# Black Friday Sales data analysis
## SI 618: Data Manipulation & Analysis Final Project

**Yung-Chun Lee**
yungclee@umich.edu

## Abstract

In this report, we explored the sales data set by implementing series of statistical methods and visualizations to present our final result which can be an insight for some or other retailers in interested to increase or build novel strategy for this once every year big sale event.

## 1 Motivation

Black Friday is the Friday following by the Thanksgiving, celebrated each year in the third Thursday in November, in the United State. Once every year, retailers and store owners offer great promotions for almost every items to boost the sale amount and take this chance to clear inventory. For buyers, this is the deal of the year that will save you a lot of money to purchase new furniture, electronics, or even Christmas gift! Literally anything you can think of. From a seller's perspective, the sales record can be renewed if they are able to come up with a better propaganda or a tailor-made selling strategy for each customer by utilizing the sales record from the past. For consumers, will more likely to know what are the hot products in this season and get an idea of what to get before the deals are gone. In either way, both parties will get benefit from a proper analysis.
We constructed 4 distinct question in this project:

- **Question 1:** SALES
  What is the most profitable product / category among all the data?
- **Question 2:** CUSTOMERS
  What is the targeted customer (profile: Gender/Age/Occupation/...)if the retailer were to make more profit (sell more products to the largest share of customers / smallest share of customer)?
- **Question 3:** SOURCES
  Is there any evidence to suggest that there exist geographical differences for total sales (purchase)? If so, rank the cities by their sales and comment on whether you should stock more goods or not.
- **Question 4:** PREDICTIONS
  What is the expected amount spent from 20 years old, married women who have lived in the city C for more than 10 consecutive years?

## 2 Data Source

The data set is a public comma separated value file (.csv) available on Kaggle,
`https://www.kaggle.com/mehdidag/black-friday`.
This data set consists of 537,577 observations of the transaction records that took place in that certain day from a certain retailer. Variables include user ID, to identify if certain purchase were made by same customer, Product ID, Gender, Age, Occupation, city category and years the customer stay in the city. All the categories and customer information has been masked and pre-processed to be untraceable and insensitive for public studying purposes.

| Variable | Cont. | Bin. | Cat. | Ord. | NAs | UNIQ |
|---:|---|---|---|---|---:|---:|
| User ID | | | - | | | 5,891 |
| Product ID | | | - | | | 3,623 |
| Gender | | v | | | | 2 |
| Age | | | | v | | 7 |
| Occupation | | | v | | | 21 |
| City Category | | | v | | | 3 |
| Stay In Current City Years | | | | v | | 5 |
| Marital Status | | v | | | | 2 |
| Product Category$_1$ | | | v | | | 18 |
| Product Category$_2$ | | | v | | -166,986 | 18 |
| Product Category$_3$ | | | v | | -373,299 | 16 |
| Purchase | v | | | | | 17,959 |

Table 1: Data Set Variables

Some entries contains missing values, mainly because some products do not have secondary or third product categories. There are no missing values in other variables. As for the unique values, from Table 1, the column UNIQ represnets the counts of unique value presented in the data set, which of all 537,577 transactions there were only 5,891 distinct customers.

## 3 Methods

In this section, we briefly described the statistical approaches, methods or manipulation we implemented on the data set for each of the questions in interest.

(a) How did you manipulate the data to prepare it for analysis?
(b) How did you handle missing, incomplete, or noisy data?
(c) How did you perform data analysis in code, i.e. Briefly describe the workflow of your source code
(d) What challenges did you encounter and how did you solve them?

### 3.1 Question 1: SALES

*What is the most profitable product category among all the data?*

- Aggregation was performed on the data set with respect to the `Product ID` and `Product Category 1`(since not all products have secondary or more than that labels) and sorted on a descending order by the total amount of `Purchase`, after the process, we should be able to see the ranking of the sales amount by product ID and Product category respectively.
- Variables we used for this specific question does not have missing values.
- We utilized the methods from `pandas` module. `groupby()` and `sort_values()` were chained for a *pythonic* style coding.

