E-Commerce (Walmart) Sales

Under the guidance of
Dr. Riyaz Sikora

GROUP 6
Devansh Pandey
Greeshma Gajula
Kaustubh Ashok Gawande
Vishwaksen Reddy Manda
Yuvarekha Mahendran

UTA

Introduction

- **Walmart - A Global Retail Leader**

Walmart is a globally recognized retail giant known for its extensive range of products and competitive pricing.

The company is committed to delivering convenience and value, serving millions of customers across the globe.

- **Why Analyze the Walmart Dataset?**

Analyzing this dataset provides a deep dive into customer behaviors and purchasing patterns at Walmart.

It reveals significant insights into customer demographics, product preferences, and overall spending patterns.

- **Key Insights from the Dataset:**

Customer Demographics: Understand who Walmart's customers are and how different segments shop.

Product Preferences: Analyze which product categories are most popular.

Spending Patterns: Examine how much customers spend, and what factors influence this behavior.

Applications of Insights

Strategic Applications of the Data:

- Inventory Management: Tailoring stock levels based on demand patterns.
- Targeted Marketing: Reaching out to specific customer segments with personalized promotions.
- Customer Relationship Management (CRM): Building stronger relationships by understanding customer needs.

The insights from this dataset support strategic decision-making in key areas:

- Marketing Strategies: Designing campaigns based on actual customer behavior.
- Customer Segmentation: Identifying distinct customer groups for focused engagement.
- Product Demand Forecasting: Improving predictions for high-demand products.
- Product Recommendation Systems

Data Description and Source

- **Data Source:** Kaggle.com
- **URL**: https://www.kaggle.com/datasets/devarajv88/walmart-sales-dataset
- This dataset includes 550,068 Walmart transactions, capturing key customer purchase details.
- **Target Variable** – Purchase

Analyzing the purchase amount helps in:

- Forecasting demand.
- Personalizing marketing strategies.

Categorical Variables:

- User_ID: Unique identifier for each customer.
- Product_ID: Unique identifier for each product purchased.
- Gender: Gender of the customer (Male/Female).
- Age: Age group of the customer (e.g., 18-25, 26-35).
- City_Category: City type where the customer resides (A, B, C).
- StayInCurrentCityYears: Number of years the customer has lived in their current city.
- Marital_Status: Binary categorical variable, '0' (Unmarried) and '1' (Married)
- ProductCategory: Masked identifier for the product category.

Numerical Variables:

- Purchase: The transaction amount spent on a product (continuous variable).
- Occupation: Masked identifier for customer's occupation (treated as categorical for grouping purposes).

Methodology-Preprocessing

Step	Action Taken	Purpose
Outlier Removal	Removed rows with values outside IQR range for numerical columns. After removal, the shape is (409468, 10)	To ensure data integrity and remove extreme anomalies.
Categorical Encoding	Converted Gender (M/F → 1/0) and City_Category (A/B/C → 0/1/2) using label encoding.	To make data suitable for machine learning models.
Scaling	Standardized numerical features like Purchase using StandardScaler.	To bring features to a uniform scale for better model accuracy.
Train-Test Split	Split data into 80% training and 20% testing sets.	To evaluate the model on unseen data.

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

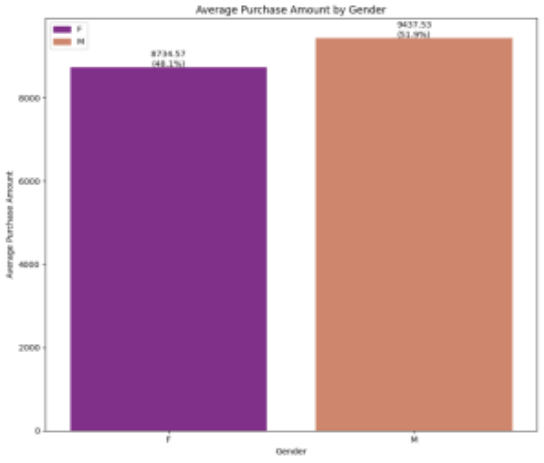
```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 550068 entries, 0 to 550067
Data columns (total 10 columns):
 #   Column                      Non-Null Count   Dtype
---  ------                      --------------   -----
 0   User_ID                     550068 non-null  int64
 1   Product_ID                  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
```

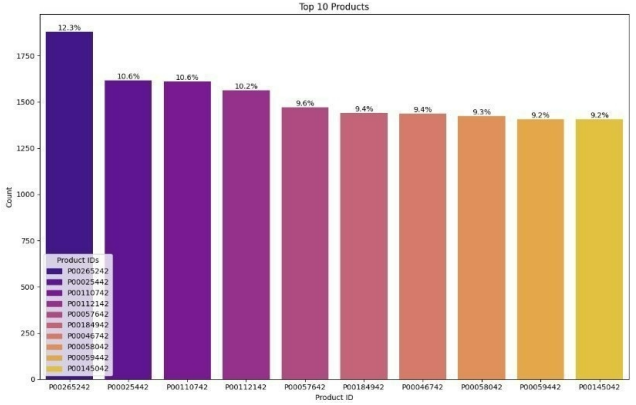
```
#Outlier Removal
for feature in walmart_df.select_dtypes(include=['number']).columns:
    q1 = walmart_df[feature].quantile(0.25)
    q3 = walmart_df[feature].quantile(0.75)
    iqr = q3 - q1
    lower_bound = q1 - (1.5 * iqr)
    upper_bound = q3 + (1.5 * iqr)
    walmart_df = walmart_df[(walmart_df[feature]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

#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)
```

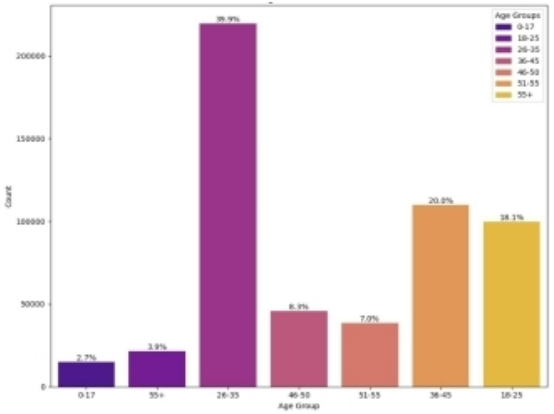
Exploratory Data Analysis (EDA)

