BLACK FRIDAY SALES PREDICTION

GROUP - 5

TEAM MEMBERS

SANJANA ATHREYA

ASHIQUE NAWAZ CHOUDHURY

HARSH KUMTHEKAR

Table of Contents

ABSTRACT

PROBLEM STATEMENT:-

The dataset comprises of sales transactions captured at a retail store. It's a classic dataset to explore and expand our feature engineering skills and day to day understanding from multiple shopping experiences. This is a regression problem. The dataset has 537,577 rows and 12 columns.

This project analyzes the Black Friday sales data and tries to answer these key business questions :

1. What are maximum products sold?
2. Which Product category has highest sales?
3. Finding the buyer's age group and their product of interest.
4. Finding the marital status of the buyers.
5. Analyzing the gender group, which has high interest in the sales

This can be used to understand the customer purchase behaviour (specifically, purchase amount) against various features like Products of different Categories, Gender, Age, Occupation of Customer, etc.

This project also aims at creating a simple predicting model to predict the purchase amount of customer against various products which will help them to create personalized offer for customers against different products.

METHODS USED:-

1. **MULTIPLE LINEAR REGRESSION**

   In our dataset, there are more than one independent variables. Hence, we have performed the multiple linear regression using the OLS(Ordinary Least Squares) Model.

2. **RIDGE REGRESSION**

   Ridge Regression helps reduce Variance by shrinking parameters and making predictions less sensitive. It can find solution with Cross validation and Ridge Regression Penalty. It is better when all variables are useful for making predictions. In our dataset, we could see that all the variables are important hence we use Ridge for regularization.

RESULTS OBTAINED:-

1. MULTIPLE LINEAR REGRESSION

- p-value ≤ 0.05 indicates strong evidence against the null hypothesis, so you reject the null hypothesis.
- All variables except the Product_Category_1 are positively associated with Sales.
- R-squared score is **0.757**, which means that this model explains 75% of the total variance
- MSE - 26978364.965550404
- MAE – 4059.7367343106857

2. RIDGE REGRESSION

- MSE - 24084394.75832766

MATERIALS AND METHODS

DATASET DESCRIPTION:-

The dataset comprises of sales transactions captured at a retail store. It's a classic dataset to explore and expand our feature engineering skills and day to day understanding from multiple shopping experiences. This is a regression problem. The dataset has 537,577 rows and 12 columns.

Data Dictionary :

Variable	Definition
User_ID	User ID
Product_ID	Product ID
Gender	Sex of the User
Age	Age of the User(in bins)
Occupation	Occupation of the User (Masked)
City_Category	Category of City(A,B,C)
Stay_In_Current_City_Years	Number of years of stay in Current City
Marital_Status	Marital Status of the User
Product_Category_1	Product Category 1(Masked)
Product_Category_2	Product Category 2(Masked)
Product_Category_3	Product Category 3(Masked)
Purchase	Purchase Amount(Target Variable

TOOLS AND TECHNIQUES:-

- Python
- Tableau

RESULTS AND DISCUSSIONS

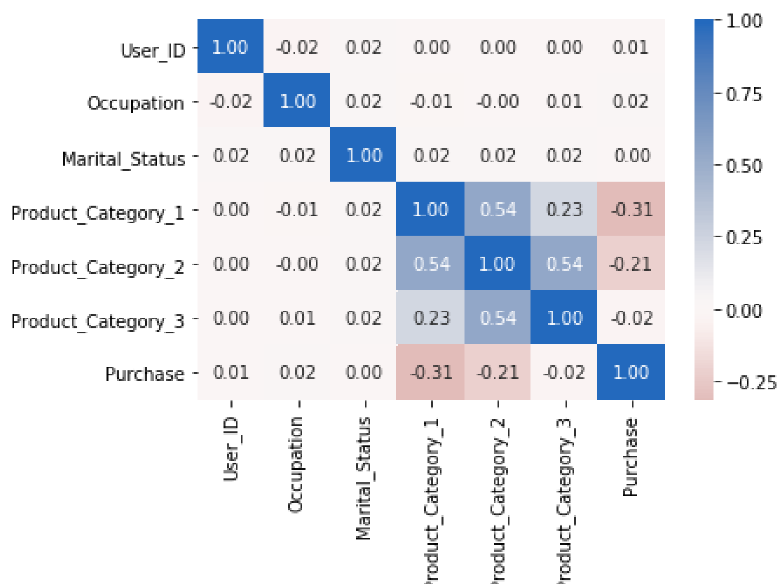
STATISTICAL ANALYSIS:-

Below shows the statistical description of the dataset :

