Introduction

The goal of this report is to give practicable suggestions on how to gain more high-quality users, increase product sales, and then make business plans for future marketing of the E-commerce with multidimensional analysis, through analyzing the purchases of our customers for 1 year in the America E-commerce dataset.

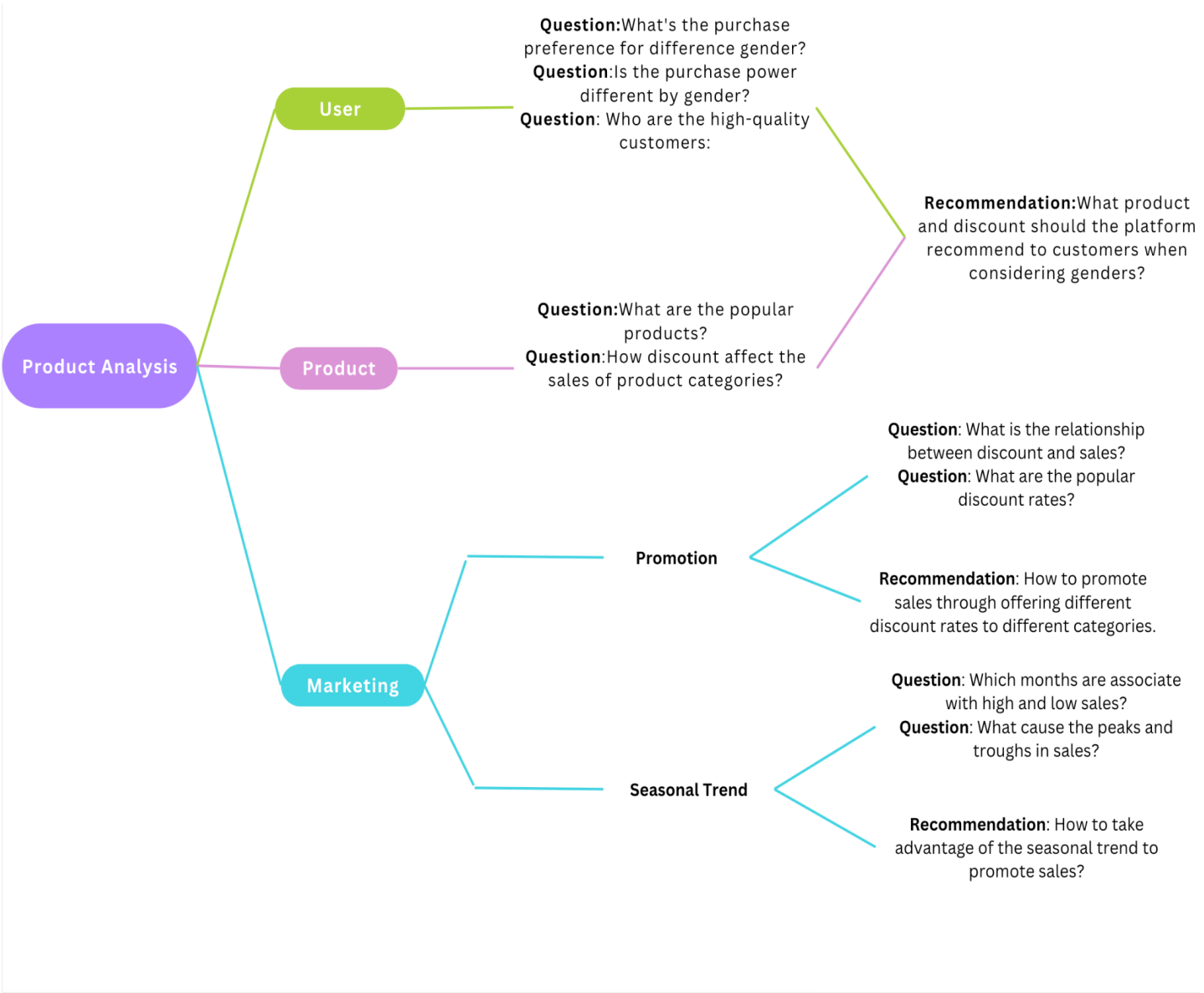
About Dataset

Column	Description
Order_Date	The date the product was ordered
Aging	The time from the day the product is ordered to the day it is delivered
Customer_id	Unique id created for each customer
Gender	Gender of customer
Device_Type	The device the customer uses to actualize the transaction (Web/Mobile)
Customer*Login* Type	The type the customer logged in. Such as Member, Guest etc.
Product_Category	Product_Category
Product	Product
Sales	Sales
Quantity	Unit amount of product
Discount	Percent discount rate
Profit	Profit
Shipping_cost	Shipping_cost
Order_Priority	Order priority. Such as critical, high etc.
Payment_method	Payment method

Purpose

The goal of this report is to give practicable suggestions with statistical evidence on how to gain more high-quality users, promote product sales, and then make business plans for the future marketing of E-commerce.

Mind Map



Data Exploratory

1. Descriptive Statistics

Profit	
Mean	70.4071578
Standard Error	0.21517562
Median	59.9
Mode	17
Standard Deviation	48.7300331
Sample Variance	2374.61612
Kurtosis	-1.4648453
Skewness	0.26096079
Range	167
Minimum	0.5
Maximum	167.5
Sum	3610971.9
Count	51287

Quantity	
Mean	2.50301246
Standard Error	0.00667587
Median	2
Mode	1
Standard Deviation	1.51185881
Sample Variance	2.28571705
Kurtosis	-1.2829514
Skewness	0.46419519
Range	4
Minimum	1
Maximum	5
Sum	128372
Count	51287

Sales	
Mean	152.344356
Standard Error	0.29361694
Median	140
Mode	228
Standard Deviation	66.4943516
Sample Variance	4421.4988
Kurtosis	-1.4400224
Skewness	-0.0879037
Range	217
Minimum	33
Maximum	250
Sum	7813285
Count	51287

Delivery_day	
Mean	5.07884809
Standard Error	0.03432406
Median	5
Mode	1
Standard Deviation	3.06081326
Sample Variance	9.36857782
Kurtosis	-1.3205651
Skewness	0.11353372
Range	9
Minimum	1
Maximum	10
Sum	40387
Count	7952

Shipping	
Mean	7.59264025
Standard Error	0.11262987
Median	7.3
Mode	1.8
Standard Deviation	4.9646653
Sample Variance	24.6479015
Kurtosis	-1.5131511
Skewness	0.10046633
Range	16.7
Minimum	0.1
Maximum	16.8
Sum	14752.5
Count	1943

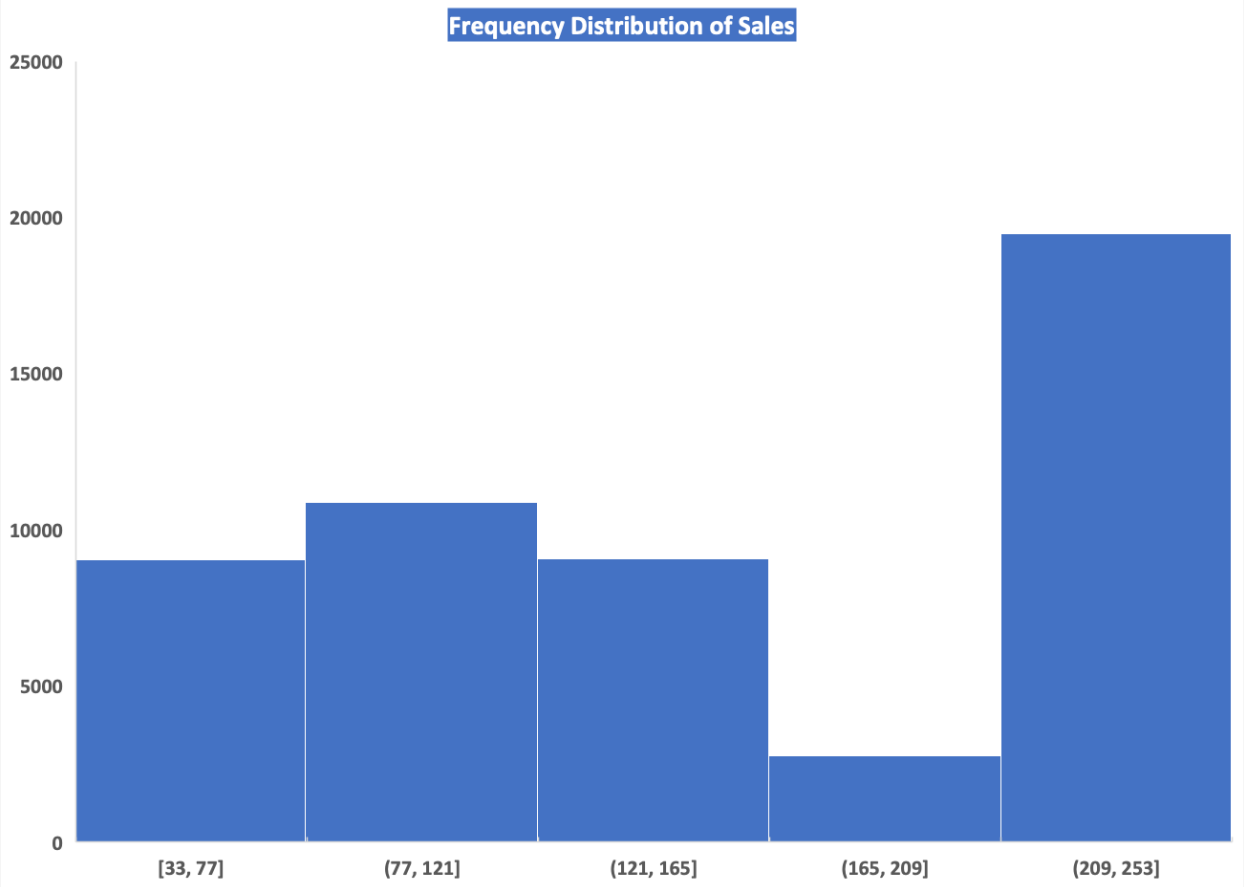
Discount	
Mean	0.30382553
Standard Error	0.00057858
Median	0.3
Mode	0.3
Standard Deviation	0.13102905
Sample Variance	0.01716861
Kurtosis	-1.1227905
Skewness	0.03304111
Range	0.5
Minimum	0
Maximum	0.5
Sum	15582.3
Count	51287

From a business perspective, let's take a closer look at the sales and profit distribution of E-commerce to have an overview of the performance of the business.