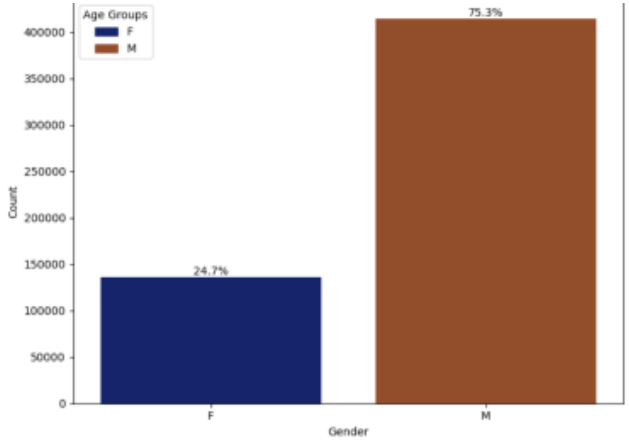
Key Points:

- Demographic patterns (gender, age distribution)
- Product preferences (top categories/products)
- Purchase patterns based on key variables (e.g., gender, city category)

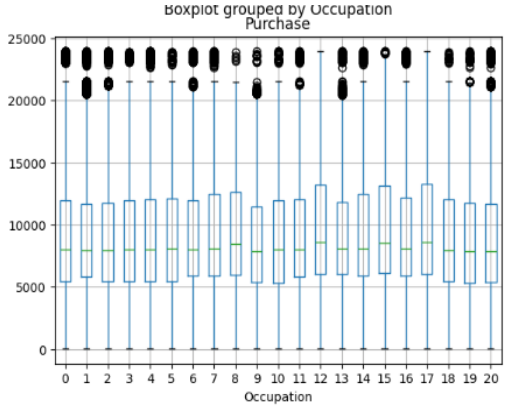


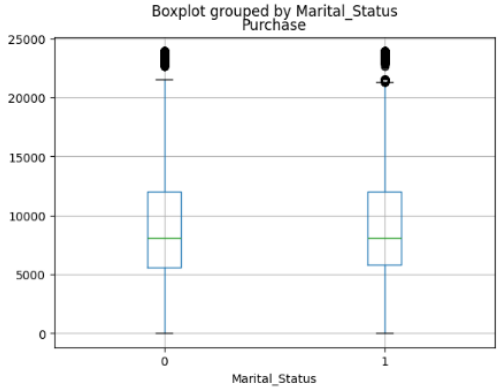


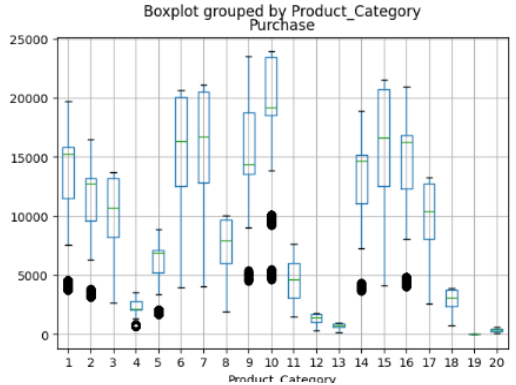


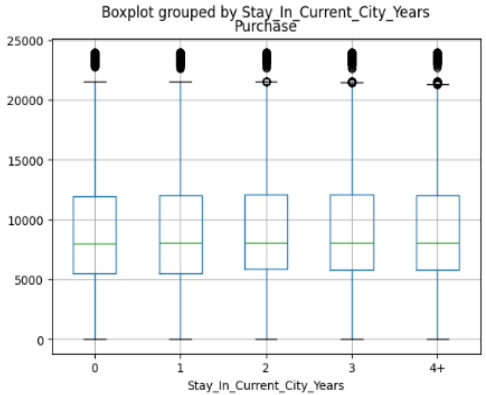


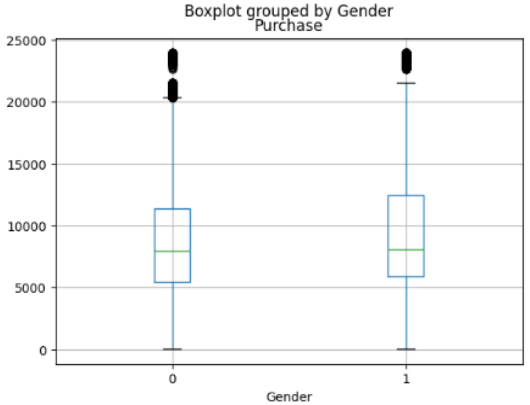
Feature-Wise Analysis of Purchase Trends





