**University of Texas at Dallas**

**Report**

**Analytics Practicum – BUAN 6390.501**

**Milestone 2 -- Analysis & Model Development.**

**Professor: Sunela Thomas**

**Group: 01**

**Melody Tavallaee**

**Harshita Yakkantii**

**Bhavith Akula**

**Introduction:**

This report begins our deeper-level analysis on our data to understand the segments in the customer base seen in our dataset. Our objective in this report is to gain an understanding of our customer data so we can begin the process of constructing meaningful and actionable business recommendations in the later phases of our project. Our approach for this report started with conducting Exploratory Data Analysis (EDA). This first level of analysis will help us have a big picture understanding of the data we are looking at. This is valuable knowledge that will help us consider the context of our data and help us to fully understand the impacts of business decisions in our customers. Following our EDA step, we moved on to applying analytical techniques to help us segment our customers into clusters. We also built linear regression models to help us predict customer behavior. These clusters will help us create meaningful segments to understand customers’ behavior.

**Exploratory Data Analysis (EDA):**

Each visualization offers valuable insights into the business's operations and customer behaviors. The significant drop in sales in January 2023 and the spending patterns could be key areas to focus on for strategic planning and operational adjustments. These charts together can help tailor marketing efforts and product offerings to better match customer demographics and spending habits.

1. **Age Distribution:** This bar chart with a line graph overlay shows the frequency of customers by age, smoothed to highlight trends. There’s a relatively even distribution across age groups, with slight peaks around ages 30 and 50. This suggests a diverse customer base, possibly indicating that the business appeals to a wide demographic.

A graph of a number of customers

AI-generated content may be incorrect.

1. **Gender Distribution:** Overall, there are more female customers than male, as shown in this bar chart. The distribution is not highly skewed, but it does show a noticeable preference, which could align with the previous chart's insights into category preferences by gender.

**A graph showing a number of people in different colors

AI-generated content may be incorrect.**

1. **Most popular product:** This bar chart compares the number of purchases across various categories by gender. Overall, clothing tends to be the most popular category across all customers. Females tend to dominate all the categories, particularly in cosmetics and clothing.

A bar graph with numbers and text

AI-generated content may be incorrect.

1. **Total Spending Distribution**: This histogram displays the frequency of transactions across different spending amounts. Most transactions are concentrated at the lower end of spending, with a significant peak at around $0 to $1,000. There are periodic smaller peaks, suggesting that specific spending ranges might correspond to popular product price points or promotional pricing tiers.

.**A graph of a distribution of spending

AI-generated content may be incorrect.**

1. **Monthly Sales Trends:** This graph shows the monthly sales trends over a period from January 2021 to January 2023. Sales are relatively stable with minor fluctuations until a sharp decline is observed in January 2023. The y-axis represents total sales on a logarithmic scale (1e7 indicates $10,000,000), emphasizing the magnitude of sales values. This dramatic drop could indicate an issue such as a data entry error, significant market change, or operational problem that needs further investigation.

**A graph with blue lines and numbers

AI-generated content may be incorrect.**

**Outlier Analysis**

Addressing these data outlier challenges is essential for ensuring accurate customer segmentation and predictive modelling. By implementing techniques like **outlier handling, feature selection, Ridge Regression, and the Elbow Method**, the dataset was refined for better analysis.

#### **Key Takeaways:**

* **Data Cleaning & Outlier Handling Improved Model Accuracy.**
* **Ridge Regression was the Best Fit for Predicting Spending.**
* **Customer Segmentation with K-Means Provided Actionable Insights.**

Future steps involve **enhancing predictive models with advanced time-series forecasting** and **further refining customer segmentation using additional behavioral metrics**.

**Introduction**

For this part of the report, we will be analyzing and addressing outliers in three categories: Age, Quantity, and Price. Outliers are data points that significantly deviate from the dataset and may indicate errors or unique cases. Therefore, we would need to properly handle any outliers to ensure our analysis is not biased or skewed from abnormal data points.

1. **Age Outliers**

Data Points: The data shows the ages of our customers.

Approach: We will check “Age” for any abnormal looking points. For example, "1" is a potential outlier, since it is not feasible to assume any of the customers would be 1 years old. Ages 20-70 would be more typical to see in the context of our data.

