BLACK FRIDAY PROJECT

Submitted by:

HARSH NEMA

ACKNOWLEDGMENT

I would like to express my profound gratitude to the **Flip Robo team**, who has given me this opportunity to deal with a beautiful dataset and it has helped me to improve my analytical skills. I would like to express my special thanks to our mentor **Mr. Shwetank Mishra Sir** (SME Flip Robo) for their contributions to the completion of my project titled **'Black Friday Project'**.

I am eternally grateful to **"Datatrained"** for giving me the opportunity for an internship at Flip Robo. Last but not least I have to thank my parents and wife for their love and support during my project.

References:

1. SCIKIT Learn Library Documentation
2. Datatrained recorded videos for Data Science Course
3. Hands-on Machine learning with Scikit learn and tensor flow by Aurelien Geron

1. INTRODUCTION

- ## Business Problem Framing

The Black Friday dataset contains information about the transactions that occurred during the Black Friday sales event. To frame a business problem using this dataset, we can focus on recommender systems that can help retailers improve customer experience by suggesting products that are likely to interest customers based on their shopping history and preferences. Using the Black Friday dataset, retailers can develop a recommender system that suggests products to customers during the sales event.

- ## Conceptual Background of the Domain Problem

A retail company “ABC Private Limited” wants to understand the customer purchase behaviour (specifically, purchase amount) against various products of different categories. They have shared purchase summary of various customers for selected high-volume products from last month. The domain problem of Black Friday sales is about understanding consumer behaviour, optimizing sales and revenue, managing inventory, segmenting customers, and using recommender systems to personalize the shopping experience.

- ## Review of Literature

Several studies have used this dataset to gain insights into consumer behaviour during Black Friday sales. For example, one study analysed the relationship between demographic factors and purchase behaviour, and found that men tend to spend more than women during Black Friday sales. Another study analysed the impact of online reviews on purchase decisions during Black Friday sales, and found that positive reviews had a significant impact on sales.

Another commonly analysed dataset related to Black Friday is the "Black Friday Ads Dataset" also available on Kaggle. This dataset contains information about the deals and discounts offered by various retailers during the 2015 Black Friday sale in the United States. The dataset includes information such as the retailer, product category, and discount percentage.

Studies have used this dataset to gain insights into the pricing strategies employed by retailers during Black Friday sales. For example, one study analysed the relationship between discount percentage and sales volume, and found that products with higher discounts tend to sell more. Another study analysed the impact of advertising on sales volume, and found that retailers who advertised their Black Friday deals earlier tended to have higher sales volumes.

In conclusion, there have been several studies that have used Black Friday datasets to gain insights into consumer behaviour and retail pricing strategies during Black Friday sales. The availability of these datasets has enabled researchers to analyse large amounts of data and draw meaningful conclusions about the shopping event.

- **Motivation for the Problem Undertaken**

Some potential motivations for analysing the Black Friday dataset :

a) To identify trends and patterns in customer purchasing behaviour during the Black Friday sales event.
b) To determine which products are most popular among different demographic groups.
c) To help retailers optimize their pricing and product selection strategies for the Black Friday sales event.
d) To gain insights into consumer preferences and behaviour that can be used to inform marketing and advertising campaigns throughout the year.

2. Analytical Problem Framing

- **Mathematical/ Analytical Modeling of the Problem**

One possible mathematical model is to use linear regression to predict sales based on various factors, such as gender, age, occupation, and city category. We can use the dataset to train a linear regression model that predicts sales based on these factors, and then use the model to make predictions about future sales.

Another possible modelling technique is to use clustering to group shoppers into different segments based on their shopping behaviour. We can use the dataset to cluster shoppers based on factors such as their purchase history, product preferences, and demographics. This will allow retailers to tailor their marketing and promotional strategies to specific shopper segments.

