Black Friday Sales Prediction

Made By:- Parv Rajesh Thakkar

Problem Statement

- The problem is to understand the customer purchase behavior(specifically, purchase amount) against various products of different categories. The data include various customer demographics, product details and total purchase amount.
- The company's agenda is to predict the purchase amount of customer against various products which can help them create personalized offer for customers against different products.

Understanding Data

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 550068 entries, 0 to 550067
Data columns (total 12 columns):
 #   Column                      Non-Null Count   Dtype  
---  ------                      --------------   -----  
 0   User_ID                     550068 non-null  int64  
 1   Product_ID                  550068 non-null  object 
 2   Gender                      550068 non-null  object 
 3   Age                         550068 non-null  object 
 4   Occupation                  550068 non-null  int64  
 5   City_Category               550068 non-null  object 
 6   Stay_In_Current_City_Years  550068 non-null  object 
 7   Marital_Status              550068 non-null  int64  
 8   Product_Category_1          550068 non-null  int64  
 9   Product_Category_2          376430 non-null  float64
 10  Product_Category_3          166821 non-null  float64
 11  Purchase                    550068 non-null  int64  
dtypes: float64(2), int64(5), object(5)
memory usage: 50.4+ MB
```

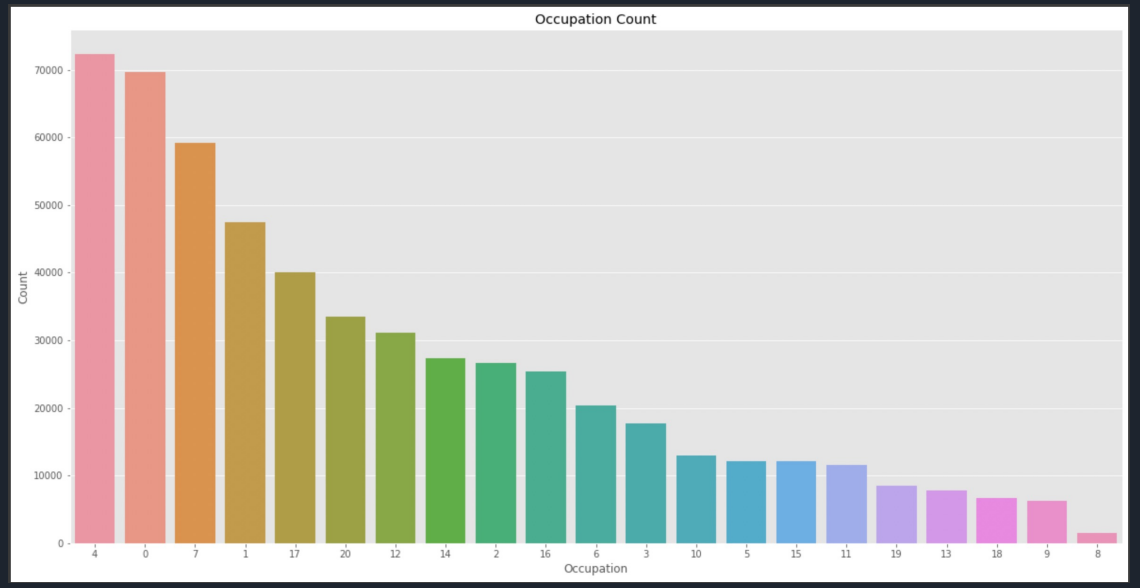
- The first step was to understand the data. I saw that the product_category_2 and product_category_3 had a lot of missing values.
- We can see that from the image on the left that the data has multiple NaN values.
- To solve this problem, we tried to fill the null values by replacing it with 0, mode and median.
- Out of all three mode gave the best results.

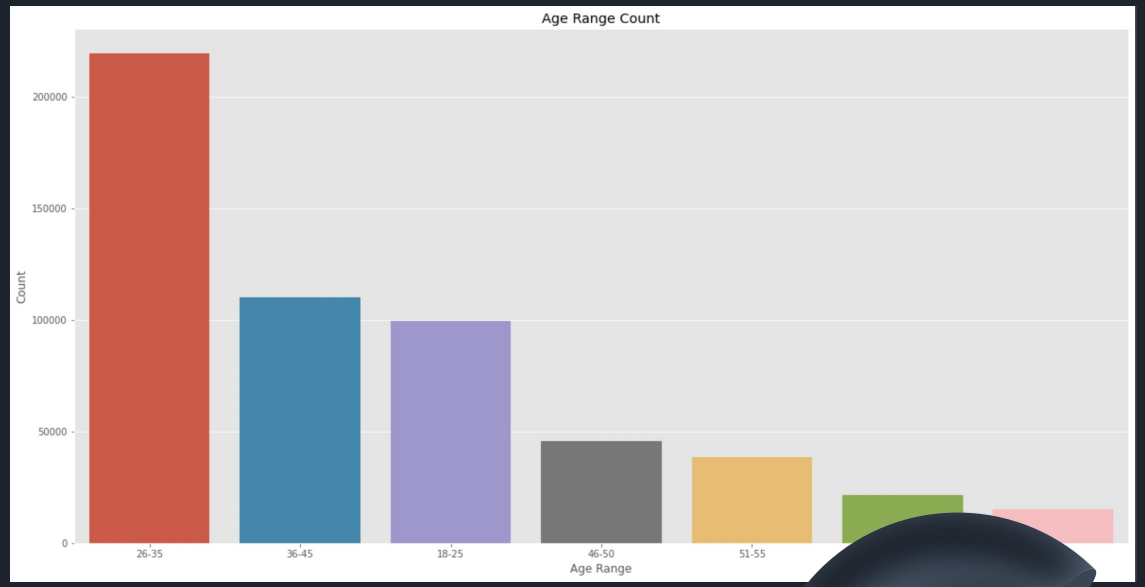
Exploratory Data Analysis

Next to better understand the data we plot the count plots, box plots to analyze the fields better.

From the first graph we can see that the people having occupation 4 visited the store most of the times.

