**Optimizing Sales and Profitability through Predictive Analysis and Customer Segmentation in Retail**

## 1. Executive Summary

**Primary Goal**

The primary goal of this project is to enable a retail company to forecast and optimize future sales and profitability across its various product categories. By leveraging advanced analytics and customer segmentation, we aim to uncover insights that will refine the company's strategic decision-making.

**Approach**

Our approach utilizes predictive analytics to forecast sales trends and identify key factors impacting sales volumes and amounts. We also examine customer purchasing behaviors, demographics, and product preferences to tailor marketing efforts and inventory management.

**Expected Outcomes**

The expected outcomes include more accurate inventory planning, improved pricing strategies, and enhanced targeted marketing campaigns. These outcomes will drive business value by aligning product offerings with market demand, optimizing resource allocation, and fostering deeper customer engagement, thereby bolstering the company's competitive stance in the dynamic retail landscape.

## 2. Background, Context, and Domain Knowledge

Istanbul's shopping malls serve as vital economic and social landmarks. They are diverse ecosystems where consumer patterns unfold and where understanding these patterns is key for businesses to thrive.

Spanning two years of rich transactional data, the dataset provides the information into the shopping habits of Istanbul's citizens. With data points ranging from personal demographics to details of each purchase, we are equipped to delve deep into the analysis of consumer behaviors. The dataset comprises of the following:

* Invoice and customer identifiers
* Demographic details (age and gender)
* Transaction specifics (product categories, quantities, prices)
* Payment methods used
* Dates and locations of transactions

The dataset provides a clear map of shopping in Istanbul's malls. It uses simple facts about shoppers and their buying habits to help us understand how people decide what to buy. This information comes from retail analytics and consumer economics, giving us a peek into the choices shoppers make and how stores can use this to better serve them and grow their business.

## 3. Industry’s Traditional Approaches to the Problem

In the retail industry, traditional approaches to forecasting and sales optimization often rely on historical sales data, seasonal trends, and broad market research. Retailers typically observe past performance during comparable sales periods to anticipate future demand, employing standard replenishment and marketing strategies across various product categories.

This conventional methodology, while practical, does not account for rapid shifts in consumer behavior or real-time market dynamics. Strategic decision-making tends to follow a reactive pattern, with adjustments made post facto based on observed sales outcomes. Consequently, such strategies may lead to either overstocking or stockouts, inefficient resource allocation, and missed opportunities for personalized customer engagement. This aligns with the classic business model focused on volume and efficiency but often lacks the granularity provided by modern data analytics, potentially leaving retailers a step behind in an increasingly data-driven and customer-centric marketplace.

## 4. Analyses

We aim to break down the main business questions into three parts, in which we will apply different machine learning techniques.

