Milestone Report

1. Problem statement

Based on sales observations in a retail store during Black Friday, the objective is to make decisions on strategy of targeted advertising. For a specific group of people, predict the categories of products they are most likely to buy and include the categories in the advertising email or other advertising method.

The client of this analysis would be the owner of the retail store. With the results, the client would be able to determine their advertising strategy to promote sales. They would be able to decide target customers for different categories of products.

2. Description of the dataset

The raw data files in this project are acquired from Kaggle at https://www.kaggle.com/mehdidag/black-friday.

The raw csv file is BlackFriday.csv. The file includes dataset of 537,577 observations about the balck Friday in a retail store. It contains different kinds of variables either numerical or categorical including custormer information and purchase information. The variables include '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'.

3. Data wrangling

3.1 Understanding the dataset

This is a dataset with 537,577 rows and 12 columns. Characteristics of the dataset include:

1) Numerical columns:

'Stay_In_Current_City_Years', 'Purchase'

2) Categorical columns:

'User_ID', 'Product_ID', 'Gender', 'Age', 'Occupation', 'City_Category', 'Marital_Status', 'Product_Category_1', 'Product_Category_2', 'Product_Category_3'

3) Missing values:

Missing values occur in columns of 'Product_Category_2' and 'Product_Category_3', because some products belong to one or two categories, instead of three.

3.2 Data wrangling strategies

1) Set uniform format for column names:

Change all column names as lowercase names to make sure uniform format.

2) Add a new column ‘product_category_num’ to df:

A new Series 'product_category_num' containing the number of product categories is created and appended to the main set df.

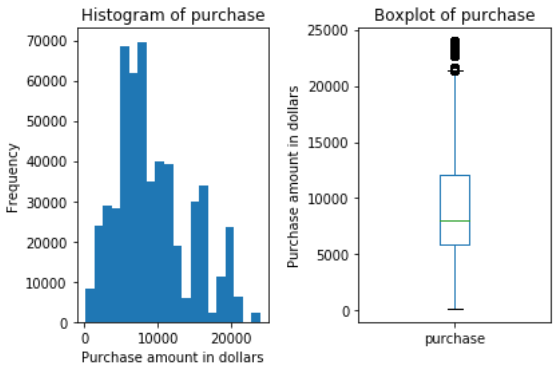
3) Deal with missing data in df:

Missing data occur in columns of 'product_category_2' and 'product_category_3'. All values in 'product_category_1', 'product_category_2', 'product_category_3' range from 1 to 18 except for missing value. Missing value indicates some products only belong to one or two categories instead of three. So it's better to replace all NaN values into 0 to make data clean.

4) Convert data types in columns of 'product_category_2' and 'product_category_3':

The data type of column 'product_category_1' is int64. Data types of columns 'product_category_2' and 'product_category_3' are float64. But columns of 'product_category_2' and 'product_category_3' only contain integer values from 1 to 18, similar as in column 'product_category_1'. Converting data type in 'product_category_2' and 'product_category_3' from float to int will help save up memory.

5) Exam column 'purchase' for outliers:



6) Save cleaned dataset to CSV file:

The cleaned dataset is saved to ‘blackfriday_clean.csv’.

4. Initial findings from exploratory analysis

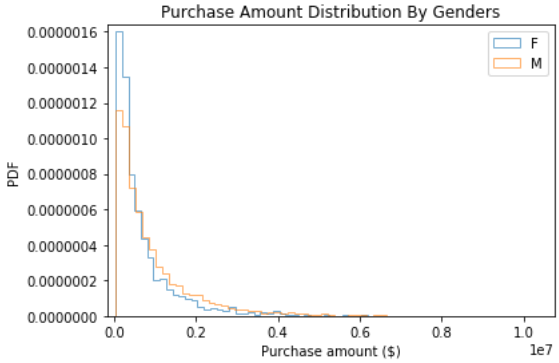
4.1 Purchase amounts variation among different parameters

To understand how purchase amounts vary with different parameters, purchase amounts are investigated from two perspectives:

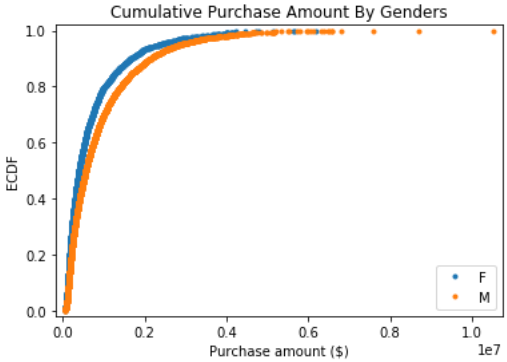
Histograms are plotted to show purchase amount distributions by parameters;

Cumulative density function and boxplot are also plotted to provide quantitative comparisons among parameters.