- **Data Sources and their formats**

The dataset has following columns:

• User ID

• Product ID

• Gender

• Age

• Occupation

• Category of the City (A, B, C)

• Stay_In_Current_City_Years

• Marital Status

•Product_Category_1

• Product_Category_2

• Product_Category_3

- Purchase (Target Variable)

There were 11 features in dataset including target feature ‘Purchase’

- Train dataset

```
In [44]: df_train = pd.read_csv('blackFriday_train.csv')
```

```
In [3]: df_train.head()
```

Out[3]:

	User_ID	Product_ID	Gender	Age	Occupation	City_Category	Stay_In_Current_City_Years	Marital_Status	Product_Category_1	Product_Category_2	Product
0	1000001	P00069042	F	0-17	10	A	2	0	3	NaN	
1	1000001	P00248942	F	0-17	10	A	2	0	1	6.0	
2	1000001	P00087842	F	0-17	10	A	2	0	12	NaN	
3	1000001	P00085442	F	0-17	10	A	2	0	12	14.0	
4	1000002	P00285442	M	55+	16	C	4+	0	8	NaN	

- Test dataset

```
In [45]: df_test = pd.read_csv('blackFriday_test.csv')
```

```
In [6]: print("\033[1m" + 'Number of rows in the given test dataset:' + "\033[0m")
        print(df_test.shape[0])

        print("\033[1m" + 'Number of columns in the given test dataset:'  + "\033[0m")
        df_test.shape[1]
```

```
Number of rows in the given test dataset:
233599
Number of columns in the given test dataset:
```

```
Out[6]: 11
```

```
In [46]: df = df_train.append(df_test) # Merging train and test data
```

```
In [8]: df.head()
```

Out[8]:

	User_ID	Product_ID	Gender	Age	Occupation	City_Category	Stay_In_Current_City_Years	Marital_Status	Product_Category_1	Product_Category_2	Product
0	1000001	P00069042	F	0-17	10	A	2	0	3	NaN	
1	1000001	P00248942	F	0-17	10	A	2	0	1	6.0	
2	1000001	P00087842	F	0-17	10	A	2	0	12	NaN	
3	1000001	P00085442	F	0-17	10	A	2	0	12	14.0	
4	1000002	P00285442	M	55+	16	C	4+	0	8	NaN	

- Merging train and test data into one data frame

```
In [9]: print("\033[1m" + 'Number of rows in the given final dataset:' + "\033[0m")
        print(df.shape[0])

        print("\033[1m" + 'Number of columns in the given final dataset:'  + "\033[0m")
        df.shape[1]
```

```
Number of rows in the given final dataset:
783667
Number of columns in the given final dataset:
```

```
Out[9]: 12
```

```
In [10]: df.describe().T
```

Out[10]:

	count	mean	std	min	25%	50%	75%	max
User_ID	783667.0	1.003029e+06	1727.266668	1000001.0	1001519.0	1003075.0	1004478.0	1006040.0
Occupation	783667.0	8.079300e+00	6.522206	0.0	2.0	7.0	14.0	20.0
Marital_Status	783667.0	4.097774e-01	0.491793	0.0	0.0	0.0	1.0	1.0
Product_Category_1	783667.0	5.366196e+00	3.878160	1.0	1.0	5.0	8.0	20.0
Product_Category_2	537685.0	9.844506e+00	5.089093	2.0	5.0	9.0	15.0	18.0
Product_Category_3	237858.0	1.266860e+01	4.125510	3.0	9.0	14.0	16.0	18.0
Purchase	550068.0	9.263969e+03	5023.065394	12.0	5823.0	8047.0	12054.0	23961.0

- There seem missing values in 'Product_Category_2', 'Product_Category_3' and 'Purchase'. Missing values in 'Purchase' is due to test dataset merging with train dataset
- The mean purchase amount is **9263.96**

- **Data Pre-processing**

The dataset is large and it may contain some data errors. To make clean and error-free data, some data cleaning & data pre-processing was performed on the scraped dataset.