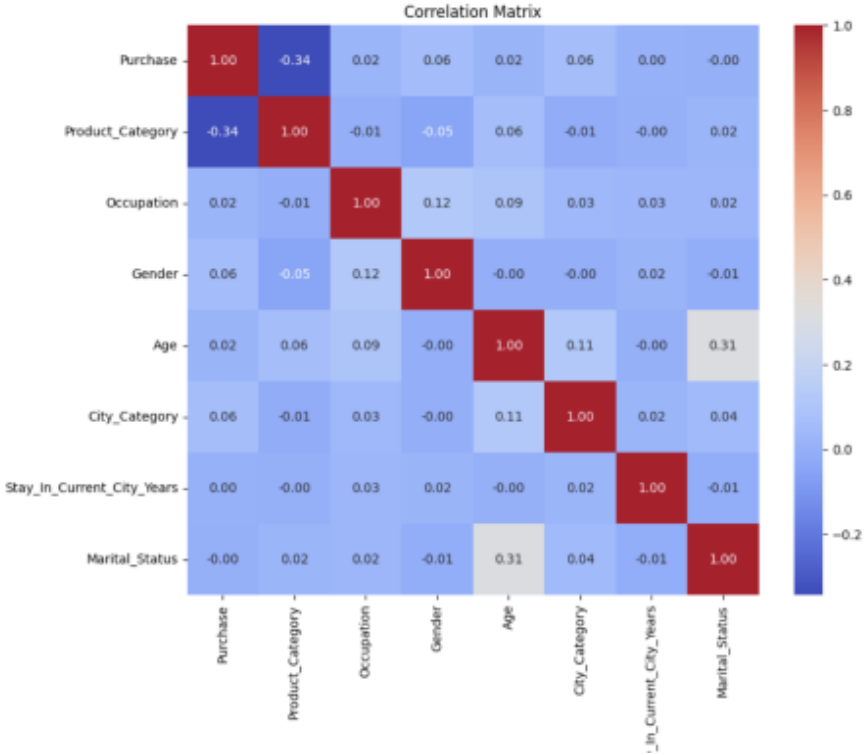






Correlation Analysis

- **Purpose**: The heatmap visualizes correlations among numeric features, highlighting their relationships with Purchase.
- **Key Insight:** Product_Category has a moderate negative correlation (-0.34) with Purchase, indicating its significant influence on purchase amounts.
- **Weak Relationships:** Other features like Gender, Age, and City_Category show weak positive correlations (~0.06) with Purchase.
- **No Multicollinearity:** Correlations between independent features are low, ensuring minimal redundancy in the dataset.
- **Impact:** Guides the selection of Product_Category as a primary feature for further analysis and predictive modeling.



Model Comparison and Metrics Insights

Models Used:

Decision Tree Regressor:

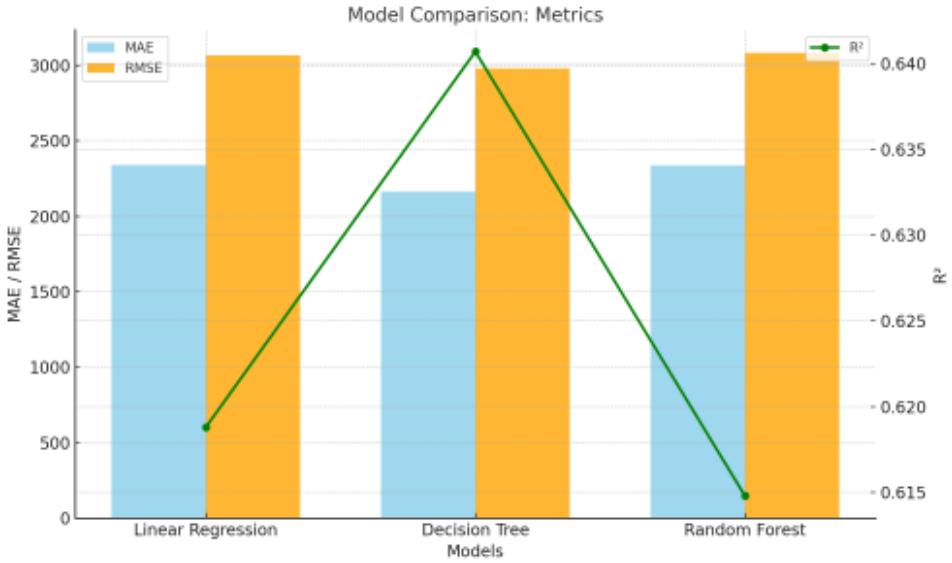
- Captures non-linear patterns effectively.
- Provides interpretable rules for understanding purchase behavior.
- Performs better in explaining variability with an R^2 of 0.64 and lower RMSE of 2979.04.

Random Forest Regressor:

- Combines multiple decision trees for more stable predictions.
- Reduces overfitting through ensemble learning.
- Slightly lower performance (R^2 of 0.61, RMSE of 3084.51) but more robust to variability.

Linear Regression :

- Simplest model for identifying linear relationships.
- Performs well for baseline predictions but struggles with complex patterns.



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

Research Questions

1.Key Factors Driving High-Value Purchases Among Walmart Customers

Applying Clustering for Customer Segmentation

Algorithm: K-Means Clustering.

Number of Clusters: Set to 3, based on domain understanding.

Preprocessing Steps:

- Dropped User_ID as it's not relevant for clustering.
- Standardized features (e.g., Purchase, Age, Occupation) using StandardScaler to ensure uniform scaling.

Variables Used for Clustering:

- Purchase (spending behavior).
- Age (grouped into categories).
- Occupation (categorical, encoded numerically).
- City_Category and Stay_In_Current_City_Years (one-hot encoded).
- Marital_Status (binary).

Visual Representation:

- Box Plot: Displaying the Purchase distribution across the three clusters.

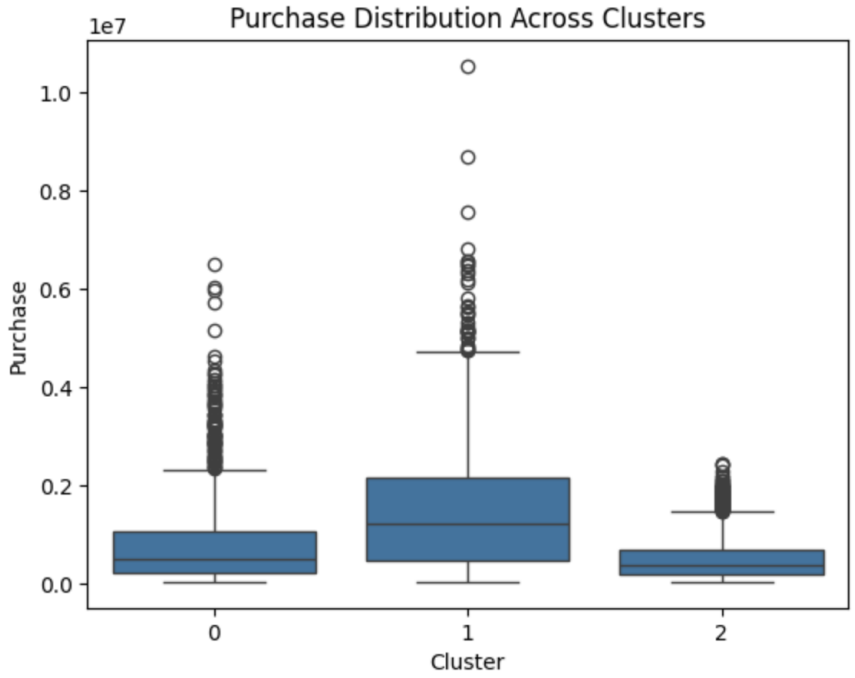
Clusters identified distinct segments of customers based on their purchasing patterns and demographic characteristics.

```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())
```

Analysis and Insights

Cluster	Median Purchase	Characteristics
Cluster 1	High	High-value spenders, wide variability
Cluster 0	Moderate	Moderate spending, some high-value outliers
Cluster 2	Low	Low spending, narrow range



Identifying Key Factors with Random Forest

Purpose: To analyze the importance of factors influencing high-value purchases.

