Customer Purchase Behavior

Do women spend more on Black Friday than men?

Jessica Pham

Introduction

This study aims to explore customer purchase behavior during Black Friday at Walmart Inc., focusing on demographic factors such as gender, age, city category, and occupation. Understanding these patterns is crucial for identifying key trends and differences in spending habits among various customer segments. By analyzing the data, the study seeks to uncover insights that can guide targeted marketing strategies and enhance overall business performance. The findings will provide a detailed examination of how different groups of customers engage with the brand, highlighting areas for potential growth and optimization.

Content

- Exploratory Data Analysis
  - Data Description
  - Data Inspection
  - Statistical Summary
  - Data Cleaning
- Data Visualization
  - Analysis of Purchase Behavior by Gender and Other Features
  - Key Insights
- Hypothesis Testing
  - T Test
- Confidence Intervals Analysis
  - Overall Confidence Interval for Purchase Amounts
  - Group-Specific Confidence Intervals
  - Key Insights
- Conclusions
- Recommendations

Programming Language: Python

Visualization Tool: Power BI

Statistical Techniques: T-Test & Confidence Interval

GitHub Repository: View here

Source of Data: Kaggle - Link to the Dataset

Exploratory Data Analysis (EDA)

Analysis of Purchase Behaviour by Gender and Other Features

Data Description

The company collected the transactional data of customers who purchased products from Walmart Stores during Black Friday.

- User_ID: Unique identifier for each user.
- Product_ID: Unique identifier for each product.
- Gender: Gender of the user.
- Age: Age group of the user (e.g., "0-17", "55+").
- Occupation: User's occupation represented by a numerical code.
- City_Category: Category of the city (e.g., "A", "B", "C").
- Stay_In_Current_City_Years: Duration of stay in the current city (e.g., "2", "4+").
- Marital_Status: Marital status of the user (0 = Unmarried, 1 = Married).
- Product_Category: Category of the product represented by a numerical code.
- Purchase: Purchase amount.

Data Inspection

The first few rows & the last few rows

	User_ID	Product_ID	Gender	Age	Occupation	City_Category	Stay_In_Current_City_Years	Marital_Status	Product_Category	Purchase
0	1000001	P00069042	F	0-17	10	A	2	0	3	8370
1	1000001	P00248942	F	0-17	10	A	2	0	1	15200
2	1000001	P00087842	F	0-17	10	A	2	0	12	1422
3	1000001	P00085442	F	0-17	10	A	2	0	12	1057
4	1000002	P00285442	M	55+	16	C	4+	0	8	7969

	User_ID	Product_ID	Gender	Age	Occupation	City_Category	Stay_In_Current_City_Years	Marital_Status	Product_Category	Purchase
550063	1006033	P00372445	M	51-55	13	B	1	1	20	368
550064	1006035	P00375436	F	26-35	1	C	3	0	20	371
550065	1006036	P00375436	F	26-35	15	B	4+	1	20	137
550066	1006038	P00375436	F	55+	1	C	2	0	20	365
550067	1006039	P00371644	F	46-50	0	B	4+	1	20	490

Unique Values

Unique Values of User_ID: 5891

Unique Values of Product_ID: 3631

Unique Values of Gender: 2

Unique Values of Age: 7

Unique Values of Occupation: 21

Unique Values of City_Category: 3

Unique Values of Stay_In_Current_City_Years: 5

Unique Values of Marital_Status: 2

Unique Values of Product_Category: 20

Unique Values of Purchase: 18105

Summary Statistics

```
<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)
```

- There are no missing values in the dataset.
- The minimum purchase amount recorded is $12.
- The maximum purchase amount recorded is $23,961.
- The majority of customers are aged 26-35, more than any other age group.
- The total number of purchases is highest in City B.
- The dataset includes 414,259 male customers, with the remainder being female.

