Superstore Data Analysis

Financial Statistics

HOSSAIN MUNTASIR 5/22/23 FIN10002

Executive Summary

This report presents an analysis of a supermarket dataset covering a four-year period. The analysis provides valuable insights into customers' purchasing behavior, shipping preferences, sales frequency, consumer demographics, regional sales, product performance, sales item quantity, impact of discounting on sales, and profitability of different product categories. Key findings include the observation of distinct sales patterns across different months, with September standing out as a month of high sales. The majority of customers preferred the standard class shipping mode, and office supplies had the highest sales contribution. Additionally, customers tended to purchase smaller quantities more frequently, and discounts of 20.0% had the most significant impact on sales. The regression analysis revealed a positive relationship between sales and profit, with sales explaining approximately 44.65% of the variation in profit. These insights can guide decision-making processes to enhance sales and profitability for the supermarket.

Table of Contents

Introduction

This report analyzes the Supermarket data from January 2014 to December 2017, focusing on 200 randomly selected transactions. The report utilizes various techniques and visualizations to gain insights into customers' purchasing behavior, shipping preferences, sales frequency, consumer demographics, regional sales, product performance, sales item quantity, impact of discounting on sales, and profitability of different product categories. Additionally, the report includes confidence intervals and hypothesis testing for customer ordering behavior and profitability across product categories. Finally, a regression analysis is conducted to investigate the relationship between Sales and Profit, including scatterplots, a linear regression model, coefficients of correlation and determination, and a hypothesis test.

Analysis

Based on the Supermarket data spanning from January 2014 to December 2017, which includes 9,994 transactions and nine columns of variables (Order date, Ship Mode, Segment, Region, Category, Sales, Quantity, Discount, Profit), I utilized various techniques and visualization approaches to analyze the data for our shop. Through the use of charts, graphs, and other visual tools, I successfully obtained significant insights. In the following sections, I will present and discuss these insights.

Sample Selection

To simplify the analysis, I randomly selected 200 data points from the original dataset of 9,994 tuples, which will serve as a representative sample for the entire dataset.

Descriptive Statistics

Customers’ Purchasing Behaviour on Different Months

Analyzing the sales data throughout the year revealed distinct peaks and valleys, indicating fluctuations in customer purchasing behavior.

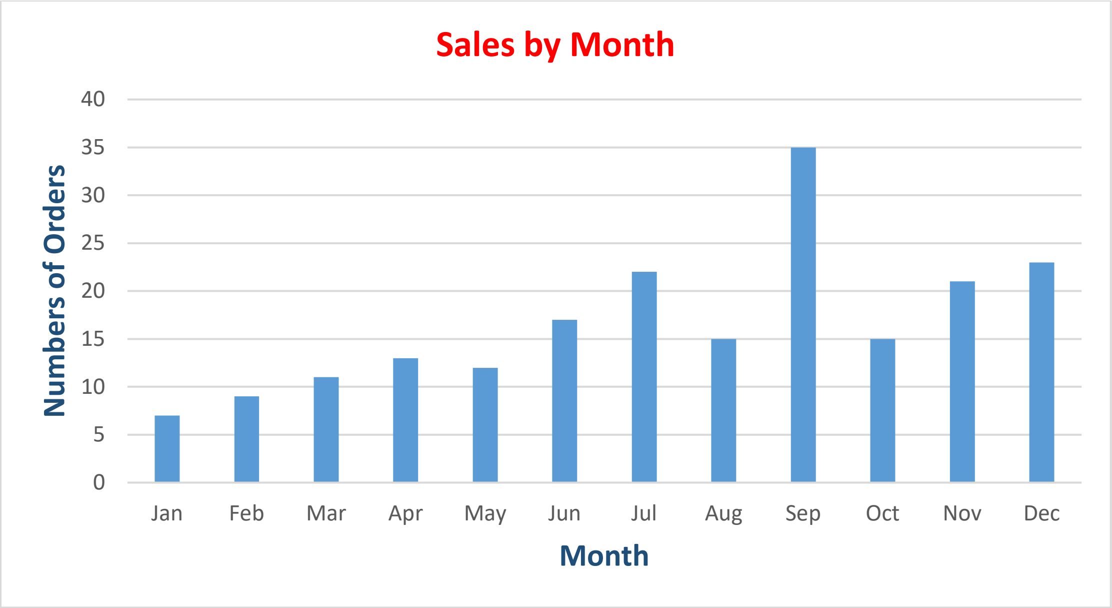


Figure 1 Sales by Month

Analyzing customers' purchasing behavior across different months revealed interesting patterns and trends. For instance, September consistently stood out as a month with high sales, indicating that customers tend to make more purchases during that time. This could be attributed to factors such as back-to-school shopping or seasonal promotions. On the other hand, January showed the lowest sales, suggesting a decline in customer spending, possibly due to the post-holiday period or financial constraints after the festive season.

Customers' Shipping Mode Preferences

I analyzed the "Orders shipping preferences" variable to showcase the percentage distribution and highlight the most popular shipping modes among customers.

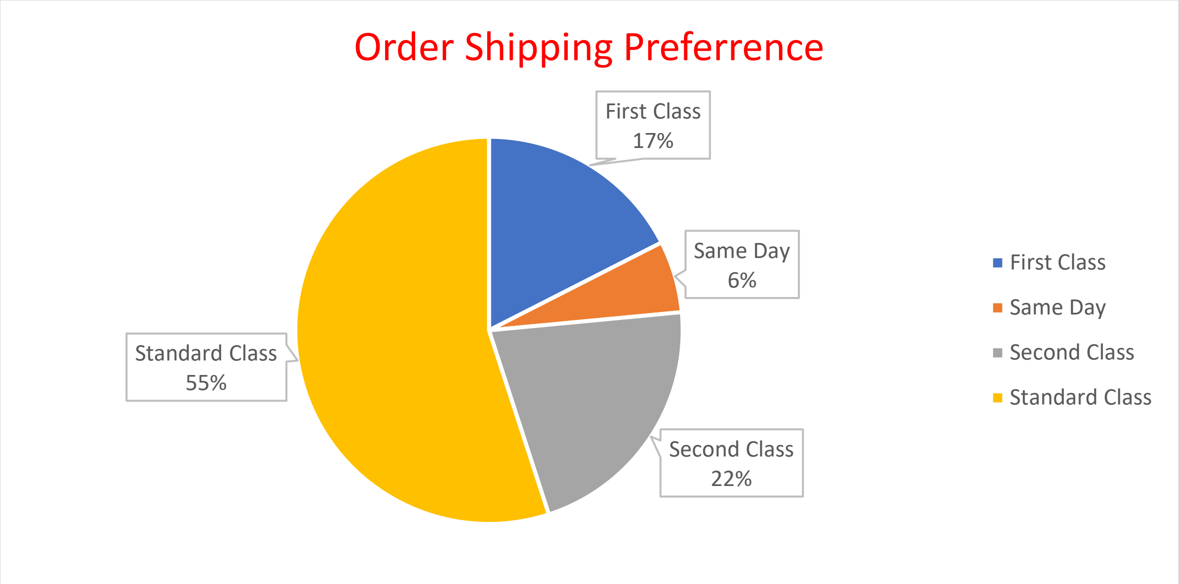


Figure 2 Customer shipping preferences

The analysis indicates that the majority of customers, accounting for 55.00% of preferences, opted for the standard class shipping mode. This suggests that customers value cost-effective and reliable shipping options for their orders. Second class was the next popular choice, with a preference rate of 21.50%, followed by first class at 17.50%. The same-day shipping option had the lowest preference, accounting for 6.00% of orders.

Sales Frequency Analysis by Amount

In this section of the report, we present an analysis of sales frequency based on different sale amounts. The objective was to understand the distribution of sales across various price ranges and identify patterns in customer purchasing behavior.

Sales

Bins	Freq
$ 50.00	97
$ 100.00	31
$ 150.00	12
$ 200.00	7
$ 250.00	9
$ 300.00	5
$ 350.00	3
$ 400.00	8
$ 450.00	2
$ 500.00	1
$ 4,500.00	25

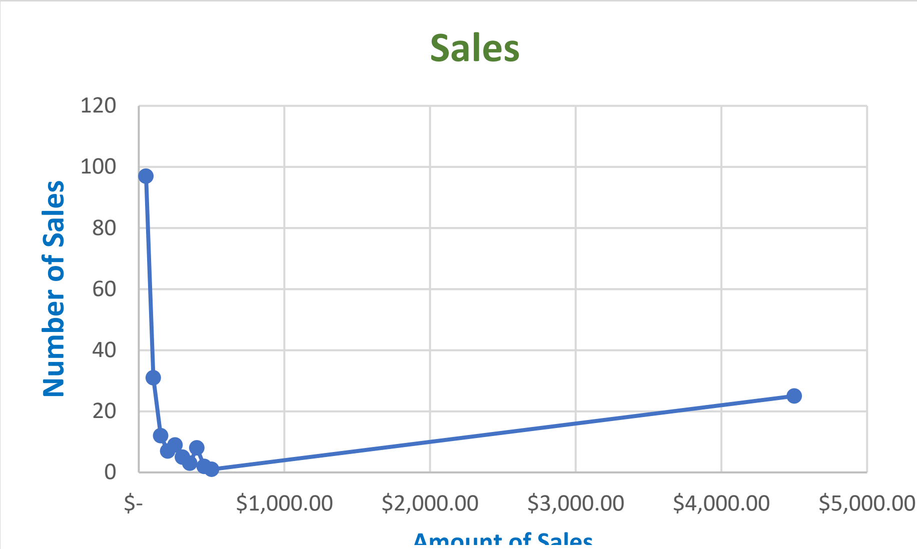


Figure 3 Sales Frequency

The sales data analysis reveals that the majority of sales (97 out of 200) fall within the range below $50. As the sales amount increases, the frequency gradually decreases. The mean sales amount of $219.73 indicates the average purchase amount in sales. The median sales amount of $54.78 suggests that half of the sales amounts are below $54.78 and half are above it.

Consumer Demographic

In this section, I analyzed the consumer type section to determine the percentage of customers belonging to different demographic categories.

