SI 618 Project 2 Report: Exploratory Data Analysis for the Retailer on Black Friday

---

Yonnie Chan (yonniech@umich.edu)
Data Science
November 28, 2022

1 Motivation

In the United States, Black Friday is, without a doubt, one of the largest days for shopping. Coming right after Thanksgiving, Black Friday starts the holiday shopping season by providing shopping bargains for both in-store and online customers. There was research stated that Americans spent over $717 billion on the day after Thanksgiving last year. That's over $1,000 per person and an increase of more than $35 billion from 2017[1].

For retailers, it would be prodigal and unsustainable business practices if they make a wrong promotion, including useless price cutting, and targeting the wrong customers instead of rewarding valuable customers. Therefore, conducting Black Friday sales analytics helps us to get a better understanding of the customer purchase behavior (specifically, purchase amount) against various products of different categories, especially during the season, to obtain further insights for retailers to develop better data-driven strategies for the future.

If possible, we could further build a model to predict the purchase amount of customers against various products which will help them to create personalized offers for customers against different products.

2 Datasets

For this project, I had access to the dataset available on Kaggle.com (https://www.kaggle.com/datasets/sdolezel/black-friday) from a retail company "ABC Private Limited". The dataset contains up to 550K transaction records, more than 3K unique products, and 5K+ unique customers. Information includes different aspects of customer demographics (age, gender, marital status, city category, stay in current city years), product details (productid and product category), and total purchase amount, which is given in **.csv** format.

```
<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
```

---

[1] https://executiveeducation.wharton.upenn.edu/thought-leadership/wharton-at-work/2019/12/black-friday-needs-to-go-proves-analytics/

```
 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)
```

	Column Name	unique values
0	User_ID	5891
1	Product_ID	3631
2	Gender	2
3	Age	7
4	Occupation	21
5	City_Category	3
6	Stay_In_Current_City_Years	5
7	Marital_Status	2
8	Product_Category_1	20
9	Product_Category_2	17
10	Product_Category_3	15
11	Purchase	18105

From the figure on the left we could observed:

- 5,981 unique customers.
- 3,631 unique products.

Fig. 1: Table for the count of unique values in each column

3 Method

- Handle missing data

  There are some missing values in *Product_Category_2* and *Product_Category_3*, which are supposed to be 550,068 but 376,430 and 166,821 non-null values respectively.

  Most of the cases, we will tend to delete variables if the data is missing for more than 60% because these features might be useless. On the other hand, we would try to use an imputation technique to fill in the missing values in that column.

	Total	Percent
Product_Category_3	383247	0.696727
Product_Category_2	173638	0.315666
User_ID	0	0.000000
Product_ID	0	0.000000
Gender	0	0.000000
Age	0	0.000000
Occupation	0	0.000000
City_Category	0	0.000000
Stay_In_Current_City_Years	0	0.000000
Marital_Status	0	0.000000

Fig. 2: Table for the percentage of missing values in each column

  For this case, I would drop *Product_Category_3* because missing data in this column is more than 60% of observations. Then for *Product_Category_2*, I would utilize **simpelImputer** from **Sklearn** to fill in missing data.

- Detect outliers

  I generate a line chart and box plot to detect outliers in purchases. Then I drop records that are not in [Q1 (first quartile) - 1.5 * IQR (interquartile range), Q3 (third quartile) + 1.5 * IQR (interquartile range)] (Interquartile range: Q3 - Q1).

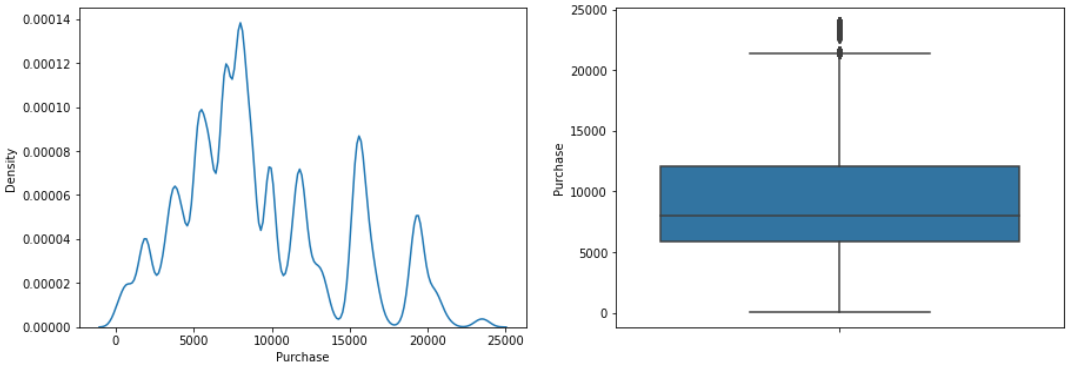

Fig. 3: Kdeplot and boxplot for distribution of purchases

4 Analyses and business suggestions

Question 1: Who are our core customers?