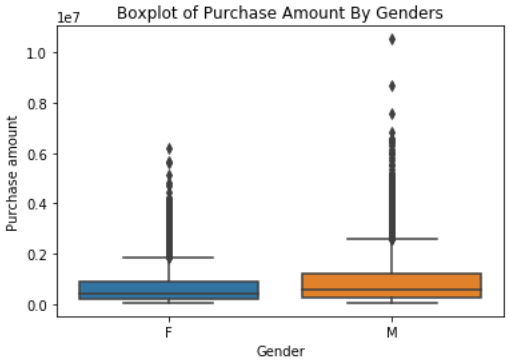
4.1.1 Purchase amounts variation among genders



This plot shows purchase amount histograms by female customers and male customers. For smaller purchase amounts, there are more female customers than male customers. For larger purchase amounts, there are more male customers than female customers.

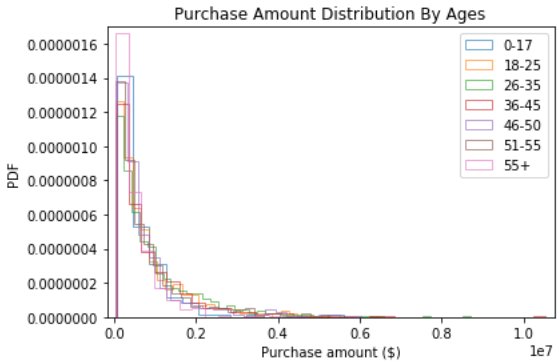


This plot shows cumulative distribution function of purchase amount by female customers and male customers. For smaller purchase amounts, there are more female customers than male customers. For larger purchase amount, there are more male customers than female customers. As purchase amount reaches 4,000,000, ECDF becomes flat. Small portion of customers spend more than 4,000,000. Above 6,000,000, there are several male customers but almost no female customers.

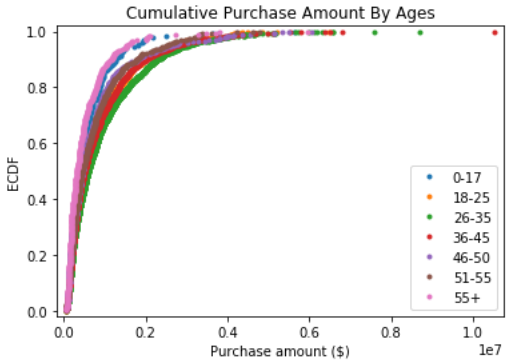


This boxplot shows comparison of purchase amount distribution between female customers and male customers. Totally, there are 5891 users, among them 1666 are females, 4225 are males. The number of male users is about 2.5 times the number of female users. The mean of purchase amount by male is about 1.3 times that by female. The median of purchase amount by male is about 1.4 time that by female.

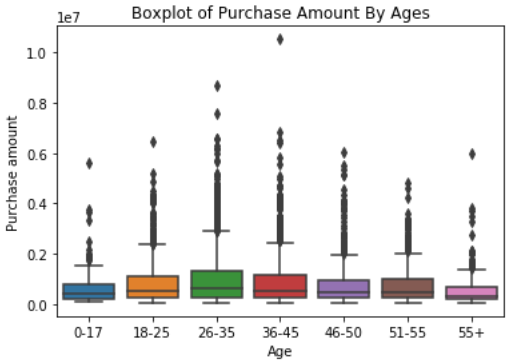
4.1.2 Purchase amounts variation among ages



This plot shows purchase amount histograms by customers of different ages. For smaller purchase amounts, there are most 55+ customers.

