BLACK FRIDAY SALES PROJECT

SUBMITTED BY:
ATISH KALANGUTKAR

ACKNOWLEDGMENT

I would like to express my deepest gratitude to my SME (Subject Matter Expert) Shwetank Mishra as well as Flip Robo Technologies who gave me the opportunity to do this project on Black Friday sales, which also helped me in doing lots of research wherein I came to know about so many new things.

References:

I have also used few external resources that helped me to complete this project successfully. I ensured that I learn from the samples and modify things according to my project requirement. All the external resources that were used in creating this project are listed below:

1. https://www.google.com/
2. https://scikit-learn.org/stable/index.html
3. https://github.com/
4. https://www.analyticsvidhya.com/
5. www.researchgate.net/

INTRODUCTION

Business Problem Framing:

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

Conceptual Background of the Domain Problem:

The concept of Black Friday originated in the United States and refers to the day after Thanksgiving, which traditionally marks the beginning of the holiday shopping season. Retailers offer steep discounts to attract customers, with the goal of boosting sales and clearing out inventory before the end of the year.Over time, Black Friday has evolved into a global phenomenon, with retailers across the world participating in the event.

Black Friday presents both opportunities and challenges for retailers. On one hand, it can be a lucrative sales event that generates significant revenue and attracts new customers. On the other hand, it requires careful planning and execution to avoid stock outs, manage logistics, and

optimize pricing and promotion strategies. Additionally, Black Friday is a highly competitive and dynamic environment, with retailers vying for the attention and wallets of consumers.

To succeed in Black Friday sales, retailers need to understand consumer behavior and preferences, identify trends and patterns, and develop targeted marketing and sales strategies. With the increasing availability of data and advanced analytics tools, retailers can leverage machine learning and statistical techniques to gain insights and make data-driven decisions that can improve their bottom line and enhance the shopping experience for their customers. This project aims to explore a dataset of Black Friday sales to uncover such insights and develop strategies that can help retailers optimize their performance during this critical sales event.

Review of literature:

Black Friday is a well-established phenomenon in the retail industry, and there has been a growing body of literature that explores various aspects of the event. Researchers have examined consumer behavior and preferences, marketing and promotion strategies, and the impact of Black Friday on sales and revenue.

One common theme in the literature is the importance of price discounts in driving consumer behavior during Black Friday. Studies have found that consumers are highly price-sensitive during this event, and that deeper discounts are associated with higher sales volume. Additionally, research has shown that consumers are more likely to make unplanned purchases during Black Friday, suggesting that effective

marketing and merchandising strategies can play a significant role in driving sales.

Another area of research has focused on the impact of Black Friday on the overall retail landscape. Some studies have suggested that Black Friday may cannibalize sales from other periods, while others have argued that it can stimulate overall consumer spending and contribute to economic growth. There has also been a debate about the ethics and sustainability of Black Friday, with some researchers questioning the environmental and social impact of the event.In recent years, the rise of e-commerce has had a significant impact on Black Friday sales. Online retailers have increasingly taken part in the event, and mobile shopping has become a key driver of sales growth. This has led to a greater emphasis on digital marketing and personalized promotions, as well as challenges related to logistics and inventory management.

Overall, the literature suggests that Black Friday is a complex and dynamic event that requires careful analysis and strategic planning. By leveraging data and advanced analytics techniques, retailers can gain insights into consumer behavior and preferences, identify opportunities for optimization, and develop effective marketing and sales strategies that can help them succeed during this critical sales period.

Motivation for the Problem Undertaken:

The motivation for the Black Friday Sales Project is to increase revenue during the Black Friday period, which is a crucial time for retailers. Black Friday is known for generating high sales volume, and businesses that do not capitalize on this opportunity may miss out on significant revenue.

Furthermore, the COVID-19 pandemic has shifted consumer behavior towards e-commerce, with more people shopping online. This trend is expected to continue, making it even more important for businesses to have a strong online presence during the Black Friday period.

