferred product categories?

Hypothesis

1. **Age and Purchasing Quantity**:

- Middle-aged customers (31–50 years) are likely to purchase the highest quantity of products as they tend to shop for family needs, essentials, and bulk items.
- Young adults (18–30 years) may buy fewer products but could focus on specific, trendy categories.

2. **Age and Product Preferences**:

- Young adults (18–30 years) are expected to prefer fashion-oriented categories like Clothing and Technology.
- Middle-aged customers (31–50 years) might prioritize Food & Beverage or Household items due to family responsibilities.
- Seniors (51+ years) could lean toward Souvenirs and products related to leisure or personal care.

Significance

1. **Targeted Marketing Campaigns**:

- Understanding how product preferences vary by age group allows businesses to design highly personalized marketing strategies. For example:
    - ▪ Offer discounts on Food & Beverage for middle-aged customers.
    - ▪ Promote trendy gadgets to young adults.
    - ▪ Highlight Souvenirs for older customers, especially during holidays or travel seasons.

2. **Optimized Inventory and Product Placement**:

- Insights into the most purchased categories by age group enable businesses to:
    - ▪ Allocate shelf space more effectively.
    - ▪ Stock age-specific items in higher quantities.

    - Position relevant products in stores and online for higher visibility.

3. **Enhanced Customer Retention**:

- By tailoring promotions and product offerings to age-group preferences, businesses can improve customer satisfaction and loyalty.

Potential: The question has the ability to help understand customer behaviours by discovering patterns in purchase behaviours across age groups. Insights collected can help firms modify product offers, marketing strategies, and inventory management to fit the tastes and wants of specific demographic groupings, resulting in increased sales and customer satisfaction.

Popularity: The question of how age groups influence product quantities and favoured categories is a common one in consumer behaviour research. Research has demonstrated that different age demographics have varied purchasing behaviours and product preferences, which are critical for efficient market segmentation and targeted marketing strategies.
https://www.bls.gov/opub/btn/volume-4/consumer-expenditures-vary-by-age.htm

Background: The question investigates how different age groups influence purchasing patterns and product preferences, which is a major emphasis in consumer behaviour research. According to research, age has a considerable impact on purchasing behaviours, with each demographic having unique tastes and spending patterns.
https://www.edgecrm.app/articles/how-demographics-influence-customer-behavior-and-buying-habits

Controversy: Relying entirely on age-based segmentation has caused discussion, as it risks overgeneralization and overlooking specific customer peculiarities. Critics contend that such division perpetuates preconceptions and fails to account for the diversity of behaviours within age groups.
https://www.skyword.com/contentstandard/closing-the-age-gap-why-marketers-shouldnt-just-rely-on-generational-segmentation/

Data Complexity: The question includes several variables, such as age, product categories, and purchase quantities, which can be studied over time to detect patterns. To develop meaningful linkages across complex datasets, tables including demographic and sales data may need to be joined.

Question Depth: The question addresses conditions such as how age effects purchasing patterns and product preferences, which may involve several theories. For example, specific age groups may favor certain categories or purchase larger quantities, resulting in a series of hypotheses to test and validate.

Creativity: The question takes a novel approach by investigating the subtle link between age groups and their purchasing habits, which is sometimes missed in traditional analysis. Its surprise aspect is the ability to disclose unanticipated tastes and purchasing habits among varied demographics, prompting novel marketing methods.

Q4. What customer behaviours (age, gender, price sensitivity, and quantity purchased) predict product category preferences?

Hypothesis:

- Customers who purchase high quantities and expensive products are more likely to prefer categories such as Technology or Clothing.
- Gender influences category preference, with women favouring categories like Clothing and Cosmetics, and men leaning towards Technology and Food & Beverages.

- Age plays a role, with younger customers favouring Toys and Fashion, while older customers might prefer Souvenirs or Household items.

Significance:

Personalized Product Recommendations:
By understanding the factors influencing category preferences, retailers can recommend products tailored to individual customer profiles, increasing the likelihood of purchase.

Targeted Promotions and Discounts:
Businesses can design promotions (e.g., discounts on cosmetics for women or tech gadgets for younger men) based on behavioural trends, boosting the effectiveness of promotional efforts and overall customer satisfaction.

Potential: This question has great promise because it immediately influences targeted marketing strategies and product recommendations, allowing firms to better align their offers with client preferences. Research by Kotler and Keller (2021) emphasizes the importance of segmentation and tailored techniques in increasing sales and customer satisfaction.

Popularity: The question is extensively researched in marketing analytics and consumer behaviour research. According to a McKinsey & Company (2022) analysis, personalization using behavioural data greatly enhances conversion rates, making this a topic of interest in both academics and industry.

Background: Consumer behaviour research has long sought to understand how demographic aspects (e.g., age and gender) and economic characteristics (e.g., price sensitivity) influence purchasing decisions (Solimon, 2018). This extends previous theories such as Maslow's hierarchy of requirements, demonstrating how preferences correlate with individual and economic conditions.

Controversy: The question of whether tastes are influenced more by demography or individual personality qualities is hotly debated. Critics contend that relying too much on demographic data might lead to stereotyping, as emphasized by Noble and Noble (2019), who propose for a more nuanced approach that incorporates behavioural and psychographic data.

Data Complexity: Answering this question necessitates combining complicated statistics from several sources, such as purchase history, demographic profiles, and behavioural trends, which frequently necessitate advanced machine learning models for effective analysis. Data sparsity and balance biases add to the analysis's complexity.

Question Depth: The subject is quite deep, since it investigates the interaction of several aspects such as age, gender, and price sensitivity with category preferences, necessitating an interdisciplinary approach that blends economics, psychology, and data science insights.

Creativity: The topic demonstrates ingenuity in integrating standard demographic and economic indicators with behavioural insights to forecast preferences, enabling the adoption of novel data modelling tools and yielding meaningful commercial insights.

Approach:

Data Retrieval:
We processed the unstructured data obtained from Kaggle (https://www.kaggle.com/datasets/mehmettahiraslan/customer-shopping-dataset) to ensure its quality for analysis.
Duplicate Removal: Identified and eliminated redundant records.
Missing values: Addressed null entries by either removing incomplete records or imputing missing data based on statistical measures.