To find out the core customers of the retail store, I visualized their demographic and personal information from the purchasing records by generating six count plots to present the percentage of customers' information. For customers, I got six demographic information from the dataset: Gender, Age, Occupation, City, Years in the current city, and Marital status.

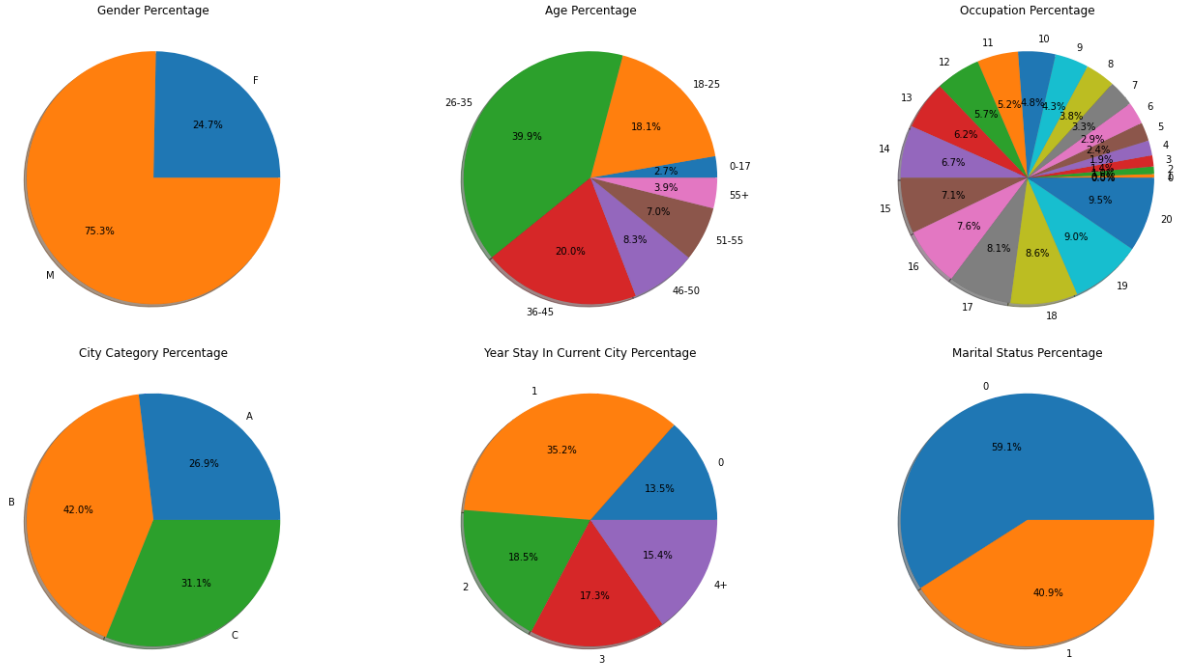

Fig. 4: Pie chart for the percentage of the number of transactions in each demographic feature

For the **number of transactions**, it is salient that the stores got more male customers (75.3% male and 24.7% Female). Up to 40% of purchase records are 26-35 age group and 20% are from the 36-45. City 'B' has a higher visiting rate. And most buyers have stayed in the current city for more than '1' year. Unmarried customers count more than married customers.

From the chart on the right, single people have purchased more than married people and in both categories' men, have purchased more than women. Especially, single men occupied a relatively large proportion.