	User_ID	Occupation	Marital_Status	Product_Category_1	Product_Category_2	Product_Category_3	Purchase
count	5.375770e+05	537577.00000	537577.000000	537577.000000	370591.000000	164278.000000	537577.000000
mean	1.002992e+06	8.08271	0.408797	5.295546	9.842144	12.669840	9333.859853
std	1.714393e+03	6.52412	0.491612	3.750701	5.087259	4.124341	4981.022133
min	1.000001e+06	0.00000	0.000000	1.000000	2.000000	3.000000	185.000000
25%	1.001495e+06	2.00000	0.000000	1.000000	5.000000	9.000000	5866.000000
50%	1.003031e+06	7.00000	0.000000	5.000000	9.000000	14.000000	8062.000000
75%	1.004417e+06	14.00000	1.000000	8.000000	15.000000	16.000000	12073.000000
max	1.006040e+06	20.00000	1.000000	18.000000	18.000000	18.000000	23961.000000

Correlation Plot :

Let us plot a correlation plot to check the correlation between the variables in the data set.



From the above plot, we can see that, the correlation between Product Category_1, Product Category_2 and Product Category_3 is greater than 0.5 which means these **variables are moderately correlated**.

ADDRESSING NA AND NULL VALUES:-

- There were 166986 null values in Product_Categorgy_2 and 373299 null values in Product_Category_3.
- The proportion of missing values in the dataset were upto 70%.
- Removing the NaN values will result in 70% loss of data from the data set. This results in the model being biased and it would be underfit.The available alternate approaches are replacing the missing values with mean, mode or fill it with 0.
- The values in Product_Catgeory_2 and Product_Catgeory_3 coulmns are interlinked with values present in Product_Catgeory_1, hence replacing it with mean/mode is not a good strategy.
- Thus, we decided to fill the Nan values with 0.

EXPLORATORY DATA ANALYSIS:-

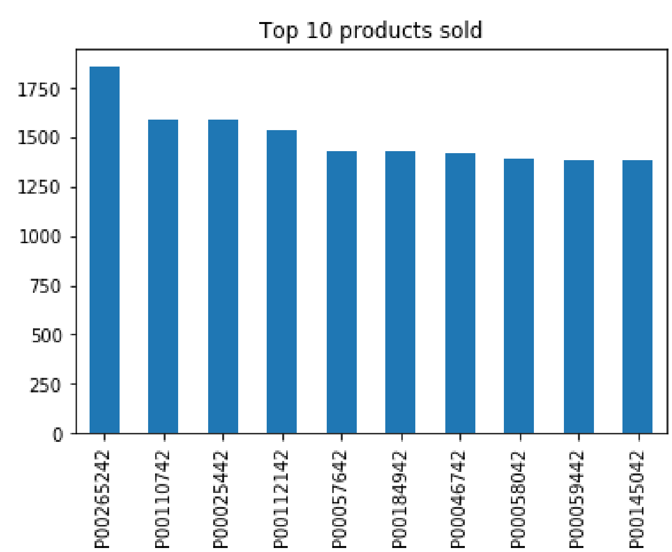
Let us now see how the independent variables are related to the dependent variable by plotting graphs using the matplotlib library in Python.