### 3.2 Question 2: CUSTOMERS

*What is the targeted customer (profile: Gender/Age/Occupation/...)if the retailer were to make more profit?*

- For this question, we constructed a summary table that is aggregated on the customer ID. Since we want to understand the background profile of the customer source. Some aggregation and reduction was necessary for a cleaner layout. For example; Sum, mean and count were reduced from the full transaction history a customer made, and the most frequent product category was created as a profile as the customer's preference.
- Variables we used for this specific question does not have missing values.
- We utilized the methods from `pandas` module. `groupby()` and `agg()` are used to create the main data frame of the reduced data. `.apply()` and `lambda` is used to gain the customer preference profile.

### 3.3 Question 3: SOURCES

*Is there any evidence to suggest that there exist geographical differences for total sales purchase?*

- We used the count that the each customer made within Black Friday for this question. We 'binned' the count data into 20 categories and performed chi-square test on the data to exam if the sales count is independent from the city the customer located in.
- Variables we used for this specific question does not have missing values.
- The `scipy.stats` module provides a variety of useful tools for statistical modeling and testing, `chi2_contingency` is really useful for the independence testing.
- First we exam the distribution of for the sales amount but found out that it is not normally distributed and most statistical comparison such as ANOVA is based on the normality assumption. We used the count data for the independence test and Yete's continuity correlation to account for the fact that some of the cells contain '0'.

### 3.4 Question 4: PREDICTIONS

*What is the expected amount spent from 20 years old, married women who have lived in the city C for more than 10 consecutive years?*

- We constructed a regression model for this question. With certain description of a customer we should be able to predict the amount of a customer will more likely to spend on Black Friday.
- Variables we used for this specific question does not have missing values.
- We utilized the methods from `statsmodel` module. we constructed a naive model with non-transformed data and then do residual

analysis if the residuals satisfy the normality assumption, and use **log-**transformation on the response to get our final model.

# 4 Analysis & Result

In this section, we performed statistical analysis, visualizations and modeling on the data set to exam each of the aspects we posted in the questions.

## 4.1 SALES

For a retailer to maximize its profit from sales, they need to know where the main source of the sales, which ever the product or the category will be helpful for the in-store display or more targeting promotion. We started off with some exploratory visualization to check the how the sales are distributed among different aspects.

### 4.1.1 Category

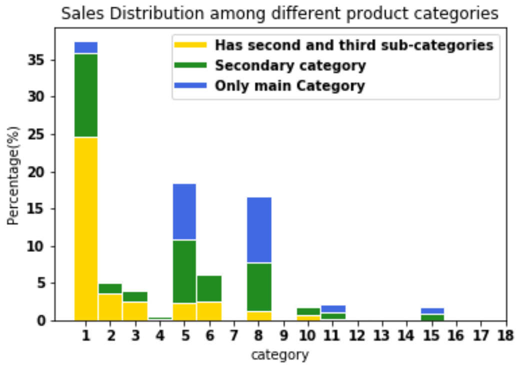

Figure 1: Sales Distribution among Product category

From Figure 1, each stack bar represent the percentage that the product category contributed. The color yellow indicates that the product has both second and third product categories other than the main one, and green represents if the product has a secondary category index and blue for no further category than the first one.

### 4.1.2 Product

In Figure 2, we displayed the top 25 products with the largest total purchase total, which we can also observe that most of them are also from 'Product Category' **group 1**. The product with the most sale has the total sale of 27,532,426, which is about 0.548% of the total sale from 3,623 product options.

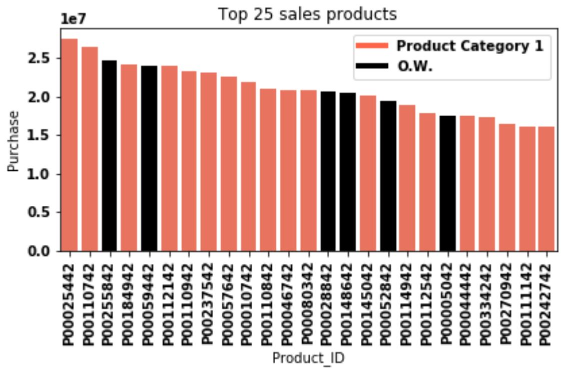

Figure 2: Top 25 Products

From either category or product perspective, we have observed that the sales amount in product category 1 is the most popular and also contributed 37.52% of the total sales.