Desired Interval	5		
Interval Width	44	←	(Bigest Sales - Smallest Sales)/Desired Interval (5)

Interval Start	Interval End	Requency	Relative Frequency
33	77	9051	0.18
77	121	10892	0.21
121	165	9079	0.18
165	209	2777	0.05
209	253	19487	0.38

COUNTIFS
Ex. Count the number of sales requency for sales in the range: sales >= $209 and <=$253

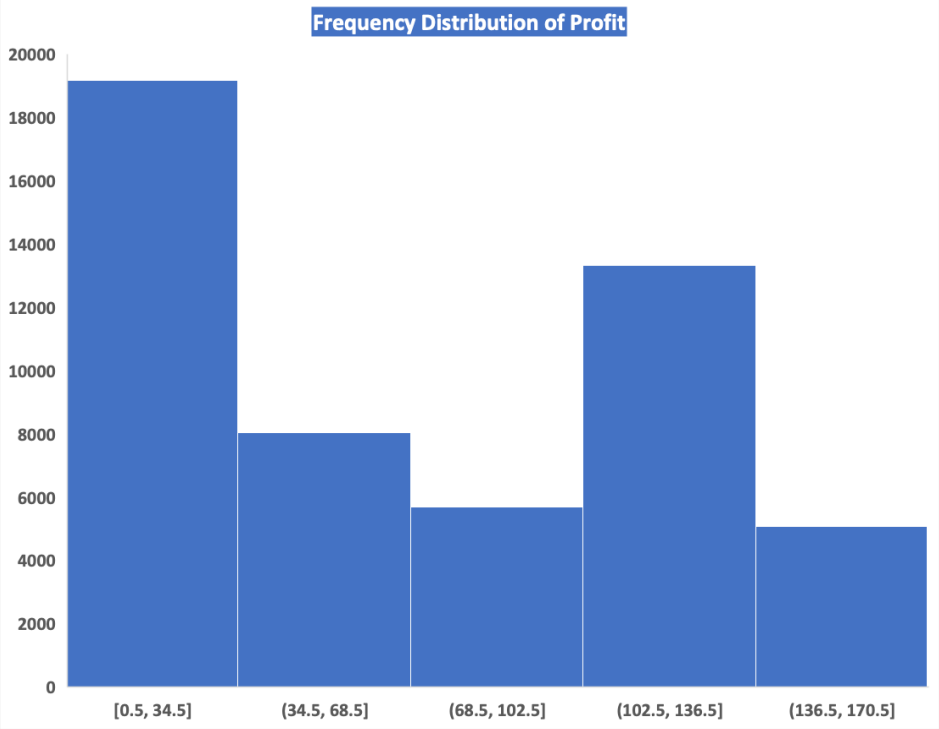


After splitting the sales into 5 intervals, we can see that 38% of the sales are from orders ranging from $209 to $253. It is worth investigating which products are associated with these big orders, so we can have a better idea of the preference of profitable customers.

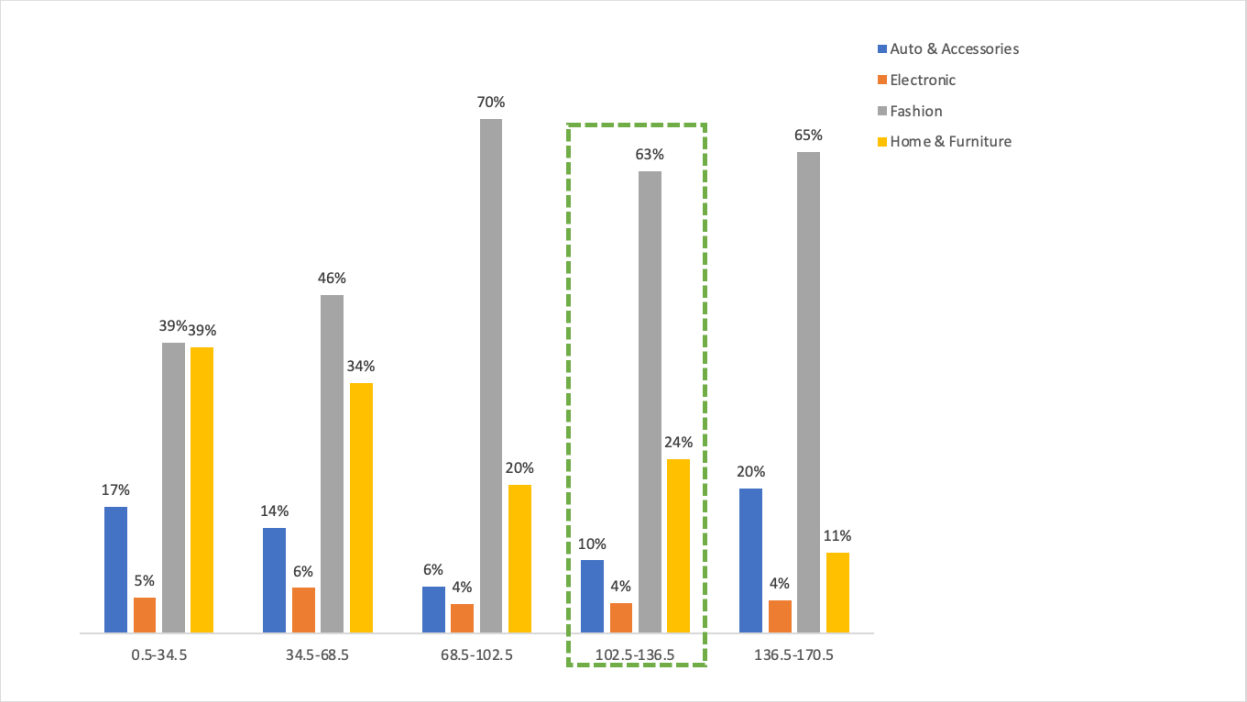
After filtering out only the large orders (orders ranging from $209 to $253), the above plot shows the top 10 most frequently bought products of the large orders. Products, such as Fossil Watch, Jeans, Shirts, T-shirts and Suits are among the top of most frequently bought products and introducing more relevant products to give customers more shopping options would promote large sales.Now, let’s take a look at the overview of the profit.

Desired Interval	5					
Interval Width	34	←	(Highest Profit- Lowerest Profit)/Desired Interval (5)			
Interval Start	Interval End	Requency	Relative Frequency			
0.5	34.5	19171	0.37	}		
34.5	69	8095	0.16	}	53%	
69	103	5687	0.11			
103	137	13624	0.27			
137	171	4774	0.09			
		↑				

COUNTIFS
Ex. Count the number of proft requency for profit in the range: profit >= $137 and <=$171

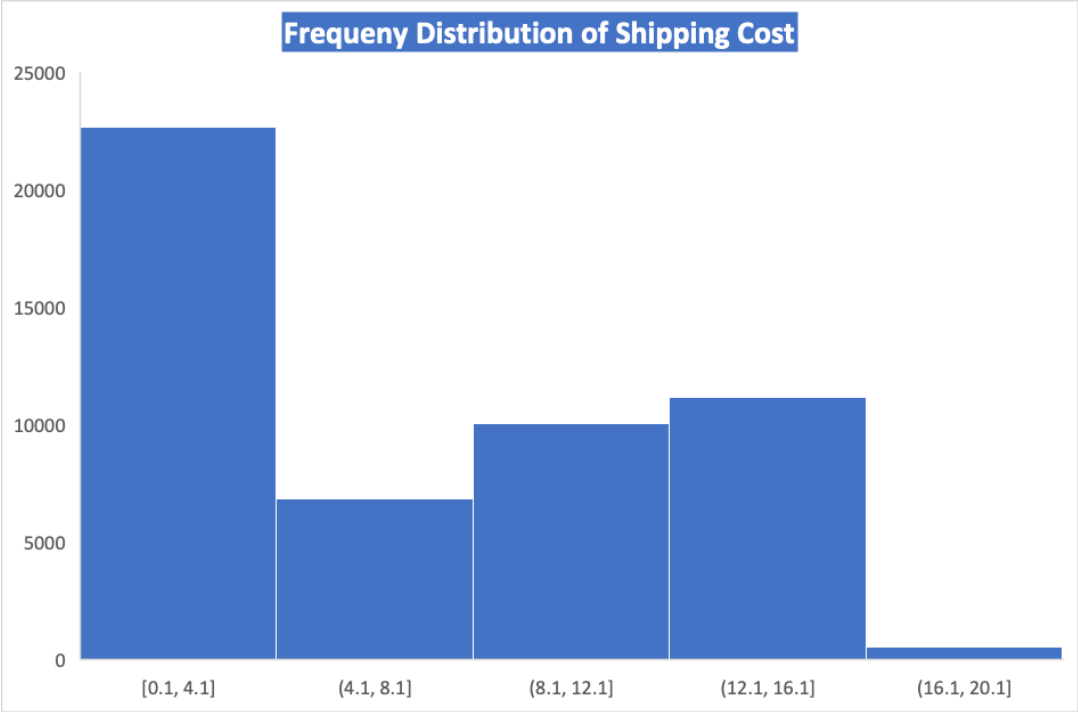


One finding worth investigating further is that 27% of the orders bring in profit between $103 and $137. It is worth investigating to see what orders and products are associated with this large profit.



53% of the orders result in profit under $69. My first assumption is that low profit may result from high cost. Let's explore the shipping cost to see if we can validate the assumption as the shipping cost is the only cost data we have.

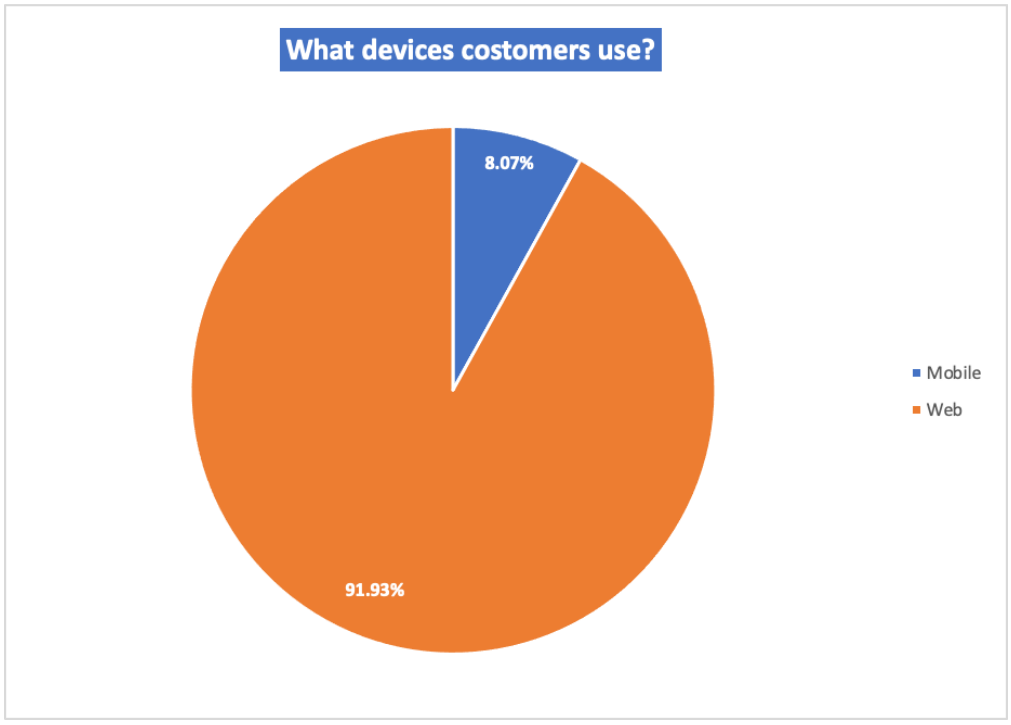
Desired Interval	5							
Interval Width	4	←	(Highest Shipping Cost- Lowerest Shiping cost)/Desired Interval (5)					
Interval Start	**Interval End**	**Requency**	**Relative Frequency**					
0.1	4.1	22681	0.442					
4.1	8.1	7008	0.137					
8.1	12.1	10057	0.196					
12.1	16.1	11642	0.227	}	44%			
16.1	20.1	741	0.014					
		↑						
	COUNTIFS Ex. Count the number of shipping cost requency for cost in the range:							



44% of the orders are associated with shipping costs above $8.1, which indicates that there is a potential need to lower the shipping cost in order to increase profit. However, we only have cost data regarding shipping cost, so a better recommendation can be made if we had more cost data.