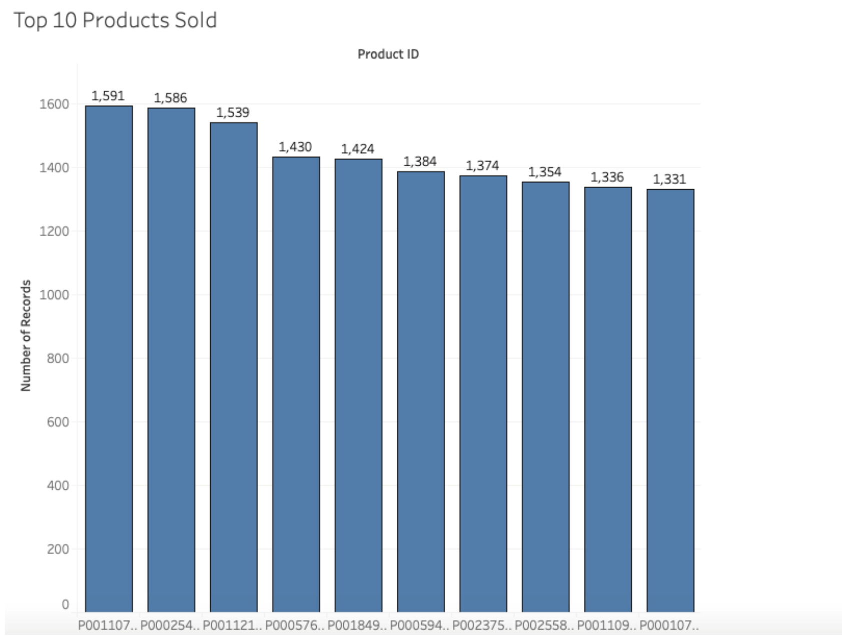
User_ID and Product_ID :
The nunique() method gives us the uniques values present in the column.

From the User_ID, we can conclude that in this specific retail store, during Black Friday, 5,891 different customers have bought something from the store.

Also, from the Product_ID, we can see that there are 3,623 different products that have been sold.

Most Purchased Products :



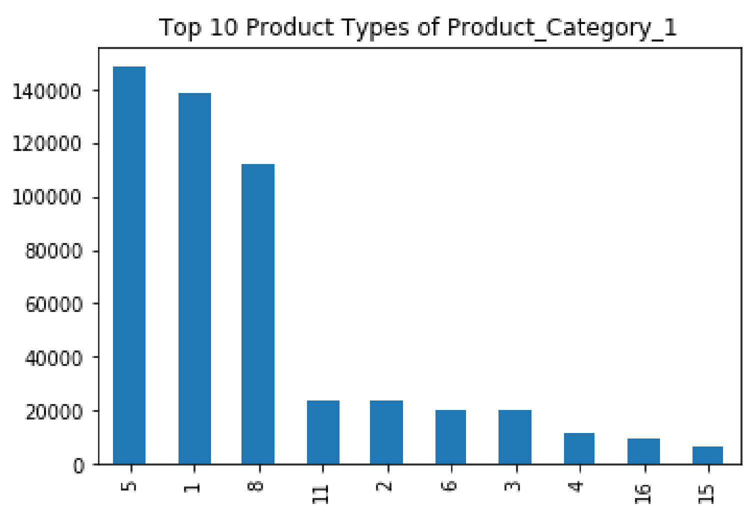


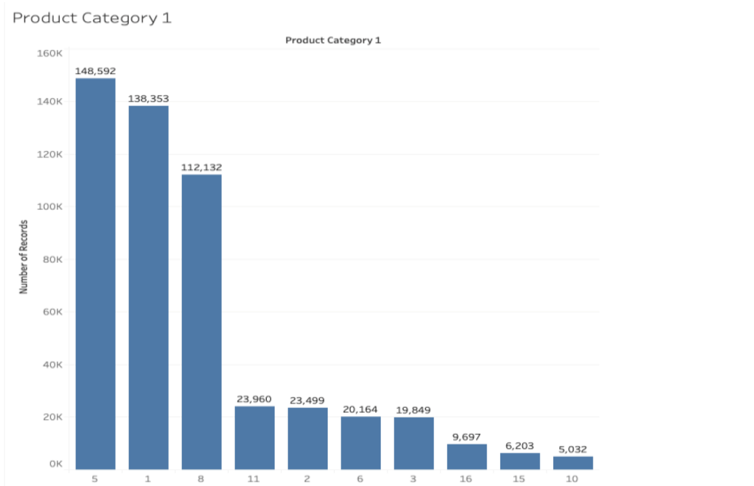
In Tableau

We can see from the above graph that, the top 10 products are sold more than 1200 in quantity. The description of the products, is however not present in the dataset. But let us see what all product category that interested the people.