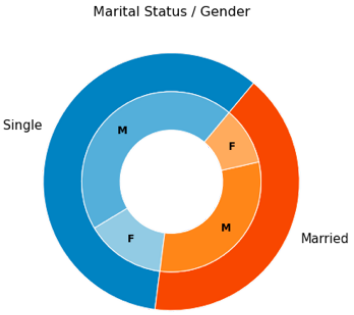

Fig. 5: Pie chart for the percentage of marital status and gender

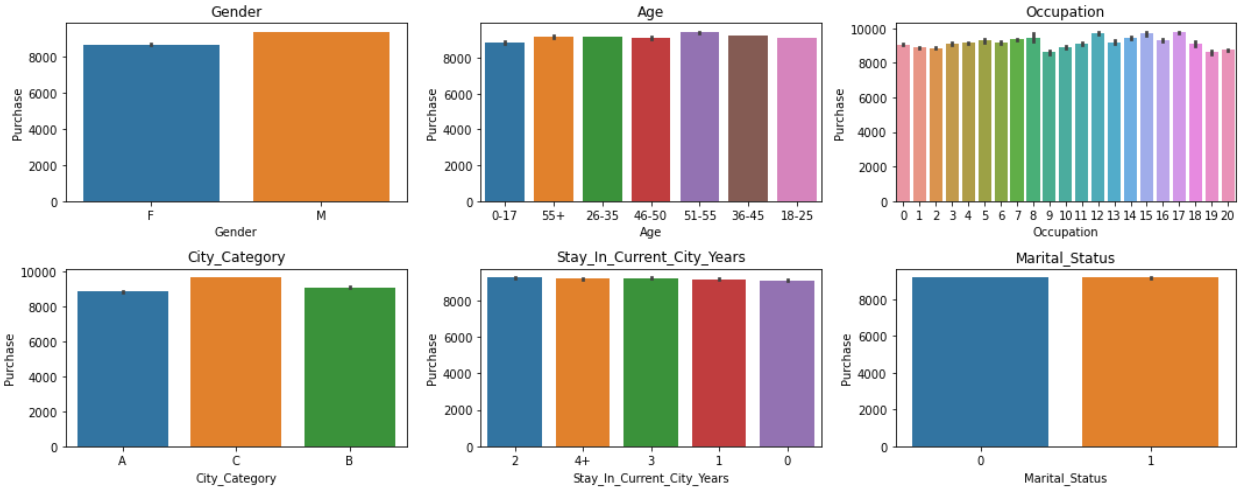

Fig. 6: Bar charts for the sum of purchases in each demographic feature

However, when it comes to the **sum of purchases**, the means of male customers is also higher than female, which is 9367.72 and 8671.05 respectively, but the difference is not large. In other demographic features, we couldn't observe a significant difference between different labels.

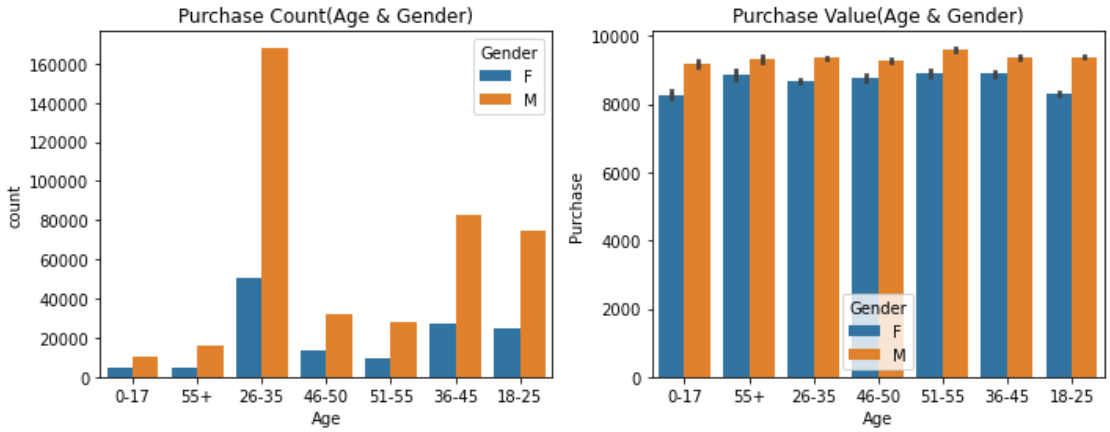

Fig. 7: Bar charts for the number of transactions and sum of purchases for age and gender

In the above plot, we could ensure the most purchases was made by 26-35 male, while those customers don't really purchase a lot or purchase high-value items within a visit.

This result represents that purchasing value could have a large improvement in males because they actually purchase in the shop much more frequently, and so does those aged between 26 and 35. From those plots, we could know more about our customers to give our directions for targeting our audience.

Business suggestion: We should put more emphasis on 26-35 males, rather than females in the next marketing campaign to improve the overall profit.

Question 2: What are our best-selling products and product categories?