index	User_ID	Product_ID	Gender	Age	Occupation	City_Category	Stay_In_Current_City_Years	Marital_Status	Product_Category	Purchase
count	550068.0	550068	550068	550068	550068.0	550068	550068.0	550068.0	550068.0	550068.0
unique	NaN	3631	2	7	NaN	3	NaN	NaN	NaN	NaN
top	NaN	P00265242	M	26-35	NaN	B	NaN	NaN	NaN	NaN
freq	NaN	1880	414259	219587	NaN	231173	NaN	NaN	NaN	NaN
mean	1003028.8424013031	NaN	NaN	NaN	8.076706879876669	NaN	1.8584175047448679	0.40965298835780306	5.404270017525106	9263.968712959126
std	1727.5915855305516	NaN	NaN	NaN	6.522660487341824	NaN	1.2894425553200026	0.49177012631733	3.936211369201389	5023.065393820582
min	1000001.0	NaN	NaN	NaN	0.0	NaN	0.0	0.0	1.0	12.0
25%	1001516.0	NaN	NaN	NaN	2.0	NaN	1.0	0.0	1.0	5823.0
50%	1003077.0	NaN	NaN	NaN	7.0	NaN	2.0	0.0	5.0	8047.0
75%	1004478.0	NaN	NaN	NaN	14.0	NaN	3.0	1.0	8.0	12054.0
max	1006040.0	NaN	NaN	NaN	20.0	NaN	4.0	1.0	20.0	23961.0

Data Cleaning

There is no duplicate value in the dataset

```
# Checking duplicate values in the dataset
df.duplicated(subset=None,keep='first').sum()
```

```
0
```

The 'Stay_In_Current_City_Years' column: removing the '+' symbol and converting it to a numeric format

```
#Removing the '+' symbol and converting it to a numeric format for 'Stay_In_Current_City_Years' column
df.Stay_In_Current_City_Years.unique()
```

```
array(['2', '4+', '3', '1', '0'], dtype=object)
```

```
# Removing "+" symbol
df.Stay_In_Current_City_Years=df.Stay_In_Current_City_Years.str.replace("+","")
```

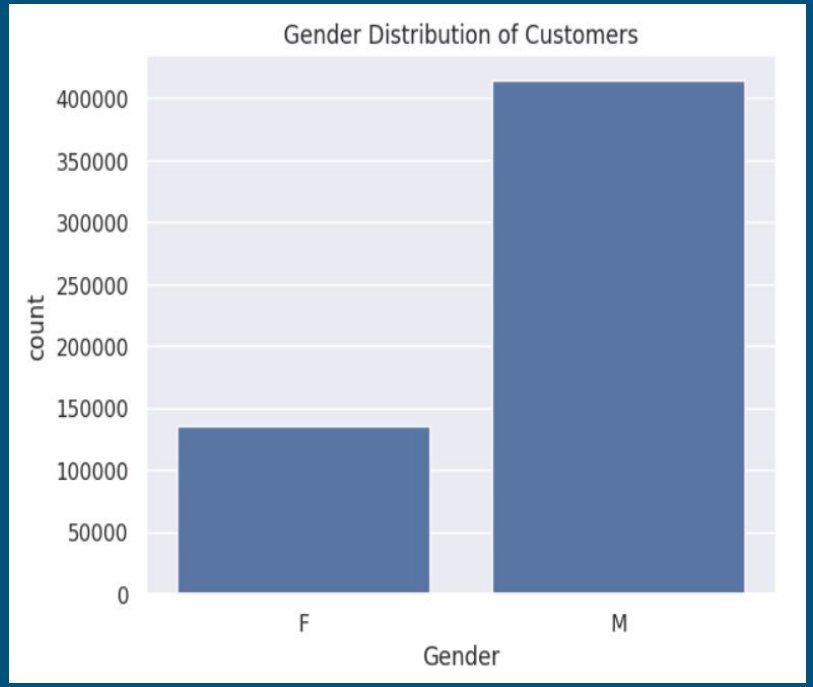
```
#checking after removing
df.Stay_In_Current_City_Years.unique()
```

