Black Friday Suggestion System Prediction

Submitted in partial fulfillment of the requirements

for the award of the degree of

Bachelor of Computer Applications

Guide:

Ms. Leena Gupta

Submitted by:

Simran Bagga

07013702020

Institute of Information Technology
&Management,New Delhi – 110058
Batch (2020-2023)

Certificate

I, **Simran Bagga (07013702020)** certify that the Major Project Report (BCA-356) entitled "**Black Friday suggestion prediction**" is done by me and it is an authentic work carried out by me at **Institute of Information Technology & Management** (Name of the organization or of the Institute). The matter embodied in this project work has not been submitted earlier for the award of any degree or diploma to the best of my knowledge and belief.

Signature of the Student

Date:

Certified that the Project Report (BCA-356) entitled **"Black Friday suggestion prediction"** done by the above student is completed under my guidance.

Signature of the Guide:

Date:

Name of the Guide:

Ms. Leena Gupta

Designation:

Countersign HOD

Countersign Directo

SELF CERTIFICATE

This is to certify that the dissertation/project report entitled **"Black Friday suggestion prediction"** done by me is an authentic work carried out for the partial fulfilment of the requirements for the award of the degree of Bachelor of Computer Applications under the guidance of **Ms. Leena Gupta.** The matter embodied in this project work has not been submitted earlier for award of any degree or diploma to the best of my knowledge and belief.

Signature of the student

Name of the Student:-

Simran Bagga

Roll No.:-

07013702020

Acknowledgement

Place: INSTITUE OF INFROMATION TECHNOLOGY AND MANAGEMENT

Date:

I would like to express my profound gratitude to **Prof. (Dr.) Rachita Rana, Director IITM Janakpuri** and **Prof. (Dr.) Sudhir Kumar Sharma, HOD (Computer Science) IITM Janakpuri** department for their contributions to the completion of my project titled “**Black Friday suggestion prediction**”.

I would like to express my special thanks to my Guide **Ms. Leena Gupta,** for his time and efforts he provided throughout the training. Your useful advice and suggestionswere really helpful to me during the project’s completion. In this aspect, I am eternally grateful to you.

I would like to acknowledge that this project was completed entirely by me and not by someone else.

Signature

Simran Bagga

C O N T E N T S

List of Tables/Figures/Symbols

SYNOPSIS

Title: Black Friday Suggestion prediction

Introduction

Black Friday, the legendary shopping extravaganza following Thanksgiving, has become a pivotal event for retailers and consumers alike. In recent years, the power of data and predictive analytics has emerged as a game-changer for businesses aiming to forecast Black Friday sales. By analyzing historical data, consumer behavior, and market trends, retailers can gain valuable insights into future outcomes and optimize their strategies.In this project, explores the significance of Black Friday suggestion suggestion prediction, examining data sources, algorithms, and key factors influencing sales. By leveraging data-driven insights, businesses can make informed decisions and maximize profitability during this high-stakes shopping event.

Statement about the problem

The main problem is that retailers struggle to predict how much they will sell on Black Friday. This uncertainty makes it difficult for them to make smart decisions about things like how much stock to order or how to price their products. If they don't get these decisions right, they can end up with too much or too little inventory, which can lead to financial losses or disappointed customers. So, the challenge is to find a way to accurately predict Black Friday sales, so retailers can plan better and make the most of this important shopping day.

Why is the particular topic chosen?

The topic of Black Friday suggestion prediction was chosen because it addresses a crucial challenge faced by retailers. Black Friday is a significant shopping event, but the uncertainty of sales makes it difficult for businesses to plan effectively. By studying sales prediction, retailers can make better decisions about pricing and inventory, leading to improved profitability and customer satisfaction. Additionally, this topic aligns with the increasing importance of data analytics and machine learning, providing valuable insights for businesses looking to leverage data for competitive advantage.

Objective and scope of the project

The objective of this project is to develop a reliable and accurate Black Friday suggestion prediction model using data analytics and machine learning techniques. The project aims to provide retailers with actionable insights to optimize their strategies, make informed decisions about pricing, inventory management, and marketing campaigns, and ultimately maximize profitability during the Black Friday shopping event.

Scope of this project is to collect the data from source, clean the data by removing null or missing values,then apply visualization and implement different machine learning models.

Non-Functional Modules:

Various Requirement:

- **Security:** It will increase the security of cctv cameras
- **Accuracy**: The accuracy will be managed for sure
- **Reliability:** The system will perform consistently its intended function

Tools and Platforms

Hardware Components:

- Hard Disk: 512 GB
- Processor: INTEL CORE i5
- RAM: 8 GB
- System Type: 64-bit operating system, x64 based processor

Methodology:

Stage of Iterative Model:

SDLC (Software Development Life Cycle)

SDLC processes generally number at 6 distinct stages: planning, analysis, designing, development and testing, implementation, and maintenance. Each of them is briefly explained below.

Phases of Life Cycle Model:

- **Planning and requirement:** All possible requirements of the system to be developed are captured in this phase and documented in a requirement specification document.
- **Analysis and design:** The requirement specifications from first phase are studied in this phase and system design is prepared. It helps in specifying hardware and system requirements and also helps in defining overall system architecture.
- **Implementation:** With inputs from system design, the system is first developed in small programs called units, which are integrated in the next phase. Each unit is developed and tested for its functionality which is referred to as Unit Testing.
- **Testing:** All the units developed in the implementation phase are integrated into a system after testing of each unit. Post integration the entire system is tested for any faults and failures.
- **Evaluation:** This will evaluate the outcome of the processing whether the application is running successfully or not.

Summary

This project aims to develop a Black Friday suggestion prediction model using data analytics and machine learning techniques. The focus is on collecting and analyzing historical data related to Black Friday sales, customer behavior, and market trends. The data is preprocessed, and relevant features are selected and engineered to capture insights. Various machine learning algorithms are evaluated, and the best-performing model is selected for prediction. The deployed model provides actionable insights and recommendations to retailers for pricing, inventory management, and marketing strategies specific to Black Friday. Continuous monitoring and improvement of the model ensure its accuracy and reliability over time.

What contribution would the project make

- Data analysis: It can help analyze historical Black Friday sales data to identify patterns, trends, and correlations. By examining past sales performance, we can uncover insights that inform predictions for future sales.
- Predictive modeling: Using machine learning algorithms and statistical techniques, It can assist in developing predictive models to forecast Black Friday sales. This involves training models on historical data and using them to make predictions based on various factors such as product categories, pricing strategies, promotional campaigns, and customer behavior.
- External factors analysis: It can gather and analyze relevant external factors that influence consumer behavior during Black Friday, such as economic indicators, social media trends, and competitor activities. This information can be integrated into the predictive models to improve accuracy.
- Customer segmentation: It can help segment customers based on their purchasing patterns, preferences, and demographics. By understanding different customer segments, businesses can tailor their marketing strategies and promotional offers to effectively target specific groups.
- Insights and recommendations: It can generate insights and recommendations based on the analysis of historical data and current market trends. These insights can guide retailers in making informed decisions regarding inventory management, pricing strategies, marketing campaigns, and resource allocation.
- Scenario analysis: It can simulate various scenarios based on different assumptions and parameters. For example, we can explore the impact of specific marketing strategies or pricing changes on sales forecasts. This allows businesses to evaluate the potential outcomes and make data-driven decisions accordingly.
- Content generation: It can help generate reports, summaries, and presentations summarizing the analysis and findings. This can be useful for communicating predictions and recommendations to stakeholders and decision-makers.

Chapter-1

Software Project Planning

1.1.General Introduction

1.1.1. Description of the Software System under Study

The day after Thanksgiving Black Friday, has been regarded as the beginning of the United States Christmas shopping season since 1952, although the term "Black Friday" did not become widely used until more recent decades. Many stores offer highly promoted sales on Black Friday and open very early, such as at midnight, or may even start their sales at some time on Thanksgiving. Our project deals with determining the product prices based on the historical retail store sales data. After generating the predictions, our model will help the retail store to decide the price of the products to earn more profits.