## 4.2 CUSTOMERS

Knowing customer who is willing to spend money is a good way to increase the profit a company can make so does knowing where the potential market is, in either way, it is key to know what are the features the customer have. We took a look into customers available information from the data set, which will help us have a better understanding of the customers and their profiles or construct preference based on the purchase history.

### 4.2.1 Gender & Age

Women and shopping, is that just an stereotype or just marketing skills?
There is no doubt that this kind of stereotype has exist for long but is that even true? We used the data and tried to discover if there is certain correlation or differences between women and men. We further created a cross table between gender and age group to see if there exist certain pattern among customer.

First we plotted the bar-plot for the age and gender group, notice that we used the total *amount* of purchase in Figure 3. The result is quite surprising, not only men contributed more sales in total, it also uniformly surpass the percentage in every age group.

Since it might be affected by the difference of measurement standards, since we use the total price instead of the purchase *count*, we further exam if using different standard will affect the result shown in Figure 3.

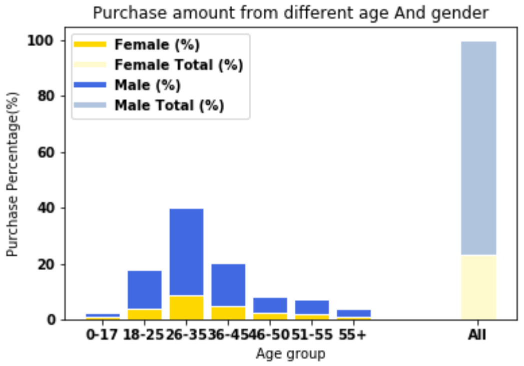

Figure 3: Age and Gender distribution for total amount of purchases

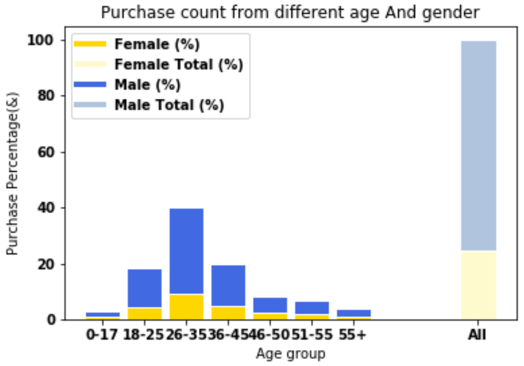

Figure 4: Age and Gender distribution for total counts of purchases

There is merely no difference between using amount(fig 3) or count(fig 4) for the total purchase both measuring standards suggested that men actually purchased and contributed more sales on Black Friday given our data set.

### 4.2.2 Occupation & Gender

We plotted another plot (Figure 5) with two variables: occupation code and gender.

We can see the distribution for occupation and gender in figure 5, unfortunately, the data set does not have decoding for the job category we cannot make inference from the occupation code.

### 4.2.3 Result

One thing worth noticing that the conclusion we have drawn above is based on this data set, which the ratio for male and female customers are imbalanced (4225:1666) which can also be explained as why we observed the dominating trend in both amount and count in figure 3 and 4, since the gender ratio in the original data is about 71:29 which is fairly close to to the overall percentage we got from both.

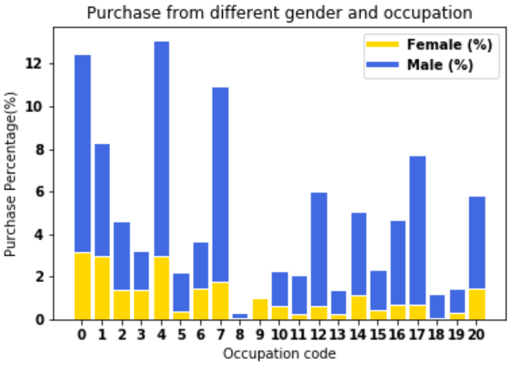

Figure 5: Occupation and Gender distribution for total amount of purchases

## 4.3 SOURCES

We are interested in knowing whether there exist sales difference among different locations. We attempted to use ANOVA but was not valid approach dued to the violation and ended up using chi-square independence test to test if there exist statistical significant correlation between the city the customer belong and the purchase total.