Overall, the motivation for the project is to ensure that the business is well-positioned to take advantage of the opportunities presented during Black Friday, while also addressing the challenges that come with increased demand and changing consumer behavior.

ANALYTICAL PROBLEM FRAMING

Mathematical/ Analytical Modeling of the Problem:

We need to develop an efficient and effective machine learning model which predicts the purchased amount of customers against various products. So 'Purchase' is our target variable which is continuous in nature. So it is a Regression problem where we need to use regression algorithms to predict the results.

Data Sources and their formats:

Data set provided by Flip Robo was in the format of CSV (Comma Separated Values). The dimension of the dataset is 550068 rows and 12 columns including target variable "Purchase". The dataset contains both categorical and numerical data type. There were two datasets that were provided. One is training data and one is testing data.

1) Train file will be used for training the model, i.e., the model will learn from this file. It contains all the independent variables and the target variable. Size of training set: 550068 records.

2) Test file contains all the independent variables, but not the target variable. We will apply the model to predict the target variable for the test data. Size of test set: 233599records. Data description is as follows:

Variables	Definition
User_ID	It gives the information about the User
Product_ID	It gives the information about the Product
Gender	It gives the information about the gender or sex of the user
Age	It gives the information about the Age of the user
Ocuupation	It gives the information about the occupation of the user which is in a encoded form
City_Category	It gives the information about the categories of the city
Stay_in_current_city_years	It gives the information about how long a user staying in a particular city
Marital_status	It gives the information about the marital status of the user which is in a encoded form
Product_category_1	It gives the information about the product which belongs to category 1 which is in a encoded form
Product_category_2	It gives the information about the product which belongs to category 2 which is in a encoded form
Product_category_3	It gives the information about the product which belongs to category 3 which is in a encoded form

Data Pre-Processing Done:

Data pre-processing is the process of converting raw data into a well readable format to be used by machine learning model. I have used following pre-processing steps:

- Importing required libraries and loading the dataset.
- Checked some statistical information such as shape of the dataset,unique values and number of unique values present, type of data present in the columns etc.
- Checked for null values and found that there were about 173638 null values present in the column product category 2 and 383247 null values in column product category 3. Further by using median method null values were treated from the dataset.
- Also there was one unwanted column named User ID and Product ID which was dropped as it was just having numerical values which was irrelevant at the time of prediction.
- While checking for duplicates it was found that there were 5261 duplicates present in the dataset. Further it was dropped from the dataset.
- Done feature engineering on column 'Age' where 55+ was replaced with age category 55-60 and after that all the age categories was replaced by 'Child' for 0-17, 'Teen' for 18-25, 'Adult' for 26-35, 36-45, 46-50 and for 'Old' for 51-55 and 55-60. And also on column stay in current city years it was having + sign, because of that it was showing data type as object. So replaced them by empty spaces and converted the data type to integer.
- Than separated categorical and numeric features according to the data types.

- Performed uni-variate, bi-variate and multi-variate analysis by plotting Dist plot, Count plot, Pie plot, Bar plot and Line plot.
- Than checked for Top 100 highest priced products from the dataset and Cheapest 100 priced products.
- Encoded the categorical features by using Label encoded method.
- Used bar plot to check the correlation between the features and the label and after with the help of heatmap checked whether there is multicollinearity present or not among the features. So found that there was no multicollinearity problem within the features.
- Checked whether there are any outliers present or not using box plot on numerical data columns and found that in column Product category 1 and product category 3 was having outliers so by using z-score method removed outliers.
- Checked for skewness on numerical data columns and found that column Product category 1 and product category 3 was having skewness, so by using power transform method treated the skewness of that particular columns.
- Separated feature and label data and than feature scaling was done by using power transform method to avoid any kind of data biasness.
- After plotting heatmap, it was seen than there was no multicollinearity problem within the features, but just to cross verify vif method was used. After using vif method it was found that all the columns values were below 5. So it was confirmed that there was no multicollinearity problem within the features.

