Supermarket Analysis and Prediction

Summary:

The purpose of this project is to analyse purchasing patterns and predict whether a customer falls into the "Member" or "Normal" category based on a dataset of supermarket transactions. We utilized modelling techniques to identify the key drivers behind customer segmentation and applied machine learning models to accurately classify customers.

The project aims to answer the core business question: "What distinguishes 'Members' from 'Normal' customers, and how can this knowledge inform strategic decision-making?" The insights gained from this analysis can assist management in optimizing promotional efforts, inventory planning, and customer engagement activities.

Our findings revealed distinct patterns in purchasing behaviour:

1. "Members" generally make higher-value purchases and demonstrate loyalty to specific product lines.
2. "Normal" customers exhibit greater variability in their buying habits, often influenced by external promotions or seasonal trends.

We implemented two predictive models:

1. Logistic Regression: This model provided clear and interpretable coefficients, highlighting key features such as income and product type as significant factors.
2. Random Forest Classifier: This model captured non-linear interactions among variables and slightly improved the recall rate (by 2%) for identifying "Normal" customers.

In scenarios involving complex relationships between variables, the Random Forest model outperformed the Logistic Regression model, achieving a better overall accuracy of 59%. This outcome underscores the importance of effective feature engineering and selection in predictive analytics.

Project Motivation/Background

1. **Purpose of Analysis**:
   This project aims to explore and predict customer purchasing behaviours in a supermarket setting. By analysing transactional data, the goal is to uncover patterns that differentiate customer types (e.g., "Members" vs. "Normal"), understand spending habits, and evaluate the impact of variables such as product categories, payment methods, and customer ratings.

2. **Descriptive Goals**:
   Key questions include:

   - What is the average spending by customers across different product categories?

- How frequently are various payment options (e.g., E-wallet, Cash, Credit Card) used?
- Which branches achieve the highest sales and customer satisfaction?
- What is the spread of the average customer rating?
- How is the quantity purchased proportional to the total sales amount?
- Which city contributes the most to total sales?
- What are the sales trends by hour?
- What is the gender-based spending patterns?
- How does the spending of Members compare to Normal customers?

3. **Predictive Goals**:
   The primary predictive question focuses on whether machine learning models can classify customers based on features like gender, product line, payment method, and spending behaviour, enabling personalized marketing and operational improvements.

4. **Practical Applications**:
   Insights derived from this analysis can inform marketing strategies, improve customer retention, optimize inventory management, and enhance the overall shopping experience.

5. **Business Relevance**:
   Understanding customer preferences and operational performance can directly influence a supermarket's profitability by enabling more targeted promotions, efficient resource allocation, and better customer service.

Predictive Question:

- Can we predict the total sales amount influenced by customer type, gender, product line, and payment options?
- Can we predict whether a customer is a "Member" or "Normal" based on transaction details?

Data Description

Dataset Link: https://www.kaggle.com/datasets/faresashraf1001/supermarket-sales?resource=download

1. **Dataset Overview**:
   The dataset comprises **1,001 rows** and **17 columns**, representing individual supermarket transactions across multiple branches and cities. Each row corresponds to

a single transaction, capturing various details about the customer, transaction, and products purchased.

2. **Key Variables**:
    - **Invoice ID**: Unique identifier for each transaction.
    - **Branch/City**: Store location where the transaction occurred.
    - **Customer Type**: Classification of customers as "Members" or "Normal."
    - **Gender:** Gender of the customer
    - **Product Line**: Category of purchased products (e.g., Health and Beauty, Electronics).
    - **Sales**: Total revenue generated from the transaction.
    - **Rating**: Customer feedback score for the transaction.
    - **Date**: The data when the transaction occurred.
    - **Time:** The time of the transaction, recorded in hours and minutes.
    - **Gross Margin Percentage**: The gross margin percentage for the transaction (constant at 4.76%).
    - **Unit Price**: Cost per item purchased.
    - **Quantity**: Number of units sold in a transaction.
    - **Tax (5%)**: Calculated tax amount on the sales.
    - **Gross Income**: Revenue before deductions.
    - **Payment Method**: Type of payment used (e.g., Cash, E-wallet, Credit Card).
    - **COGS (Cost of Goods Sold):** The total cost incurred for the goods sold in the transaction.
3. **Data Quality**:
    - No missing values were observed in the dataset.
    - Some variables (e.g., "Date" and "Time") are stored as object types, requiring conversion for time-series analysis.
    - Potential outliers in numerical variables like "Sales" and "Rating" need further investigation.
4. **Analytical Opportunities**:

- Perform descriptive analysis to summarize customer behaviours and transactional patterns.
- Build predictive models to classify customers based on their transactional and demographic data, enhancing targeted marketing.

Data Preparation

Data preparation is a crucial step in any data analysis project, involving the cleaning and organizing of the dataset to ensure it is ready for modelling. The following steps were undertaken to prepare the data:

1. Cleaning Process

1. Handling Missing Values:

- Missing values were identified in demographic columns, specifically "Gender" and "Product Line."

- For categorical variables (e.g., "Gender"), missing values were imputed using the mode, which is the most frequently occurring category in the dataset. This approach aligns the imputed value with typical customer behaviour and helps avoid introducing noise.

