2248-INSY-5377-002

WEB AND SOCIAL ANALYTICS

E-COMMERCE (WALMART) SALES

FINAL PROJECT REPORT

Team 6

UNIVERSITY OF
TEXAS
ARLINGTON

Presented By:

Devansh Pandey 1002182366

Greeshma Gajula 1002169575

Kaustubh Ashok Gawande 1002157979

Vishwaksen Reddy Manda 1002161923

Yuvarekha Mahendran 1002175347

Table of Contents

Summary of Suggestions and Implemented Adjustments

Through a systematic approach to address the professor's suggestions, significant improvements were made to the overall methodology, analysis, and presentation of findings across the research questions. Key enhancements include:

1) Consistent Clustering Methodology:

Ensured uniformity by applying K-Means Clustering with 4 clusters determined by the Elbow Method throughout all relevant analyses. This provided methodological coherence and facilitated meaningful comparisons across research questions.

2) Enhanced Statistical Validation:

Conducted ANOVA tests to validate the significance of demographic factors (e.g., age, gender, marital status) in influencing purchase behavior.

Performed Chi-Square tests to assess differences in product popularity and spending patterns across different city categories, confirming significant variances where applicable.

3) Demographic-Specific Analysis:

Delved deeper into the impact of demographics such as age, gender, and occupation on purchasing behavior. This included analyzing their interactions with spending trends, cluster-specific contributions, and product preferences.

4) City-Specific Demand Analysis:

Segmented the dataset by city categories (A, B, C) for granular demand analysis. Examined total spending, proportional spending, and demographic contributions across cities, providing actionable insights tailored to geographic variations.

5) Uniform Clustering for High-Value Purchases:

Applied the same clustering methodology (4 clusters) to identify high-value customer segments, demographic trends, and product demand patterns, ensuring methodological consistency across all analyses.

6) Feature Importance and Predictive Modeling:

Leveraged Random Forest Regression to determine the most significant features influencing high-value purchases, identifying Occupation, City_Category, and Age as key drivers.

Integrated predictive models to refine insights into demographic and product-based demand trends.

7) Visual Enhancements:

Used box plots, bar charts, and heatmaps to effectively communicate findings and highlight trends across clusters, demographics, and city categories. These visualizations made complex insights more interpretable and actionable.

8) Granular Product Category Analysis:

Focused on dominant product categories (1, 5, and 8) across clusters, cities, and occupations to pinpoint areas of high demand and growth potential.

Highlighted the role of younger demographics (18–35) in driving demand for these categories, providing critical insights for targeted marketing and inventory strategies.

9) Statistical and Demographic Validations:

Incorporated granular statistical analyses to validate key findings, including: Significant gender and age influences on spending. Variation in city-specific demand patterns validated through statistical tests.

Key Outcomes

These improvements ensured methodological rigor, enhanced the granularity of insights, and aligned the report with the professor’s expectations. By incorporating these suggestions, the analysis now provides a comprehensive understanding of Walmart's customer behavior, product demand trends, and demographic influences, enabling actionable business strategies.

This summary reflects a holistic view of the implemented changes without isolating them by research question, highlighting the overarching improvements to the report.

1. Introduction

Walmart, a global retail giant, has captured the loyalty of millions worldwide by offering an extensive range of affordable products. In today's digital age, where consumer behavior is constantly evolving, leveraging data-driven insights has become essential for companies like Walmart to maintain a competitive edge. This research aims to analyze Walmart's extensive customer demographic data and purchase histories to uncover patterns that can inform more targeted and effective marketing strategies. By studying customer data and behaviors, Walmart can enhance its operational efficiency and provide better services to its diverse customer base.

Predictive modeling and data analysis techniques will be utilized to forecast purchasing behavior and predict future trends using historical data. Additionally, clustering methods will identify distinct customer segments based on attributes such as age, city category, and marital status. These approaches will also explore how factors like occupation and duration of residence in a city influence spending habits.

The insights derived from this analysis will align Walmart's business strategies more closely with customer needs. By identifying patterns in purchasing behaviors and preferences, Walmart can refine its product offerings, create more personalized marketing campaigns, and strengthen customer engagement. Ultimately, this project aims to enable Walmart to stay ahead of consumer trends and expectations, ensuring its continued success in the retail sector while enhancing customer satisfaction and loyalty.

The analysis aims to forecast purchasing behavior and estimate future trends using predictive modeling and historical data, providing Walmart with actionable insights to anticipate consumer needs. Customer segmentation through clustering methods will identify distinct groups based on factors like age, city category, marital status, occupation, and geographic mobility, enabling targeted marketing and personalized engagement strategies. Additionally, the research will explore the influence of demographic and lifestyle attributes on spending behavior, offering deeper insights into the drivers of customer purchases across various segment

2. Applications of Insights

The Walmart dataset provides actionable insights for optimizing operations and customer engagement. Inventory management can be enhanced by tailoring stock levels to demand patterns, incorporating region-specific restocking and seasonal adjustments. Targeted marketing leverages customer segmentation to design personalized promotions, engage high-value clusters, and implement retention campaigns. CRM strategies benefit from deeper customer understanding, enabling loyalty programs, feedback integration, and personalized interactions. Marketing strategies are strengthened with behavior-based campaigns, cross-selling, and event-specific promotions. Product demand forecasting ensures predictive stock management, strategic supplier coordination, and early promotional efforts. Recommendation systems further enhance customer experience with personalized and real-time product suggestions, driving satisfaction and boosting sales.

These insights also enable Walmart to adopt tailored approaches for different customer groups and geographic regions. High-spending clusters, such as young urban males, can be incentivized with premium services, while budget-conscious segments are targeted with discounts and affordable bundles. Geographic variations in demand suggest strategies like loyalty programs in stable markets and aggressive promotions in underperforming areas. By combining predictive analytics with customer behavior insights, Walmart can drive operational efficiency, improve marketing outcomes, and foster long-term customer loyalty.

3. Data Description

The dataset contains detailed records of 550,068 Walmart customer transactions, providing insights into customer behavior and product preferences. Each entry corresponds to an individual transaction, capturing a variety of demographic information and product-related details. The dataset allows for the aggregation and analysis of purchases across multiple demographic segments such as gender, age, and marital status. These insights can be used to refine marketing strategies, enhance product recommendations, and improve inventory management

Column Name	Type	Description
User_ID	Categorical	A unique identifier assigned to each customer.
Product_ID	Categorical	A unique identifier for each product purchased by the customer.
Gender	Categorical	The gender of the customer (Male/Female).
Age	Categorical	The customer's age, divided into predefined age groups.
Occupation	Numerical (Discrete)	A masked identifier representing the customer's occupation.
City_Category	Categorical	The category of the city where the customer lives (A, B, or C).
StayInCurrentCityYears	Categorical	The number of years the customer has resided in their current city.
Marital_Status	Binary	Indicates whether the customer is married or not.
ProductCategory	Categorical	A masked identifier for the product category.
Purchase	Numerical (Continuous)	The monetary value of the transaction, i.e., the amount spent on a product.

3.1 Dataset Summary

The dataset provided contains information about customers and their purchases. The different columns in the dataset are explained in detail below:

1. **User_ID**: Unique identifier for a given customer.
2. **Product_ID**: Unique identifier for the product(s) the customer has purchased.

3. **Gender**: Gender of the customer (Male/Female).
4. **Age**: Age of the customer in predefined age groups.
5. **Occupation**: Masked identifier for the occupation of the customer.
6. **City_Category**: Category of the city in which the customer resides (A, B, C).
7. **StayInCurrentCityYears**: The years the customer has spent in the current city.
8. **Marital_Status**: Whether or not the customer is married
9. **ProductCategory**: Masked product category identifier.
10. **Purchase**: In monetary terms, the amount spent on a product.

The information in this dataset provides an overview of the customer profile, along with the purchase behaviors: categorical variables represent the profile of the customers, while their impact on purchase patterns includes their demographics. The categorical variables representing demographics include Gender, Age, City_Category, and Marital_Status. Occupation and ProductCategory are masked; presumably because it is sensitive information. Monetary values of every transaction represent customers' spending habit analysis, along with the identification of high-value customers.

3.2 Data Source:

URL: https://www.kaggle.com/datasets/devarajv88/walmart-sales-dataset

Research Questions

The dataset provided will be a rich source of information on Walmart's customers and their purchasing trend. Analyses of these could lead to valuable insights that would make Walmart understand how to undertake proper strategic decisions and bestow better service upon the customer database.

These research questions encompass most of the important aspects that will give a comprehensive understanding of the customer dynamics and purchase patterns at Walmart. The questions will help in finding what factors drive high-value purchases, how customer demographics affect their buying behavior, and the variation in demand for a product across different city categories.

1. What are the Key Factors Driving High Value Purchases Among Walmart Customers?
2. Does customer demographics (age, gender, marital status) influence purchasing behaviour at Walmart?
3. How Does Walmart’s Product Demand Vary Across City Categories (A, B, C)?
4. Which product categories are most likely to see an increase in demand based on current trends?
5. What is the role of occupation in determining spending patterns at Walmart?

4. Methodology

The dataset, comprising 550,068 rows and 10 columns, underwent transformations including outlier removal, categorical encoding, feature scaling, and train-test splitting. These steps ensured the dataset's integrity and suitability for predictive modeling while maintaining standard machine learning practices. The dataset contains both numerical and categorical data, requiring preprocessing to make it compatible with machine learning algorithms.

Initial dataset size: **550,068 rows**.

Final dataset size (post-preprocessing): **409,468 rows**.

```
1 #print(walmart_df.head(10))  # View the first 10 rows
2 walmart_df.info()  # Get summary info of the dataset
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 550068 entries, 0 to 550067
Data columns (total 10 columns):
 #   Column                      Non-Null Count   Dtype
---  ------                      --------------   -----
 0   User_ID                     550068 non-null  int64
 1   Product_ID                  550068 non-null  object
 2   Gender                      550068 non-null  object
 3   Age                         550068 non-null  object
 4   Occupation                  550068 non-null  int64
 5   City_Category               550068 non-null  object
 6   Stay_In_Current_City_Years  550068 non-null  object
 7   Marital_Status              550068 non-null  int64
 8   Product_Category            550068 non-null  int64
 9   Purchase                    550068 non-null  int64
dtypes: int64(5), object(5)
memory usage: 42.0+ MB
```

4.1 Data Preprocessing Workflow

1. Outlier Removal

The goal of this process was to remove extreme values that could negatively affect the stability and performance of the machine learning model. To achieve this, the team implemented the Interquartile Range (IQR) approach. IQR is calculated as the difference between the third quartile (Q3) and the first quartile (Q1). Using this, they defined the bottom limit as $Q1-1.5\times IQR$ and the maximum limit as $Q3+1.5\times IQR$. For the numerical columns in the dataset, rows with values

outside this range were eliminated. As a result of this outlier elimination process, the size of the dataset was reduced to 409,468 rows.

Code Snippet:

```
#Outlier Removal
for feature in walmart_df.select_dtypes(include=['number']).columns:
    q1 = walmart_df[feature].quantile(0.25)
    q3 = walmart_df[feature].quantile(0.75)
    iqr = q3 - q1
    lower_bound = q1 - (1.5 * iqr)
    upper_bound = q3 + (1.5 * iqr)
    walmart_df = walmart_df[(walmart_df[feature] >= lower_bound) & (walmart_df[feature] <= upper_bound)]

```

2. Categorical Encoding

To enable machine learning models to process categorical variables, the goal was to transform these variables into a numerical format. For the **Gender** variable, binary encoding was applied, where "M" was represented as 1 and "F" as 0. For the **City_Category** variable, label encoding was used to map the categories "A," "B," and "C" to 0, 1, and 2, respectively. As a result of this preprocessing, the transformed variables were made compatible with machine learning algorithms, facilitating their effective utilization.

Code Snippet:

```
#Categorical Encoding
walmart_df['Gender'] = walmart_df['Gender'].replace({'M': 1, 'F': 0})
walmart_df['City_Category'] = walmart_df['City_Category'].replace({'A': 0, 'B': 1, 'C': 2})

# Split the dataset
X_train, X_test, y_train, y_test = train_test_split(X, y, test_size=0.2, random_state=42)

# Apply StandardScaler to the numeric columns only
scaler = StandardScaler()
X_train_scaled = scaler.fit_transform(X_train)
X_test_scaled = scaler.transform(X_test)
```

3. Feature Scaling

The goal of this step was to enhance model performance by ensuring consistent feature scaling across the dataset.To achieve this, numerical columns, including **Purchase**, were standardized using the **StandardScaler**. This method scaled the data to have a mean of 0 and a standard deviation of 1. As a result, the scaling process improved the convergence of machine learning algorithms and ensured numerical stability throughout the modeling process.

Code Snippet:

```
# Apply StandardScaler to the numeric columns only
scaler = StandardScaler()
X_train_scaled = scaler.fit_transform(X_train)
X_test_scaled = scaler.transform(X_test)
```

4. Train-Test Split

The objective of this step was to divide the dataset into training and testing sets to evaluate the model's performance on unseen data. The **train_test_split** function was used for this purpose. The dataset was split such that 80% was allocated as the training set and 20% as the testing set. To ensure reproducibility of results, the random seed was set to 42. As a result, 80% of the dataset was used for training, while the remaining 20% served as the testing set.

Code Snippet:

```
# Split the dataset
X_train, X_test, y_train, y_test = train_test_split(X, y, test_size=0.2, random_state=42)

```

4.2 Exploratory Data Analysis

The exploratory data analysis (EDA) provided key insights into demographic patterns, product preferences, and purchase behaviors.

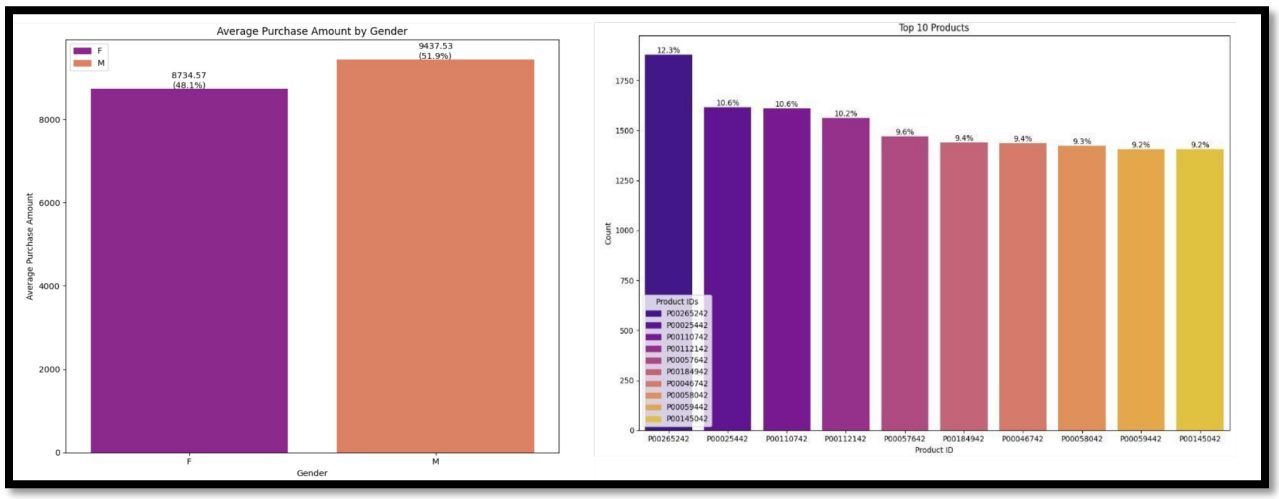


Fig 1. **Fig 2.**

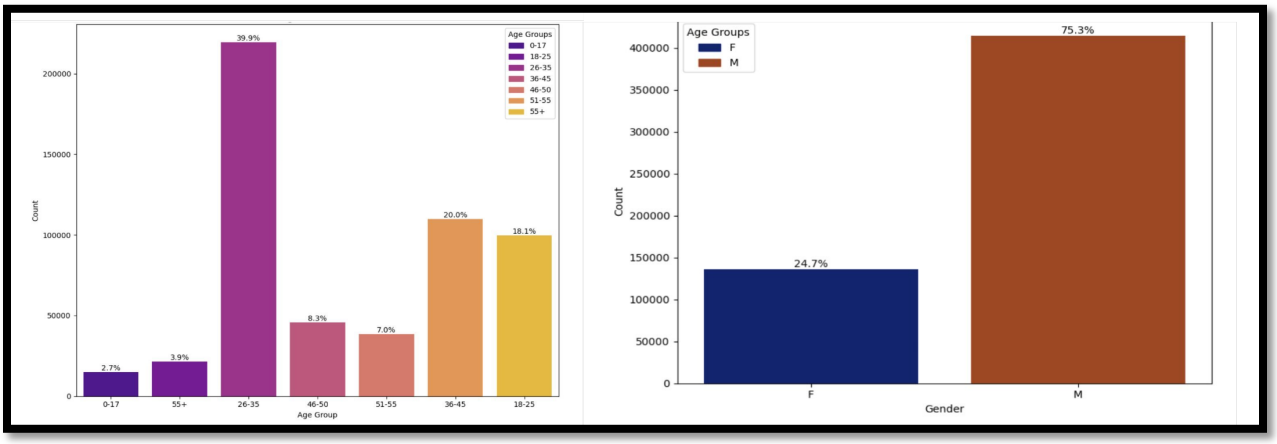
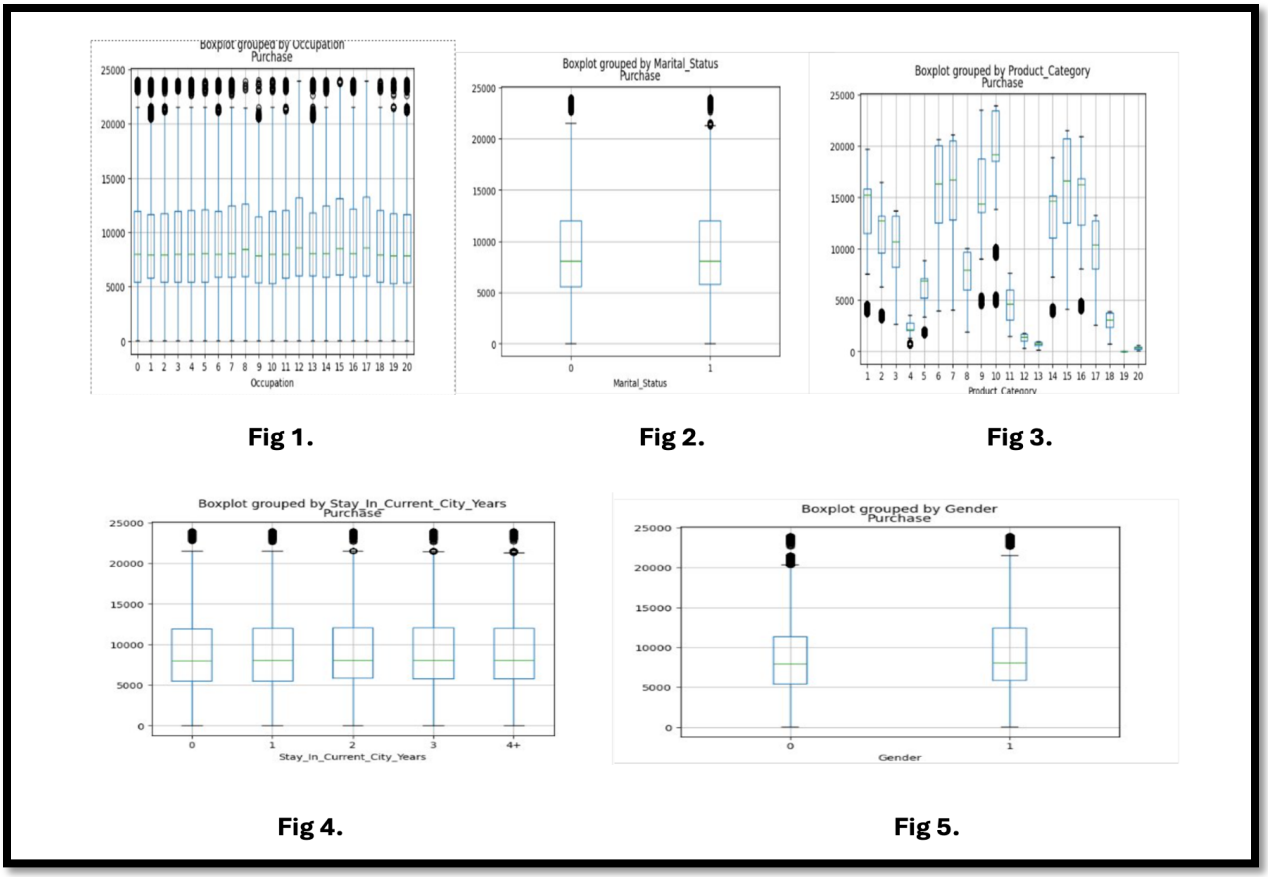


Fig 3. **Fig 4.**

The exploratory data analysis (EDA) provided key insights into demographic patterns, product preferences, and purchase behaviors. Fig. 1: The gender gap in spending patterns was highlighted by the gender analysis, which showed that men made slightly larger average purchases than women. Figure 2 displays the top 10 products in terms of product preference, with the highest product accounting for 12.3% of all purchases.The main finding is that, in Figure 4, 75.3% of the purchasers were men, but, in Figure 3, the largest proportion of buyers were between the ages of 26 and 35 (39.9%), followed by those between the ages of 18 and 25 and 36 and 45. This implies that the consumer base is dominated by youthful and middle-aged people.