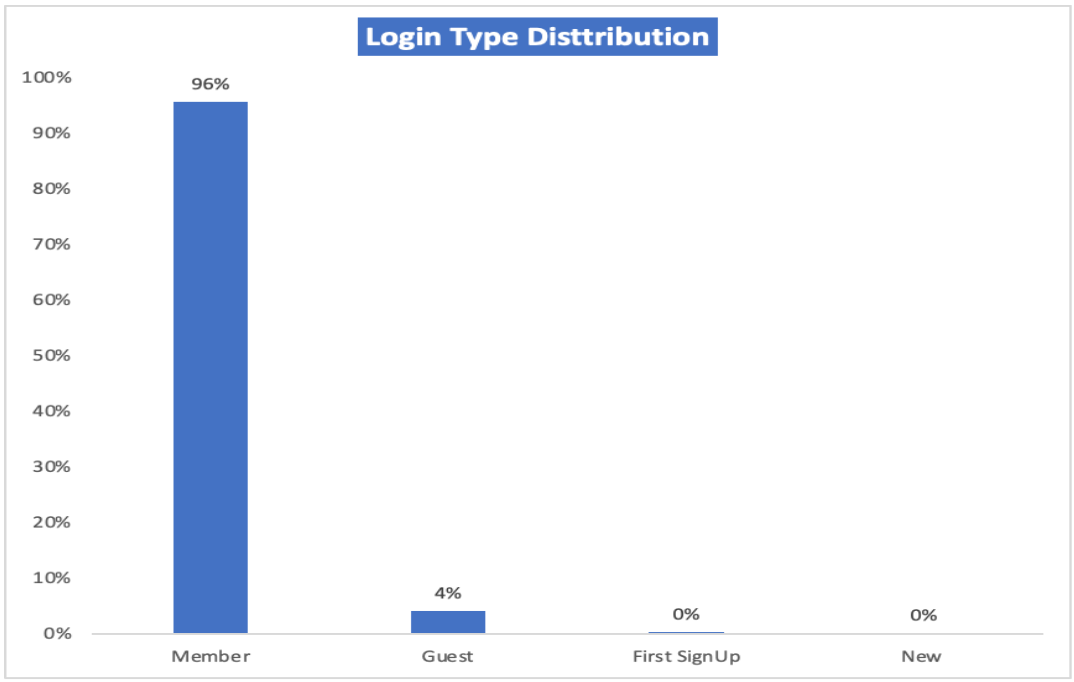
Overview of the Customer Base

What devices do customers use to make purchases?



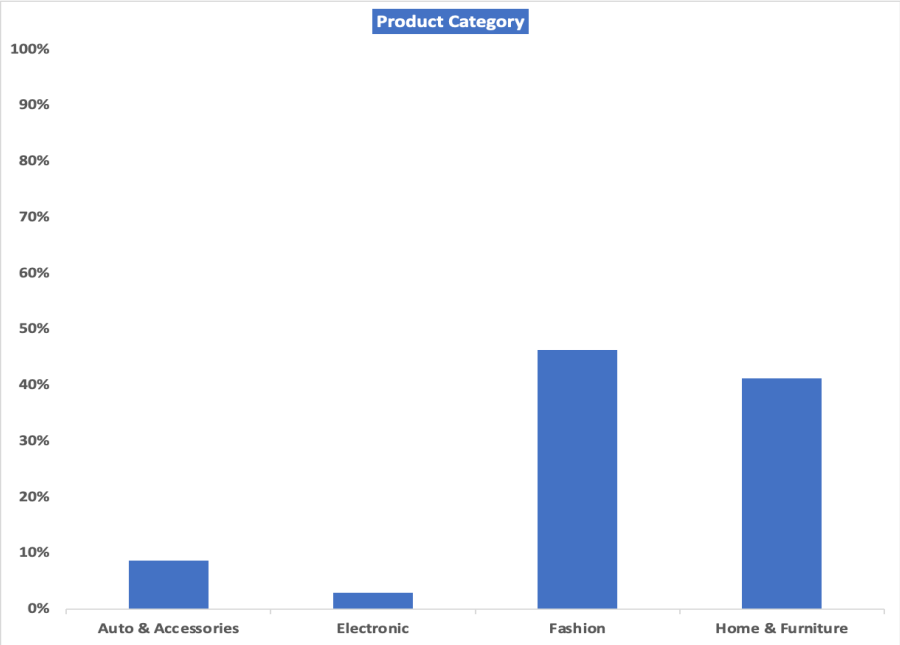
- 91.93% of the customers use Web to make purchases

Who is the customer base?



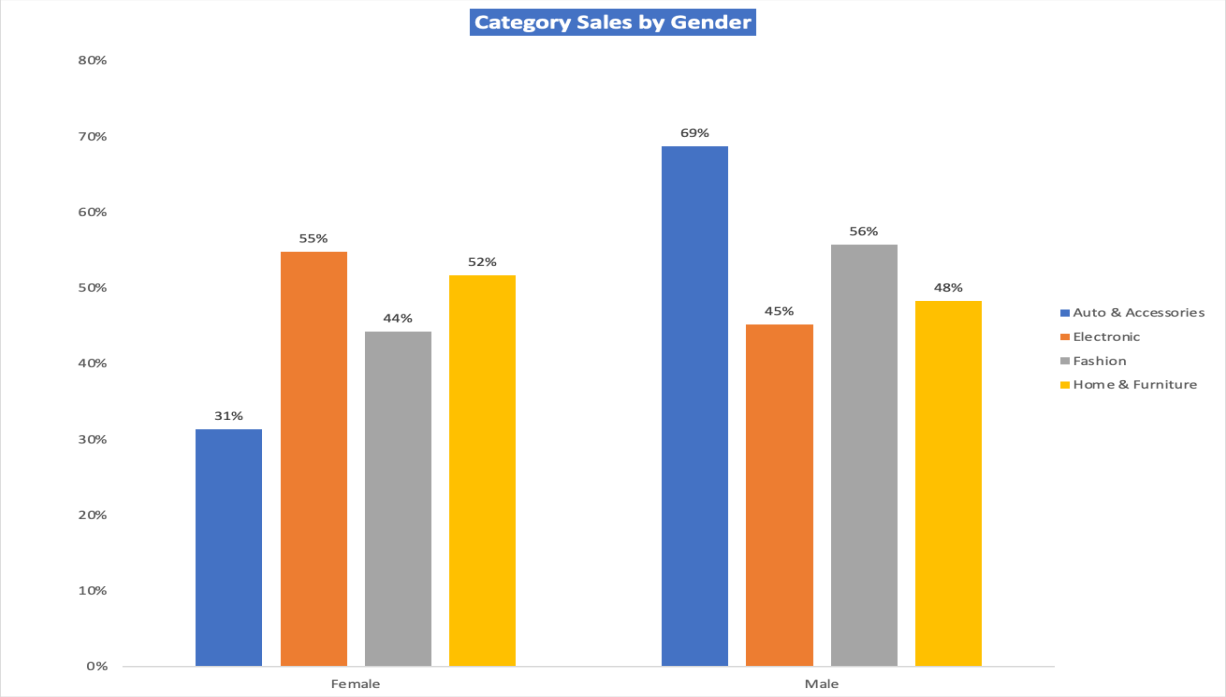
- 96% of the customers are members who have signed up for membership and have made at least one purchase already
- Very few new customers make purchase from the e-commerce

What product categories is e-commerce selling?



- Account 0% of the purchase made from Fashion and Home & Furniture categories

Which product categories do I sell to whom? (Gender Distribution by Category)



- Female and male customers show different preference for each category
- The most popular category for female customer is Electronic and the most popular category for male customer is Auto & Accessories

Data Visualization and Analysis

After exploring the dataset to have an overview of the platform, we see that female and male customers show different preferences toward different product categories, and there are high quality customers who made large purchases. These findings lead us to ask a few questions to dig into the details of each finding and test out the hypothesis to get the valid conclusions before making any recommendations.

Question 1: How Does Gender Influence Sales? What Product should the Platform Recommend to Customers When Considering Genders?

Hypothesis testing: Is there a significant difference in the buy amount between the female and male customers?

Null Hypothesis: $\mu_m = \mu_f$
Alternative Hypothesis: $\mu_m > \mu_f$

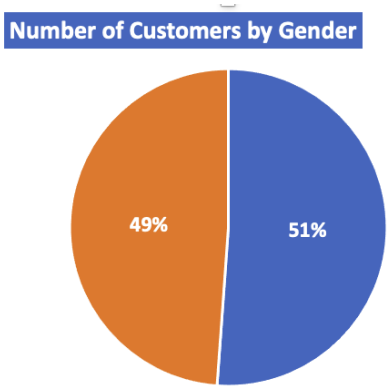
Step one: Check if there is significant difference between the variance of the two samples to decide which test method to use for the hypothesis test.

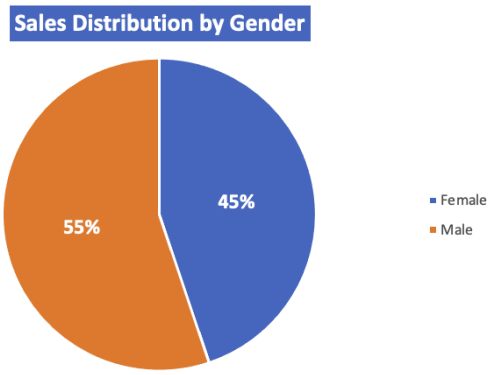
According to the F-test, there is no statistical significant difference between the variance of sales by gender. Thus, I will be running a T-test for two samples assuming equal variances.