Data Inputs-Logic-Output Relationships:

The dataset consists of label and features. The features are independent and label is dependent.

I have checked the correlation between the label and the features using heatmap and bar plot in which both positive and negative correlation between the label and features were there.

Hardware and Software Requirements and Tools Used:

To build a machine learning projects it is important to have the following hardware and software requirements tools.

Hardware Required:

- RAM: 8GB
- CPU: AMD RYZEN 5 4600H
- GPU: AMD Radeon ™ Vega 8 Graphics and NVIDIA GeForce GTX 1650 Ti

Software Required:

- Distribution: Anaconda Navigator
- Programming Language: Python
- Browser based language shell: Jupyter Notebook

Libraries Required:

Pre-processing and Visualization

```
#importing required libraries
import pandas as pd
import numpy as np
import seaborn as sns
import matplotlib.pyplot as plt
import warnings
warnings.filterwarnings('ignore')
```

- Encoding

```
#Importing required libraries
from sklearn.preprocessing import LabelEncoder
```

- To remove outliers

```
#Importing required libraries
from scipy.stats import zscore
```

- To remove skewness

```
#Importing required libraries
from sklearn.preprocessing import power_transform
```

- To standardize the data

```
#Standardiziing the data
#Importing require libraries
from sklearn.preprocessing import StandardScaler
```

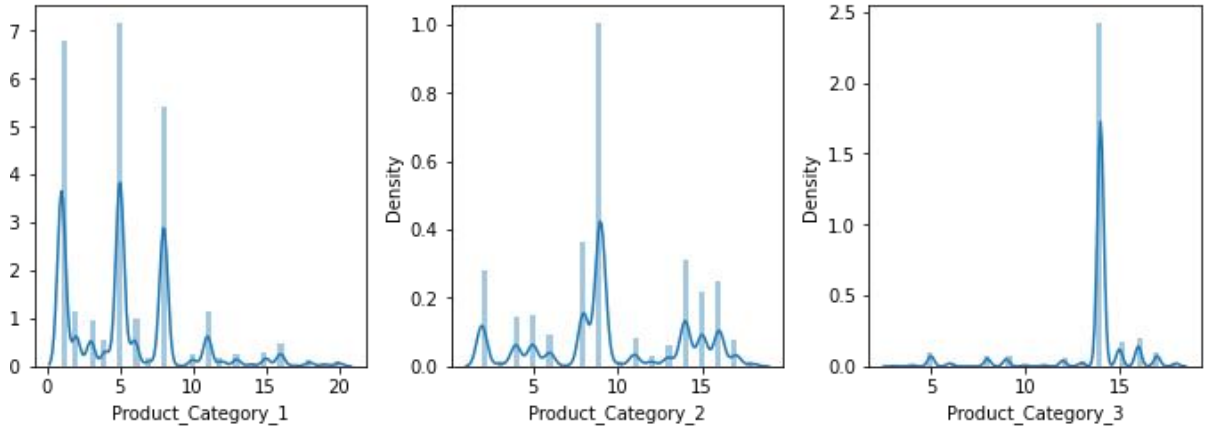
- Vif method

```
#Importing require libraries
from statsmodels.stats.outliers_influence import variance_inflation_factor
```

Visualization:

I used pandas profiling to get the over viewed visualization on the pre processed data. Pandas is an open source python module with which we can do an exploratory data analysis to get detailed description of the features and it helps in visualizing and understanding the distribution of each variable. I have analyzed the data using distribution plot, pie plot,count plot, bar plot, line plots and box plots to get the relation between the features and label.