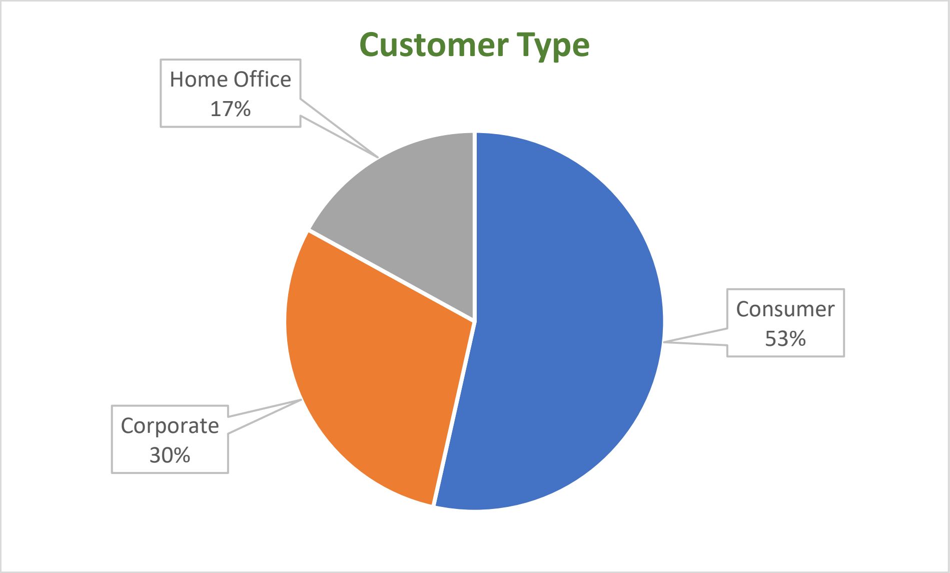


Figure 4 Consumer Demographic

The analysis indicates that the majority of customers, representing 53% of the total, belong to the consumer category. The corporate segment accounts for 30% of customers, while the home office category comprises 17% of the customer base.

Sales by Region

The analysis of sales by region revealed the following sales figures:

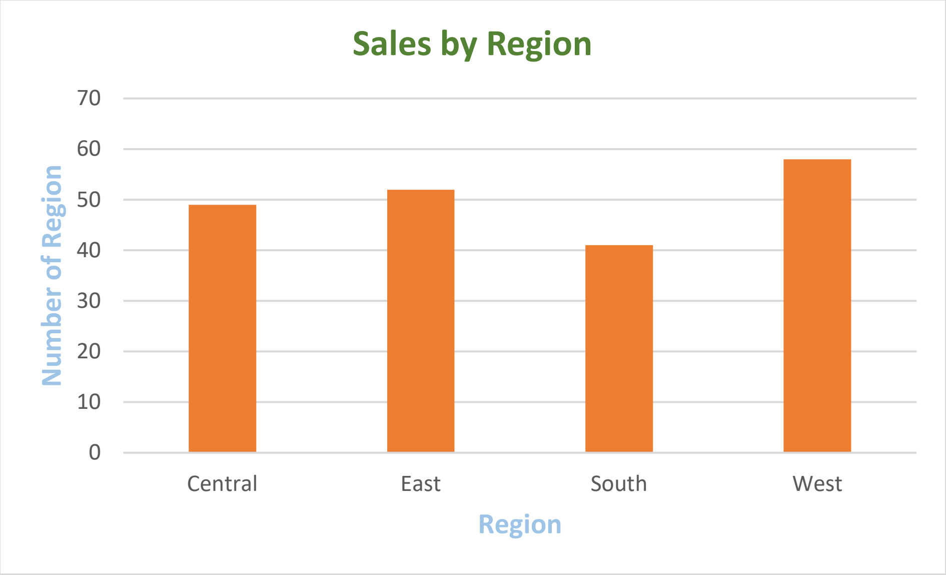


Figure 5 Sales by Region

These numbers indicate the total sales volume for each respective region. The West region has the highest sales figure with 58, followed by the East region with 52. The Central region and the South region have sales figures of 49 and 41, respectively.

Performance of Different Product Type

In this section, I present an analysis of sales by product category to gain insights into the distribution of sales across different categories and understand their respective contributions to overall sales.

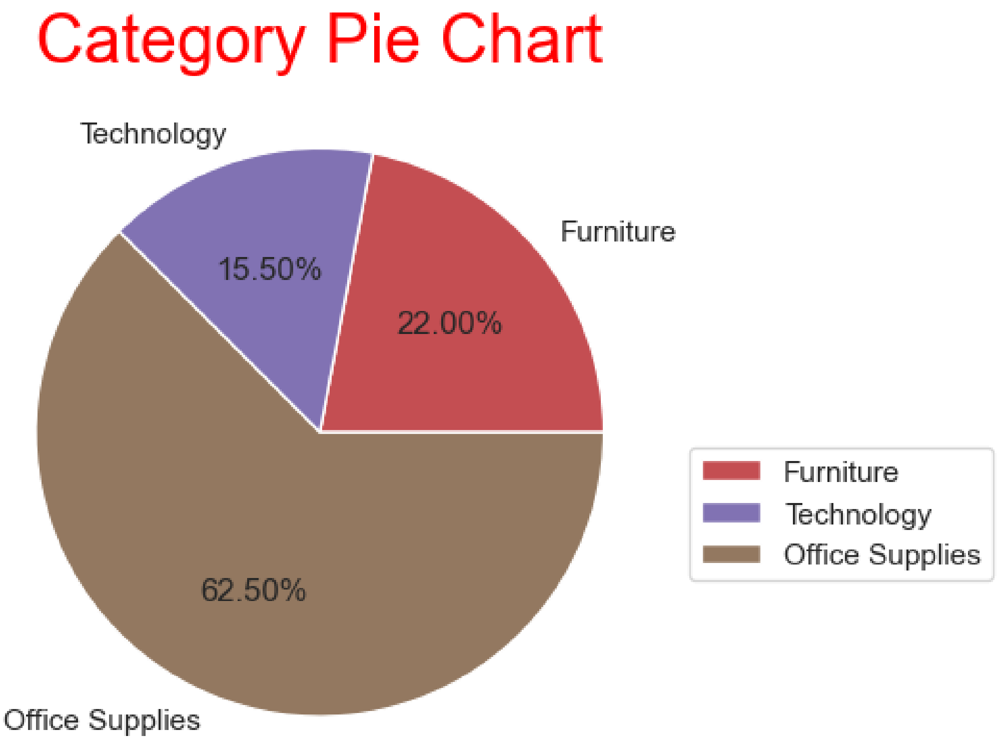


Figure 6 Performance of different product type

The data indicates that office supplies have the highest sales contribution, accounting for 62.50% of total sales. Furniture follows with a significant share of 22% of total sales, while technology products have a relatively lower contribution at 15.50% of total sales.

Analysis of Sales Item Quantity

In this section, we analyze the quantity of sales items to understand the frequency of different quantities sold.

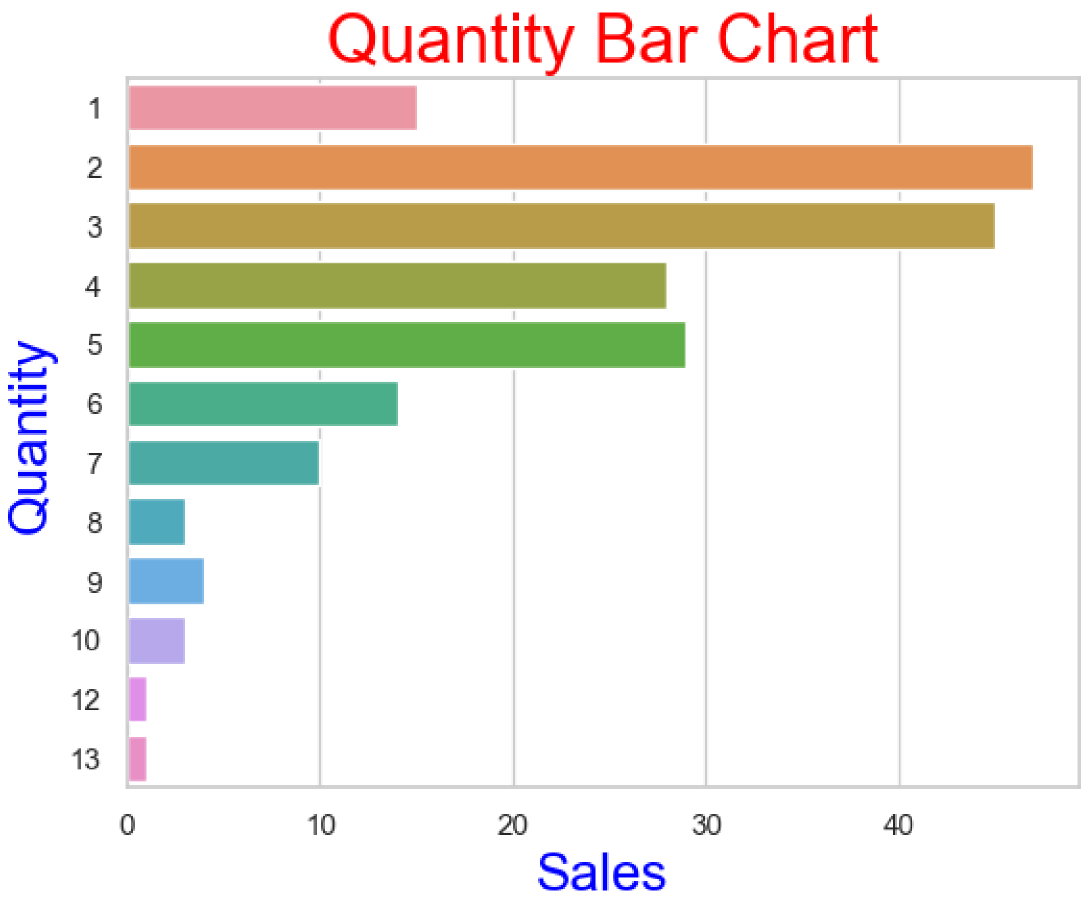


Figure 7 Sales by Quantity

Analyzing sales item quantities provides insights into customer purchasing behavior and preferences. The findings suggest that customers tend to purchase smaller quantities more frequently, with quantities of 2 and 3 being the most popular. As the quantity increases beyond 5, the frequency gradually decreases, indicating that larger quantities are less common.

Impact of discounting on sales

In this section, we analyze the distribution of discounts applied to sales transactions.