### 4.3.1 Continuity

We first used violin plot (Figure 6) and faceted the data with different city, it can be observed that the tail behaviors among the cities are similar so did the gender-wise comparison within each city. However, the biggest concern we had is that the purchase amount is not normally distributed so that most statistical comparison for mean such as ANOVA or t-test will not work from the violation. We will use the purchase count per customer as our alternative.

### 4.3.2 Binning

For simplicity, we choose BINS=20 (bin range 50) to be our binning range (Table 2), there's certain impact on the binning choice we made with respect to the independence test, we will not discuss the binning procedure and the consequences in this report. By setting the bins to 20 we got a decent representation of the purchase amount while still not having many zeros in all of the cells.

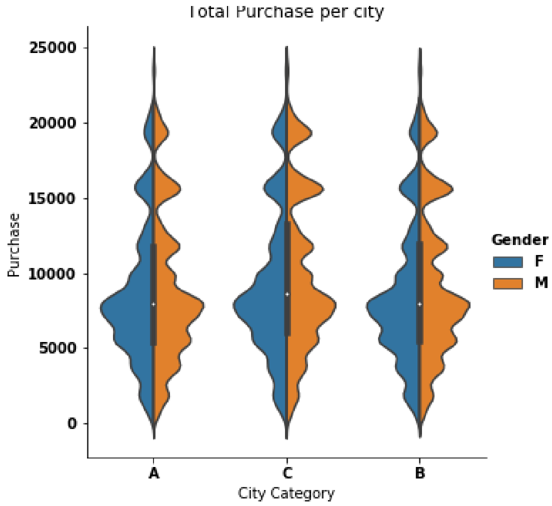

Figure 6: Total Purchase per City

| City | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 | 11 | 12 | 13 | 14 | 15 | 16 | 17 | 18 | 19 | 20 |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| A | 392 | 226 | 98 | 82 | 60 | 57 | 21 | 24 | 18 | 19 | 10 | 11 | 4 | 11 | 5 | 2 | 1 | 2 | 1 | 1 |
| B | 568 | 329 | 204 | 153 | 169 | 102 | 70 | 42 | 36 | 25 | 7 | 2 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| C | 1875 | 806 | 316 | 136 | 6 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |

Table 2: Data binning

### 4.3.3 Yete's correction

We used Yete's continuity correction for the chi-square test since some of the cells are zeros. We finally perform a chi-square test to test whether the city(geographical variable) is independent with the purchase.

### 4.3.4 Result

under the hypothesis
$H_0$: City and purchase amount are independent.
$H_A$: City and purchase amount are not independent.
The test statistic from the chi-square independence test(Table 3) suggests that city and purchase amount are not independent.

| Test Statistic | DOF | P-value |
|---|---|---|
| 1215.554 | 38 | 2.28e-230 |

Table 3: Chi-square independence test

## 4.4 PREDICTIONS

The last question is useful for predicting the customer's expected spend or can be useful to exam the importance among the variables.

### 4.4.1 Training and test

Instead of duplicating the full data we only created two lists, one for training index and test index, this approach will save more memory space. We used 80/20 for training/test respectively.

### 4.4.2 Original Data

We first fit a simple linear regression model with variables we have `Age,Gender Stay,Marital_Status,Most_Freq` as explanatory variables and `Purchase` as response variable and did the residual analysis afterward.

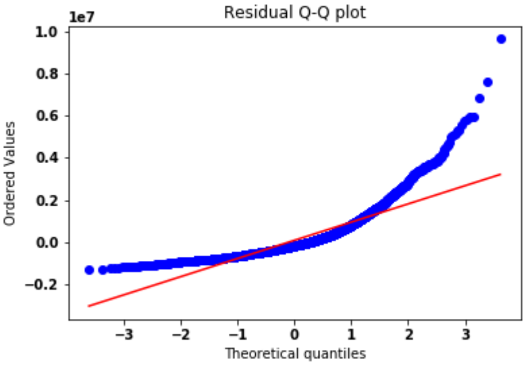

Figure 7: First model residual Q-Q plot

The residual in our first model is not normally distributed(Figure 7) which violated the residual normality assumption hence we performed a log-transformation on the response and construct a second model.