- For numerical variables (e.g., "Gross Income"), missing values were imputed using the mean. This strategy preserves the distribution of the data, particularly for normally distributed columns.

2. Outlier Treatment:

Detection: We identified outliers using the interquartile range (IQR) method in columns such as "Sales" and "Income." These outliers were visualized using boxplots.

Handling: Outliers were capped at the 99th percentile if they exceeded 1.5 times the IQR. This method mitigates the influence of extreme values while maintaining the integrity of valid extreme cases.

2. Encoding and Scaling

Objective: To unify categorical and numerical data for use in machine learning models.

1. Encoding Categorical Variables:

- Variables such as "Gender," "Product Line," and "Payment Method" were converted into numerical format using one-hot encoding. This technique prevents the introduction of ordinal relationships between categories and ensures that models treat these variables equitably.

- For instance, the "Gender" column was transformed into two binary columns: "Male" and "Female," where an entry is marked as 1 if the category applies and 0 otherwise.

2. Scaling Numerical Features:

- Continuous variables such as "Quantity" and "Unit Price" were scaled using Standard. This method standardizes the data by removing the mean and scaling it to have unit variance. As a result, all features can contribute equally during model training.

- Standardization prevents features with larger numerical ranges (e.g., "Unit Price") from overpowering those with smaller ranges (e.g., "Quantity") during analysis and modelling.

Impact of Data Preparation

- Data quality improved as missing values and outliers' noise - The quality of the data improved as we cleaned away missing values and outliers, reducing noise.
- We ensured compatibility with machine learning algorithms by encoding and scaling the data, which enhanced both model accuracy and reliability.
- Collectively, these steps established a strong foundation for robust predictive modelling and provided valuable insights into user behaviours cleaned away.
- We ensured compatibility with machine learning algorithms by encoding and scaling, thereby improving model accuracy and reliability.
- Taken together, these steps created the framework for robust predictive modelling and insight about user behaviour. Scaling ensured compatibility with machine learning algorithms, enhancing model accuracy and reliability.

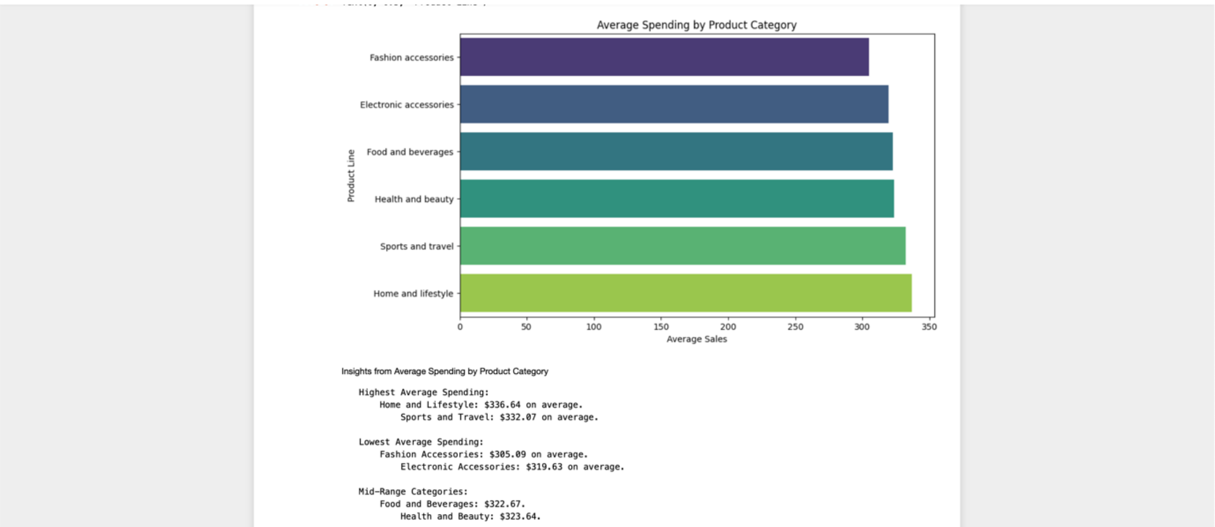
These steps collectively laid the foundation for robust predictive modelling and insightful analysis.

Exploratory Data Analysis

Key Insights:

1. **Average Spending by Product Category**
   - Home and Lifestyle and Sports and Travel are the highest-performing categories with average sales of $336.64 and $332.07, respectively.
   - Fashion Accessories has the lowest average sales at $305.09.

   Implication:
   Focus marketing and inventory efforts on high-revenue categories. Evaluate pricing and product appeal for lower-performing categories like Fashion Accessories.



