Consumer Spending Habits: Maximizing Profits on Black Friday

Black Friday - Maddie Adelman, Danling Ma & Xiaohui Liao

12/5/2018

The URL for our Google drive document is {https://docs.google.com/document/d/1T3_oHoBMlnA8Nxgh2FqpNqtH2oy3_LRf452Zlt3FbtU/edit (https://docs.google.com/document/d/1T3_oHoBMlnA8Nxgh2FqpNqtH2oy3_LRf452Zlt3FbtU/edit)} and our Team GitHub repository is {https://github.com/bladechildrenlxh/-blackfridayproject (https://github.com/bladechildrenlxh/-blackfridayproject)}.

Two kernels we used: https://www.kaggle.com/gloriousc/black-friday-analysis/ (https://www.kaggle.com/gloriousc/black-friday-analysis/) https://www.kaggle.com/monethong/who-bought-what (https://www.kaggle.com/monethong/who-bought-what)

The data set we used: https://www.kaggle.com/mehdidag/black-friday/downloads/BlackFriday.csv/1 (https://www.kaggle.com/mehdidag/black-friday/downloads/BlackFriday.csv/1)

Introduction

Every year the Friday after Thanksgiving is a day of shopping. Known as Black Friday, the day is characterized by high demand for retail goods. This paper aims to understand the average shopper and the customer behavior on Black Friday for one retail store. There are two hypotheses this paper will study:

Hypothesis 1: Average shopper purchase of people living in City A is more than those living in cities B and C.

Hypothesis 2: The most profitable product categories are categories 1, 5 and 8.

Through the analysis of the above hypotheses, this paper will provide suggestions to the retail store on which segment, demographic and city category the store can focus its marketing efforts on. Lastly, using predictive analytics this report suggests the customers the retail firm can target to maximize Black Friday profits.

Summary Statistics

The analysis of the Black Friday data set used in this paper includes 12 variables: `Product_ID` , `Gender` , `Age` , `City_Category` , `Stay_In_Current_City_Years` , `Marital_Status` , `Product_Category_1` and `Purchase` . This data set is cleaned from the original data set which contained 3 additional variables - `Product_Category_2` and `Product_Category_3` . These three variables are excluded from this analysis because `Product_Category_2` and `Product_Category_3` contain missing observations. Excluding `Product_Category_2` and `Product_Category_3` will not affect the results we have drawn. Although logically a retail product may or may not belong to multiple product categories, for ease these two variables are not analyzed in this paper. For easy interpretation, Product_Category_1 was subsequently renamed Product_Category. The clean Black Friday data set yields 537,577 entities observed for each of the 12 variables.

The Black Friday data set is further cleaned by stating the format of each variable. This included changing `Product_ID` , `Gender` , `Age` , `City_Category` , Marital_Status and Product_Category from character variables to factors. These variables are nominal since they do not have order or a distance metric. 'Age' is ordinal because the age groups are increasing with actual age, but are not equally spaced.

For the purposes of this report, `City_Category` is assumed the three levels `A` , `B` , and `C` can refer to either type of city (urban/city versus suburban versus rural) or a specific store's location. However, it is important to note that the category that refers to type of city is unknown and is masked from the Kaggle where the authors downloaded the data set. Thus, the specific levels of this factor variable cannot be determined nor can the authors confirm which specific city category details has a relationship with the remaining variables in the Black Friday data set.

In order to estimate causal relationships, two subsets of the original Black Friday data are created to understand how distinct shoppers and product categories can be segmented. Doing so provides insight to the managers of the retail store about where the shoppers to target are, which demographics these shopper's have as well as the products and preferences of different segments. Using these new data sets, revenues by product category and distinct user are calculated. The variables quantity of items bought is also calculated for each shopper and product category. Doing so allows for deeper analysis for the retail to use in its targeted marketing efforts.

Table 1: Gender

Gender	Number of Distinct shoppers
F	1666
M	4225

Given the cleaned Black Friday data set, the retail store has 5,891 shoppers making purchases, 3,623 products sold in store and total revenue of $5,017,668,378. Of the total shoppers at all the stores on Black Friday, about 72% of the shoppers are men and the remaining 28% of shoppers are women. The skewness of the gender of shoppers might lead to a larger total revenue for male than that of female. Therefore, we adopted average Shopper Purchase to reflect the purchase power of each individual shopper. These statistics are further analyzed below.

Table 2: Summary by Each City

City_Category	Product Revenue	Number of Distinct shoppers	Avg. Purchase per Shopper	Number of Product	Avg. Unit Price per Product	Product per Shopper
A	1295668797	1045	1239874.4	144638	8958.011	138

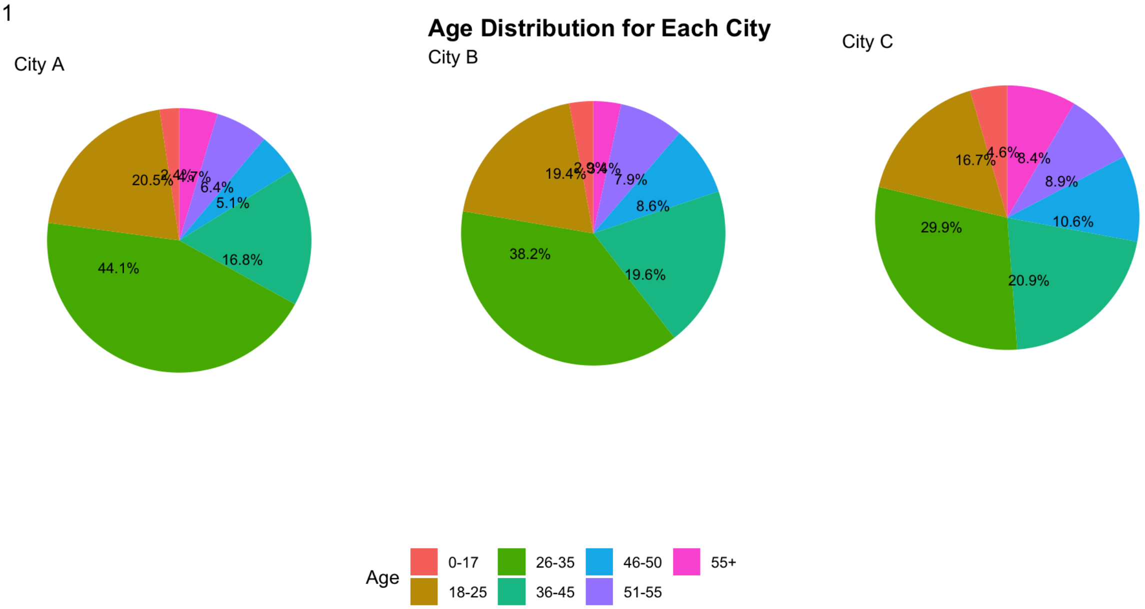
City_Category	Product Revenue	Number of Distinct shoppers	Avg. Purchase per Shopper	Number of Product	Avg. Unit Price per Product	Product per Shopper
B	2083431612	1707	1220522.3	226493	9198.658	133
C	1638567969	3139	522003.2	166446	9844.442	53

The above frequency table demonstrates how sales change across different city categories. **Table 2** shows city C attracts the most unique shoppers, 3,139; which is larger than the number of shoppers in cities A and B combined. The total revenue brought by city C residents, approximately 1.64 billion dollars, falls in between cities A ($1.3 billion) and B ($2.1 billion). However, the average purchase per unique shopper in city C is $522,003.20, which is much smaller than cities A and B.