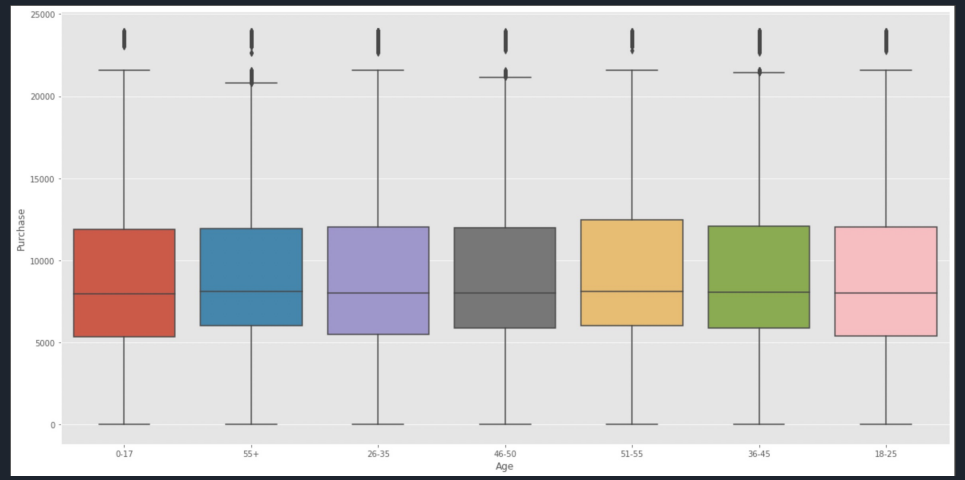
From the second graph we see that people having an age in between 26-35 visit the store the most.

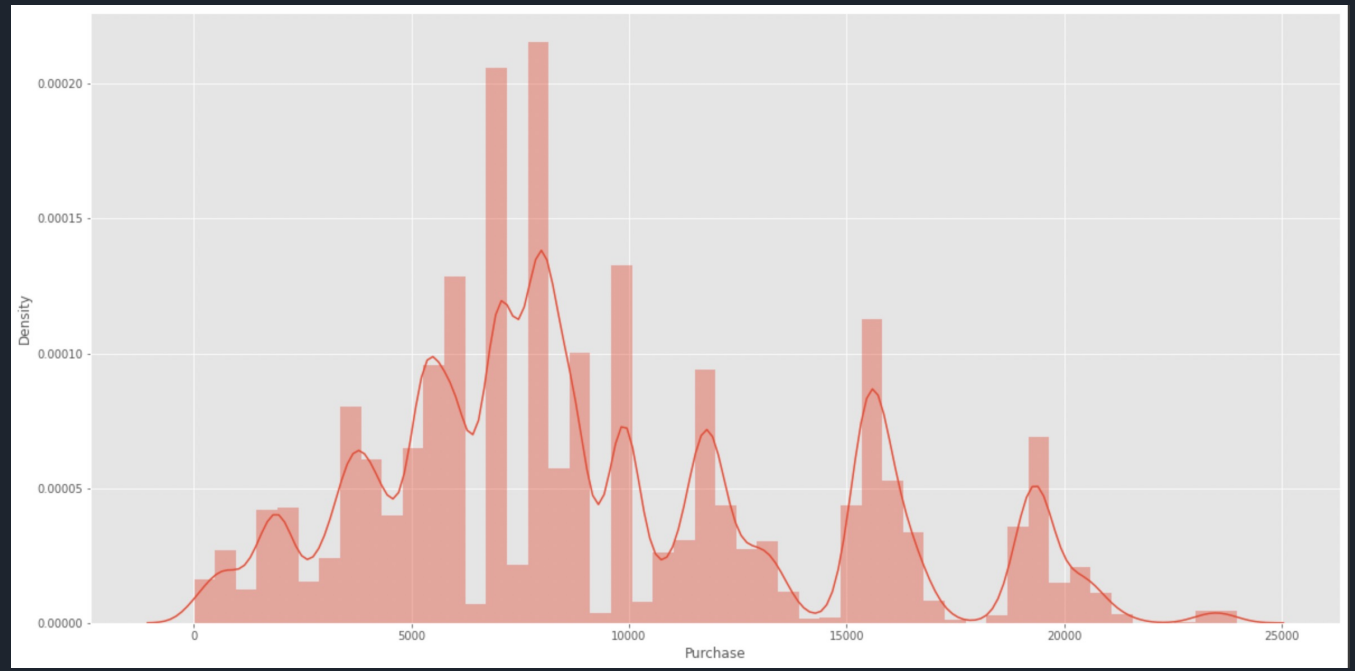




Exploratory Data Analysis Continue..

- From the Age box plot we can conclude that the Age of the customer has very low impact on their purchasing power.
- From the density plot we can see that the purchase amount is normally distributed which means it does neither too low nor too high.





Label Encoding

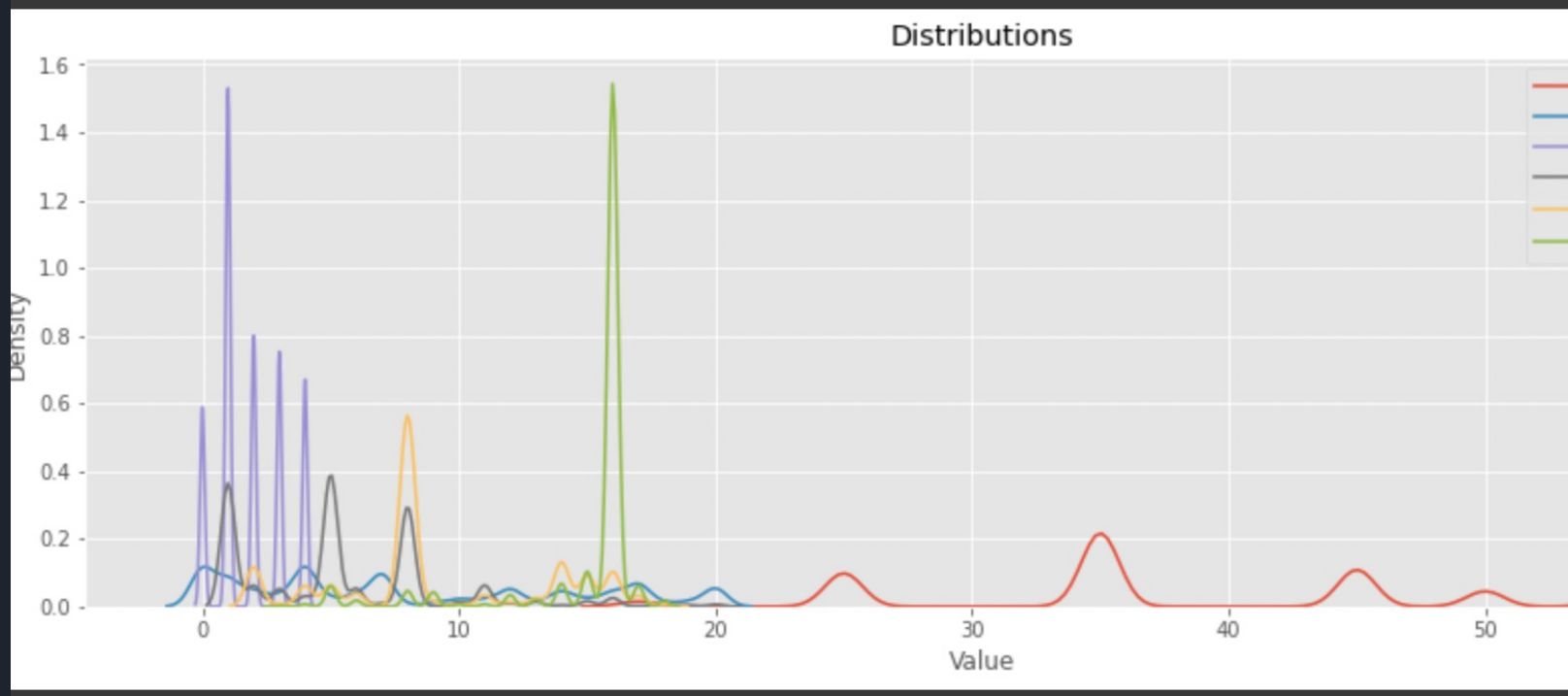
- As we know machine learning model does not work well with strings. Therefore, encoding is necessary to make the model run better.
- The image shows how the encoding has been done in the code.

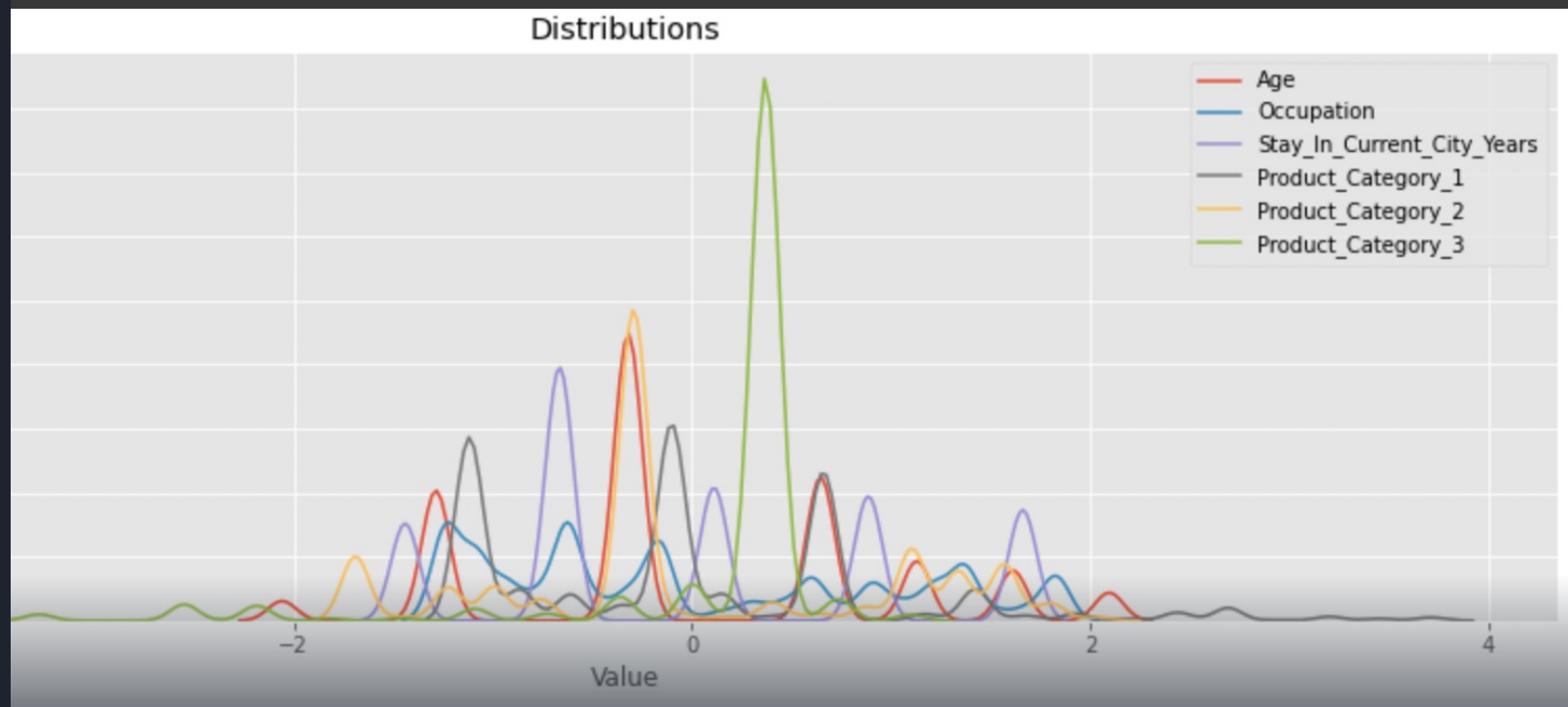
```
def label_encoding(df):
    df['Age'] = df['Age'].replace('0-17', 17)
    df['Age'] = df['Age'].replace('18-25', 25)
    df['Age'] = df['Age'].replace('26-35', 35)
    df['Age'] = df['Age'].replace('36-45', 45)
    df['Age'] = df['Age'].replace('46-50', 50)
    df['Age'] = df['Age'].replace('51-55', 55)