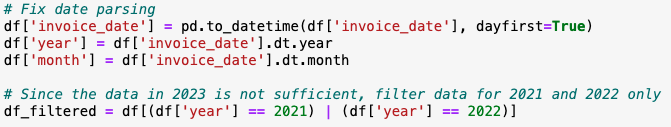
1. Predict Future Sales Trends

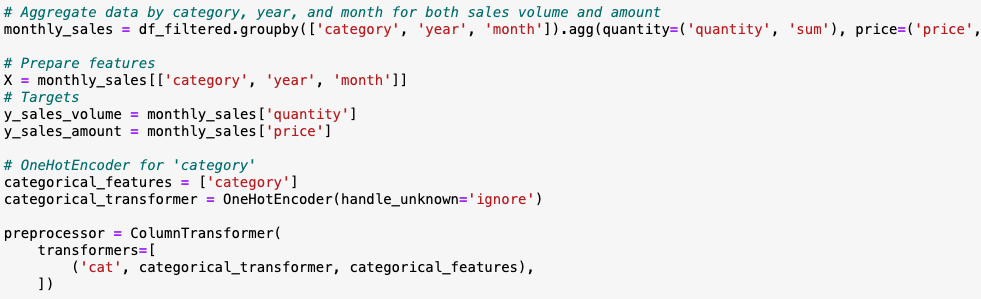
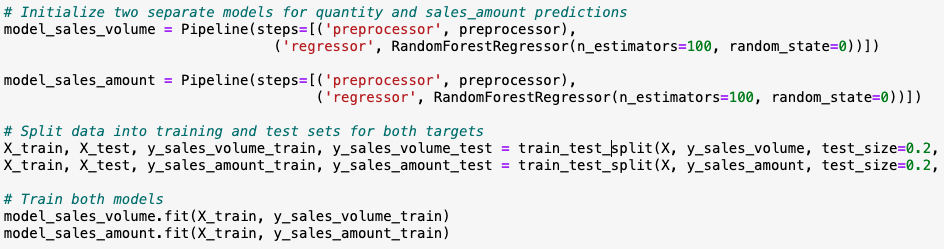
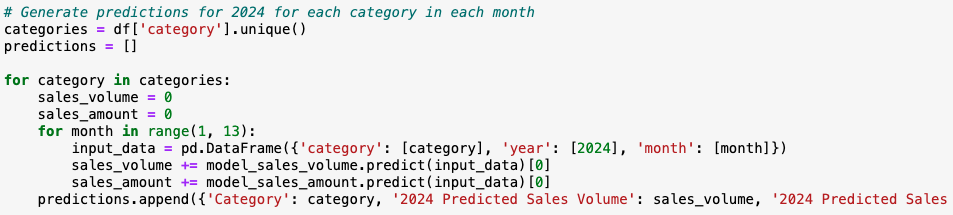
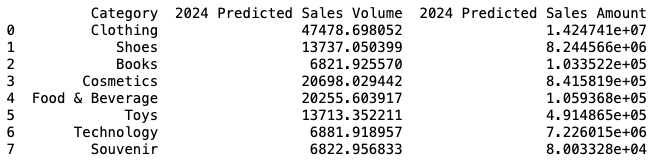
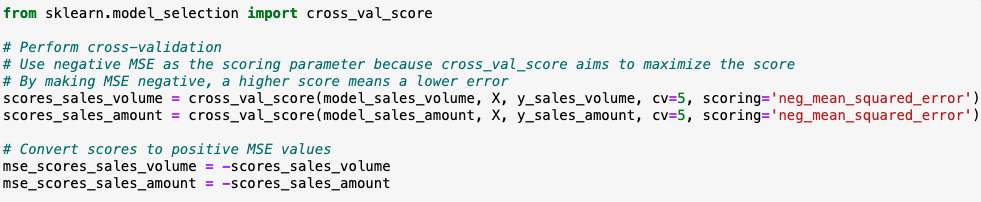
1. Methodology: Random Forest Regressor

The RandomForestRegressor model was chosen as the approach for predicting future sales volume and sales amount for its several advantageous features in handling non-linear relationships, feature interaction, and outliers.

1. Code Explanation and Value of the Procedure

* Data Preparation: The code begins by parsing dates correctly and filtering the dataset for the years 2021 and 2022, focusing on the complete years of data to train the model.



* Feature Engineering: It aggregates sales data monthly by category. This step transforms transactional data into a form where trends can be more easily identified and predicted.  
  
* Model Training: Separate RandomForestRegressor models are trained for predicting sales volume and amount. This dual-model approach allows for targeted predictions of both quantity and revenue.  
  
* Predictive Modeling: The model predicts sales for each category in 2024. This method translates complex, detailed models into actionable business insights, predicting annual sales performance by category.  
  
* Business Insights: The final output is a DataFrame showing predicted sales volume and amount by category for 2024.  
  
* Cross-validation: Finally, use cross-validation to ensure that the model's performance is consistent across different subsets of the data. The main advantages of cross-validation include providing a more robust measure of a model's predictive accuracy and helping to ensure that the model is not overly tailored to a specific set of data.  
  