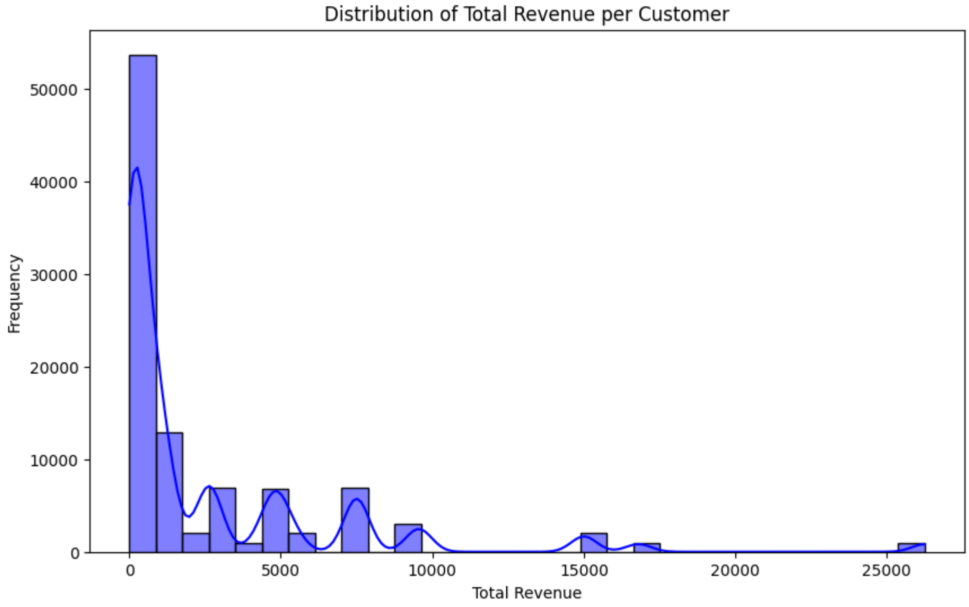
String Standardization: Resolved inconsistencies in categorical data formats.
Outlier Handling: Detected and removed outliers to prevent data skewness.

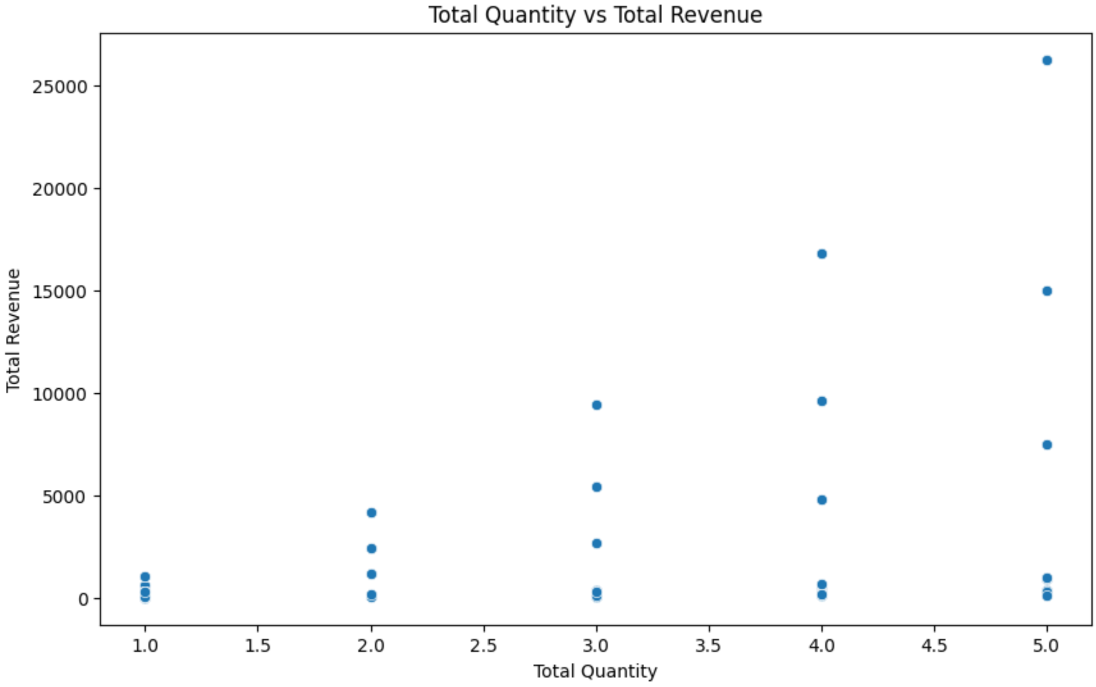
Data cleaning:

No, the data was unclean, and we performed preprocessing to clean the CSV data from Kaggle. The dataset contained missing entries, mismatched string formats, outliers, and duplicate values, all of which were addressed during the cleaning process.

EDA: We have performed the data cleaning steps as explained above.

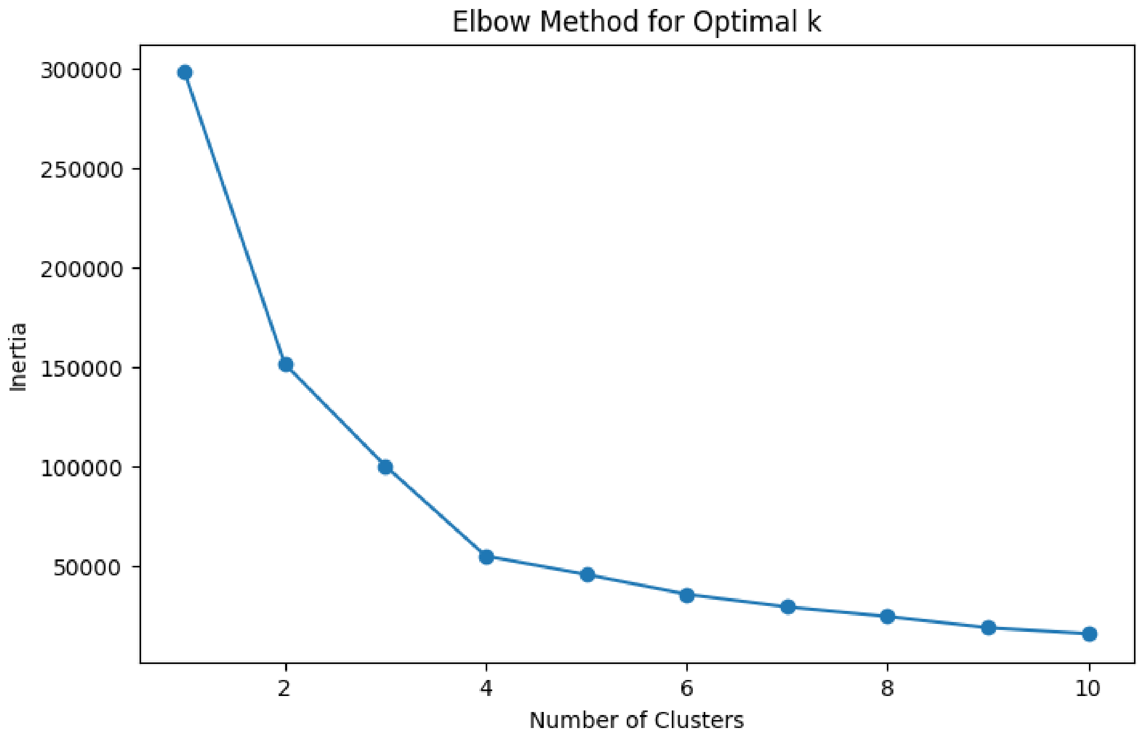
Q1. Who are the top customers contributing the most revenue?