F-Test Two-Sample for Variances		
	Male	*Female*
Mean	153.144274	151.365773
Variance	4391.70353	4456.0561
Observations	28134	23151
df	28133	23150
F	0.9855584	
P(F<=f) one-	0.12305571	
F Critical one	0.97958126	

t-Test: Two-Sample Assuming Equal Variances		
	Male	*Female*
Mean	153.147041	151.36891
Variance	4391.76287	4456.09149
Observations	28135	23152
Pooled Variance	4420.80201	
Hypothesized Mean Difference	0	
df	51285	
t Stat	3.01388966	
P(T<=t) one-tail	0.00129024	
t Critical one-tail	1.64488334	
P(T<=t) two-tail	0.00258048	
t Critical two-tail	1.96001024	

Since the p-value is 0.001 <0.05, we reject the null hypothesis. We can conclude there's a significant difference in the mean of purchase amount between female and male customers.





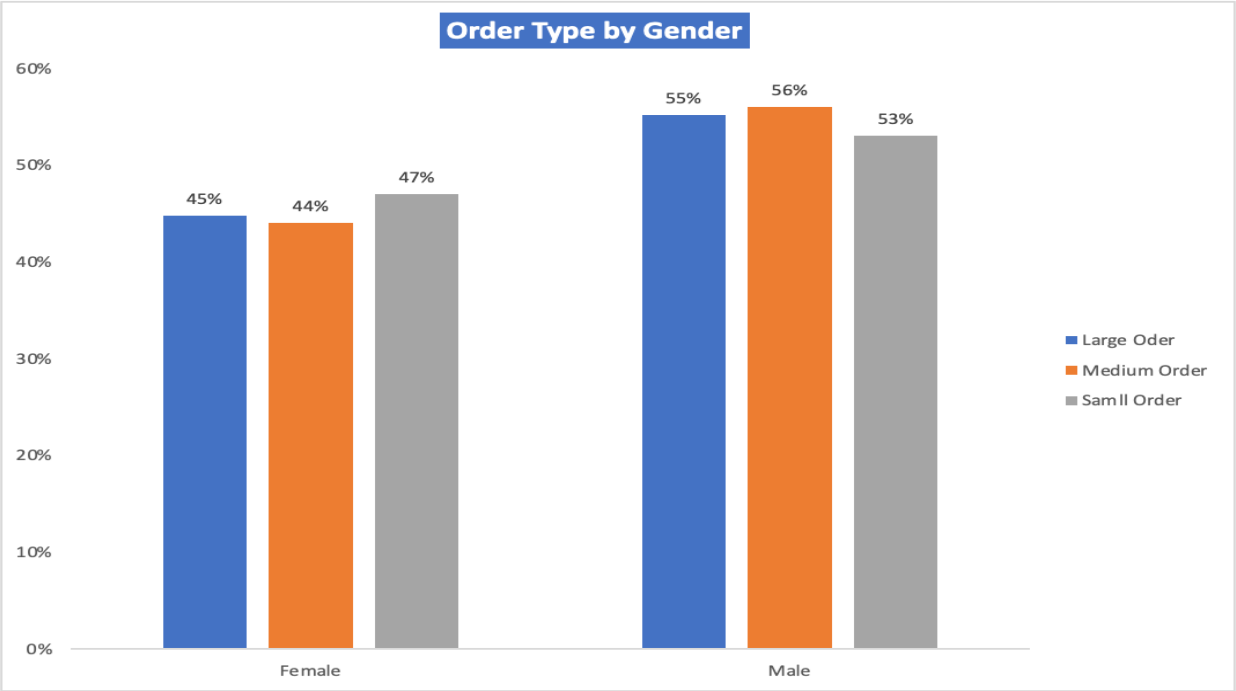
The difference in the number of customers between genders is insignificant:49% for male customers, 51% for female customers. However,

the sales amount in percentage between male and female customers are significantly different: 55% for males and 45% for females, which aligns with our hypothesis that male customers purchased more than females customers on average.

Follow-up Question：Why is the buy amount for males more than females given that the difference in the number of customers between genders is insignificant?

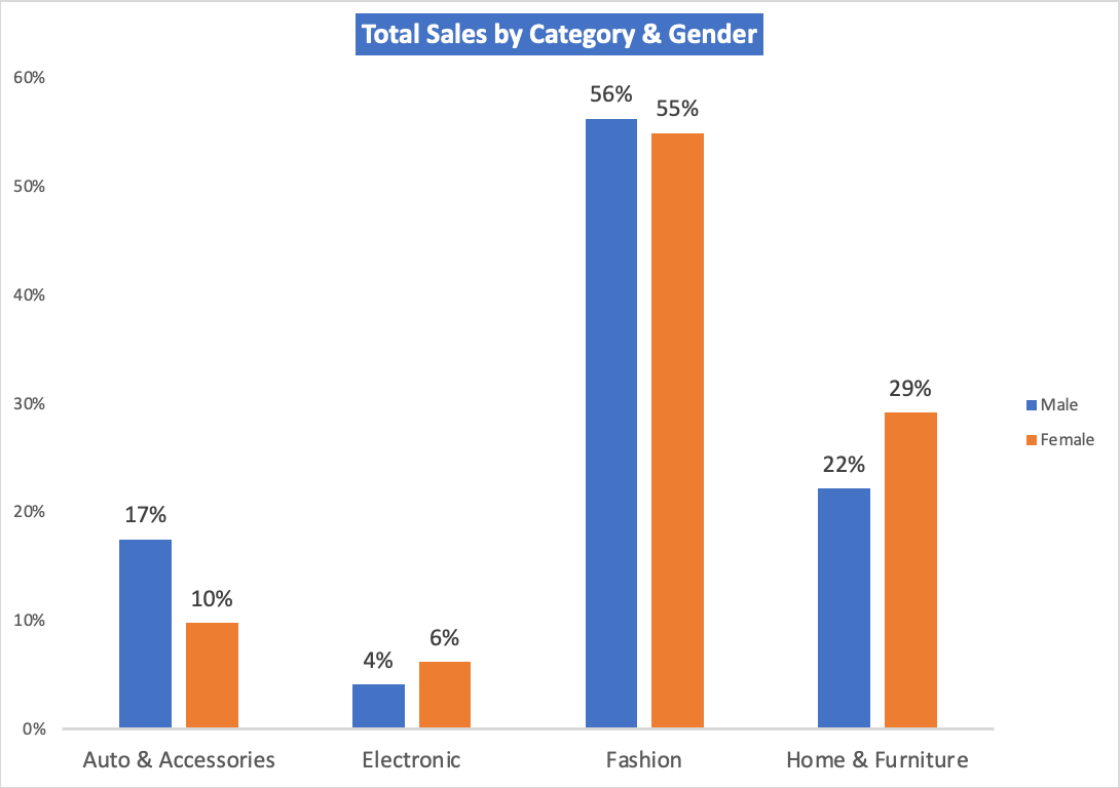
The hypothesis at the user level:
There are more high-quality users among male customers.

To investigate what the sale amount for males is significantly larger than females, I have split the sales amount into 3 ranges: Small order, Medium order, and Large order.
We have sales ranging from $33 to $250. Small orders are orders that range from $33 to $105. Medium orders are orders that range from $105 to $177. Large orders are orders that are larger than $177.

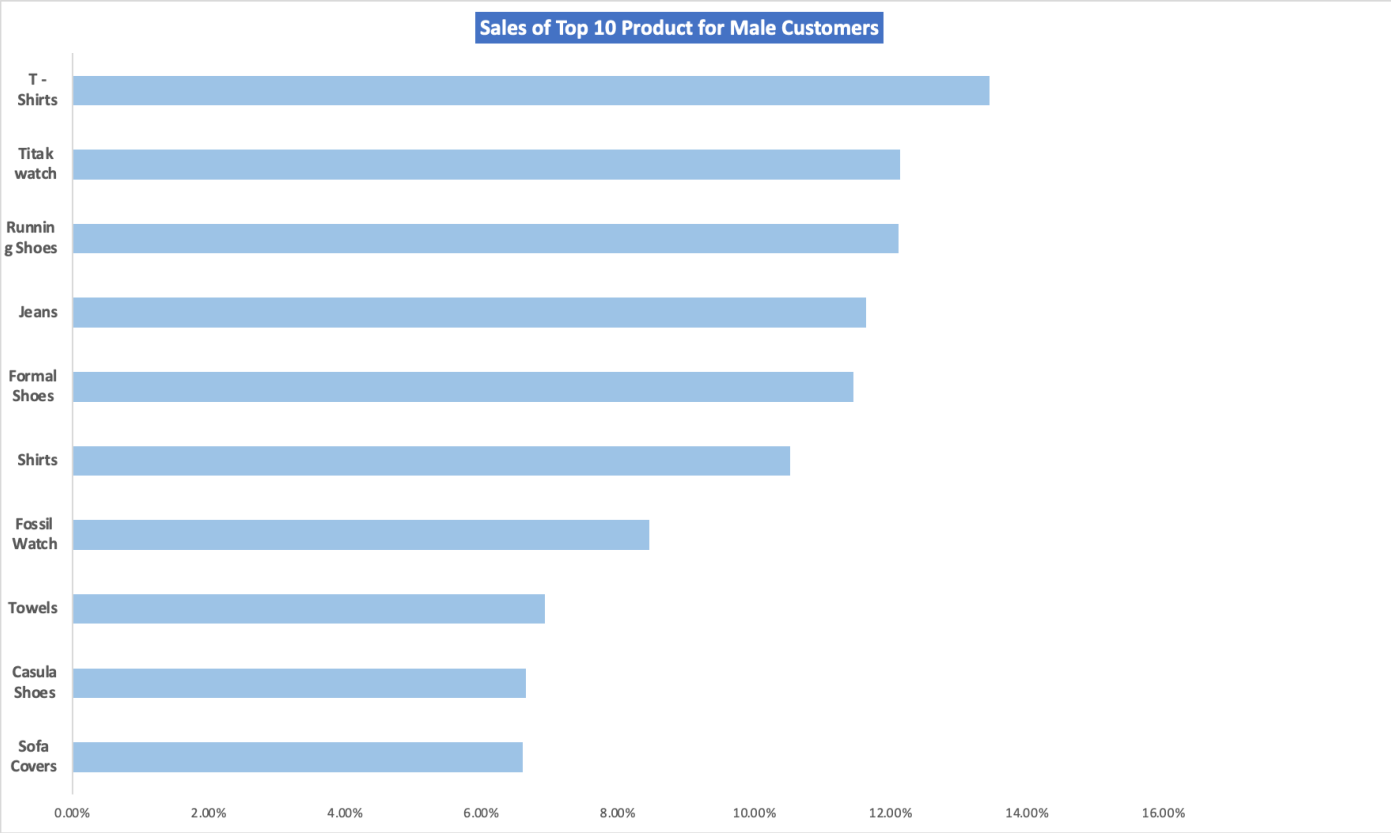


Overall, male customers made more sales than female customers. Based on the information the platform has, there are 55% large orders made from males and 45% made from females. The large order amount for males' orders is 10% greater than that of females' orders. Therefore, the result supports the hypothesis and we can conclude that there are more high-quality users among male customers.