2. Variable Selection and Feature Importance

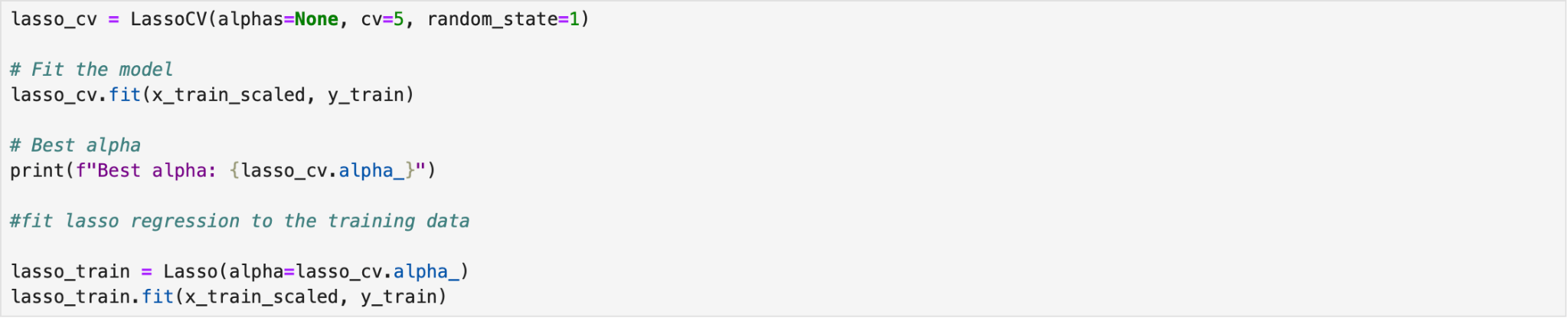
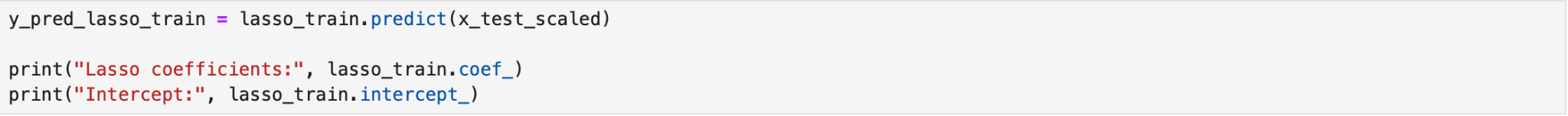
1. Methodology: Lasso Regression  
   Through the regularization and feature selection properties of lasso regression, we can identify the product categories or customer segments that have the most significant impact on sales amount. Features with large, non-zero coefficients have strong, positive associations with sales amount while features with zero-coefficients are deemed less influential and excluded from the model.
2. Code Explanation and Value of the Procedure