1.1.2. Problem Statement

We want to understand the customer purchase behaviour (specifically, purchase amount) against various products of different categories. We have shared purchase summary of various customers for selected high-volume products from last month. The data set also contains customer demographics (age, gender, marital status, city_type, stay_in_current_city), product details (product_id and product category) and Total purchase_amount from last month.

Now, we want to build a model to predict the purchase amount of customer against various products which will help us to create personalized offer for customers against different products.

1.1.3. Intended Operations to be performed

1. Used LabelEncoder for encoding the categorical columns like Age, Gender and City_Category
2. Used get_dummies form Pandas package for converting categorical variable State_In_Current_Years into dummy/indicator variables.
3. Filled the missing values in the Product_Category_2 and Product_Category_3

1.1.4. End Users

- Retailers and E-commerce Companies: Retailers and e-commerce companies are the primary end users of Black Friday suggestion prediction. They can leverage the insights and predictions generated by this project to optimize their inventory management, pricing strategies, and marketing campaigns. By accurately forecasting sales, they can ensure they have sufficient stock of popular items, offer competitive prices, and effectively target their marketing efforts to maximize revenue and customer satisfaction.
- Marketing and Sales Teams: Marketing and sales teams within retail organizations can benefit from the project's predictions to plan and execute targeted campaigns. They can use the insights and recommendations provided to create personalized offers, design effective promotional strategies, and allocate resources to drive customer engagement and increase sales during the Black Friday period.
- Supply Chain and Operations Teams: Supply chain and operations teams can utilize the sales predictions to optimize their logistical operations. By having accurate forecasts, they can plan their procurement, transportation, and warehousing activities more effectively, ensuring the availability of products and minimizing the risk of stockouts or excess inventory.
- Executive Management and Decision-Makers: The executive management and decision-makers in retail organizations can use the predictions and insights from this project to make informed strategic decisions. They can assess the potential impact of different scenarios and allocate resources accordingly, based on the projected sales numbers and market trends.
- Industry Researchers and Analysts: Researchers and analysts in the retail industry can benefit from this project by gaining insights into consumer behavior, market trends, and the effectiveness of different marketing strategies during the Black Friday period.

The findings and analysis from this project can contribute to industry research, benchmarks, and best practices.

1.2. Data Collection

The dataset is acquired from an online site for data called Kaggles. The data contained features like age, gender, marital status, categories of products purchased, city demographics, purchase amount etc. The data consists of 12 columns and 537577 records. Our model will be predicting the purchase amount of the products.

1.3.. Phases of Analysis

1.3.1. Block Diagram

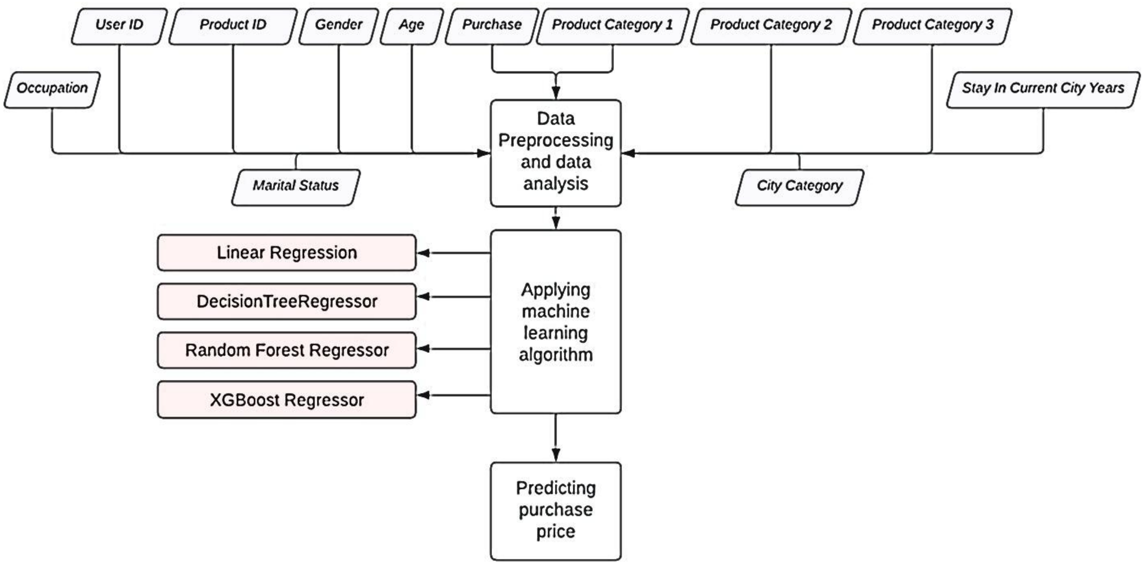

Fig 1

1.3.2. Attributes considered for studying

In the current dataset, there are 11 features and one binary target which are considered as attributes for studying . A brief information about the features is given below:

1. .User_ID: User ID

2. Product_ID: Product ID
3. Gender: Sex of User
4. Age: Age in bins
5. Occupation: Occupation (Masked)
6. City_Category: Category of the City (A,B,C)
7. Stay_In_Current_City_Years: Number of years stay in current city
8. Marital_Status: Marital Status
9. Product_Category_1: Product Category (Masked)
10. Product_Category_2: Product may belongs to other category also (Masked)
11. Product_Category_3: Product may belongs to other category also (Masked)
12. Purchase: Purchase Amount (Target Variable)

1.4. Tools/Platforms

1.4.1. Hardware Specifications

Processor	Core i5
RAM	8.00 GB
Memory	512 GB
System type	64-bit operating system, x64-based processor

Table No 1.1: Hardware Specifications

1.4.2. Software Specifications

OS	Windows 11
Language	Python
Software Development Kit	Jupiter Lab

Table No 1.2: Software Specifications

1.4.3. Packages to be imported

- Pandas: A powerful data manipulation and analysis library that provides data structures like DataFrames, which are essential for handling and analyzing structured data.
- NumPy: A fundamental package for scientific computing in Python, providing support for large, multi-dimensional arrays and mathematical functions.
- Matplotlib: A widely-used plotting library for creating visualizations and graphs.
- Seaborn: A statistical data visualization library built on top of Matplotlib that provides additional high-level plotting functions and improved aesthetics.
- XGBoost Gradient boosting frameworks that excel in handling tabular data and can be used for building powerful prediction models.

1.5. Project Planning Activities

1.5.1. Gantt Chart

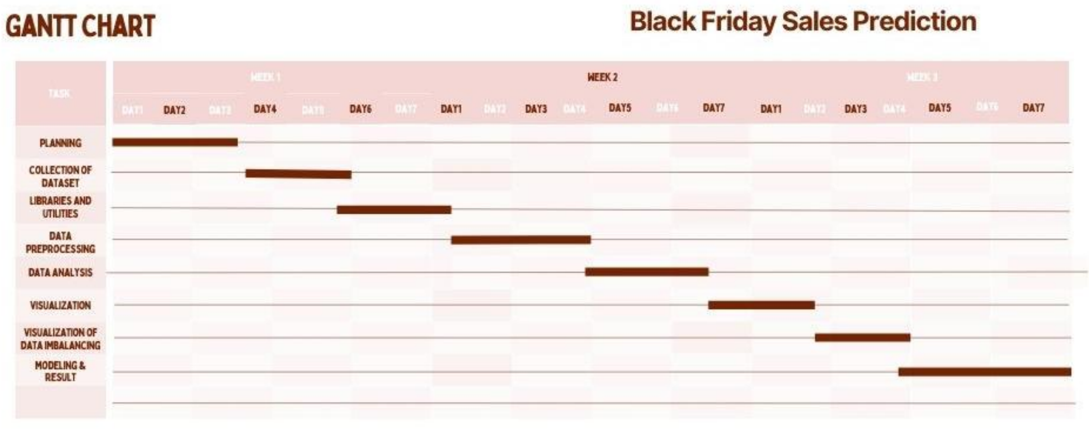

Fig 2

Chapter-2:
Literature Review

2.1. Summary of Paper Studies

Machine Learning Application for Black Friday Suggestion Prediction Framework

By: H V Ramachandra; G Balaraju; A Rajashekar; Harish Patil

Understanding the purchase behavior of various customers (dependent variable) against different products using their demographic information (IS features where most of the features are self -explanatory. This dataset consist of null values, redundant and unstructured data. Machine learning is the most common applications in the domain retail industry. This concept helps to develop a predictor that has a distinct commercial value to the shop owners as it will help with their inventory management, financial planning, advertising and marketing. This entire process of developing a model includes preprocessing, modelling, training testing and evaluating. Hence, frameworks will be developed to automate few of this process and its complexity will be reduced. The algorithm we proposed was Random Forest regressor that performed an average accuracy of 83.6% and with minimum RMSE (Root Mean Squared Error) value of 2829 on tire Black Friday sales dataset.