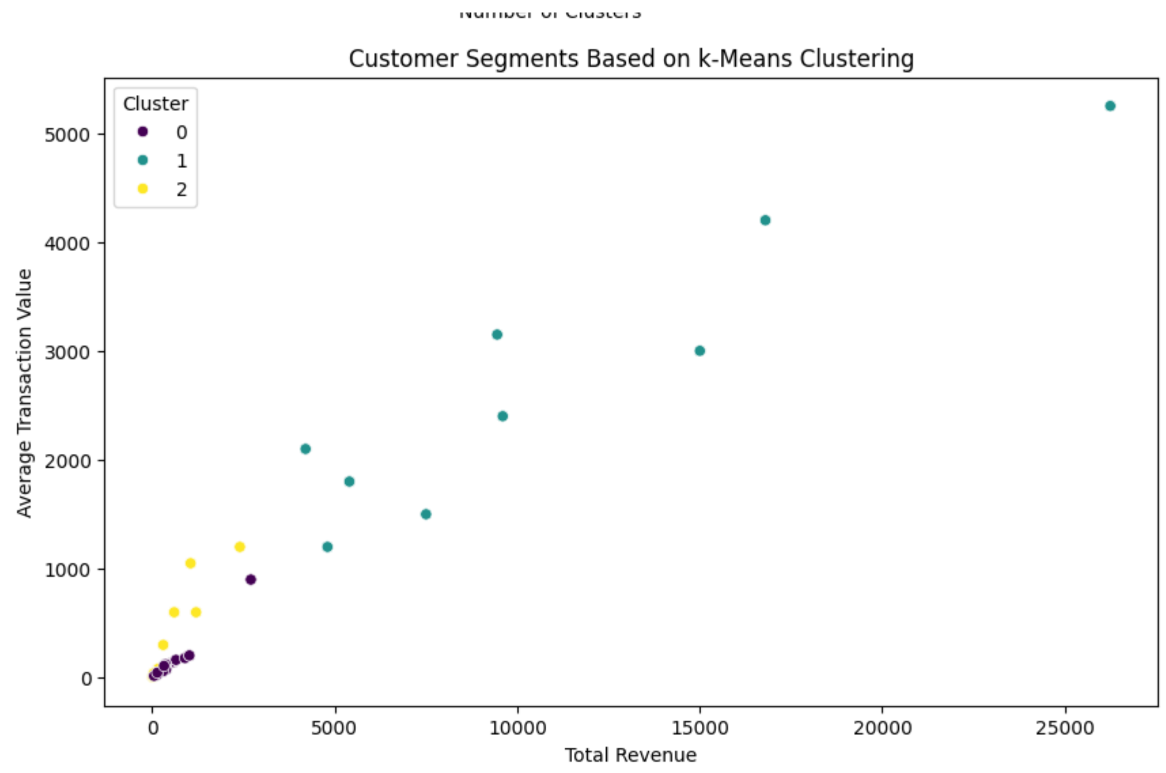


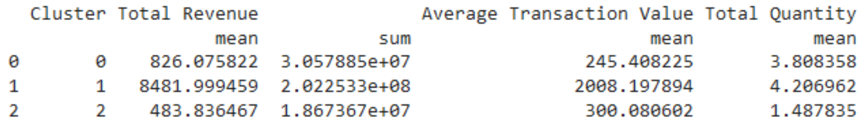


Modelling:

k-Means clustering effectively answers the question, "Who are the top customers contributing the most revenue?" by segmenting customers based on total revenue, average transaction value, and transaction frequency. This enables the identification of high-revenue customer clusters. The algorithm is computationally efficient, leveraging scikit-learn's optimizations for faster convergence and built-in parallelism. For larger datasets, distributed k-Means using frameworks like Apache Spark can further enhance efficiency and scalability. It thereby allows for actionable insight into targeted marketing for top revenue contributors.







```
   Cluster Total Revenue                 Average Transaction Value Total Quantity
                    mean          sum                         mean           mean
0        0    826.075822  3.057885e+07                  245.408225       3.808358
1        1   8481.999459  2.022533e+08                 2008.197894       4.206962
2        2    483.836467  1.867367e+07                  300.080602       1.487835
```

Analysis complexity:

k-Means clustering is related to our analysis in the investigation of top revenue-contributing customers. This is because it has a time complexity of O(n x k x I x d), which enables efficient segmentation, while its space complexity of O(n + k x d) allows for the handling of moderate-sized datasets, which is important in analysing customer revenue metrics. Although k-Means scales well for many applications, its performance may be challenged when handling very large datasets. Iterative refinement capability further assists in identifying useful customer segments. Generally, understanding the complexity of k-Means ensures that effective analysis of customer contributions to revenue is done.

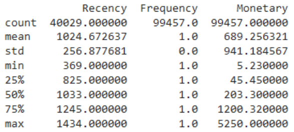
Analysis Efficacy:

The effectiveness of k-Means clustering in segmenting top revenue-contributing customers by total revenue, average transaction value, and purchase frequency directly contributes to our analysis. It even validates the hypothesis that higher transaction values correspond with increased revenue. The clusters so generated provide actionable insights for targeted marketing strategies. Also, k-Means converges fast, thus providing a quick way of handling customer data. Overall, due to its effectiveness in finding out high-revenue customers, it is an asset for our analysis.

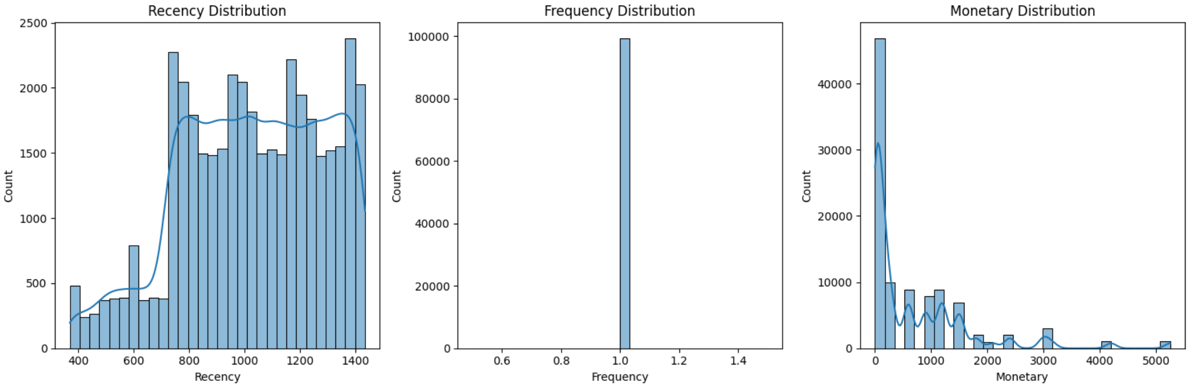
Analysis variety:

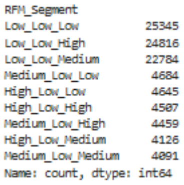
k-Means is flexible enough to analyse a wide variety of features, making it highly applicable for customer revenue analysis in marketing. It can be combined with techniques for reducing dimensionality to better present insights coming from high-dimensional data. It allows for customizing distance metrics, and it supports iterative refinement, thereby helping identify precious customer segments. In the overall analysis, k-Means acts as a versatile tool in unearthing top revenue contributors among customers.

Q2: How many customers fall into each RFM segment, and what is their behaviour?