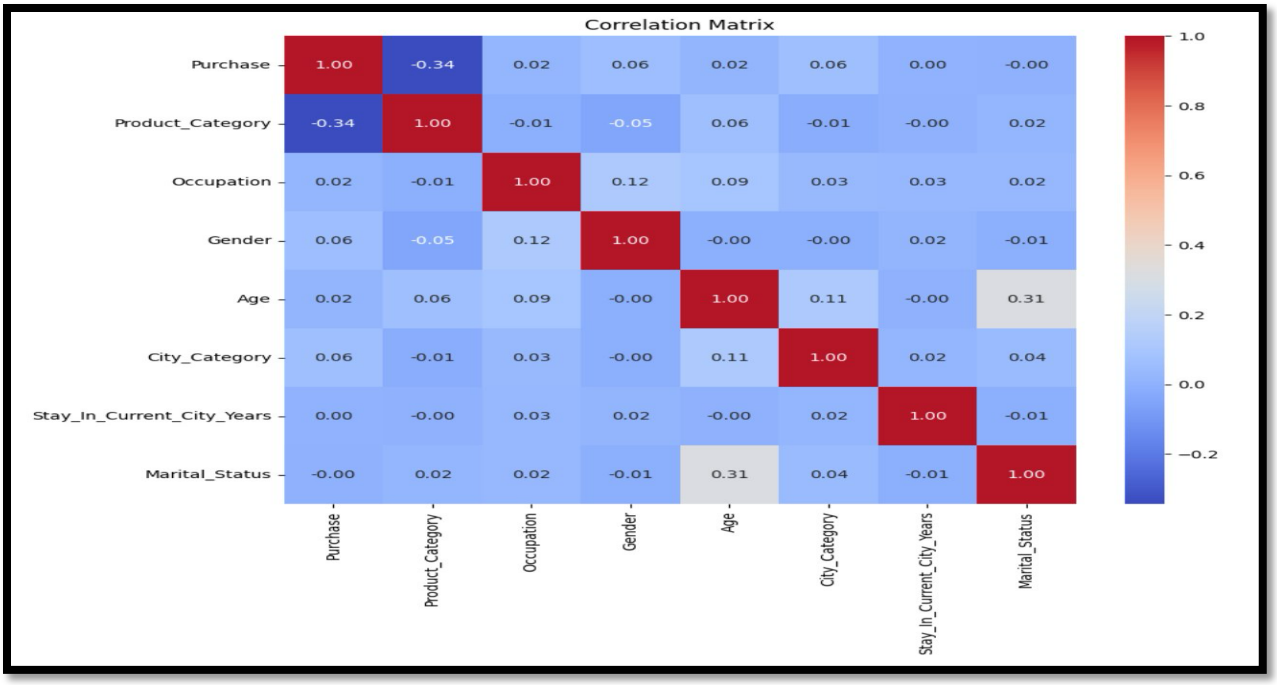
4.3 Feature-Wise Analysis of Purchase Trends

The feature-wise analysis revealed interesting patterns in purchase behavior.



Considering occupation in Fig 1, there is a marked difference in the amount purchased by each occupational category, indicating that the type of employment does have a bearing on the amount spent. In Fig 2, we find Marital status to have very little influence, which suggests it does not bear much impact on purchase behavior. Further, in Fig. 3, the purchase trends across product categories were very varied, with some categories having a wider range of purchase amounts, including notable outliers that reflect high-value or premium purchases. Interestingly, in Fig. 4, length of stay in the current city seemed to positively influence spending, wherein customers who lived longer in a city showed slightly higher purchase values, probably due to increased local familiarity. Fig 5: Gender analysis showed that males had a tendency to spend more than females on average.

4.4 Correlation Analysis



An understanding of how factors relate to the quantities of purchases made was made possible by the correlation matrix. Product_Category and purchase amount had the highest value, exhibiting a moderately negative correlation of -0.34. This would imply that spending is actually determined by the kind of product purchased. The weak positive correlations of ~0.06 between variables like gender, age, and city category indicate that those factors have little direct impact on consumer behavior. No multicollinearity or significant correlation between independent variables was present in the dataset, thus making it apt for predictive modeling. Generally, the product categories were the most important to predict purchase volumes.

5. Research Questions

5.1 What are the Key Factors Driving High Value Purchases Among Walmart Customers?

Step 1: Clustering for Customer Segmentation

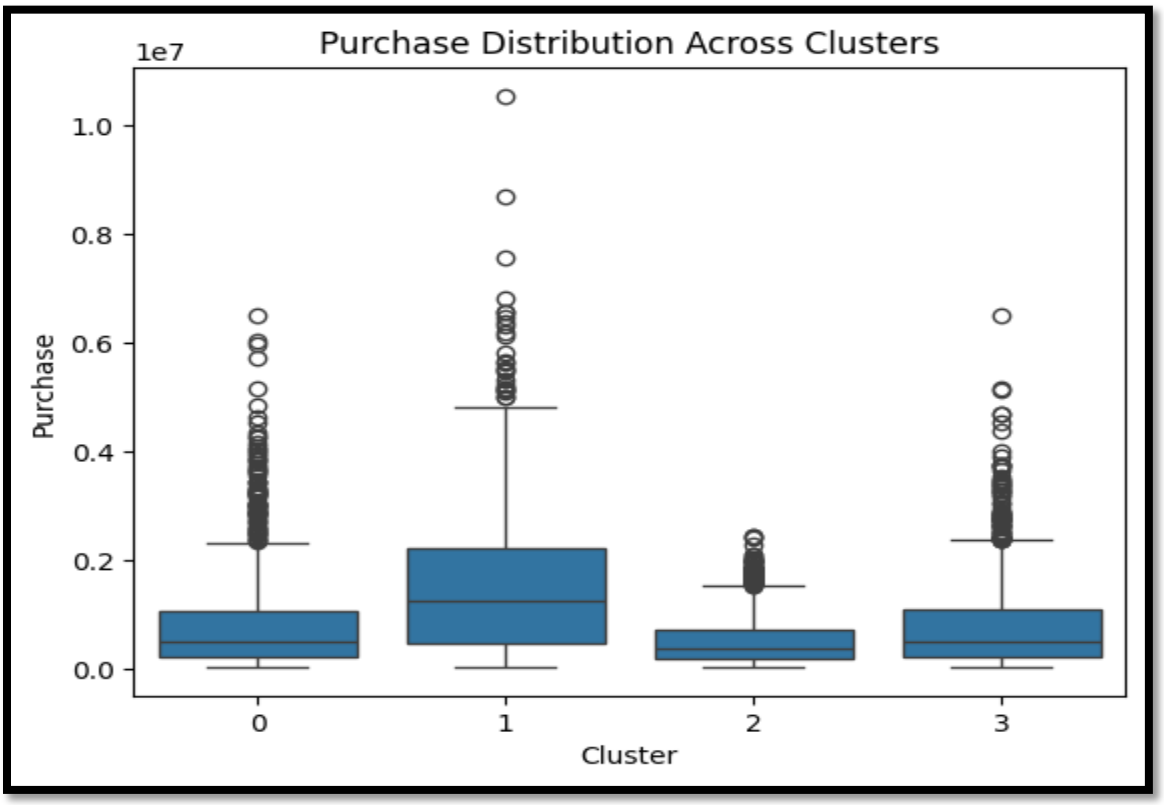
```python
import pandas as pd

# Assuming your dataset is named 'user_data'
summary_data = user_data.groupby('User_ID').agg({
    'Purchase': 'sum',  # Total spending per user
    'Age': 'first',  # Assuming age is constant per user
    'Occupation': 'first',  # Assuming occupation is constant per user
    'City_Category': 'first',  # Assuming city category is constant per user
    'Marital_Status': 'first',  # Assuming marital status is constant per user
    'Stay_In_Current_City_Years': 'first'  # Assuming this is constant per user
}).reset_index()

# Check the first few rows of the aggregated data
print(summary_data.head())

```

```python
from sklearn.cluster import KMeans
from sklearn.preprocessing import StandardScaler

# Prepare data for clustering
X = summary_data.drop('User_ID', axis=1)  # Drop 'User_ID' as it is not needed for clustering
scaler = StandardScaler()
X_scaled = scaler.fit_transform(X)

# Apply K-means clustering (you can adjust the number of clusters)
kmeans = KMeans(n_clusters=3, random_state=42)
summary_data['Cluster'] = kmeans.fit_predict(X_scaled)

# Check the assigned clusters
print(summary_data.head())

```

- **Algorithm**: K-Means Clustering
- **Number of Clusters**: 3 (selected based on domain understanding)
- **Preprocessing**:
    - Dropped the User_ID column as it is irrelevant for clustering.
    - Standardized features using StandardScaler to ensure uniform scaling.
- **Variables Used for Clustering**:
    - **Numerical**: Purchase (spending behavior), Age (categorized).
    - **Categorical**: Occupation, City_Category (one-hot encoded), Stay_In_Current_City_Years, Marital_Status.
- **Results**:
    - The clusters revealed distinct customer groups based on demographic and purchasing patterns.
    - A box plot visualized purchase distribution across the three clusters.

Step 2: Analysis and insights

The clustering analysis reveals distinct patterns in purchase behavior across the identified customer segments. The mean purchase amounts for each cluster are summarized as follows:
Cluster 0 8.645593e+05
Cluster 1 1.542957e+06
Cluster 2 5.200578e+05
Cluster 3 8.163649e+05



- **Cluster 1** probably has valuable clients who make This segment may benefit from much larger purchases and specialized tactics like loyalty programs or exclusive deals.
- Due to their somewhat diverse distributions and somewhat near average purchases, **Clusters 0 and 3** are considered middle-tier clients. Therefore, they will require separate strategies to try to elevate the spending by these groups.
- **Cluster 2** represents customers who have lower values because their spending is smaller. The effort to engage this group could focus on upselling, cross-selling, or encouraging purchases.
- The high-variability segmentation in customer purchase behavior allows for the creation of strategies tailored to each segment in order to maximize engagement and revenue.

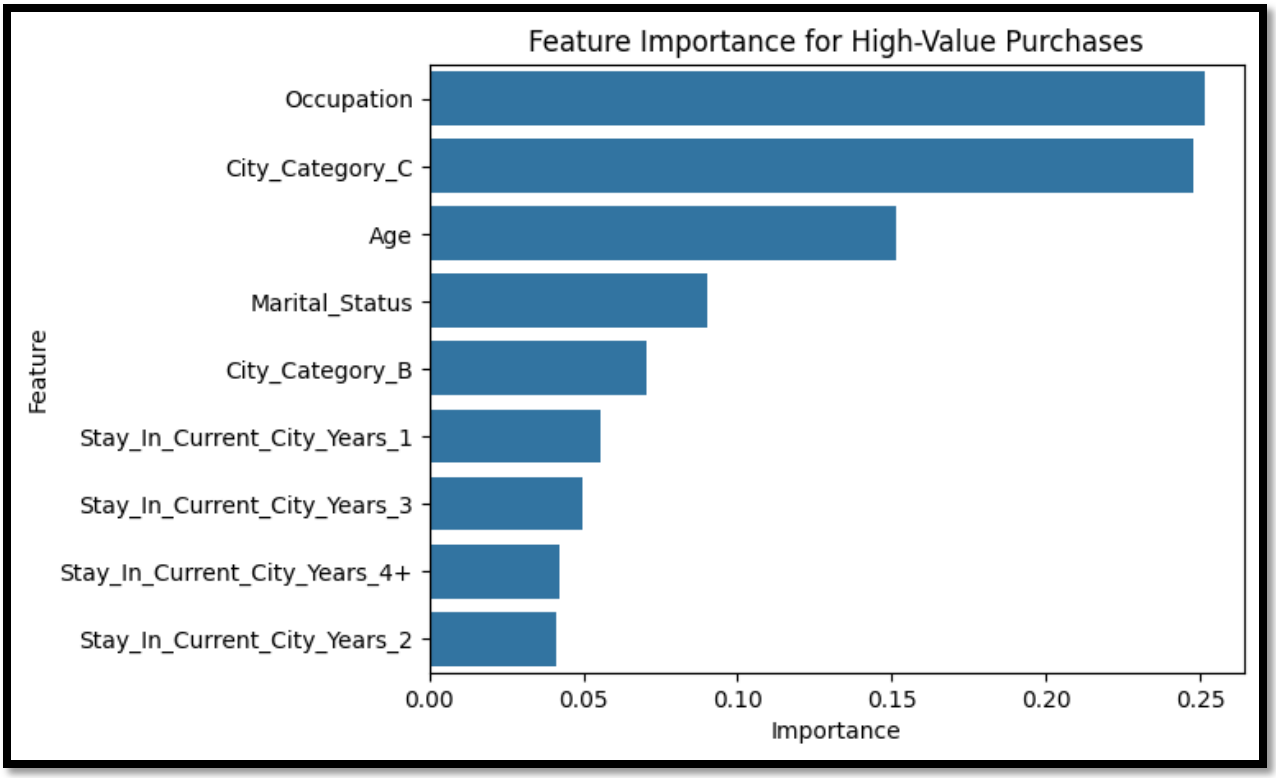
Prioritizing high-value customers (Cluster 1) while designing tailored approaches for mid- and low-value segments can yield significant business benefits.

Step 3: Identifying Key Factors with Random Forest

```python
from sklearn.ensemble import RandomForestRegressor

# Prepare data for Random Forest
X = summary_data.drop(['User_ID', 'Purchase', 'Cluster'], axis=1)
y = summary_data['Purchase']

# Fit a Random Forest model
rf = RandomForestRegressor(random_state=42)
rf.fit(X, y)

# Feature importance
importance = rf.feature_importances_
features = X.columns

# Create a DataFrame to visualize importance
feature_importance_df = pd.DataFrame({'Feature': features, 'Importance': importance})
feature_importance_df = feature_importance_df.sort_values(by='Importance', ascending=False)

# Plot the importance of each feature
sns.barplot(x='Importance', y='Feature', data=feature_importance_df)
plt.title('Feature Importance for High-Value Purchases')
plt.show()

```

- **Purpose**: To determine the importance of features influencing high-value purchases.
- **Model**: Random Forest Regressor
- **Data Preparation**:
  - Target Variable: Purchase.
  - Excluded columns: User_ID, Purchase, and Cluster.
  - Features considered: Age, Occupation, Marital_Status, City_Category, Stay_In_Current_City_Years (one-hot encoded).
- **Steps**:
  - Fit the Random Forest model with preprocessed data.
    Calculated feature importance scores and visualized them using a bar chart.

Step 4: Top Key Factors Influencing High-Value Purchases



1. Occupation

Occupation was the most relevant feature that determined high-value purchases with the highest feature importance score. This shows that a person's occupation is very important in determining their purchasing power. There could be several ways to develop campaigns for target revenue generation-based-on-occupational-groups.

2. City Category

City_Category_C had the second largest impact. It showed there were clear spend patterns between different geographic city categorizations. Customers of City Category C have highly strong associations in high value, while the category City_Category_B also influences fairly moderate. Since location-specific factors influence the consumer buying habits, they must be taken into consideration-when-framing-targeted-advertisements-as-well

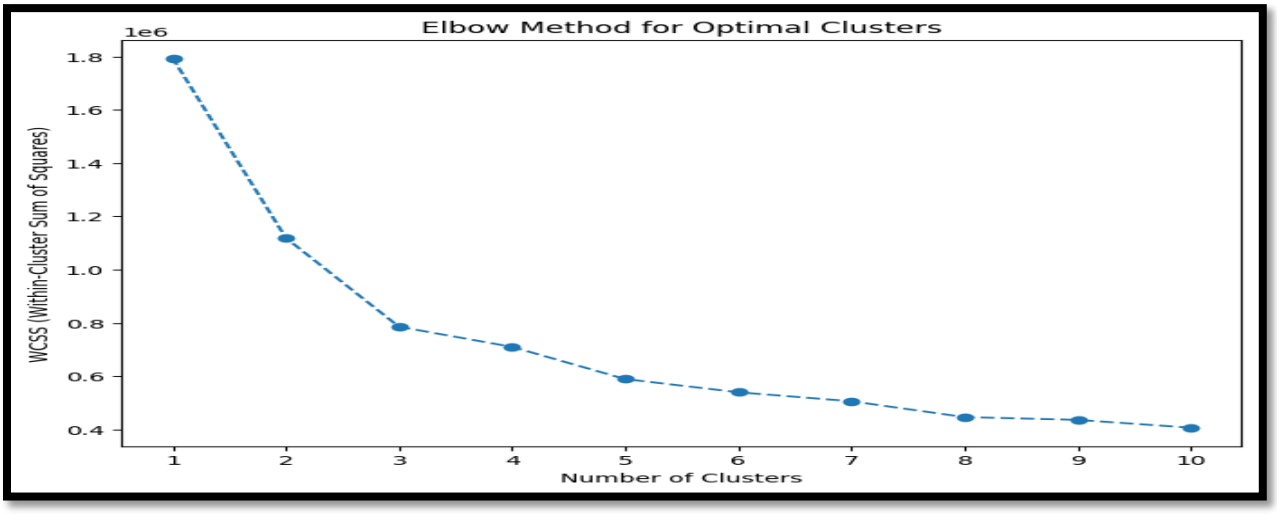
3. Age: Age emerged as the third important factor, which proves that age factor plays a very crucial role in influencing buying behavior. Since spending patterns vary with age, effective client engagement may only be ensured through targeting age-specific advertisements

4. Marital Status: The fact that marital status was given a moderate amount of weight suggests that a customer's marital status does affect their purchasing patterns.

5.2 Does customer demographics (age, gender, marital status) influence purchasing behaviour at Walmart?

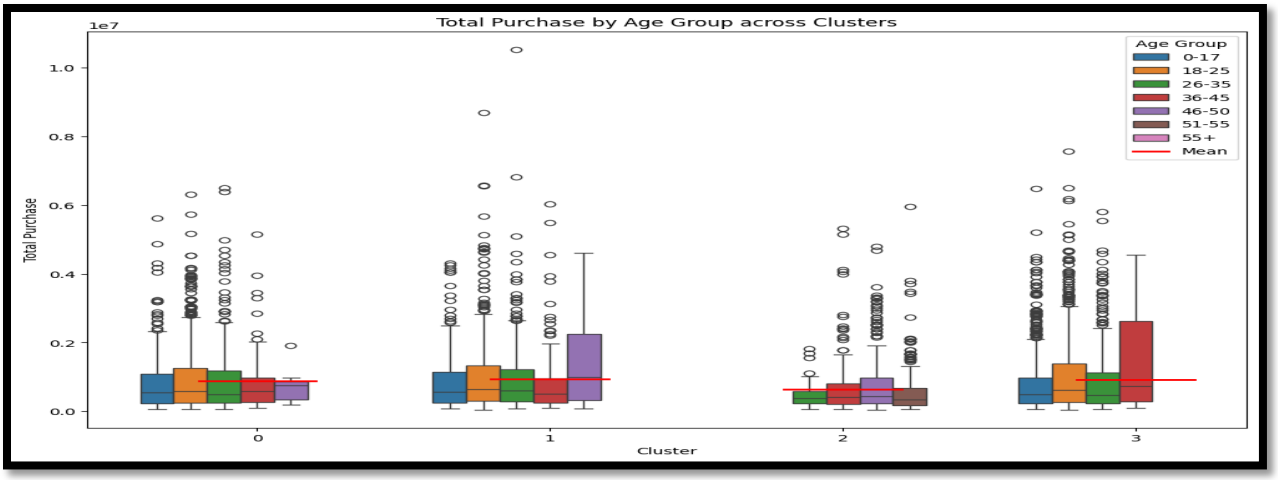
Step 1: Cluster Analysis

The key objectives of the analysis are to determine the critical customer demographics and behavioral patterns that drive high-value purchases and enable the creation of specific strategies to optimize customer engagement, thus maximizing revenue.



This analysis above was done for selecting the optimal number of clusters from a dataset using the Elbow method. The Elbow Method is essentially a well-known heuristic technique which measures the variation within cluster sum of squares against the number of clusters. We have chosen 4 as optimal number of clusters.

Total Purchase by Age Group across Clusters:



1. Distribution Patterns

- The box plots depict the distribution of total purchase amounts in various age groups across the clusters, ranging from 0 to 3.
- All clusters exhibit significant variability, with a large number of outliers across age groups, particularly in **Cluster 1** and **Cluster 3**.

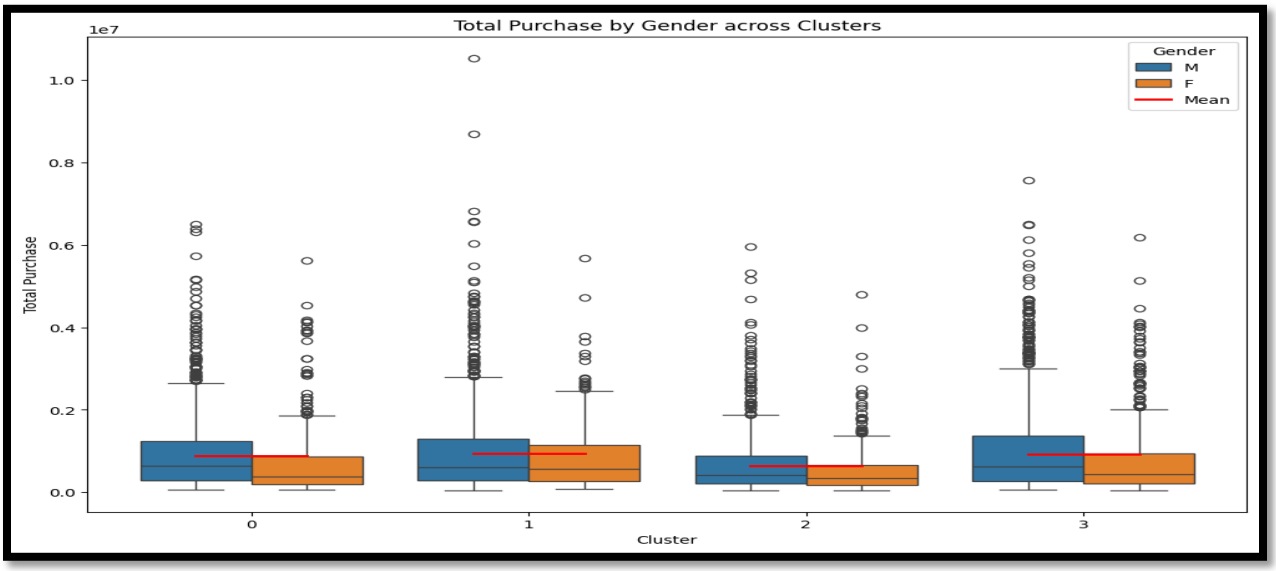
2. Cluster-Specific Trends

- **Cluster 0**: All age groups have relatively small total purchases, with slight differences across the groups.
- **Cluster 1**: Variability is more pronounced, with several outliers, especially in older age groups (46–55 and 55+).
- **Cluster 2**: This cluster has the least variability in purchase values among the age groups compared to other clusters.
- **Cluster 3**: Contains the highest purchase amounts, particularly for older age groups (46–50, 51–55, and 55+).