Follow-up Question: What can be recommended to high-quality customers?



Summary: Fashion and Home & Furniture categories are popular among female and male customers. Male customers bought more Auto & Accessories and Fashion products than female customers and less Home & Furniture and Electronic products than female customers.



Suggestions:

1. The top 10 products sold to male customers are shown above. The most popular items, such as T-shirts, Titak watches, etc, are under Auto & Accessories and Fashion categories. The platform should bring in more Auto & Accessories and Fashion products to attract high-quality users (male customers given their stronger purchasing power than female customers).

2. Improve after-sale service: follow up high-quality users and recommend products under Auto & Accessories and Fashion to male customers and Home & Furniture and Electronic products to female customers.

3. Design questionnaires survey form at product and marketing dimension to find the reason why male customers have such strong purchasing power.

Question 2: How Does Discount Influence Sales? Which Discount should the Platform offer to Customers When Considering Product?

ANOVA Analysis on sales at each discount level:

Null Hypothesis: There is no difference in sales among different discount rates: μ**0.1** =μ**0.2**= μ**0.3** =μ**f** =μ**0.4** =μ**0.5**

Alternative Hypothesis: There is at least one discount type causes difference in sales amount

Anova: Single Factor

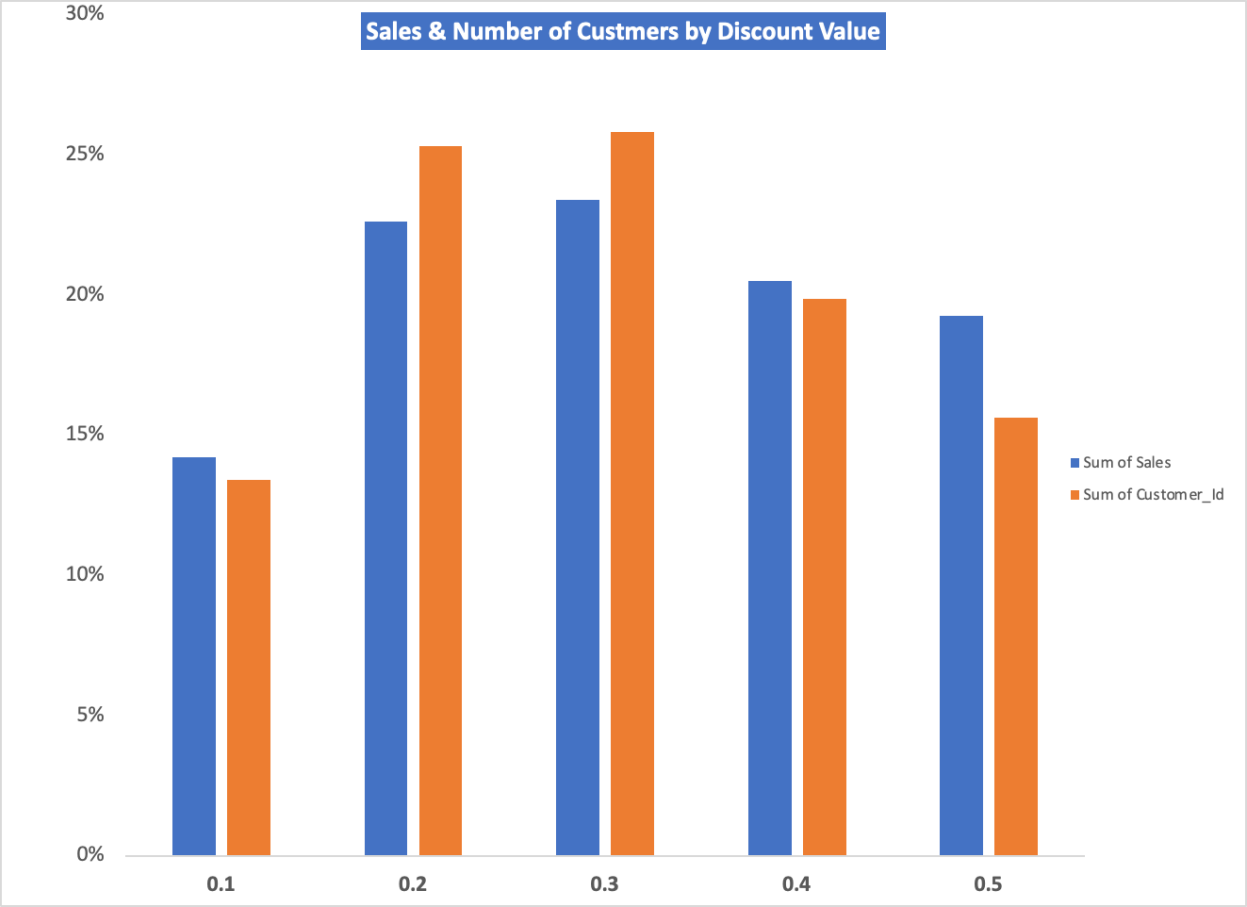
SUMMARY

Groups	*Count*	*Sum*	*Average*	*Variance*
0.1 discount	7207	1110978	154.152629	4330.10548
0.2 discount	12251	1766951	144.229124	4536.1084
0.3 discount	12401	1827261	147.347875	4502.27801
0.4 discount	10222	1603246	156.842692	4338.93524
0.5 discount	9204	1504537	163.465558	4054.28057

ANOVA

ource of Variat	*SS*	*df*	*MS*	*F*	*P-value*	*F crit*
Between Gr	2485172.24	4	621293.06	142.068027	5.374E-121	2.37210544
Within Grou	224258116	51280	4373.2082			
Total	226743289	51284				

As the result of ANOVA analysis showed, the p-value is 5.374E-121, which is extremely small and approximate to 0, so we reject the null and conclude that there is at least one discount type that causes difference in sales amount.



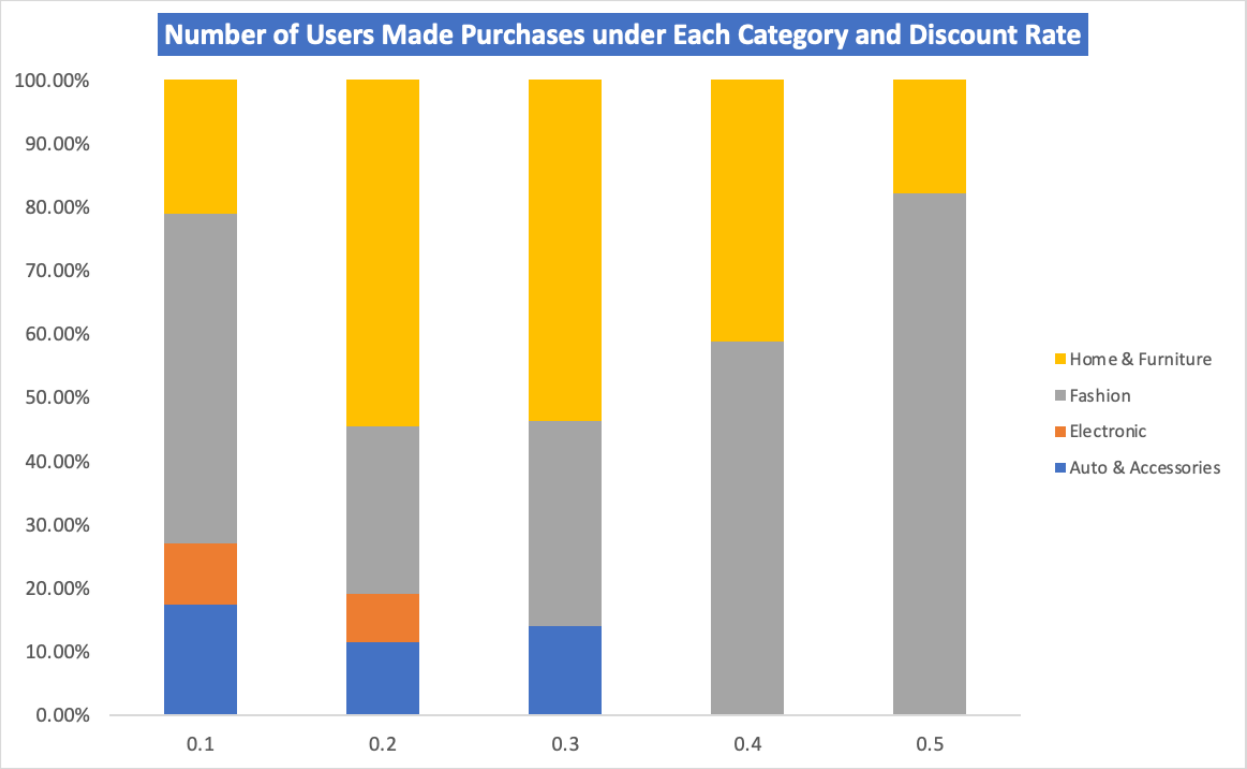
From the above plot, we can see that discount rates at 20% and 30% result in high purchases and number of customers. Half of the customers made purchases at a 20% and 30% discount rate.

Now we have two follow-up questions after having an overview of sales at each discount rate.

Follow-up Question 1: Why do the discount rates at 20% and 30% have a large user size?

a) The hypothesis at the product level:

Discount rates at 20% and 30% are associated with the popular products that customers love to buy.



The most frequently bought categories are Home & Furniture and Fashion according to the figure above. A reasonable inference can be made that It seem like different categories offer discount rates at different levels and Auto & Accessories and Electronic do not offer discounts above 30%, so there is no purchase made at 40% and 50% discount for Auto & Accessories and Electronic.

Let's take a closer look at whether there is a difference in the discount rate among different categories.