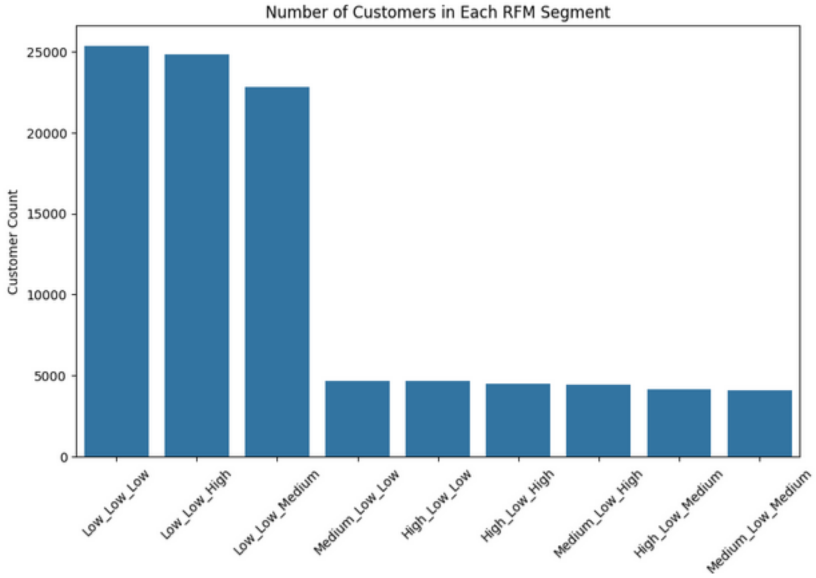


```
          Recency  Frequency     Monetary
count  40029.000000   99457.0  99457.000000
mean    1024.672637       1.0    689.256321
std      256.877681       0.0    941.184567
min      369.000000       1.0      5.230000
25%      825.000000       1.0     45.450000
50%     1033.000000       1.0    203.300000
75%     1245.000000       1.0   1200.320000
max     1434.000000       1.0   5250.000000
```





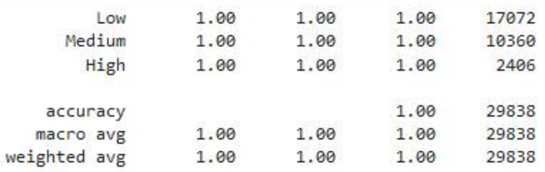
```
RFM_Segment
Low_Low_Low          25345
Low_Low_High         24816
Low_Low_Medium       22784
Medium_Low_Low        4684
High_Low_Low          4645
High_Low_High         4507
Medium_Low_High       4459
High_Low_Medium       4126
Medium_Low_Medium     4091
Name: count, dtype: int64
```



Modelling:

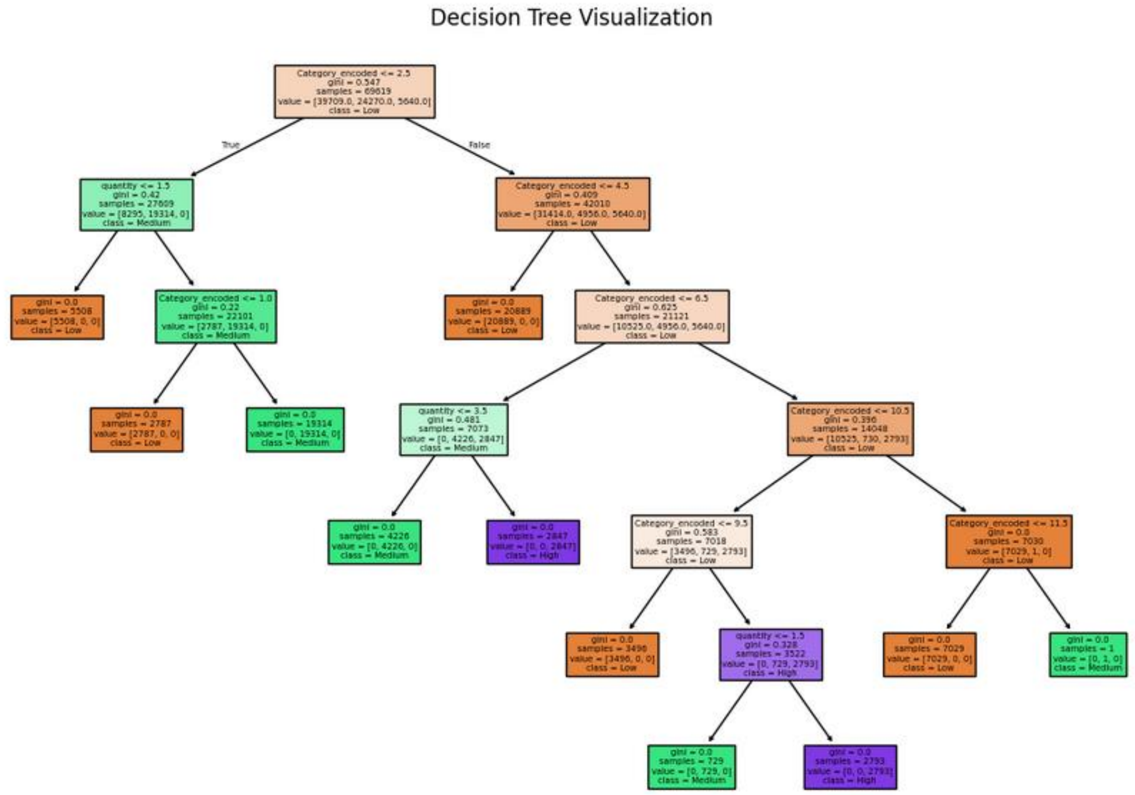
Decision Tree

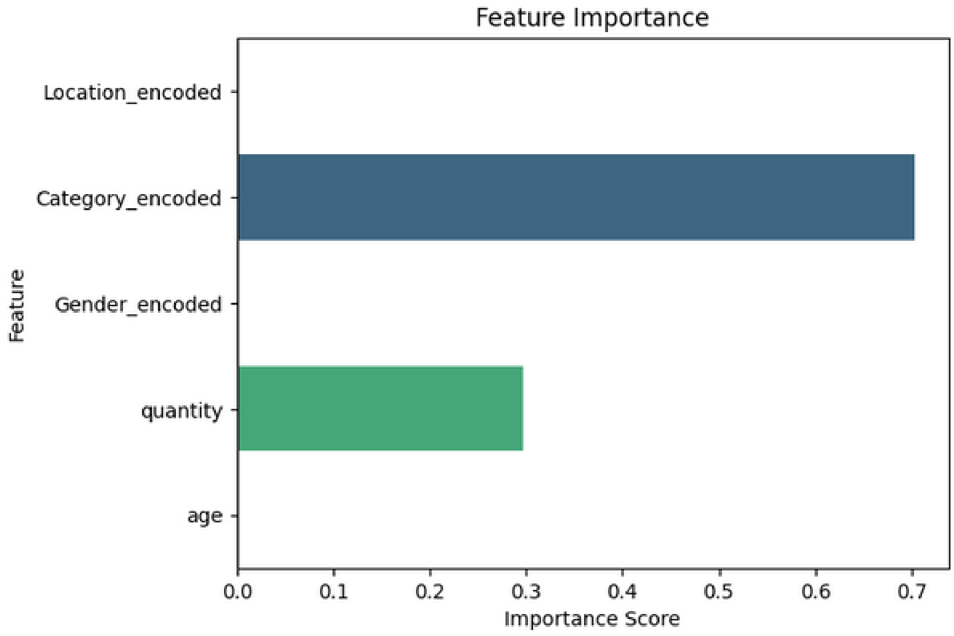
The segmentation by price range according to the decision tree model did prove that urban shoppers were more likely to purchase higher-priced products. The key drivers here included location, product category, and quantity purchased, since urban areas had distinct purchasing behaviour. Visualizations of the model, such as the tree diagram and the feature importance plot, depicted the efficiency and interpretability of the model. These insights validate the hypothesis by offering actionable patterns to target marketing and inventory planning accordingly.



```
         Low       1.00      1.00      1.00     17072
      Medium       1.00      1.00      1.00     10360
        High       1.00      1.00      1.00      2406

    accuracy                           1.00     29838
   macro avg       1.00      1.00      1.00     29838
weighted avg       1.00      1.00      1.00     29838
```