<u>Maximum sold Product Category :</u>

Product_Category_1 :

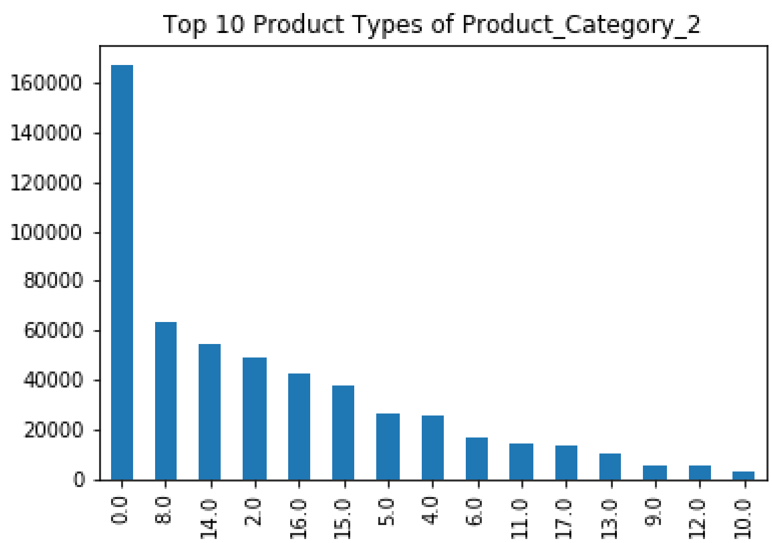


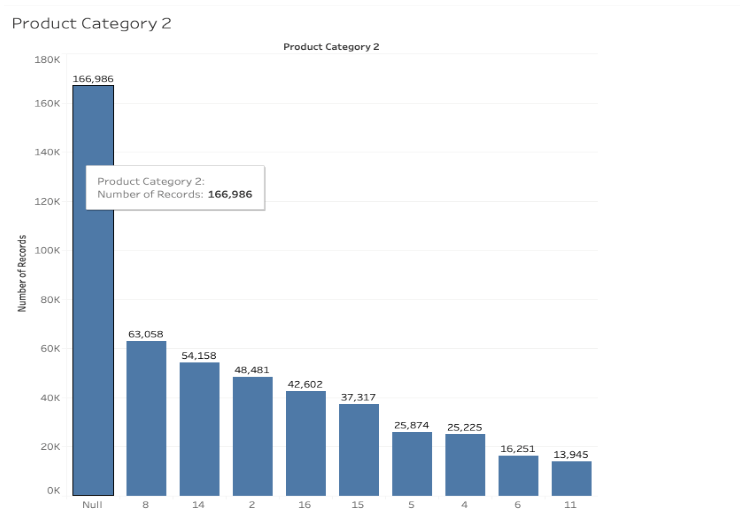


In Tableau

The highest selling product types of **Product_Category_1 are 5, 1, and 8** which are worth more than 100k.

Product_Category_2 :