Model: Random Forest Regressor for feature importance analysis.

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:

- Fitted the Random Forest model with preprocessed data.
- Calculated feature importance scores.
- Visualized results using a bar plot for better understanding.

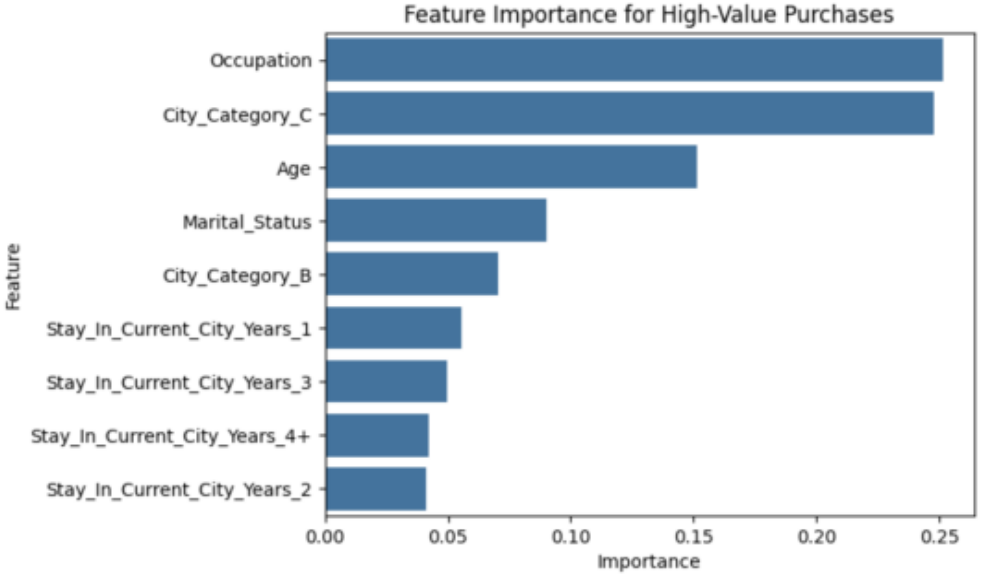
Key Findings:

- Age: Strongly associated with purchase behavior.
- Occupation: Significant driver of high-value purchases.
- City_Category: Highlights urban vs. rural differences in spending.

```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()
```

Top Key Factors Influencing High-Value Purchases:

- **Occupation**: Occupation emerges as one of the top factors affecting high-value purchases. This suggests that customers with specific professions or income levels may have different purchasing behaviors.
- **City Category:** The city or region where the customers reside plays a significant role in driving high-value purchases. Urban areas may exhibit different purchasing patterns compared to rural or suburban regions.
- **Age:** Age is also a crucial factor, with different age groups having distinct preferences and purchasing power. Younger or older customers may be more likely to make high-value purchases depending on their needs and disposable income.

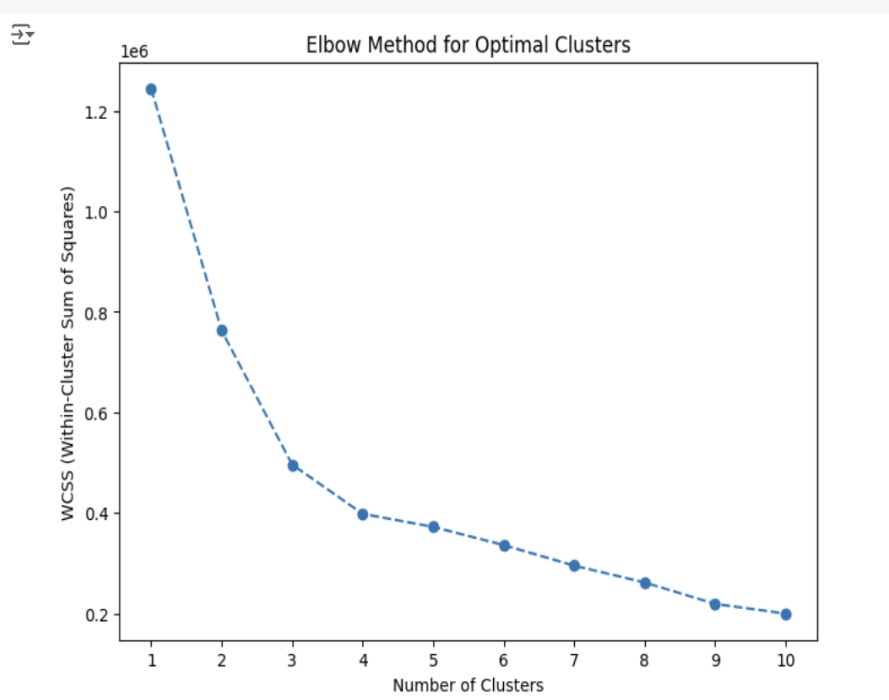


2. customer demographics (age, gender, marital status) influence purchasing behaviour at Walmart

Objective: To identify customer demographics and behavioral patterns influencing high-value purchases.

Approach:

- Preprocessed customer demographics and purchase data.
- Applied K-Means clustering to segment customers.
- Analyzed clusters to uncover spending patterns by gender, age, and marital status.



Insights from Customer Segmentation

Cluster 0 (High spenders):

- Predominantly males (81%), aged 26-35.
- Balanced marital status distribution.
- Highest total and average spending.