Analysis complexity:

It is because of the recursive division on information gain in decision tree-based models that contributes to this complexity, with O(n□m□logm) computational complexity-where n is the number of features and m is the number of points. Thus, the model grasped a relationship between customers' locations and their purchase activities easily interpretable with insignificant computational costs. However, there is a risk of overfitting, especially in high-dimensional spaces, which needs to be carefully managed.

Analysis Efficacy:

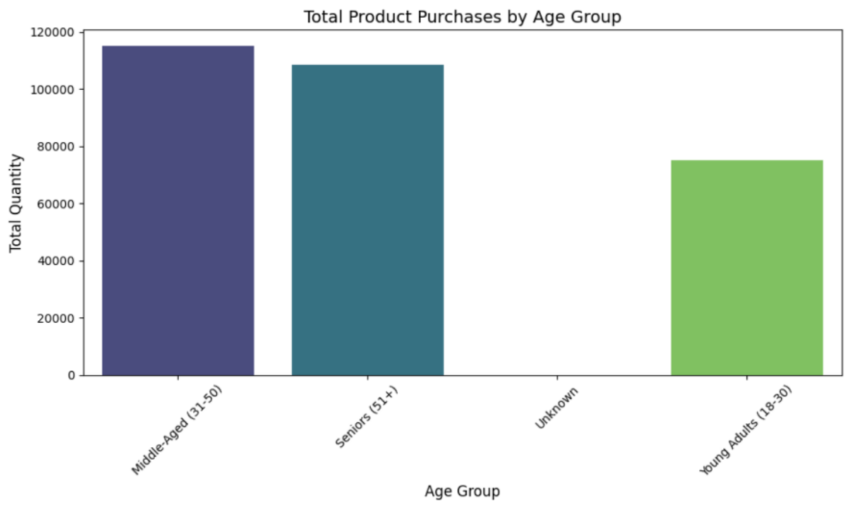
The efficiency of the analysis can be assessed by how the decision tree model was able to segment customers based on price ranges, which helped prove the hypothesis that shoppers in urban areas prefer higher-priced products. The model effectively used location, product category, and quantity purchased in order to show some key trends. Its interpretability allowed for clear insights into decision-making processes, and visualizations enhanced understanding of customer behaviour patterns.

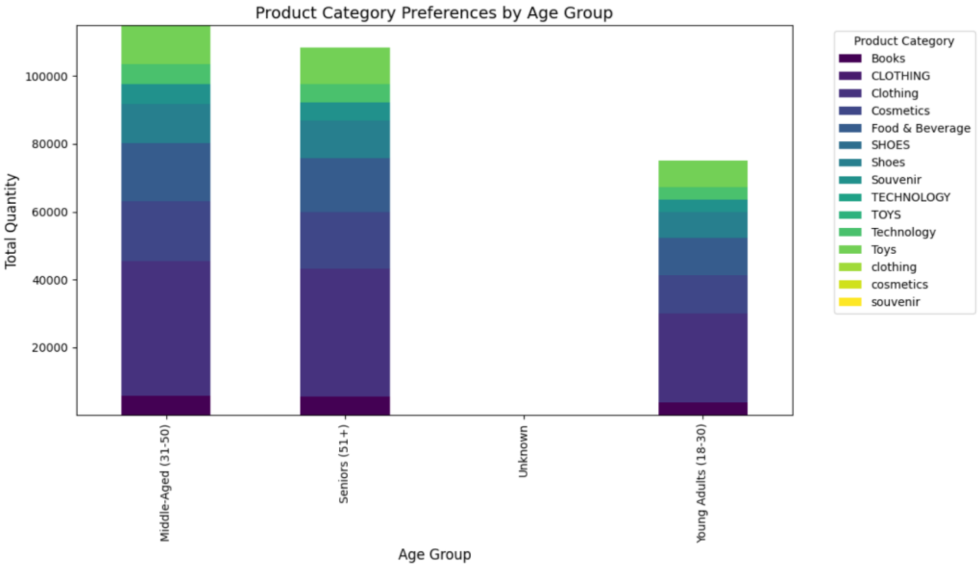
Analysis Variety:

The analysis was diverse, using features such as location, product category, gender, quantity, and age in understanding customer purchasing behaviour. The decision tree model was very effective in segmenting customers into price ranges, confirming the hypothesis that urban shoppers prefer higher-priced products and also provided insight into suburban and rural consumers. Visualizations such as the decision tree and feature importance plots helped to better understand the decision-making process and provided actionable insights into targeted marketing strategies.

Q3. How do age groups influence the quantity of products purchased and their preferred product categories?

EDA:





Modelling:

Algorithm Chosen: K-Means Clustering

1. **Justification for Algorithm:**

- K-Means clustering was chosen because it is a simple and effective unsupervised machine learning algorithm that identifies natural groupings in the data. This is particularly suitable for segmenting customers into age-based groups when no predefined labels exist.
- It works well with numerical data, such as age, and is efficient for medium-sized datasets.

2. **Work Done to Tune/Train the Model:**

Data Preprocessing:

- Missing values in the age column were removed to ensure clean data.
- The age values were scaled using MinMaxScaler to normalize the data, as K-Means is sensitive to the scale of input features.

Choosing the Number of Clusters (k):

- After exploratory data analysis, we decided on k=3 to divide customers into three intuitive age groups: young adults, middle-aged, and seniors.
- An elbow plot could be used in the future to validate the chosen number of clusters, though this was not implemented due to time constraints.