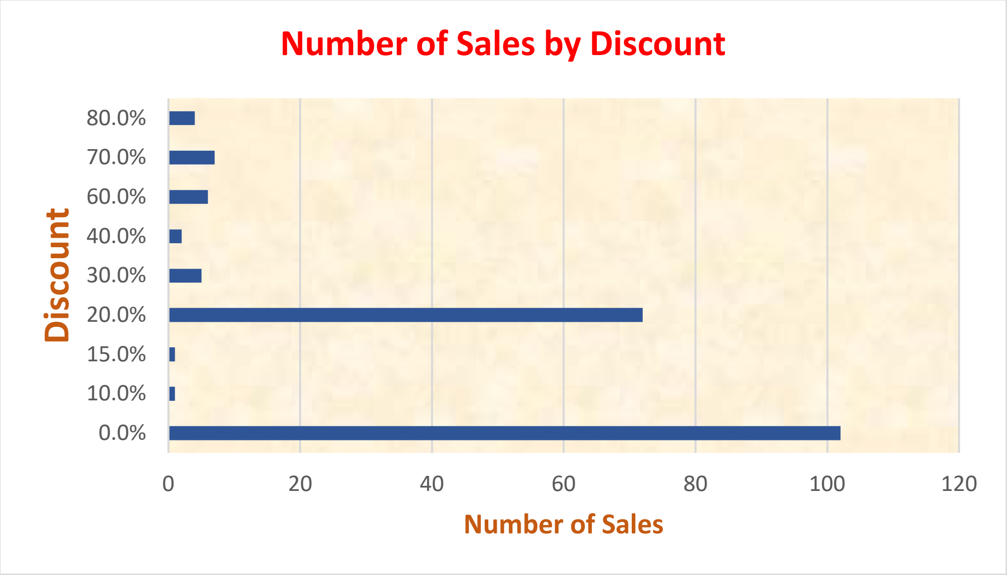


Figure 8 Number of Sales by discount

The analysis of discount distribution in sales transactions reveals that the majority of transactions (102 out of 200) did not have any discounts applied. Among the transactions with discounts, discounts of 20.0% were the most common, occurring 72 times. Lower discount percentages, such as 10.0% and 15.0%, were less prevalent, each occurring only once. Higher discount percentages, ranging from 60.0% to 80.0%, had moderate occurrences ranging from 4 to 7. These findings highlight the significance of discounts in influencing customer purchasing decisions, with 20.0% discounts being more impactful.

Profitability of Different Product Categories

In this section, we analyze the profitability of different product categories.

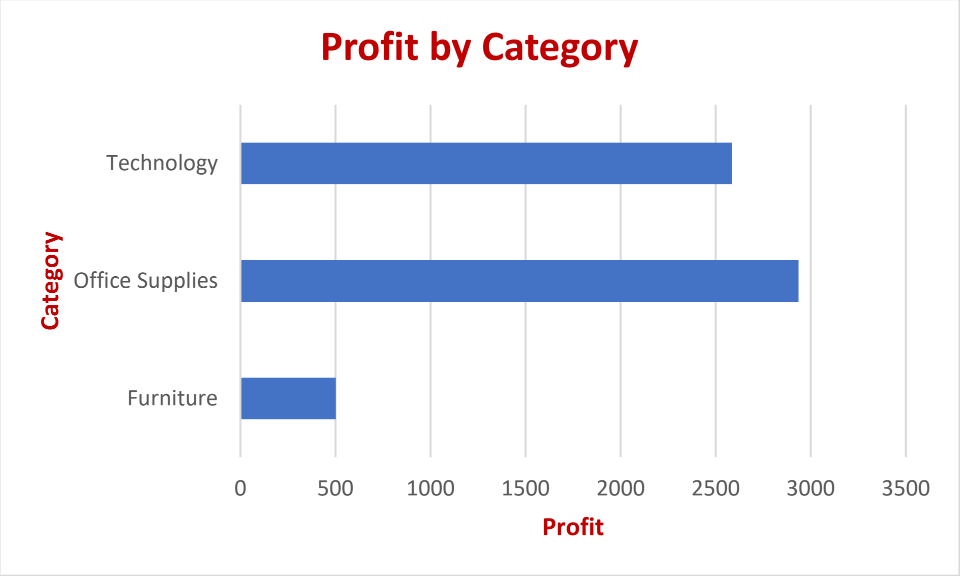
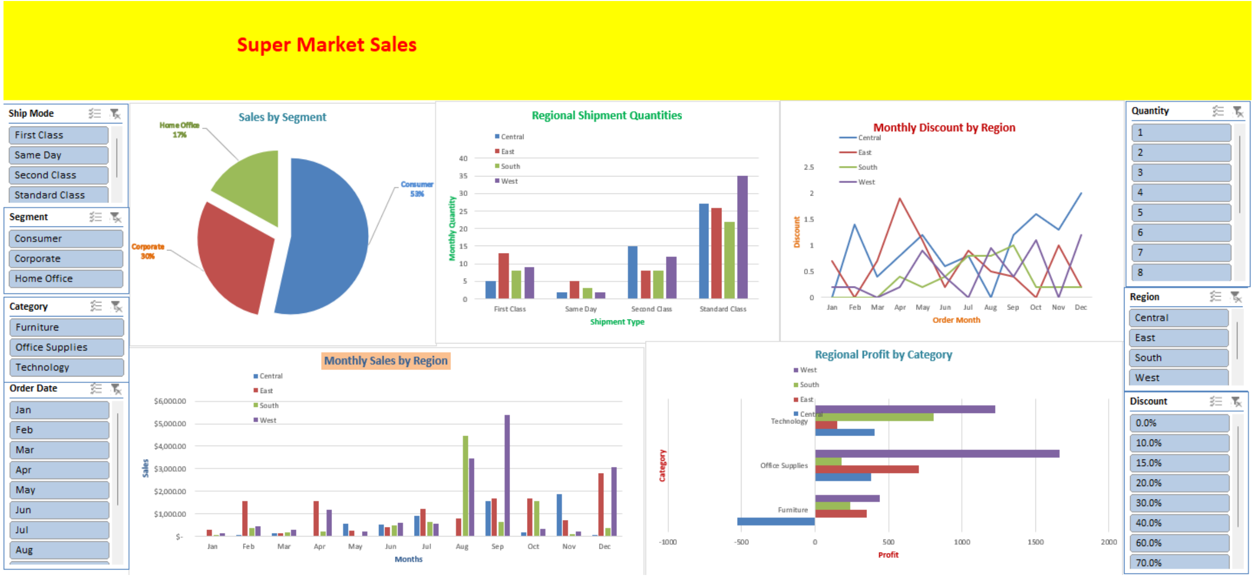


Figure 9 Profit by Category

Office supplies generated the highest profit of $2,934.77, followed by technology with $2,585.37. Furniture had a lower profit of $501.50. This highlights the varying levels of profitability among the categories. Businesses can focus on high-profit categories and explore opportunities to improve lower-performing categories for overall profitability and growth.

Dashboard



Confidence Interval

We are 95% confident that the average sales for the Consumer sector falls within the interval of $131.36 to $341.12. This means that if we were to repeat the sampling process multiple times and construct confidence intervals, approximately 95% of those intervals would contain the true population mean of consumer sales.

In comparing the confidence interval with the true population mean of $223.73, we can see that the interval of $131.36 to $341.12 encompasses the true mean. This indicates that our sample accurately estimates the average sales for the Consumer sector.

See Appendix D for calculation.

Hypothesis Testing

Hypothesis Test 1: Investigating customer ordering behavior:

Customers in the Corporate segment are often believed to place larger orders compared to their Home Office counterparts. To test this contention, carry out an appropriate hypothesis test to determine if there is a statistically significant difference in the average quantity of orders between these two customer segments.

Hypothesis Test 2: Examining profitability across product categories:

It is often felt that the average profit per transaction would be different between the Furniture and Technology categories. Test if there is a difference in the average total profit for these two categories using an appropriate hypothesis test.

Hypothesis Testing 1

Based on the results of the hypothesis test 1, we cannot reject the null hypothesis (Ho: µHome - µCorp ≤ 0) at a significance level of 0.05. The p-value obtained from the test is 0.212, which is greater than the chosen significance level. This indicates that there is not enough evidence to support the claim that the average quantity ordered for corporate customers is greater than the average quantity ordered for home office customers.

Therefore, we do not have sufficient statistical evidence to conclude that customers in the Corporate segment are more likely to place larger orders than their home office counterparts. Further analysis or additional data may be necessary to make a definitive determination in this regard.

See Appendix D for calculation.

Hypothesis Testing 2

Based on the results of the hypothesis test 2, we can reject the null hypothesis (Ho: µFurn - µTech = 0) at a significance level of 0.05. The p-value obtained from the test is 0.042, which is less than the chosen significance level. This suggests that there is evidence to support the claim that there is a difference in the average total profit between the furniture and technology categories.

Therefore, based on the available data, we can conclude that the average profit per transaction is indeed different between the furniture and technology categories. Further analysis can be conducted to investigate the specific nature and magnitude of this difference.

See Appendix D for calculation.

Correlation and Regression

In this section, we will investigate the relationship between Sales and Profit and develop a regression model to predict Profit from Sales. We will conduct a full regression analysis and discuss the results.

Scatterplot: We will create a scatterplot of Sales and Profit to visualize the relationship between the two variables. The scatterplot will also include a line of best fit, which represents the regression model.

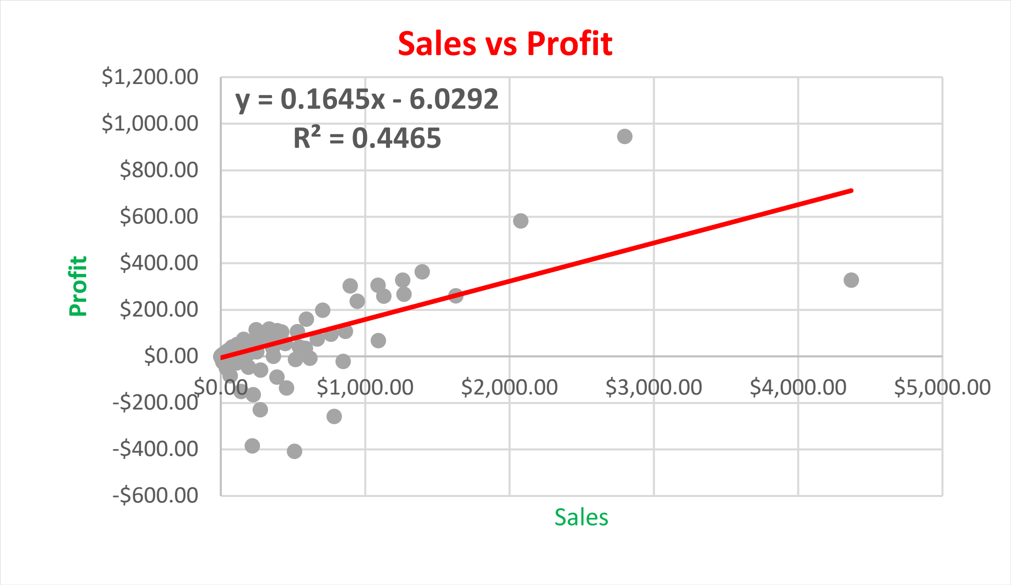


Figure 5 Scatterplot

Linear regression model: The linear regression model for predicting Profit from Sales is as follows:

Profit = 0.164463 * Sales - 6.02922

The coefficient of Sales (0.164463) indicates that for every unit increase in Sales, we can expect an increase of 0.164463 units in Profit.