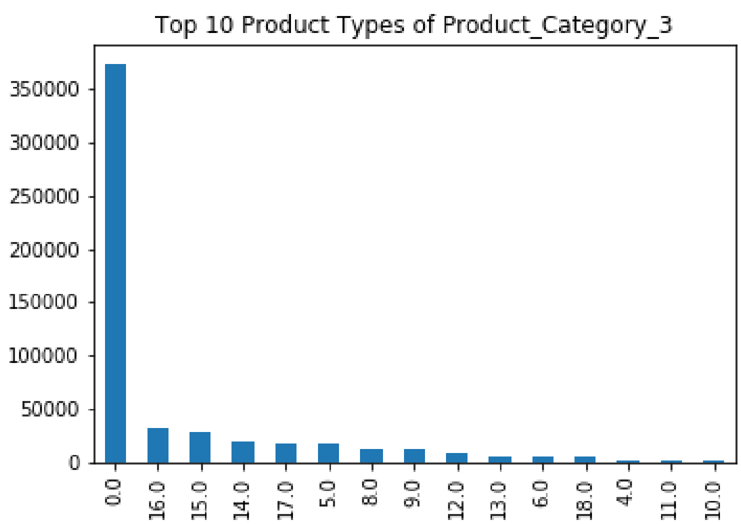


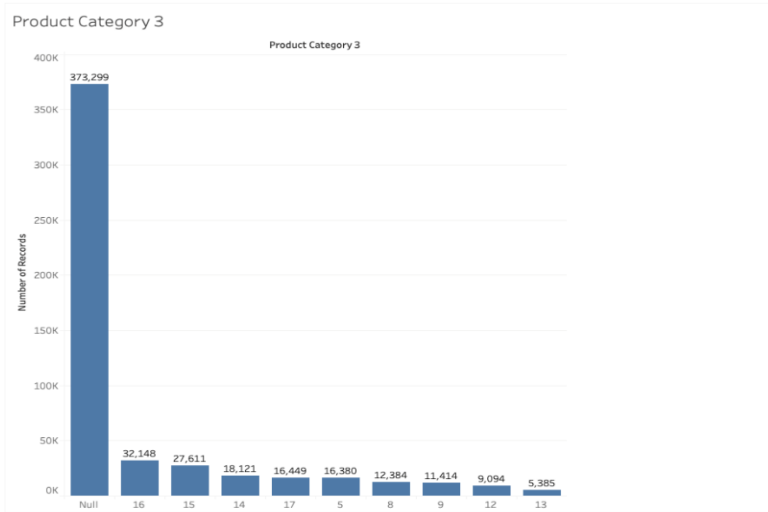


In Tableau

The highest selling product types of **Product_Category_2 is 0**, which is worth upto 160k. Product types 7, 13 and 1 have sold upto 50k.

Product_Category_3 :

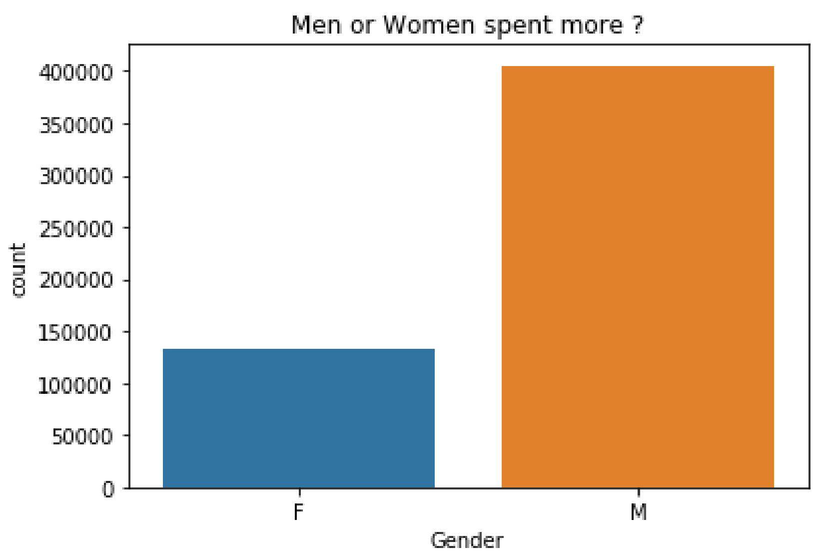




In Tableau

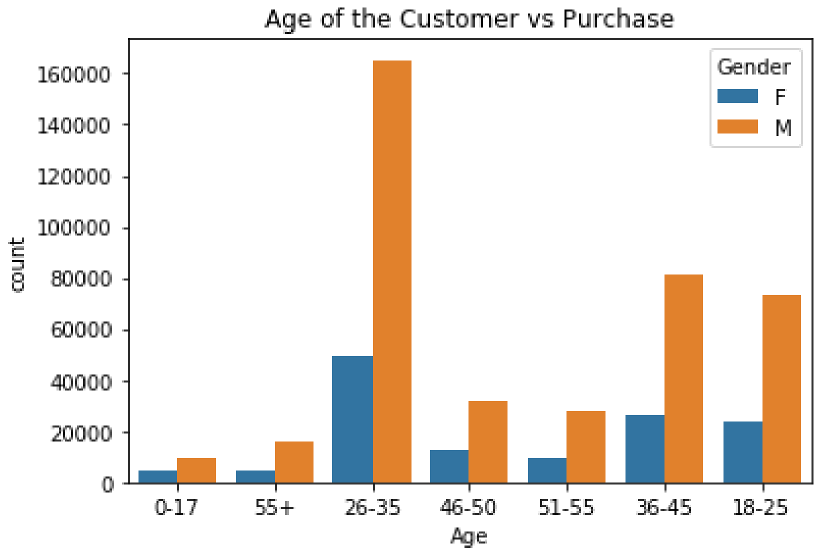
The highest selling product types of **Product_Category_3 is 0**, which is worth upto 3500k. Product types 13, 12 and 11 have sold upto 40k.