Predictive Analytics for Black Friday Sales using Machine Learning Technique

By: Saravanan Alagarsamy; K. Ganesh Varma; K. Harshitha; K. Hareesh; K. Varshini

In the day to day life, many products are launched in the market. Many products get successful in the market and some of the products get fail. Predicting the behavior of customers towards various products leads to make successful in sales. Different machine learning techniques are used to predict the behavior of different customers in purchasing various products. The prediction model with the combination of different regression and Xbooster will lead to identifying the customer demands. The above-said process leads to an increase in the profit of retail stores with an accuracy of 99.21%. The customers can identify black Friday, where the products can purchase very cheaply according to interest also based on the various features like product price and comparison with the existing one.

Black Friday Sales Prediction using Supervised Machine Learning

By: Shambhavi Patil; Om Nankar; Renuka Agrawal; Kanhaiya Sharma; Shashank Awasthi; Neha Jha

Machine learning has developed as one of the most influential research domains in the last decade with reasonable doubt. The emphasis on "learning" in machine learning enables computers to judge better. Based on previous experiences, machine learning models are able to judge better and predict future outcomes precisely. Recent advancements in machine learning have promoted efficient intelligence in business decisions and have further made systems capable of a wide range of applications from facial recognition to natural language processing. Prediction models are put to use in businesses in order to determine the most likely outcomes based on the data that is presented. Understanding and predicting the future purchase pattern of discrete customers against different products based on their demographic information of the features is the motive behind the work. The ideology discussed in this work helps to design and develop a predictor model which will be of much assistance to sales administration at the time of Black Friday. The developed model before implementation is tested with different classification techniques. Random Forest regression-based approach used to predict black Friday sales.

Black Friday Sales Prediction using Machine Learning

By: C. M. Wu et al

Black Friday sales prediction model to analyze the customer's past spending and predict the future spending of the customer. The dataset referred is Black Friday Sales Dataset from analytics vidhya. They have machine learning models such as Linear Regression, MLK classifier, Deep learning model using Keras, Decision Tree, and Decision Tree with bagging, and XGBoost. The performance evaluation measure Root Mean Squared Error (RMSE) is used to evaluate the models used. Simple problems like regression can be solved by the use of simple models like linear regression instead of complex neural network models.

2.2. Integrated summary of the Literature studied

Black Friday marks the beginning of the Christmas shopping festival across the US. The product categories range from electronic items, Clothing, kitchen appliances, Décor. Research has been carried out to predict sales by various researchers. The analysis of this data serves as a basis to provide discounts on various product items. With the purpose of analyzing and predicting the sales, we have used three models. The dataset Black Friday Sales Dataset available on Kaggle has been used for analysis and prediction purposes. The models used for prediction are linear regression, Decision Tree Regressor, Random Forest Regressor and XGBoost Regressor. Mean Squared Error (MSE) is used as a performance evaluation measure. XGBoost Regressor outperforms the other models with the least MSE score.

Chapter-3

Implementation and Results

3.1. Phase 1: Installing Packages

```
import numpy as np
import pandas as pd
import matplotlib.pyplot as plt
import seaborn as sns
import xgboost as xgb
```

3.2. Phase 2: Importing and reading data

```
data = pd.read_csv("BlackFriday.csv")

data.head()
```

	User_ID	Product_ID	Gender	Age	Occupation	City_Category	Stay_In_Current_City_Years	Marital_Status	Product_Category_1	Product_Category_2	Product_Category_3	Purchase
0	1000001	P00069042	F	0-17	10	A	2	0	3	NaN	NaN	8370
1	1000001	P00248942	F	0-17	10	A	2	0	1	6.0	14.0	15200
2	1000001	P00087842	F	0-17	10	A	2	0	12	NaN	NaN	1422
3	1000001	P00085442	F	0-17	10	A	2	0	12	14.0	NaN	1057
4	1000002	P00285442	M	55+	16	C	4+	0	8	NaN	NaN	7969

Table No 3.2: black Friday data

```
data.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 537577 entries, 0 to 537576
Data columns (total 12 columns):
 #   Column                      Non-Null Count   Dtype
---  ------                      --------------   -----
 0   User_ID                     537577 non-null  int64
 1   Product_ID                  537577 non-null  object
 2   Gender                      537577 non-null  object
 3   Age                         537577 non-null  object
 4   Occupation                  537577 non-null  int64
 5   City_Category               537577 non-null  object
 6   Stay_In_Current_City_Years  537577 non-null  object
 7   Marital_Status              537577 non-null  int64
 8   Product_Category_1          537577 non-null  int64
 9   Product_Category_2          370591 non-null  float64
 10  Product_Category_3          164278 non-null  float64
 11  Purchase                    537577 non-null  int64
dtypes: float64(2), int64(5), object(5)
memory usage: 49.2+ MB
```

Age should be treated as a numerical column
City_Category we can convert this to a numerical column and should look at the frequency of each city category.
Gender has two values and should be converted to binary values
Product_Category_2 and Product_Category_3 have null values

Checking Null values

```
data.isnull().sum()
```

```
Out[7]: 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            166986
        Product_Category_3            373299
        Purchase                           0
        dtype: int64
```

```
data.isnull().sum()/data.shape[0]*100
```

```
Out[8]: User_ID                        0.000000
        Product_ID                     0.000000
        Gender                         0.000000
        Age                            0.000000
        Occupation                     0.000000
        City_Category                  0.000000
        Stay_In_Current_City_Years     0.000000
        Marital_Status                 0.000000
        Product_Category_1             0.000000
        Product_Category_2            31.062713
        Product_Category_3            69.441029
        Purchase                       0.000000
        dtype: float64
```

Unique elements in each attributes

```
data.nunique()
```

```
Out[9]: User_ID                        5891
        Product_ID                     3623
        Gender                            2
        Age                               7
        Occupation                       21
        City_Category                     3
        Stay_In_Current_City_Years        5
        Marital_Status                    2
        Product_Category_1               18
        Product_Category_2               17
        Product_Category_3               15
        Purchase                      17959
        dtype: int64
```

Chapter-4

Implementation and Visualisations

4.1. Visualization through graphs

Target Variable Purchase

```
sns.distplot(data["Purchase"],color='r')
plt.title("Purchase Distribution")
plt.show()
```

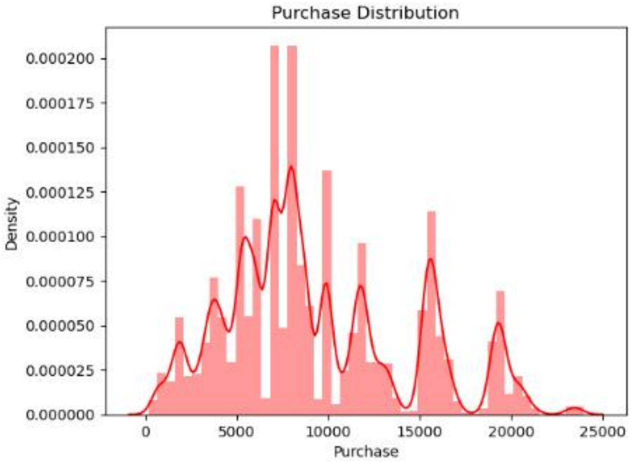

fig 4.1

We can observe that purchase amount is repeating for many customers.This may be because on Black Friday many are buying discounted products in large numbers and kind of follows a Gaussian Distribution.

```
sns.boxplot(data["Purchase"])
plt.title("Boxplot of Purchase")
plt.show()
```