Coefficients of correlation and determination: The coefficient of correlation (R) is 0.668186, indicating a moderate positive correlation between Sales and Profit. The coefficient of determination (R Square) is 0.446473, which means that 44.65% of the variation in Profit can be explained by Sales.

Hypothesis test: We conducted a hypothesis test to determine if there is a linear relationship between Sales and Profit. The null hypothesis (Ho) is that the coefficient of Sales (β1) is equal to zero, indicating no linear relationship. The alternative hypothesis (Ha) is that β1 is not equal to zero, suggesting a linear relationship. The test resulted in a p-value of 3.14E-27, which is less than the significance level of 0.05. Therefore, we reject the null hypothesis and conclude that there is evidence of a linear relationship between Sales and Profit.

Overall, the regression analysis indicates that Sales have a significant impact on Profit. The positive coefficient suggests that an increase in Sales is associated with an increase in Profit. The coefficient of determination indicates that Sales explain approximately 44.65% of the variation in Profit. These findings provide insights into the relationship between Sales and Profit and can guide decision-making in maximizing profitability.

Conclusion

In conclusion, the analysis of the Supermarket data provided valuable insights into customer purchasing behavior, shipping preferences, profitability of different product categories, and the relationship between sales and profit. The sample of 200 randomly selected transactions from January 2014 to December 2017 revealed seasonal fluctuations and a preference for standard class shipping. The population mean of sales was estimated with a confidence interval, and hypotheses testing showed no significant difference in the average quantity ordered between corporate and home office customers, but a difference in profit between furniture and technology categories. The regression analysis demonstrated a linear relationship between sales and profit, with sales explaining 44.65% of profit variation. Despite these findings, it is important to consider limitations such as the small sample size, specific time period, and limited variables analyzed, which warrant further analysis for a comprehensive understanding of the supermarket's operations and profitability.

Appendixes

Appendix-A (Random Sample)

Order Date	Ship Mode	Segment	Region	Category	Sales	Quantity	Discount	Profit
30/08/2016	First Class	Consumer	East	Furniture	$786.74	4	30.0%	-$258.50
27/06/2014	Standard Class	Consumer	South	Technology	$223.96	4	0.0%	$53.75
12/04/2016	Standard Class	Corporate	South	Office Supplies	$129.55	3	20.0%	-$22.67
28/06/2016	Second Class	Consumer	Central	Technology	$359.98	3	20.0%	$36.00
6/06/2014	Standard Class	Consumer	South	Office Supplies	$62.02	2	20.0%	$22.48
13/04/2017	First Class	Corporate	West	Office Supplies	$895.92	5	20.0%	$302.37
17/06/2016	First Class	Consumer	East	Office Supplies	$32.40	5	0.0%	$15.55
4/09/2017	Standard Class	Consumer	West	Office Supplies	$217.85	5	0.0%	$65.36
1/07/2016	Same Day	Consumer	South	Office Supplies	$12.96	2	0.0%	$6.22
5/04/2015	Same Day	Consumer	West	Office Supplies	$23.84	8	0.0%	$6.44
29/10/2016	Standard Class	Corporate	East	Office Supplies	$11.67	3	0.0%	$3.03
26/06/2017	Standard Class	Corporate	East	Office Supplies	$28.40	5	0.0%	$8.24
6/12/2016	Standard Class	Consumer	West	Office Supplies	$35.89	1	0.0%	$16.15
11/03/2014	Second Class	Corporate	South	Office Supplies	$146.76	3	0.0%	$38.16
26/09/2016	Standard Class	Corporate	Central	Furniture	$454.97	5	30.0%	-$136.49
12/08/2014	Standard Class	Home Office	South	Office Supplies	$31.10	6	20.0%	$10.89
1/09/2015	Standard Class	Home Office	East	Office Supplies	$114.60	5	0.0%	$51.57
18/10/2015	Standard Class	Home Office	Central	Technology	$27.70	3	20.0%	$3.46
20/07/2015	Second Class	Consumer	Central	Office Supplies	$26.40	5	0.0%	$11.88
7/11/2015	Second Class	Corporate	East	Office Supplies	$7.30	2	0.0%	$3.43
28/11/2015	Standard Class	Consumer	East	Office Supplies	$3.01	2	20.0%	$0.56
8/07/2016	Second Class	Corporate	Central	Technology	$863.64	9	20.0%	$107.96
28/06/2014	Standard Class	Consumer	East	Office Supplies	$335.52	4	20.0%	$117.43
6/12/2014	First Class	Corporate	West	Office Supplies	$1,261.33	7	0.0%	$327.95
13/06/2015	First Class	Consumer	West	Office Supplies	$36.62	3	20.0%	$13.73
3/07/2016	Standard Class	Corporate	East	Office Supplies	$706.86	7	0.0%	$197.92
13/11/2017	First Class	Corporate	East	Technology	$60.86	4	20.0%	$9.13
23/12/2016	Second Class	Corporate	Central	Furniture	$2.33	2	60.0%	-$0.76
20/10/2017	Second Class	Corporate	West	Office Supplies	$20.93	4	20.0%	$7.59
1/12/2017	Standard Class	Consumer	West	Office Supplies	$45.36	7	0.0%	$21.77
18/11/2014	Standard Class	Corporate	Central	Office Supplies	$14.48	5	80.0%	-$23.89
5/03/2015	Standard Class	Consumer	Central	Office Supplies	$7.10	6	20.0%	$2.49
30/10/2017	Standard Class	Consumer	West	Office Supplies	$43.86	6	0.0%	$20.61
9/07/2015	Standard Class	Consumer	East	Office Supplies	$43.68	6	0.0%	$21.40
5/03/2015	Standard Class	Consumer	Central	Office Supplies	$60.69	7	0.0%	$16.39
24/09/2017	Same Day	Corporate	East	Technology	$391.98	2	0.0%	$109.75
23/08/2017	Second Class	Consumer	South	Technology	$4,367.90	13	20.0%	$327.59
8/09/2017	First Class	Home Office	South	Office Supplies	$61.68	5	20.0%	$5.40
2/12/2016	Standard Class	Home Office	East	Office Supplies	$2,079.40	5	0.0%	$582.23
22/09/2015	Standard Class	Home Office	South	Office Supplies	$12.00	4	20.0%	$4.20
18/10/2014	Second Class	Corporate	South	Technology	$1,394.95	5	0.0%	$362.69
14/12/2016	First Class	Corporate	West	Furniture	$81.42	2	20.0%	-$9.16
10/03/2015	Same Day	Corporate	East	Office Supplies	$89.82	6	0.0%	$25.15
22/10/2015	Standard Class	Consumer	Central	Office Supplies	$5.18	4	80.0%	-$7.76