Nonetheless, the average purchase per unique shopper in city B is $1,220,522.30. In total, all of the 1,707 unique shoppers from city B spent over 2 billion dollars on the holiday. City A has the fewest distinct shoppers (1,045) and lowest product revenue, yet the higest average purchase per shopper at $1,239,874.40. The difference in revenue between A and B may result from the difference in shoppers populations, assuming there is no significant difference in their purchase behaviors.

With respect to the quantity of items purchased in each city, city B shoppers buy the most items. The average item price for a product in city B is $9,198.66, less than city C but more than city A. City C, with the most shoppers, has the highest average item price. **Table 2** demonstrates to a manager the need to examine City A. How does an additional item purchased increase revenue per shopper? What demographics describe the shoppers in each city? Will demographics or City impact total expected profit? If so, which maximize expected profits? The answers to these questions will provide the retail store with the best strategy to form its marketing campaigns.

Figure 1



There are more shoppers who are 26-35 years old in each city than any other age group. The second largest age segment are shoppers who are 36-45 years. Together, these shoppers who are 26-45 years old comprise 50% of the total customers at the store.

Is it best for the retail store to market shoppers with the highest average purchase or the store with the most unique shopper traffic? Should the marketing strategy to be increase average purchase amount in stores with the most shoppers or to increase the number of shoppers in the store with highest grossing average purchase? To answer these questions, we created **Hypothesis 1**: Average shopper purchase of people living in City A is more than those living in cities B and C.

Figure 2

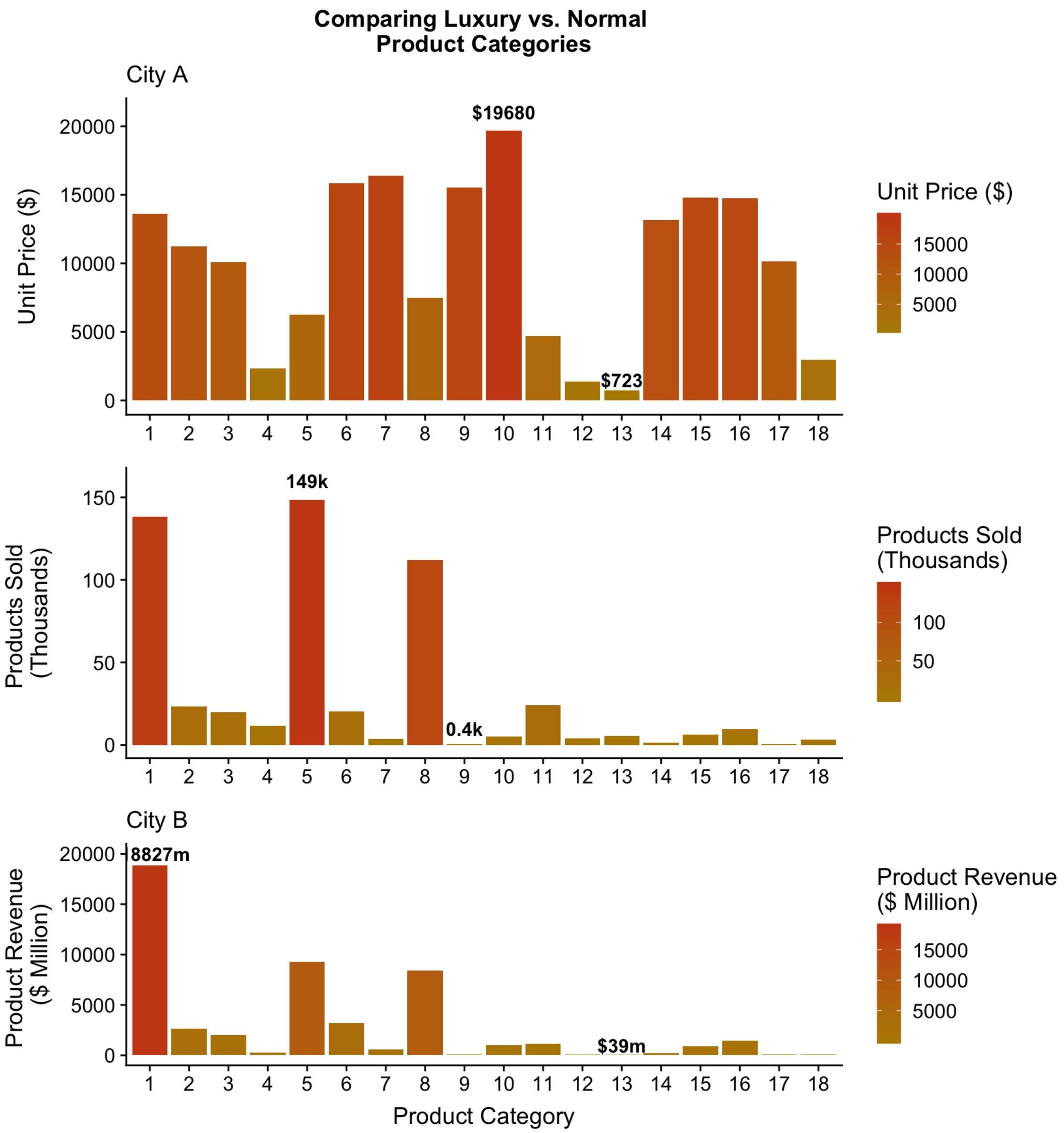


Table 3: Product Price

Product_Category	Product Revenue	Products Sold	Unit Price ($)
1	1882666325	138353	13607.70
2	264497242	23499	11255.68
3	200412211	19849	10096.84
4	26937957	11567	2328.86
5	926917497	148592	6238.00
6	319355286	20164	15837.89
7	60059209	3668	16373.83
8	840693394	112132	7497.35
9	6277472	404	15538.30
10	99029631	5032	19679.97
11	112203088	23960	4682.93
12	5235883	3875	1351.20
13	3931050	5440	722.62
14	19718178	1500	13145.45