Cluster 1 (Regular shoppers):

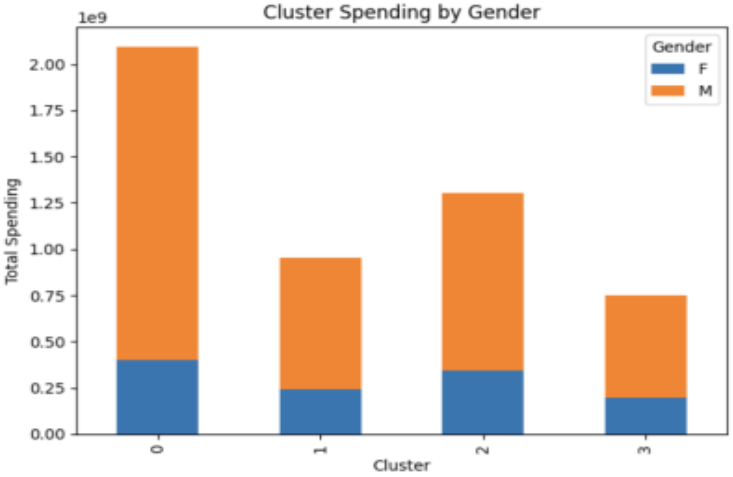
- Males aged 46-50 dominate.
- Frequent but lower-value transactions.

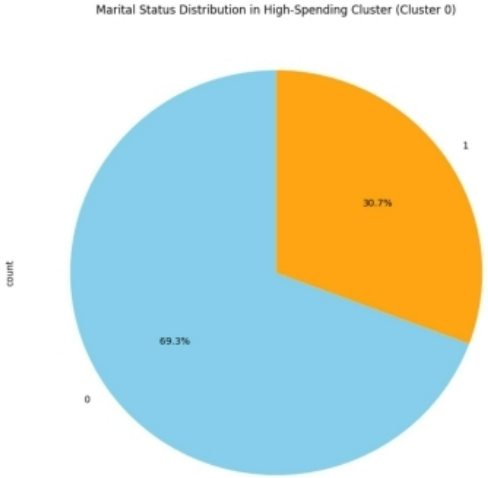
Cluster 2 (Young, budget-conscious):

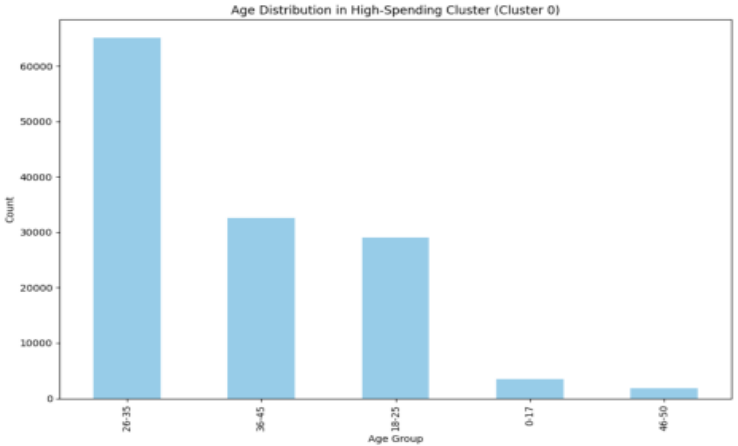
- Primarily unmarried males aged 18-25.
- Moderate spending.

Cluster 3 (Low-value shoppers):

- Balanced gender, aged 36-45.
- Lowest spending.

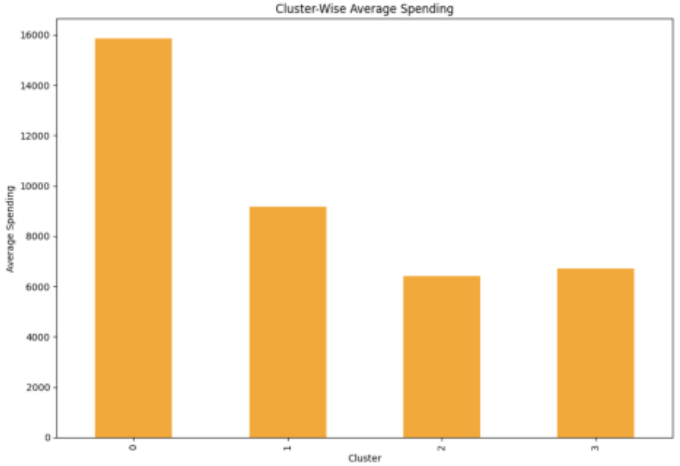






Actionable Insights: Leveraging Demographics for Purchasing Behavior

- Target age and gender-specific campaigns to maximize high-value purchases.
- Leverage insights from Cluster 0 to replicate high-spending behavior in other clusters.



3. Analyzing Walmart's Product Demand Across City Categories (A, B, C)

Objective:

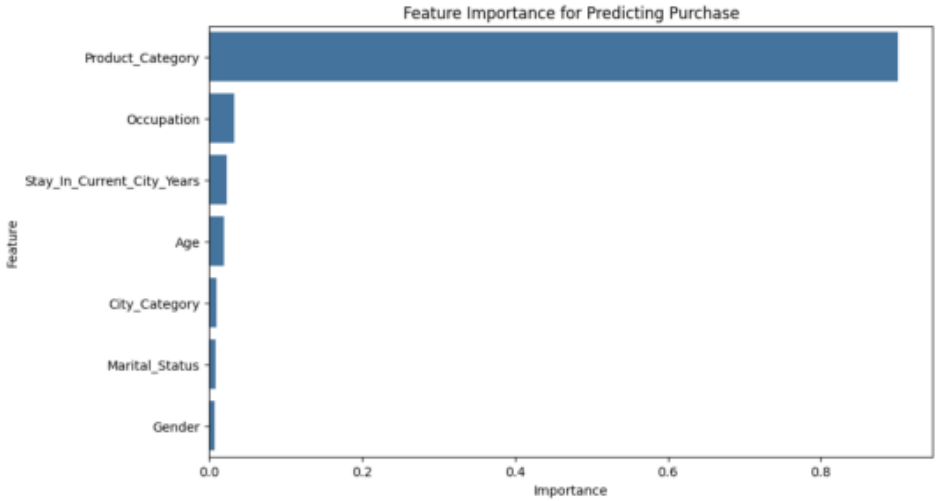
- Investigate how Walmart's product demand varies across city categories A, B, and C.

Approach:

- EDA: Explore demand trends across categories.
- ANOVA: Test statistical significance in demand variations.
- Random Forest Modeling: Identify key factors driving purchase behavior.

Why This Matters:

- Insights help Walmart adopt city-specific strategies for inventory, pricing, and marketing.