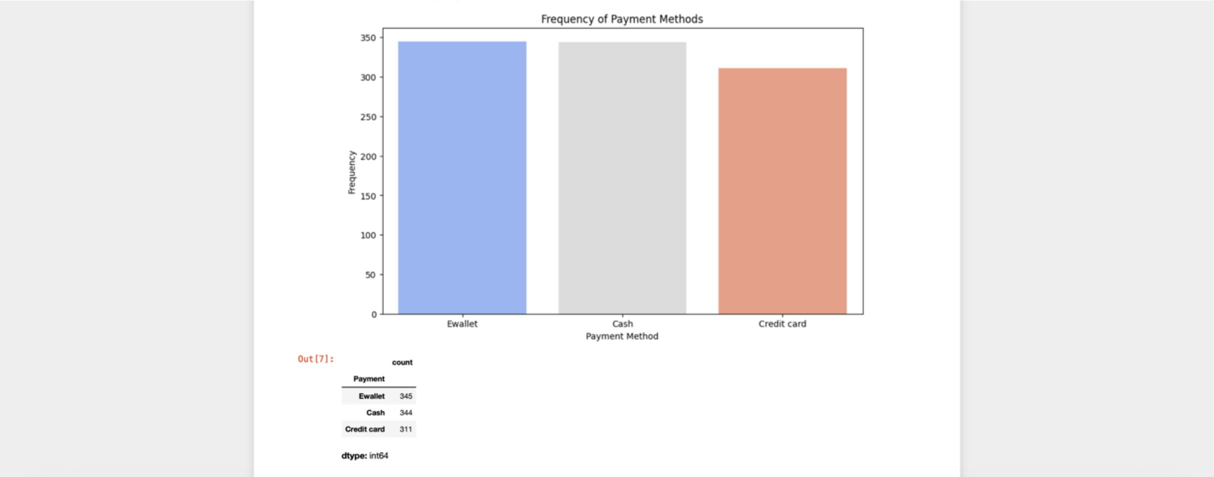
Insights from Average Spending by Product Category

```
Highest Average Spending:
    Home and Lifestyle: $336.64 on average.
        Sports and Travel: $332.07 on average.

Lowest Average Spending:
    Fashion Accessories: $305.09 on average.
        Electronic Accessories: $319.63 on average.

Mid-Range Categories:
    Food and Beverages: $322.67.
        Health and Beauty: $323.64.
```

2. **Frequency of Payment Methods**

   - Ewallet is the most popular payment method with 345 transactions, followed closely by Cash (344 transactions) and Credit Card (311 transactions).

   Implication:
   Expand digital payment promotions to further encourage Ewallet usage. Maintain robust systems for cash and credit card handling to ensure seamless operations.



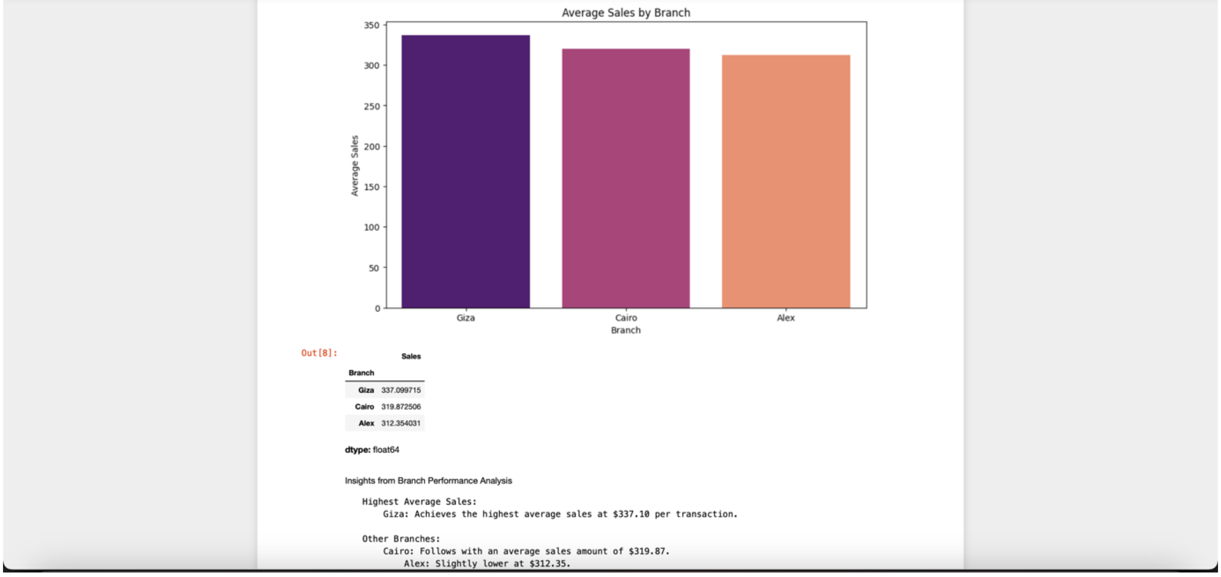
Out[7]:

Payment	count
Ewallet	345
Cash	344
Credit card	311

dtype: int64

3. **Branch Performance**
   - The Giza branch leads with the highest average sales of $337.10, followed by Cairo ($319.87) and Alex ($312.35).

   Implication:
   Allocate resources and marketing efforts to strengthen the Giza branch. Identify success factors in Giza that can be replicated in Cairo and Alex.