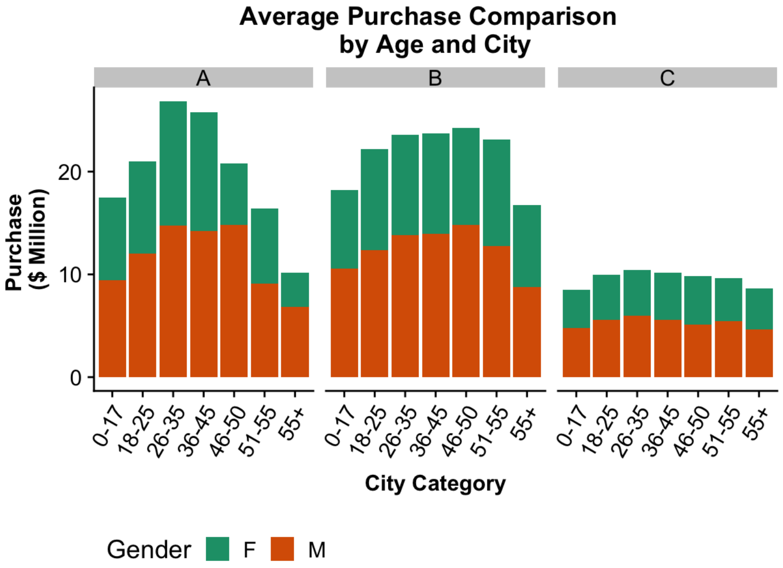
Product_Category	Product Revenue	Products Sold	Unit Price ($)
15	91658147	6203	14776.42
16	143168035	9697	14764.16
17	5758702	567	10156.44
18	9149071	3075	2975.31

Analyzing the average unit price of each product category in **Table 3**, the maximum average price is $19,679.97 for products in Category 10 while the minimum price is 722.62 dollars for category 13. Category 5 sells the most products, 148,592 items, while category 9 sells the least amount of products, 404 items. **Figure 2** above demonstrates that normal products are those in categories with low unit prices and sell a high quantity of items whereas product categories with higher average unit price sell smaller quantities of products refer to luxury products. This initial analysis indicates managers should focus their attention on the product categories that sell normal goods - categories 5 and 8.

Below, **Figure 2** analyzes the different product categories at the retail store and the revenue per shopper in each product category. This allows us to understand the average shopper's purchasing habits by product category. The bar graph demonstrates the revenue of product categories by shopper in a clear manner because product category is a categorical variable that was converted into a factor. Initially **Figure 2** was simply a bar graph; to enhance the graph, it now contains a dodged fill aesthetic by gender, bolded titles and axes titles.

On average, product categories 1, 5 and 8 are the most popular product categories amongst shoppers. Category 12 and 13 yield the least revenue per shopper. Using **Figure 2** and **Table 3**, the authors derived **Hypothesis 2**: The most profitable product categories are categories 5 and 8.

Figure 3



It is important to note that **Figure 3** is derived from the kernel (https://www.kaggle.com/monethong/who-bought-what (https://www.kaggle.com/monethong/who-bought-what)). We adjust the kernel's graph from box plot to bar graph, grouping by city, gender and age; then adjusted the variables on each axis. These changes allow for analysis specifically on city category, age and gender's effects on purchasing.

If we analyze **Figure 3** purchasing habits by age and gender, male shoppers buy more than their female counterparts in all age groups and all cities. The data portrays a trend of purchasing habits - men purchase significantly more than women in age groups 18-25, 26-35 and 36-45. Shoppers in City B purchase more than shoppers in cities A and C in all age groups except 26-35 and 36-45 years old. Across age groups, shoppers in City C purchase the least regardless of age.

Figure 3 depicts purchasing revenue by gender and city category. This graph helps us to understand the product purchasing habits of men and women shoppers in different cities. The above bar graph demonstrates the revenue by gender and city in a concise way. Both city category and gender are categorical variables that were previously converted into factors. Initially, **Figure 3** was simple without the fill aesthetic for gender. The current **Figure 3** scales the y-axes of purchases to millions of dollars in order to avoid the axis' intervals being denoted by scientific notation. Additionally, there are bolded titles and axes titles.

Figure 4

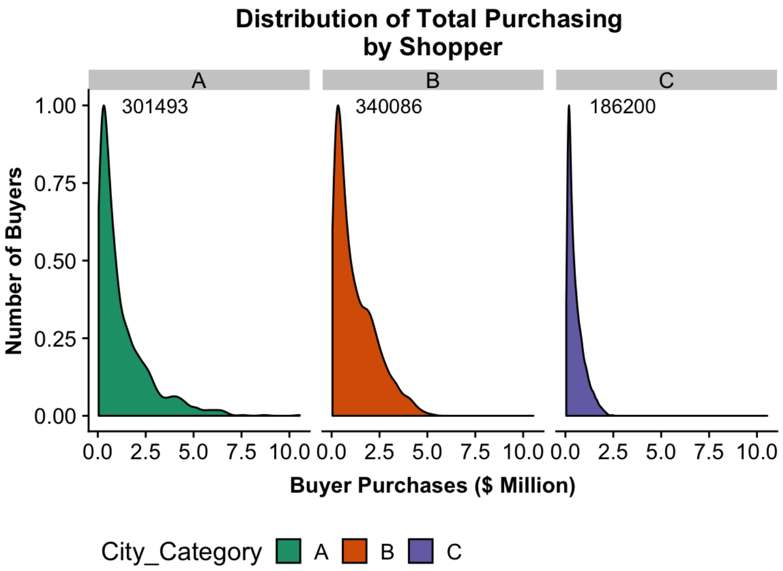


Figure 4 was derived from Kernel “https://www.kaggle.com/gloriousc/black-friday-analysis (https://www.kaggle.com/gloriousc/black-friday-analysis).” This paper expands the distribution analysis by segmenting buyers and purchases by city. The kernel displayed the distribution as a histogram, here we use a density plot to examine the number of buyers at the retail store and their spread of purchase amounts.

This illustration of buyer purchases shows a downward slope. A portion of shoppers from city B spend more than their peers in cities A and C. Additionally, not many shoppers in city C lie in the right tail - there are fewer customers spending more on purchases. City C is positively skewed, meaning there are more shoppers in C spent less per transaction. City A’s spending habits answers our question in **Hypothesis 1** that City A does not spend significantly differently than other cities in the retail firm. This is because city A is less positively skewed, meaning there are more shoppers in A who spend more per transaction. Also, most of shoppers in City B spent $340,086, while most of shoppers in City A and C spent $301,493 and $186,200, respectively.