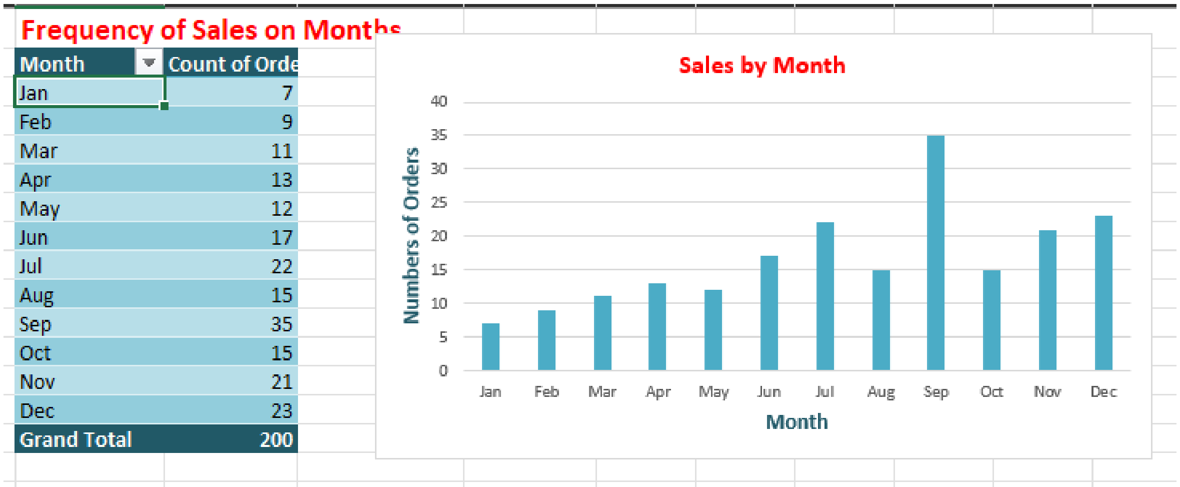
26/05/2016	Same Day	Consumer	Central	Furniture	$388.43	5	30.0%	-$88.78
29/05/2016	Standard Class	Consumer	South	Office Supplies	$4.45	2	20.0%	$0.33
16/09/2017	Second Class	Home Office	South	Technology	$18.00	1	0.0%	$3.24
20/12/2014	First Class	Home Office	South	Office Supplies	$122.48	2	0.0%	$0.00
21/08/2015	Standard Class	Corporate	West	Furniture	$586.40	6	15.0%	$34.49
19/10/2015	First Class	Consumer	East	Office Supplies	$34.44	3	0.0%	$17.22
23/11/2014	Standard Class	Corporate	East	Office Supplies	$62.81	3	20.0%	$21.20
14/02/2014	Second Class	Consumer	Central	Office Supplies	$16.18	3	20.0%	$6.07
21/07/2016	Standard Class	Consumer	South	Office Supplies	$6.26	3	20.0%	$2.04
18/06/2015	First Class	Corporate	South	Office Supplies	$20.74	4	20.0%	$7.26
22/02/2016	Standard Class	Consumer	West	Technology	$445.96	5	20.0%	$55.74
4/08/2014	Second Class	Consumer	West	Office Supplies	$1,089.75	3	0.0%	$305.13
10/09/2017	Second Class	Consumer	Central	Office Supplies	$24.18	2	0.0%	$7.25
28/05/2017	Second Class	Corporate	Central	Furniture	$106.87	3	30.0%	-$29.01
25/07/2015	Standard Class	Home Office	East	Office Supplies	$25.18	4	70.0%	-$18.46
10/11/2016	Standard Class	Corporate	Central	Office Supplies	$81.96	2	0.0%	$39.34
10/12/2016	First Class	Home Office	West	Office Supplies	$80.28	12	0.0%	$36.93
5/07/2014	First Class	Consumer	South	Office Supplies	$19.44	3	0.0%	$9.33
20/07/2017	Standard Class	Corporate	Central	Office Supplies	$14.62	2	0.0%	$6.87
20/06/2016	First Class	Consumer	West	Office Supplies	$21.78	2	0.0%	$5.66
20/06/2014	Standard Class	Consumer	Central	Office Supplies	$11.65	2	20.0%	$4.08
9/02/2015	Second Class	Corporate	Central	Technology	$20.80	2	20.0%	$6.50
20/04/2017	First Class	Home Office	East	Technology	$122.38	3	40.0%	-$24.48
24/12/2014	First Class	Consumer	South	Office Supplies	$9.57	2	20.0%	$3.47
2/10/2014	First Class	Corporate	West	Office Supplies	$4.67	2	20.0%	$1.46
4/09/2017	Standard Class	Consumer	West	Office Supplies	$421.10	2	0.0%	$105.28
9/01/2015	Standard Class	Consumer	South	Office Supplies	$51.55	5	0.0%	$24.23
8/04/2017	Standard Class	Home Office	West	Office Supplies	$244.55	5	0.0%	$114.94
11/07/2017	Second Class	Consumer	East	Technology	$132.60	6	0.0%	$17.24
20/09/2017	Standard Class	Consumer	East	Technology	$59.97	3	0.0%	$20.39
15/09/2016	Standard Class	Corporate	West	Furniture	$1,128.39	3	0.0%	$259.53
12/12/2016	Standard Class	Consumer	South	Technology	$249.95	5	0.0%	$20.00
21/08/2017	Standard Class	Consumer	Central	Office Supplies	$37.24	4	0.0%	$10.80
21/07/2016	Standard Class	Consumer	South	Furniture	$363.92	5	20.0%	$0.00
8/09/2014	Standard Class	Consumer	Central	Office Supplies	$275.93	3	20.0%	-$58.63
24/09/2017	Standard Class	Home Office	South	Office Supplies	$14.28	4	0.0%	$3.71
9/09/2014	Second Class	Consumer	East	Furniture	$17.47	3	20.0%	$5.02
6/02/2014	Second Class	Consumer	Central	Office Supplies	$8.95	2	80.0%	-$14.77
21/08/2015	First Class	Home Office	West	Furniture	$544.01	3	20.0%	$40.80
11/04/2015	Standard Class	Consumer	South	Furniture	$67.36	2	20.0%	$10.10
22/08/2016	Standard Class	Corporate	West	Office Supplies	$5.76	2	0.0%	$1.67

27/12/2014	Standard Class	Home Office	West	Office Supplies	$37.06	3	20.0%	$13.90
23/11/2015	Standard Class	Corporate	Central	Office Supplies	$335.52	4	20.0%	$117.43
13/04/2014	Second Class	Corporate	East	Office Supplies	$509.97	10	70.0%	-$407.98
10/11/2016	Second Class	Consumer	West	Office Supplies	$67.71	3	0.0%	$32.50
11/05/2014	Standard Class	Consumer	Central	Furniture	$66.11	4	60.0%	-$84.29
6/02/2017	First Class	Consumer	South	Furniture	$359.97	3	0.0%	$79.19
23/09/2016	Second Class	Consumer	South	Furniture	$368.97	3	0.0%	$81.17
16/03/2015	Second Class	Consumer	West	Furniture	$171.96	2	0.0%	$44.71
3/05/2015	Standard Class	Consumer	East	Office Supplies	$7.97	2	20.0%	$2.89
2/03/2014	Standard Class	Home Office	East	Office Supplies	$36.40	5	0.0%	$17.47
6/10/2017	Standard Class	Home Office	West	Office Supplies	$37.94	2	0.0%	$18.21
9/06/2015	Second Class	Consumer	West	Furniture	$355.36	4	0.0%	$92.39
1/01/2017	First Class	Home Office	Central	Office Supplies	$3.60	2	0.0%	$1.73
30/05/2016	Standard Class	Corporate	West	Office Supplies	$14.95	2	70.0%	-$11.96
22/07/2016	Standard Class	Corporate	East	Office Supplies	$51.84	8	0.0%	$24.88
1/08/2014	Standard Class	Corporate	South	Office Supplies	$17.54	3	20.0%	$5.92
9/01/2017	Standard Class	Consumer	East	Office Supplies	$274.49	3	70.0%	-$228.74
5/04/2015	Standard Class	Home Office	East	Technology	$41.99	2	40.0%	-$9.80
15/04/2016	Standard Class	Corporate	Central	Office Supplies	$33.49	7	20.0%	$5.86
22/07/2014	Second Class	Consumer	West	Office Supplies	$236.50	10	0.0%	$68.59
29/01/2017	Standard Class	Consumer	West	Office Supplies	$119.62	8	20.0%	$40.37
29/11/2015	Standard Class	Corporate	Central	Office Supplies	$19.92	4	0.0%	$9.36
6/02/2015	Standard Class	Consumer	East	Furniture	$1,268.82	9	0.0%	$266.45
8/06/2017	Second Class	Corporate	Central	Office Supplies	$85.06	3	20.0%	$28.71
6/02/2015	Standard Class	Consumer	East	Furniture	$283.92	4	0.0%	$82.34
2/09/2016	Same Day	Consumer	Central	Office Supplies	$1.81	1	0.0%	$0.65
16/12/2017	Standard Class	Consumer	Central	Office Supplies	$10.80	5	0.0%	$5.18
11/11/2017	Second Class	Home Office	West	Furniture	$34.92	4	0.0%	$11.87
8/09/2016	Second Class	Consumer	Central	Technology	$57.58	2	20.0%	$0.72
16/02/2017	Standard Class	Consumer	Central	Office Supplies	$18.37	2	20.0%	$6.20
15/09/2017	Standard Class	Consumer	West	Furniture	$529.90	5	0.0%	$105.98
28/09/2015	Second Class	Consumer	East	Technology	$307.98	2	0.0%	$89.31
3/07/2017	Standard Class	Home Office	East	Technology	$258.90	10	0.0%	$93.20
20/04/2017	First Class	Home Office	East	Office Supplies	$848.54	4	20.0%	-$21.21
14/10/2014	First Class	Consumer	East	Furniture	$1,628.82	9	0.0%	$260.61
21/10/2014	Standard Class	Corporate	South	Office Supplies	$2.84	1	0.0%	$0.88
28/12/2017	Standard Class	Corporate	Central	Office Supplies	$1.68	5	80.0%	-$2.69
18/09/2016	Standard Class	Consumer	South	Office Supplies	$3.00	1	20.0%	$1.05
9/08/2014	Standard Class	Home Office	West	Technology	$1,091.17	4	20.0%	$68.20
30/03/2017	Standard Class	Consumer	East	Office Supplies	$5.72	5	70.0%	-$4.76
16/12/2017	Second Class	Corporate	West	Furniture	$81.57	2	20.0%	$9.18

13/10/2016	Standard Class	Consumer	Central	Furniture	$139.92	5	60.0%	-$150.41
19/11/2017	First Class	Consumer	Central	Furniture	$191.06	3	30.0%	-$46.40
25/08/2014	Standard Class	Corporate	West	Furniture	$6.28	1	0.0%	$2.64
5/03/2015	Second Class	Consumer	Central	Office Supplies	$60.42	2	20.0%	$6.04
20/11/2016	Standard Class	Consumer	Central	Technology	$944.93	7	0.0%	$236.23
1/12/2017	First Class	Consumer	East	Office Supplies	$37.39	3	20.0%	$2.34
12/12/2014	Second Class	Consumer	West	Furniture	$764.69	6	20.0%	$95.59
26/08/2016	Standard Class	Corporate	East	Office Supplies	$5.47	3	20.0%	$1.64
25/07/2016	Standard Class	Consumer	West	Furniture	$255.76	4	0.0%	$81.84
27/07/2017	Second Class	Consumer	South	Furniture	$194.85	4	20.0%	$12.18
6/05/2014	Standard Class	Home Office	West	Office Supplies	$5.78	2	0.0%	$2.72
20/04/2017	First Class	Corporate	East	Furniture	$51.97	2	20.0%	$10.39
6/09/2014	First Class	Corporate	West	Furniture	$41.88	6	0.0%	$12.15
25/09/2015	Standard Class	Corporate	South	Office Supplies	$10.48	1	20.0%	$3.80
26/05/2014	Standard Class	Corporate	West	Technology	$201.58	2	20.0%	$20.16
20/01/2014	Standard Class	Consumer	West	Office Supplies	$19.36	2	0.0%	$9.29
8/02/2015	First Class	Consumer	Central	Office Supplies	$5.81	1	0.0%	$1.80
13/12/2016	Standard Class	Corporate	West	Office Supplies	$6.10	2	20.0%	$2.21
26/04/2017	First Class	Consumer	Central	Furniture	$1.99	1	60.0%	-$1.44
30/05/2017	First Class	Consumer	South	Furniture	$8.01	3	0.0%	$3.04
6/03/2017	Second Class	Corporate	West	Office Supplies	$67.78	2	0.0%	$16.95
15/09/2014	Standard Class	Consumer	East	Office Supplies	$14.94	3	0.0%	$7.02
24/11/2017	Second Class	Corporate	South	Technology	$79.10	2	0.0%	$39.55
27/06/2017	Second Class	Corporate	West	Furniture	$126.30	3	0.0%	$40.42
11/08/2015	Standard Class	Consumer	South	Furniture	$46.15	3	20.0%	$12.11
9/07/2014	Standard Class	Home Office	West	Office Supplies	$23.92	4	0.0%	$4.07
25/11/2014	Second Class	Corporate	West	Office Supplies	$26.76	4	0.0%	$12.31
23/11/2015	Second Class	Consumer	East	Office Supplies	$13.12	4	0.0%	$5.64
5/09/2016	Second Class	Consumer	Central	Office Supplies	$70.95	3	0.0%	$20.58
3/05/2016	First Class	Consumer	East	Technology	$224.94	3	70.0%	-$164.95
14/07/2016	Standard Class	Consumer	South	Office Supplies	$36.40	5	0.0%	$17.11
24/09/2017	First Class	Consumer	South	Office Supplies	$40.29	3	0.0%	$10.07
5/05/2017	Same Day	Consumer	East	Office Supplies	$6.68	1	0.0%	$3.21
28/10/2016	Same Day	Consumer	South	Furniture	$165.05	3	20.0%	$41.26
15/12/2015	Standard Class	Home Office	East	Office Supplies	$3.28	1	0.0%	$1.41
27/03/2016	Standard Class	Corporate	South	Furniture	$20.24	1	0.0%	$8.70
20/06/2016	Standard Class	Corporate	Central	Furniture	$57.69	3	0.0%	$23.65
21/09/2014	Standard Class	Consumer	West	Technology	$239.98	2	20.0%	$24.00
10/11/2015	Standard Class	Consumer	West	Technology	$79.90	2	0.0%	$35.16
6/12/2016	Standard Class	Home Office	West	Office Supplies	$13.86	7	0.0%	$0.00
28/07/2014	Same Day	Consumer	South	Office Supplies	$14.32	5	20.0%	$5.19