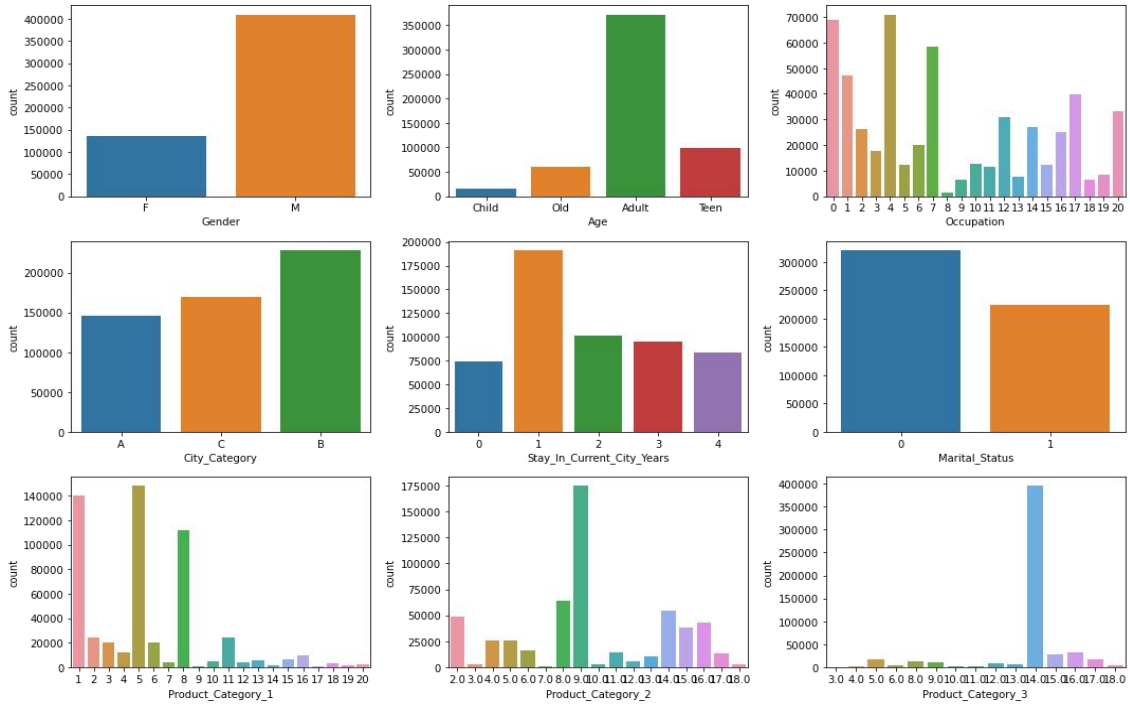
Distribution Plots:



In above plots we can see that:

- So in all three column i can see skewness as th data is not normally distributed. So to cross verify it io further i will check the skewness.
- It can can be seen that outliers also can be present in all three columns which will be cross verify with the help of box plots.