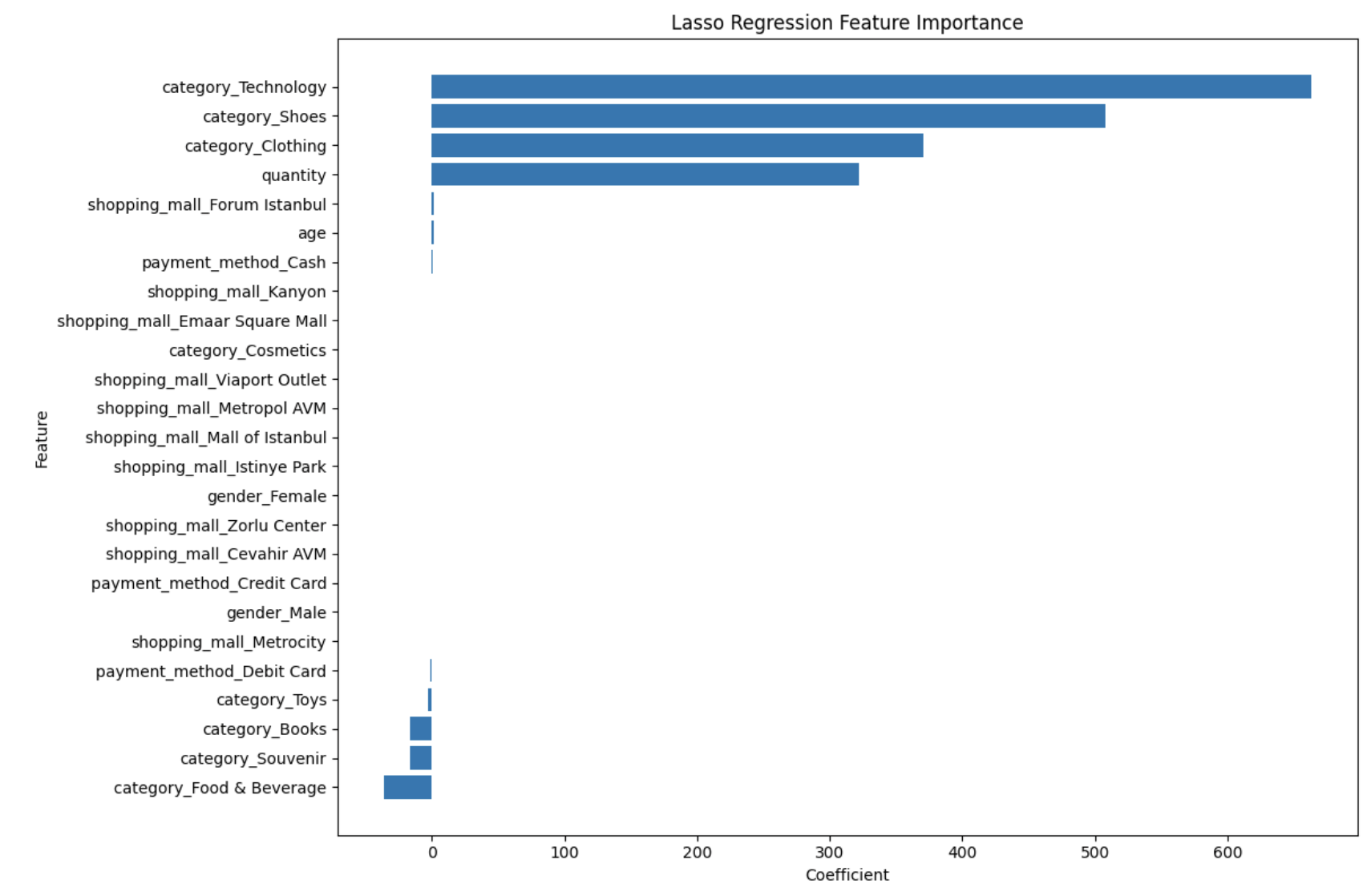
* Data Preparation: First, we convert the categorical variables into dummy variables. The variables gender, category, payment method and shopping mall are the primary segments of interest.



After creating dummy variables, they are included in the independent variable “x”. The dependent variable “y” is represented by the “Price” as it represents the sales amount of the transaction. The data is then split to form a training set and a testing set. This allows for the evaluation of the model’s performance on out-of-sample data. The features are also standardized to ensure that features are on a similar scale.



* Fitting the Model: The lasso model is fit to the scaled training data. Using cross validation, the best alpha is determined to minimize the mean squared error. The resulting lasso model is fit with this best alpha, enhancing its predictive performance on the training data.
* Prediction and Feature Importance: The trained lasso model is used to predict prices based on the scaled testing data. The coefficients from the trained model reflect the impact of each feature on the predicted prices. 

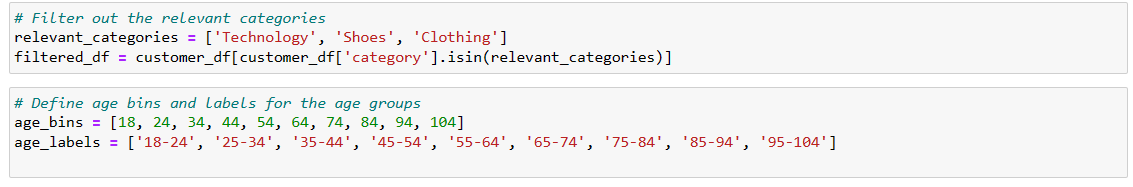


* Evaluate the Model: The model is evaluated using mean squared error (MSE), root mean squared error (RMSE) and R-squared. Lower MSE and RMSE values suggest that the model’s predictions are closer to the actual prices, on average. Approximately 83.7% of the variability in prices is captured by the lasso model. 