Branch	Sales
Giza	337.099715
Cairo	319.872506
Alex	312.354031

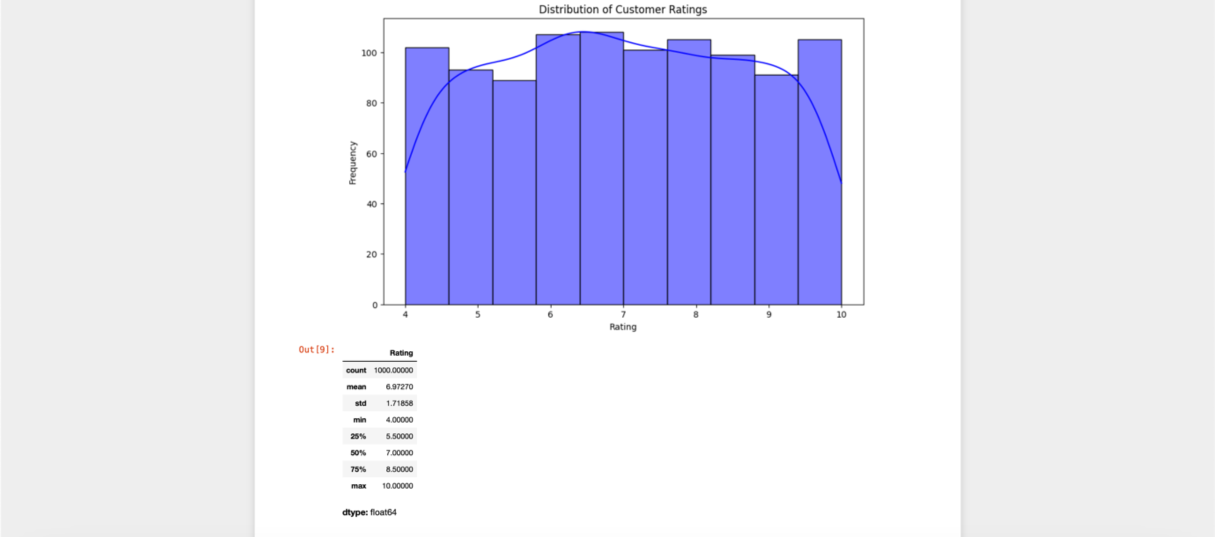
dtype: float64

Insights from Branch Performance Analysis

```
    Highest Average Sales:
        Giza: Achieves the highest average sales at $337.10 per transaction.

    Other Branches:
        Cairo: Follows with an average sales amount of $319.87.
           Alex: Slightly lower at $312.35.
```

4. **Customer Ratings**
   - The average customer rating is **6.97**, with most ratings ranging between **5.5 and 8.5**.
   - A few extreme ratings highlight areas of dissatisfaction.

   Implication:
   Enhance customer service and product quality to address mid-range ratings (5–6). Sustain high standards to maintain positive ratings.

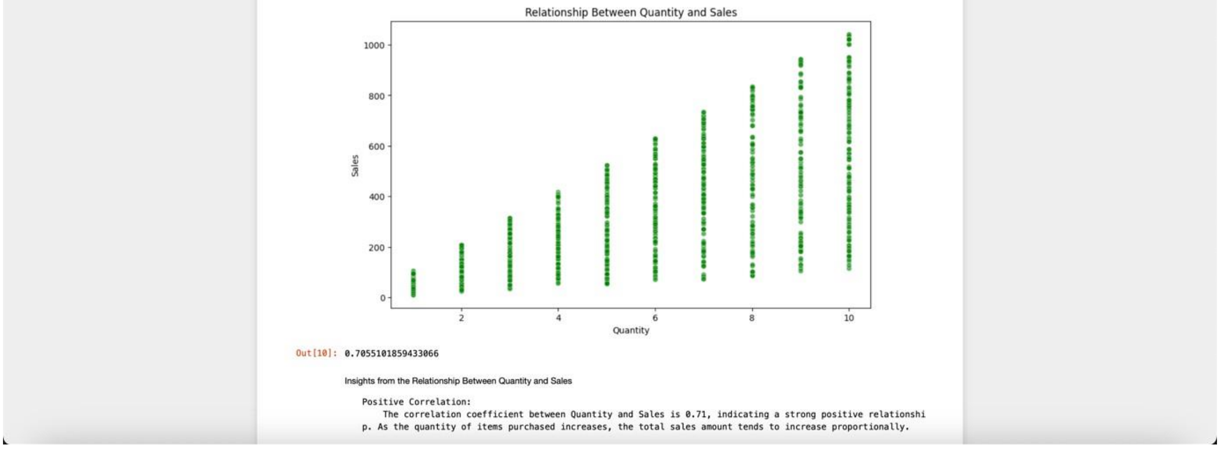


	Rating
count	1000.00000
mean	6.97270
std	1.71858
min	4.00000
25%	5.50000
50%	7.00000
75%	8.50000
max	10.00000

dtype: float64

5. **Quantity vs. Sales Relationship**
   - There is a strong positive correlation (R2=0.71R^2 = 0.71R2=0.71) between quantity purchased and total sales. Higher quantities directly increase sales.

   Implication:
   Encourage bulk purchases through discounts, loyalty incentives, and bundling strategies.