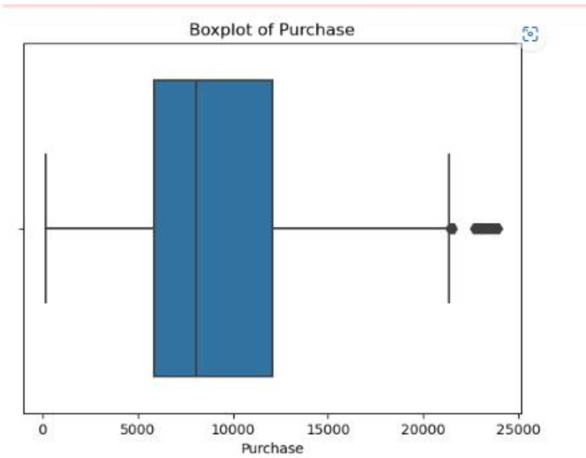

fig 4.2

```
data["Purchase"].skew()
```

```
Out[12]: 0.6242797316083074
```

```
data["Purchase"].kurtosis()
```

```
Out[13]: -0.34312137256836284
```

```
data["Purchase"].describe()
```

```
Out[14]: count    537577.000000
         mean       9333.859853
         std        4981.022133
         min         185.000000
         25%        5866.000000
         50%        8062.000000
         75%       12073.000000
         max       23961.000000
         Name: Purchase, dtype: float64
```

The purchase is right skewed and we can observe multiple peaks in the distribution we can do a log transformation for the purchase.

Gender

```
sns.countplot(data['Gender'])
plt.show()
```

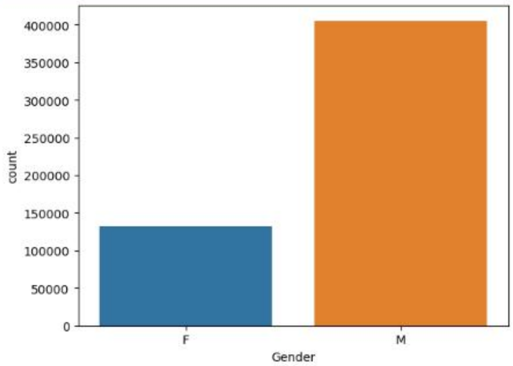

fig 4.3

```
data['Gender'].value_counts(normalize=True)*100
```

```
Out[16]: M    75.408732
         F    24.591268
         Name: Gender, dtype: float64
```

There are more males than females

```
data.groupby("Gender").mean()["Purchase"]
```

```
Out[17]: Gender
         F    8809.761349
         M    9504.771713
         Name: Purchase, dtype: float64
```

On average the male gender spends more money on purchase contrary to female, and it is possible to also observe this trend by adding the total value of purchase.

Marital Status

```
sns.countplot(data['Marital_Status'])

plt.show()
```

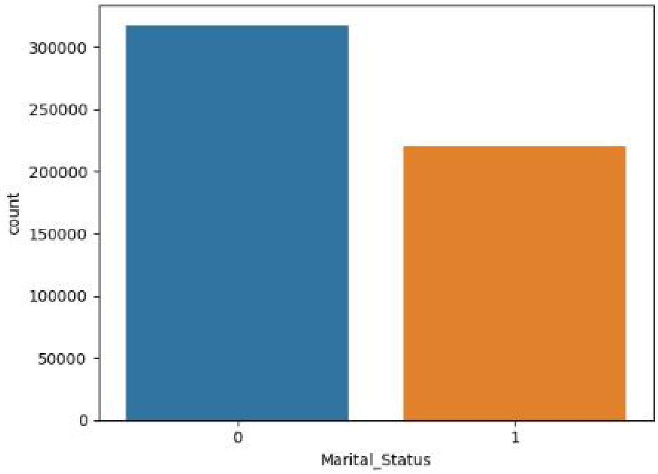

fig 4.4

There are more unmarried people in the dataset who purchase more

```
data.groupby("Marital_Status").mean()["Purchase"]
```

```
Out[19]: Marital_Status
         0    9333.325467
         1    9334.632681
         Name: Purchase, dtype: float64
```

```
data.groupby("Marital_Status").mean()["Purchase"].plot(kind='bar')
plt.title("Marital_Status and Purchase Analysis")
plt.show()
```

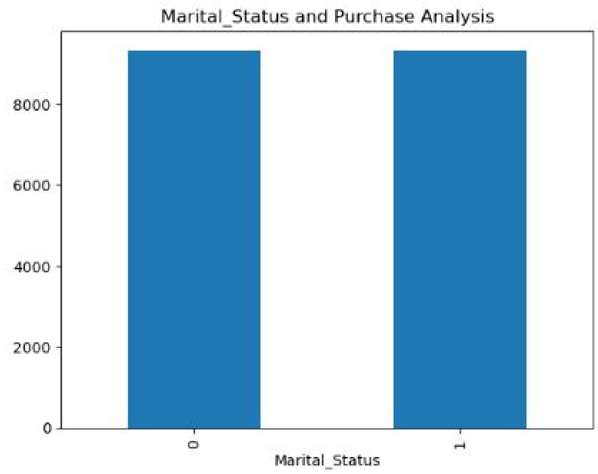

fig 4.5

```
plt.figure(figsize=(18,5))
sns.countplot(data['Occupation'])
plt.show()
```

This is interesting though unmarried people spend more on purchasing, the average purchase amount of married and unmarried people are the same.

Occupation

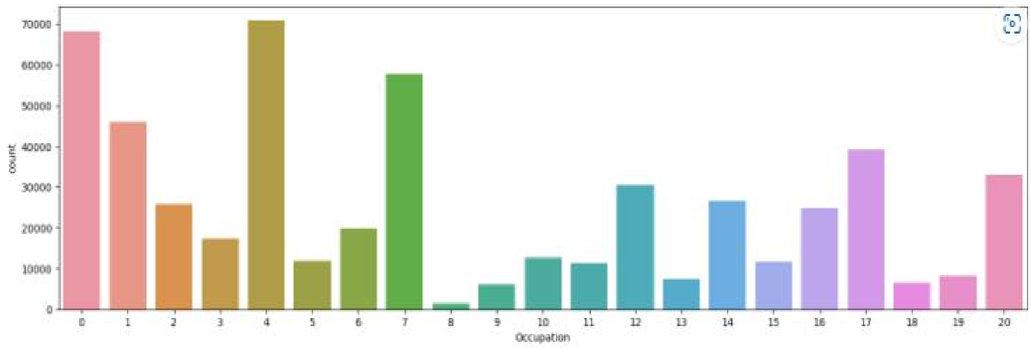

fig 4.6

```
occup = pd.DataFrame(data.groupby("Occupation").mean()["Purchase"])
occup
```

There are at least 20 different occupations in our dataset, but we don't know which number corresponds to each occupation. This makes it challenging to analyze the data effectively. Unfortunately, we don't have any other option but to work with these 20 occupations since we can't reduce their number.