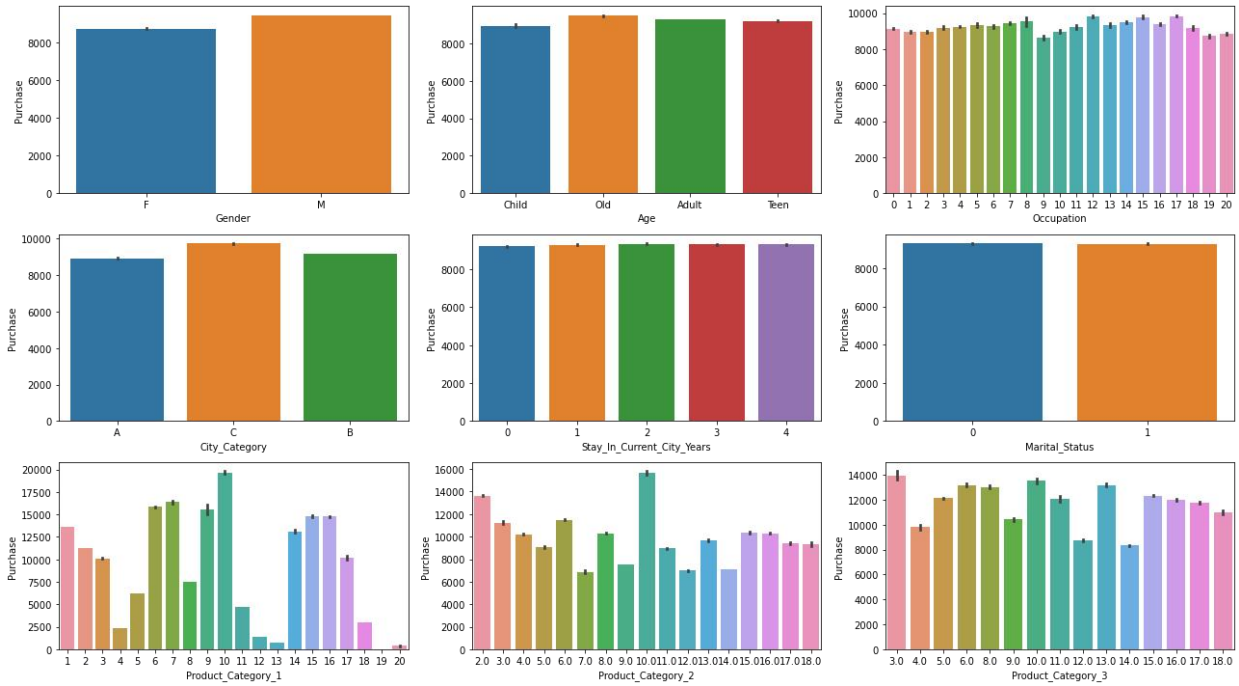
Count Plots:



In above plots we can see that:

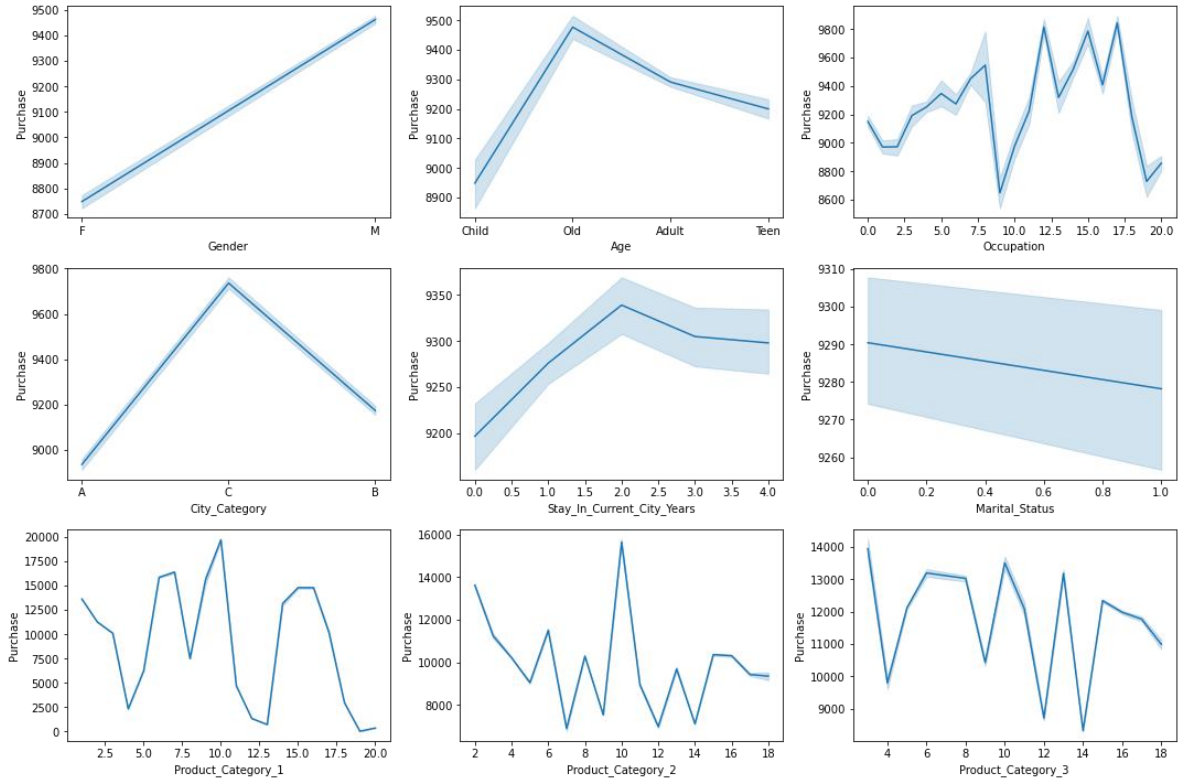
- In plot Gender we can see that the male users are more than the female users.
- In plot Age we can see that most of the users are Adult as compared to other age category.
- In plot Occupation most of the users h=are having value as 4.
- In the plot City category we can see that most of the users are from city category B.
- In the plot stay in current city years most of the users are staying in a particular city from past one year.
- In the plot product category 1, most of the users are purchasing product having value as 5 followed by 1 and 8.
- In the plot product category 2, most of the users are purchasing product having value as 9.0.
- In the plot product category 3, most of the users are purchasing product having value as 14.0.