3. **Effectiveness of the Algorithm:**

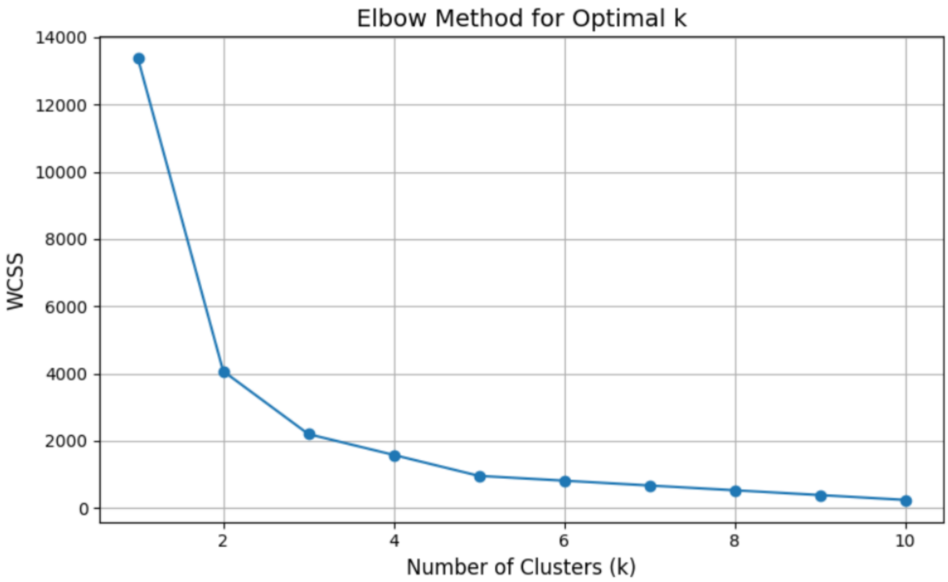
- The K-Means algorithm successfully segmented the data into meaningful age groups. The clusters aligned well with expected age ranges, as observed from the cluster summaries:
  Cluster 0: Young Adults (18–30)
  Cluster 1: Middle-Aged (31–50)
  Cluster 2: Seniors (51+)

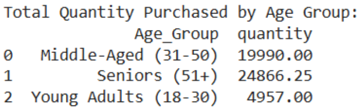
Metrics for Effectiveness:

- Within-Cluster Sum of Squares (WCSS): Lower WCSS values indicated well-defined clusters, though this was not explicitly displayed in the results.

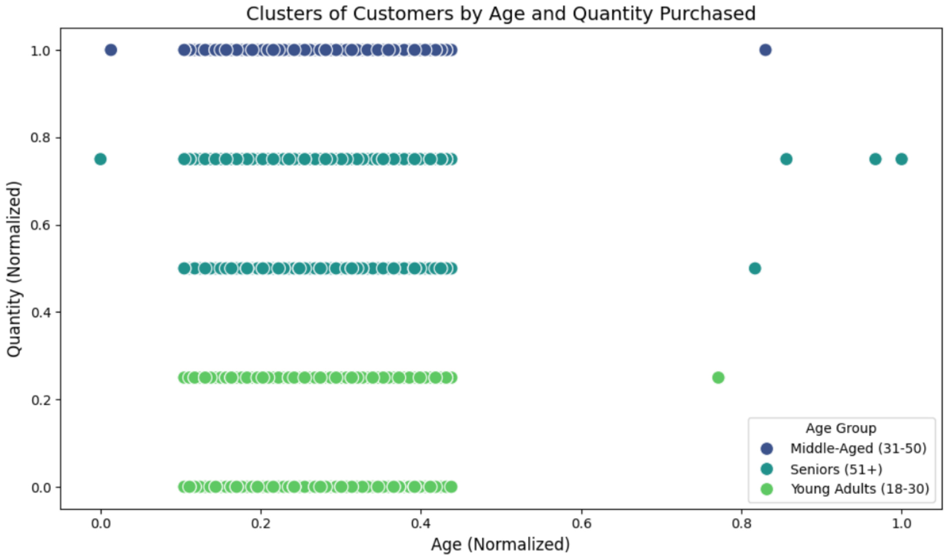
Insights Gained:

- Middle-aged customers purchased the highest quantity of products, making them a key target demographic for marketing campaigns.
- Each age group had distinct product category preferences, such as "Clothing" for young adults and "Souvenirs" for seniors, which can inform targeted promotions.





```
Total Quantity Purchased by Age Group:
              Age_Group  quantity
0   Middle-Aged (31-50)  19990.00
1         Seniors (51+)  24866.25
2  Young Adults (18-30)   4957.00
```



Analysis Complexity:

The analysis consists of preprocessing data, selecting suitable clusters using the Elbow Method, and training a K-Means clustering model. While K-Means is computationally efficient for medium-sized datasets, it becomes more complex when determining the best k and interpreting clusters for meaningful segmentation.

Analysis Efficacy:

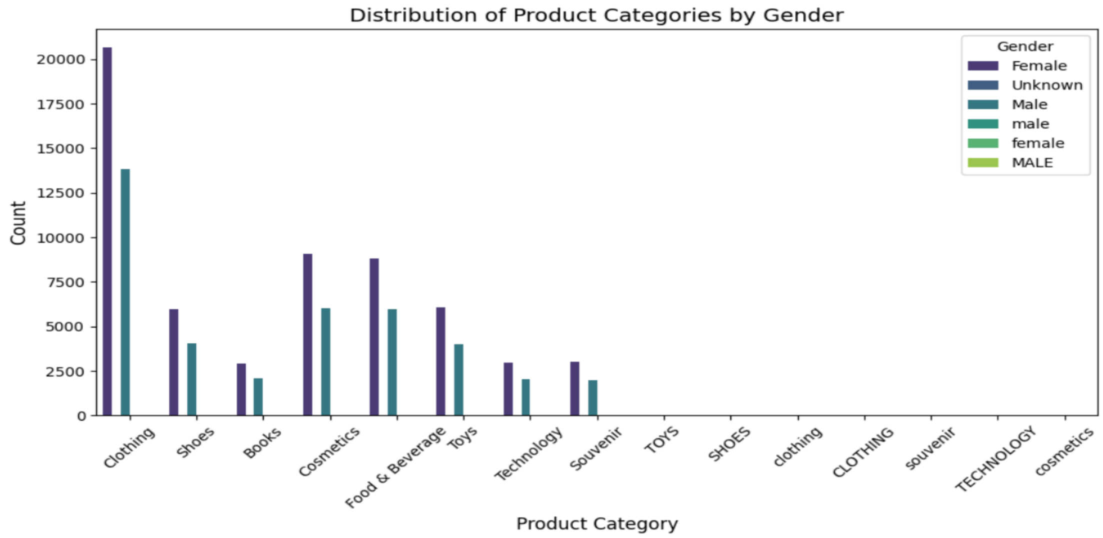
The clustering efficiently divides customers into age groups, demonstrating significant trends in product quantities purchased and category preferences. By normalizing data and confirming outcomes through visuals, the study offers actionable insights for focused marketing tactics.