	Purchase
Occupation	
0	9186.946726
1	9017.703095
2	9025.938982
3	9238.077277
4	9279.026742
5	9388.848978
6	9336.378620
7	9502.175276
8	9576.508530
9	8714.335934
10	9052.836410
11	9299.467190
12	9883.052460
13	9424.449391
14	9568.536426
15	9866.239925
16	9457.133118
17	9906.378997
18	9233.671418
19	8754.249162
20	8881.099514

```
occup.plot(kind='bar',figsize=(15,5))
plt.title("Occupation and Purchase Analysis")
plt.show()
```

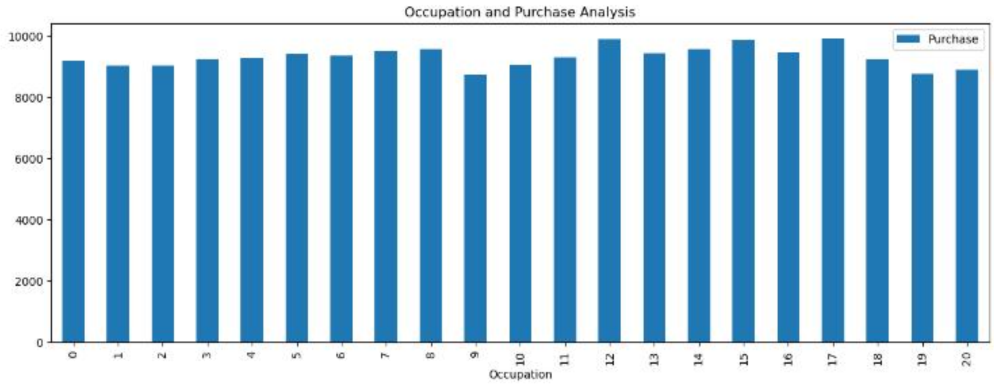

fig 4.7

```
sns.countplot(data['City_Category'])
plt.show()
```

Although there are some occupations which have higher representations, it seems that the amount each user spends on average is more or less the same for all occupations. Of course, in the end, occupations with the highest representations will have the highest amounts of purchases.

City_Category

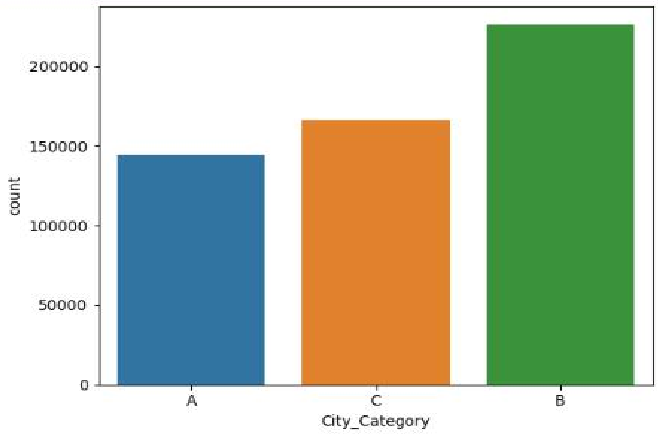

fig 4.8

```
data.groupby("City_Category").mean()["Purchase"].plot(kind='bar')
plt.title("City Category and Purchase Analysis")
plt.show()
```

It is observed that city category B has made the most number of puchases.

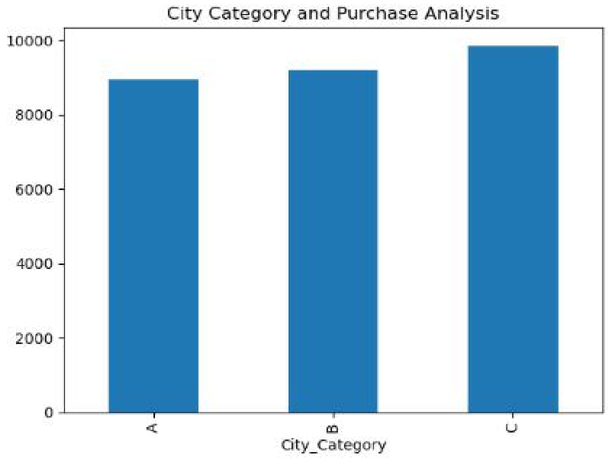

fig 4.9

```
sns.countplot(data['Stay_In_Current_City_Years'])
plt.show()
```

However, the city whose buyers spend the most is city type 'C'.

Stay_In_Current_City_Years

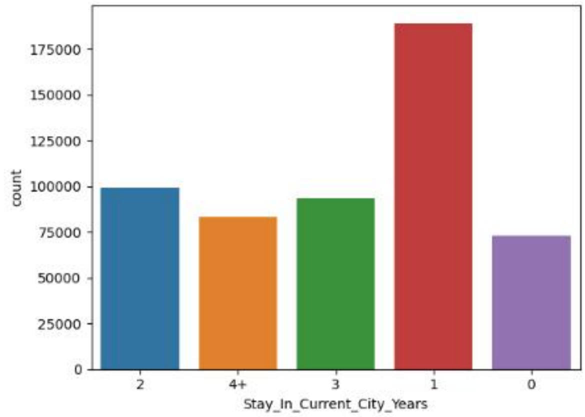

fig 4.10

It looks like the longest someone is living in that city the less prone they are to buy new things. Hence, if someone is new in town and needs a great number of new things for their house that they'll take advantage of the low prices in Black Friday to purchase all the things needed.

```
data.groupby("Stay_In_Current_City_Years").mean()["Purchase"].plot
(kind='bar')
plt.title("Stay_In_Current_City_Years and Purchase Analysis")
plt.show()
```

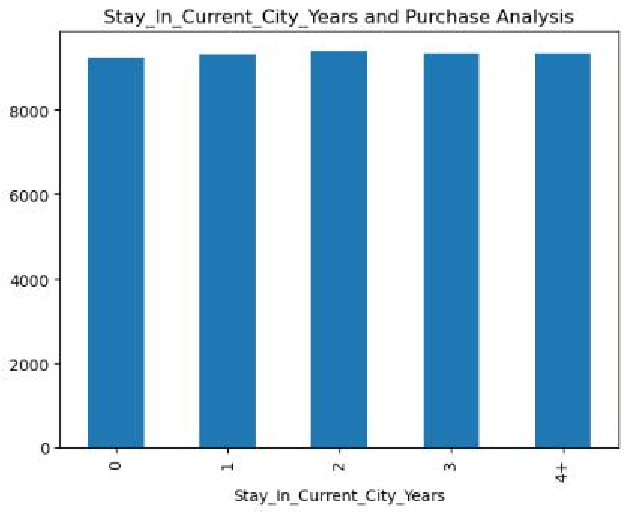

fig 4.1 1

We see the same pattern seen before which show that on average people tend to spend the same amount on purchases regardeless of their group. People who are new in city are responsible for the higher number of purchase, however looking at it individually they tend to spend the same amount independently of how many years the have lived in their current city.

```
sns.countplot(data['Age'])
plt.title('Distribution of Age')
```

```
plt.xlabel('Different Categories of Age')
plt.show()
```

Age

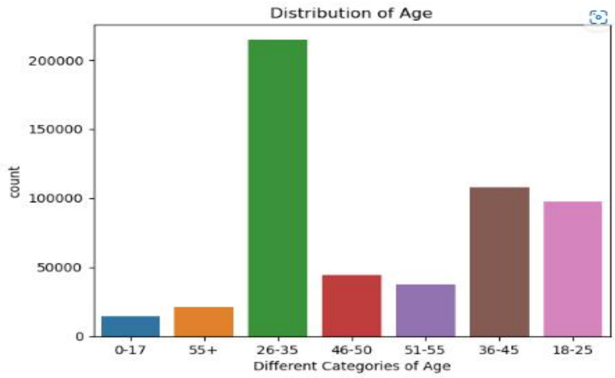

fig 4.12

Age 26-35 Age group makes the most no of purchases in the age group.

```
data.groupby("Age").mean()["Purchase"].plot(kind='bar')
```