Bar Plots:



- In the above plot Gender vs Purchase we can see that Male users have purchased more than the Female users.
- In the above plot Age vs Purchase we can see that users which are old have purchased more as compared to other age category users.
- In the above plot Occupation vs Purchase we can see that users with occupation value as 12, 15 and 17 have purchased more.
- In the above plot City category vs Purchase we can see that users from city category C have purchased more than that of other city categories.
- In the above plot stay in current city in years vs Purchase we can see that users staying their particular cities from past 0, 1, 2, 3, and 4 years have purchased equally.
- In the above plot Marital status vs Purchase we can see that both the users who are married as well as who are not married have purchase equally.
- In the above plot Product category 1 vs Purchase users have purchased more with value 10 from product category 1.
- In the above plot Product category 2 vs Purchase users have purchased more with value 10 from product category 2.
- In the above plot Product category 3 vs Purchase users have purchased more with value 3.0 from product category 3.

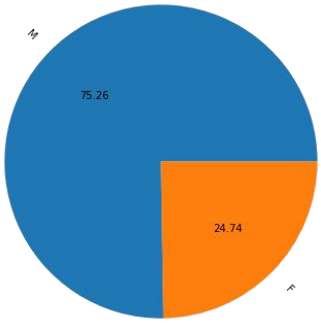
Line Plots:



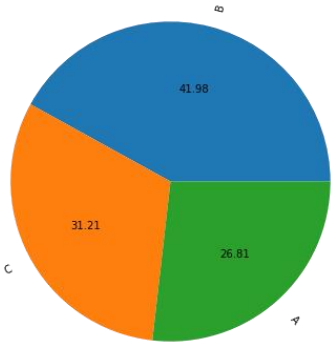
- In the above plot Gender vs Purchase we can see that Male users have purchased more than the Female users.
- In the above plot Age vs Purchase we can see that users which are old have purchased more as compared to other age category users.
- In the above plot Occupation vs Purchase we can see that users with occupation value as 12, 15 and 17 have purchased more.
- In the above plot City category vs Purchase we can see that users from city category C have purchased more than that of other city categories.
- In the above plot stay in current city in years vs Purchase we can see that users staying in their particular cities from past 2 years have purchased more.

- In the above plot Marital status vs Purchase we can see that both the users who are not married have purchase more than the users which are married.
- In the above plot Product category 1 vs Purchase users have purchased more with value 10 from product category 1.
- In the above plot Product category 2 vs Purchase users have purchased more with value 10 from product category 2.
- In the above plot Product category 3 vs Purchase users have purchased more with value 3.0 from product category 3.