3. Mean Comparison

- The red line indicates the mean purchase value across clusters and age groups.
- The mean of total purchase value increases almost smoothly from **Cluster 0** through **Cluster 3**, reflecting that older age groups in **Cluster 3** contribute more to the total amount purchased.

Total Purchase by Gender across Clusters:



1. Distribution Patterns

- The total amount of purchases across genders shows many outliers in each cluster of variability in the purchasing behaviour.
- In most of the clusters, the male (blue) consistently has a little higher median and IQR compared to females (orange).

2. **Cluster-Specific Trends**

- **Cluster 0 and Cluster 1**: The median total purchases for males and females are nearly identical, with males showing a slightly higher median in both clusters.
- **Cluster 2 and Cluster 3**: A noticeable gap is observed, where males consistently exhibit higher purchase medians and a wider IQR compared to females, indicating greater variability in male purchasing behavior in these clusters.

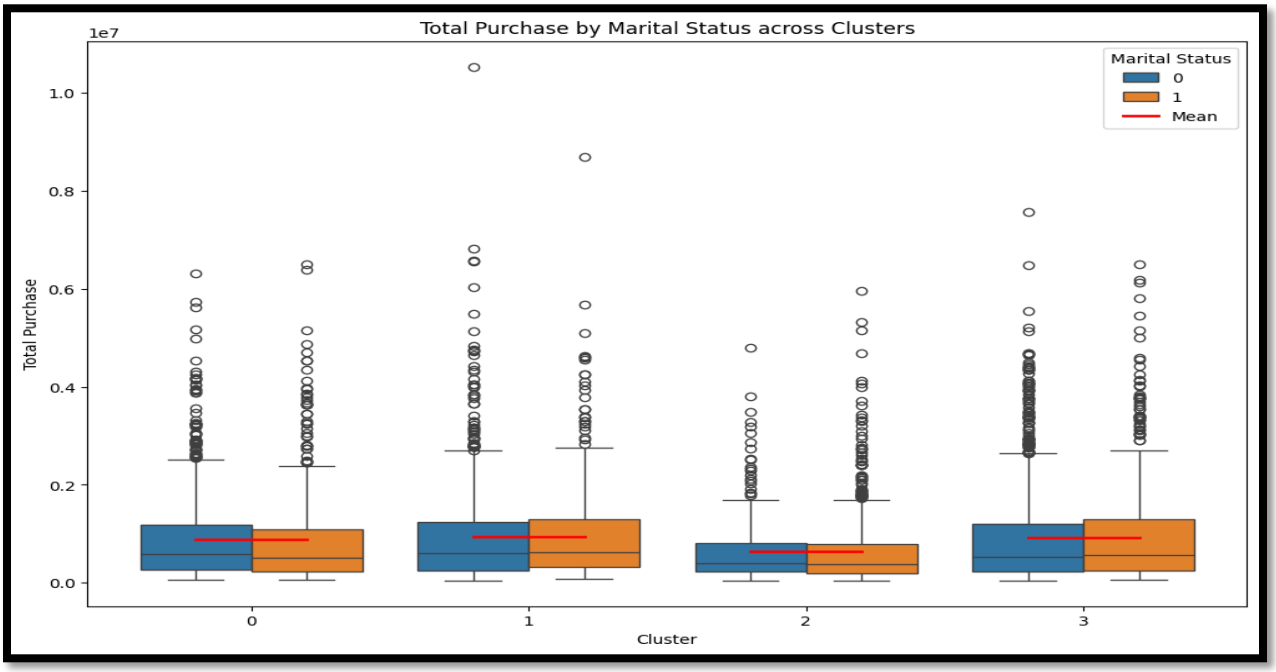
3. **Mean Comparison**

- The mean total purchase (indicated by the red line) is consistently higher for males across all clusters, aligning with the observed trend in medians and confirming a consistent gender-based disparity.

Conclusion

The analysis has shown that gender significantly impacts the total purchase patterns of males, who consistently spend more and have higher variability in purchasing behavior, especially in Cluster 2 and Cluster 3. These might suggest some opportunities for gender-specific strategies that could meet these differences in purchasing.

Total Purchase by Marital Status across Clusters:



1. Distribution Patterns

- The distribution of the total purchases across the clusters contains a lot of outliers, hence showing variability within the data.
- The IQR remains constant for all clusters; only the marital status - 0 and 1 - had a moderate difference in dispersion.

2. Cluster-Specific Trends

- **Cluster 0 and Cluster 1:** The medians of total purchases are nearly identical for both marital statuses, showing minimal variation between the groups.
- **Cluster 2:** A slight divergence is noticeable, where individuals with marital status 0 (unmarried) exhibit a marginally higher median total purchase than those with marital status 1(married).

- **Cluster 3:** This cluster shows **increased variability** in total purchases for both marital statuses, alongside **slightly higher median values** compared to other clusters.

3. Mean Comparison

- The **mean total purchase** (represented by the red line) is consistently close to the median across all clusters, indicating **minimal skewness** in the purchase data distribution.

Conclusion

The analysis shows that all distributions of total purchases are well-consistent among clusters, with the most variability in Cluster 3, and slight differences between marital statuses at Cluster 2 and Cluster-3.

These findings should, therefore, be tacked onto potential areas requiring a deeper dig into the behavioral determinant for such variations.

Step 2: Other Analysis and Findings

ANOVA Test results:

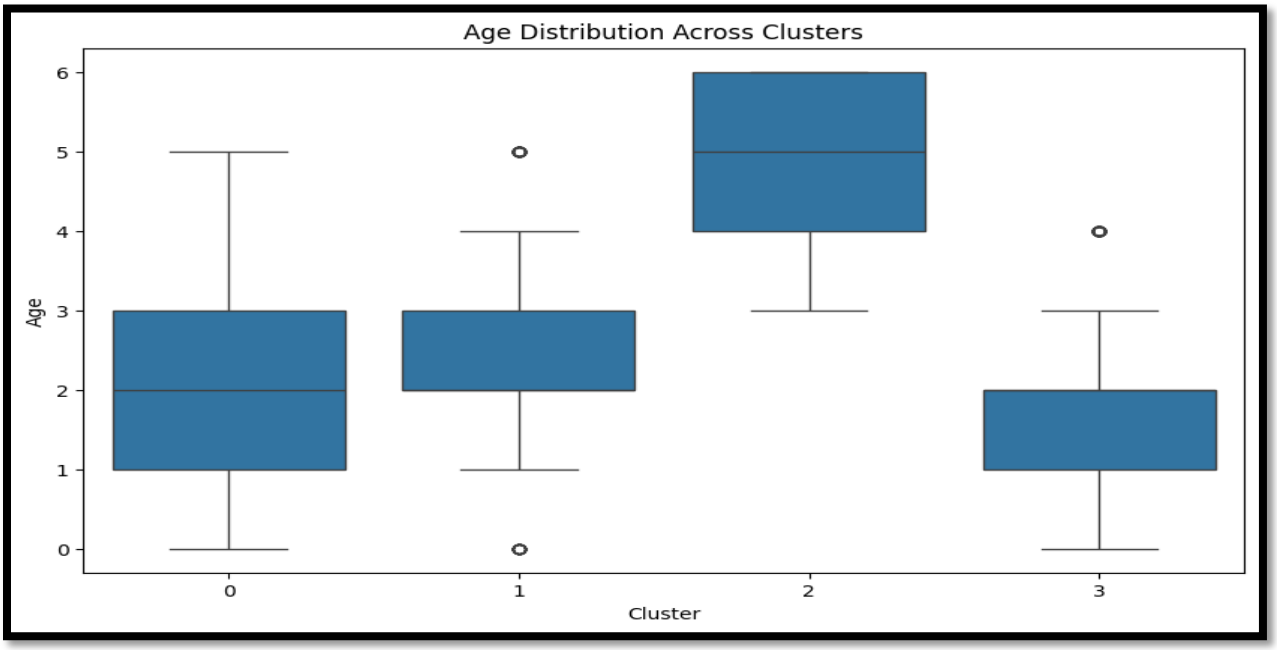
```
ANOVA Results for Age Group:
F-statistic: nan, p-value: nan

ANOVA Results for Gender:
F-statistic: 21.514637958662913, p-value: 7.358014490757776e-29

ANOVA Results for Marital Status:
F-statistic: 12.800997363667685, p-value: 2.0140744359547477e-16
```

These ANOVA results indicate that gender and marital status are significant, age group could be determined if there is sufficient data, on the totals of purchase. These will serve as a guide for further investigation and target strategies to leverage the demographic insights.

Step 3: Cluster Results and Other Visualizations

Cluster	Age	Gender	Marital_Status
0	2	M	0
1	2	M	0
2	5	M	1
3	2	M	0

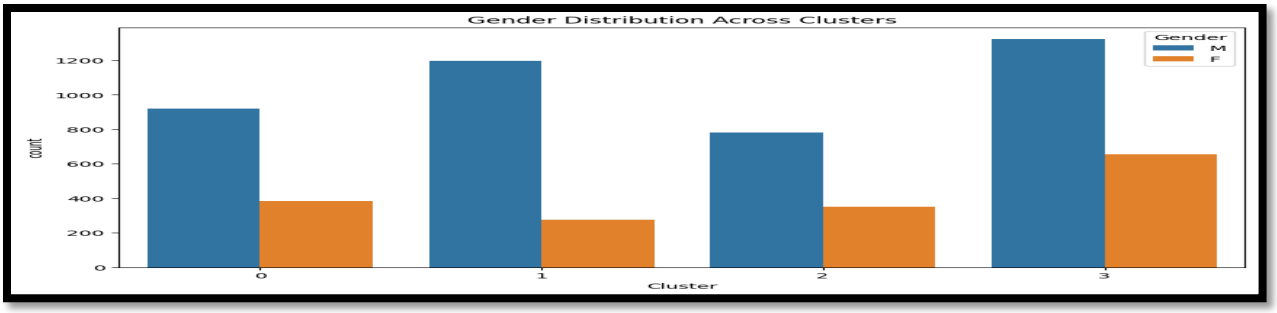
- **Cluster 0**: Includes individuals with varying Age values, where both male (M) and female (F) genders are present, along with both married (M) and non-married statuses.
- **Cluster 1**: Shows a similar distribution of Age, Gender, and Marital Status, but specific patterns might differ in terms of dominance of certain demographics.
- **Cluster 2**: Contains fewer combinations of demographic attributes compared to Clusters 0 and 1, indicating more homogeneity in the group.
- **Cluster 3**: Has the highest diversity or variability in demographic attributes, particularly in terms of the Age and Gender categories.



Age Distribution Across Clusters (Box Plot):

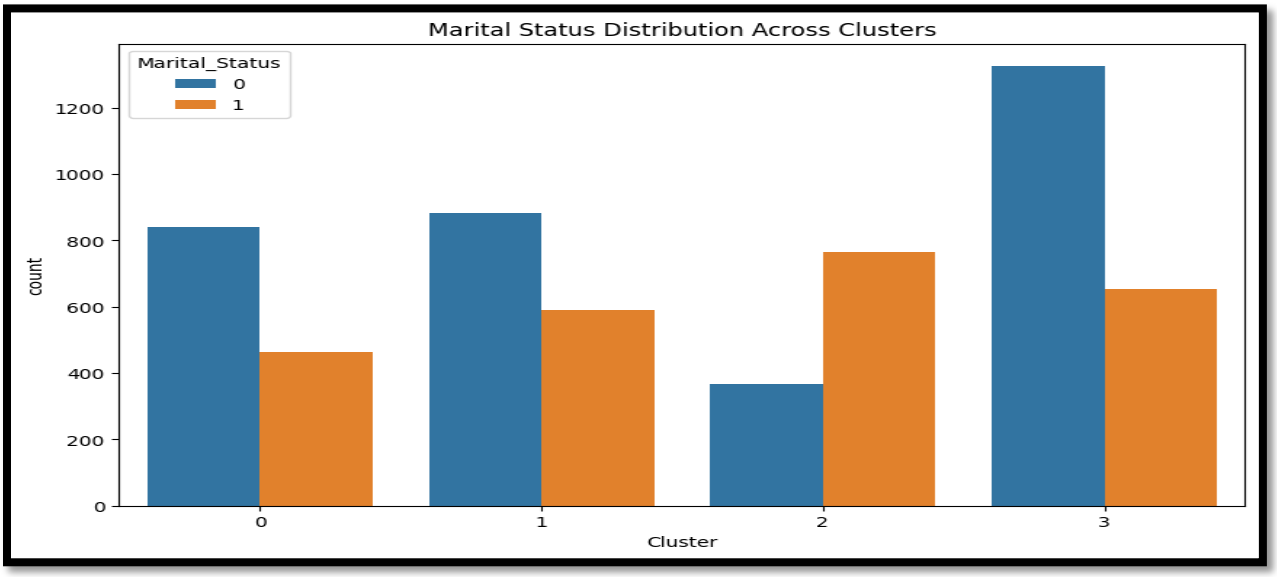
- **Cluster 0**: Contains a relatively smaller range of age values. Shows some outliers, indicating individuals outside the typical age range of the group.
- **Cluster 1**: Exhibits a wider range of age values with some outliers.
- **Cluster 2**: Has a narrow range of age values with minimal variability, suggesting greater homogeneity within this cluster.
- **Cluster 3**: Displays the largest age range among all clusters, with notable variability and a few outliers.

Gender Distribution Across Clusters (Box Plot):



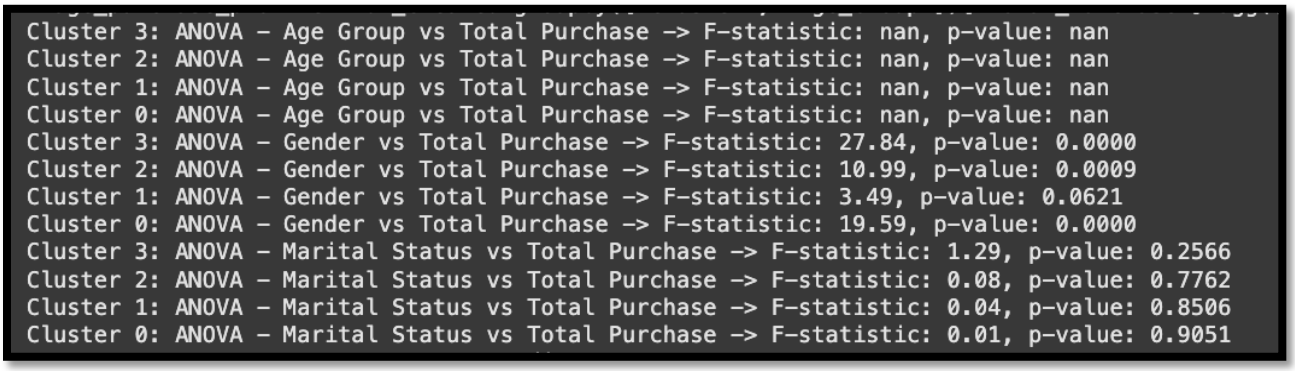
- **Cluster 0**: Male representation is significantly higher than female representation.
- **Cluster 1**: Similar to Cluster 0, males dominate, but females have a noticeable presence.
- **Cluster 2**: Male representation continues to dominate with limited female presence.
- **Cluster 3**: Shows the highest counts for both genders, with males having a larger share compared to females.

Marital Status Distribution Across Clusters (Box Plot):



- **Clusters 0 and 1** are dominated by unmarried individuals.
- **Cluster 2** is dominated by married individuals.
- **Cluster 3** has the highest total counts for both groups, with unmarried individuals leading.

ANOVA Test results:



With highly significant p-values highlighting its dependability as a crucial demographic element, the ANOVA analysis showed a number of encouraging trends, most notably the strong impact of gender on total purchases across the majority of clusters (0, 2, and 3). Actionable insights for customized tactics are highlighted by cluster-specific observations, such as the high influence of gender in Cluster 3 (F-statistic: 27.84, $p < 0.0000$). While the nonsignificant results for marital status help in focusing on other important variables, the constancy of the effect of gender across clusters underlines its importance. Moreover, the granular cluster-level analysis of ANOVA guarantees accurate and focused insights into demographic trends in consumer behavior.

```
                              mean    median
Cluster Age_Group
0       0-17        7.869046e+05  544711.0
        18-25       9.710532e+05  589412.5
        26-35       8.852094e+05  499883.0
        36-45       8.281331e+05  590083.5
        46-50       7.242910e+05  744021.0
        51-55                NaN       NaN
        55+                  NaN       NaN
1       0-17        8.753327e+05  560903.5
        18-25       9.747456e+05  631980.0
        26-35       9.281572e+05  603718.0
        36-45       8.974685e+05  514305.5
        46-50       1.402563e+06  990099.0
        51-55                NaN       NaN
        55+                  NaN       NaN
2       0-17                 NaN       NaN
        18-25                NaN       NaN
        26-35       4.973114e+05  375272.0
        36-45       6.340302e+05  408401.0
        46-50       7.251992e+05  436502.0
        51-55       5.396972e+05  332731.5
        55+                  NaN       NaN
3       0-17        8.078445e+05  484953.0
        18-25       1.014443e+06  619392.0
        26-35       8.681864e+05  479709.5
        36-45       1.518746e+06  728690.0
        46-50                NaN       NaN
        51-55                NaN       NaN
        55+                  NaN       NaN
                           mean    median
Cluster Gender
0       F        7.108348e+05  372967.5
        M        9.649510e+05  645579.0
1       F        8.458806e+05  575162.0
        M        9.701168e+05  612813.5
2       F        5.308046e+05  337687.0
        M        6.794855e+05  421288.5
3       F        7.530906e+05  426707.5
        M        1.002495e+06  613486.5
                                    mean    median
Cluster Marital_Status
0       0                 892515.790476  577696.5
        1                 885950.810753  518384.0
1       0                 942853.307692  600566.0
        1                 952848.461017  615422.0
2       0                 624948.385246  406730.5
        1                 637608.588773  384625.5
3       0                 902035.229842  536818.0
        1                 956094.022971  562743.0
```

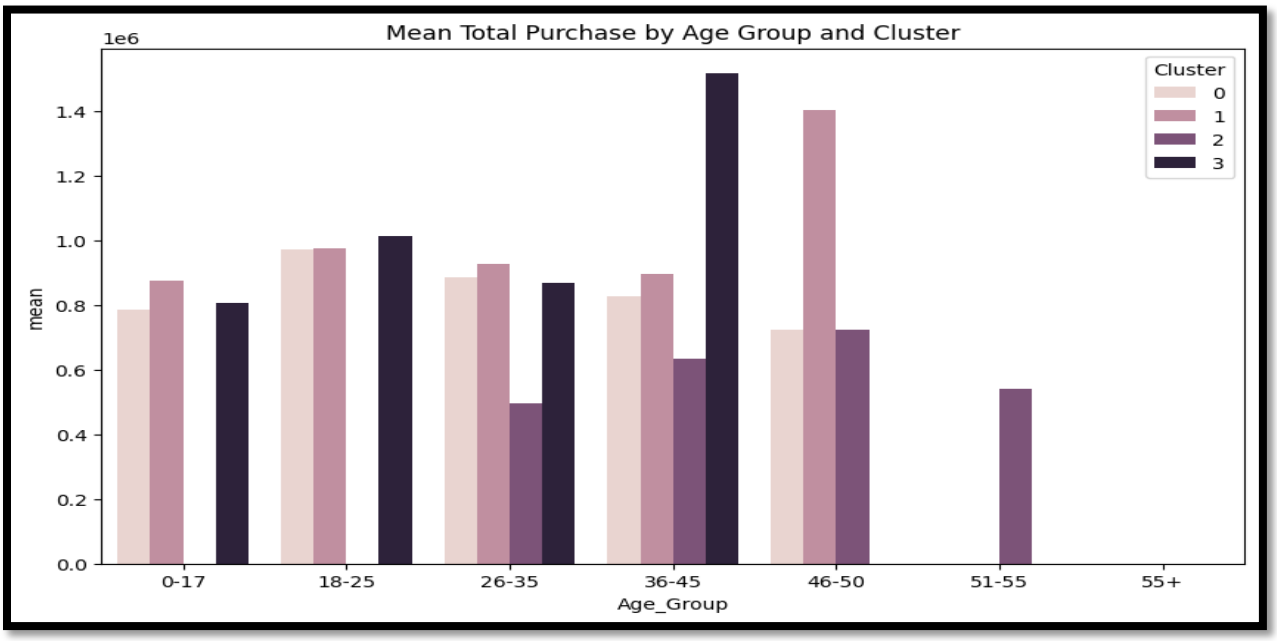
The analysis of data shows interesting trends in the total purchase by clusters based on age, gender, and marital status. From the age group, in general, the younger segments 18–25 and 26–35 have higher mean and median purchase values in Cluster 0, 1, and 3, while older age groups 46–50 have values spiking at Cluster-1. Gender analysis shows that across all clusters, males are outperforming females in terms of mean and median purchase values.

The gap is greatest in Cluster 3, where the male mean purchases have exceeded one million. The analysis reveals that marital status, particularly in the "married" category, shows relatively higher mean and median purchase values in Clusters 1 and 3, indicating a minor influence.

Overall, these findings emphasize that gender and certain age groups significantly impact purchasing trends, while marital status plays a secondary role.