Men or Women, who are likely to spend more?



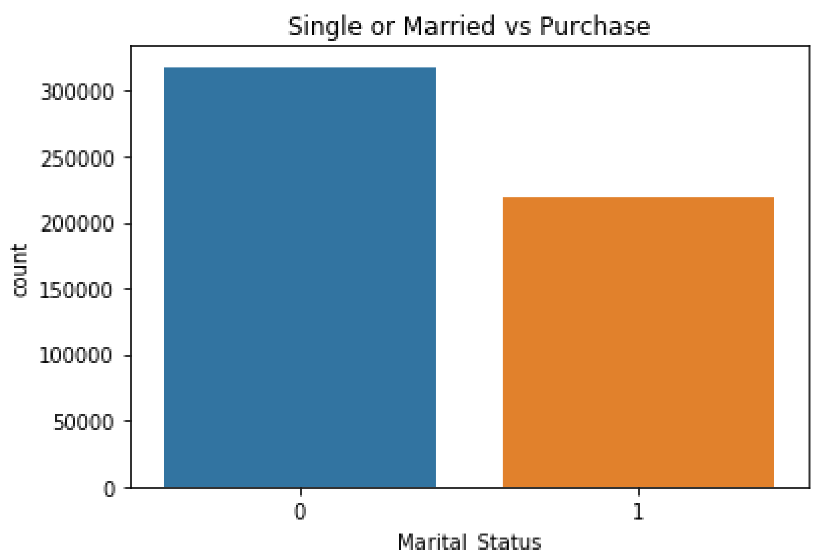
From this, we can see that the number of male customers is almost 3 times higher than the number of female customers. This could mean that, Men are most likely to buy during the Black Friday sales.

Age of the Customers :



From the graph, we see that the Majority of customers are from the age group of 26-35. We can also check the majority of a gender among the age groups by adding a hue. And as seen above, more Men spent in the sale than Women.

Married or Individuals, who spends more?



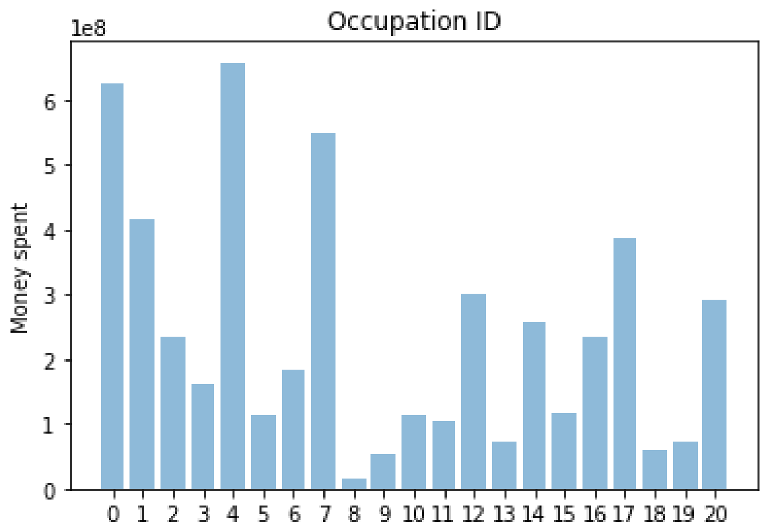
From the above graph, we can see that, Single customers purchased more than Married Customers.

Occupation of the Customers :

We can see there are 21 different occupation ID's are registered during the shopping day.

The Occupation number could represent different professions of customers: for example, number 1 could be an engineer, number 2 a doctor, number 3 an artist, etc.

It would be also interesting to see how much money each costumer group (grouped by occupation ID) spent. To do that, we can use a for loop and sum the spent money for each individual occupation ID.