ANOVA Analysis on Discount Rate at category level:

Null Hypothesis: The discount rate is no difference among all categories:

$\mu_h = \mu_f = \mu_e = \mu_a$

Alternative Hypothesis: There is at least one category has different discount rate

Anova: Single Factor

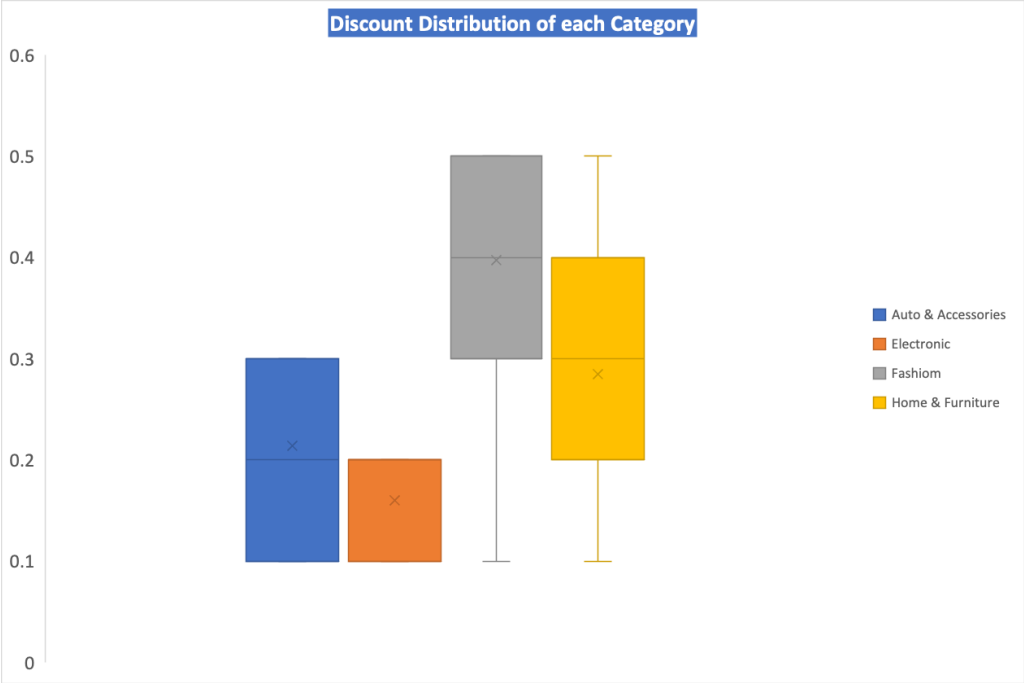
SUMMARY

Groups	*Count*	*Sum*	*Average*	*Variance*
0.1	7500	1606.9	0.21425333	0.00650304
0.1	2700	431.2	0.1597037	0.00240673
0.5	25645	9130.4	0.35603042	0.01837154
0.2	15437	4412.9	0.28586513	0.0105452
	0	0	#DIV/0!	#DIV/0!

ANOVA

ource of Variat	*SS*	*df*	*MS*	*F*	*P-value*	*F crit*
Between Gro	191.12701	4	47.7817526	3555.2164	0	2.37210545
Within Grou	689.157749	51277	0.0134399			
Total	880.28476	51281				

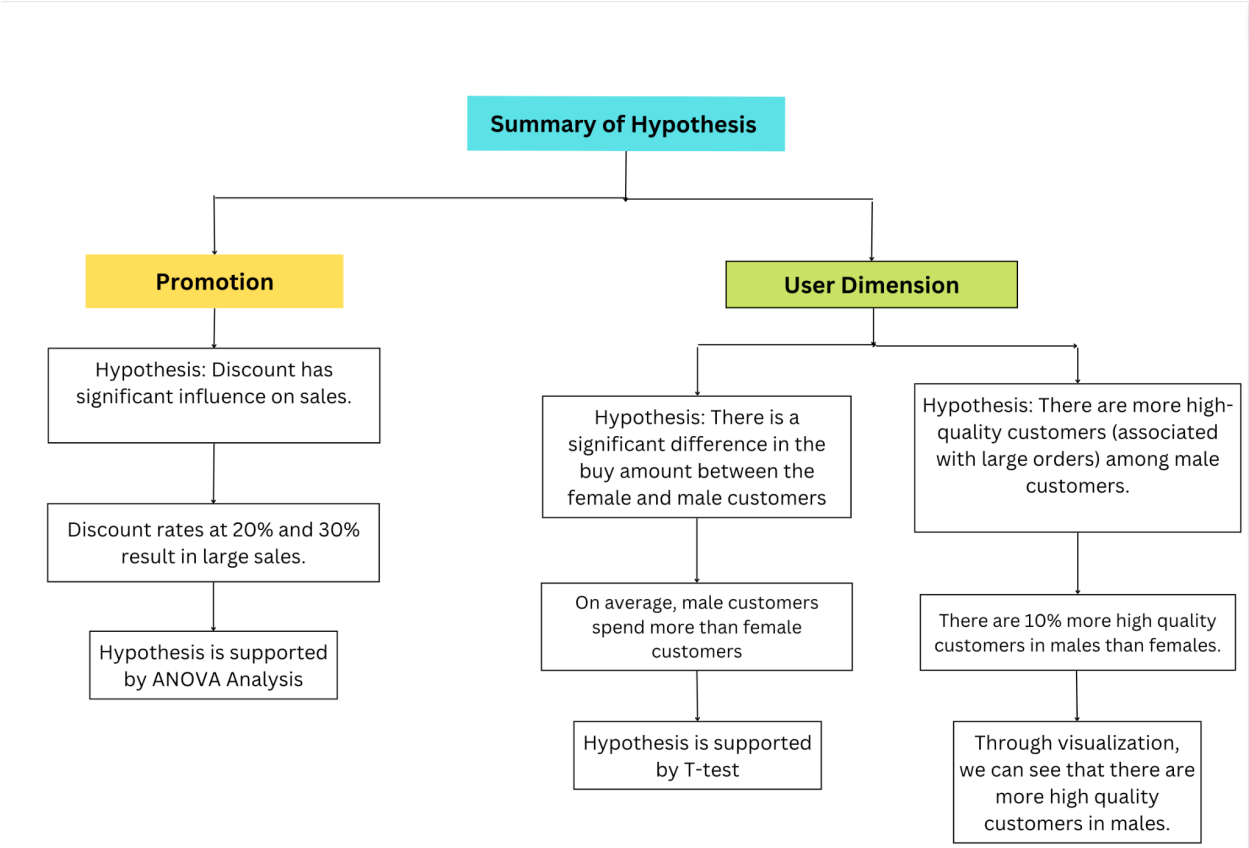
As the result of the ANOVA analysis shown above, the p-value is 0 and we can reject the null hypothesis and conclude that there is at least one discount type that causes a difference in sales amount.



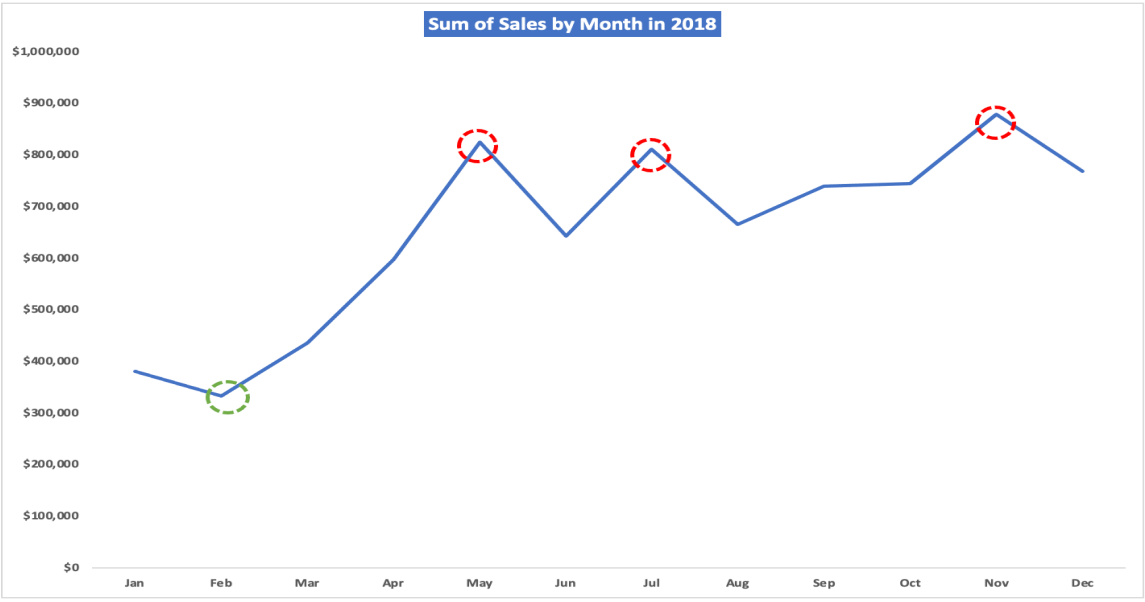
The box-plot of the distribution of discount rate by each category shows clearly that Home & Furniture and Fashion categories have higher discount rates compared with auto & Accessories and electronics. Therefore, a reasonable reference can be drawn: the discount rate plays a critical role in higher sales for Home & Furniture and Fashion categories.

Recommendation

Discount plays a critical role in promoting sales. Therefore, e-commerce can offer higher discount rate to customer if it consider to increase the sales of auto & Accessories and electronics, or introduce more products of Home & Furniture and Fashion to promote overall sales by providing more options to customers



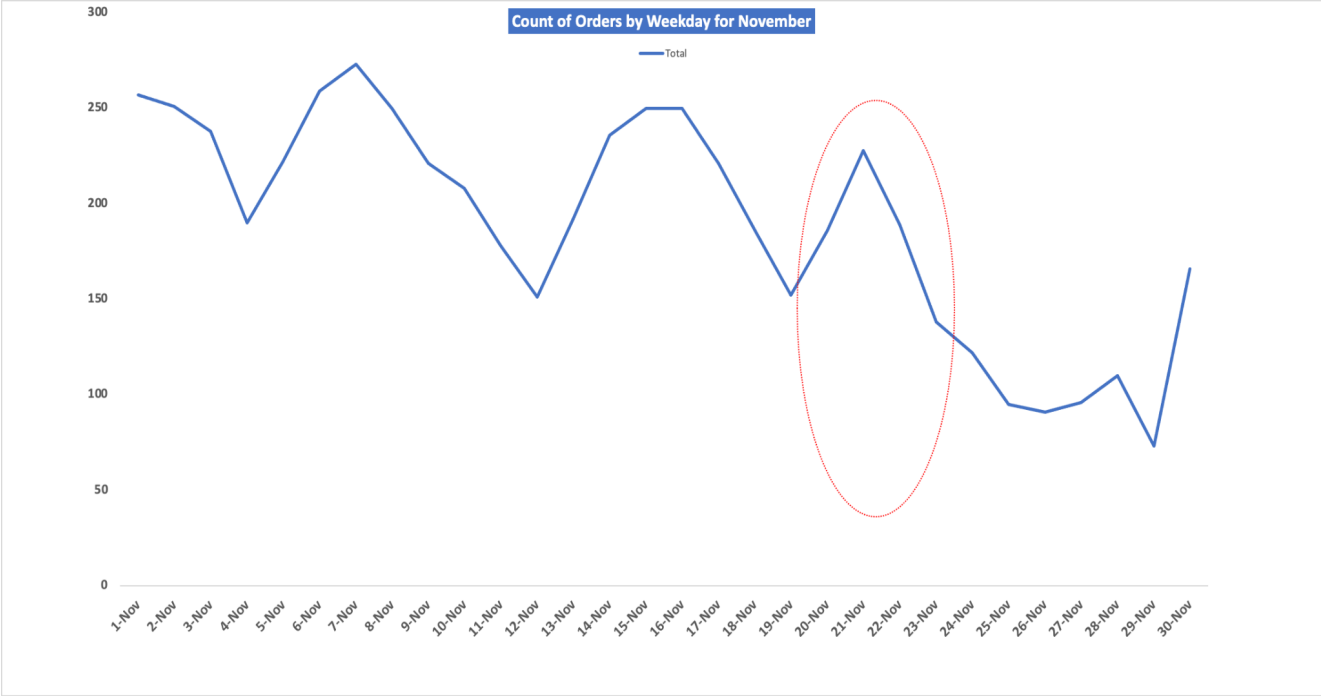
Seasonal Change and Sales



When looking at the trend plot above, we can see that there is an increase in sales in May, July, and August in 2016. Now we can start our exploration by asking several questions:

1. What causes the sales peak in May, July, and November?
2. What causes the business rough (lower sales) in February?
3. Which days of week have more customers shopping online?