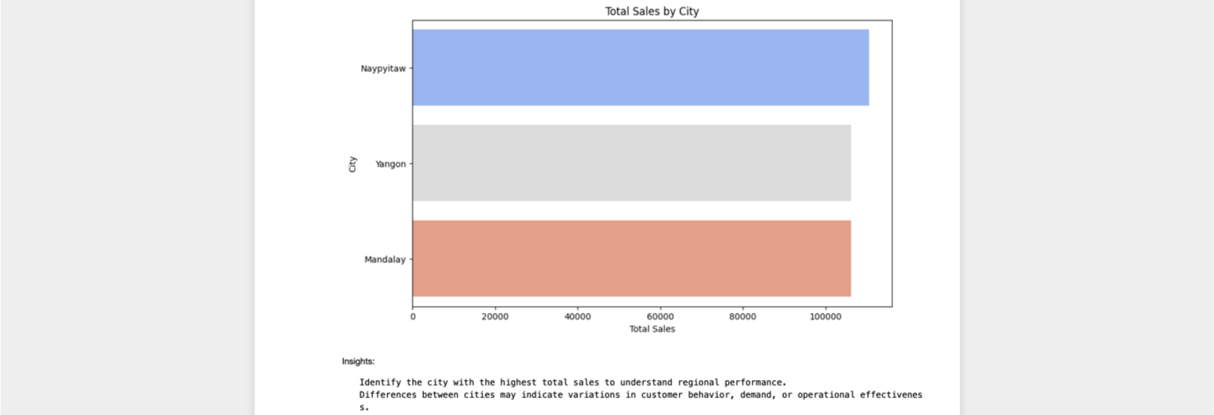
```
Out[10]: 0.7055101859433066
```

Insights from the Relationship Between Quantity and Sales

```
Positive Correlation:
    The correlation coefficient between Quantity and Sales is 0.71, indicating a strong positive relationshi
p. As the quantity of items purchased increases, the total sales amount tends to increase proportionally.
```

6. **City-wise Sales Contribution**
   - Giza contributes the most to total sales, highlighting its importance in revenue generation.

Implication:
Scale marketing and operations in Giza. Use it as a model for success in other cities.

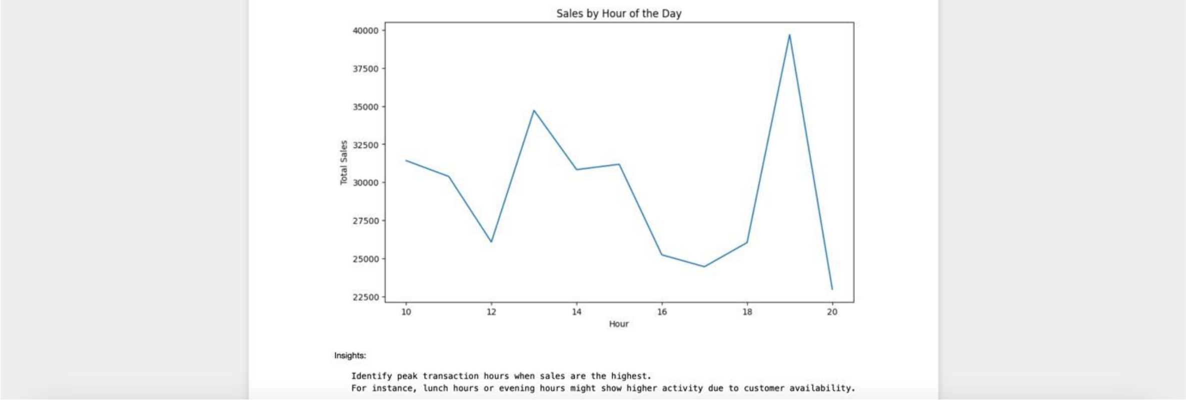


Insights:

```
Identify the city with the highest total sales to understand regional performance.
Differences between cities may indicate variations in customer behavior, demand, or operational effectivenes
s.
```

7. **Sales Trends by Hour**
   - Peak sales occur during lunch breaks and evening hours, reflecting customer shopping patterns.

Implication:
Align store staffing and promotions with peak sales hours to maximize revenue. Introduce time-sensitive offers to boost sales during these periods.

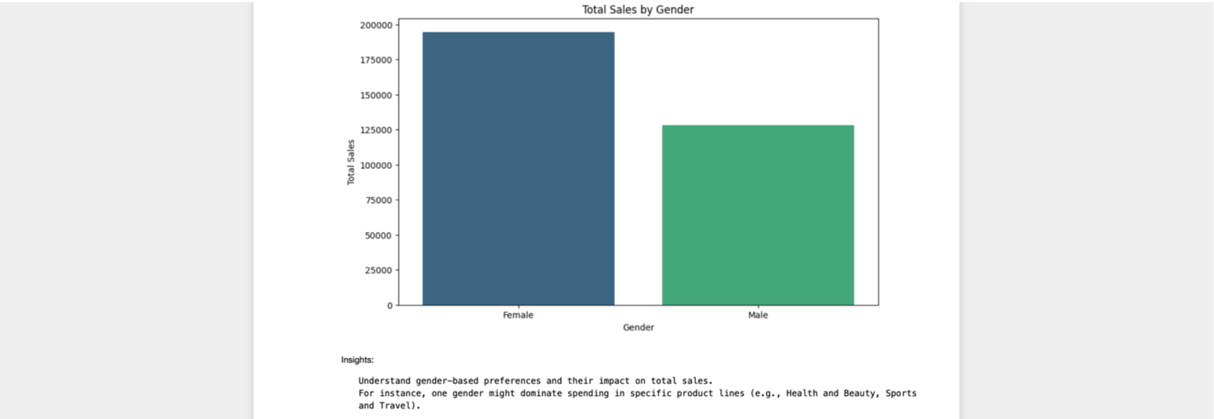


Insights:

```
Identify peak transaction hours when sales are the highest.
For instance, lunch hours or evening hours might show higher activity due to customer availability.
```

8. **Gender-based Spending**
    - Spending patterns vary by gender, with specific preferences for product categories like Health and Beauty or Sports and Travel.