- Missing Value Imputation

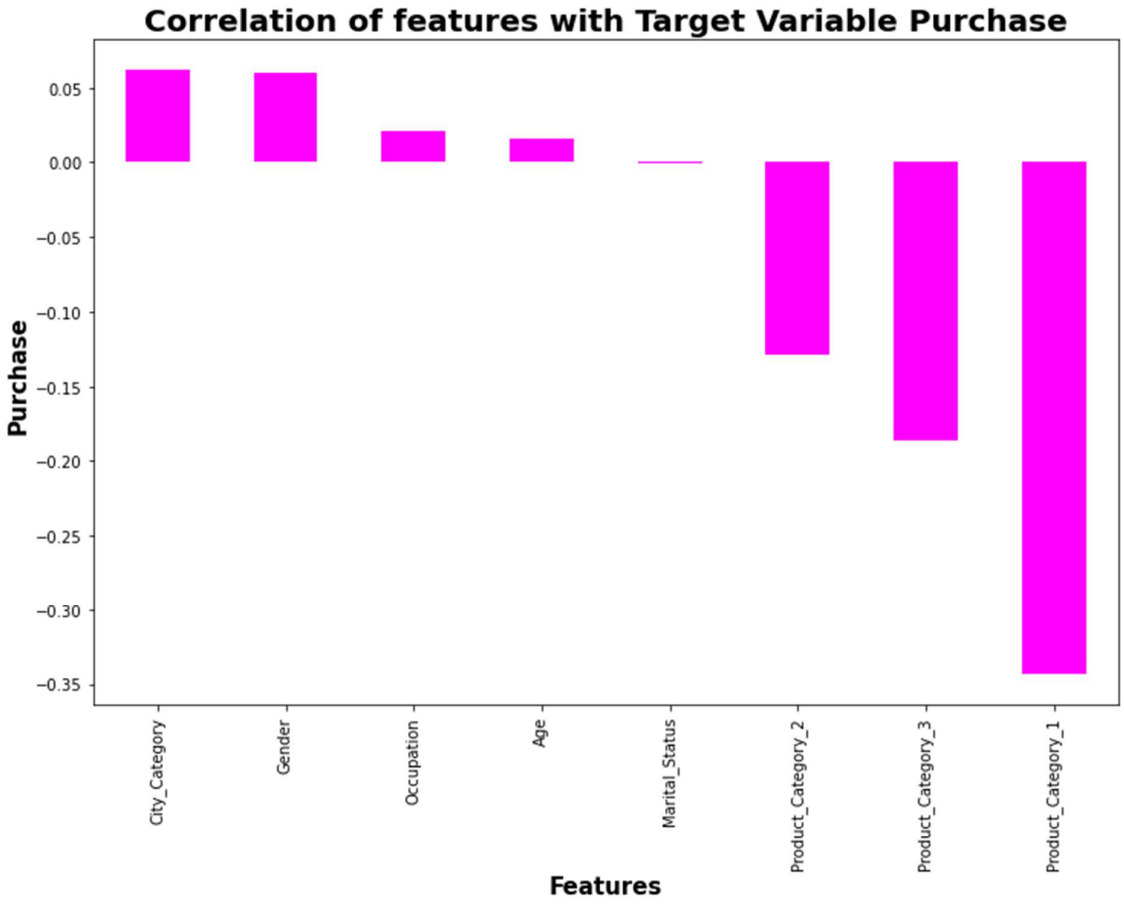
```
In [12]: # Checking missing values in the dataset

         df.isnull().sum()

Out[12]: User_ID                        0
         Product_ID                     0
         Gender                         0
         Age                            0
         Occupation                     0
         City_Category                  0
         Stay_In_Current_City_Years     0
         Marital_Status                 0
         Product_Category_1             0
         Product_Category_2        245982
         Product_Category_3        545809
         Purchase                  233599
         dtype: int64
```

It is a process of replacing missing or null values in a dataset with substituted values. The goal of this process is to obtain a complete and accurate dataset for further analysis or modelling. Lot of missing values were found in the dataset and hence proper imputation method was used to get the clean data.