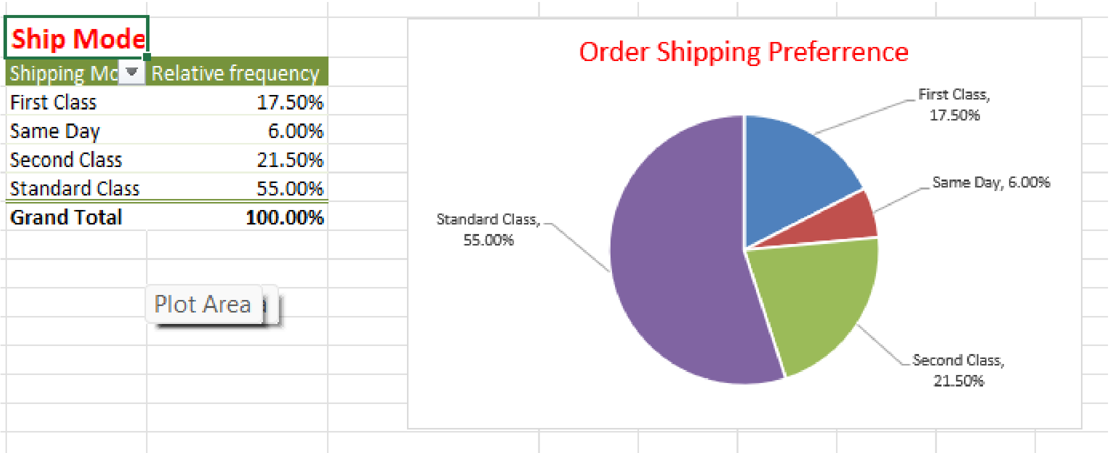
9/04/2016	Second Class	Home Office	Central	Office Supplies	$5.28	2	0.0%	$2.43
14/07/2016	First Class	Consumer	East	Office Supplies	$11.63	2	20.0%	$1.02
18/03/2017	Standard Class	Consumer	West	Office Supplies	$46.20	4	0.0%	$21.25
1/09/2016	Standard Class	Consumer	Central	Office Supplies	$376.74	4	10.0%	$71.16
12/07/2015	Second Class	Consumer	Central	Furniture	$7.76	1	60.0%	-$2.13
20/09/2016	Standard Class	Corporate	South	Furniture	$98.39	1	20.0%	-$11.07
27/11/2017	Standard Class	Corporate	Central	Office Supplies	$158.28	6	0.0%	$72.81
4/09/2017	Standard Class	Consumer	Central	Office Supplies	$10.19	7	20.0%	$3.19
4/09/2016	Same Day	Consumer	West	Technology	$2,799.96	5	20.0%	$944.99
19/08/2017	Standard Class	Consumer	West	Office Supplies	$102.72	3	20.0%	$37.24
8/05/2016	Same Day	Consumer	East	Office Supplies	$10.37	2	20.0%	$3.63
1/12/2016	Second Class	Consumer	East	Office Supplies	$88.08	6	0.0%	$40.52
7/09/2015	Standard Class	Corporate	East	Technology	$90.00	5	0.0%	$16.20
11/09/2016	Standard Class	Consumer	Central	Office Supplies	$99.57	2	20.0%	$33.60
27/12/2015	Standard Class	Corporate	West	Technology	$668.16	9	20.0%	$75.17
8/08/2016	Standard Class	Home Office	West	Office Supplies	$15.24	5	20.0%	$5.33
28/11/2017	Standard Class	Home Office	East	Furniture	$516.49	7	20.0%	-$12.91
2/06/2014	Standard Class	Home Office	West	Office Supplies	$59.81	3	20.0%	$19.44
8/12/2017	Second Class	Home Office	East	Office Supplies	$592.74	6	0.0%	$160.04
13/09/2015	Standard Class	Consumer	Central	Technology	$199.96	4	0.0%	$16.00
5/10/2017	Standard Class	Corporate	West	Office Supplies	$15.80	4	0.0%	$4.11
21/06/2017	Standard Class	Home Office	East	Office Supplies	$6.24	2	0.0%	$3.06
29/06/2016	Second Class	Consumer	South	Office Supplies	$191.88	6	0.0%	$19.19
26/11/2017	First Class	Consumer	Central	Furniture	$126.30	3	0.0%	$40.42
24/11/2016	First Class	Home Office	East	Office Supplies	$40.75	3	20.0%	$15.28
7/07/2016	Standard Class	Consumer	West	Office Supplies	$45.98	2	0.0%	$12.87
1/11/2014	Standard Class	Corporate	South	Office Supplies	$7.52	5	20.0%	$1.41
24/01/2017	Standard Class	Consumer	South	Office Supplies	$5.67	3	0.0%	$0.11
14/09/2014	Same Day	Consumer	East	Office Supplies	$68.46	7	0.0%	$31.49
2/10/2017	Standard Class	Consumer	West	Furniture	$217.76	6	70.0%	-$384.72
22/09/2015	First Class	Consumer	East	Technology	$617.98	3	20.0%	-$7.72
23/01/2017	Standard Class	Corporate	West	Office Supplies	$6.48	1	0.0%	$3.11
29/12/2014	Standard Class	Consumer	Central	Furniture	$38.98	3	60.0%	-$50.67

Appendix-B (Excel Output)



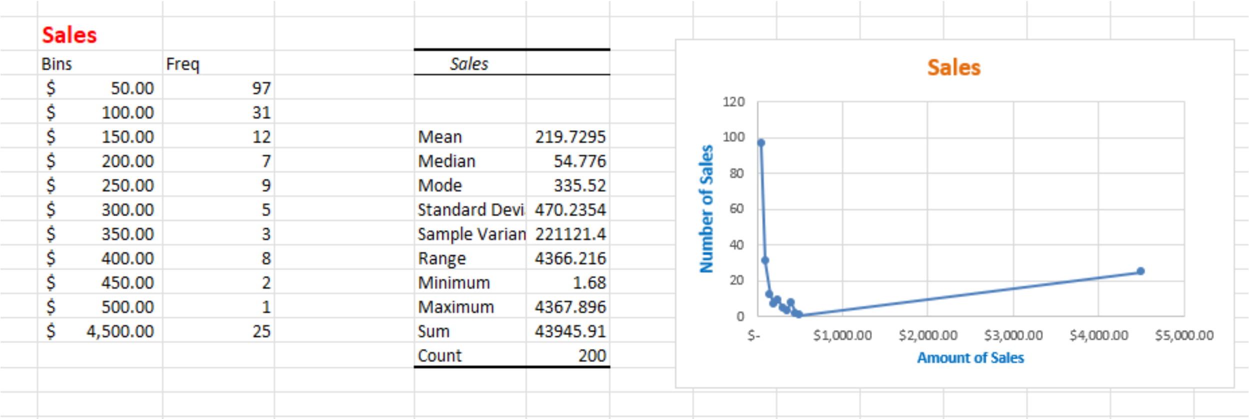
Frequency of Sales on Months

Month	Count of Orde
Jan	7
Feb	9
Mar	11
Apr	13
May	12
Jun	17
Jul	22
Aug	15
Sep	35
Oct	15
Nov	21
Dec	23
Grand Total	200



Ship Mode

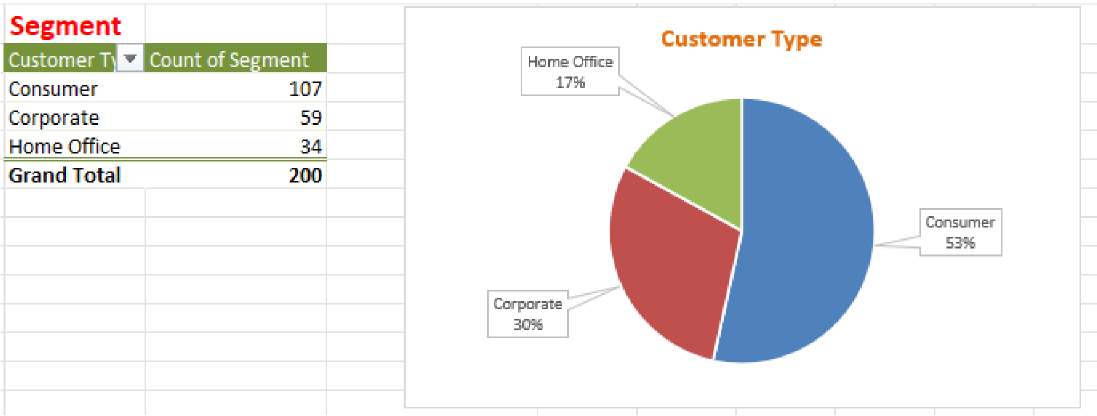
Shipping Mo	Relative frequency
First Class	17.50%
Same Day	6.00%
Second Class	21.50%
Standard Class	55.00%
Grand Total	**100.00%**