	Product	Purchase
#1	P00265242	1880
#2	P00025442	1615
#3	P00110742	1612
#4	P00112142	1562
#5	P00057642	1470
#6	P00184942	1440
#7	P00046742	1438
#8	P00058042	1422
#9	P00059442	1406
#10	P00145042	1406

Fig. 8: Table for the sum of purchase for top 10 products

I calculated profit from each category and found that profit from the product of category 2 is two times higher than from category 1 (5,404,346 for category 2 and 2,944,386 for category 1). By generating two bar charts based on their purchasing value for two product categories, we could find out the best-selling product in each category.

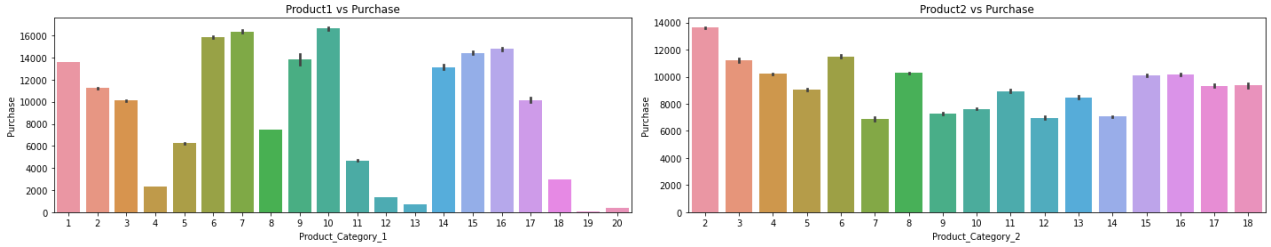

Fig. 9: Bar charts for the sum of purchase in two product categories

In the first category, there are several top sellers which have considerable sums of purchase, which are 10, 7, 6, 16, 15, 1, 9, 14 and they all have over 12,000 sums of purchase. We could also point out that some categories (19, 20, 13, 12, 4) have poor sales performance.

In the second category, there isn't a significant difference between different categories. Category 2 is the highest among all categories.

	0	1	2	3	4	5	6	7	8	9	10
#1	P00265242	P00265242	P00265242	P00265242	P00265242	P00265242	P00265242	P00265242	P00112142	P00034742	P00145042
#2	P00110742	P00220442	P00025442	P00117942	P00110742	P00114942	P00110742	P00110742	P00242742	P00265242	P00242742
#3	P00025442	P00110742	P00058042	P00025442	P00112142	P00251242	P00058042	P00025442	P00127842	P00117442	P00112142
#4	P00057642	P00046742	P00110842	P00110842	P00025442	P00110742	P00031042	P00112142	P00117942	P00000142	P00025442
#5	P00112142	P00025442	P00059442	P00110742	P00237542	P00112542	P00255842	P00184942	P00114942	P00102642	P00255842

11	12	13	14	15	16	17	18	19	20
P00265242	P00057642	P00265242	P00265242	P00025442	P00265242	P00057642	P00265242	P00265242	P00265242
P00025442	P00112142	P00010742	P00184942	P00110742	P00046742	P00025442	P00010742	P00237542	P00059442
P00059442	P00265242	P00317842	P00025442	P00265242	P00255842	P00112142	P00080342	P00058042	P00220442
P00117942	P00025442	P00080342	P00237542	P00059442	P00025442	P00110742	P00058042	P00059442	P00110742
P00148642	P00242742	P00085242	P00110742	P00112142	P00034742	P00237542	P00057642	P00112142	P00025442

Fig. 10: Table for top 5 sellers in each occupation

Table above represents the top 5 seller products based on the buyers' occupation (the same products id would have the same background color).

It is noticed 'P00265242' is the most-purchased product for 15 out of 21 occupations and an interesting fact is that this product is not presented in the top-5 products of occupations 8, 10, and 17.

Moreover, from the top 5 products of occupation 9, one of them is 'P00265242' and present in most of the other top 5s, one of them is only present in occupation 16's list and the rest are not repeated in any other lists. Combined with the insights from the previous chart, this was the only occupation with more women than men (even though the total number of men in the dataset was higher), which makes occupation 9 a unique occupation on the list.

Business suggestion: We could focus on highlighting our best-selling products in our marketing campaign, and product design as well. Advertising for the right target audience based on their occupation and shopping behaviors.

Question 3: Create clusters to show customer segments.