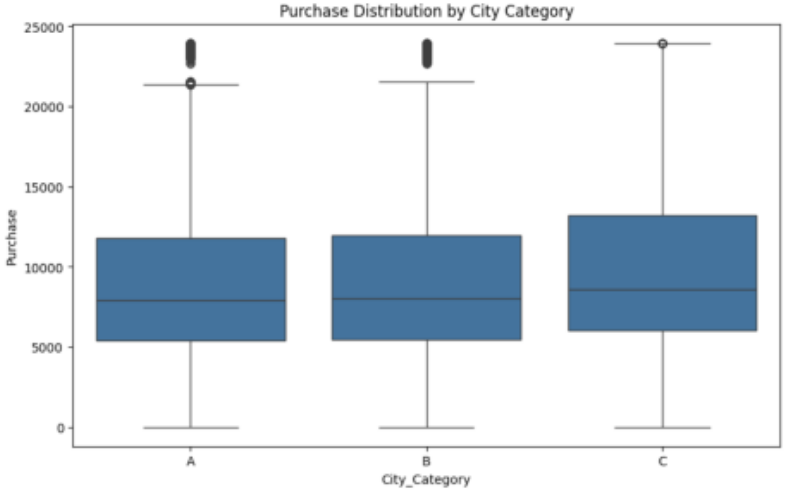
Demand Patterns and Statistical Insights

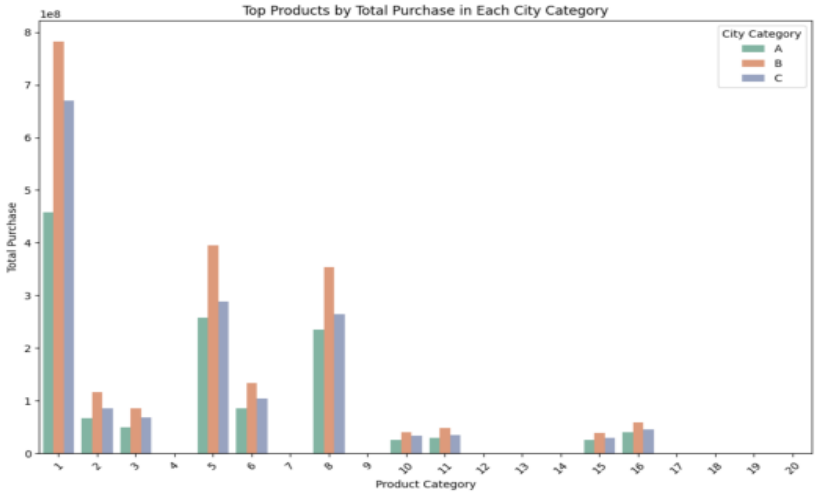
Key Observations:

- **City B**: Highest total purchases, especially in high-demand categories (1, 5, 8).
- **City C**: Moderate, stable demand; balanced consumer base.
- **City A**: Lowest total purchases; requires targeted promotions.

Statistical Testing (ANOVA):

- Confirmed significant differences in demand across city categories for most product categories.
- **P-value < 0.05**: Geographic factors heavily influence demand.






```
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
```

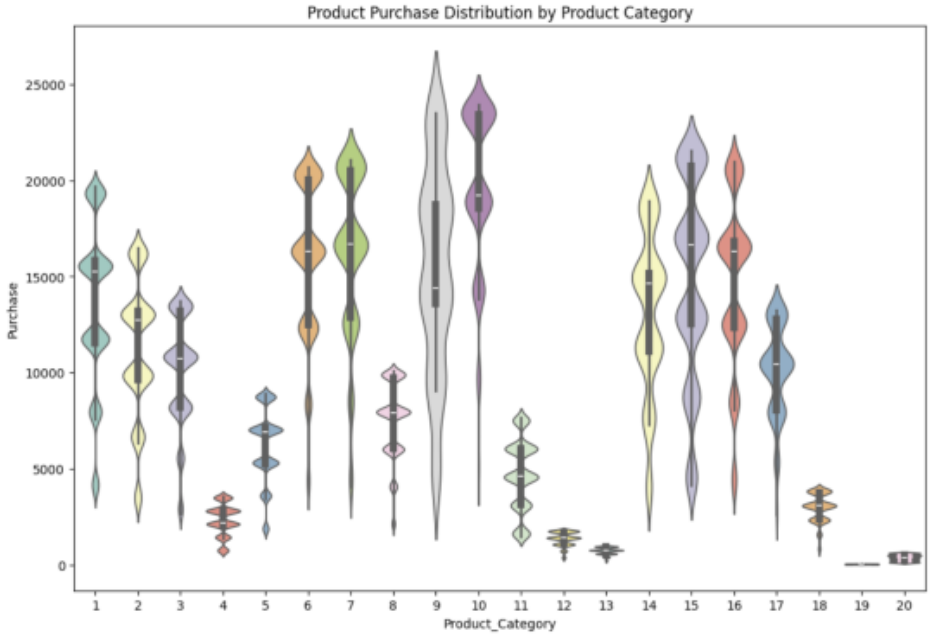
Insights and Recommendations

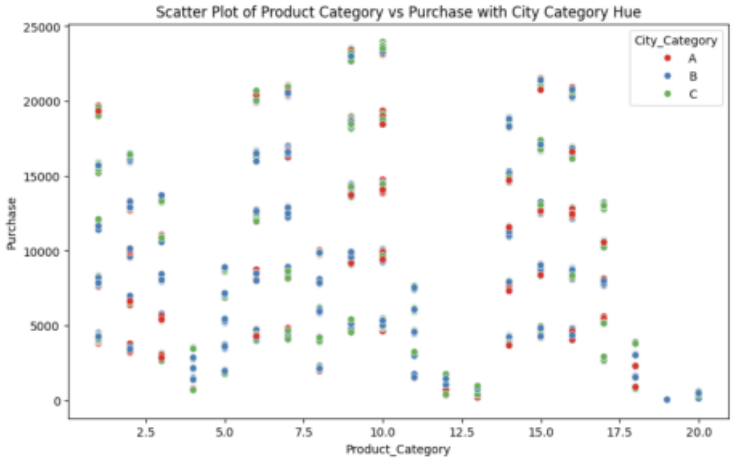
Insights:

- City B: Lucrative market; focus on high-demand categories and niche products.
- City C: Stable market; prioritize retention through loyalty programs.
- City A: Promotional efforts required to boost demand and engagement.

