

Analyzing Customer Purchase Behavior on Black Friday using Excel

Introduction

Black Friday, an annual shopping event occurs after Thanksgiving has evolved into a global retail phenomenon. This day marks the start of Christmas shopping season where ecommerce companies offer high promoted sales at discounted prices. It is also the busiest shopping day of the year. Businesses prepare for this day to boost sales and attract customers.

Scenario

As a data analyst working in the marketing analytics team at ALLMart Company, your role is important in interpreting patterns of customer purchase behavior during black Friday. You are tasked with finding the preferences and tendencies of our diverse customer base. Your goal is to discover insights that will not only enhance our understanding of black Friday dynamics but also discover insights that will help ALLMart tailor its offerings and services to meet the unique needs of our customers.

Data Collection

The dataset was gotten from [Kaggle](#) and, it contains the following information:

- User_ID: Unique identifier for each customer.
- Product_ID: Unique identifier for each product.
- Gender: Gender of the customer.
- Age: Age of the customer in bins.
- Occupation: Occupation of the customer (masked).
- City_Category: Category of the City(A,B,C).
- StayInCurrentCityYears: number of years of stay the customer has been living in their current city.
- Marital_Status: Marital status of the customer.
- Product_Category: Category of product purchased by the customer (masked).
- Purchase: Purchase Amount in USD

This analysis is divided into the following stages:

Data cleaning, Data Exploration & Analysis, Insights, Recommendations, and Visualization.

Data Processing & Cleaning

Loading the data in Excel;

User_ID	Product_ID	Gender	Age	Occupation	City_Category	Stay_In_Current_City_Years	Marital_Status	Product_Category	Purchase
1000001	P00069042	F	0-17	10 A		2	0	3	8370
1000001	P00248942	F	0-17	10 A		2	0	1	15200
1000001	P00087842	F	0-17	10 A		2	0	12	1422
1000001	P00085442	F	0-17	10 A		2	0	12	1057
1000002	P00285442	M	55+	16 C		4	0	8	7969
1000003	P00193542	M	26-35	15 A		3	0	1	15227
1000004	P00184942	M	46-50	7 B		2	1	1	19215
1000004	P00346142	M	46-50	7 B		2	1	1	15854
1000004	P0097242	M	46-50	7 B		2	1	1	15686
1000005	P00274942	M	26-35	20 A		1	1	8	7871
1000005	P00251242	M	26-35	20 A		1	1	5	5254
1000005	P00014542	M	26-35	20 A		1	1	8	3957
1000005	P00031342	M	26-35	20 A		1	1	8	6073
1000005	P00145042	M	26-35	20 A		1	1	1	15665
1000006	P00231342	F	51-55	9 A		1	0	5	5378
1000006	P00190242	F	51-55	9 A		1	0	4	2079
1000006	P0096642	F	51-55	9 A		1	0	2	13055
1000006	P00058442	F	51-55	9 A		1	0	5	8851
1000007	P00036842	M	36-45	1 B		1	1	1	11788
1000008	P00249542	M	26-35	12 C		4	1	1	19614
1000008	P00220442	M	26-35	12 C		4	1	5	8584

The dataset contains 10 columns and 550069 rows. The data looks good, I'll need to work on Gender, Age and Marital_Status table so that they can be useful for analysis.

1. I used Removed duplicates to check for duplicates, the data did not contain duplicates.

User_ID	Product_ID	Gender	Age	Occupation	City_Category	Stay_In_Current_City_Years	Marital_Status	Product_Category	Purchase
1000001	P00069042	F	0-17	10 A		2	0	3	8370
1000001	P00248942	F	0-17	10 A		2	0	1	15200
1000001	P00087842	F	0-17	10 A		2	0	12	1422
1000001	P00085442	F	0-17	10 A		2	0	12	1057
1000002	P00285442	M	55+	16 C		4	0	8	7969
1000003	P00193542	M	26-35	15 A		3	0	1	15227
1000004	P00184942	M	46-50	7 B		2	1	1	19215
1000004	P00346142	M	46-50	7 B		2	1	1	15854
1000004	P0097242	M	46-50	7 B		2	1	1	15686
1000005	P00274942	M	26-35	20 A		1	1	8	7871
1000005	P00251242	M	26-35	20 A		1	1	5	5254
1000005	P00014542	M	26-35	20 A		1	1	8	3957
1000005	P00031342	M	26-35	20 A		1	1	8	6073
1000005	P00145042	M	26-35	20 A		1	1	1	15665
1000006	P00231342	F	51-55	9 A		1	0	5	5378
1000006	P00190242	F	51-55	9 A		1	0	4	2079
1000006	P0096642	F	51-55	9 A		1	0	2	13055
1000006	P00058442	F	51-55	9 A		1	0	5	8851
1000007	P00036842	M	36-45	1 B		1	1	1	11788
1000008	P00249542	M	26-35	12 C		4	1	1	19614
1000008	P00220442	M	26-35	12 C		4	1	5	8584
1000008	P00156442	M	26-35	12 C		4	1	8	9872
1000008	P00213742	M	26-35	12 C		4	1	8	9743