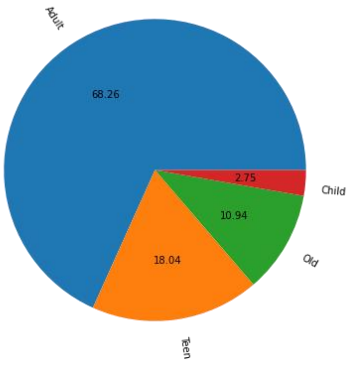
Pie Plots:



- From the above plot we can see that Male users are more than that of female users.

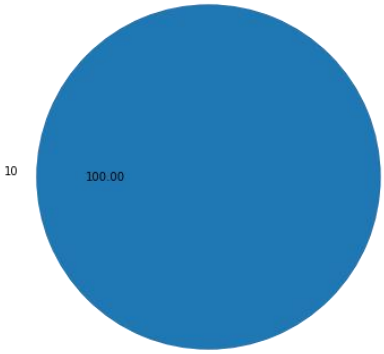


- From the above we can see that Users are more from B city category.

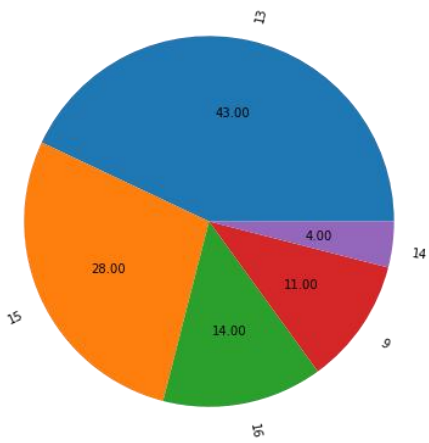


- From the above plot we can see that Adult users are more than that of other age category users.

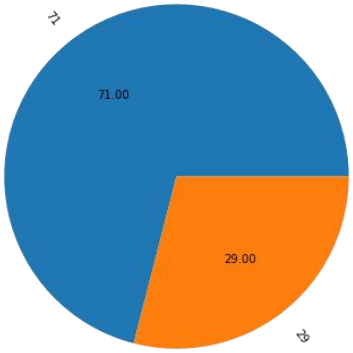
Top 100 highest priced products Pie Plots:



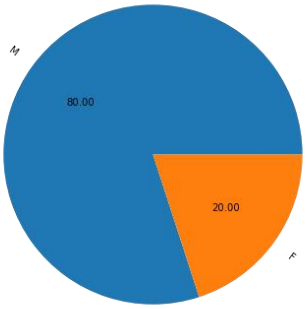
- From the above plot we can see most of the highest priced product from product category 1 is with value 10.



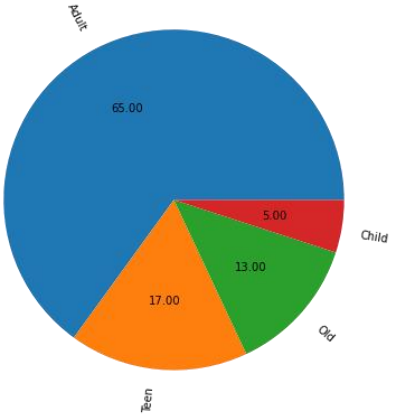
- From the above plot we can see most of the highest priced product from product category 2 is with value 13.



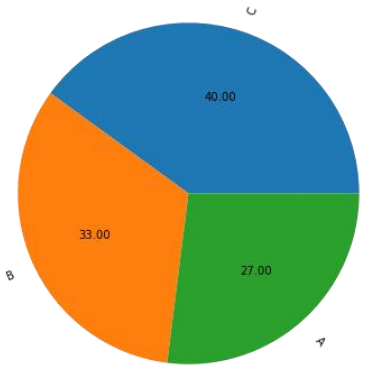
- From the above plot we can see most of the highest priced product from product category 3 is with value 71.



- From the above plot we can see most of the highest priced product are purchased more by male users.

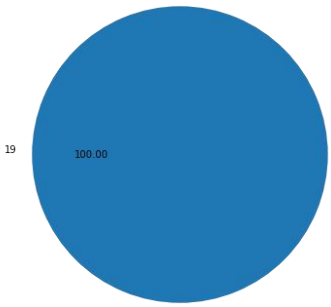


- From the above plot we can see most of the highest priced product are purchased more by Adult users.