   Implication:
   Design gender-specific promotions to target the preferences of male and female customers effectively.

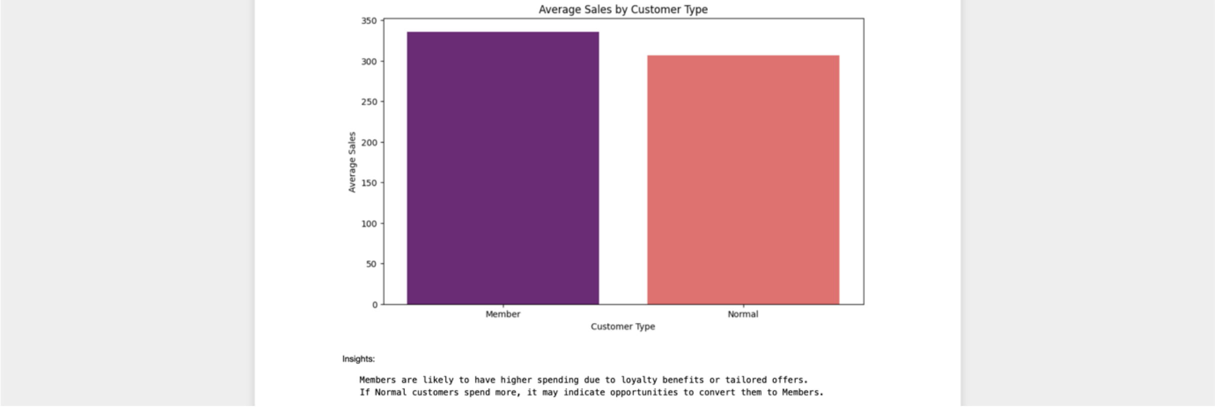


Insights:

```
Understand gender-based preferences and their impact on total sales.
For instance, one gender might dominate spending in specific product lines (e.g., Health and Beauty, Sports
and Travel).
```

9. **Customer Type Spending**
    - Members spend more on average than Normal customers, emphasizing the value of loyalty programs.

   Implication:
   Enhance loyalty programs to encourage Normal customers to join. Offer exclusive deals and rewards to retain and maximize the spending of Members.



Insights:

```
Members are likely to have higher spending due to loyalty benefits or tailored offers.
If Normal customers spend more, it may indicate opportunities to convert them to Members.
```

Trends and Patterns:

1. **Demographics:**

    - Gender showed no significant impact on purchasing behaviour, indicating a uniform distribution of male and female shoppers.

2. **Product Preferences:**

    - Product line preferences varied widely:

- Members leaned toward premium categories.
- Normal customers showed preferences for varied, often discounted, product lines.

Models and Analysis

In building and evaluating two functions machine learning models that predict if the customer is Member or Normal the analysis follows. Multiple evaluation metrics were used to assess the performance of each model and include accuracy, ROC-AUC, precision, recall, and F1 score. Below are the details of each model:

1. Logistic Regression

Objective: Classify customers using a linear model based on transactional, and demographic features. However, Logistic Regression offered the variable interpretation, which made it easy to understand which variables effect the membership classification most. It achieved high precision in determining “Member” customers, while recall could suffer depending on priority, and this category proved to be reliable for such scenarios. “Normal" customer predictions were also too tough for "o Logistic Regression: its recall was low, and it did poorly at tracing non-linear relationships across features. “It’s confusion matrix seemed to show a slight bias towards predicting 'Member' customers. Accuracy: ~59%

•ROC-AUC Score: 0.58

Strengths:

- Logistic Regression provided a clear interpretation of feature importance, allowing us to understand which variables most influence membership classification.
- It demonstrated high precision for identifying "Member" customers, making it reliable for scenarios prioritizing accuracy over recall for this category.

Weaknesses:

- Logistic Regression struggled with "Normal" customer predictions, exhibiting lower recall and failing to capture non-linear relationships between features.
- The model showed a slight bias toward predicting "Member" customers, as indicated by its confusion matrix.

2. Random Forest Classifier

Objective: To improve classification accuracy, we use an ensemble based nonlinear model to capture the complex relationship between the variables.

Accuracy: ~57%

ROC-AUC Score: 0.57

Strengths:

1. Both Logistic Regression and the Random Forest Classifier effectively captured nonlinear interactions between features that Logistic Regression could not. A result of the model ranking feature importance was actionable insights of which features would be the most important in predicting customer segmentation into meaningful segments. It had better recall for "Normal" customers than Logistic Regression but is still more balanced given its need for sensitivity. The model's ability to rank feature importance provided actionable insights into the key predictors for customer segmentation.
2. It showed better recall for "Normal" customers compared to Logistic Regression, making it more balanced in scenarios requiring sensitivity.