This report is trying to determine the customers and demographics that maximize profits on Black Friday. The questions that help this report arrive at profit maximization are 1) Do people living in District A spend differently than those living in District B and C? and 2) Which product are the most profitable? To answer the first question, the variables Shopper Purchase, City_Category, Gender and Age are selected to conduct a two-sample t-test. Based on the categorical nature of City_Category, Gender and Age (measured as factor levels of age ranges), the two-sample t-test was chosen.

```
## 
##  Welch Two Sample t-test
## 
## data:  district_a and district_b
## t = 0.39273, df = 1801.8, p-value = 0.3473
## alternative hypothesis: true difference in means is greater than 0
## 95 percent confidence interval:
##  -61740.8      Inf
## sample estimates:
## mean of x mean of y 
##   1239874   1220522
```

```
## 
##  Welch Two Sample t-test
## 
## data:  district_a and district_c
## t = 16.765, df = 1111.6, p-value < 2.2e-16
## alternative hypothesis: true difference in means is greater than 0
## 95 percent confidence interval:
##  647380.1      Inf
## sample estimates:
## mean of x mean of y 
## 1239874.4  522003.2
```

```
##
##  Welch Two Sample t-test
##
## data:  district_b and district_c
## t = 26.244, df = 2008.9, p-value < 2.2e-16
## alternative hypothesis: true difference in means is greater than 0
## 95 percent confidence interval:
##  654719.2      Inf
## sample estimates:
## mean of x mean of y
## 1220522.3  522003.2
```

In the tests above, we are testing whether there is a difference between the purchase amount of two cities at a time using a paired two sample t-test. Based on average purchase per shopper, the hypothesis is *People living in City A spend more than their peers in City B and C.* The results yield that shoppers in city A did not spend significantly more than their peers in city B but did spend more than peers in city C. Shoppers in city B spent significantly more than their peers in city C.

Therefore, the shopping habits of customers in city A are significantly different from the remaining cities. The results also indicate shoppers in city C spend significantly less than cities A and B. Based on this point, the stores can focus more on learning the spending habits in city C to further understand factors that may lead to the low purchases in city C. For example, is it because that the income level (`Occupation`), that shoppers in city C are new to the city and are not familiar with local stores (`Stay_In_Current_City_Years`) or that the products sold (`Product_Category`) do not meet the needs of shoppers in city C?

Understanding Shopper Purchases

Apriori: Segmenting by City

Starting with apriori segmentation of the Black Friday data set, the model used is:

$$log(ShopperPurchase) = NumberofProductsSold + factor(Age) + factor(Gender) + factor(CityCategory) + factor(Occupation)$$

(1)

The dependent variable is the `Shopper Purchase` which is the total amount each unique shopper purchases, measured in dollars. The independent variables in **Equation 1** are `Number of Products Sold` , `City Category` , `Age` , `Gender` , and `Occupation` . `Number of Products Sold` is the total number of items a shopper purchases and is a continuous variable. `City Category` is a categorical variable with three levels and is chosen as the variable to conduct the apriori segmentation. `Gender` (male or female) and `Occupation` (21 categories) are categorical variables and `Age` (years) is an ordinal variable, but all are treated the same and factored in our model. To better understand how the independent variables influence `Shopper Purchase` , the model is enhanced by regressing the logarithm of `Shopper Purchase` on all independent variables. This will allow for a proportional analysis of the increase in shopper purchasing amounts across all independent variables. **Equation 1** yields an AIC of 7199.758, the lowest AIC of all iterations of possible models given the initial variables in the Black Friday data set. This model excludes the variables `Stay_In_City_Years` and `Marital_Status` . To achieve the Apriori segmentation, this model is further divided into three segments based on City_Category:

Equation 1 regresses average shopper purchase on the number of products sold given demographic actors of age, gender and occupation. This model is then segmented by shoppers residing in District A, District B and District C. The apriori segmentation of **Equation 1** uses a subset of the Black Friday data set collapsed on unique User ID.

	Dependent variable:		
	log(`Shopper Purchase`)		
	(1)	(2)	(3)
`Number of Products Sold`	0.006***	0.008***	0.017***
	(0.0001)	(0.0001)	(0.0002)
factor(Age)18-25	-0.073	-0.012	0.075
	(0.201)	(0.110)	(0.051)
factor(Age)26-35	-0.006	0.043	0.092*
	(0.201)	(0.110)	(0.051)
factor(Age)36-45	0.027	0.014	0.091*
	(0.204)	(0.112)	(0.052)
factor(Age)46-50	-0.200	-0.002	0.100*
	(0.215)	(0.116)	(0.054)
factor(Age)51-55	-0.065	0.115	0.073
	(0.213)	(0.117)	(0.055)
factor(Age)55+	-0.257	0.019	0.021
	(0.221)	(0.129)	(0.056)
factor(Gender)M	0.130***	0.148***	0.081***
	(0.043)	(0.027)	(0.015)
factor(Occupation)1	-0.039	0.015	0.037
	(0.081)	(0.053)	(0.029)
factor(Occupation)2	0.142	-0.065	0.036
	(0.089)	(0.064)	(0.039)
factor(Occupation)3	0.067	0.136*	0.050
	(0.115)	(0.078)	(0.043)
factor(Occupation)4	0.103	0.061	0.051*

	(0.074)	(0.050)	(0.030)
factor(Occupation)5	0.005	-0.004	0.065
	(0.168)	(0.083)	(0.051)
factor(Occupation)6	-0.008	-0.026	-0.021
	(0.132)	(0.066)	(0.037)
factor(Occupation)7	0.031	-0.031	0.027
	(0.077)	(0.051)	(0.027)
factor(Occupation)8	-0.317	-0.555	-0.081
	(0.302)	(0.343)	(0.109)
factor(Occupation)9	-0.100	0.079	-0.044
	(0.232)	(0.104)	(0.052)
factor(Occupation)10	-0.069	0.032	0.087
	(0.211)	(0.117)	(0.055)
factor(Occupation)11	0.201	0.024	0.087*
	(0.147)	(0.080)	(0.048)
factor(Occupation)12	0.097	0.150***	0.095***
	(0.087)	(0.057)	(0.032)
factor(Occupation)13	-0.370*	-0.244**	0.056
	(0.192)	(0.103)	(0.045)
factor(Occupation)14	0.097	0.053	0.080**
	(0.097)	(0.063)	(0.034)
factor(Occupation)15	0.240*	0.126	0.076*
	(0.131)	(0.081)	(0.046)
factor(Occupation)16	0.214*	-0.092	0.013
	(0.116)	(0.071)	(0.036)
factor(Occupation)17	0.077	0.086	0.071**
	(0.093)	(0.054)	(0.028)
factor(Occupation)18	-0.045	-0.062	-0.013
	(0.231)	(0.143)	(0.055)
factor(Occupation)19	0.082	-0.018	-0.047
	(0.170)	(0.119)	(0.061)
factor(Occupation)20	0.026	0.044	0.042
	(0.094)	(0.058)	(0.039)
Constant	12.575***	12.418***	11.763***
	(0.205)	(0.111)	(0.053)
Observations	1,045	1,707	3,139
R^2	0.719	0.783	0.809
Adjusted R^2	0.711	0.779	0.808
Residual Std. Error	0.593 (df = 1016)	0.479 (df = 1678)	0.356 (df = 3110)
F Statistic	92.654*** (df = 28; 1016)	215.957*** (df = 28; 1678)	471.768*** (df = 28; 3110)
Note:			*p<0.1;* ***p<0.05;*** p<0.01