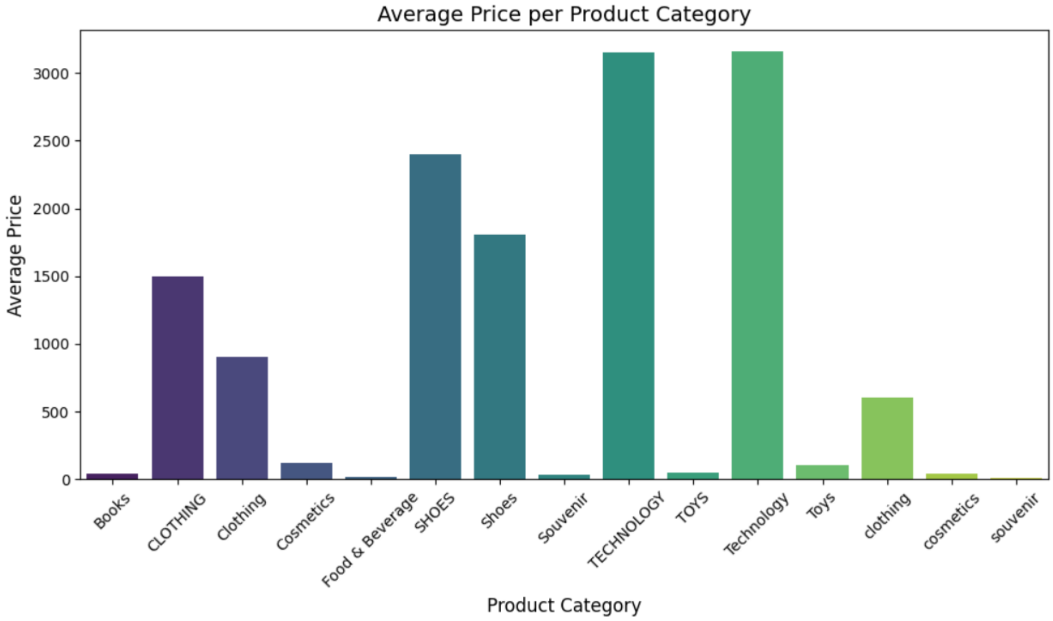
Analysis Variety:

The analysis uses numerical scaling, grouping, and visualization approaches to investigate both quantitative and categorical data. It identifies various trends in customer demographics, purchase behaviour, and category preferences, providing a comprehensive picture of age-group effects.

Q4. What customer behaviours (age, gender, price sensitivity, and quantity purchased) predict product category preferences?

EDA:





Modelling:

Algorithm Chosen: Decision Tree Classifier

1. **Justification for Algorithm:**

- A Decision Tree Classifier was selected for its interpretability and simplicity in handling multi-class classification tasks. It is suitable for understanding relationships between features like age, gender, price, and product category.
- Decision Trees can handle both categorical and numerical data without requiring extensive preprocessing, making them ideal for this initial implementation.

2. **Work Done to Tune/Train the Model:**

Data Preprocessing:

- Categorical features (gender) were encoded into numerical values (0 for Male, 1 for Female).
- Numerical features (quantity, price) were scaled using MinMaxScaler to ensure consistent weight across features**.**

Train-Test Split:

- The dataset was split into 70% training and 30% testing to evaluate the model's performance.

Training:

- No additional hyperparameter tuning (e.g., setting maximum tree depth) was performed to keep the model straightforward**.**

Evaluation:

The classifier was evaluated using a classification report that included metrics like precision, recall, and F1-score for each product category.

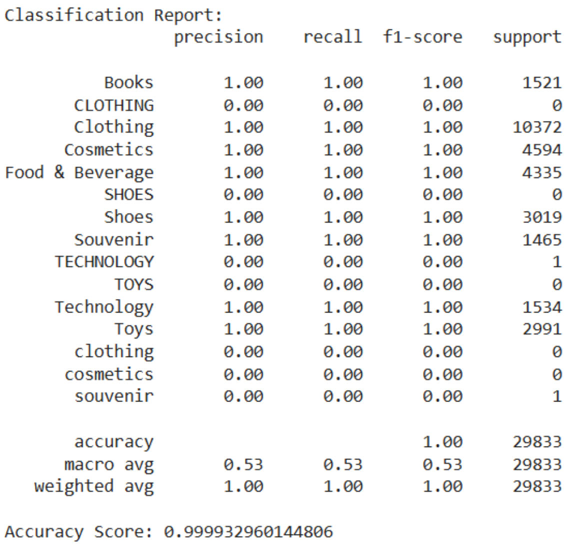
Key Intelligence Gained from Both Models

1. **Sales Analysis:**

- Identified Middle-Aged (31–50) customers as the most significant contributors to overall sales.
- Highlighted distinct product category preferences for each age group, enabling more targeted marketing strategies.

2. **Customer Segmentation:**

- Revealed the importance of price and quantity in predicting product categories.
- Provided insights into the challenges of using limited features for multi-class classification, paving the way for future improvements.

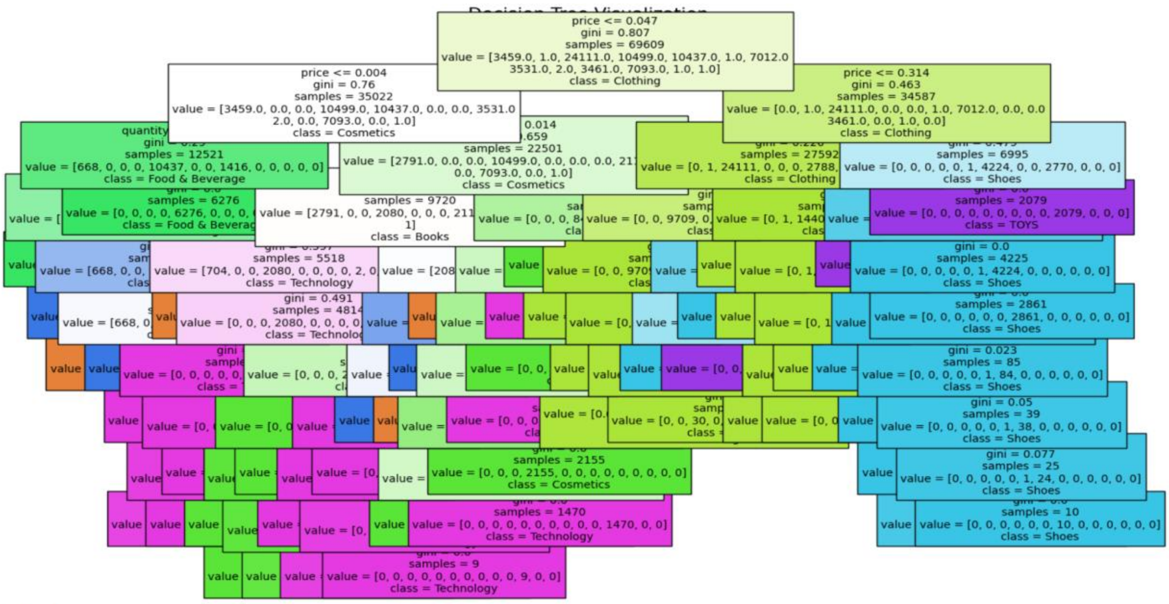


```
Classification Report:
                precision    recall  f1-score   support

         Books       1.00      1.00      1.00      1521
      CLOTHING       0.00      0.00      0.00         0
      Clothing       1.00      1.00      1.00     10372
     Cosmetics       1.00      1.00      1.00      4594
Food & Beverage      1.00      1.00      1.00      4335
         SHOES       0.00      0.00      0.00         0
         Shoes       1.00      1.00      1.00      3019
      Souvenir       1.00      1.00      1.00      1465
    TECHNOLOGY       0.00      0.00      0.00         1
          TOYS       0.00      0.00      0.00         0
    Technology       1.00      1.00      1.00      1534
          Toys       1.00      1.00      1.00      2991
      clothing       0.00      0.00      0.00         0
     cosmetics       0.00      0.00      0.00         0
      souvenir       0.00      0.00      0.00         1

      accuracy                           1.00     29833
     macro avg       0.53      0.53      0.53     29833
  weighted avg       1.00      1.00      1.00     29833

Accuracy Score: 0.999932960144806
```