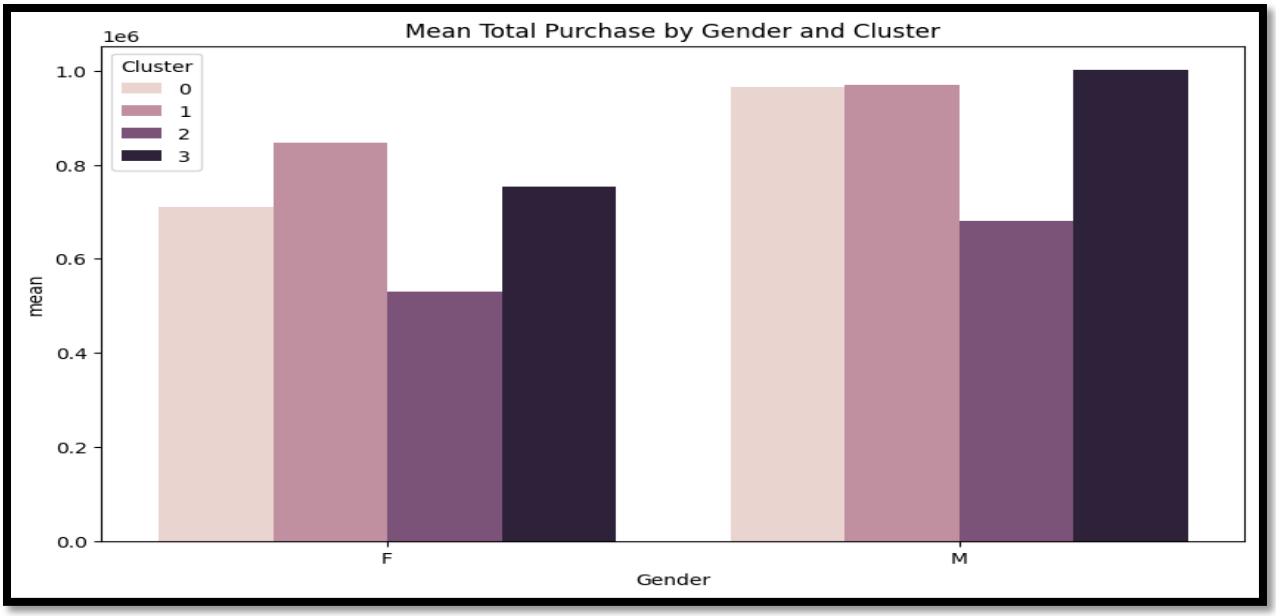
Step 4: Bar plot showing the mean values:

Mean Total Purchase by Age Group and Cluster:



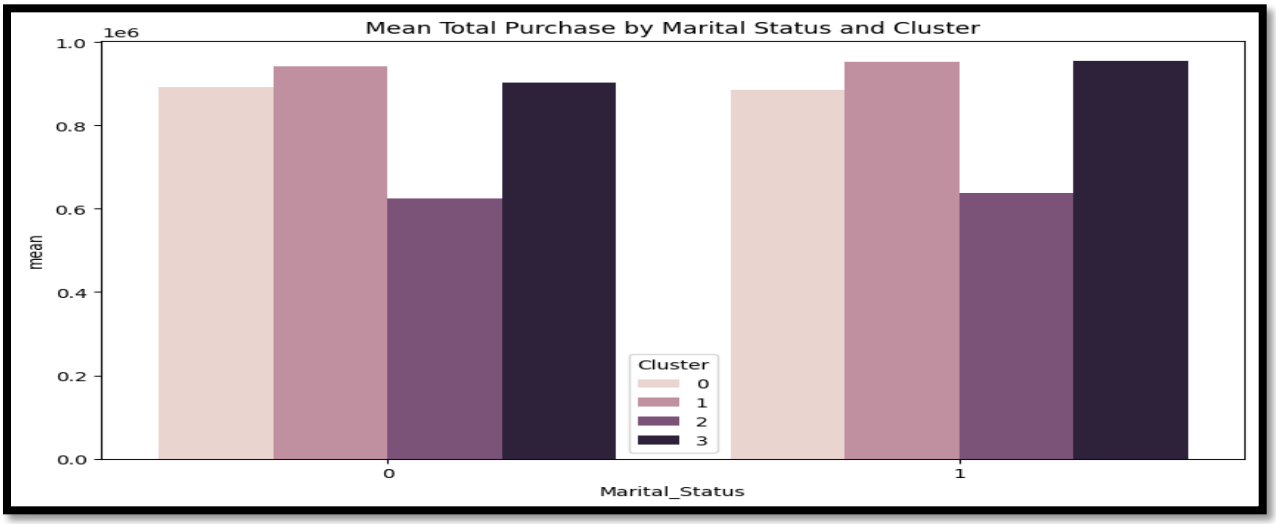
The chart also indicates that the age group 36–45 has the highest mean purchase value in one of the clusters, which might represent the most active or high-spending demographic. Other age groups, such as 18–25 and 26–35, have shown moderate purchase levels, while other age groups like 51–55 and 55+ show much lower mean purchase values across all clusters. This visualization helps show the variations in spending behaviors among different age groups in each cluster.

Mean Total Purchase by Gender and Cluster:



The chart shows that, for most clusters, the mean purchase value for males is always higher than that of females. Specifically, the purchasing values for males in Cluster 3 are the highest among all gender-cluster combinations. Females have a rather consistent pattern across clusters, with a slightly lower mean purchase compared to their male counterparts. This chart highlights the differences in spending patterns by gender within distinct customer segments.

Mean Total Purchase by Marital Status and Cluster



This chart indicates that in most clusters, marital status 1 (married) usually has a slightly higher average purchase value compared to status 0 (single). In Cluster 3, the purchasing values are highest for both categories of marital status, although the difference is in favor of married with a fair margin. This chart highlights how marital status can influence purchasing behavior across different customer clusters. This chart emphasizes how marital status can influence purchasing behavior within different customer clusters.

5.3 How Does Walmart's Product Demand Vary Across City Categories (A, B, C)?

The analysis is divided into two parts:

1. **Descriptive Analysis**: Exploration of demand trends and identification of feature importance through the application of statistical and machine learning models
2. **Clustering Analysis**: Explores more deeply the patterns in proportional spending, demographic contributions, and city-specific product reliance.

Method 1: Descriptive Analysis:

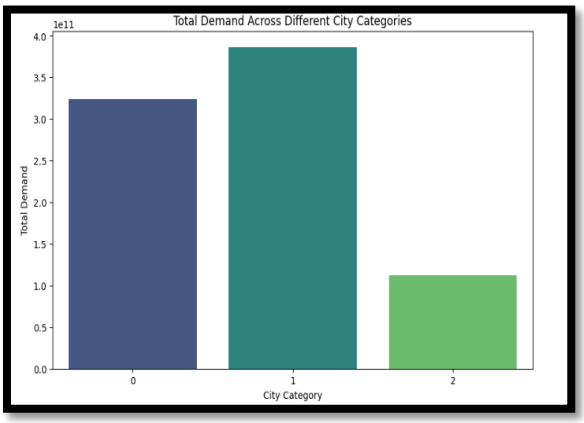
Methodology:

1. Aggregated spending metrics by city and product categories.
2. Conducted feature importance analysis to identify drivers of purchasing behavior.
3. Performed statistical analyses, including ANOVA and Chi-Square tests, to assess variations in spending patterns and product preferences across different cities.

Findings:

- **Total Spending**:
  - City B has the highest total spending across most product categories.
  - City A leads, with City C trailing marginally in overall demand.
- **Key Product Categories**:
  - Product Category 1 is the most demanded across all cities (35–40% of total spending).
  - Categories 5 and 8 always come in second and third in demand, especially in Cities A and B.
- **Demographic Insights**:
  - Customers in the 18–35 age group constitute the bulk of demand, particularly in Product Category 1.
  - Older age groups (46+) contribute much less to total spending.
- **Feature Importance**:
  - In the purchase behavior, the most important feature is Product Category, followed by Occupation and City Category.
- **Statistical Tests**:
  - ANOVA shows no significant difference in total spending across cities (p-value > 0.05).

- Chi-Square confirms that the difference in product popularity between cities is significant (p-value=0.0).




```
ANOVA Results for Product Demand Across City Categories:
Product Category: 3, F-Statistic: 19.28, p-value: 0.0000
Product Category: 1, F-Statistic: 854.19, p-value: 0.0000
Product Category: 12, F-Statistic: 12.85, p-value: 0.0000
Product Category: 8, F-Statistic: 337.79, p-value: 0.0000
Product Category: 5, F-Statistic: 426.17, p-value: 0.0000
Product Category: 4, F-Statistic: 36.79, p-value: 0.0000
Product Category: 2, F-Statistic: 96.98, p-value: 0.0000
Product Category: 6, F-Statistic: 49.86, p-value: 0.0000
Product Category: 14, F-Statistic: 13.48, p-value: 0.0000
Product Category: 11, F-Statistic: 64.72, p-value: 0.0000
Product Category: 13, F-Statistic: 32.62, p-value: 0.0000
Product Category: 15, F-Statistic: 7.58, p-value: 0.0005
Product Category: 7, F-Statistic: 1.52, p-value: 0.2199
Product Category: 16, F-Statistic: 65.49, p-value: 0.0000
Product Category: 18, F-Statistic: 2.92, p-value: 0.0542
Product Category: 10, F-Statistic: 2.71, p-value: 0.0664
Product Category: 17, F-Statistic: 3.03, p-value: 0.0492
Product Category: 9, F-Statistic: 1.45, p-value: 0.2364
Product Category: 20, F-Statistic: 2.60, p-value: 0.0743
Product Category: 19, F-Statistic: 1.15, p-value: 0.3176
```

Insights:

1. City B has the highest purchasing power and consumer demand.
2. Product Category 1 significantly influences expenditure trends throughout various urban areas.
3. Younger demographics (18–35) are critical drivers of Walmart's sales across all categories.

Method 2: Clustering Analysis

Objective: Use clustering to identify deeper patterns of demand and spending across cities.

Methodology:

1. Quantified demand indicators include overall expenditure, relative expenditure, and frequency of purchases.
2. Analyzed proportional reliance on product categories across cities.
3. Visualized proportional spending patterns and demographic breakdowns.

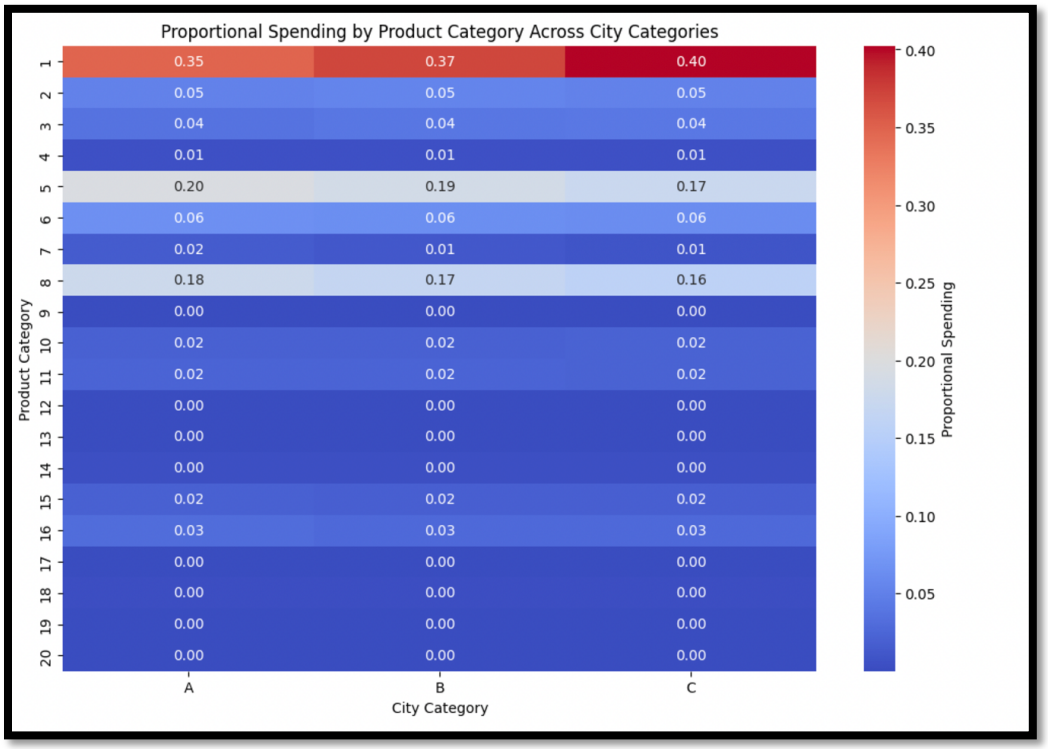
Findings:

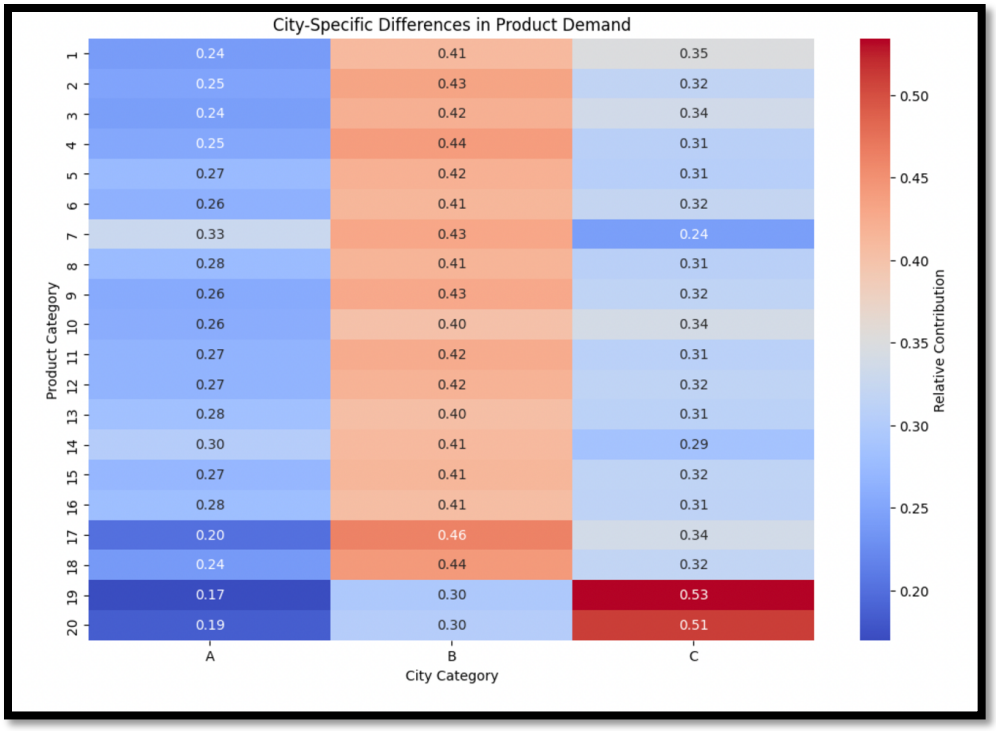
- **Total Spending by City**:
  - City B is the highest spender, especially in Categories 1, 5, and 8.
  - City A spends fairly in all departments.
  - City C relies most heavily on Product Category 1 (40% of total spending).
- **Proportional Spending**:
  - Categories 1, 5, and 8 dominate spending across all cities, with slight variations:
    - City C has the highest dependence on Category 1.
    - Cities A and B are more spread out

- **Demographic Breakdown**:
    - Customers aged 26–35 old rule spending across all cities, especially in Category1
    - Spending by older age groups (46+) is minimal, indicating a weaker influence.
- **Statistical Analysis**:
    - ANOVA: No significant differences in total spending across cities (p-value > 0.05).
    - Chi-Square: Significant differences in product popularity (p-value = 0.0).

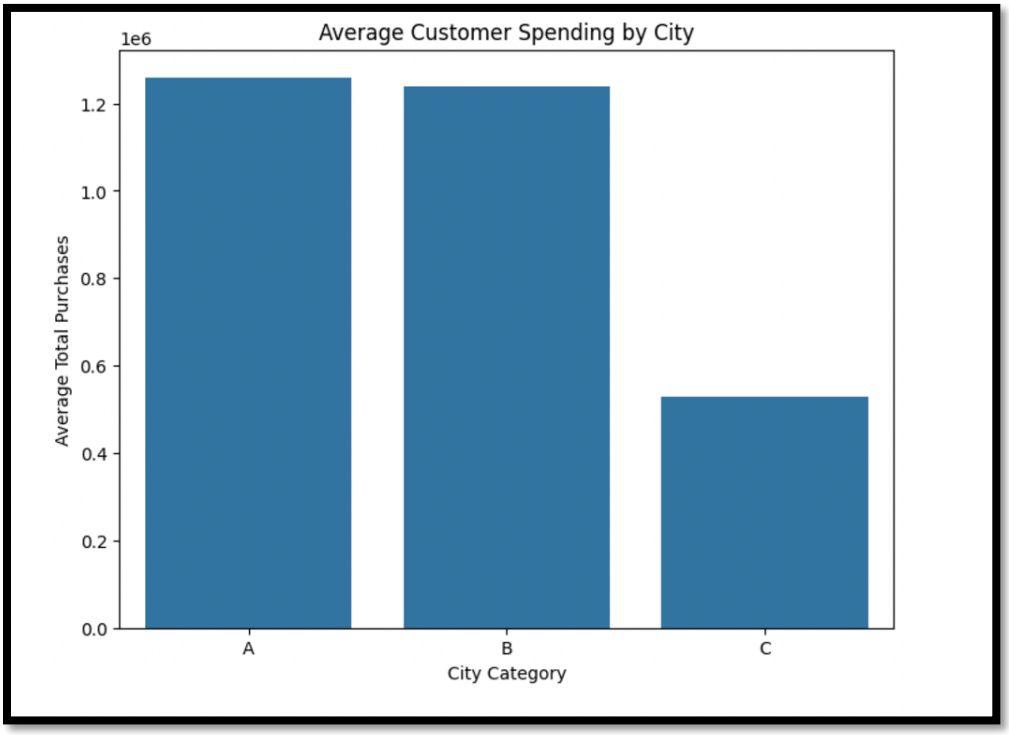
Visual Insights:

1. The heatmap of proportional spending shows dominance of Categories 1, 5, 8

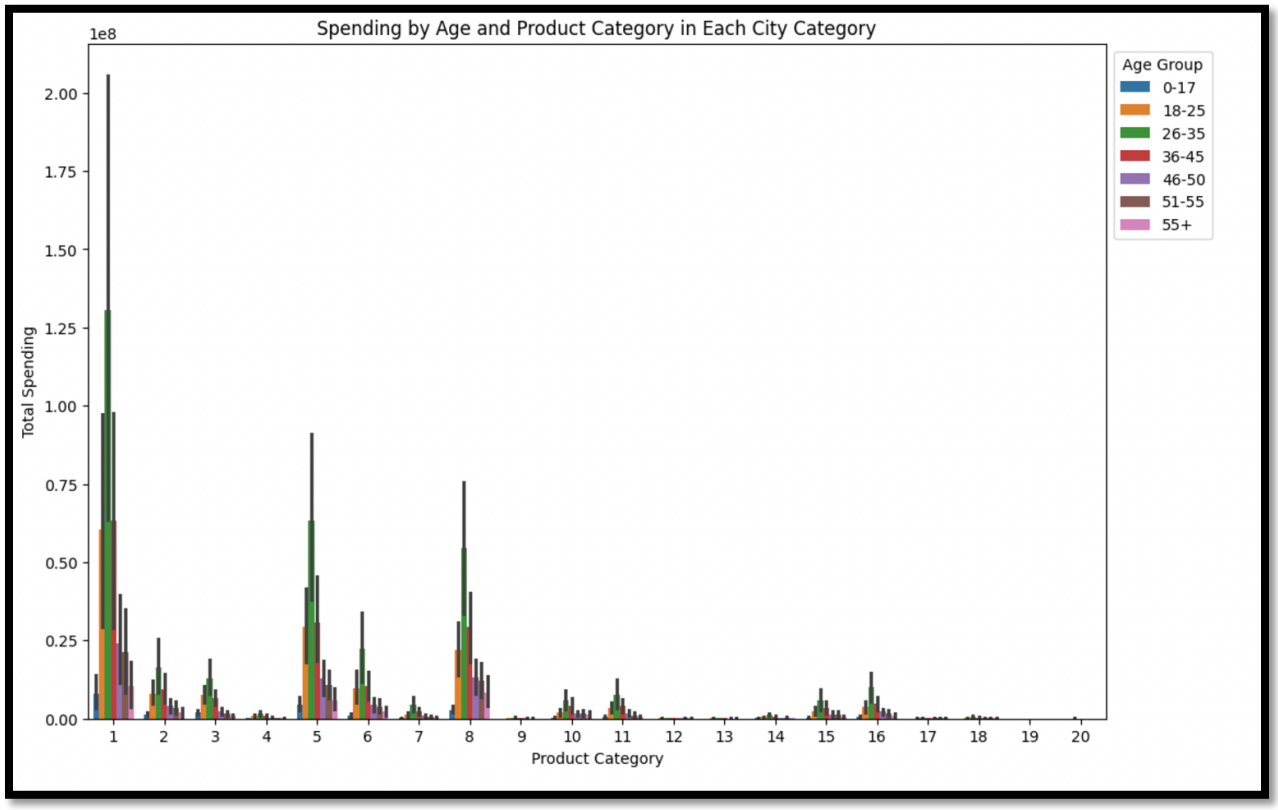




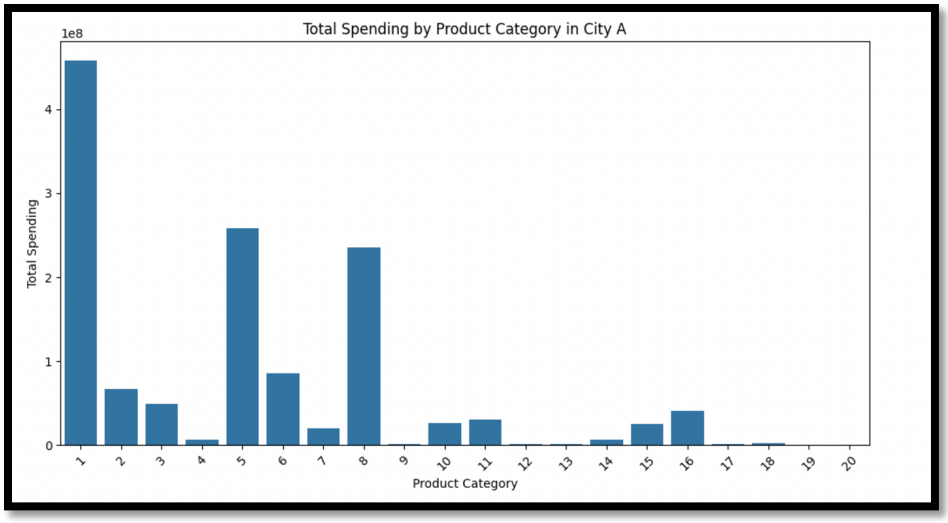
2. Bar plots reveal City B's dominance in total spending, followed by A and C.