Overall, the proportion of the variance in Shopper Purchase is explained by the about 71.1% of the independent variables in City A, 77.9% of independent variables in City B and 80.8% of independent variables in City C. All of which is analyzed using the Adjusted R2, with respect to the addition of one more variable to the model. This relationship is statistically significant across all City Categories as the F-statistics are all statistically significant at the 0.01% level. Considering of the marketing cost, we suppose that the retail store should only focus on implementations with more than 95% significant level.

The above segments reveal interesting patterns. We find -

1. `gender` and `Number of Products Sold` affects all segments. All these coefficients are significant at the 99% confidence level. Each additional product sold will significantly increases the `Shopper Purchase` by 0.6% in city A, 0.8% in city B and 1.7% in city C. So for city C, managers should focus on developing strategies that boost sales and increase the number of products sold. With respect to `gender` , a male shopper will significantly increase the `Shopper Purchase` by 13% in city A, 14.8% in city B and 8.1% in city C. While the effect of female shoppers is significant, the magnitude of the increase in `Shopper Purchase` of a woman is smaller than that of a man's. So managers should adjust the supply that meets the need of male shoppers, especially in city B.
2. With respect to `age` , shoppers who are between 26-50 years old in city C have a significant impact on the `Shopper Purchase` . They increase the `Shopper Purchase` amount by 9.2%-10%. Therefore, for city C, managers can launch new promotions that increase the number of shoppers in age range 26-50 respectively.
3. With respect to `Occupation` , occupation 12 increases the `Shopper Purchase` of both city B and C by 15% and 9.5%. Managers should further analyze the purchase behaviors for this type of job, making marketing strategy more personal. For city C, occupation 14 and 17 increase the Shopper Purchase by 8% and 7.1%. Occupation 13 has a negative impact on `Shopper Purchase` in cities A and B, -37% and -24.4% respectively. Therefore, we suggest managers either do not spend any of the marketing budget on these segments, or further analyze the shopper profiles to better understand the reason of the negative effect.

Post-Hoc: Segmenting by Clusters

The relative quality of the apriori segmentation models, as mentioned above is determined by **Equation 1**'s AIC. The AIC for segmentation by City_Category is 7,199.758. To test whether **Equation 1** is a good model for estimating shopper purchasing amounts, we used post-hoc segmentation as a comparison. The equation remains the same but the segments differ. By scaling the Black Friday data, K-means clustering suggests 6 clusters may be a good alternative to apriori segmentation by City_Category. Using **Equation 1**, post-hoc analysis is conducted with 6 segments using the following equation:

$$log(ShopperPurchase) = NumberofProductsSold + factor(Age) + factor(Gender) + factor(Occupation) + e$$

(2)

Equation 2 is broken down into 6 segments based upon Post-Hoc analysis via k-means clustering. All segments have more men than women. Segment 1 is called `Senior Citizens` because it contains only shoppers ages 55+. `Nomadic & Married` is the second segment because most shoppers stay in their city less than a year and are married. Segment 3 is the `Single City Dwellers` because most shoppers are single and live in their city for 4 or more years. Cluster `One and Done` refers to segment 4 because all shoppers stay in the city for just one year. Segment 5 refers to `Single Professionals` because most shoppers are single in age ranges 26-35. Lastly, segment 6 is labeled `Young & Single` because the shoppers are single and in a younger age range, 18-35 years old.