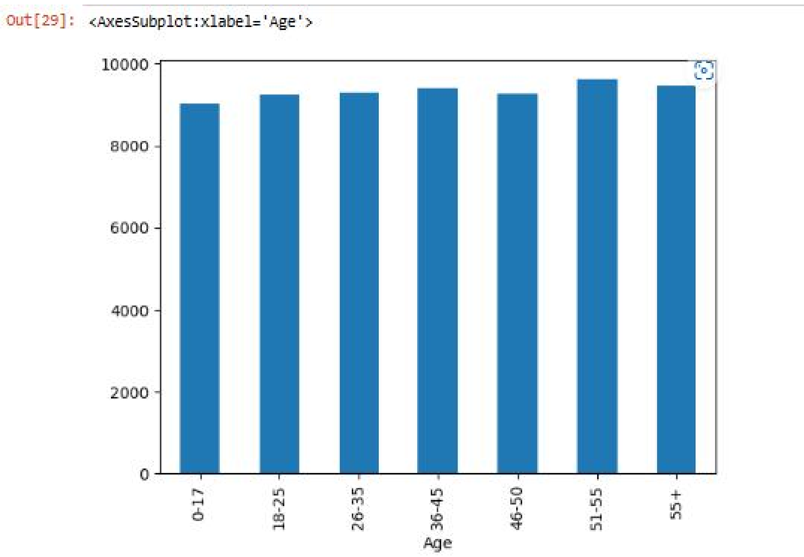

fig 4.13

```
data.groupby("Age").sum()['Purchase'].plot(kind="bar")
plt.title("Age and Purchase Analysis")
plt.show()
```

Mean puchase rate between the age groups tends to be the same except that the 51-55 age group has a little higher average purchase amount

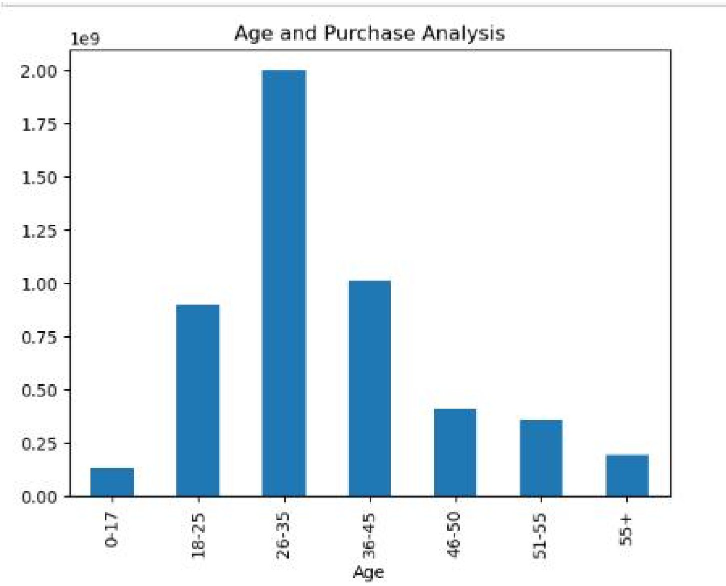

fig 4.14

```
plt.figure(figsize=(18,5))
sns.countplot(data['Product_Category_1'])
plt.show()
```

Total amount spent in purchase is in accordance with the number of purchases made, distributed by age.

Product_Category_1

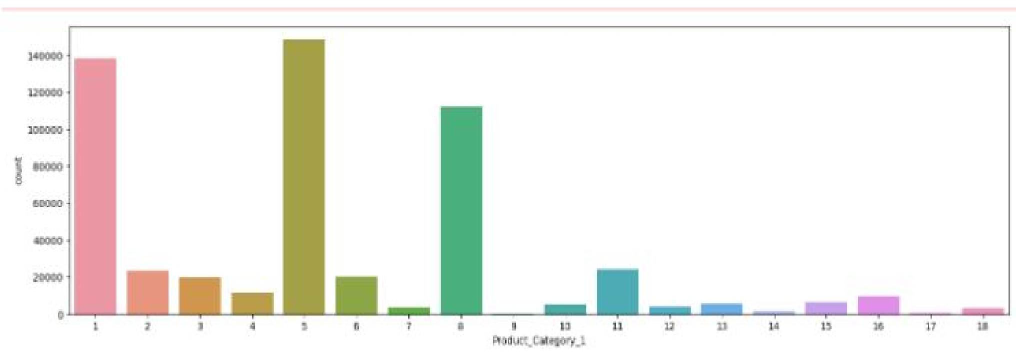

fig 4.15

It is clear that Product_Category_1 numbers 1,5 and 8 stand out. Unfortunately we don't know which product each number represents as it is masked.

```
data.groupby('Product_Category_1').mean()['Purchase'].plot(kind='b
ar',figsize=(18,5))
plt.title("Product_Category_1 and Purchase Mean Analysis")
plt.show()
```

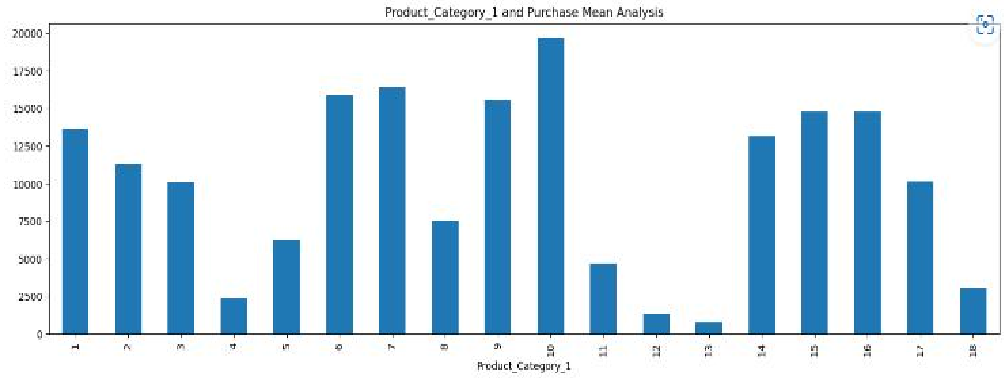

fig 4.16

```
data.groupby('Product_Category_1').sum()['Purchase'].plot(kind='ba
r',figsize=(18,5))
plt.title("Product_Category_1 and Purchase Analysis")
plt.show()
```

If we see the value spent on average for Product_Category_1 you see that although there were more products bought for categories 1,5,8 the average amount spent for those three is not the highest. It is interesting to see other categories appearing with high purchase values despite having low impact on sales number.

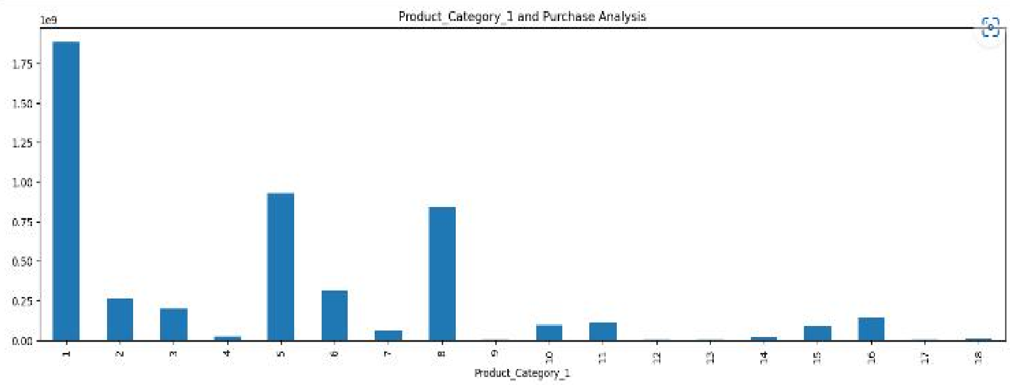

fig 4.17

The distribution that we saw for this predictor previously appears here. For example, those three products have the highest sum of sales since their were three most sold products.