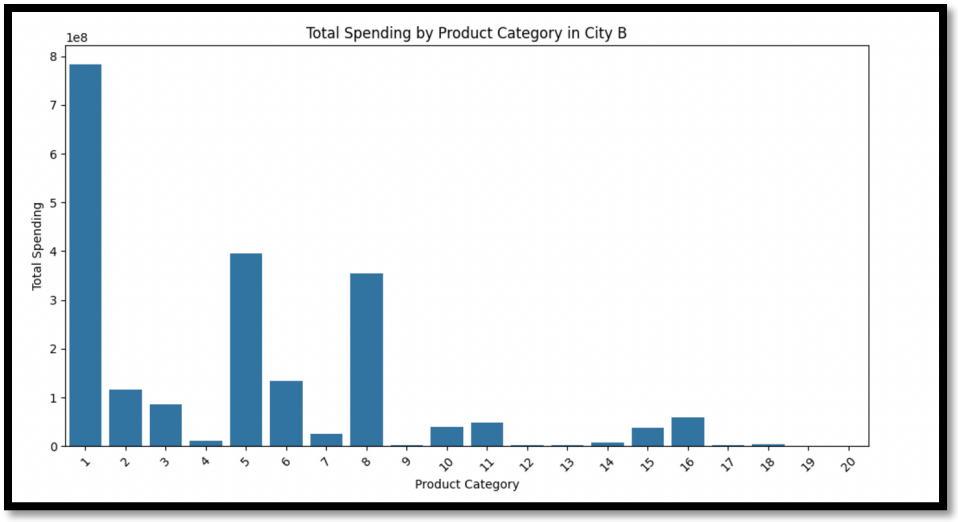


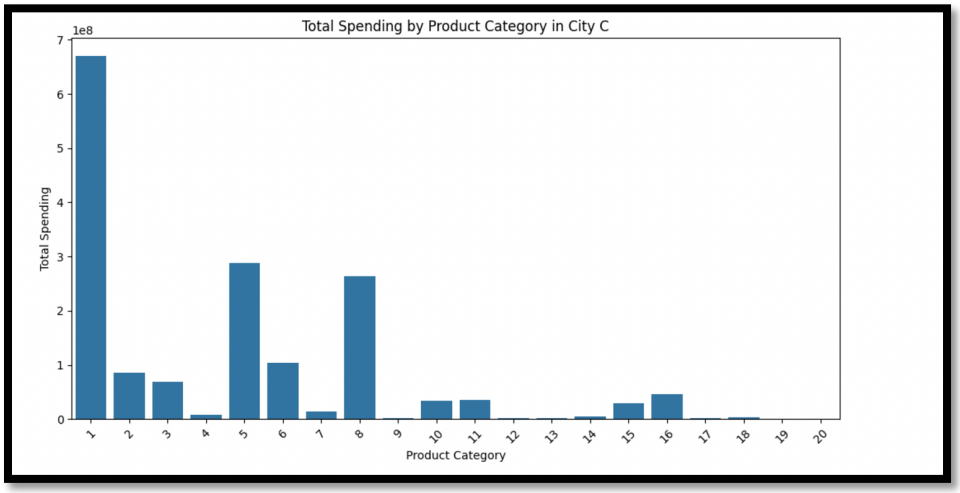
3. Age-based bar plots confirm the importance of younger demographics in driving demand.



4. Total Spending by Product Category: The bar graphs showing the total spending in each city show the following: • Spending in City A is lower than in City B but higher than in City C. • City B is the highest spending city, notably in Product Categories 1, 5, and 8. • City C shows a similar trend in spending to City A, though it reports slightly Reduced-overallspending-across-most-categories







Insights:

1. City B's higher total spending and broader product demand suggest it is Walmart's strongest market.
2. City C shows a high concentration in Category 1 and therefore needs focused approaches.
3. Younger customers (18–35) are pivotal for all cities, requiring age-targeted marketing.

4. Recommendations

1. **City-Specific Strategies**:
    a. Focus on City B as a key growth market.
    b. Give promotional efforts in City A and City C based on product preferences.
2. **Product Category Optimization**:
    a. Prioritize inventory for Categories 1, 5, and 8 across all cities.
    b. Address City C's higher dependence on Category 1 with targeted marketing.

3. **Demographic Targeting**:
    a. Develop age-specific campaigns focusing on younger customers (18–35).
    b. Leverage digital platforms to engage this demographic effectively.
4. **Data-Driven Insights**:
    a. Use clustering insights to refine city-specific inventory and marketing decisions.
    b. Incorporate demand metrics into pricing and promotion strategies.

5. Conclusion

Walmart's demand for products shows up with consistent patterns across city categories, with significant demographic and city-specific variations. City B looks like the strongest market, while City C represents a concentrated reliance on specific categories. Younger customers (18–35) are the primary drivers of demand across all cities; this would call for tailored marketing and inventory strategies.

5.4 Which product categories are most likely to see an increase in demand based on current trends?

Method 1: Spending-Based Analysis

Approach:

The data was analyzed to identify product categories with the highest total spending by aggregating the total_spent_category for each product category.

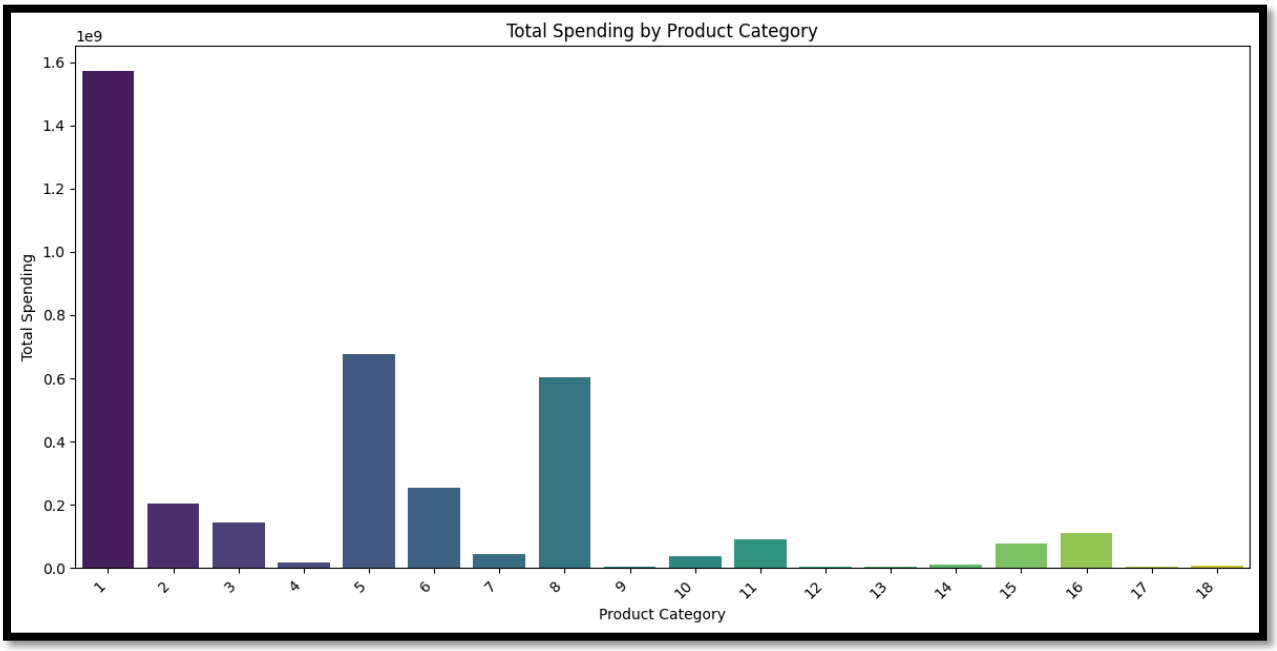
```
# Load the dataset
data = pd.read_csv('/content/drive/MyDrive/Web&Social/cleaned_walmart.csv')
# Group by User_ID and calculate the sum of spending, count of product categories, and unique product categories
user_data = data.groupby('User_ID').agg(
    total_spent=('Purchase', 'sum'),  # Sum of spending (assuming 'purchase' is the amount spent)
    num_products=('Product_Category', 'count'),  # Count of total purchases
    unique_categories=('Product_Category', 'nunique')  # Count of unique product categories
).reset_index()

# Optionally, you can add recency, frequency, and monetary (RFM) features if you have date data:
# user_data['recency'] = (current_date - user_data['last_purchase_date']).dt.days
# user_data['frequency'] = user_data['purchase_count']
# user_data['monetary'] = user_data['total_spent']

# Merge product category data back with user data for further analysis
category_data = data.groupby(['User_ID', 'Product_Category']).agg(
    total_spent_category=('Purchase', 'sum'),
    num_purchases_category=('Purchase', 'count')
).reset_index()

# Merge this back with the user data
user_data = pd.merge(user_data, category_data.groupby('User_ID').agg(
    category_trend=('Product_Category', 'count')
).reset_index(), on='User_ID', how='left')

# Check the resulting aggregated user data
print(user_data.head())
```

```
   User_ID  total_spent  num_products  unique_categories  category_trend
0  1000002       810353            76                  5               5
1  1000003       341635            29                  6               6
2  1000004       205987            13                  1               1
3  1000005       821001           106                 12              12
4  1000007       234427            16                  3               3
```

The aggregated data was then sorted in descending order to rank product categories based on their contribution to total spending.

```
Top Categories by Total Spending:
   Product_Category  total_spent_category
0                 1            1572382609
4                 5             677177151
7                 8             602636323
5                 6             253046186
1                 2             203972569
```

A bar plot was generated to visualize the spending distribution across all product categories, making it easier to identify the top-performing categories.



Key Insights:

- **Top Product Categories:**
  - Category 1 leads in total spending, exceeding 1.57 billion, significantly outpacing all other categories.
  - Category 5 ranks second with total spending nearing 677 million, followed by Category 8 at 602 million.
  - Together, these three categories account for the majority of spending, highlighting their popularity among consumers

- **Visualization Findings:**

- The bar chart shows that spending is concentrated on a few categories, with Category 1 standing out as exceptionally good performance.

  Categories with lower spending still have potential but may require targeted marketing strategies in order to increase their share.

Business Implications:

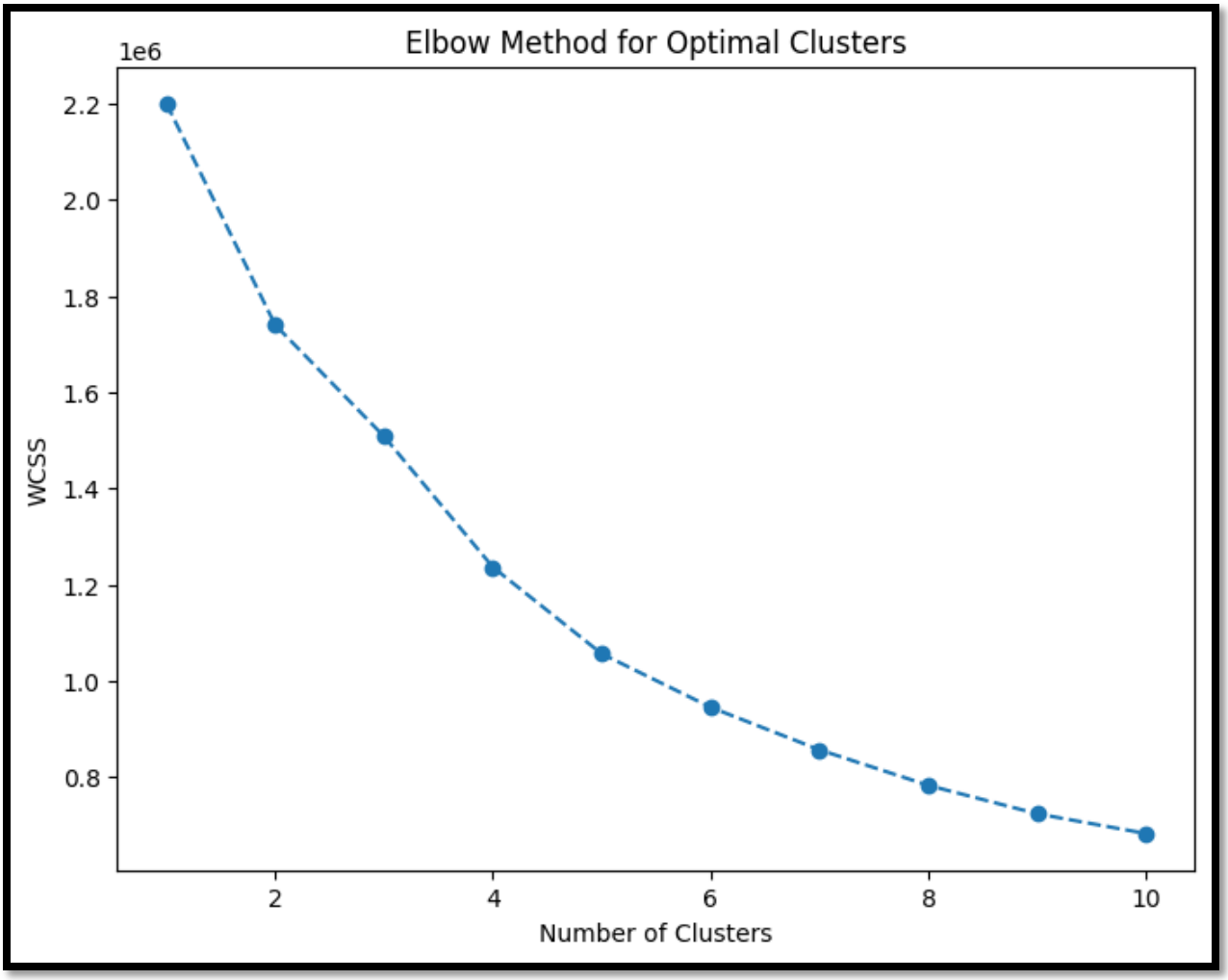
1. **High Demand Prediction:**
   a. Categories with higher spending (e.g., Categories 1, 5, and 8) are more likely to see sustained or increasing demand due to their popularity.
   b. These categories could be prioritized for inventory management and marketing campaigns to maximize returns.
2. **Customer Preference Insights:**
   a. The higher spending in these categories closely aligns with consumer preferences and purchasing behavior. This provides corporations an opportunity to analyze and improve their products or services in these areas
3. **Resource Allocation:**
   a. Marketing budgets, advertising strategies, and promotional efforts can be allocated to these high-performing categories to maximize their demand potential.

Method 2: Clustering Analysis

Methodology and Findings

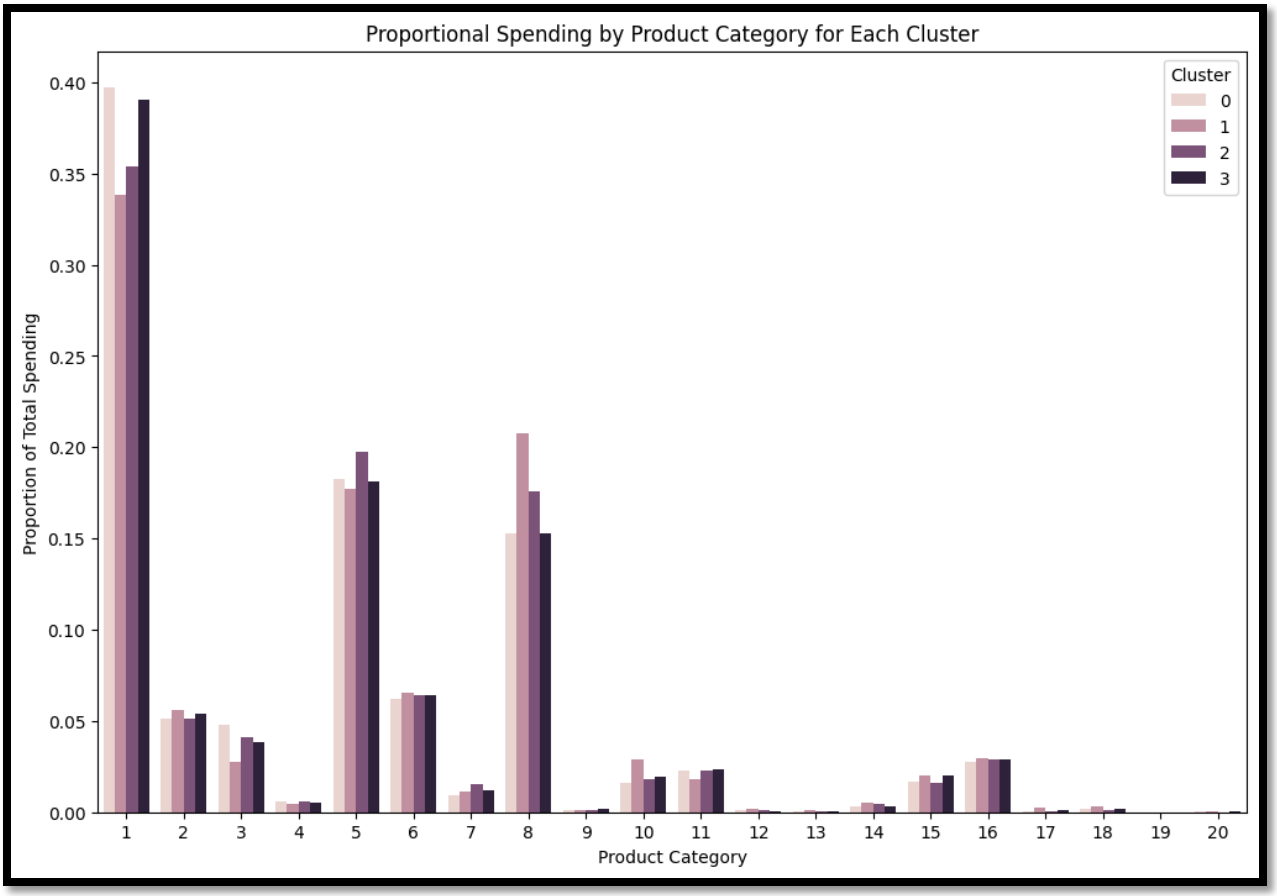
Step 1: Elbow Method for Cluster Formation

- **Analysis**:
  - The Elbow Method was used in order to find the optimal number of clusters. The WCSS plot showed that 4 clusters were optimal for customer segmentation.
- **Relevance**:
  - Well-defined clusters allow an aggregation of customers showing similar behaviors, hence a profound analysis of demand within segments.



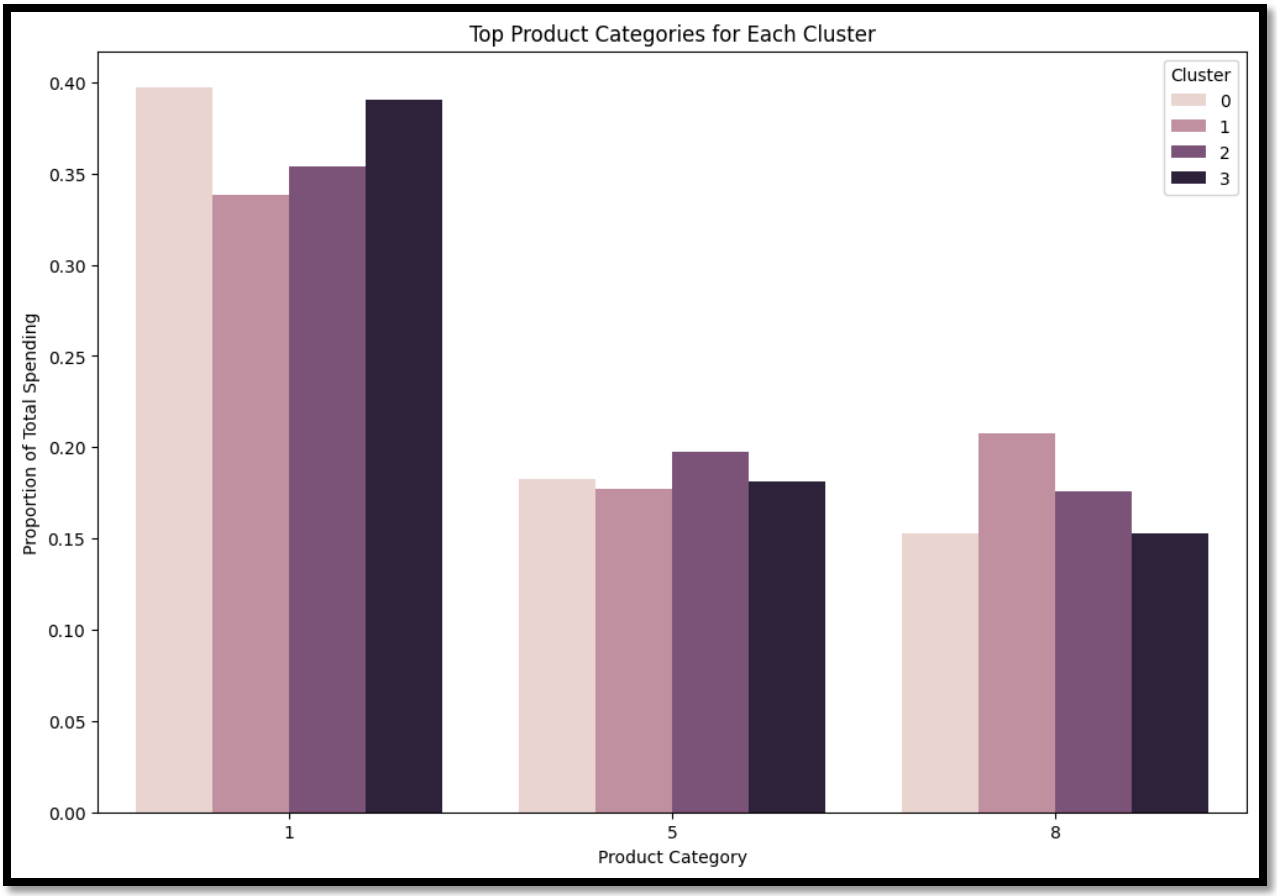
Step 2: Proportional Spending by Product Category Across Clusters

- **Analysis**:
  - Product Categories 1, 5, and 8 consistently show the highest proportional spending across all four clusters.
  - Category 1 accounts for approximately 35–40% of spending in each cluster.
- **Relevance**:
  - These categories are the most popular and therefore likely to experience sustained or increased demand.