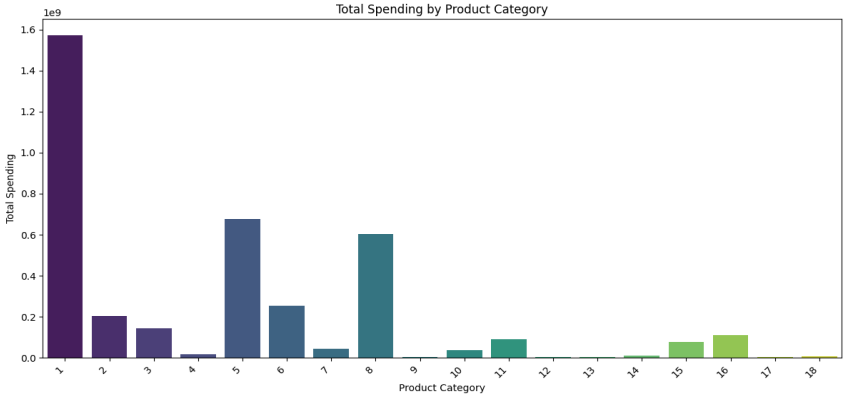
Conclusion:

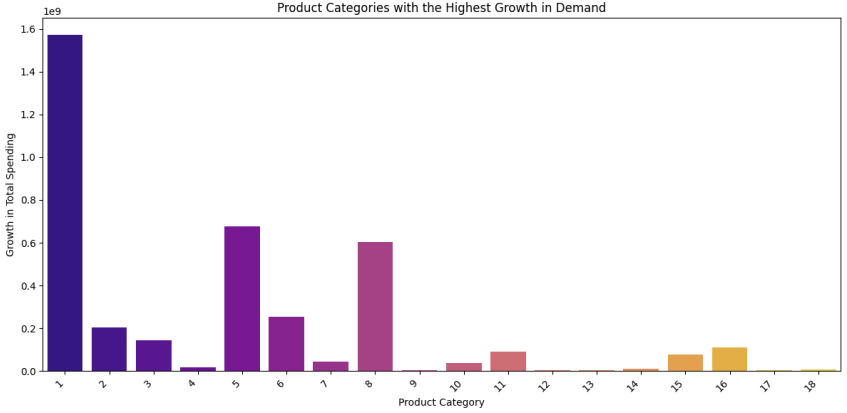
- Demand varies significantly across cities due to geographic and demographic factors.
- Tailored strategies for inventory, marketing, and pricing are essential for success.





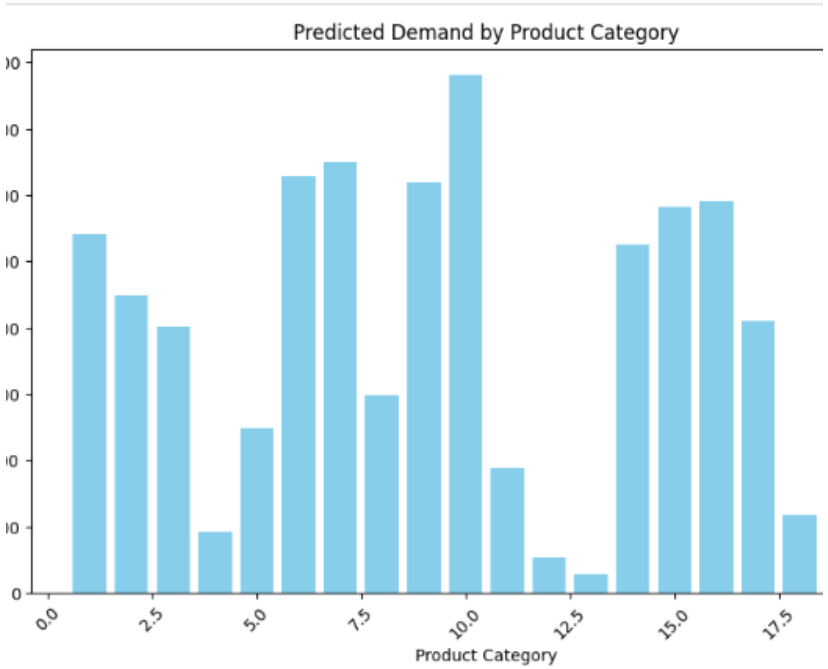
4. product categories are most likely to see an increase in demand based on current trends





- To analyze current trends, we aggregated total spending (total_spent_category) and calculated growth in spending (total_spent_growth) for each product category over time.
- This helped identify both high-performing and emerging categories. The results revealed that Product Category 1 leads in both total spending and growth, indicating its strong current popularity.
- Categories 5 and 6 also showed high spending and growth, confirming their rising demand. These findings highlight categories that are critical for immediate inventory and marketing strategies.