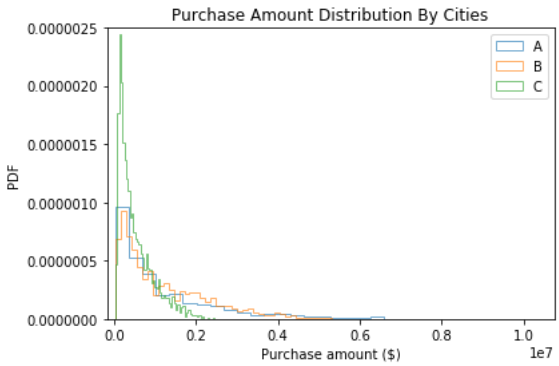


This plot shows cumulative distribution function of purchase amount by customers of different ages. For smaller purchase amounts, there are most 55+ customers, second most are 0-17 customers. As purchase amount reaches 4,000,000, ECDF becomes flat. Small portion of customers spend more than 4,000,000. Above 6,000,000, there are most 26-35 customers and 36-45 customers.

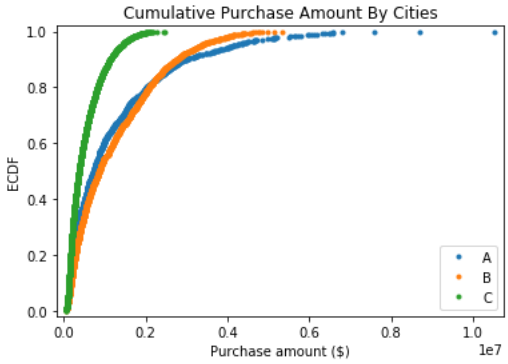


This boxplot shows comparison of purchase amount distribution among customers of different ages. Age 26-35 has the largest number of customers. Age 0-17 has the smallest number of customers. The second least customers are in the age range of 55+. The number of 26-35 customers is 9.4 times that of 0-17 customers, 5.5 times that of 55+ customers. The mean of purchase amounts by 26-35 customers is the largest, while the mean of purchase amounts by 55+ customers is the smallest. The median of purchase amounts by 26-35 customers is the largest, while the median of purchase amounts by 55+ customers is the smallest.

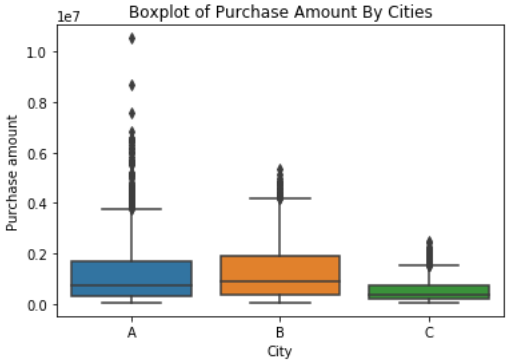
4.1.3 Purchase amounts variation among cities



This plot shows purchase amount histograms by customers from city A, B, and C. For smaller purchase amounts, there are more customers from city C than customers from city A or city B. For larger purchase amounts, there are more customers from city A or city B than customers from city C.

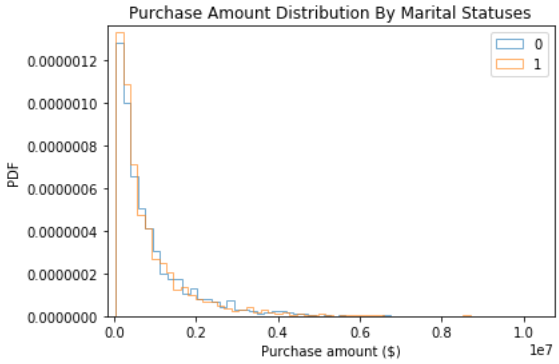


This plot shows cumulative distribution function of purchase amount by customers from different cities. For smaller purchase amounts, there are most customers from city C. For smaller purchase amounts, there are more customers from city A than city B. However, between 3,000,000 and 6,000,000 there are more customers from city B than city A. Above 6,000,000, there are only customers from city A. City A has the largest range of purchase amounts.