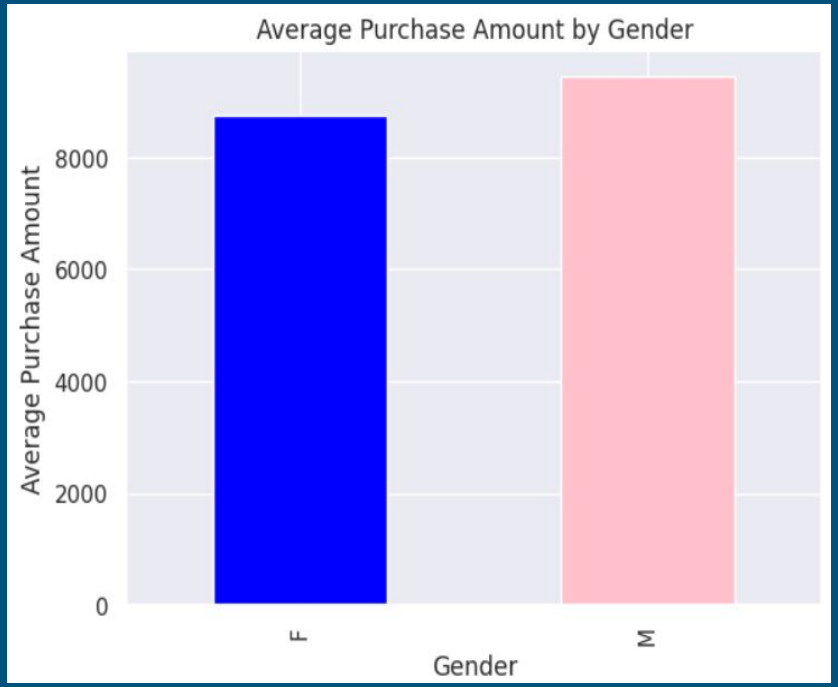
```
array(['2', '4', '3', '1', '0'], dtype=object)
```

Visualization

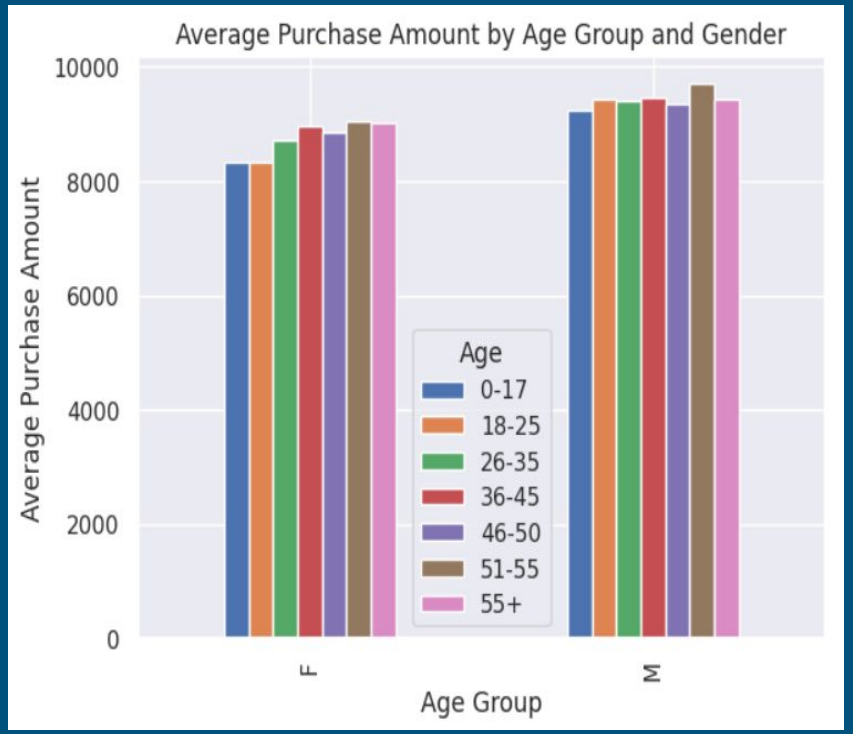
Analysis of Purchase Behaviour by Gender and Other Features