Question 1: What causes the sales peak in May, July, and November?

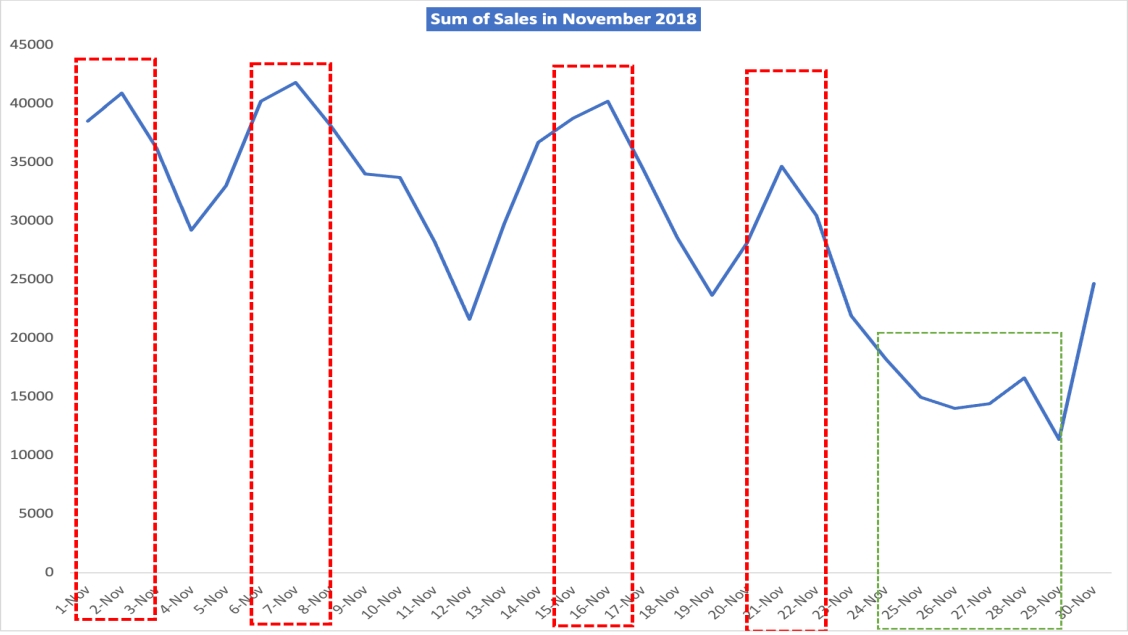


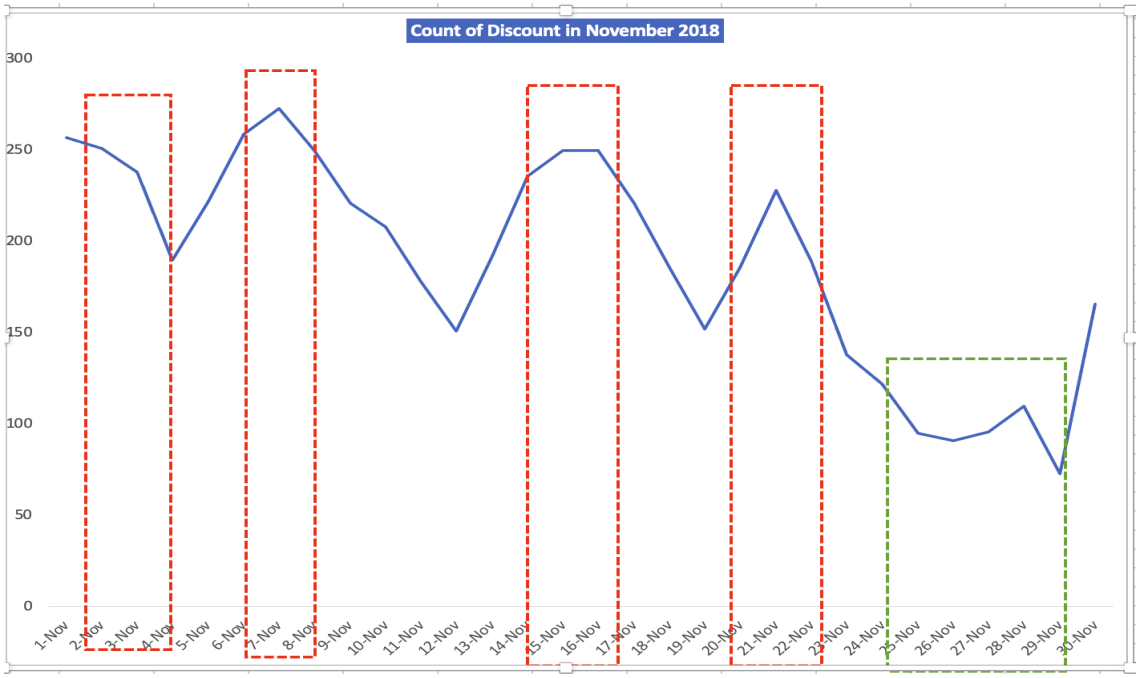
My initial expectation was that there would be an increasing trend of sales in November as November is normally a major promotion season. Looking at the plot above showing the count of orders in November, there are peaks

and troughs going on instead of an overall increasing trend. This brings us to the next question:

Follow-up Question: Why is there not an increasing trend shown in November?

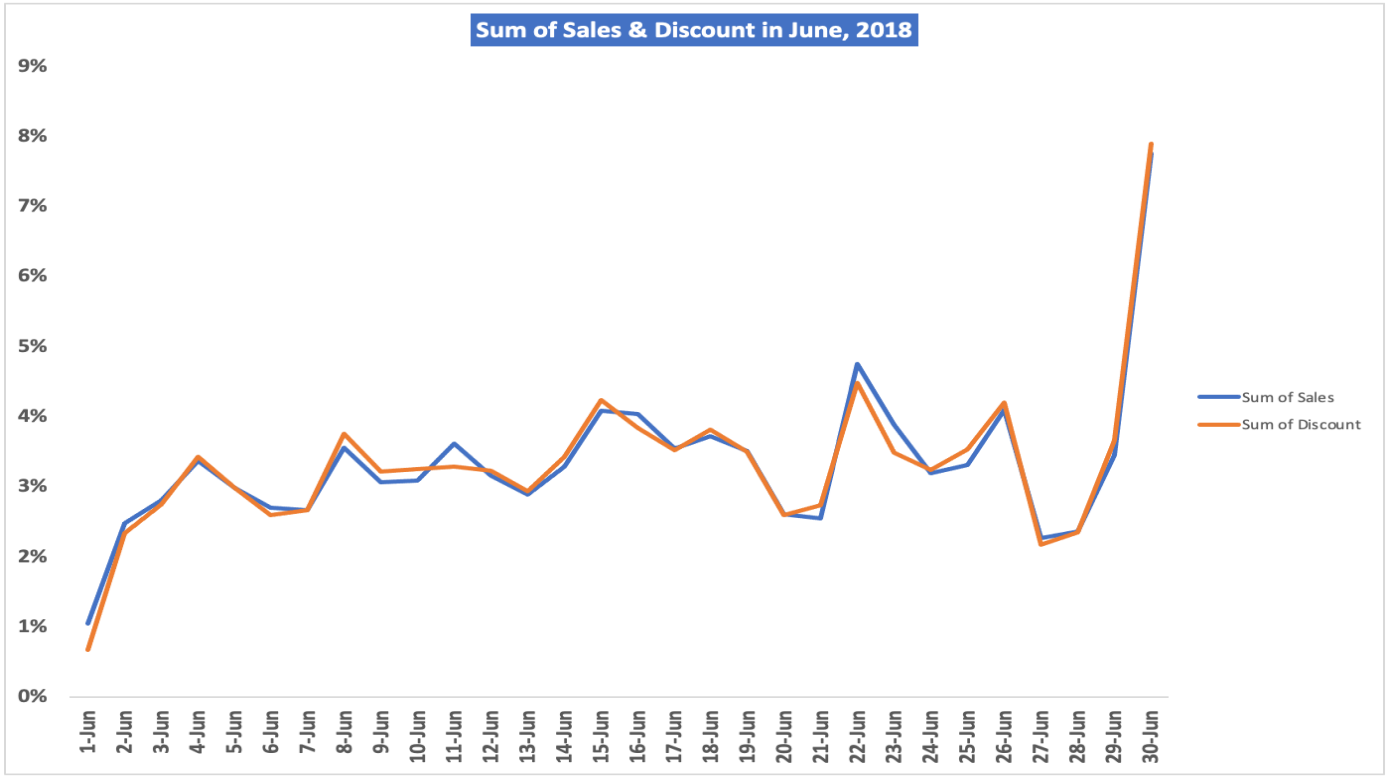
***Hypothesis**: Discount does not offer consistently in November, which causes the peaks and troughs in sales.*