	Dependent variable:					
	log(Shopper.Purchase)					
	(1)	(2)	(3)	(4)	(5)	(6)
Number.of.Products.Sold	0.007***	0.008***	0.007***	0.008***	0.008***	0.009***
	(0.0002)	(0.0002)	(0.0002)	(0.0001)	(0.0002)	(0.0002)
factor(Age)18-25	-0.307		0.119	0.049		
	(0.276)		(0.254)	(0.141)		
factor(Age)26-35	-0.225		0.129	0.058		
	(0.228)		(0.239)	(0.137)		
factor(Age)36-45	-0.137		0.181	0.021		0.258**
	(0.229)		(0.234)	(0.141)		(0.101)
factor(Age)46-50	-0.179		0.156	-0.003		0.227**
	(0.234)		(0.238)	(0.148)		(0.104)
factor(Age)51-55	-0.168		0.137	0.175		0.268**
	(0.237)		(0.239)	(0.149)		(0.105)
factor(Age)55+	-0.262		0.022	-0.107		0.160
	(0.239)		(0.242)	(0.173)		(0.107)
factor(Occupation)1	-0.017	0.010	-0.076	0.017	0.133	-0.032
	(0.083)	(0.074)	(0.098)	(0.071)	(0.110)	(0.062)
factor(Occupation)2	0.045	0.059	0.004	-0.009	-0.098	0.110
	(0.104)	(0.087)	(0.126)	(0.081)	(0.109)	(0.093)
factor(Occupation)3	0.121	0.017	-0.159	0.251**	0.122	0.067
	(0.127)	(0.098)	(0.143)	(0.108)	(0.146)	(0.095)
factor(Occupation)4	0.031	0.030	-0.054	0.134*	0.059	0.152
	(0.107)	(0.069)	(0.136)	(0.069)	(0.064)	(0.168)
factor(Occupation)5	0.142	0.032	-0.759***	0.121	-0.074	0.137
	(0.158)	(0.119)	(0.209)	(0.098)	(0.149)	(0.140)
factor(Occupation)6	-0.250**	-0.029	-0.140	0.112	-0.276	-0.149*
	(0.108)	(0.097)	(0.123)	(0.087)	(0.203)	(0.082)
factor(Occupation)7	-0.202***	0.010	-0.082	0.008	0.154	0.005
	(0.077)	(0.062)	(0.092)	(0.068)	(0.154)	(0.060)
factor(Occupation)8		0.103	0.496	-0.543	-0.850	-0.222
		(0.231)	(0.560)	(0.344)	(0.554)	(0.217)
factor(Occupation)9	-0.313**	-0.022	-0.203	0.118	0.235	-0.079
	(0.155)	(0.138)	(0.167)	(0.153)	(0.283)	(0.123)
factor(Occupation)10	-0.246		0.145	0.019	0.065	0.238**
	(0.250)		(0.264)	(0.152)	(0.127)	(0.113)
factor(Occupation)11	-0.063	0.210*	0.052	0.024	0.253	-0.005
	(0.163)	(0.115)	(0.152)	(0.101)	(0.216)	(0.104)
factor(Occupation)12	-0.012	0.072	0.005	0.197***	0.205**	0.210**
	(0.091)	(0.073)	(0.117)	(0.074)	(0.097)	(0.082)
factor(Occupation)13	-0.108		-0.209	-0.176		0.033
	(0.175)		(0.133)	(0.137)		(0.082)
factor(Occupation)14	-0.063	-0.024	-0.085	0.211**	-0.014	0.137*
	(0.096)	(0.080)	(0.122)	(0.085)	(0.113)	(0.083)
factor(Occupation)15	0.234*	0.054	0.053	0.111	0.177	0.065
	(0.142)	(0.100)	(0.160)	(0.106)	(0.169)	(0.109)
factor(Occupation)16	0.073	-0.041	-0.011	-0.179*	-0.047	0.056
	(0.105)	(0.117)	(0.117)	(0.102)	(0.154)	(0.074)
factor(Occupation)17	-0.026	0.041	0.013	0.137*	0.243**	-0.010
	(0.084)	(0.069)	(0.108)	(0.070)	(0.105)	(0.067)
factor(Occupation)18	-0.009	0.075	-0.284	0.098	-0.202	-0.138
	(0.217)	(0.135)	(0.258)	(0.176)	(0.204)	(0.134)
factor(Occupation)19	0.077	-0.064	0.534*	0.094	0.025	0.205
	(0.233)	(0.181)	(0.304)	(0.154)	(0.119)	(0.172)
factor(Occupation)20	0.049	-0.112	0.063	0.146*	-0.021	-0.145
	(0.110)	(0.087)	(0.105)	(0.076)	(0.107)	(0.107)
factor(Gender)M	0.168***	0.162***	0.170***	0.138***	0.120***	0.103***
	(0.045)	(0.040)	(0.051)	(0.037)	(0.043)	(0.035)
factor(City_Category)B	0.232***		0.167***		0.089	
	(0.058)		(0.056)		(0.056)	
factor(City_Category)C	0.019	0.032	0.004		0.072	0.115**

	(0.057)	(0.039)	(0.068)		(0.048)	(0.048)
Constant	12.600***	12.340***	12.196***	12.352***	12.208***	11.921***
	(0.235)	(0.062)	(0.234)	(0.140)	(0.077)	(0.115)
Observations	817	1,142	642	947	918	1,425
R^2	0.713	0.670	0.734	0.787	0.699	0.585
Adjusted R^2	0.703	0.664	0.721	0.780	0.691	0.577
Residual Std. Error	0.549 (df = 787)	0.555 (df = 1120)	0.550 (df = 611)	0.476 (df = 918)	0.550 (df = 894)	0.559 (df = 1397)
F Statistic	67.453*** (df = 29; 787)	108.524*** (df = 21; 1120)	56.239*** (df = 30; 611)	121.103*** (df = 28; 918)	90.169*** (df = 23; 894)	72.891*** (df = 27; 1397)
Note:						*p<0.1;* ***p<0.05;*** p<0.01

Across segments, managers should focus on increasing the number of products sold, but specifically on segment 6, the `Young & Single`. When one additional product is sold to the `Young & Single`, `Shopper Purchase` increases by 0.9%, holding other factors constant at the 0.01% significance level. Most statistically significant age range for the Black Friday shoppers is to focus on the older customers in the `Young & Single` segment. Specifically shoppers ages 36-45 years old have the largest magnitude impact, ranging in a 22.7-26.8% increase in `Shopper Purchase` amount, holding other factors constant. Ages 18-55 are not statistically relevant for “Senior Citizens”, “Nomadic & Married” and “One and Done”.

With respect to gender, both men and women have a statistically significant relationship with average purchase amount per shopper in each segment. However, across segments women have a smaller impact on `Shopper Purchase` than men. A male `Single City Dweller` increases the `Shopper Purchase` by 17%. A female `senior citizen` has the largest impact on `Shopper Purchase`, with respect to all women in the other segments.

Occupation categories that do not have a statistically significant relationship with `Shopper Purchase` amount in any cluster segment are categories 1, 2, 8, 13, 14, 16 and 18. Relevant in segment 1 only is Occupation categories 6, 7, 9, 15, and 21. Occupations 6, 7 and 9 negatively impact `Shopper Purchase` by an average of about 25%. Occupation category 5 in `Single City Dwellers` decreases the amount a shopper buys by about 76%. Consequently, no effort or funds should be spent marketing towards this segment. In the same segment, however, Occupation 19 positively increases `Shopper Purchase` by 53.4%, holding other factors constant. Occupation 12 is significant in segments `One and Done`, `Single Professionals` and `Young & Single`. These results indicate that a unique marketing approach based on the Occupation of the customer might be necessary.

Lastly, City A is statistically significant predictor of a positive increase in `Shopper Purchase` amount in all 6 segments. `Senior Citizens` living in City B increase the amount shoppers spend by 23.2% and `Single City Dwellers` living in city B increase customer spending by 16.7%. `Young & Single` customers residing in city C increase `Shopper Purchase` by 11.5%.

Model Selection

In terms of model selection, apriori regression versus post-hoc regression, a comparison of the AIC’s above suggests the quality regression model is the Apriori segmentation. Post-Hoc segmentation yields AIC=9,663.19 whereas Apriori segmentation yields AIC=7,199.76. Therefore, the managers should use the below analysis of the Apriori shopper segmentation to inform marketing policy.

Shopper Purchases by City and Product Category

The final analysis examines shopper purchases on the product categories of the retail firm. The model omits the demographic variables of Marital_Status and Stay_In_City_Years. This is because the smallest AIC (AIC= -28,687.26) is derived using the equation below:

$$log(ShopperPurchase) factor(ProductCategory) + NumberofProductsSold + factor(Gender) + factor(Age) + factor(CityCategory) + factor(Occu$$

(3)

Model 3 is segmented by city category because Apriori analysis resulted in a smaller AIC as determined by the comparison of apriori and post-hoc AIC. User ID is used to match the clusters to the user category subset of Black Friday data.