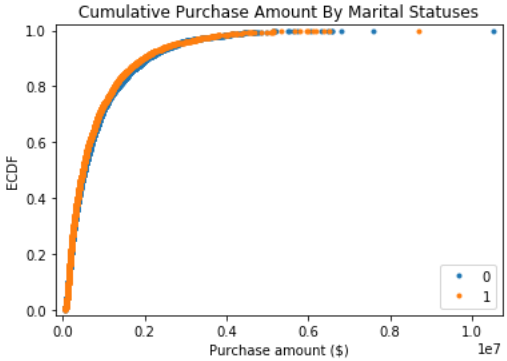


This boxplot shows comparison of purchase amount distribution among customers of different cities. City C has the largest number of customers but the mean or median of the purchase amount is the least. The number of customers of city C is about 1.8 times that of city B, 3 times that of city A. The mean purchase amount of city C is about 0.4 times that of city B or city A. The median purchase amount of city C is 0.4 times that of city B, 0.5 times that of city A. City A has the smallest number of customers, but the largest purchase amounts occur in city A. City A and city B has similar mean value and similar median value.

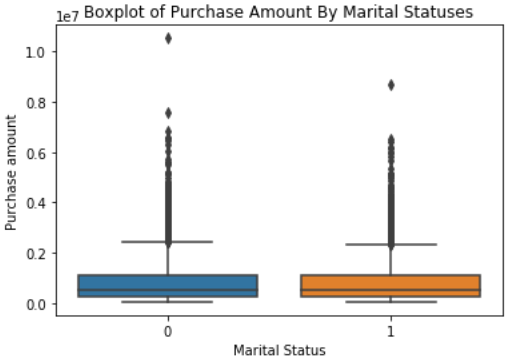
4.1.4 Purchase amounts variation among marital statuses



This plot shows purchase amount histograms by married customers and single customers. They have very similar distributions.

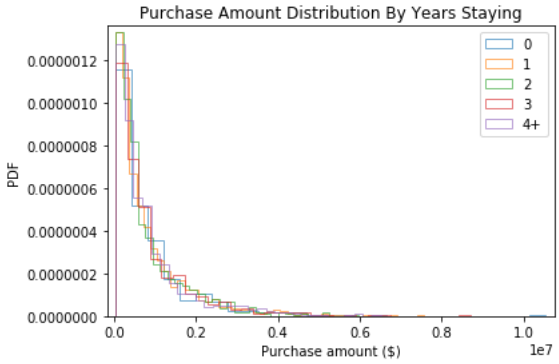


This plot shows cumulative distribution function of purchase amount by married customers and single customers. They have very similar cumulative distributions.

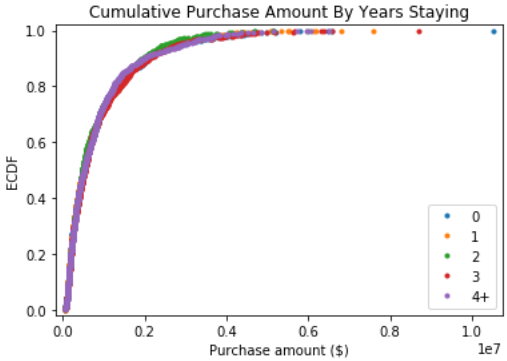


This boxplot shows comparison of purchase amount distribution among customers of different marital statuses. Marital status 0 and 1 have very similar distributions and similar statistics.

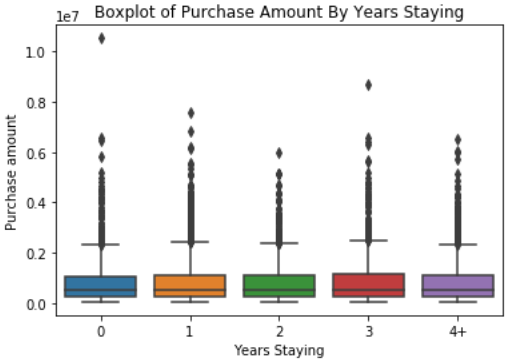
4.1.5 Purchase amounts variation among years staying in current city



This plot shows purchase amount histograms by customers who stay in current city for different years. They have very similar distributions.



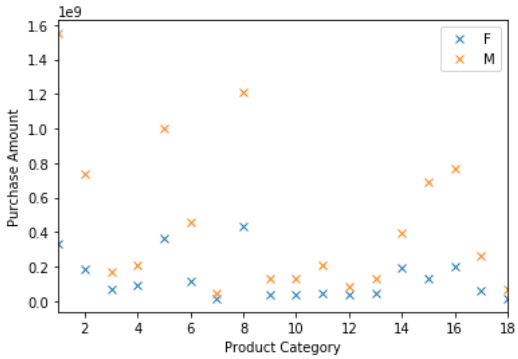
This plot shows cumulative distribution function of purchase amount by customers who stay in current city for different years. They have very similar cumulative distributions.