Step 3: Top Categories by Cluster

- **Analysis**:
  - Cluster-specific analysis shows consistent preferences:
    - **Cluster 0**: Categories 1 (39.7%), 5 (18.3%), and 8 (15.3%).
    - **Cluster 1**: Categories 1 (33.8%), 8 (20.8%), and 5 (17.0%).
    - **Clusters 2 and 3**: Similar to Cluster 0, with Category 1 dominating.
- **Relevance**:
  - Across all clusters, Categories 1, 5, and 8 still show dominance, proving their wide appeal to the varying customer segments.



Step 4: Spending by Demographics

- **Analysis**:
  - Younger customers (ages 18–35) are the largest contributors to spending.
  - Product Categories 1, 5, and 8 dominate spending across all age groups.
- **Relevance**:
  - Demand in high-demand categories is being driven by younger demographics. Marketing campaigns should focus on this segment to capitalize on demand trends.

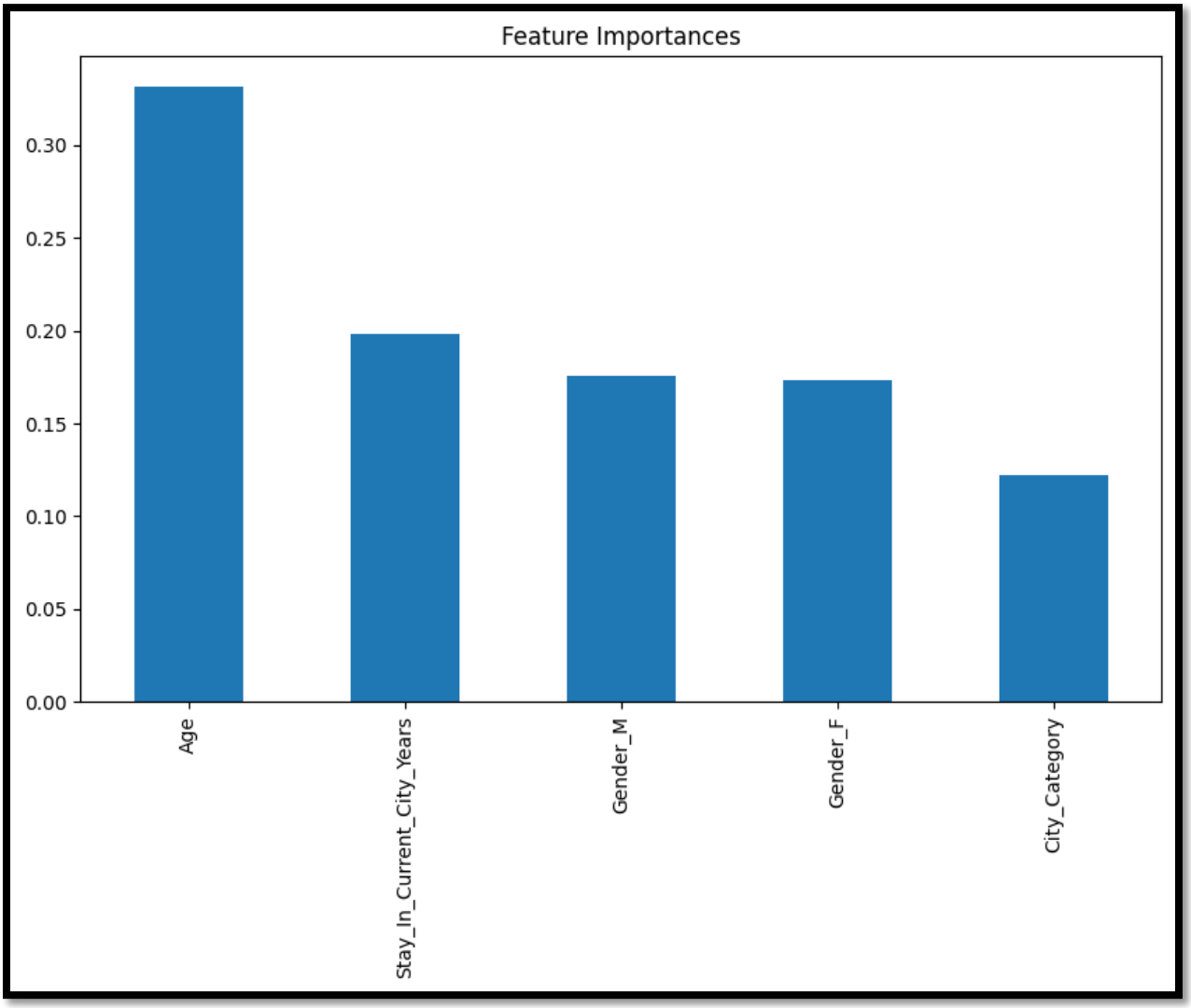
Step 5: Feature Importance from Predictive Modeling

- **Analysis**:

Using a Random Forest Classifier, the most influential features for predicting product category preferences were:

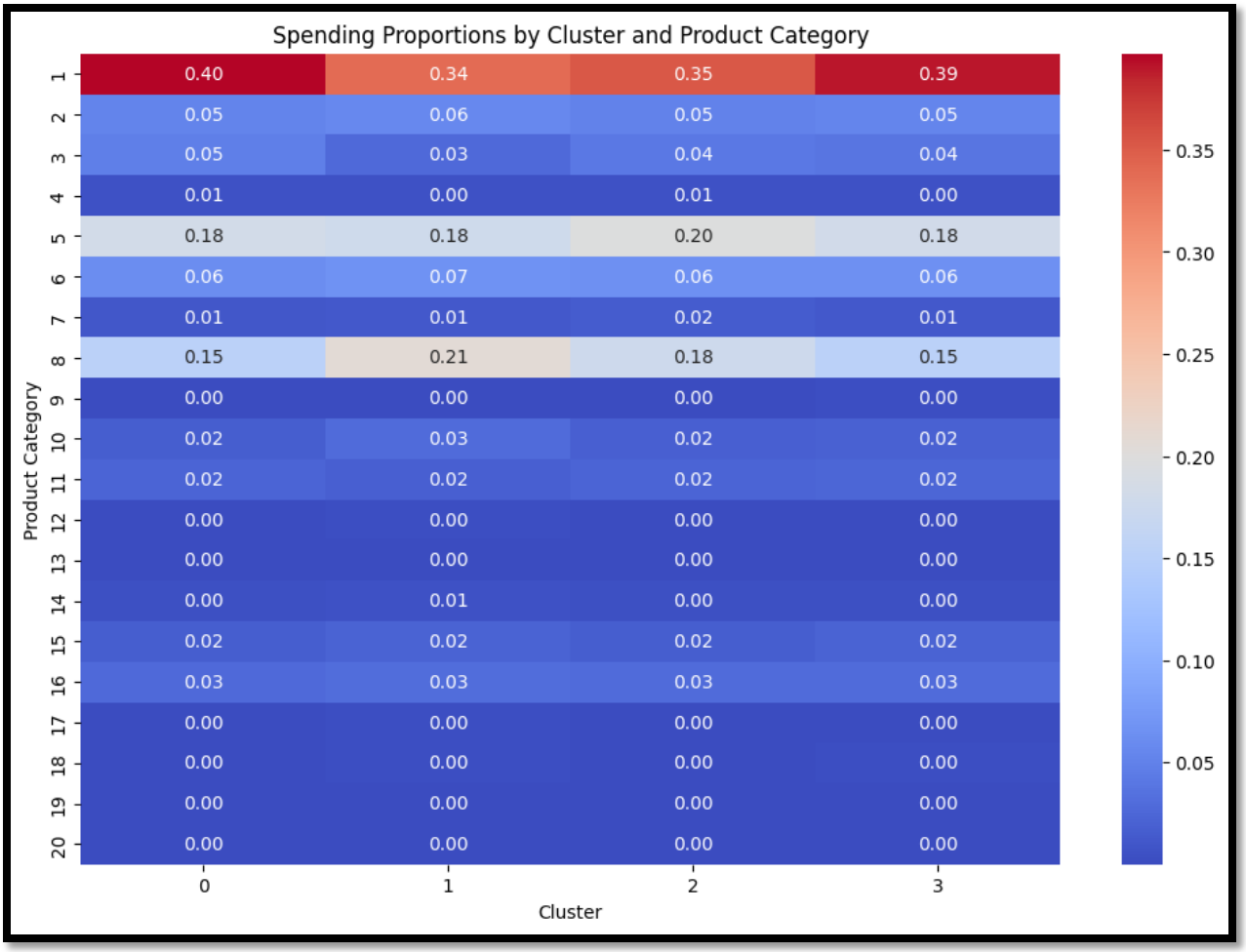
- **Age** (most significant).
- **Stay_In_Current_City_Years** and **Gender** also play critical roles.

- **Relevance**:
    - Understand the demographic factors behind demand for practical knowledge on how to tailor marketing strategies and product positioning.



Step 6: Spending Heatmap

- **Analysis**:
    - The heatmap visually confirms that Categories 1, 5, and 8 dominate spending across all clusters.
    - Spending proportions for other categories are negligible.
- **Relevance**:
    - The heatmap supports the claim that these categories are indeed the key areas of focus for growth and increasing demand.



Key Findings

1. **Dominant Product Categories**:
    a. Categories 1, 5, and 8 consistently account for the highest spending proportions across clusters and demographics.
    b. Category 1 represents the highest overall demand, followed closely by Categories 5 and 8.

2. **Demographic Trends**:
    a. These categories, in essence, attract mostly young consumers aged between 18–35. Expenditure patterns exhibit consistency across both genders, albeit with minor differences in secondary preferences.
3. **Cluster-Specific Insights**:
    a. Clusters follow the same trends for Categories 1, 5, and 8, showing similarities in demand trends between these clusters.

Actionable Insights

1. **Focus on High-Demand Categories**:
    a. Invest in inventory and marketing for Categories 1, 5 and 8 to meet increasing demand.
2. **Target Younger Audiences**:
    a. Focus marketing campaigns on younger customers (18–35), who contribute the most to spending in these categories.
3. **Personalized Cluster Strategies**:
    a. Create tailored promotions for each cluster:
        i. Cluster 0: highlight Categories 1 and 5.
        ii. Cluster 1: Highlight Categories 8 and 1.
        iii. Clusters 2 and 3: Mirror Cluster 0's strategy.
4. **Seasonal Trends**:
    a. Analyze if seasonal demand patterns exist for any of these categories to optimize stock and marketing plans.
5. **Demographic-Specific Marketing**:
    a. Use the learned feature importances to create targeted offers based on age, gender, and city of residency.

Conclusion

Demand will surely rise for Product Categories 1, 5, and 8, as these have shown overall dominance across most of the clusters and demographic groups. Most of the products under these categories are mainly purchased by young buyers who are the primary target market for promotional and marketing campaigns. From these insights, Walmart can enhance inventory planning and modify marketing strategies to gain the most in sales growth.

5.5 What is the role of occupation in determining spending patterns at Walmart?

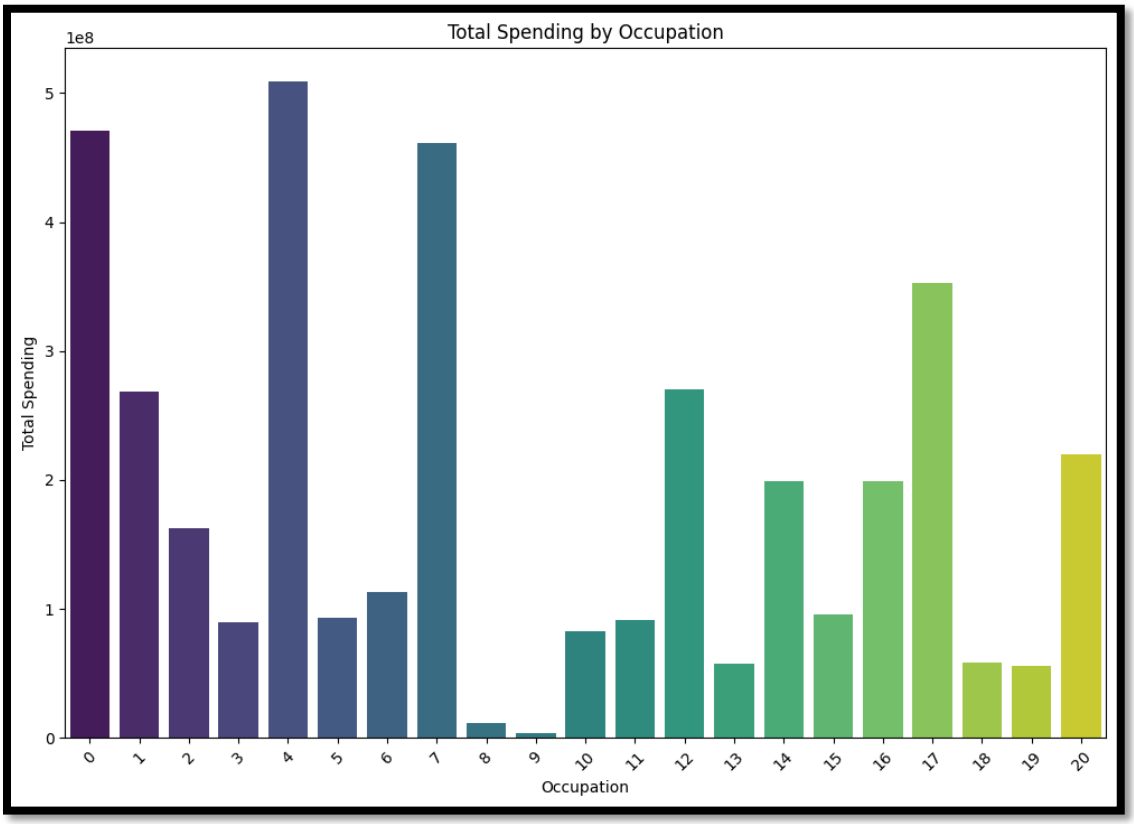
Analysis and Findings

Step 1: Total Spending by Occupation

- **Observation**:

- Occupations 4, 0, 7, and 17 have the highest total spending, with Occupation 4 leading
- Occupations 9 and 8 have lowesr spending, which indicates limited purchasing activity.

- **Insight**:
  - Certain occupations contribute disproportionately to the revenue of Walmart, indicating that occupation plays a significant role in determining spending behavior.



Step 2: Average Spending and Purchase Frequency

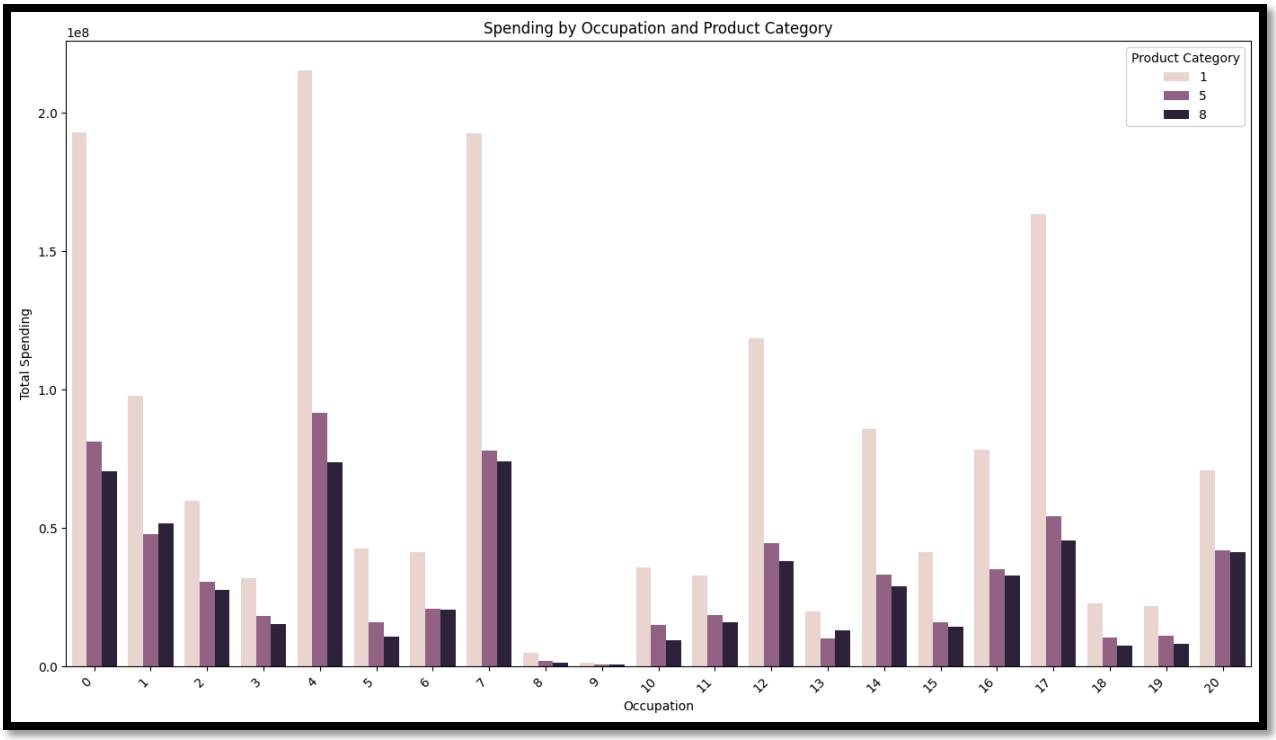
- **Observation**:
  - Occupation 4 has the highest total spending and also shows high average spending per user and a high purchase frequency.
  - Occupations 0, 7 and 17 are of a similar nature with high average spending and frequent purchases.
- **Insight**:
  - These jobs probably have people with more money to spend or who buy things regularly, making them important targets for marketing.

```python
# Group by Occupation to get total and average purchase amount per occupation
occupation_spending = data.groupby('Occupation').agg(
    total_spent=('Purchase', 'sum'),
    average_spent=('Purchase', 'mean'),
    num_purchases=('Purchase', 'count')
).reset_index()

# Sort occupations by total spending
occupation_spending = occupation_spending.sort_values('total_spent', ascending=False)

# Display top occupations by total spending
print("Top Occupations by Total Spending:")
print(occupation_spending.head())
```

```
Top Occupations by Total Spending:
    Occupation  total_spent  average_spent  num_purchases
4            4    509387740    9446.401231          53924
0            0    471048528    9232.986946          51018
7            7    461694565    9507.909244          48559
17          17    352547530    9883.032350          35672
12          12    270321872    9893.564835          27323
```

Step 3: Spending Patterns by Product Category

- **Observation**:
  - Categories 1, 5, and 8 dominate spending across all occupations.
  - Occupation 4 spends significantly on Category 1
  - Occupation 0 is most strongly associated with Category 1 but also holds substantial interests in Categories 5 and 8.
- **Insight**:
  - Spending changes a little by job, but people still like Categories 1, 5, and 8. This proves that these categories fit customer needs for different types of jobs.



Step 4: Proportional Spending Visualization

- **Observation**:
  - Bar plots and heatmaps obviously show that Occupations 4, 0 and 7 have the largest spending.
  - Spending in other industries is more diffused but has a lesser overall impact.
- **Insight**:
  - Occupations with the most spending probably represent some groups of people or economic situations that Walmart can focus on to increase profits.

Key Findings

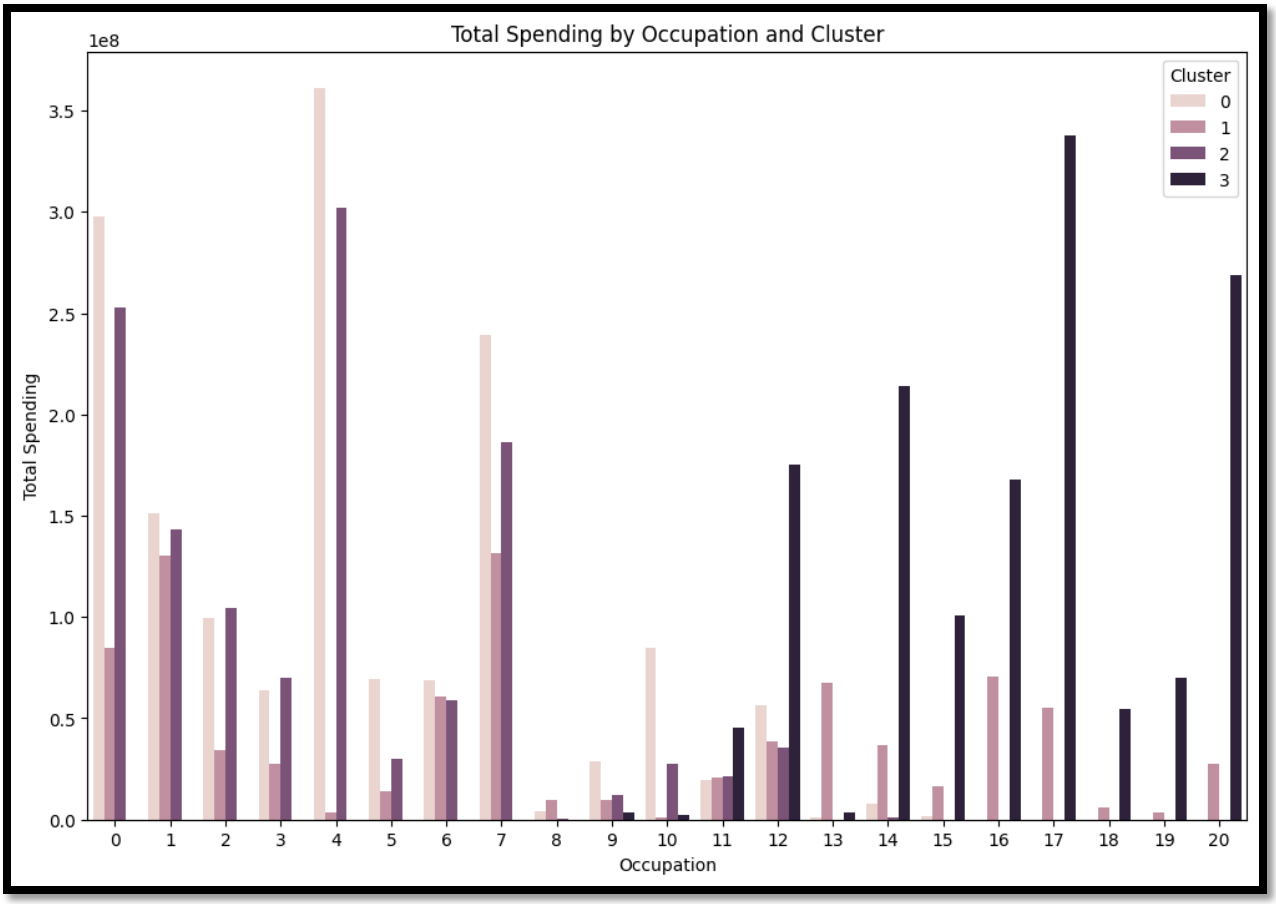
1. **High-Spending Occupations**:
   a. Occupations 4, 0, 7, and 17 are the best customer segments based on total and average spending.
   b. These jobs make up most of Walmart's income and should be focused on in marketing and inventory plans.
2. **Category Preferences**:
   a. Categories 1, 5, and 8 are popular across all occupations, with Occupation 4 leading in spending for these categories.
   b. This shows the importance of these categories in Walmart's product mix.