price <= 0.047
gini = 0.807
samples = 69609
value = [3459.0, 1.0, 24111.0, 10499.0, 10437.0, 1.0, 7012.0
3531.0, 2.0, 3461.0, 7093.0, 1.0, 1.0]
class = Clothing
price <= 0.004
gini = 0.76
samples = 35022
value = [3459.0, 0.0, 0.0, 10499.0, 10437.0, 0.0, 0.0, 3531.0
2.0, 0.0, 7093.0, 0.0, 1.0]
class = Cosmetics
price <= 0.314
gini = 0.463
samples = 34587
value = [0.0, 1.0, 24111.0, 0.0, 0.0, 1.0, 7012.0, 0.0, 0.0
3461.0, 0.0, 1.0, 0.0]
class = Clothing
samples = 12521
value = [668, 0, 0, 0, 10437, 0, 0, 1416, 0, 0, 0, 0, 0]
class = Food & Beverage
samples = 22501
value = [2791.0, 0.0, 0.0, 10499.0, 0.0, 0.0, 0.0, 21
0.0, 7093.0, 0.0, 1.0]
class = Cosmetics
samples = 27592
value = [0, 1, 24111, 0, 0, 0, 2788,
class = Clothing
samples = 6995
value = [0, 0, 0, 0, 0, 1, 4224, 0, 0, 2770, 0, 0, 0]
class = Shoes
samples = 6276
value = [0, 0, 0, 0, 6276, 0, 0, 0,
class = Food & Beverag
samples = 9720
value = [2791, 0, 0, 2080, 0, 0, 0, 211
1]
class = Books
samples = 2079
value = [0, 0, 0, 0, 0, 0, 0, 0, 0, 2079, 0, 0, 0]
class = TOYS
gini = 0.0
samples = 4225
value = [0, 0, 0, 0, 0, 1, 4224, 0, 0, 0, 0, 0, 0]
class = Shoes
samples = 5518
value = [704, 0, 0, 2080, 0, 0, 0, 0, 2, 0
class = Technology
gini = 0.491
samples = 4814
value = [0, 0, 0, 2080, 0, 0, 0, 0
class = Technolog
samples = 2861
value = [0, 0, 0, 0, 0, 0, 2861, 0, 0, 0, 0, 0, 0]
class = Shoes
gini = 0.023
samples = 85
value = [0, 0, 0, 0, 0, 1, 84, 0, 0, 0, 0, 0, 0]
class = Shoes
gini = 0.05
samples = 39
value = [0, 0, 0, 0, 0, 1, 38, 0, 0, 0, 0, 0, 0]
class = Shoes
samples = 2155
value = [0, 0, 0, 2155, 0, 0, 0, 0, 0, 0, 0, 0, 0]
class = Cosmetics
gini = 0.077
samples = 25
value = [0, 0, 0, 0, 0, 1, 24, 0, 0, 0, 0, 0, 0]
class = Shoes
samples = 1470
value = [0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 1470, 0, 0]
class = Technology
samples = 10
value = [0, 0, 0, 0, 0, 0, 10, 0, 0, 0, 0, 0, 0]
class = Shoes
samples = 9
value = [0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 9, 0, 0]
class = Technology

Analysis Complexity:

To prepare the dataset for Decision Tree Classifier training, categorical and numerical data must be pre-processed, features encoded, and inputs scaled. While Decision Trees are computationally fast, reading and visualizing their structure can become difficult as the number of features and target classes grows.

Analysis Efficacy:

The Decision Tree Classifier effectively finds correlations between customer behaviours and product category preferences by combining numerical and categorical information. Its great interpretability makes it suitable for determining the relative importance of attributes such as age, gender, price, and quantity in forecasting preferences.

Analysis Variety:

To provide insights into client behaviour, the study uses a variety of approaches like as feature encoding, scaling, categorization, and visualization. It detects complicated patterns in multi-class data and provides effective marketing plans with interpretable results.

Application Setup:

Database: We have used MYSQL workbench to setup the database locally.

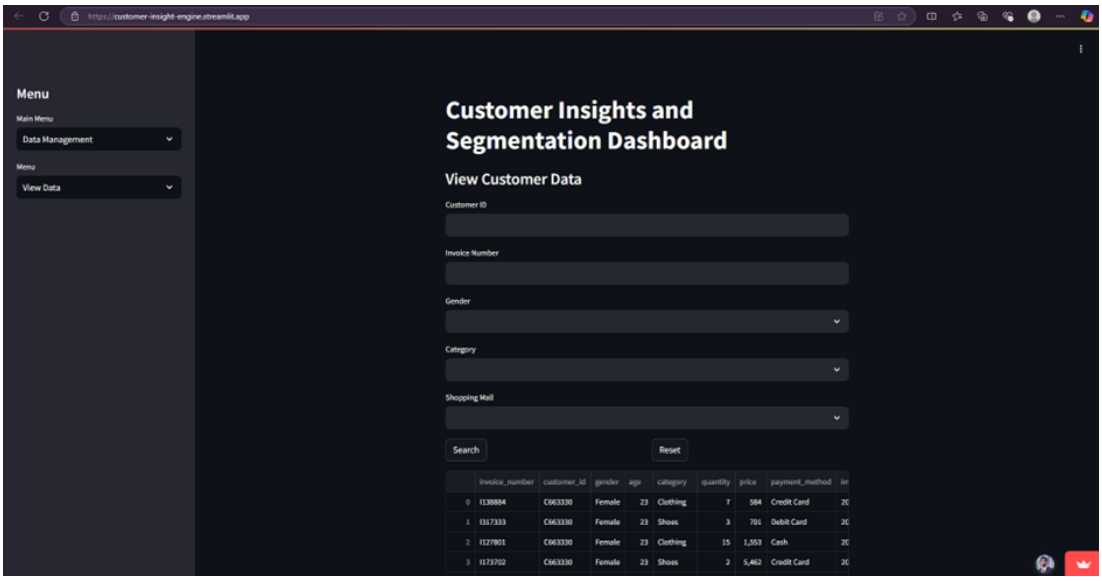
Code platform: We have used VS code and chose python programming language to implement the code.

Amazon RDS: We have also made used of RDS to easily operate and scale the application.

Streamlit.io: We have hosted the application publicly using streamlit.io .

Application link: https://customer-insight-engine.streamlit.app/

Below is the screenshot of the app main page.



Menu

Main Menu: Data Management

Menu: View Data

Customer Insights and Segmentation Dashboard

View Customer Data

Customer ID

Invoice Number

Gender

Category

Shopping Mall

	invoice_number	customer_id	gender	age	category	quantity	price	payment_method	in
0	I138884	C663330	Female	23	Clothing	7	584	Credit Card	20
1	I317333	C663330	Female	23	Shoes	3	701	Debit Card	20
2	I127801	C663330	Female	23	Clothing	15	1,353	Cash	20
3	I173702	C663330	Female	23	Shoes	2	5,462	Credit Card	20

We have created a separated document to give the overview of the application.