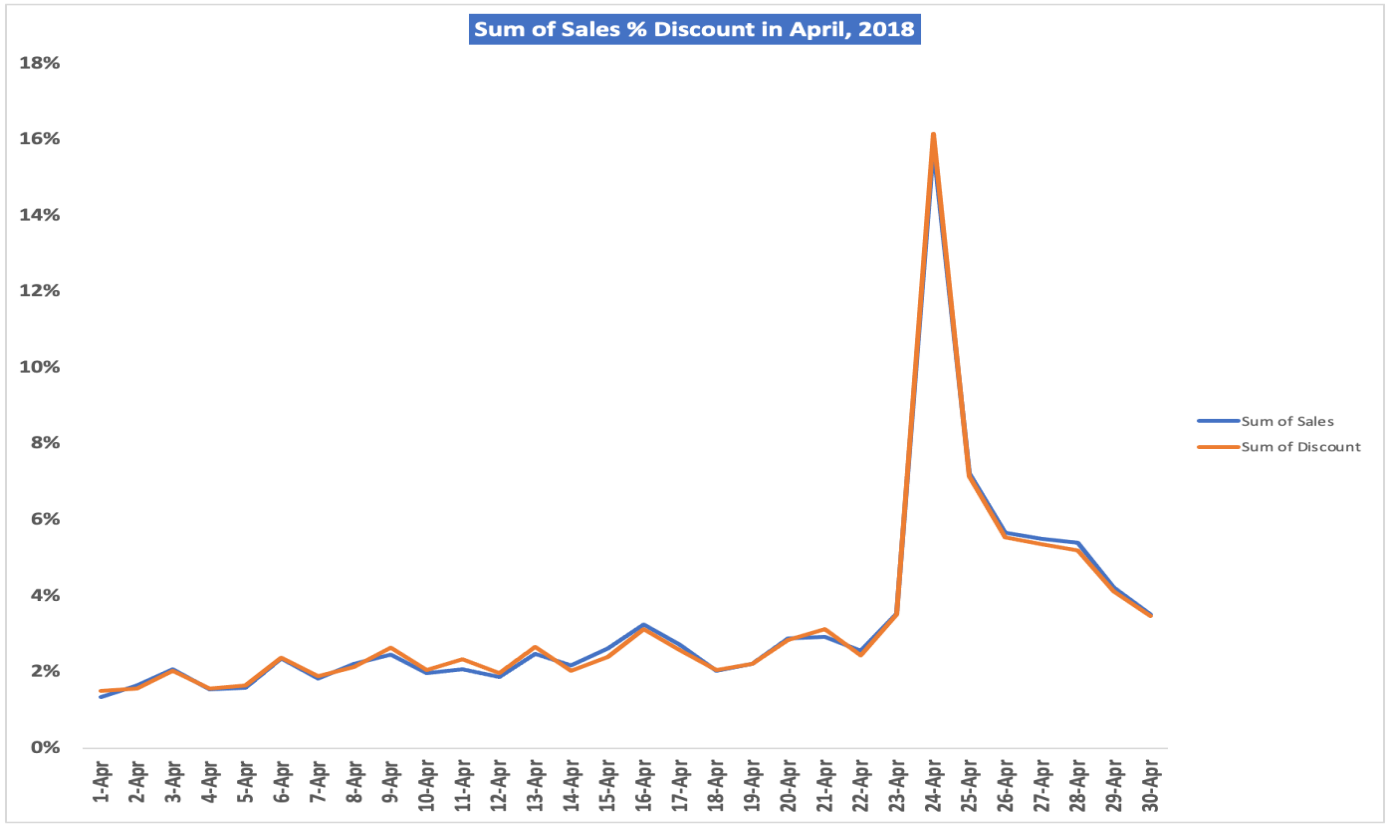




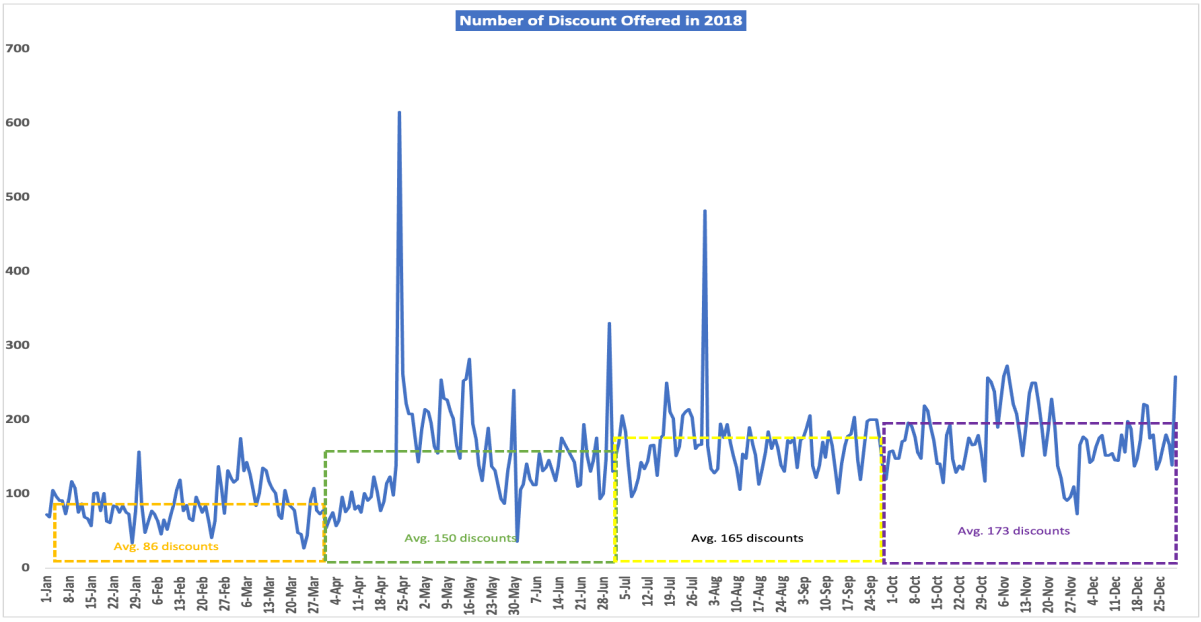
The sum of Sales amount matches the number of discounts offered in November. Sales increase whenever there are more discounts offered and vice versa. Also, there is evidence supporting the hypothesis that discount does not offer consistently over the month, which causes the peaks and troughs in sales in November.

Another interesting finding is that, there is no peak at ‘Black Friday’ on 25, November. One possible assumption is that there are major promotions in physical stores on Black Friday. Our studied object is an e-commerce platform and its major promotions tend to cover before and after the big sales offered by physical stores as customers tend to go out to shop during the period. We can see that there is an increasing trend in discounts starting on 29, November.


Sum of Sales & Discount in June, 2018
9%
8%
7%
6%
5%
4%
3%
2%
1%
0%
1-Jun
2-Jun
3-Jun
4-Jun
5-Jun
6-Jun
7-Jun
8-Jun
9-Jun
10-Jun
11-Jun
12-Jun
13-Jun
14-Jun
15-Jun
16-Jun
17-Jun
18-Jun
19-Jun
20-Jun
21-Jun
22-Jun
23-Jun
24-Jun
25-Jun
26-Jun
27-Jun
28-Jun
29-Jun
30-Jun
Sum of Sales
Sum of Discount

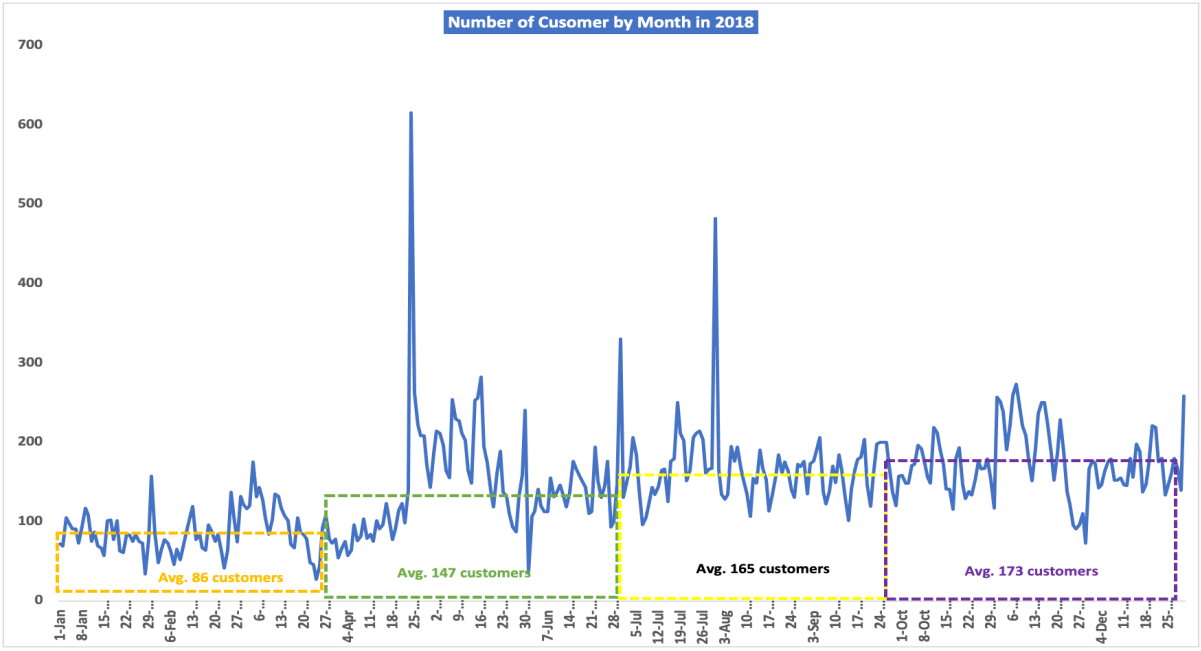

Sum of Sales % Discount in April, 2018
18%
16%
14%
12%
10%
8%
6%
4%
2%
0%
1-Apr
2-Apr
3-Apr
4-Apr
5-Apr
6-Apr
7-Apr
8-Apr
9-Apr
10-Apr
11-Apr
12-Apr
13-Apr
14-Apr
15-Apr
16-Apr
17-Apr
18-Apr
19-Apr
20-Apr
21-Apr
22-Apr
23-Apr
24-Apr
25-Apr
26-Apr
27-Apr
28-Apr
29-Apr
30-Apr
Sum of Sales
Sum of Discount

As the above plots show, sales amount in July and April show the same pattern that sales increase whenever there are more discounts offered and vice versa.



Question 2: What causes the business trough (lower sales) in February?

Hypothesis: *There are less customers making purchases after the major promotion period in November*



Summary: Looking at the count of the number of customers above, we can see an obvious increasing trend in the average number of customers after aggregating the data by quarter. There are on average 86 customers making purchases during the first quarter in 2018 and the number of customers increases to 173 customers in the fourth quarter in 2018.

We only have purchase information in 2018 so a conclusion would be made on a reasonable assumption that there are less customers making purchases after the major promotion period in the fourth quarter of the year. Also, there is an increasing trend in the average number of discounts offered during the year with the first quarter having the least average number of discounts and the fourth quarter having the largest average number of discounts, which brings us to the second assumption that less discount offered in the first quarter causes the business rough in February. We would need more data, such as data from 2019 to further validate our assumptions.

Suggestions:

1. There’s a lot of orders in May, July, and November and as the major promotion events present. It is recommended to adjust the stock in February to avoid too many overstocked products.
2. Prepare enough products in May, July, November according to the sales in last year, especially for products under the categories of Fashion and Home & Furniture.