3. **Spending Consistency**:
    a. Those occupations with higher average spending per user also have higher purchase frequencies, indicating loyal customer segments with predictable buying patterns.
4. **Low-Spending Occupations**:
    a. Occupations 9 and 8 contribute the least, indicating less engagement or spending power. Walmart could look at specific ways to increase spending in these segments.

Method 2: Clustering Analysis

1. Total Spending by Occupation Across Clusters

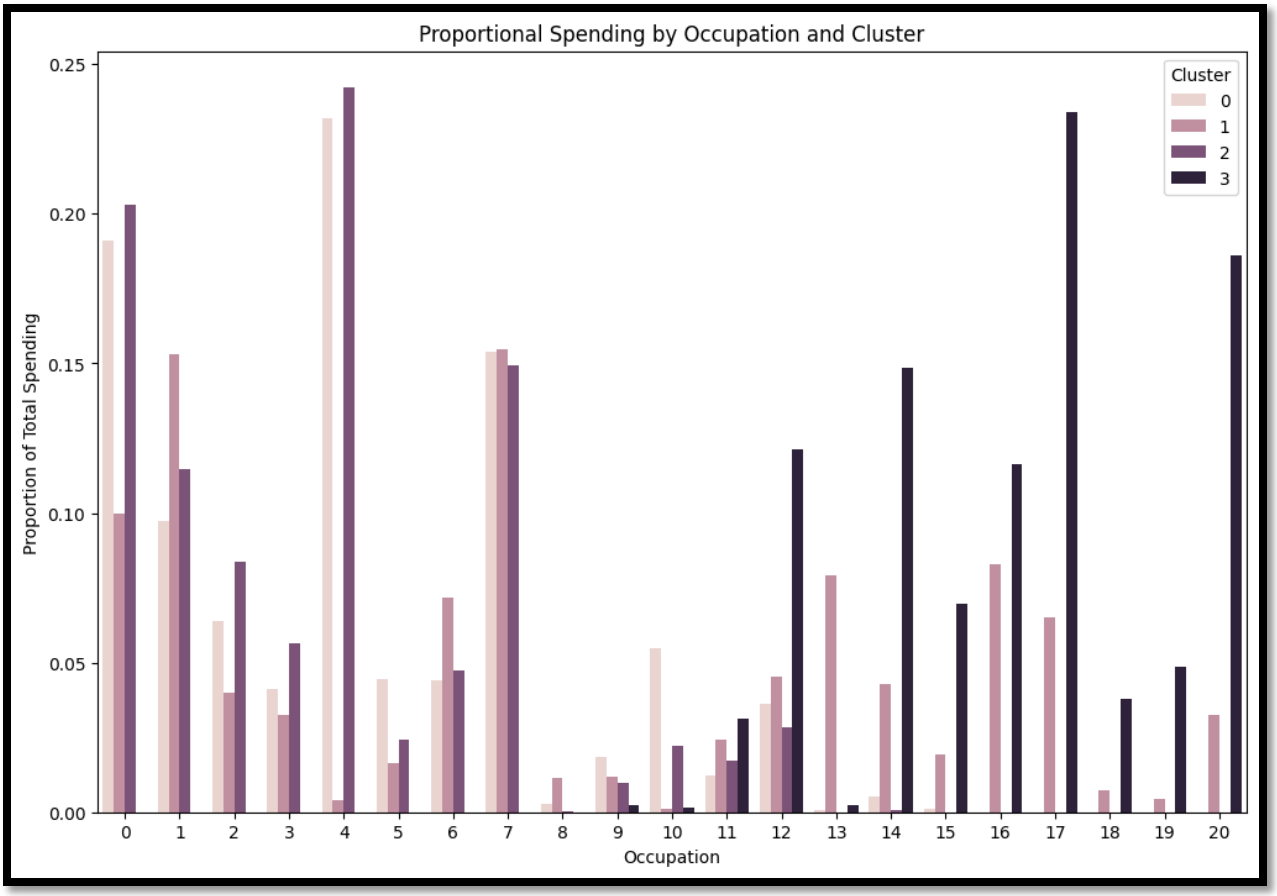
- **Observation**:
    - Occupations 4 and 7 are strong in spending, particularly in Clusters 0 and 1.
    - Occupation 20 exhibits higher spending in Cluster 3, suggesting unique shopping patterns.
- **Insight**:
    - Spending patterns across clusters vary significantly by occupation, reflecting differences in demographics or lifestyle preferences associated with each occupation
- **Relevance**:
    - Occupations 4 and 7 represent significant customer segments for Walmart, suggesting that targeted marketing efforts should be directed toward these groups



2. Proportional Spending by Occupation and Cluster

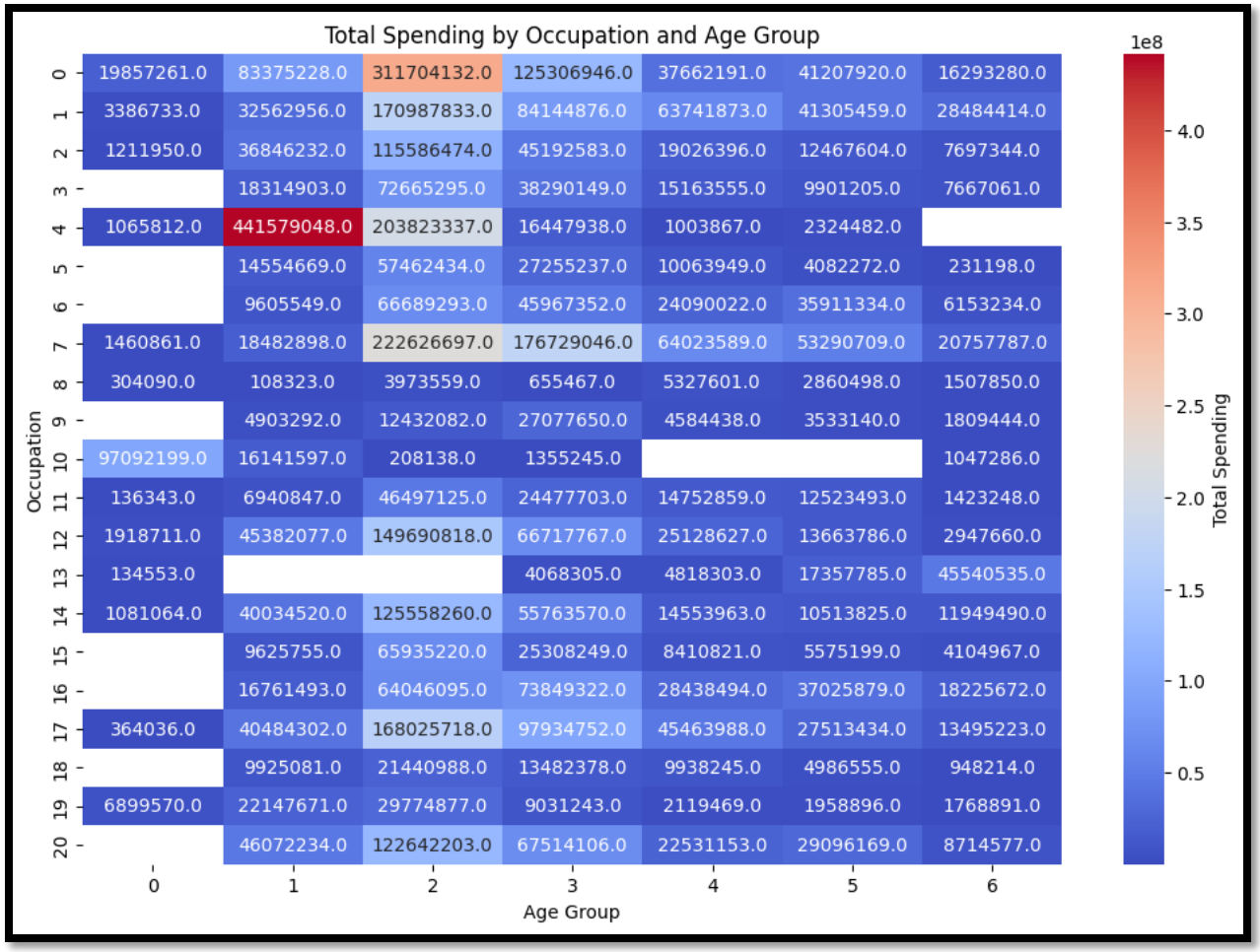
- **Observation**:
  - Occupation 4 has the highest proportional spending in Cluster 0, while Occupation 7 leads in Cluster 1.
  - Occupation 20 stands out in Cluster 3, indicating specialized shopping habits in this group.
- **Insight**:

  The varying amounts of money spent on different jobs also correspond quite well with the characteristics of certain customer groups.
- **Relevance**:
  - The information can be used by Walmart to create special offers for each job in its chosen group.



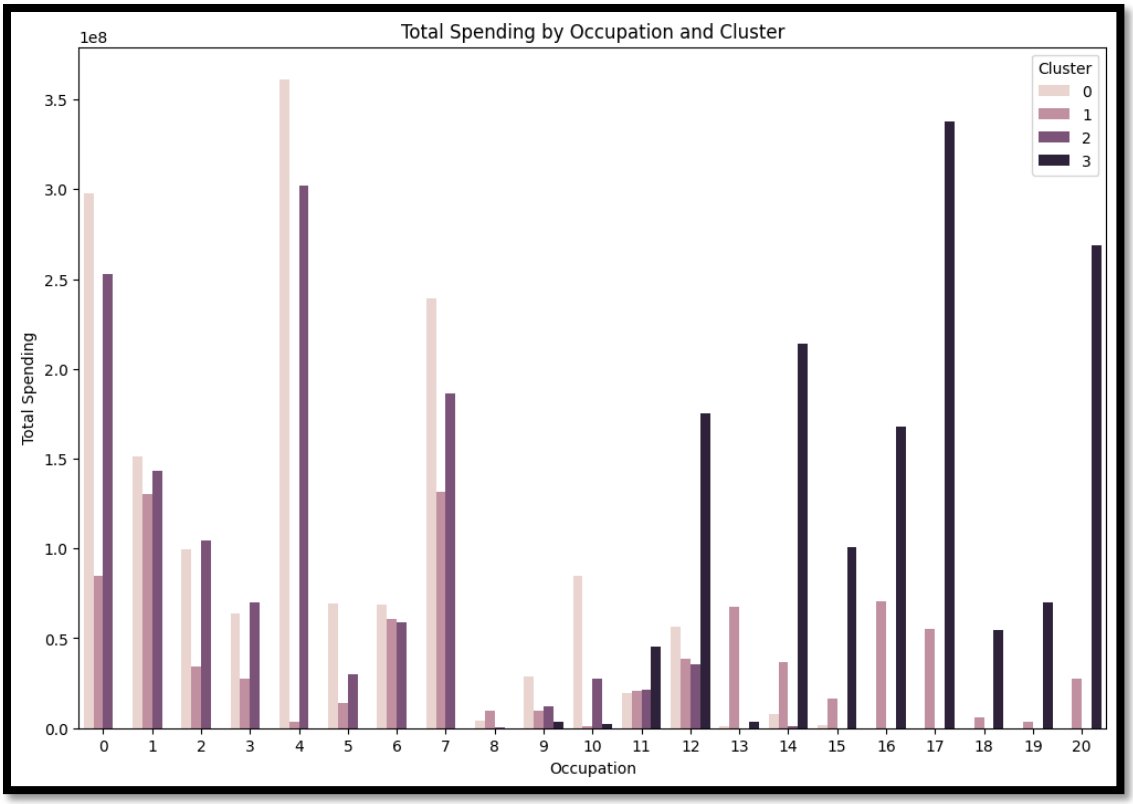
3. Spending by Occupation and Age Group

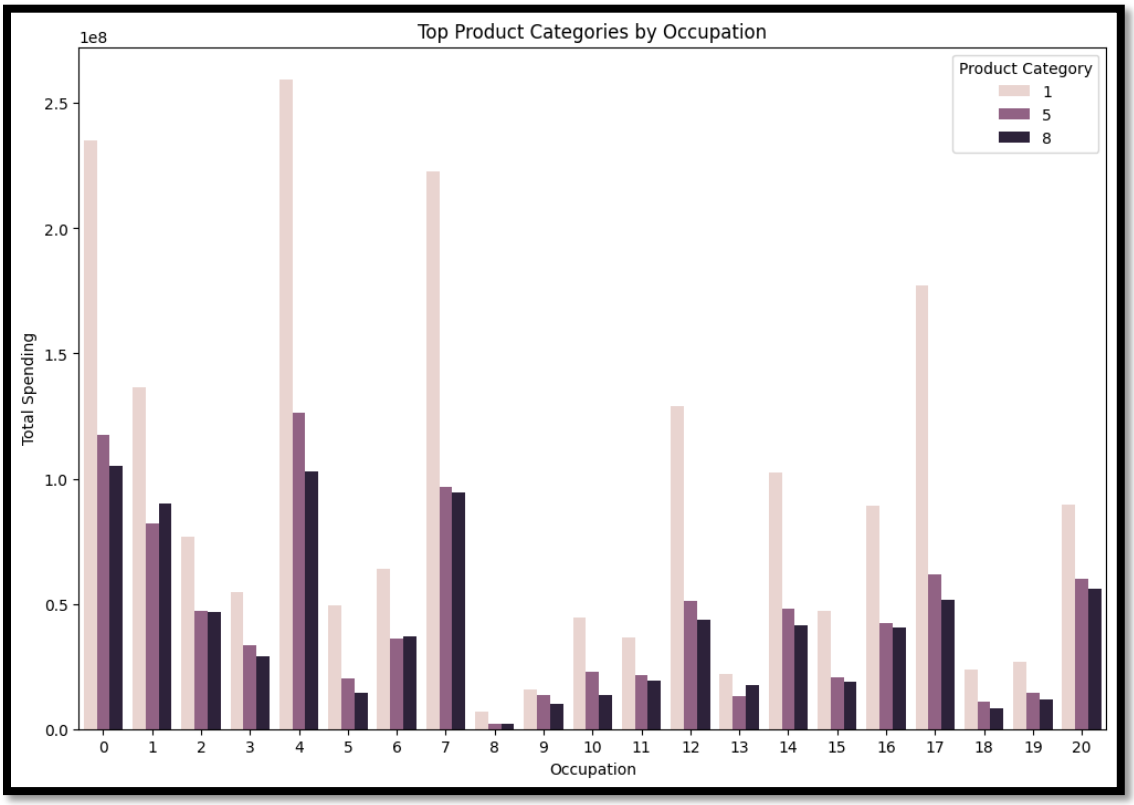
- **Observation**:
  - Occupation 4, combined with age groups (18–25 years), shows the highest spending.
    Younger people spend more money in most jobs, and spending goes down as people get older
- **Insight**:
  - Occupations interact with age to create distinct spending profiles, highlighting generational preferences.
- **Relevance**:
  - Walmart can target younger generations that hold high-paying jobs with discounts on things such as technology and apparel.

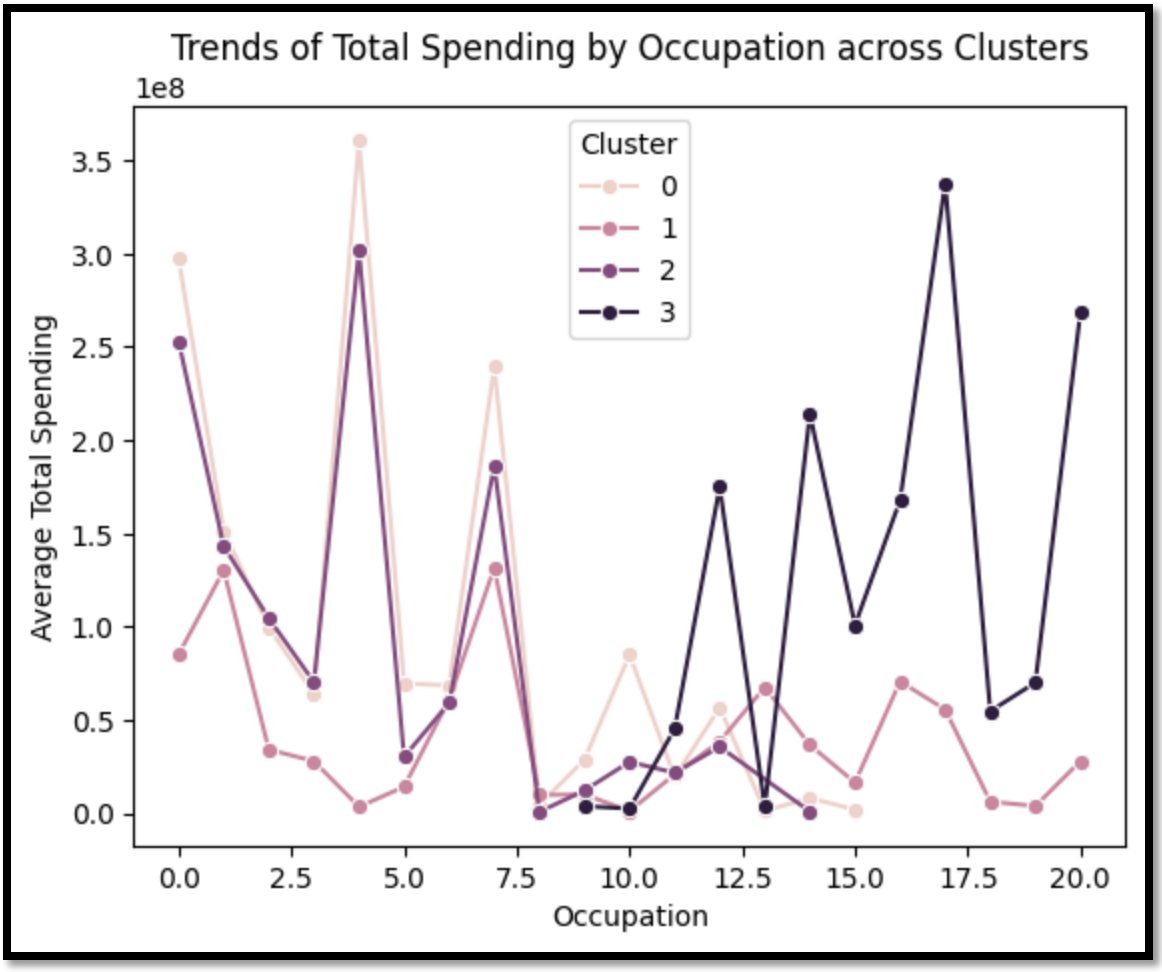


4. Top Product Categories by Occupation

- **Observation**:
  - Product Category 1 dominates spending across all occupations, followed by Categories 5 and 8.
  - Occupations 4 and 7 show slightly higher spending on Categories 5 and 8 compared to others.
- **Insight**:
  - Occupational roles influence product category preferences. Walmart can optimize inventory and promotions for Categories 1, 5, and 8, especially for high-spending occupations.