3. Customer Segmentation

1. Methodology: This step is primarily descriptive and exploratory, aiming to provide insights into average spending within customer segments. By segmenting customers and calculating average spending within each segment for selected categories, the analysis provides a detailed look at spending patterns. This method accounts for the variability in spending within segments and categories, offering a nuanced view that simple averages or total sums might miss.
2. Code Explanation and Value of the Procedure

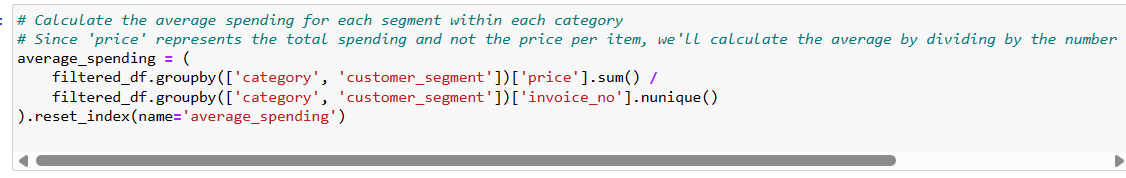
* Data Loading and Filtering: The customer shopping data is loaded into a DataFrame, and entries are filtered to focus on the most impactful categories (Technology, Shoes, Clothing) identified in the previous step. The narrowing down makes the subsequent analysis more targeted and manageable.



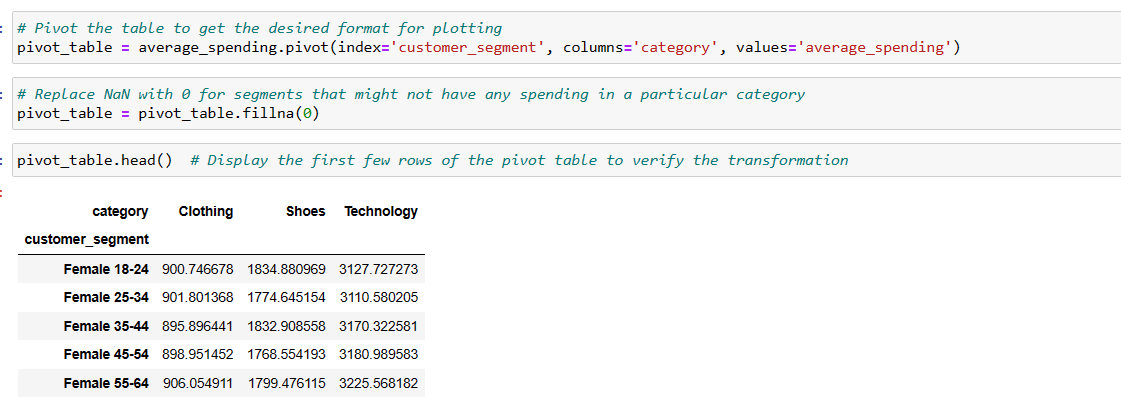
* Customer Segmentation: Customers are segmented into groups based on age and gender. Age bins are defined, and customers are categorized into these bins. A new segmentation label combining gender and age group is created. Segmentation allows for a more nuanced understanding of spending habits.