2. I checked for blanks, the data did not contain blanks

The screenshot shows an Excel spreadsheet titled 'Black Friday Sales.xlsx'. The data table has the following columns: User_ID, Product_ID, Gender, Age, Occupation, City_Category, Stay_In_Current_City_Years, Marital_Status, Product_Category, and Purchase. A dialog box with a yellow warning icon and the text 'No cells were found.' is centered over the data. The status bar at the bottom indicates 'Ready' and 'Accessibility: Investigate'.

Its time to verify the data and make sure its in the right format for analysis. I used IF Statements to perform the following:

- Change 'M' to Male and 'F' to Female from the Gender to the Sex column so it can be understood in data visualization

The screenshot shows the same Excel spreadsheet, but now with a new column 'Sex' added. The formula bar shows the formula: `=IF(C2="F","Female","Male")`. The data table now includes the 'Sex' column, with 'F' converted to 'Female' and 'M' converted to 'Male'. The status bar at the bottom indicates 'Ready' and 'Accessibility: Investigate'.

- Changed Marital_Status '0' to Single and '1' to Married in the MaritalStatus column for the same reason

	A	B	C	D	E	F	G	H	I	J	K	L	M	N
	User_ID	Product_ID	Gender	Sex	Age	Occupation	City_Category	Stay_In_Current_City_Years	Marital_Status	MaritalStatus	Product_Category	Purchase		
2	1000001	P00069042	F	Female	0-17	10 A		2	0	Single	3	8370		
3	1000001	P00248942	F	Female	0-17	10 A		2	0	Single	1	15200		
4	1000001	P00087842	F	Female	0-17	10 A		2	0	Single	12	1422		
5	1000001	P00085442	F	Female	0-17	10 A		2	0	Single	12	1057		
6	1000002	P00285442	M	Male	55+	16 C		4	0	Single	8	7969		
7	1000003	P00193542	M	Male	26-35	15 A		3	0	Single	1	15227		
8	1000004	P00184942	M	Male	46-50	7 B		2	1	Married	1	19215		
9	1000004	P00346142	M	Male	46-50	7 B		2	1	Married	1	15854		
10	1000004	P0097242	M	Male	46-50	7 B		2	1	Married	1	15686		
11	1000005	P00274942	M	Male	26-35	20 A		1	1	Married	8	7871		
12	1000005	P00251242	M	Male	26-35	20 A		1	1	Married	5	5254		
13	1000005	P00014542	M	Male	26-35	20 A		1	1	Married	8	3957		
14	1000005	P00031342	M	Male	26-35	20 A		1	1	Married	8	6073		
15	1000005	P00145042	M	Male	26-35	20 A		1	1	Married	1	15665		
16	1000006	P00231342	F	Female	51-55	9 A		1	0	Single	5	5378		
17	1000006	P00190242	F	Female	51-55	9 A		1	0	Single	4	2079		
18	1000006	P0096642	F	Female	51-55	9 A		1	0	Single	2	13055		
19	1000006	P00058442	F	Female	51-55	9 A		1	0	Single	5	8851		
20	1000007	P00036842	M	Male	36-45	1 B		1	1	Married	1	11788		
21	1000008	P00249542	M	Male	26-35	12 C		4	1	Married	1	19614		
22	1000008	P00220442	M	Male	26-35	12 C		4	1	Married	5	8584		

- I created a new column called Age_Group from the Age column by replacing the age groups to Children & Teenagers, Young Adults, Adults, Middle Aged Adults, Older Adults, Young Seniors and Seniors.

Analysis

These are the questions that will guide my analysis:

1. What is the average amount spent by each customer?
2. What's the total amount spent by customers?
3. What Gender bought more on black Friday?
4. What is the average purchase amount for different age group and how does it relate to overall average customer purchase?
5. How does marital status affect purchase?
6. What is the average purchase amount for high-value products on black Friday?
7. Is there any relationship between Stay_In_Current_City_Years and purchase?
8. How does City_Category influence customer purchase?

Pivot tables was used for analysis:

1. The average amount spent by each customer is \$9264.

	A
1	Average of Purchase
2	9264
3	

2. The total amount spent by customers was \$5,095,812,742.

3	
4	Sum of Purchase
5	5095812742
6	

3. Males bought more on black Friday accounting for 80% of total purchase. This suggests that gender plays an important role in purchases.

Row Labels	Sum of Purchase
Female	20%
Male	80%
Grand Total	100.00%

4. Average Purchase gradually increases across age groups with children spending the least at \$8933 and young seniors spending the most at \$9535.
Interestingly, all age groups except Middle Aged Adults, Seniors and Young Seniors spend below the overall purchase average.