This boxplot shows comparison of purchase amount distribution among customers who stay in current city for different years. For different years staying in current city, purchase amounts have very similar distributions and similar statistics. Possible reasons are the retail store is a chain store. It's easy for people living in different cities to get access.

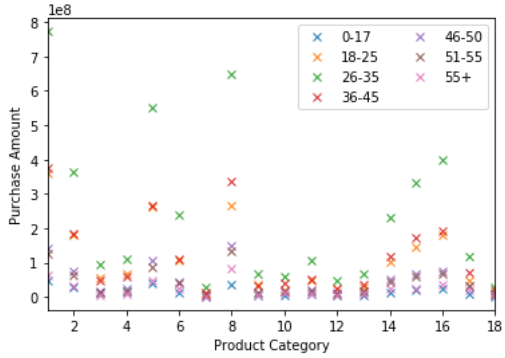
4.2 Rank purchase amounts of different product categories by genders

4.2.1 Purchase amount comparison among different product categories by genders



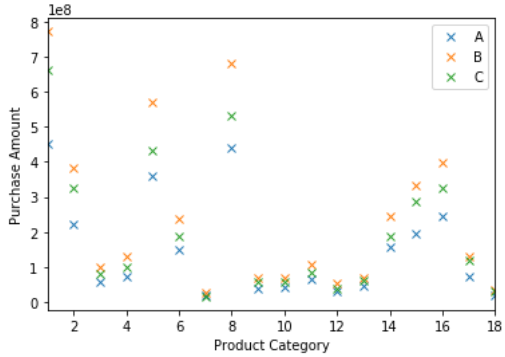
This plot shows purchase amount comparison among different product categories by genders. Overall, rankings for different product categories are similar between females and males. For most product categories, purchase amount by males is slight larger than that by females. For several product categories, purchase amount by males is much larger than that by females.

4.2.2 Purchase amount comparison among different product categories by ages



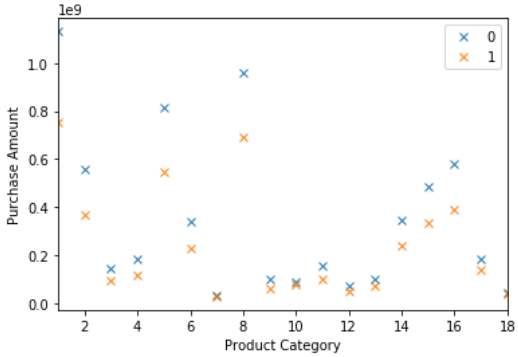
This plot shows purchase amount comparison among different product categories by ages. Overall, rankings for different product categories are similar among different ages. For several product categories, purchase amount by 26-35 customers is much larger than that by other ages. This is consistent with the total purchase amount comparison.

4.2.3 Purchase amount comparison among different product categories by cities



This plot shows purchase amount comparison among different product categories by cities. Rankings of purchase amounts for different product categories are very similar for different cites. For most product categories, difference of purchase amounts among cities is small. For several product categories, difference of purchase amounts among cities is large.

4.2.4 Purchase amount comparison among different product categories by marital statuses



This plot shows purchase amount comparison among different product categories by marital statuses. Rankings of purchase amounts for different product categories are very similar for different marital statuses. Overall, purchase amount of single customers for different product categories is larger than that of married customers. For most product categories, purchase amounts are similar between married customers and single customers. For several product categories, purchase amount of single customers is obviously larger than that of married customers.

4.3 Test of independence between product category and different parameters

Based on the Chi-squared test, product category is dependent on different parameters, including gender, age, occupation, city, years of staying in current city, and marital status.

4.4 Test of independence between purchase level and different parameters

In order to conduct the Chi-square test of independence between purchase levels and different parameters, total purchase amounts are categorized into different categories. In this case, categories are chosen based on every 10000 interval of purchase amounts, 0, 1, 2, 3, 4 correspond to purchase amounts of 0-1,000,000, 1,000,000-2,000,000, 2,000,000-3,000,000, 3,000,000-4,000,000, above 4,000,000, repectively.

Based on the Chi-squared test, purchase level is dependent on different parameters, including gender, age, occupation, city, years of staying in current city, and marital status.