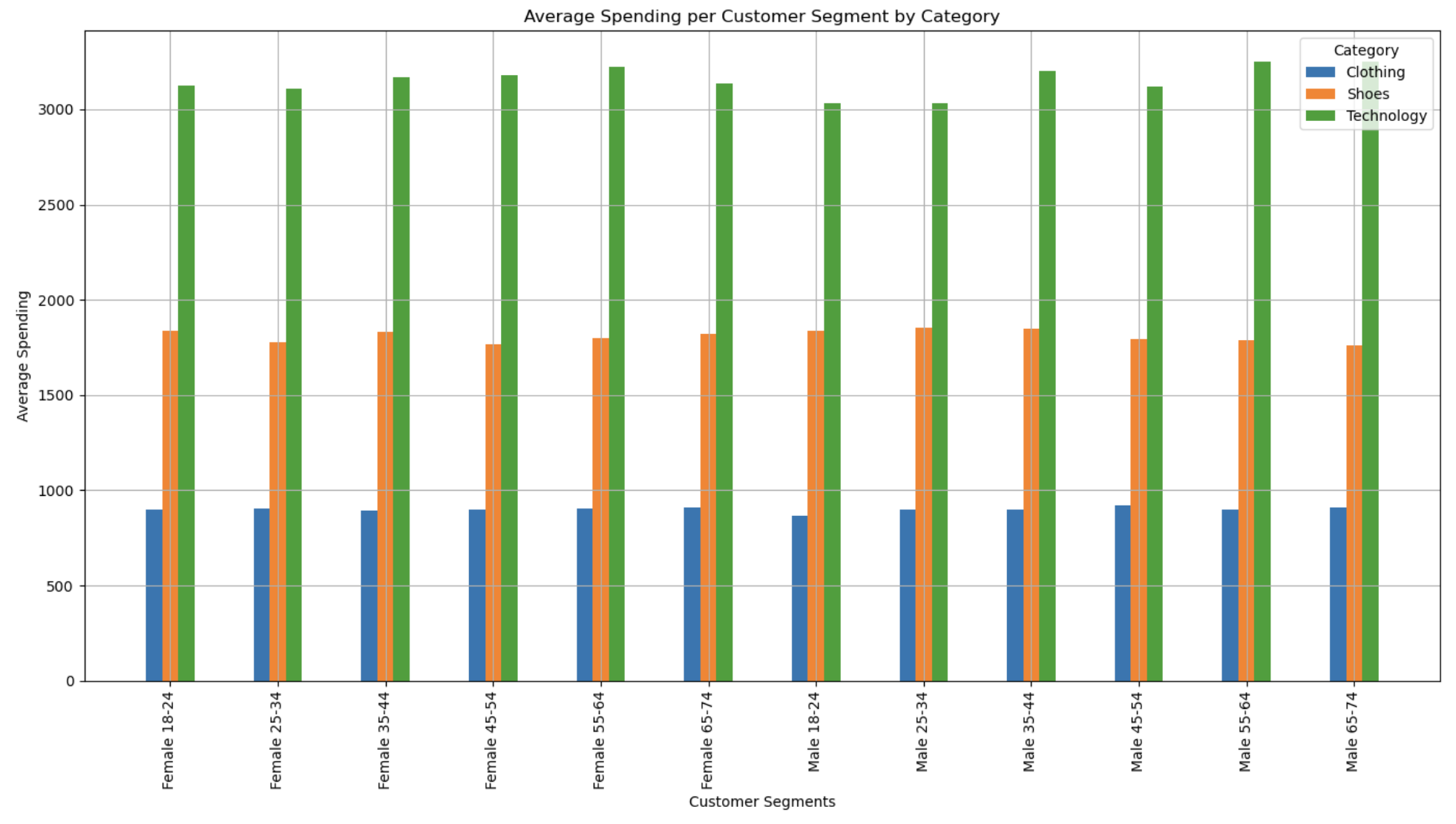
* Average Spending Calculation: The code calculates the average spending per customer segment within each category by dividing the total spent by the number of unique invoices. This approach provides insight into how much, on average, segments are spending in each category.



* Data Transformation: The data is transformed into a pivot table format, with customer segments as rows and product categories as columns. This format facilitates easy comparison of average spending across segments and categories, making it easier to visualize and understand the spending patterns.



* Visualization: A bar chart is generated to visually compare the average spending of different customer segments across the selected product categories, helping in quickly identifying which customer segments are spending more in each category, highlighting opportunities for targeted marketing and decision-making



## 5. Recommendations and Business Value

The outcomes from each of the three analyses can be assessed to drive decisions that aim to maximize the value of the business.