Row Labels	Average of Purchase
Children & Teenagers	8933
Young Adults	9170
Older Adults	9209
Adults	9253
Middle Aged Adults	9331
Seniors	9336
Young Seniors	9535
Overall Average	9264

5. Single customers spent more than married customers accounting for 59% of purchase.

Row Labels	Sum of Purchase	Sum of Purchase2
Married	1193715426	41%
Single	1741052221	59%
Grand Total	2934767647	100.00%

6. The criteria I used to get high-value products are the top 10 products by purchase. I filtered the top 10 products with the highest average and I used AVERAGE formula to find the average.

Row Labels	Average of Purchase
17	10171
2	11252
14	13142
1	13606
16	14766
15	14780
9	15537
6	15838
7	16366
10	19676
Grand Total	13831 AVERAGE 14513

The average purchase amount for high-value products is \$14,513.

7. The longer a customer stays in their current city in Years the less they spend on black Friday , except customers who spent less than a year who had the least purchase and customers who spent a year had the highest purchase.

Row Labels	Sum of Purchase
0	388,692,831
1	1,028,199,602
2	552,155,121
3	514,487,719
4	451,232,374
Grand Total	2,934,767,647

8. To determine how City_Category influence customer purchase, I started by finding the number of customers in different cities

The screenshot displays the Microsoft Excel interface. The top ribbon shows the 'Formulas' tab with the 'AutoSum' button. Below the ribbon, the 'PivotTable Fields' task pane is open on the right side. The task pane shows a list of fields: 'User_ID', 'Product_ID', 'Gender', 'Age_Group', 'Occupation', 'City_Category', 'Stay_In_Current_City_Years', and 'Marital_Status'. The 'City_Category' field is selected and placed in the 'Rows' area. The 'User_ID' field is selected and placed in the 'Values' area. The 'Count of User_ID' is shown in the 'Values' area. The 'City_Category' field is also shown in the 'Rows' area. The 'Defer Layout Update' checkbox is checked at the bottom of the task pane.

Row Labels	Count of User_ID
A	54,565
C	68,984
B	88,634
Grand Total	212,183

city B had the highest population of customers. This city is likely a metropolitan area with high economic development.

city C had a smaller population as compared to city B. This city is likely an urban area.

city A had the least population of customers. This city is likely a small city or town with less economic development and infrastructure compared to other cities.

I then found the sum of purchase by city

Row Labels	Sum of Purchase
A	730,736,234
C	993,132,033
B	1,210,899,380
Grand Total	2,934,767,647

Customers from city A had the least amount of purchase followed by customers in city C and customers from city B had the highest overall purchase.

This suggests that city category influences customer purchase, this may be attributed to factors such as closeness to warehouse or competition.

Insights

1. The number of customers that participated in this edition of black Friday was 550,069.
2. The average amount spent by each customer is \$9264.
3. The total amount spent by customers was \$5,095,812,742.
4. Average Purchase gradually increases across age groups with children spending the least at \$8933 and young seniors spending the most at \$9535.
Interestingly, all age groups except Middle Aged Adults, Seniors and Young Seniors spent below the overall average purchase.
5. Single customers spent more than married customers accounting for 59% of purchase.
6. The average purchase amount for high-value products is \$14,513, this indicates that customers are willing to spend more for premium products, providing an opportunity for targeted marketing.
7. The longer a customer stays in their current city in years the less they spend on black Friday except customers who spend less than a year who spent the least and customers who spend exactly a year had the highest purchase. This indicates that the number of years a customer spends in their city plays an important role in purchase.
8. City B leads in both population and purchase, city C follows with a smaller population and spending, while city A had the least population and purchase.

Recommendations

1. **Age Group Segmented Campaigns:** Tailor marketing campaigns to cater specifically to age groups that spend below the overall average. More analysis needs to be done to identify products and deals that resonate with their preferences.
2. **Family-Centric Promotions to Boost Spending:** Introduce family-centric promotions and deals to encourage increased spending among Married customers as it is a market segment with potential. Highlight packages that cater to family needs and convenience.
3. **Loyalty Programs for Single Customers:** Develop and promote loyalty programs to reward and retain single customers enhancing their overall shopping experience and encouraging repeat business.
4. **Premium Products Promotion:** Since customers are willing to spend more for premium products, intensify marketing efforts emphasizing their unique features and benefits.
5. **Residence Duration Incentives:** Introduce incentives for customers with shorter residence durations, aiming to boost spending in this demographic. Consider loyalty programs that offer advantages for both short -term and long-term customers.
6. **City-Specific Strategies:** More data needs to be collected to determine how the distance from home to store in various cities to affects purchase.
Develop marketing strategies to address the varying population sizes and spending patterns. Implement targeted promotions in areas with lower customer engagement to boost participation and sales.

Dashboard