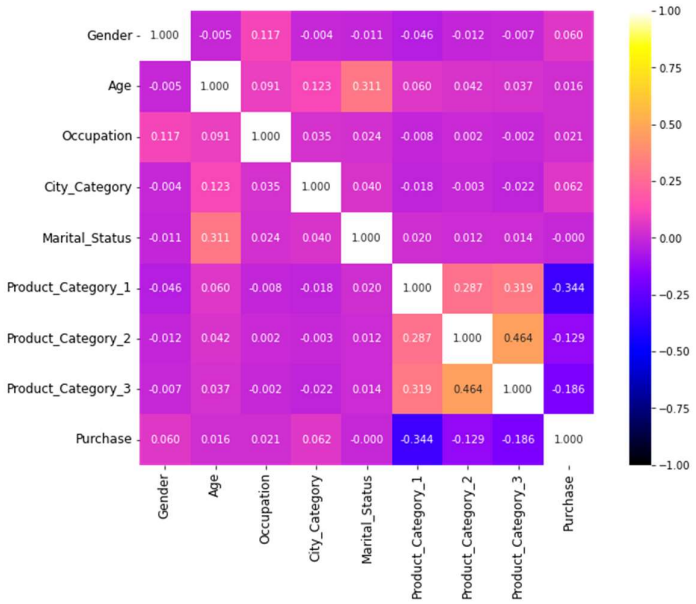
```
In [13]: df['Product_Category_2'].unique()

Out[13]: array([nan,  6., 14.,  2.,  8., 15., 16., 11.,  5.,  3.,  4., 12.,  9.,
                10., 17., 13.,  7., 18.])

In [47]: df['Product_Category_2'] = df['Product_Category_2'].fillna(df['Product_Category_2'].mode()[0])

In [15]: df['Product_Category_2'].isnull().sum()

Out[15]: 0

In [16]: df['Product_Category_3'].unique()

Out[16]: array([nan, 14., 17.,  5.,  4., 16., 15.,  8.,  9., 13.,  6., 12.,  3.,
                18., 11., 10.])

In [48]: df['Product_Category_3'] = df['Product_Category_3'].fillna(df['Product_Category_3'].mode()[0])

In [18]: df['Product_Category_3'].isnull().sum()

Out[18]: 0
```

Mode imputation is a method used to replace missing values in a dataset with the most common value (i.e., mode) of that variable. This is a simple and commonly used method of imputation, especially for categorical variables. To perform mode imputation, the missing values in a variable are replaced with the value that occurs most frequently in that variable.

- **Data Inputs- Logic- Output Relationships**



A correlation heat map is plotted to gain an understanding of the relationship between target features & independent features.



- **<u>Hardware and Software Requirements and Tools Used</u>**

  Hardware Used -

  1. Processor — Intel(R) Core(TM) i7-9700K CPU @ 3.60GHz 3.60 GHz

  2. RAM — 16.0 GB

  3. GPU — 2GB AMD Radeon Graphics card

  Software utilized -

  1. Google Colab Notebook - to write and execute arbitrary python code through the browser

  2. Microsoft office – for making project reports and ppt

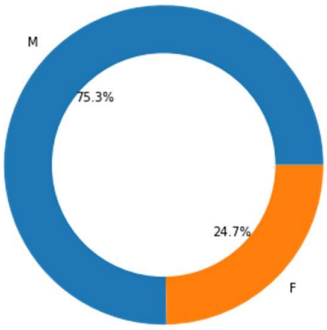
  Libraries Used – General libraries used for data wrangling

```
In [1]: import pandas as pd
        import numpy as np
        import matplotlib.pyplot as plt
        import plotly.express as px
        import seaborn as sns
        import warnings
        warnings.filterwarnings('ignore')
        %matplotlib inline
        from collections import Counter
        import matplotlib.ticker as ticker
```

3. Exploratory Data Analysis

- Visualizations

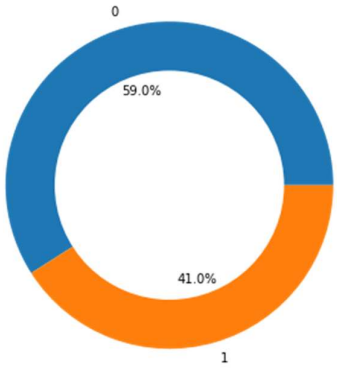
1. Gender Proportion



Gender

- The pie plot shows that males make up 75.3% of the Black Friday shoppers, while females make up 24.7%. This means that males are almost 3 times more likely to participate in Black Friday shopping compared to females.
- The gender distribution data can be used to gain insights into consumer behaviour, particularly around Black Friday shopping. For example, retailers may use this data to plan their marketing strategies and promotions for the upcoming Black Friday sales, by targeting their messaging to the male or female audience.