Forecast Future Demand

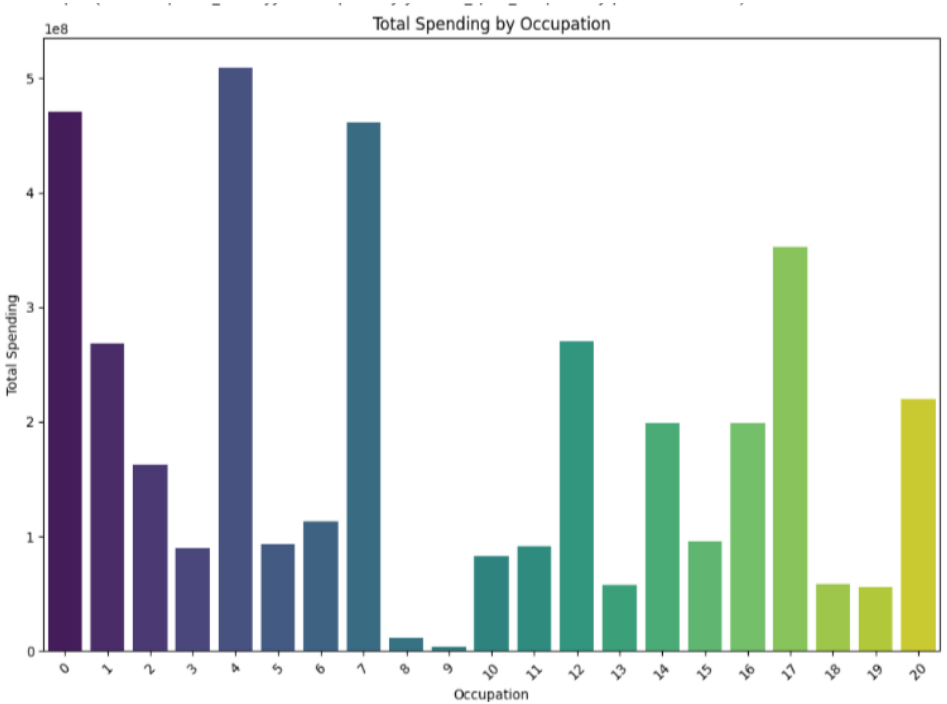


- Based on our analysis, product categories 10, 7, 5, 9, and 16 are most likely to see an increase in demand, as they exhibit the highest predicted purchase values (~19,530, ~16,328, ~15,782, ~15,495, and ~14,777, respectively).
- This result was achieved using a Random Forest Regression model trained on customer demographic and product data, including features like Product_Category, Age, Gender, and Occupation.
- The model predicted purchase amounts for the test dataset, and these predictions were aggregated by Product_Category to calculate the average predicted purchase per category.
- A bar chart was then used to visualize and identify the top categories likely to see increased demand.
- These insights suggest that focused inventory management and targeted marketing campaigns should prioritize these high-demand categories, particularly among customer groups contributing the most to their sales.

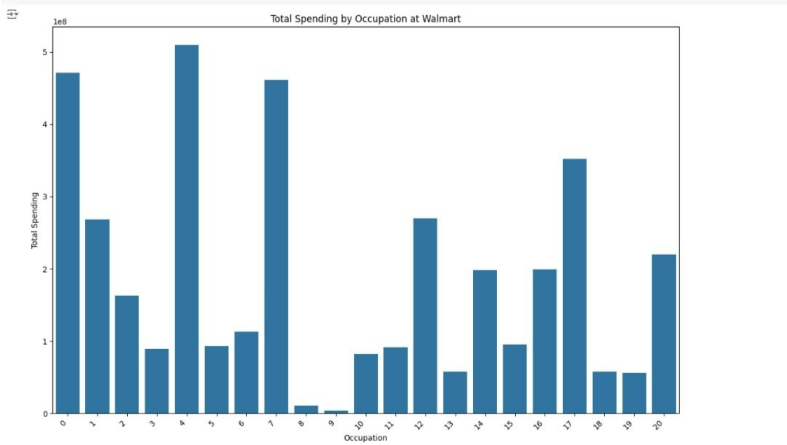
5. Role of occupation in determining spending patterns at Walmart

Occupation Summary by Spending:

	Occupation	total_spent_occupation	avg_spent_per_user	total_purchases	avg_purchases_per_user
4	4	509387740	9.948979e+05	53924	105.320312
0	0	471048528	1.019586e+06	51018	110.428571
7	7	461694565	8.678469e+05	48559	91.276316
17	17	352547530	7.994275e+05	35672	80.888889
12	12	270321872	8.191572e+05	27323	82.796970
1	1	268322538	8.545304e+05	29076	92.598726
20	20	219651445	1.120671e+06	24464	124.816327
16	16	199402091	1.072054e+06	21033	113.080645
14	14	198825797	9.204898e+05	20274	93.861111
2	2	162937795	9.698678e+05	17728	105.523810
6	6	112905069	8.752331e+05	12040	93.333333
15	15	95338052	8.512326e+05	9646	86.125000
5	5	93334364	1.166680e+06	9869	123.362500
11	11	91472239	8.629457e+05	9933	93.707547
3	3	89321700	1.240579e+06	9643	133.930556
10	10	82411042	6.540559e+05	8818	69.984127
18	18	58066671	9.216932e+05	6331	100.492063
13	13	57950154	5.415902e+05	6115	57.149533
19	19	55981580	9.996711e+05	6382	113.964286
8	8	11286549	8.061821e+05	1175	83.928571
9	9	4110299	1.370100e+06	445	148.333333



- The analysis grouped user data by occupation to calculate total and average spending, as well as purchase counts, revealing spending trends across occupations.
- Occupation 4, 0, and 7 emerged as top spenders. Further grouping by Product_Category identified category-specific preferences, showing that occupation significantly influences purchasing behavior.
- Visualizations highlighted these trends, confirming the impact of occupation on spending patterns at Walmart.



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

Results

Demographics Driving Purchases:

- City Category: Urban areas (City Category B) have the highest total purchases.
- Age: Middle-aged (26–35) and young adults (18–25) are significant contributors to high-value purchases.
- Marital Status: Married customers show a tendency for higher spending.

Product Demand Insights:

- High-demand categories include 1, 5, and 8, dominating customer preferences across all city categories.
- Categories 10 and 7 are emerging as growth segments.

Groundbreaking Findings:

- Decision Tree Regressor emerged as the most effective model with the highest R^2 (0.6407) and lowest prediction error (RMSE: 2979.04, MAE: 2163.25).
- Clustering revealed 4 distinct customer segments, with Cluster 0 (young, high-spending males aged 26–35) as the key demographic.

Conclusion

Conclusion:

- Walmart's customer purchases are significantly influenced by age, gender, and city category.
- Product demand varies geographically, requiring city-specific strategies for marketing and inventory.
- The Decision Tree Regressor is the most reliable model for predicting purchase behavior, effectively capturing non-linear relationships in the data.

Closing Note:

- This analysis provides actionable insights into Walmart's customer behavior, enabling better decision-making for **Targeted marketing, Inventory management, Customer satisfaction improvement.**

Recommendations

Recommendations:

- High-Value Customers: Focus on premium loyalty programs for high-spending males (26–35).
- Emerging Categories: Prioritize inventory for categories 1, 5, and 10 to capture growth opportunities.
- City-Specific Strategies:
- Urban regions (City B): Tailor campaigns for top-performing categories.
- City A: Deploy aggressive promotions to boost sales.