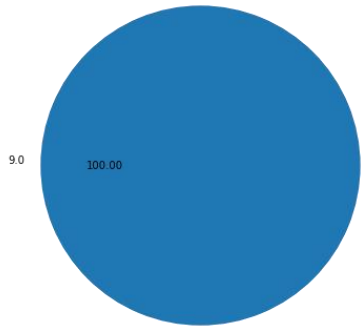


- From the above plot we can see most of the highest priced product are purchased more users from city category C.

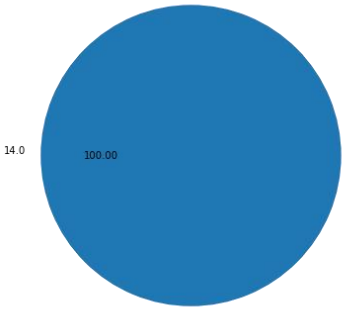
Cheapest 100 priced products Pie Plots:



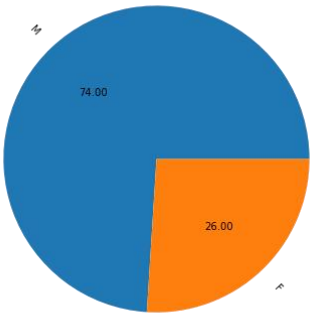
- From the above plot we can see most of the lowest priced product from product category 1 is with value 19.



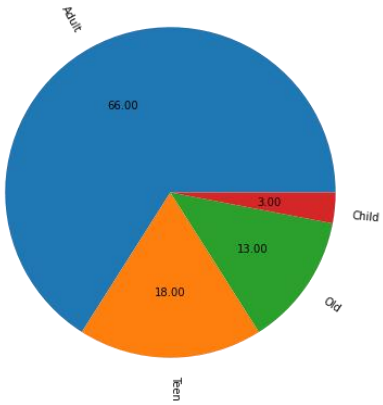
- From the above plot we can see most of the lowest priced product from product category 2 is with value 9.0.



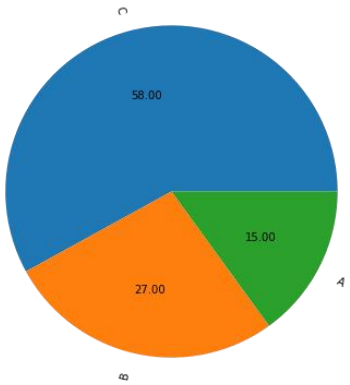
- From the above plot we can see most of the lowest priced product from product category 3 is with value 14.0.



- From the above plot we can see most of the lowest priced product are purchased more by male users.



- From the above plot we can see most of the lowest priced product are purchased more by Adult users.



- From the above plot we can see most of the lowest priced product are purchased more by users from city category C.

CONCLUSION

Key Findings and Conclusions of the Study:

From our detailed analysis we can determine the following:

Most of the users are purchasing product having value as 5 followed by 1 and 8 from the product category 1.

Most of the users are purchasing product having value as 9.0 from the product category 2.

Most of the users are purchasing product having value as 14.0 from the product category 3.

From product category 1, the purchased amount is highest of product having value as 10.

From product category 2, the purchased amount is highest of product having value as 10.0.

From product category 3, the purchased amount is highest of product having value as 3.0.

Limitations of this work and Scope for Future Work:

Limitations: Most of the data which were provided to us was already in the masked form so we were not able carry out the analysis perfectly as we were not knowing the what type of product was there in category 1 and category 2 and category 3.

There were null values present in the dataset specially in column Product category 2 and Product category 3.

And also there were more than 5000 duplicates were present in the dataset.

Future Work: I think that the data should be cleaned and there should no null values and duplicates present in the dataset and the columns from the dataset should have a non encoded values.