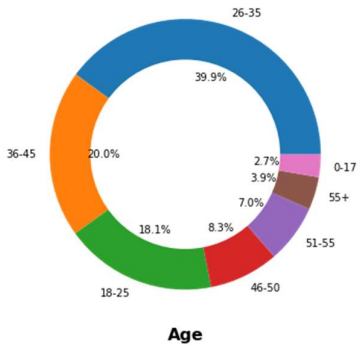
2. Marital Status



Marital Status

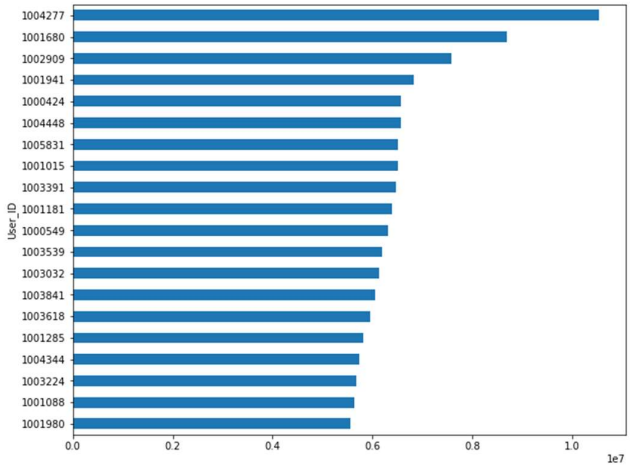
- The plot shows that the majority of shoppers (59%) were unmarried, while the remaining 41% were married. This suggests that unmarried individuals were more likely to make purchases on Black Friday compared to married individuals.
- It is possible that unmarried individuals had more time, flexibility, or disposable income to shop on Black Friday.

3. Age Groups



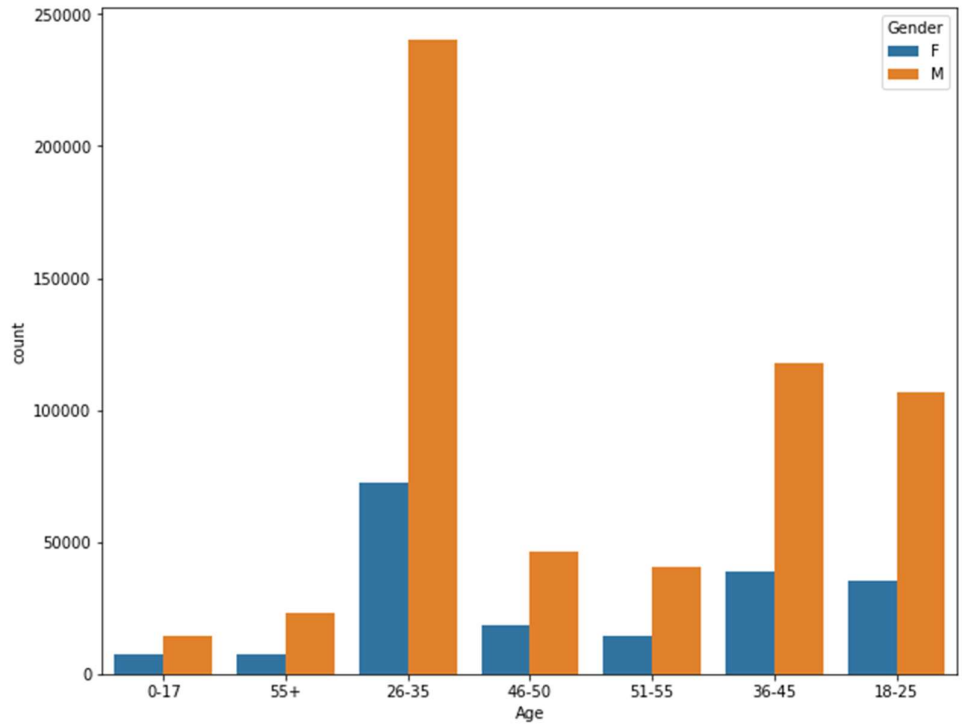
- The largest age group of Black Friday shoppers are those aged between 26 and 35 years, accounting for almost 40% of the shoppers.
- The second largest group is the 18-25 age group, making up 18.1% of the total number of shoppers.
- It can help retailers to tailor their Black Friday promotions and advertising campaigns to the age groups that are most likely to participate in the event.