Sales

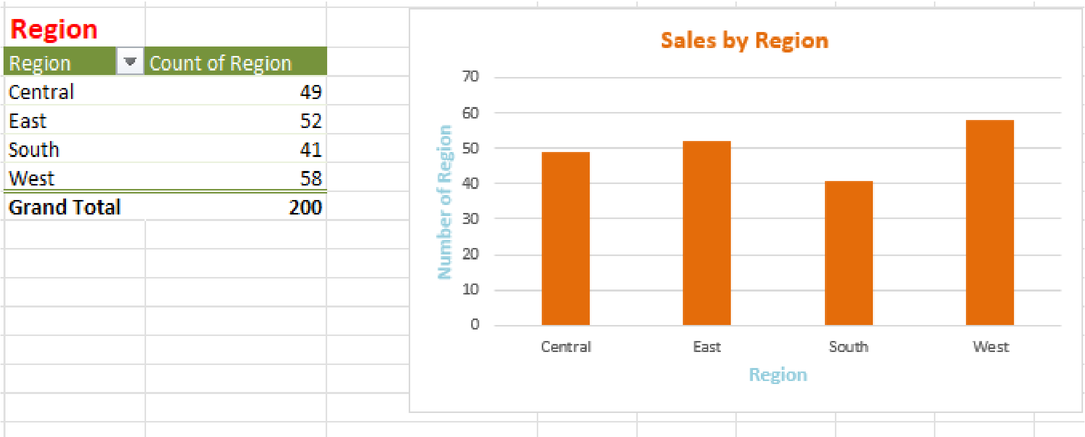
Bins	Freq
$ 50.00	97
$ 100.00	31
$ 150.00	12
$ 200.00	7
$ 250.00	9
$ 300.00	5
$ 350.00	3
$ 400.00	8
$ 450.00	2
$ 500.00	1
$ 4,500.00	25

Sales	
Mean	219.7295
Median	54.776
Mode	335.52
Standard Devi	470.2354
Sample Varian	221121.4
Range	4366.216
Minimum	1.68
Maximum	4367.896
Sum	43945.91
Count	200



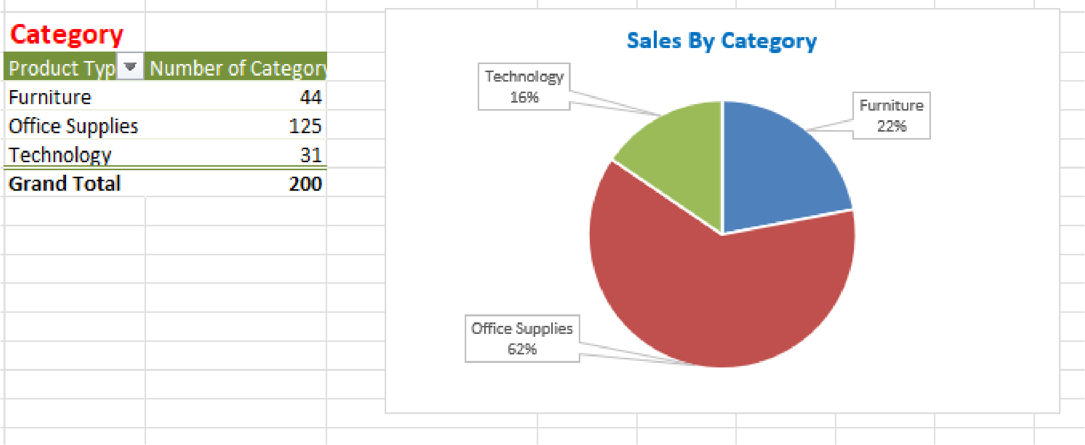
Segment

Customer Ty	Count of Segment
Consumer	107
Corporate	59
Home Office	34
Grand Total	**200**



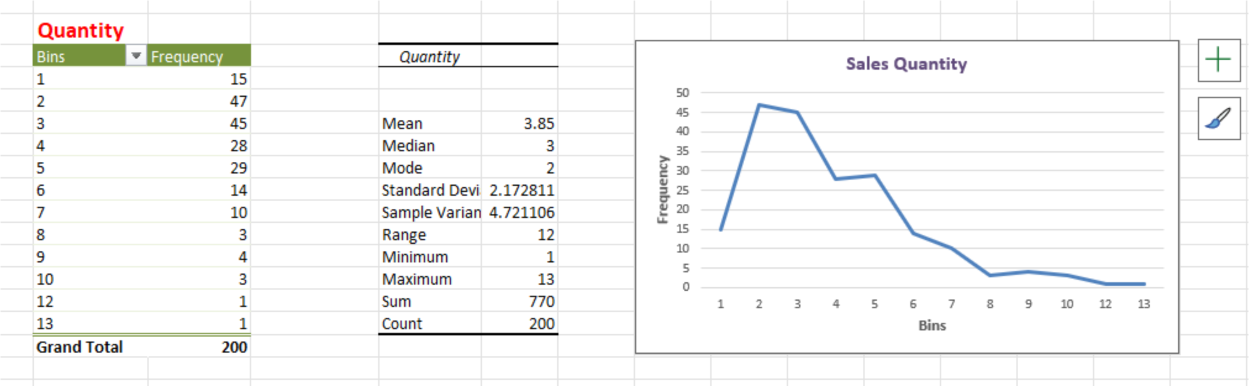
Region

Region	Count of Region
Central	49
East	52
South	41
West	58
Grand Total	**200**



Category

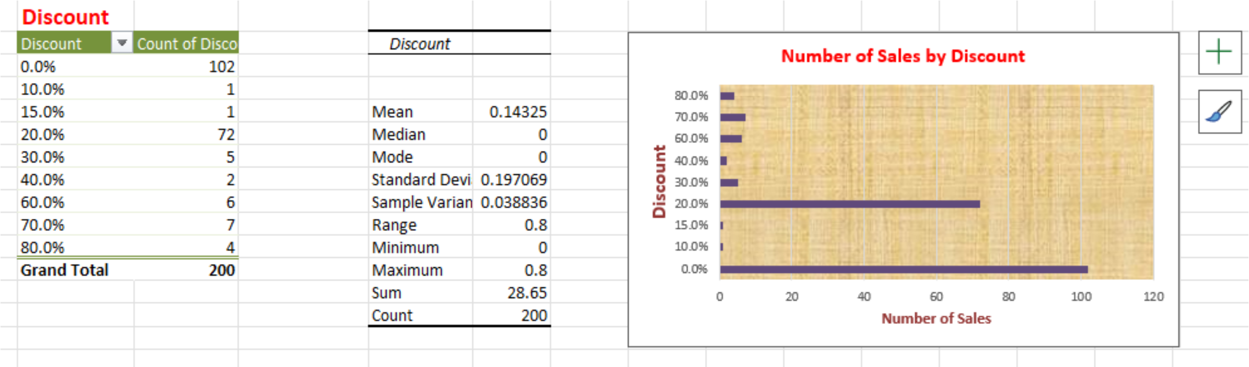
Product Typ	Number of Category
Furniture	44
Office Supplies	125
Technology	31
Grand Total	**200**



Quantity

Bins	Frequency
1	15
2	47
3	45
4	28
5	29
6	14
7	10
8	3
9	4
10	3
12	1
13	1
Grand Total	**200**

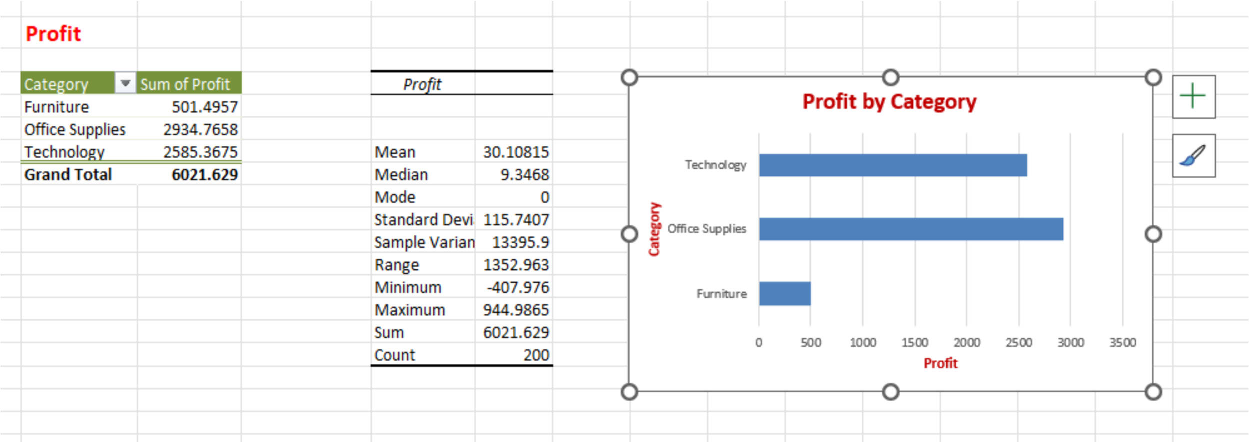
Quantity	
Mean	3.85
Median	3
Mode	2
Standard Devi	2.172811
Sample Varian	4.721106
Range	12
Minimum	1
Maximum	13
Sum	770
Count	200



Discount

Discount	Count of Disco
0.0%	102
10.0%	1
15.0%	1
20.0%	72
30.0%	5
40.0%	2
60.0%	6
70.0%	7
80.0%	4
Grand Total	**200**

Discount	
Mean	0.14325
Median	0
Mode	0
Standard Devi	0.197069
Sample Varian	0.038836
Range	0.8
Minimum	0
Maximum	0.8
Sum	28.65
Count	200