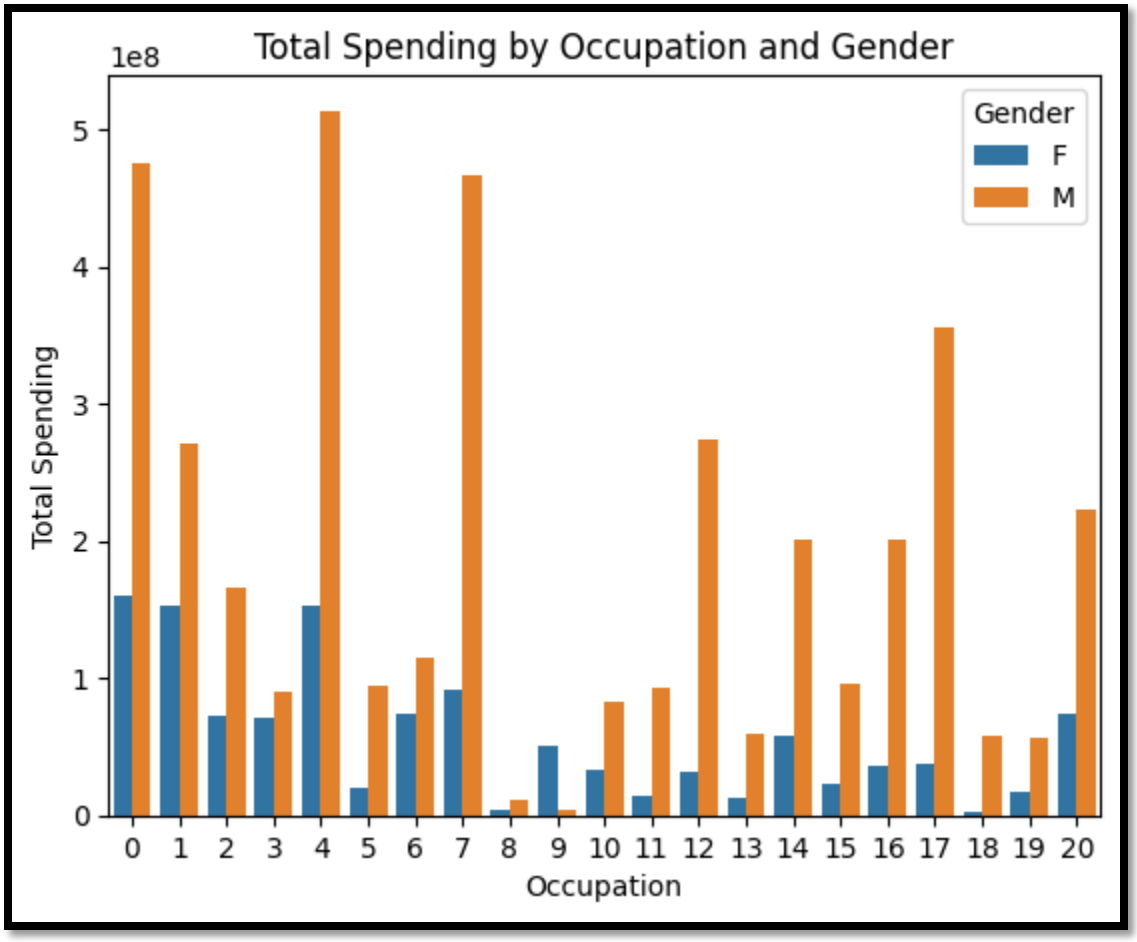

Total Spending by Occupation and Cluster
1e8
Cluster
0
1
2
3
Total Spending
Occupation
0.0
0.5
1.0
1.5
2.0
2.5
3.0
3.5
0 1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20


Top Product Categories by Occupation
1e8
Product Category
1
5
8
Total Spending
Occupation
0.0
0.5
1.0
1.5
2.0
2.5
0 1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20



5. Spending by Occupation and Gender

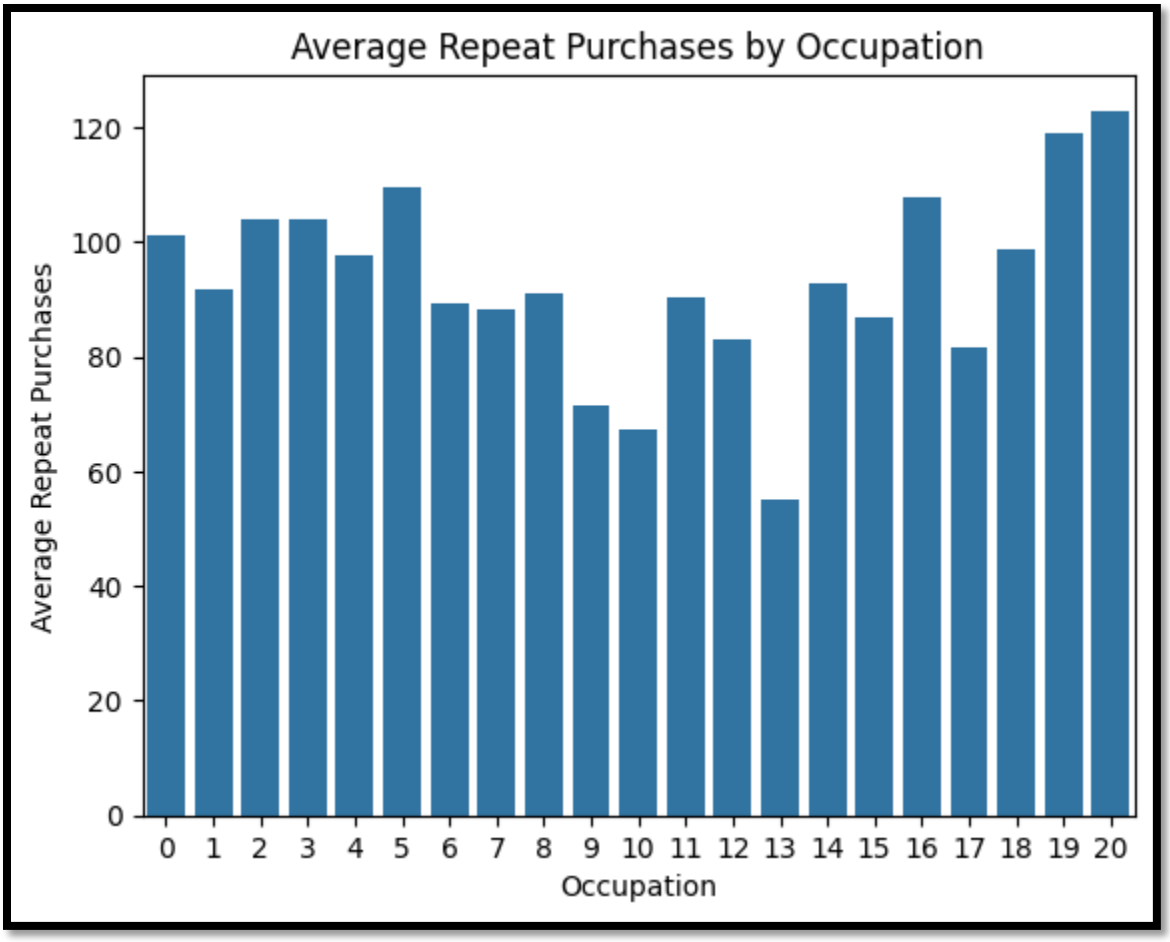
- **Observation**:
  - Males outspend females across all occupations.
  - Occupations 4, 7, and 20 show the highest male spending, reflecting gendered shopping patterns.
- **Insight**:
  - Gender differences within occupations influence overall spending patterns.
- **Relevance**:
  - Walmart can tailor gender-specific promotions within high-spending occupations to maximize engagement.



6. Average Repeat Purchases by Occupation

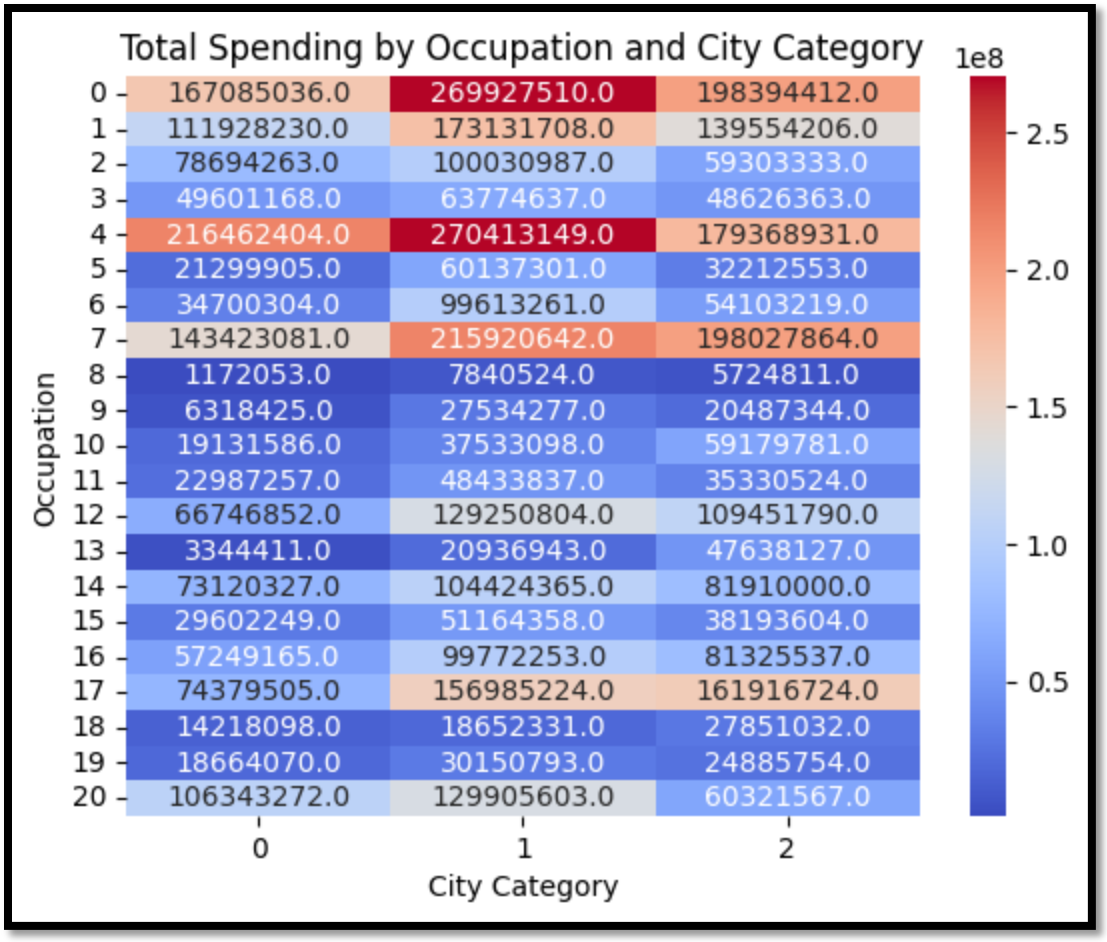
- **Observation**:
  - Occupations 19 and 20 have the most repeat purchases, indicating very loyal customers.
  - Lower repeat purchases are seen for some occupations, which may indicate less consistent shopping habits.
- **Insight**:

  Loyalty and Engagement levels vary significantly by occupation, providing Walmart with an opportunity to target frequent buyers more effectively.
- **Relevance**:
  - Starting Loyalty programs tailored to jobs with frequent repeat purchases can enhance customer retention and prolong their engagement.



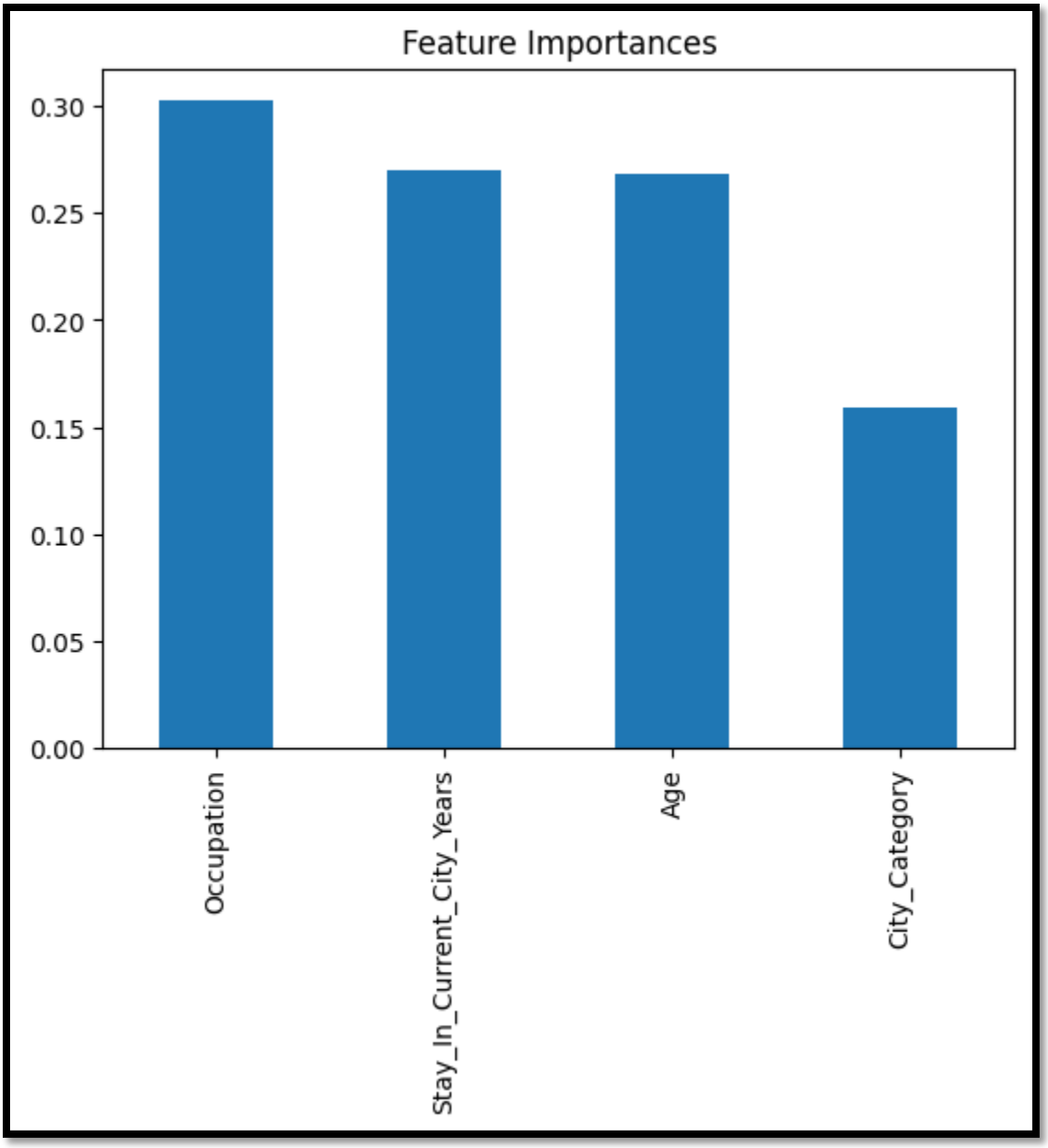
7. Spending by Occupation and City Category

- **Observation**:
  - Occupations 4 and 7 spend the most in urban areas (City Category 1), followed by little contributions from City Category 0.
- **Insight**:
  - Urban customers dominate spending across most occupations, reflecting differences in purchasing power and access.
- **Relevance**:
  - Walmart can focus on urban centers to maximize sales from high-spending occupations.



8. Feature Importance from Predictive Modeling

- **Observation**:
  - Occupation is the most significant predictor of spending, followed by Stay_In_Current_City_Years and Age.
- **Insight**:
  - The high importance of occupation shows it strongly affects the revenue patterns of Walmart.
- **Relevance**:
  - Predictive models can help improve Walmart's marketing strategies by predicting how people will spend money based on their jobs.



Key Findings

1. **High-Spending Occupations**:
    a. Occupations 4, 7, and 20 dominate spending, especially in Clusters 0, 1, and 3.
    b. These occupations are key revenue drivers for Walmart.
2. **Category Preferences**:
    a. Product Categories 1, 5, and are very popular for any job. This means they must fit the needs of the customers.
3. **Generational Trends**:

    a. Younger age groups in high-spending occupations exhibit distinct preferences, with a focus on technology, apparel, and essentials.
4. **Cluster-Specific Insights**:
    a. Occupations align with specific clusters, highlighting the importance of demographic segmentation in understanding spending behaviors.
5. **Gender Dynamics**:
    a. Males consistently outspend females across occupations, with significant differences in spending patterns.
6. **Loyalty Trends**:
    a. Occupations with high repeat purchases indicate strong engagement and present opportunities for loyalty programs.

Actionable Insights

1. **Target High-Spending Occupations**:
    a. Create special marketing plans for Occupations 4, 7, and 20 to make the most of their buying power.
2. **Focus on Dominant Product Categories**:
    a. Optimize inventory and promotions for Categories 1, 5, and 8 to cater to occupational preferences.
3. **Engage Urban Customers**:
    a. Give importance to urban centers with customized offers for high-value occupations.
4. **Leverage Predictive Models**:
    a. Use occupation as an important variable in predictive models to anticipate spending patterns and refine marketing strategies.
5. **Enhance Loyalty Programs:**
    a. Introduce occupation-specific loyalty programs targeting repeat buyers to boost customer retention.

Conclusion: Profession has a strong effect on the formation of spending habits at Walmart, including overall spending, product preferences, clustering behavior, and demographics. Using occupational trend data, Walmart can run better marketing campaigns, level up stock levels, and increase customer engagement, ultimately yielding more revenue and customer satisfaction.

6. Results and Discussion

Additional Insights of Model Comparison:

Metric	Linear Regression	Decision Tree	Random Forest
Mean Absolute Error (MAE)	2340.16	2163.25	2337.83
Mean Squared Error (MSE)	9413996.64	8874693.51	9514230.86
Root Mean Squared Error (RMSE)	3068.22	2979.04	3084.51
Mean Absolute Percentage Error (MAPE)	35.34%	35.34%	35.34%
R-squared (R^2)	0.6188	0.6407	0.6148

Among the three models being compared in the performance metrics from the table, it appears that the Decision Tree model is the best. It can be observed to have the lowest MAE of 2163.25, the lowest MSE of 8874693.51, and the lowest RMSE of 2979.04. Besides, the Decision Tree model has the highest value of R-squared, which is 0.6407, explaining a greater amount of variance than Linear Regression and Random Forest models. Overall, the Decision Tree model is seen to perform the best at making more accurate predictions or capturing the underlying patterns in the data.

6.1 Key Factors Driving High-Value Purchases

The clustering analysis uncovered distinct customer segments based on demographic and purchasing behaviors. Among the four clusters identified, Cluster 1 emerged as the highest value segment, with customers making significantly larger average purchases. These customers predominantly reside in urban areas, belong to higher occupational strata, and have lived in their current city for a considerable period.

- **Key Insights**:
  - **Occupation**: Occupation emerged as the most influential feature, with professionals in technology, finance, and related fields dominating high-value purchases.
  - **City Category**: Customers from City Category C had strong associations with high-value purchases.
  - **Age and Marital Status**: Middle-aged (26–35) and married customers were overrepresented in high-value segments, indicating these demographics as key drivers of spending.

6.2 Influence of Demographics on Purchasing Behavior

The feature-wise analysis highlighted critical demographic factors influencing purchasing behaviors:

- **Age**: Younger customers (18–35) were the largest contributors to overall spending. This pattern was consistent across clusters, reflecting the importance of this demographic group.
- **Gender**: Males consistently spent more than females across all clusters, with the largest gap observed in Cluster 3.
- **Marital Status**: Married customers exhibited marginally higher spending, particularly in high-value categories, though their influence was less significant compared to age and gender.

6.3 Variation in Product Demand by City Categories

City-specific analysis revealed stark differences in purchasing patterns:

- **City B** (Suburban): The highest overall spending, with significant demand for Product Categories 1 (electronics) and 5 (household goods).
- **City A** (Urban): Balanced demand across categories, reflecting diverse consumer preferences.
- **City C** (Rural): Heavy reliance on Product Category 1, indicating a focus on essentials and core products. Statistical tests, such as ANOVA and chi-square, confirmed significant variations in product popularity across cities but no notable differences in total spending.

6.4 Categories Likely to See an Increase in Demand

Spending-based and clustering analyses revealed consistent growth in Product Categories 1, 5, and 8 across all segments and demographics:

- **Product Category 1** consistently accounted for 35–40% of total spending, reflecting sustained and growing demand.
- **Product Categories 5** and **8** ranked as the next most popular categories. Demographic analysis showed younger customers as primary drivers of demand, reinforcing the need for age-targeted marketing strategies.

6.5 Role of Occupation in Spending Patterns

Occupation significantly influenced spending behaviors, with the following key findings:

- **High-Spending Occupations**: Occupations 4, 7, and 20 were the highest contributors to total and average spending.
- **Product Category Preferences**: Occupations consistently favored Categories 1, 5, and 8, aligning with general consumer trends.
- **Loyalty Trends**: Occupations with higher repeat purchases (e.g., Occupations 19 and 20) displayed strong engagement, highlighting opportunities for targeted loyalty programs. The analysis underscored the importance of occupation-specific marketing strategies to maximize revenue and engagement.

7. Conclusion

The analysis provided actionable insights into Walmart's customer behavior and spending patterns, answering the research questions comprehensively:

1. **Key Drivers of High-Value Purchases**: Occupation, city category, and age emerged as the most influential factors, with specific focus required on high-value clusters dominated by professionals and middle-aged customers.
2. **Demographic Influence on Spending**: Age and gender showed significant impacts on purchasing behavior, with younger males contributing the most to high-value purchases.
3. **City-Specific Variations**: City B exhibited the highest spending, while City C demonstrated reliance on specific product categories like electronics and essentials.
4. **Product Categories with Growth Potential**: Categories 1 (electronics), 5 (household goods), and 8 (health & wellness) consistently ranked as the most demanded across all clusters and demographics.
5. **Role of Occupation in Spending**: Occupational groups exhibited distinct purchasing behaviors, with high-value occupations showing greater spending and loyalty patterns.

Recommendations

- **City-Specific Strategies**: Focus on City B for promotional campaigns while addressing the unique demands of City A and City C.
- **High-Demand Categories**: Invest in inventory and marketing for Categories 1, 5, and 8 to sustain growth and meet demand.
- **Targeted Marketing**: Develop campaigns targeting younger demographics (18–35) and high-spending occupations (e.g., Occupations 4, 7, and 20).
- **Loyalty Programs**: Enhance engagement by introducing occupation-specific loyalty rewards for repeat customers.
- **Personalized Clustering Strategies**: Tailor promotional efforts based on cluster-specific preferences and demographic attributes.

By aligning business strategies with customer behavior insights, Walmart can enhance customer satisfaction, optimize inventory, and drive revenue growth across diverse market segments.

8. Acknowledgements

We would like to express our sincere gratitude to the following individuals and resources that contributed to the successful completion of this project:

- Kaggle.com for hosting the Walmart Sales Dataset, which served as the foundation for our analysis.
- The Python open-source community, particularly the developers behind tools like Pandas, Scikit-learn, Seaborn, and Matplotlib, for providing powerful tools that enabled effective data preprocessing, analysis, and visualization.
- Our team members—Devansh Pandey, Greeshma Gajula, Kaustubh Ashok Gawande, Vishwaksen Reddy Manda, and Yuvarekha Mahendran—for their dedicated collaboration and contributions to every aspect of the project.

9. References

1. Kaggle Dataset: http://www.kaggle.com/datasets/devarajv88/walmart-sales-dataset