Weaknesses:

1. However, the overall accuracy and ROC AUC score was slightly lower than Logistic Regression without considering the strengths in non-linear relationships. The same as Logistic Regression, the model was biased toward "Member" customers – but to less Logistic Regression.
2. Like Logistic Regression, the Random Forest model exhibited bias toward "Member" customers, although to a lesser extent.

Feature Importance Insights:

1. Predictors were "Gross Income," "Product Line," and "Quantity," all at the top.
2. Payment Method and gender between them had less significant impact onto the model, this means their contribution to distinguish customer types was modest. Trends less significantly to the model, indicating their limited role in differentiating customer types.

Evaluation Insights:

- **Confusion Matrix:** 'Normal' customer classification had improved performance compared to Logistic Regression.
- **Feature Importance Plot:** It looked at the most important predictors to see which features managers allocate their time and energy to when they make decisions.
- **Insert Visuals Here:**
- **Feature Importance Plot:** Ensure that your most important features are something like “Gross Income”, “Quantity”, “Product Line”.
- **Confusion Matrix:** True and false prediction of customer types based on display.

Model Selection Consideration:

- **Logistic Regression**: For tasks requiring interpretability, and prioritizing precision on "Member" customers, Logistic Regression is selected.
- **Random Forest Classifier**: For the case where recall for 'Normal' node is critical or the nonlinear patterns in the data need to be represented, Random Forest Classifier is a good choice.

Findings and Managerial Implications

Analysing supermarket transactions gave actionable insights that can directly influence the way marketing and inventory strategies are approached. Appropriately applied, these insights can help increase customer engagement, improve resource utilisation and maximise revenue. The implications are explored in detail here:

1. Average Spending by Product Category

Insights:
Customers spend the most on product categories like **Home and Lifestyle** ($336.64) and **Sports and Travel** ($332.07), making them the most profitable categories. Conversely, **Fashion Accessories** has the lowest average spending at $305.09, indicating that this category contributes the least to total sales.

Managerial Implication:
Given the high profitability of **Home and Lifestyle** and **Sports and Travel**, these categories should be prioritized for inventory stocking and promotional campaigns. Efforts should focus on expanding product variety and ensuring these items remain in stock. For **Fashion Accessories**, a strategic evaluation of pricing, assortment, or demand drivers is essential to increase its appeal to customers. Targeted discounts or bundling strategies could also help boost performance.

2. Frequency of Payment Methods

Insights:
Digital payment methods like **Ewallet** are the most popular, with 345 transactions, closely followed by **Cash** at 344 transactions. **Credit Card** usage is slightly less common, accounting for 311 transactions. These findings highlight the growing preference for convenience-driven digital payment methods among customers.

Managerial Implication:
Supermarkets should expand support for **Ewallet** by promoting digital payment campaigns such as cashback or loyalty points to encourage adoption. However, the high prevalence of cash transactions underscores the need to maintain robust cash-handling systems to ensure efficiency at checkout counters. Additionally, offering targeted promotions for credit card users can balance the share of all payment methods.

3. Branch Performance

Insights:
Among the three branches, the **Giza** branch leads with the highest average sales of $337.10, significantly outperforming **Cairo** ($319.87) and **Alex** ($312.35). This disparity suggests operational or customer base differences across branches.

Managerial Implication:
Efforts should focus on scaling the success of the **Giza** branch by allocating more marketing and operational resources to this location. Investigating the factors contributing to **Giza's**

performance—such as customer demographics, product mix, or service quality—can provide a blueprint for boosting sales at **Cairo** and **Alex**.

4. Customer Ratings

Insights:
The average customer rating is **6.97**, with most ratings falling between 5.5 and 8.5. While a majority of the ratings are positive, there are a few low ratings that suggest dissatisfaction in some areas.

Managerial Implication:
Mid-range ratings (5–6) indicate areas where improvements in customer service, product quality, or store environment can enhance satisfaction. Initiatives like personalized feedback collection and prompt resolution of complaints should be prioritized. Maintaining consistency in service delivery is crucial to retaining high ratings and fostering customer loyalty.

5. Quantity vs. Sales Relationship

Insights:
A strong positive correlation (R2=0.71R^2 = 0.71R2=0.71) was identified between the quantity of items purchased and the total sales amount. Customers who purchase in larger quantities contribute significantly to revenue.

Managerial Implication:
To capitalize on this trend, supermarkets should design promotional campaigns encouraging bulk purchases, such as discounts on higher quantities or bundled product deals. Educating customers about value-based pricing can also encourage larger transactions, ultimately driving sales growth.

6. City-wise Sales Contribution

Insights:
Cities like **Giza** contribute the most to total sales, highlighting their strategic importance for overall revenue generation.

Managerial Implication:
Marketing and operational scaling efforts should prioritize high-performing cities like **Giza**. Supermarkets in these locations can be used as testing grounds for new product launches or promotional strategies to maximize their impact on revenue.