Product_Category_2

```
plt.figure(figsize=(18,5))
sns.countplot(data['Product_Category_2'])
plt.show()
```

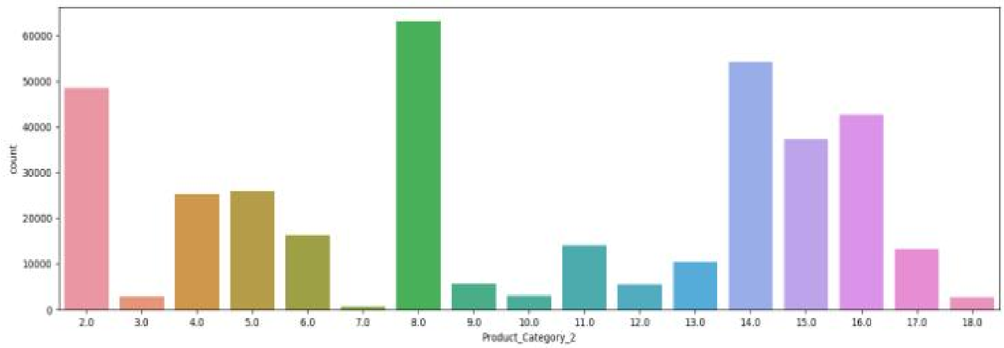

fig 4.18

Product_Category_3

```
plt.figure(figsize=(18,5))
sns.countplot(data['Product_Category_3'])
plt.show()
```

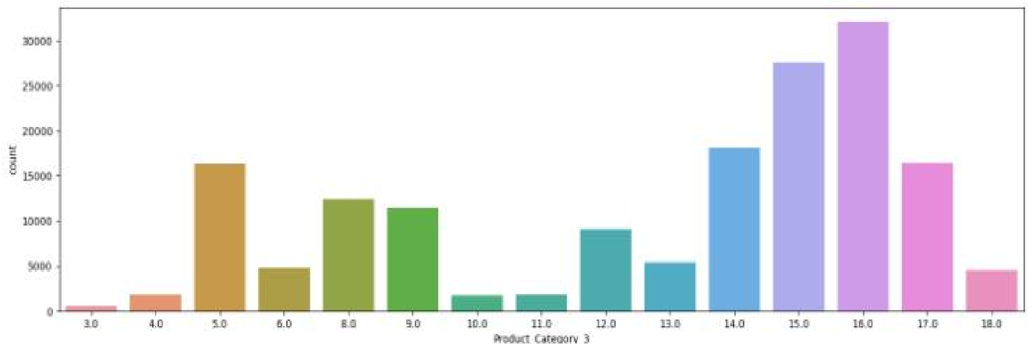

fig 4.19

HeatMap

```
data.corr()
```

Out[36]:

	User_ID	Occupation	Marital_Status	Product_Category_1	Product_Category_2	Product_Category_3	Purchase
User_ID	1.000000	-0.023024	0.018732	0.003687	0.001471	0.004045	0.005389
Occupation	-0.023024	1.000000	0.024691	-0.008114	-0.000031	0.013452	0.021104
Marital_Status	0.018732	0.024691	1.000000	0.020546	0.015116	0.019452	0.000129
Product_Category_1	0.003687	-0.008114	0.020546	1.000000	0.540423	0.229490	-0.314125
Product_Category_2	0.001471	-0.000031	0.015116	0.540423	1.000000	0.543544	-0.209973
Product_Category_3	0.004045	0.013452	0.019452	0.229490	0.543544	1.000000	-0.022257
Purchase	0.005389	0.021104	0.000129	-0.314125	-0.209973	-0.022257	1.000000

```
sns.heatmap(data.corr(),annot=True)
plt.show()
```

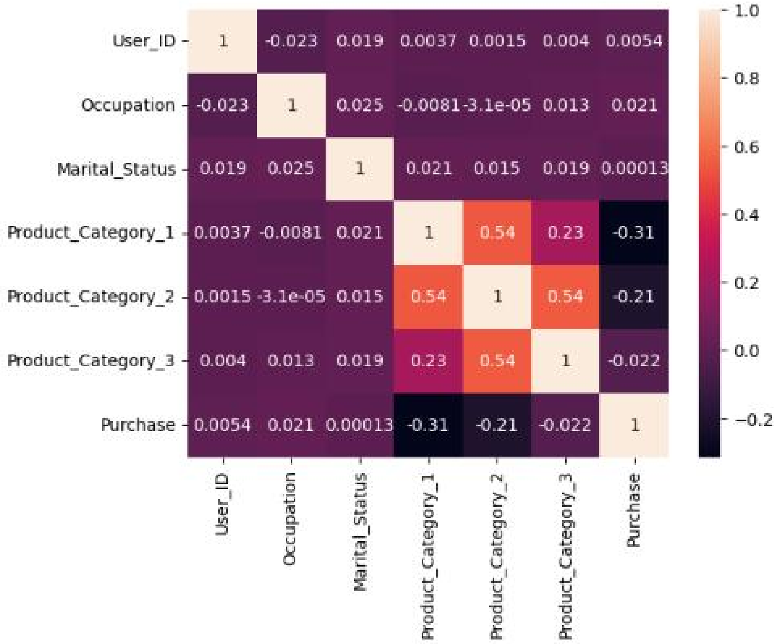

fig 4.20

There is a some corellation between the product category groups.

4.2. Implementation of machine learning models

```
df = data.copy()

#Dummy Variables:
df = pd.get_dummies(df, columns=['Stay_In_Current_City_Years'])
```

Encoding the categorical variables

```
from sklearn.preprocessing import LabelEncoder
lr = LabelEncoder()

df['Gender'] = lr.fit_transform(df['Gender'])

df['Age'] = lr.fit_transform(df['Age'])

df['City_Category'] = lr.fit_transform(df['City_Category'])

df['Product_Category_2']
=df['Product_Category_2'].fillna(0).astype('int64')
df['Product_Category_3']
=df['Product_Category_3'].fillna(0).astype('int64')

df.isnull().sum()
```

```
Out[49]: User_ID                        0
         Product_ID                     0
         Gender                         0
         Age                            0
         Occupation                     0
         City_Category                  0
         Marital_Status                 0
         Product_Category_1             0
         Product_Category_2             0
         Product_Category_3             0
         Purchase                       0
         Stay_In_Current_City_Years_0   0
         Stay_In_Current_City_Years_1   0
         Stay_In_Current_City_Years_2   0
         Stay_In_Current_City_Years_3   0
         Stay_In_Current_City_Years_4+  0
         dtype: int64
```

```
df.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 537577 entries, 0 to 537576
Data columns (total 16 columns):
 #   Column                         Non-Null Count   Dtype
---  ------                         --------------   -----
 0   User_ID                        537577 non-null  int64
 1   Product_ID                     537577 non-null  object
 2   Gender                         537577 non-null  int32
 3   Age                            537577 non-null  int32
 4   Occupation                     537577 non-null  int64
 5   City_Category                  537577 non-null  int32
 6   Marital_Status                 537577 non-null  int64
 7   Product_Category_1             537577 non-null  int64
 8   Product_Category_2             537577 non-null  int64
 9   Product_Category_3             537577 non-null  int64
 10  Purchase                       537577 non-null  int64
 11  Stay_In_Current_City_Years_0   537577 non-null  uint8
 12  Stay_In_Current_City_Years_1   537577 non-null  uint8
 13  Stay_In_Current_City_Years_2   537577 non-null  uint8
 14  Stay_In_Current_City_Years_3   537577 non-null  uint8
 15  Stay_In_Current_City_Years_4+  537577 non-null  uint8
dtypes: int32(3), int64(7), object(1), uint8(5)
memory usage: 41.5+ MB
```

Dropping the irrelevant columns

```
df = df.drop(["User_ID","Product_ID"],axis=1)
```

Splitting data into independent and dependent variables

```
X = df.drop("Purchase",axis=1)

y=df['Purchase']

from sklearn.model_selection import train_test_split
X_train, X_test, y_train, y_test = train_test_split(X, y, test_size=0.3, random_state=123)
```