The main method for clustering in this project is K-means. First, I sampled the data records for quicker computation, used **LabelEncoder** to transform categorical features to numerical form, then utilized **MinMaxScaler** to normalize the data for K-means. After that, I applied **PCA** to reduce the dimensions of the features by fitting the data with only two dimensions. Then I created a range of clusters, predicted the cluster for each data point, and calculated the mean **silhouette coefficient** for the number of clusters chosen to find the best score, which is **4 clusters**. From the following scatter plot, we could observe four clusters have separate centroids.

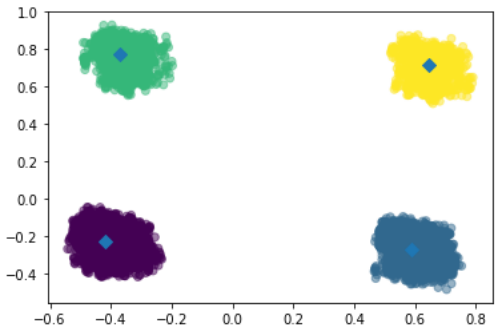

Fig. 11: Scatter plot for centroids

In the end, I generated six bar charts for each demographic characteristic for four clusters. Then we could get a preliminary segment profile for each cluster.

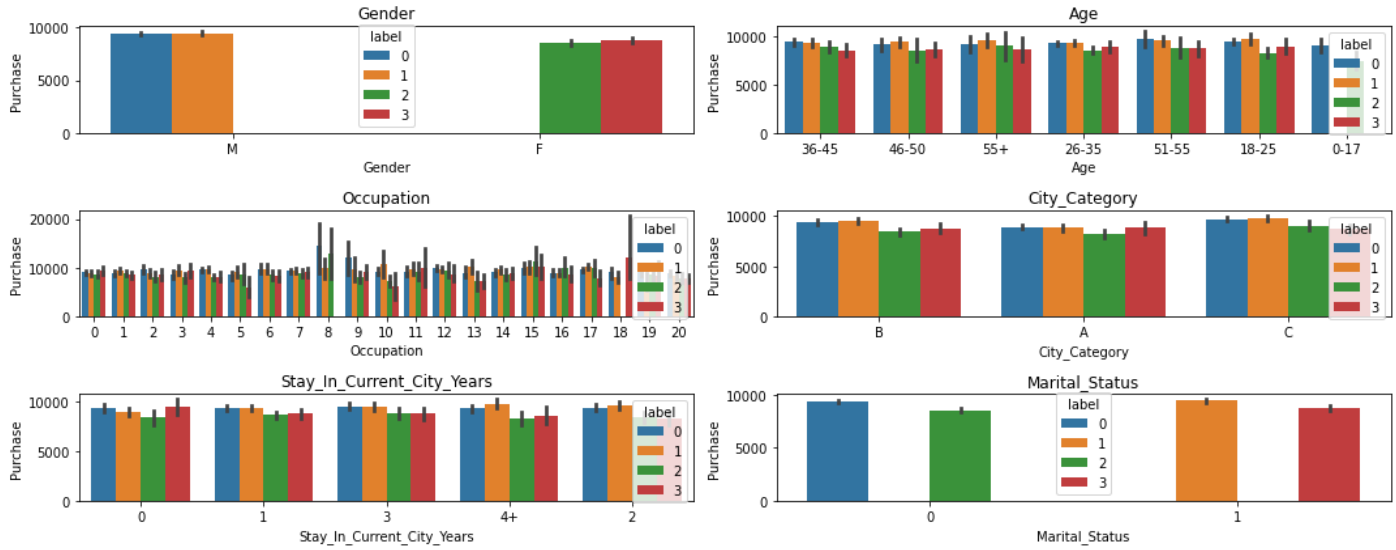

Fig. 12: Bar charts for demographic features by each cluster

From the bar charts above, we could generate the following table to illustrate our customer segments:

Cluster	**Gender**	**Age**	**Occupation**	**City category**	**Years in current city**	**Martial status**
1	Male	All	All, 8 especially	All	All	Single
2	Male	All, except for 0-17	All	All	All	Married
3	Female	All	All	All	All	Single
4	Female	All, except for 0-17	Except for 8, 18 especially	All	All	Married