### 4.4.3 Log-Transformed Data

The second model seemed to perform really well and the residual diagnostic (Figure 8, 9) showed no abnormal or sign of violation of normality assumption.

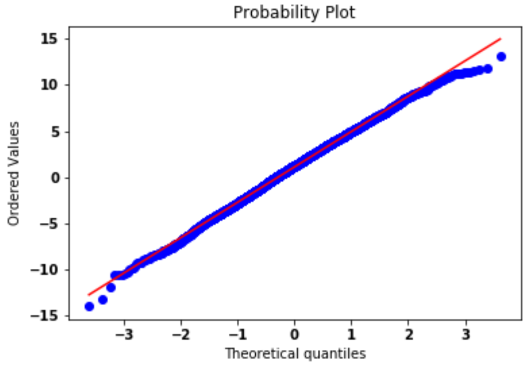

Figure 8: Second(log) model residual Q-Q plot

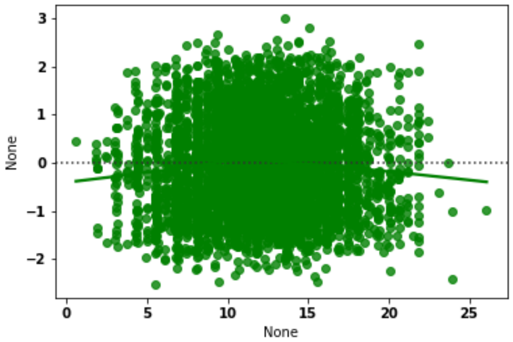

Figure 9: Second(log) model residual scatter plot

### 4.4.4 Interpretation

| Variable | coef | std err | t | P-value |
|---|---|---|---|---|
| Age | 1.2413 | 0.041 | 30.215 | 0.000 |
| Gender | 5.0020 | 0.110 | 45.298 | 0.000 |
| Stay | 1.3056 | 0.041 | 31.983 | 0.000 |
| Marital Status | 0.6618 | 0.125 | 5.295 | 0.000 |
| Most Freq | 0.5995 | 0.018 | 32.813 | 0.000 |
| $R^2 = 0.909$ | | | | |

Table 4: Coefficient for log-model

The $R^2$ is quite high for the model fitting, all explanatory variables are statistically significant. We will try to use test data and evaluate model's performance.

### 4.4.5 Prediction

We used the test data as our new data for prediction, the result is not as satisfactory as we expected. We predicted the value using the equation:

$$purchase = e^{Age+Gender+Stay+Marital+Most\_freq}$$

with each multiply by its corresponding value. We display the first few values in Table 5 the value is not quite close to the purchase as expected.

| User_ID | predict_log | predict | Original purchase | Original_log |
|---|---|---|---|---|
| 1002107 | 16.858932 | 20976882 | 152258 | 11.933332 |
| 1002109 | 13.093110 | 485585 | 3069936 | 14.937167 |
| 1002110 | 18.122581 | 74222711 | 1127264 | 13.935304 |
| 1004210 | 8.084039 | 3242 | 766243 | 13.549255 |
| 1000008 | 18.164505 | 77400625 | 796545 | 13.588039 |
| 1004213 | 14.547634 | 2079491 | 2898136 | 14.879578 |
| ... | ... | ... | ... | ... |

Table 5: Prediction for test data

## 5 Discussion

We briefly explored the data set with the four questions above. There are some concerns and some following research can be done for a more complete model. First, the gender/occupation/product category ratio is heavily imbalanced which might have been the main reason why the regression model is biased(over-fitted). Secondly, the regression might have a better estimate if we use a stepwise/ hierarchy model structure, we first predict the amount the customer might get and after we use the customer's preference to predict the total amount instead of predicting the amount directly using the one model as we suggested in this project.