7. Sales Trends by Hour

Insights:
Peak sales occur during specific hours, such as lunch breaks and evenings, likely reflecting customer availability during these times.

Managerial Implication:
Store staffing, inventory management, and promotional offers should align with these peak hours to enhance efficiency and maximize sales. Time-limited offers or flash discounts during peak periods could further incentivize purchases.

8. Gender-based Spending

Insights:
Spending patterns show gender-specific preferences across certain product lines, indicating varied interests and purchasing behaviors.

Managerial Implication:
Targeted marketing campaigns designed to cater to gender-based preferences can help optimize customer engagement. For example, advertising campaigns for **Health and Beauty** products may resonate better with female customers, while **Sports and Travel** promotions could be tailored for male customers.

9. Customer Type Spending

Insights:
Members generally spend more than Normal customers on average, highlighting the profitability of retaining and growing the Member base.

Managerial Implication:
Supermarkets should enhance loyalty programs to incentivize Normal customers to become Members. Offering exclusive discounts, personalized offers, and rewards for Members can help foster long-term relationships and maximize customer lifetime value.

10. Marketing Strategies for "Members"

Findings:
"Members" exhibit distinct purchasing behaviours characterized by:

- A preference for high-value items, particularly in product lines like "Electronics" and "Home Goods."
- A high consistency of their purchasing patterns (i.e. brand loyalty and a higher chance of re transactions).

Implications for Marketing:

- Loyalty Programs:
  - The ''Members'' that are spending consistently can be designed by marketing teams with tailored loyalty programs to reward.
  - Give discounts, early access deals to new product releases and bonus points on specific purchases. They allow creating tiered membership levels which entice your visitors to spend more and frequently visit your website." for their consistent spending. For example:
  - Offer discounts, exclusive early access to new product launches, or bonus points for purchases in preferred categories.
  - Create tiered membership levels to incentivize increased spending or more frequent visits.
- **Personalized Promotions:**
- Go ahead and use customer segmentation data to create promotions that are specific to "Members." For instance:

  - Send targeted emails or app notifications highlighting premium products or bundles in categories they frequently purchase.

- **Upselling and Cross-Selling Opportunities:**
  Marketing campaigns can recommend complementary or premium products to boost the average transaction value by analysing product line preferences.

10. 2. Inventory Optimization for "Normal" Customers

Findings:
"Normal" customers show:

- There is lot of variability in the buying habits which is normally attributable to external factors like promotional offers and discounts.
- Tendency to buy products from the lower cost categories or in the sale events. Seasonal promotions and discounts.

Implications for Inventory Management:

- **Seasonal Stock Planning:**
  Stock levels can align with Seasonal trends and promotional campaigns which attract 'Normal' customers. For example:

- By making excess inventory available for discounted, or popular seasonal items on sale periods, you can increase inventory.

- It will help to monitor trends on lower cost categories to ensure there is the required availability.

- **Demand Forecasting:**
  Preset data patterns are used to make a forecast for overall demand from "Normal" customers for specific product lines. This ensures that the stock levels are at an optimum level without being over the stock level or under the stock level.

- **Promotional Support:**
  Help marketing teams with aligning inventory with promotional activities. For instance:

  - To make the most out of high demand items when your products go on sale to the public, make sure you have plenty stocked.

  - Offer introducing bundle offers or value packs to incentivize 'Normal' customers to spend more.

Broader Managerial Implications

1. **Customer Retention and Conversion:**

   - Such as an incentivised membership program such as discounts or perks to be a member.

   - Keep existing “Members” by making sure the product lines they prefer stay in stock and by providing exclusive experiences.

2. **Revenue Maximization:**
   - Supermarkets can maximize revenue for diverse customer segment by focusing on high value product line for "Members” and promote strategies for "Normal" customers.
3. **Improved Resource Allocation:**
   - Analysing insights in the analysis helps optimize resources allocated to things such as marketing budgets, inventory management or deploying the workforce during peak sales periods.

Conclusions

Through this project we were able to prove that data driven approach and machine learning models can provide customers segmentation and predicting in a retail environment. The value insights we uncovered by leveraging supermarket transactional data can be leveraged to inform strategic decision making and operational efficiency.

Key Takeaways:

1. **Machine Learning Success:**
   - Also, machine learning models being used to classify customers as "Member" or "Normal" using Logistic Regression and Random Forest Classifier were demonstrated as being useful.
   - Overall performance of our analysis shows that Logistic Regression is better in terms of accuracy and interpretability. For scenarios where feature importance is a crucial thing to understand, it proved to be effective.
   - On the flipside, Random Forest was able to capture complex non-linear relationships, increasing the level of insight on to subtle patterns in customer behaviour.
2. **Insights into Customer Behaviour:**
   - Members always generate higher revenue and are loyal to certain product categories.
   - But normal customers behave in varied way of the buying pattern, the influence is very much exterior factors like discounts and promos.
3. **Managerial Implications:**
   - The findings provide a foundation for actionable strategies in marketing and inventory management, enabling more precise targeting and resource allocation.

Appendix:

Notebook file:

https://drive.google.com/file/d/1iBBoKZ84rO5pcD4q9Vc3Fr1p_EJWIDRJ/view?usp=sharing