4. Top 20 Spending Users



- The plot clearly shows that user ID 1004277 tops the list with the highest amount spent, followed by the next 19 users in descending order.
- This user spent a significantly higher amount than the other users, indicating that they had a substantial shopping spree on Black Friday.
- It's important for the seller to identify high quality customers.
- By understanding their needs and preferences, a seller can tailor their products, pricing, and customer service offerings to better serve these customers and increase their loyalty.

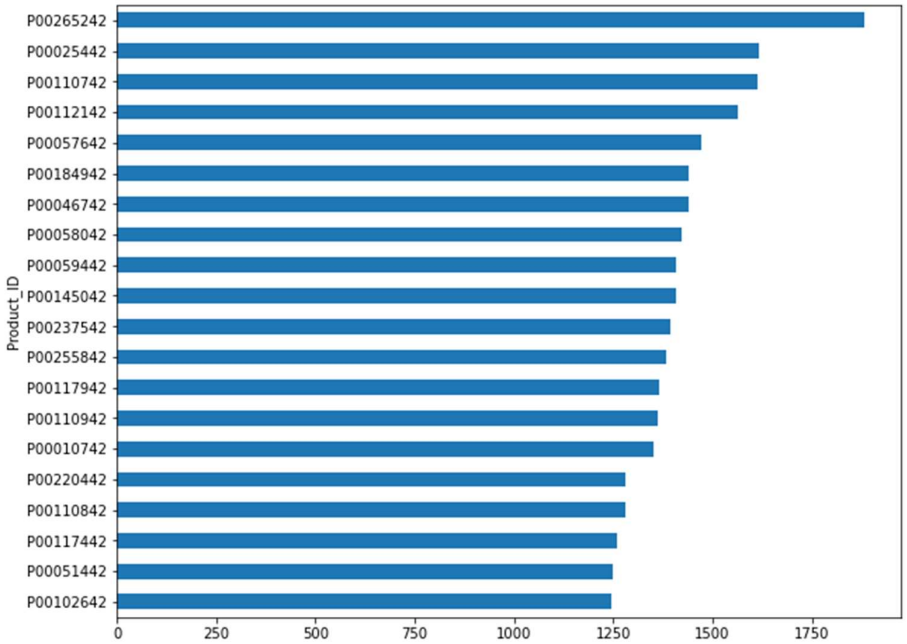
5. User distribution by age group and gender



- We observe that most of the users who participate in the Black Friday Sale are from age group 26-35, 36-45 and 18-25.
- The reason for this is that individuals in these age groups tend to have more disposable income.
- Additionally, younger consumers tend to be more tech-savvy, which makes it easier for them to shop online during Black Friday sales.