Linear Regression

```
from sklearn.linear_model import LinearRegression

lr = LinearRegression()
lr.fit(X_train,y_train)
```

```
Out[57]: LinearRegression()
```

```
lr.intercept_
```

```
Out[58]: 9392.78408085134
```

```
lr.coef_
```

```
Out[59]: array([ 481.31865517,  107.64157841,    5.13000529,  336.95273272,
                 -63.3778221 , -317.00345883,    7.9238667 ,  148.12973485,
                 -32.78694504,   -1.66930455,   34.63808922,  -12.31969823,
                  12.13785861])
```

```
y_pred = lr.predict(X_test)

from sklearn.metrics import mean_absolute_error,mean_squared_error, r2_score

mean_absolute_error(y_test, y_pred)
```

```
Out[62]: 3540.3993734221553
```

```
mean_squared_error(y_test, y_pred)
```

```
Out[63]: 21342855.359792948
```

```
r2_score(y_test, y_pred)
```

```
Out[64]: 0.13725207799200811
```

```
from math import sqrt
print("RMSE of Linear Regression Model is ",sqrt(mean_squared_error(y_test, y_pred)))
```

```
RMSE of Linear Regression Model is  4619.8328281219165
```

Decision Tree Regressor

```
from sklearn.tree import DecisionTreeRegressor

# create a regressor object

regressor = DecisionTreeRegressor(random_state = 0)


regressor.fit(X_train, y_train)
```

```
Out[72]: DecisionTreeRegressor(random_state=0)
```

```
dt_y_pred = regressor.predict(X_test)


mean_absolute_error(y_test, dt_y_pred)
```

```
Out[74]: 2403.1409470088884
```

```
mean_squared_error(y_test, dt_y_pred)
```

```
Out[75]: 11535194.335807195
```

```
r2_score(y_test, dt_y_pred)
```

```
Out[76]: 0.5337097695969879
```

```
from math import sqrt

print("RMSE of Linear Regression Model is
",sqrt(mean_squared_error(y_test, dt_y_pred)))
```

```
RMSE of Decision Tree Regressor is  3396.3501491759052
```

Random Forest Regressor

```
from sklearn.ensemble import RandomForestRegressor


# create a regressor object

RFregressor = RandomForestRegressor(random_state = 0)


RFregressor.fit(X_train, y_train)
```

```
Out[79]: RandomForestRegressor(random_state=0)
```

```
rf_y_pred = RFregressor.predict(X_test)

mean_absolute_error(y_test, rf_y_pred)
```

```
Out[81]: 2244.4372062473967
```

```
mean_squared_error(y_test, rf_y_pred)
```

```
Out[82]: 9432129.31992627
```

```
r2_score(y_test, rf_y_pred)
```

```
Out[83]: 0.6187225264053897
```

```
from math import sqrt

print("RMSE of Linear Regression Model is
",sqrt(mean_squared_error(y_test, rf_y_pred)))
```

```
RMSE of Random Forest Regressor is  3071.1771879730854
```

XGBoost Regressor

```
from xgboost.sklearn import XGBRegressor

xgb_reg = XGBRegressor(learning_rate=1.0, max_depth=6,
min_child_weight=40, seed=0)

xgb_reg.fit(X_train, y_train)
```

```
Out[86]: XGBRegressor(base_score=None, booster=None, callbacks=None,
                      colsample_bylevel=None, colsample_bynode=None,
                      colsample_bytree=None, early_stopping_rounds=None,
                      enable_categorical=False, eval_metric=None, feature_types=None,
                      gamma=None, gpu_id=None, grow_policy=None, importance_type=None,
                      interaction_constraints=None, learning_rate=1.0, max_bin=None,
                      max_cat_threshold=None, max_cat_to_onehot=None,
                      max_delta_step=None, max_depth=6, max_leaves=None,
                      min_child_weight=40, missing=nan, monotone_constraints=None,
                      n_estimators=100, n_jobs=None, num_parallel_tree=None,
                      predictor=None, random_state=None, ...)
```

```
xgb_y_pred = xgb_reg.predict(X_test)

mean_absolute_error(y_test, xgb_y_pred)
```

```
Out[88]: 2154.9546373075887
```

```
mean_squared_error(y_test, xgb_y_pred)
```

```
Out[89]: 8290522.888016781
```

```
r2_score(y_test, xgb_y_pred)
```

```
Out[90]: 0.6648699870088253
```

```
from math import sqrt

print("RMSE of Linear Regression Model is
",sqrt(mean_squared_error(y_test, xgb_y_pred)))
```

```
RMSE of XGBoost Regressor is  2879.326811603153
```

The ML algorithm that performs the best was XGBoost Regressor Model with RMSE = 2879

Chapter-5
Conclusion and Future Work

5.1. Scope of Improvement

With traditional methods not being of much help to business growth in terms of revenue, the use of Machine learning approaches proves to be an important point for the shaping of the business plan taking into consideration the shopping pattern of consumers.

Projection of sales concerning several factors including the sale of last year helps businesses take on suitable strategies for increasing the sales of goods that are in demand.

5.2. Summary

1. Approximately, 75% of the number of purchases are made by Male users and rest of the 25% is done by female users. This tells us the Male consumers are the major contributors to the number of sales for the retail store.On average the male gender spends more money on purchase contrary to female, and it is possible to also observe this trend by adding the total value of purchase.
2. When we combined Purchase and Marital_Status for analysis, we came to know that Single Men spend the most during the Black Friday. It also tells that Men tend to spend less once they are married. It maybe because of the added responsibilities.
3. For Age feature, we observed the consumers who belong to the age group 25-40 tend to spend the most.
4. There is an interesting column Stay_In_Current_City_Years, after analyzing this column we came to know the people who have spent 1 year in the city tend to spend the most. This is understandable as, people who have spent more than 4 years in the city are generally well settled and are less interested in buying new things as compared to the people new to the city, who tend to buy more.

5. When examining which city the product was purchased to our surprise, even though the city B is majorly responsible for the overall sales income, but when it comes to the above product, it majorly purchased in the city C.
6. . Splitted dataset into into random train and test subset of ratio 80:20
7. Implemented multiple supervised models such as Linear Regressor, Decision Tree Regressor, Random Forest Regressor.
8. Root Mean Square Error (RMSE) is a standard way to measure the error of a model in predicting quantitative data. It's the square root of the average of squared differences between prediction and actual observation.

5.3. Conclusion

Implanted multiple supervised models such as Linear Regressor,Decision Tree Regressor, Random Forest Regressor and XGBOOST Regressor. Out of these supervised models, based on the RMSE scores XGBRegressor/XGBOOST Regressor was the best performer with a score of 2879. Thus the proposed model will predict the customer purchase on Black Friday and give the retailer insight into customer choice of products. This will result in a discount based on customer-centric choices thus increasing the profit to the retailer as well as the customer.

References

1. H. V. Ramachandra, G. Balaraju, A. Rajashekar and H. Patil, "Machine Learning Application for Black Friday Sales Prediction Framework," 2021 International Conference on Emerging Smart Computing and Informatics (ESCI), Pune, India, 2021, pp. 57-61, doi: 10.1109/ESCI50559.2021.9396994.
2. S. Alagarsamy, K. G. Varma, K. Harshitha, K. Hareesh and K. Varshini, "Predictive Analytics for Black Friday Sales using Machine Learning Technique," 2023 International Conference on Intelligent Data Communication Technologies and Internet of Things (IDCIoT), Bengaluru, India, 2023, pp. 389-393, doi: 10.1109/IDCIoT56793.2023.10053454.
3. S. Patil, O. Nankar, R. Agrawal, K. Sharma, S. Awasthi and N. Jha, "Black Friday Sales Prediction using Supervised Machine Learning," 2023 International Conference on Artificial Intelligence and Smart Communication (AISC), Greater Noida, India, 2023, pp. 1006-1012, doi: 10.1109/AISC56616.2023.10084959.
4. Amruta Aher , Rajeswari Kannan , Sushma Vispute, 2021, Data Analysis and Price Prediction of Sales using Machine Learning Techniques, INTERNATIONAL JOURNAL OF ENGINEERING RESEARCH & TECHNOLOGY (IJERT) Volume 10, Issue 07 (July 2021),