Spending by Gender

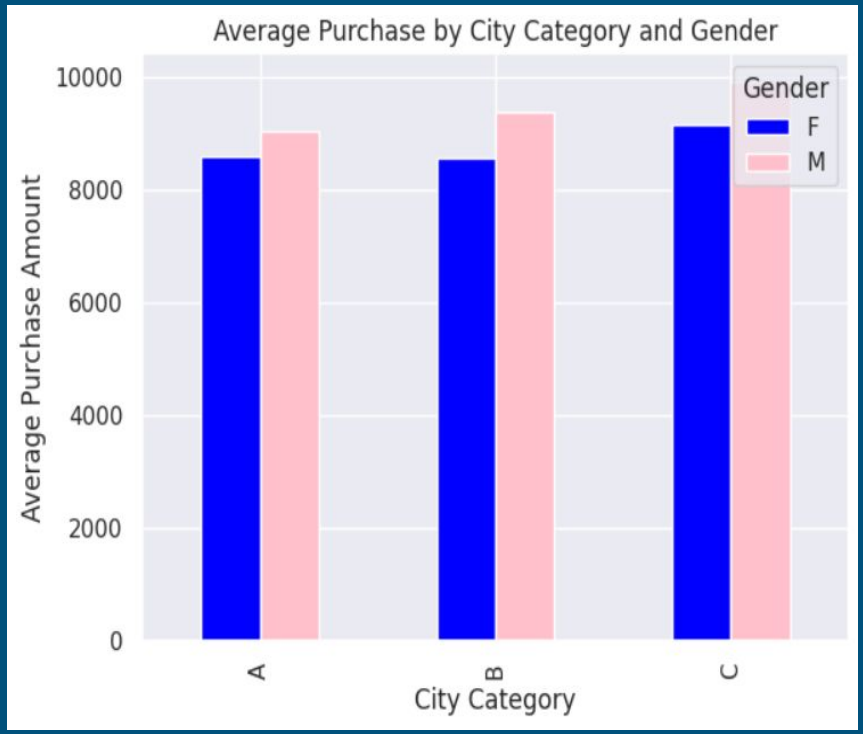




Spending by Gender & Age

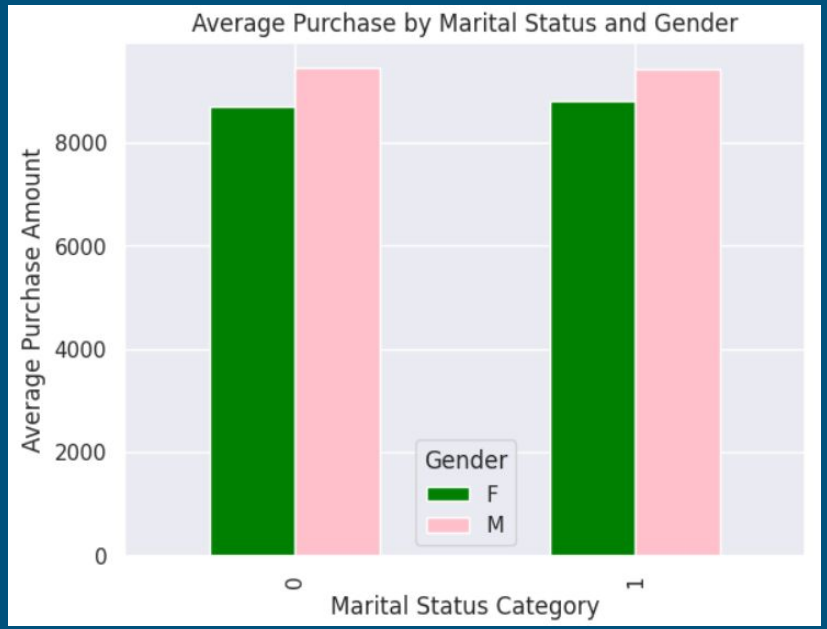


Spending by Gender & City

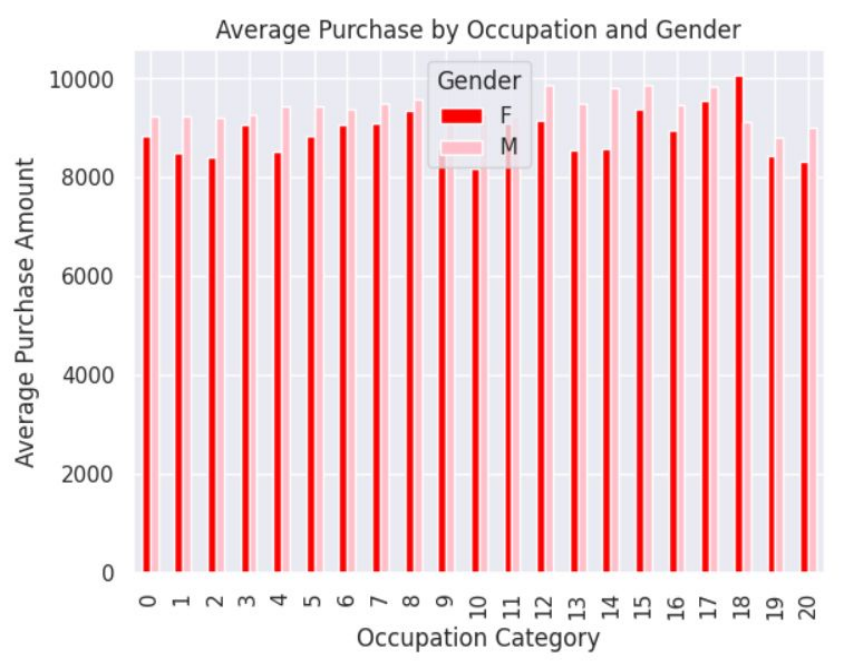


Spending by Gender & Stay_Years

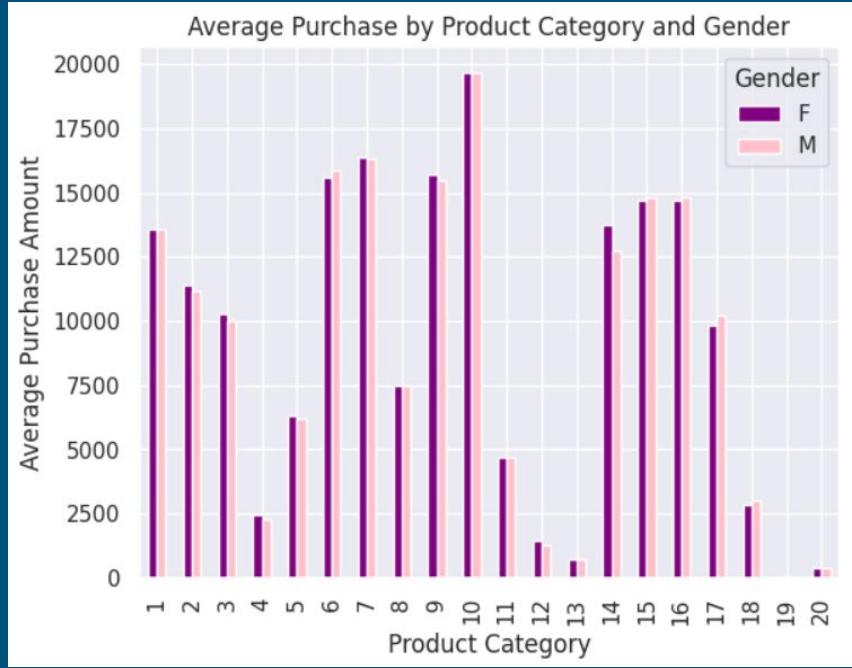
Spending by Gender & Marital Status



Spending by Gender & Occupation



Spending by Gender & Product



Key Insights

- Males generally spend more on purchases than females.
- The male age group of 51-55 years old has the highest average spending on Black Friday.
- Across all age groups, males outspend females.
- The average purchase amount in City C is higher than in Cities A and B for both males and females.
- In terms of average spending by gender across marital status, years spent in the current city, occupation and product category, males consistently spend more than females.
- An exception is observed in occupation category 18, where females slightly outspend males.
- Customers spend more on products 6, 7, 9, 10 than others.

Hypothesis Testing

T Test

Perform an independent t-test to see if the difference in spending between male and female customers is statistically significant.

Result

```
T-statistic: 44.837957934353966
P-value: 0.0
```

Statistical Insights

- T-statistic: 44.84: This value indicates the magnitude of difference between the means of male and female purchase amounts, relative to the variability of the data.
- P-value: 0.0: The p-value being 0.0 suggests that the difference in spending between male and female customers is highly statistically significant. This means that there's a very strong indication that men and women have different average purchase amounts, with this difference not being due to random chance.

Confidence Interval Analysis

Overall Confidence Interval for Purchase Amounts

Result `(9263.968712959126, 9250.694472258305, 9277.242953659947)`

The mean purchase amount across the entire dataset is approximately 9,264. The 95% confidence interval for the mean purchase amount is:

- Lower Bound: 9,251
- Upper Bound: 9,277

This means we can be 95% confident that the true mean purchase amount lies between 9,251 and 9,277.