**Recommendation**: Verify any abnormal ages, such as "1", for accuracy and decide if it should be included or treated as a special case.

1. **Quantity Outliers**

Data Points: This shows the quantity of items bought by the customers.

Approach: We will analyze the average amount of quantities, checking for any that seem to be abnormally high or low. These could indicate errors or special cases.

**Recommendation:** Investigate the context of these quantities. Correct errors or treat valid cases separately.

1. **Price Outliers**

Data Points: These show the amount of money spent by customers.

Approach: Check the data points for price. Prices such as "0" and "1" are likely outliers, that may be due to data entry errors. Higher prices (1000-5000) may also be outliers if typical prices are lower.

**Recommendation**: Validate low and high prices. Correct errors or treat valid cases separately.

**Outlier Results**

* Age: No extreme outliers are visible, but we should check for unusual values (e.g., unrealistic ages like below 10 or above 100).
* Quantity: There are no major outliers seen in our dataset.
* Price: High-price outliers suggest possible luxury or bulk purchases.

**Conclusion:** Outliers in Age, Quantity, and Price need to be validated to determine whether they are errors or valid cases.

**A diagram of a number of objects

AI-generated content may be incorrect.**

**Developing and Implementing Analytical Techniques & Machine Learning Models**

This section aims to develop and implement **analytical techniques** and **machine learning models** for customer segmentation and spending prediction. This includes:

1. **Customer Segmentation** using K-Means clustering based on RFM (Recency, Frequency, Monetary) analysis.
2. **Predictive Modeling** using Linear Regression, and ridge regression to forecast customer spending behavior.

### **Customer Segmentation Using K-Means Clustering**

#### **RFM Analysis for Segmentation**

* **Recency (R):** Number of days since the customer’s last purchase.
* **Frequency (F):** Number of transactions made by the customer.
* **Monetary (M):** Total amount spent by the customer.

A white background with black text

AI-generated content may be incorrect.

#### **Data Preparation for Clustering**

* Data was aggregated at the customer level.
* RFM values were normalized to ensure uniform clustering.
* The optimal number of clusters was determined using the **Elbow Method**, selecting **4 clusters**.

#### **Interpretation of Customer Segments**

| **Segment** | **Characteristics** |
| --- | --- |
| **High-Value Customers** | Frequent buyers with high spending. |
| **Loyal Customers** | Regular buyers with medium spending. |
| **Occasional Shoppers** | Infrequent buyers with moderate spending. |
| **One-Time Buyers** | Customers with a single or minimal purchase. |

### **Predictive Modeling for Customer Spending**

#### **Data Preprocessing for Regression Models**

* **Features Used:** Age, Quantity, Price.
* **Target Variable:** Total Spending (calculated as quantity \* price).
* **Data Split:** 80% training, 20% testing.

**Regression Models Implemented**

##### **1️. Linear Regression**

* A basic model to establish a relationship between customer features and spending.
* **Findings:** Achieved **94.5% R² score**, but prone to overfitting.

##### **2️. Ridge Regression**

* Adds L2 regularization to improve generalization.
* **Findings:** Similar R² score to Linear Regression but reduced overfitting.

#### **Initial Model Performance Results**

A black text on a white background

AI-generated content may be incorrect.

A close up of a message

AI-generated content may be incorrect.

**Key Insights**

* **Ridge Regression is the best choice** as it balances prediction accuracy and generalization.
* Spending predictions are **strongly influenced by quantity and price**.

### **Challenges & Solutions**

| **Challenge** | **Solution** |
| --- | --- |
| High multicollinearity in features | Used Ridge Regression to penalize large coefficients. |
| Outliers in price and quantity | Capped outliers using IQR method. |

### **Conclusion**

### **This milestone in our project has helped us create a deeper knowledge of our customer base, which has allowed us to begin our data analysis and model building. Customer segmentation helped categorize customers based on spending habits, and how frequently they are likely to make purchases. Our regression models successfully predicted total spending, with Ridge Regression being the most effective method. Through combining all our analysis from this report together, we will have to tools to eventually apply our findings to developing meaningful and actionable insights for the business. We now have the knowledge to better understand the segments in our audience and will be able to target these segments more efficiently.**