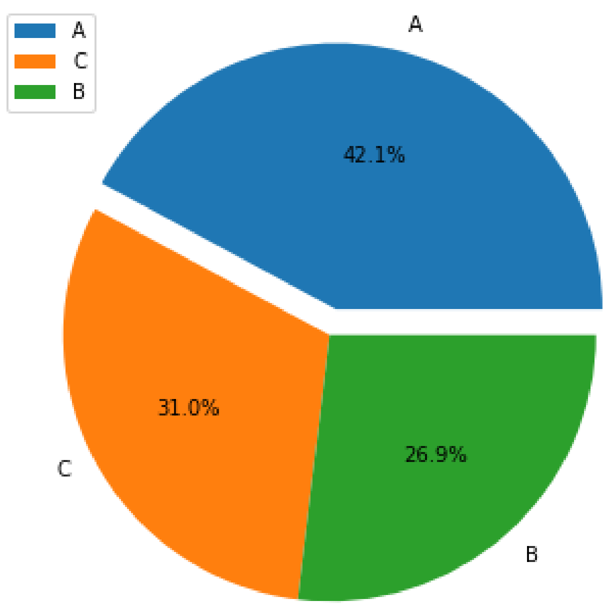


It can be easily observed that people with Occupation IDs **0 and 4 spent the most money** during Black Friday sales.

On the other hand, the people with Occupation IDs **8, 9, and 18 have spent the least** amount of money.

It can imply that these groups are the poorest ones, or contrary, the richest people who don't like to shop in that kind of retail stores. We have a deficiency with information to answer that question, and because of that, we would stop here with the analysis of the Occupation category.

__City Category :__



It is evident from the pie chart that all the three cities are almost equally represented in the retail store during Black Fridays. Maybe the store is somewhere in between these three cities, is easily accessible and has good road connections from these cities.

DATA PREPROCESSING:-

- User_ID is is the number assigned automatically to each customer, and it is not useful for prediction purposes.
- The Product_ID column contains information about the product purchased. It is not a feature of the customer. Therefore, we will remove that too.
- The data type of all the variables are different. We will convert all the variables to int to perform Linear Regression.

MULTIPLE LINEAR REGRESSION:-

Linear regression represents a very simple method for supervised learning and it is an effective tool for predicting quantitative responses.
We have performed Multiple Linear Regression using the OLS (Ordinary Least Squares) Model.

#Coefficients of the variables

```
result.params
Gender                    2918.355372
Age                       764.406935
Occupation                83.852379
City_Category             1249.388544
Stay_In_Current_City_Years    677.098363
Marital_Status            375.153063
Product_Category_1        -72.616341
Product_Category_2        101.864735
Product_Category_3        302.279087
dtype: float64
```

Summary of Multiple Linear Regression model :

OLS Regression Results

Dep. Variable:	Purchase	**R-squared (uncentered):**	0.757
Model:	OLS	**Adj. R-squared (uncentered):**	0.757
Method:	Least Squares	**F-statistic:**	1.119e+05
Date:	Tue, 19 Nov 2019	**Prob (F-statistic):**	0.00
Time:	06:51:18	**Log-Likelihood:**	-3.2183e+06
No. Observations:	322546	**AIC:**	6.437e+06
Df Residuals:	322537	**BIC:**	6.437e+06
Df Model:	9		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
Gender	2918.3554	19.339	150.909	0.000	2880.452	2956.258
Age	764.4069	6.777	112.792	0.000	751.124	777.690
Occupation	83.8524	1.395	60.123	0.000	81.119	86.586
City_Category	1249.3885	11.667	107.091	0.000	1226.522	1272.255
Stay_In_Current_City_Years	677.0984	6.647	101.859	0.000	664.070	690.127
Marital_Status	375.1531	19.596	19.145	0.000	336.746	413.560
Product_Category_1	-72.6163	2.454	-29.597	0.000	-77.425	-67.808
Product_Category_2	101.8647	1.523	66.906	0.000	98.881	104.849
Product_Category_3	302.2791	1.923	157.152	0.000	298.509	306.049