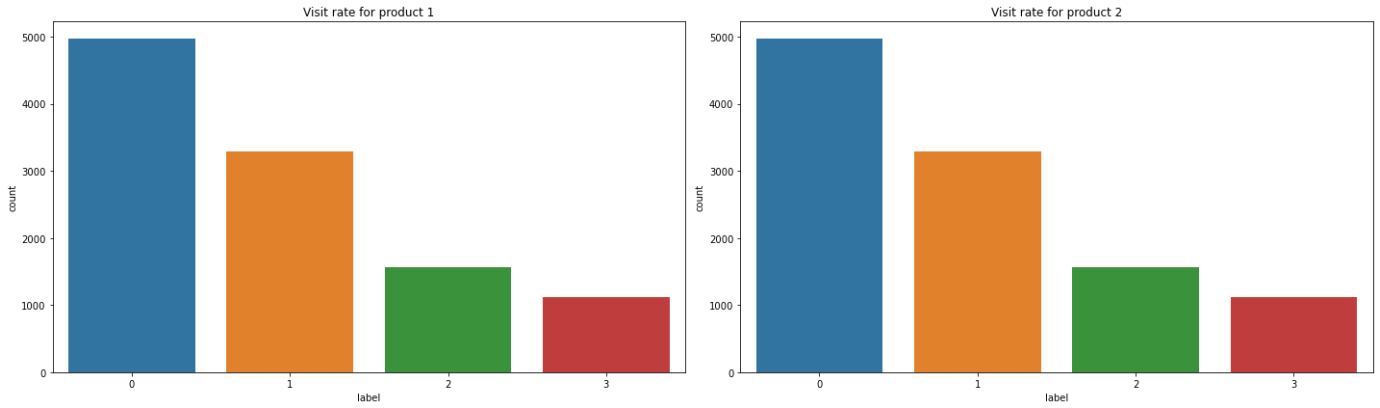

Fig. 13: Count charts for each product category by each cluster

Business suggestion: After clustering, it is evident that we should advertise and give promotion more for the cluster 0 on Black Friday, who are our main target. We've got the identical result as the previous analysis but got more information about other customer segments.

Additional Question: What variables affect the purchase (our target variable) more?

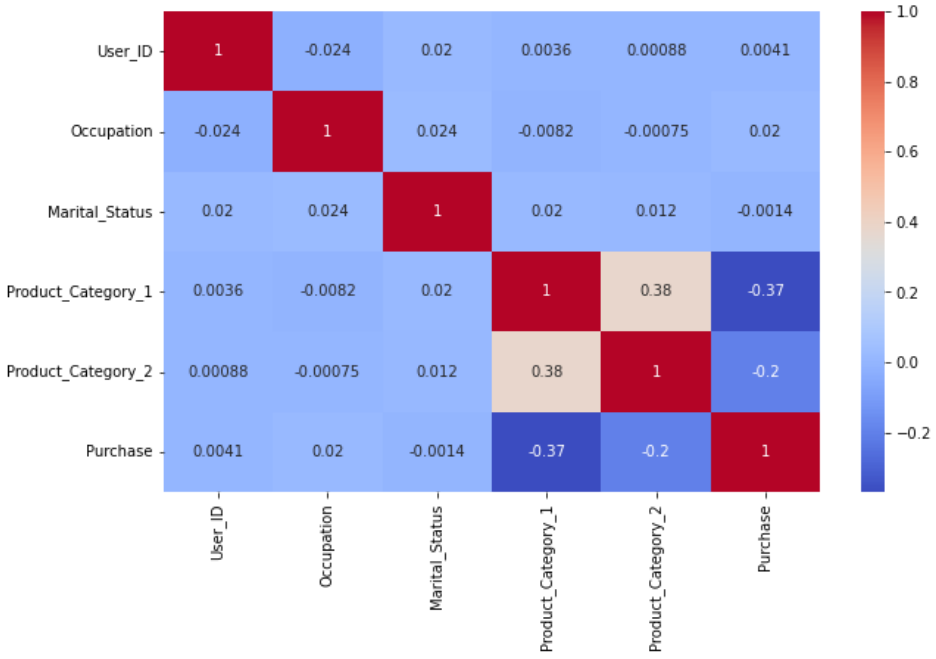

Fig. 14: Heatmap for the covariance of various features

I generated a heatmap to visualize the statistical correlation between the different variables. The purchase has a relatively higher correlation with *product_category_1* and *product_category_2* compared to other features.

Business suggestion: Different product categories should affect purchase value more instead of customers' personal information. The product designing team should come up with different ideas to attract customers. Also, for further machine learning processes to predict purchases in future data, these two variables should affect the model more.