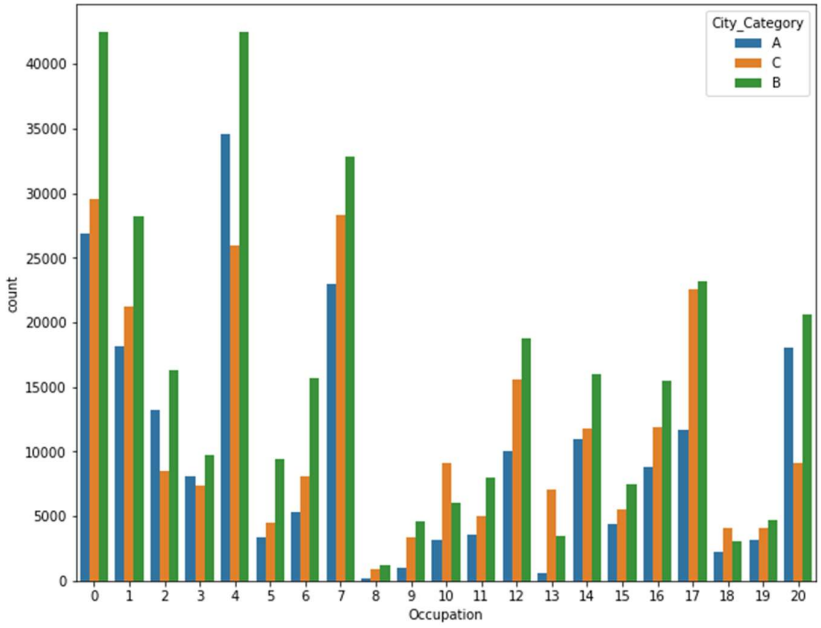
- Males are more likely to participate in Black Friday shopping compared to females in every age group.

6. Top 20 Products in Black Friday Sale



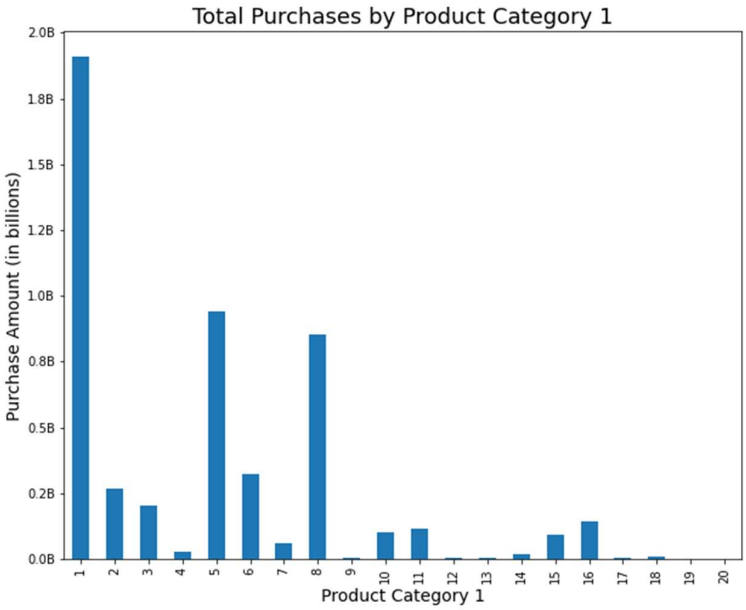
- The most popular products may help the merchant adjust their business strategy and can prepare for the next shopping season better so that to Increase revenue and profit.

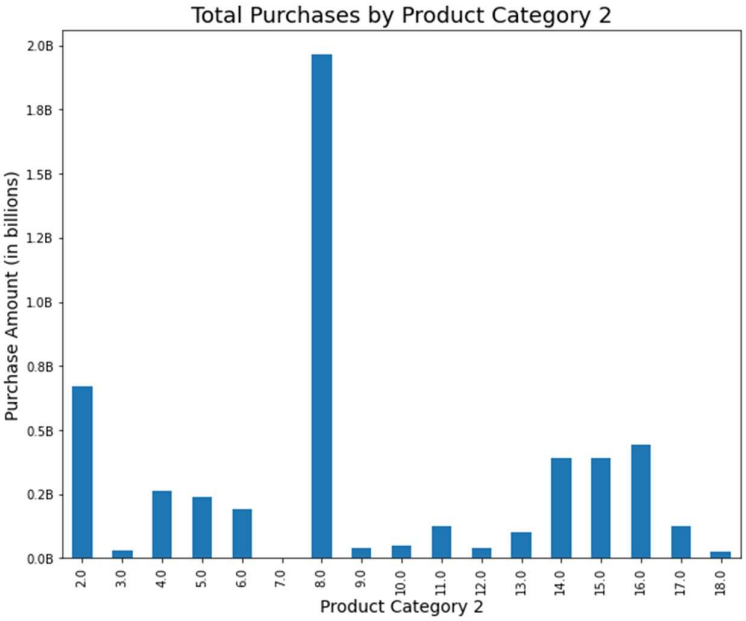
7. User distribution by Occupation and City