Omnibus:	8101.578	**Durbin-Watson:**	1.967
Prob(Omnibus):	0.000	**Jarque-Bera (JB):**	8723.541
Skew:	0.399	**Prob(JB):**	0.00
Kurtosis:	3.108	**Cond. No.**	28.6

<u>Performance Estimation of the model :</u>

In the end, it is always good to estimate our results by finding the mean absolute error (MAE) and mean squared error (MSE) of our predictions.
MAE: 4071.004455731231
MSE: 27179739.770752437

Inference from Multiple linear regression :

- p-value ≤ 0.05 indicates strong evidence against the null hypothesis, so you reject the null hypothesis.
- All variables except the Product_Category_1 are positively associated with Sales.
- R-squared score is **0.757**, which means that this model explains 75% of the total variance.
- Hence, the proposed model is a good model.

COLINEARITY AND REGULARIZATION:-

What if the independent variables are not independent of each other i.e. collinearity or multicollinearity is present among the predictor variables. We check the same using VIF(Variance inflation factor).

```
from statsmodels.stats.outliers_influence import variance_inflation_factor
from statsmodels.tools.tools import add_constant

pd.Series([variance_inflation_factor(X.values, i)
               for i in range(X.shape[1])],
              index=X.columns)
```

```
Gender                         3.350587
Age                            4.388604
Occupation                     2.482402
City_Category                  2.682763
Stay_In_Current_City_Years     2.684237
Marital_Status                 1.862445
Product_Category_1             2.321700
Product_Category_2             1.970802
Product_Category_3             1.498102
dtype: float64
```

- We can see from the VIF check that there is no collinearity among the dependent variables which is in a severe range except Age and Gender column, but it is not severe.
- When predictor variables are related, they fit well into a straight regression line that passes through many data points
- It is difficult to ascertain reliable estimates of each coefficients for the predictor variables which results in incorrect conclusions
- In these cases stated above Lasso, Ridge and Elastic Net Regression comes into play to combat variance problems

RIDGE REGRESSION:-

Ridge Regression helps reduce Variance by shrinking parameters and making predictions less sensitive. It can find solution with Cross validation and Ridge Regression Penalty. It is better when all variables are useful. In our dataset, we could see that all the variables are important hence we use Ridge for regularization.

```
from sklearn import linear_model, preprocessing
from sklearn.linear_model import Ridge
from sklearn.model_selection import train_test_split


a_train, a_test, b_train, b_test = train_test_split(X, Y, test_size=0.4)

rr = Ridge(alpha=10, normalize=True)# higher the alpha value, more restriction on the coefficients, with larger alpha
#the flexibility of the fit would be very strict.
result1=rr.fit(a_train, b_train)

b_pred = rr.predict(a_test)
```

```
mse = np.mean((b_pred - b_test)**2)
```

```
mse
```

```
24084281.247550942
```

```
prediction = rr.predict(a_test)
```

```
prediction
```

```
array([9222.72925782, 9102.69602   , 9792.19715187, ..., 9251.88157745,
       9739.26549654, 9797.32055013])
```

MSE value is lesser after ridge regression.

CONCLUSION

- We could answer all the questions in our problem statement like –

1. **What are maximum products sold?**

P00265242	1858
P00110742	1591
P00025442	1586
P00112142	1539
P00057642	1430
P00184942	1424
P00046742	1417
P00058042	1396
P00059442	1384
P00145042	1384

2. **Which Product category has highest sales?**

   Product Category 1 – 5, 1, 8
   Product Category 2 – 0, 8, 14
   Product Category 3 – 0, 16, 15

3. **Finding the buyer's age group.**

   26 – 35 Age group

4. **Finding the marital status of the buyers.**

   Single customers purchase more than the married customers.

5. **Analyzing the gender group, which has high interest in the sales.**

   Male customers purchase more than Female customers.

- Multiple Linear regression model gave us an R squared value of 0.757, which means the model explains 75% of the total variance.
- The VIF values for all the independent variables are less than 5, the variables, age and gender have a slightly high VIF and hence we appy Ridge regression for normalization of the same.
- After application of Ridge regression, we saw that MSE value was reduced slightly which means that it is definitely the preferred and/or desired choice as it shows that your data values are dispersed closely to its central moment (mean); which is usually great.