Group-Specific Confidence Intervals

Calculate the confidence interval for the mean purchase amounts based on different groups, such as by Gender, Age, and City_Category

Here are the 95% confidence intervals for the mean purchase amounts across different groups:

1. Gender

- Female (F):
  - Mean Purchase: 8,735
  - Confidence Interval: [8,709, 8,760]
- Male (M):
  - Mean Purchase: 9,438
  - Confidence Interval: [9,422, 9,453]

2. City Category

- City A:
  - Mean Purchase: 8,912
  - Confidence Interval: [8,887, 8,937]
- City B:
  - Mean Purchase: 9,151
  - Confidence Interval: [9,131, 9,172]
- City C:
  - Mean Purchase: 9,720
  - Confidence Interval: [9,695, 9,745]

Confidence Intervals by Gender:

Gender	mean	ci_lower	ci_upper
F	8734.565765	8709.211321	8759.920209
M	9437.526040	9422.019402	9453.032679

Confidence Intervals by Age:

Age	mean	ci_lower	ci_upper
0-17	8933.464640	8851.941436	9014.987845
18-25	9169.663606	9138.407569	9200.919643
26-35	9252.690633	9231.733561	9273.647705
36-45	9331.350695	9301.669084	9361.032305
46-50	9208.625697	9163.083936	9254.167458
51-55	9534.808031	9483.989875	9585.626187
55+	9336.280459	9269.295064	9403.265855

Confidence Intervals by City_Category:

City_Category	mean	ci_lower	ci_upper
A	8911.939216	8886.991621	8936.886811
B	9151.300563	9131.099743	9171.501382
C	9719.920993	9695.336934	9744.505052

Group-Specific Confidence Intervals

3. Age

- 0-17:
  - Mean Purchase: 8,933
  - Confidence Interval: [8,852, 9,015]
- 18-25:
  - Mean Purchase: 9,170
  - Confidence Interval: [9,138, 9,201]
- 26-35:
  - Mean Purchase: 9,253
  - Confidence Interval: [9,232, 9,274]
- 36-45:
  - Mean Purchase: 9,331
  - Confidence Interval: [9,302, 9,361]
- 46-50:
  - Mean Purchase: 9,209
  - Confidence Interval: [9,163, 9,254]
- 51-55:
  - Mean Purchase: 9,535
  - Confidence Interval: [9,484, 9,586]
- 55+:
  - Mean Purchase: 9,336
  - Confidence Interval: [9,269, 9,403]

Key Insights

Gender: Males tend to have a higher mean purchase amount than females.

Age: Purchase amounts generally increase with age, peaking in the 51-55 age group.

City Category: Customers in City C have the highest mean purchase amounts compared to those in Cities A and B.

Conclusions & Recommendations

Conclusion

- The analysis reveals distinct spending patterns across various demographic groups. Males consistently outspend females across most age groups, city categories, and other demographic factors. Notably, the male age group of 51-55 years old exhibits the highest average spending on Black Friday.

- Customers in City C demonstrate higher average purchase amounts than those in Cities A and B, regardless of gender.

- Additionally, while males generally spend more, females in occupation category 18 slightly outspend their male counterparts. The statistical significance of these findings is underscored by the p-value of 0.0, confirming that the observed differences in spending between male and female customers are not due to random chance.

Recommendations

Targeted Marketing Strategies:

- For Males: Given that males, particularly those aged 51-55, spend more on average, tailor marketing campaigns to appeal to this demographic. Consider promotions that resonate with their interests and preferences, especially during peak spending periods like Black Friday.
- For Females in Occupation Category 18: Explore why females in this occupation outspend males and develop targeted campaigns that could be expanded to other similar customer segments.

City-Specific Promotions:

- City C: Leverage the higher spending power in City C by offering exclusive deals or premium products to capitalize on this market's potential. Consider city-specific advertising to further engage this lucrative demographic.

Recommendations

Product Focus:

- Top-Performing Products: Since products 6, 7, 9, and 10 are top sellers, prioritize these in marketing and inventory strategies. Ensure these products are well-stocked and featured prominently in promotions.

Personalized Offers Based on Age:

- Since purchase amounts generally increase with age, especially peaking in the 51-55 age group, consider personalized offers or loyalty programs that cater to the needs and preferences of older customers.