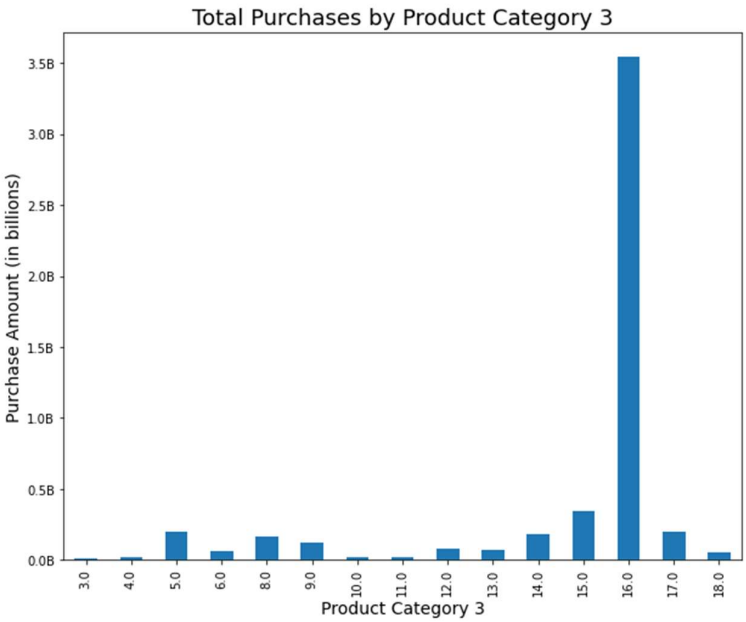


- The plot shows that for almost all Occupation Category, users from Citi B did more shopping comparing to users from City A & Citi C
- The reason is City B is larger than City A & Citi C and thus has a larger population
- And customers from occupation 0, 4, 7 did more shopping than other occupations

8. Purchase Vs Different Product Category







- In Product category 1, Product 1 was highly purchased, amounting to 1.9 billion.
- Product 8 in Product category 2 was highly purchased, amounting to 2 billion.
- Product 16 in Product category 3 was purchased up to 3.5 billion.

4. CONCLUSION

Key Findings and Conclusions of the Study

In conclusion, the Black Friday dataset provides valuable insights into consumer behaviour and purchasing patterns. Retailers can use this information to plan their marketing strategies, promotions, and product offerings to better serve their customers.

- The data shows that males are more likely to participate in Black Friday shopping than females.
- The largest age group of shoppers are those aged between 26 and 35 years.
- Unmarried individuals were also more likely to make purchases on Black Friday compared to married individuals.
- The data suggests that users from City B did more shopping compared to users from City A and Citi C. Retailers can use this information to adjust their inventory and marketing strategies based on the demographics and preferences of each city's population.
- The data reveals that users from occupations 0, 4, and 7 did more shopping compared to users from other occupations. Retailers can use this information to tailor their product offerings and marketing strategies to appeal to customers in these occupations.
- Additionally, identifying high-quality customers is crucial for sellers, as they can tailor their products, pricing, and customer service offerings to increase customer loyalty.
- The most popular products were Product 1 in Product Category 1, Product 8 in Product Category 2, and Product 16 in Product Category 3. Retailers can use this information to stock up on these products for the next Black Friday sale to meet the high demand and increase their revenue.
- Finally, the most popular products and cities can help merchants adjust their business strategies to increase revenue and profit.