Predict Future Sales in 2024 *(Unit for Sales Amount: US dollars)*

|  |  |  |
| --- | --- | --- |
| Category | Predicted Sales Volume | Predicted Sales Amount |
| Clothing | 47,479 | $ 14,247,410 |
| Shoes | 13,737 | $ 8,244,566 |
| Technology | 6,882 | $ 7,226,015 |
| Cosmetics | 20,698 | $ 841,582 |
| Toys | 13,713 | $ 491,487 |
| Food & Beverage | 20,256 | $ 105,937 |
| Books | 6,822 | $ 103,352 |
| Souvenir | 6,823 | $ 80,033 |

The analysis indicates a strategic opportunity for the retail company to enhance its sales strategy based on predictive analytics. With the expected sales volumes and corresponding sales amounts for Clothing, Shoes, and Technology, it is recommended that the company prioritizes these categories in inventory stocking and marketing efforts.

Variable Selection and Feature Importance

The feature importance derived from the lasso regression model is as follows:

|  |  |
| --- | --- |
| Category | Coefficient |
| Technology | 662.9798 |
| Shoes | 507.6683 |
| Clothing | 370.5038 |
| . | . |
| . | . |
| . | . |
| Toys | -15.9895 |
| Souvenir | -16.0532 |
| Food & Beverage | -35.8273 |

Based on the results above, the coefficients for the categories Technology, Shoes and Clothing are the highest, indicating that the quantity sold in these particular categories has the largest effect on the transaction amount. This suggests that focusing on strategies related to these three categories, also proven in predictive analytics, could potentially lead to increased transaction amounts.

Customer Segmentation *(Unit: US dollars)*

|  |  |  |  |
| --- | --- | --- | --- |
|  | Clothing | Shoes | Technology |
| Female 18-24 | $900.75 | $1,834.88 | $3,127.73 |
| Female 25-34 | $901.80 | $1,774.65 | $3,110.58 |
| Female 35-44 | $895.90 | $1,832.91 | $3,170.32 |
| Female 45-54 | $898.95 | $1,768.55 | $3,180.99 |
| Female 55-64 | $906.06 | $1,799.48 | $3,225.57 |
| Female 65-74 | $907.49 | $1,818.81 | $3,137.68 |
| Male 18-24 | $864.04 | $1,835.37 | $3,031.69 |
| Male 25-34 | $899.78 | $1,856.43 | $3,031.18 |
| Male 35-44 | $897.00 | $1,848.84 | $3,202.01 |
| Male 45-54 | $922.32 | $1,795.83 | $3,118.42 |
| Male 55-64 | $899.43 | $1,787.96 | $3,249.86 |
| Male 65-74 | $910.19 | $1,758.81 | $3,250.00 |

The data suggests a notable customer willingness to invest in Technology, with average spending in certain segments reaching $3,250.00 for males aged 65-74 and $3,127.73 for females aged 18-24. These insights should inform targeted marketing campaigns and promotional offers to maximize the revenue potential. Additionally, the Shoes and Clothing categories show substantial spending, indicating that strategic pricing in these segments can increase profitability.

Investing in customer experience, particularly for high-value segments identified through spending patterns, can significantly boost customer loyalty and lifetime value. By tailoring the product mix, store layouts, and online presence to these insights, the company can enhance its competitive edge in the market. Through these focused strategies, leveraging the robust spending capacity in key categories, the company is poised to not only meet but exceed its sales and profitability targets.

## 6. Conclusions

The project's integrated analytical approach has highlighted the benefits of combining predictive sales trend analysis with customer segmentation. Overall, a data-driven focus on key product categories and consumer demographics can significantly enhance the retail company's strategic initiatives. By adopting the recommended strategies, the company is set to strengthen its market presence, achieve higher sales volumes, and improve overall profitability. In a rapidly evolving retail environment, such an approach is both advantageous and essential to competitive advantage and fostering long-term business growth.