	Dependent variable:		
	log(`Shopper Purchase`)		
	(1)	(2)	(3)
`Number of Products Sold`	0.029***	0.036***	0.069***
	(0.0003)	(0.0003)	(0.0004)
factor(Product_Category)2	-0.743***	-0.637***	-0.670***
	(0.039)	(0.029)	(0.019)
factor(Product_Category)3	-0.948***	-0.819***	-0.856***
	(0.040)	(0.029)	(0.020)
factor(Product_Category)4	-2.705***	-2.571***	-2.547***
	(0.042)	(0.030)	(0.021)
factor(Product_Category)5	-0.762***	-0.810***	-0.721***
	(0.036)	(0.026)	(0.016)
factor(Product_Category)6	-0.422***	-0.290***	-0.374***
	(0.039)	(0.029)	(0.019)
factor(Product_Category)7	-0.958***	-0.972***	-0.761***
	(0.050)	(0.038)	(0.032)
factor(Product_Category)8	-0.592***	-0.582***	-0.594***
	(0.036)	(0.026)	(0.017)
factor(Product_Category)9	-1.825***	-1.640***	-1.158***
	(0.083)	(0.062)	(0.060)
factor(Product_Category)10	-0.877***	-0.788***	-0.531***

	(0.045)	(0.034)	(0.025)
factor(Product_Category)11	-1.692***	-1.581***	-1.651***
	(0.041)	(0.030)	(0.020)
factor(Product_Category)12	-3.512***	-3.418***	-3.160***
	(0.050)	(0.037)	(0.030)
factor(Product_Category)13	-4.017***	-3.962***	-3.819***
	(0.046)	(0.034)	(0.025)
factor(Product_Category)14	-1.502***	-1.483***	-1.127***
	(0.058)	(0.044)	(0.038)
factor(Product_Category)15	-0.993***	-1.003***	-0.785***
	(0.045)	(0.033)	(0.024)
factor(Product_Category)16	-0.825***	-0.789***	-0.683***
	(0.043)	(0.031)	(0.021)
factor(Product_Category)17	-1.945***	-1.849***	-1.406***
	(0.096)	(0.059)	(0.054)
factor(Product_Category)18	-2.778***	-2.610***	-2.399***
	(0.055)	(0.040)	(0.032)
factor(Gender)M	0.071***	0.081***	0.001
	(0.019)	(0.014)	(0.010)
factor(Occupation)1	-0.028	-0.019	-0.008
	(0.035)	(0.026)	(0.018)
factor(Occupation)2	0.035	0.033	-0.026
	(0.038)	(0.032)	(0.025)
factor(Occupation)3	0.066	0.081**	0.057**
	(0.048)	(0.038)	(0.027)
factor(Occupation)4	0.004	0.010	0.019
	(0.032)	(0.025)	(0.019)
factor(Occupation)5	0.0003	0.049	-0.054
	(0.074)	(0.040)	(0.033)
factor(Occupation)6	0.026	-0.008	-0.032
	(0.057)	(0.032)	(0.025)
factor(Occupation)7	-0.071**	-0.040	-0.029*
	(0.034)	(0.025)	(0.017)
factor(Occupation)8	-0.793***	0.245	-0.186***
	(0.161)	(0.154)	(0.071)
factor(Occupation)9	-0.083	0.059	-0.012
	(0.106)	(0.053)	(0.036)
factor(Occupation)10	-0.177*	-0.169***	-0.015
	(0.094)	(0.061)	(0.036)
factor(Occupation)11	0.031	-0.095**	0.032
	(0.065)	(0.039)	(0.030)
factor(Occupation)12	-0.126***	-0.029	0.030
	(0.039)	(0.028)	(0.021)
factor(Occupation)13	-0.252**	-0.140**	-0.022
	(0.103)	(0.055)	(0.029)
factor(Occupation)14	0.034	-0.010	0.006
	(0.042)	(0.032)	(0.022)
factor(Occupation)15	-0.003	0.009	0.042
	(0.056)	(0.039)	(0.029)
factor(Occupation)16	0.158***	0.046	-0.023
	(0.048)	(0.034)	(0.023)
factor(Occupation)17	-0.012	-0.050*	-0.004
	(0.041)	(0.027)	(0.018)
factor(Occupation)18	0.142	0.059	0.006
	(0.092)	(0.069)	(0.035)
factor(Occupation)19	0.076	0.041	-0.081**
	(0.072)	(0.056)	(0.038)
factor(Occupation)20	0.119***	0.032	0.023
	(0.039)	(0.028)	(0.025)
factor(Age)18-25	-0.045	-0.116**	-0.039
	(0.088)	(0.057)	(0.034)
factor(Age)26-35	0.046	-0.053	-0.019
	(0.087)	(0.057)	(0.034)
factor(Age)36-45	0.081	-0.035	0.022
	(0.088)	(0.058)	(0.035)
factor(Age)46-50	0.013	-0.008	0.009
	(0.094)	(0.060)	(0.036)
factor(Age)51-55	-0.044	0.001	0.033
	(0.093)	(0.060)	(0.036)
factor(Age)55+	-0.187*	-0.087	0.010
	(0.098)	(0.067)	(0.037)
Constant	11.235***	11.187***	10.676***

	(0.093)	(0.061)	(0.037)
Observations	10,412	17,390	24,569
R^2	0.769	0.795	0.809
Adjusted R^2	0.768	0.794	0.809
Residual Std. Error	0.811 (df = 10366)	0.759 (df = 17344)	0.643 (df = 24523)
F Statistic	765.087*** (df = 45; 10366)	1,491.769*** (df = 45; 17344)	2,306.433*** (df = 45; 24523)
Note:			*p<0.1;* ***p<0.05;*** p<0.01

With the result above, the managers can emphasize increasing the number of products sold in each city. Specifically, in city C, one additional product sold will result in a 6.9% increase in a shopper's purchase. With respect to gender, male shoppers result in higher average purchases. Men in city C spend much less than their peers in city A and B. If marketed to effectively, these men greatly improve sales and ultimately profit.

With respect to product category, product category 1 is dropped in the regression to create a mutually exclusive and exhaustive analysis of products. Consequently, all other products are compared to category 1 which results in negative coefficients on the other products and a decrease in shopper purchases. This is because shopper purchases for category 1 are significantly higher than most other products, as seen in **Figure 2**. Due to this, managers should understand that other products do not necessarily decrease shopper spending, but relative to category 1 do not have as great an impact on spending. To analyze the relationship of the remaining product categories, Managers should focus on the product categories whose coefficients from **Model 3** are closest to 0 (closest to the behavior of products in category 1). Therefore, in general, category 1, 2, 5, 6 and 8 perform well compared to the rest of the products. And across the cities, the store should sell more products in category 6, since they will generate the revenue similar to the products in category 1. From the luxury and normal analysis in **Figure 2**, since product category 6 might be luxury items, managers should target shoppers who are older with a decent job, because they might have higher purchasing power.