    df['Age'] = df['Age'].replace('55+', 60)
    df['Gender'] = df['Gender'].replace('M', 0)
    df['Gender'] = df['Gender'].replace('F', 1)
    df['City_Category'] = df['City_Category'].replace('A', 0)
    df['City_Category'] = df['City_Category'].replace('B', 1)
    df['City_Category'] = df['City_Category'].replace('C', 2)
    df['Stay_In_Current_City_Years'] = df['Stay_In_Current_City_Years'].replace('4+', 4)
    return df

```

Scaling

- Scaling is important to bring the data in the same range. This makes it easier for the model to work with the data and find relation between the target variable and dependent variables as the data does not vary too much.

- The two graphs show the different values of the columns before and after scaling. After, applying the standard scaler all the values mostly range from -4 to +4.





Further Analysis

- In further, analysis I plotted the correlation graph and created two new features to increase accuracy.
- Also, used groupby and mean to further analyze the data and get some new understandings.
- From the gender analysis we can see that males tend to spend more than females. Also, the people of city category C spend more compared to A & B.

```
1 city_cat_p=train.groupby(['City_Category'])['Purchase'].mean()
2 print(city_cat_p)
```

```
City_Category
0    8911.93922
1    9151.30056
2    9719.92099
Name: Purchase, dtype: float64
```

```
1 gender_p=train.groupby(['Gender'])['Purchase'].mean()
2 print(gender_p)
```

```
Gender
0    9437.52604
1    8734.56577
Name: Purchase, dtype: float64
```

```
1 avg_purchase_per_product=pd.DataFrame(train.groupby(['Product_ID'])['Purchase'].mean())
2 avg_purchase_per_product.reset_index(inplace=True)
3 avg_purchase_per_user=pd.DataFrame(train.groupby(['User_ID'])['Purchase'].mean())
4 avg_purchase_per_user.reset_index(inplace=True)
```

Modelling

- I used three Machine learning models namely Linear regression model, Random Forest model and XGBoost model.
- The accuracy of the model was compared using the root mean squared error. The accuracy is as shown in the table:

Models	Linear Regression	Random Forest	XGBoost
RMSE	2568.5666	3544.0384	2453.2947

THANKYOU!