As for age, what clashes with common assumptions is that purchase power of shoppers age 0-17 is relatively high - we expect kids and teenagers who do not earn money to spend less. As we can see, shoppers age 55 or more in city A and shoppers age 18-25 in city B will significantly lower the shopper purchase by 18.7% and 11.6% with respect to shoppers age 0-17. Commonly, it is assumed the working population spends their income on retail meaning they should have a positive influence on average shopper spending at the retail store. This might be because customer account with the store is created by a teenager but these teenagers use their parents' card to pay the bill. Shoppers age 26-50, the age range where most shoppers have their own unique credit card, living in city A increase the shopper purchase by 1-9% with respect to shoppers age 0-17. While in city C, shoppers age 36-45 and 51-55 increase the shopper purchase by 2-3%. This might be result from the fact that these shopper have work and earn money, so their purchasing power is higher than teenagers.

Occupations that are not statistically significant predictors of shopper spending habits in any city are Occupations 1, 2, 4, 5, 6, 9, 14, 15, and 18. Across the cities, shoppers with occupation 7,8 and 12 in city A, occupation 10, 11 and 13 in city B, and occupation 8 and 9 will decrease the shopper purchase. This means these shoppers are very budget conscious and do not shop. So managers could either left the market of these groups or conduct campaigns and promotions to stimulate these shoppers. Shoppers with Occupation 3 increase `Shopper Purchase` in city B and C by 8.1% and 5.7%, on average.

Predictive Analysis

To see how managers should focus marketing efforts, predictive analysis is used to answer the questions: Which segments are expected to be the most profitable? Which product categories within each segment should managers devote extra attention in order to increase profits? This is done by splitting the clustered Black Friday data into training the data and test data. The training data is used to choose the best predictive model from 16 models. The model chosen to predict shopper purchases had the smallest predictive error, or root mean square error.

Table 3 Prediction Error

Names	Error
Simple 1	81268.51
Simple 2	81273.54
Store FE 2.1	81244.94
Store FE 2.2	81241.81
Store FE 2.3	81276.16
Store FE 2.4	81291.13
Store FE 2.5	81251.07
City Interactions 3.1	79553.80
City Interactions 3.2	79560.09
City Interactions 3.3	81467.81
Polynomial 4.1	78727.25
Polynomial 4.2	62103.32
log 5.1	129917.59
log 5.2	129911.47
Log*Store FE 6.1	126994.78
Log*Store FE 6.2	126915.65

To choose the model with the lowest prediction error, we calculated the Root Mean Square Error for each model. The above predictive analysis concludes that the model with the smallest root mean square error is Model 4.2. This model is a polynomial which indicates that the relationship between `Number of Products Sold` and `Shopper Purchase` is more complex than a linear relationship. This model is the appropriate model to use when predicting future Shopper Purchase habits of shoppers.

City_Category	Predicted.Profit	LowerLimit	UpperLimit
A	475.81	459.75	491.88
B	452.05	436.02	468.09
C	217.91	202.03	233.79
Total	1145.77	1097.8	1193.75

City_Category	Product_Category	Predicted.Profit	LowerLimit	UpperLimit
A	1	127.46	126.89	128.03
A	5	90.57	89.97	91.17
B	1	128.66	128.09	129.23
B	5	84.02	83.43	84.60
C	1	75.79	75.28	76.31
C	8	22.65	22.13	23.17

Therefore, with 95% confidence, expected total profit on the test data for Black Friday lies between [$1097.8 thousand, $1193.75 thousand]. Managers can further breakdown expected profits by city and product category. If managers wish to target the most profitable city, they should spend their time building marketing strategies around shoppers in City A. Predicted Profit for shoppers living in City A is about $476 thousans; with 95$ 460 thousand, $492 thousand]. Specifically, the firm should focus its marketing efforts on the top profitable product categories - categories 5 and 1. Products in category 5 are the most profitable items in each city, the second most profitable products are in category 1 in cities B and C. In City A, the second most profitable product is category 8.

Conclusion

This report suggests the Black Friday sales data from the retail store should employ multiple marketing strategies. The overall goal should be to maximize expected profits. To maximize profits, it is necessary to increase average shopper revenue which is achieved by by increasing the number of items purchased for each shopper. Revisiting the hypotheses analyzed, the shopper data is segmented by city to focus the marketing team's efforts and establish targeted campaigns. Additionally, this report breaks down the most profitable product categories and demographics within each city.

Across all analysis, an increase in one item sold to a shopper positively impacts the total amount purchased. The results above indicate the retail store should focus its marketing efforts on both men and women who live in City A and buy Product Categories 5 and 8 because they are expected to maximize profits, holding other factors constant. This demographic has a positive effect on individual shopper purchase amounts. While the effect of female shoppers is significant in all cities, the magnitude of the increase in `Shopper Purchase` of a woman is smaller than that of a man's. So managers should adjust the supply that meets the need of male shoppers, especially in city B. Further breaking down gender by age, managers should amplify the increase in Shopper Purchase amount by targeting shoppers 18-50 years old. Managers should targeting shoppers who are older with a decent job, because they are likely to have higher purchasing power.

These customers may have certain jobs that position them to increase profits. Specifically managers should target customers who work in Occupation categories 3, 6 and 20. Shoppers with Occupation 3 increase `Shopper Purchase` in city B and C by 8.1% and 5.7%, on average. In city A, occupation category 16 increases spending by 15.8 for the average shopper and Category 20 by 11.9%.

Lastly, the marketing efforts should emphasis the product category specifically when marketing. The product category information indicates that any product category that follows the behavior purchasing behavior of category 1 will positively impact average shopper purchase amounts. The second and third product categories that yield highest profits are categories 5 and 8. Managers should create marketing strategies to increase sales of each of these categories in all cities. There is room for sales growth in product category 6. In all the cities, the retail store should sell more products in category 6 since they will generate revenue similar to the products in category 1. According to **Figure 2**, products in category 6 might be luxury items that can increase profits. Another approach is to change the pricing strategy on categories 2, 3, and 11 to maximize profits. It is important to note that without knowing what product categories or occupation are, it is hard to determine the the product types or low income occupations in each segment that negatively impact customer spending.

All of the above strategy suggestions should be factored into the marketing budget. Such strategies can come in the forms of promotions, campaigns, increasing price and/or slashing the markup of some products in order to make products affordable. Doing so, the managers can still focus on the above segment of shoppers and maintain, if not increase, purchasing habits.