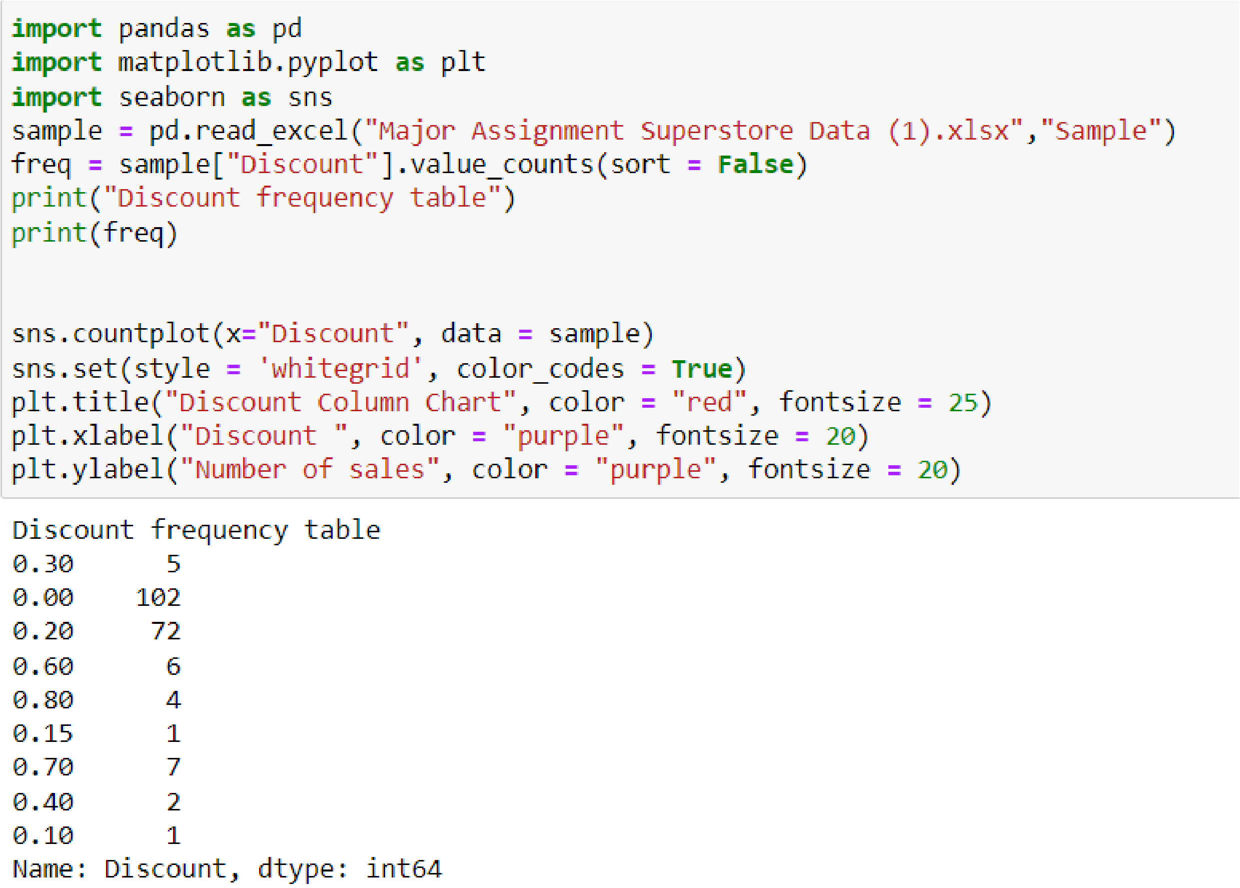
Profit

Category	Sum of Profit
Furniture	501.4957
Office Supplies	2934.7658
Technology	2585.3675
Grand Total	**6021.629**

Profit	
Mean	30.10815
Median	9.3468
Mode	0
Standard Devi	115.7407
Sample Varian	13395.9
Range	1352.963
Minimum	-407.976
Maximum	944.9865
Sum	6021.629
Count	200

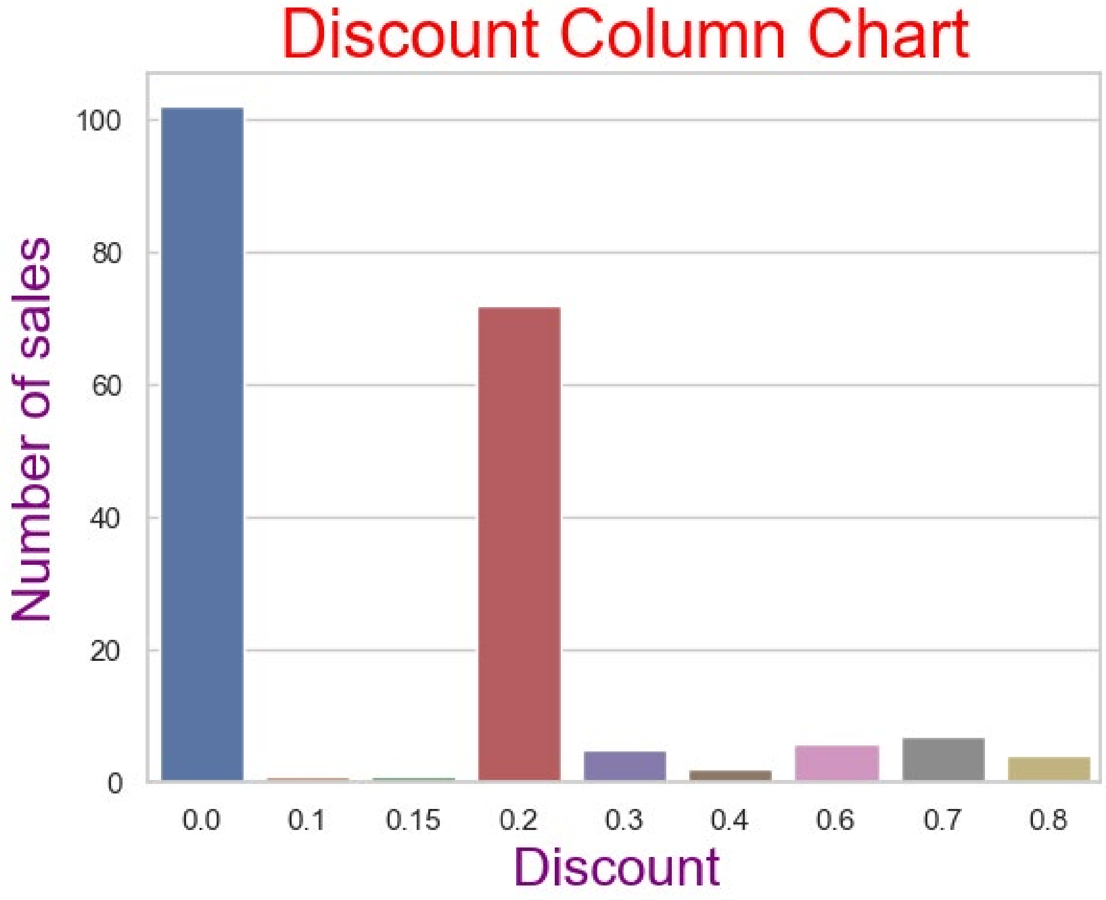
Appendix-C (Python Output and Code)

Python Code Fig 1

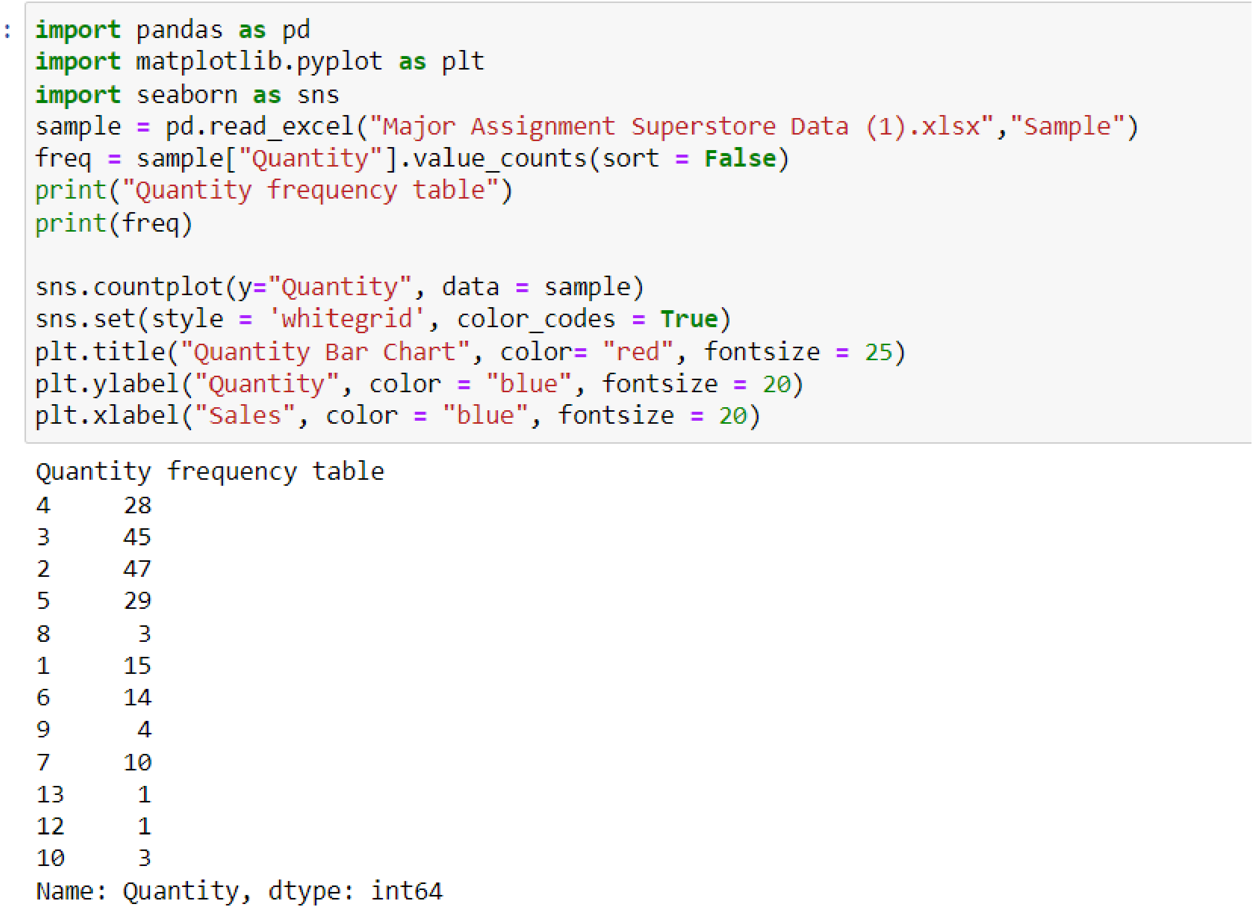


```
import pandas as pd
import matplotlib.pyplot as plt
import seaborn as sns
sample = pd.read_excel("Major Assignment Superstore Data (1).xlsx","Sample")
freq = sample["Discount"].value_counts(sort = False)
print("Discount frequency table")
print(freq)


sns.countplot(x="Discount", data = sample)
sns.set(style = 'whitegrid', color_codes = True)
plt.title("Discount Column Chart", color = "red", fontsize = 25)
plt.xlabel("Discount ", color = "purple", fontsize = 20)
plt.ylabel("Number of sales", color = "purple", fontsize = 20)
```

```
Discount frequency table
0.30      5
0.00    102
0.20     72
0.60      6
0.80      4
0.15      1
0.70      7
0.40      2
0.10      1
Name: Discount, dtype: int64
```

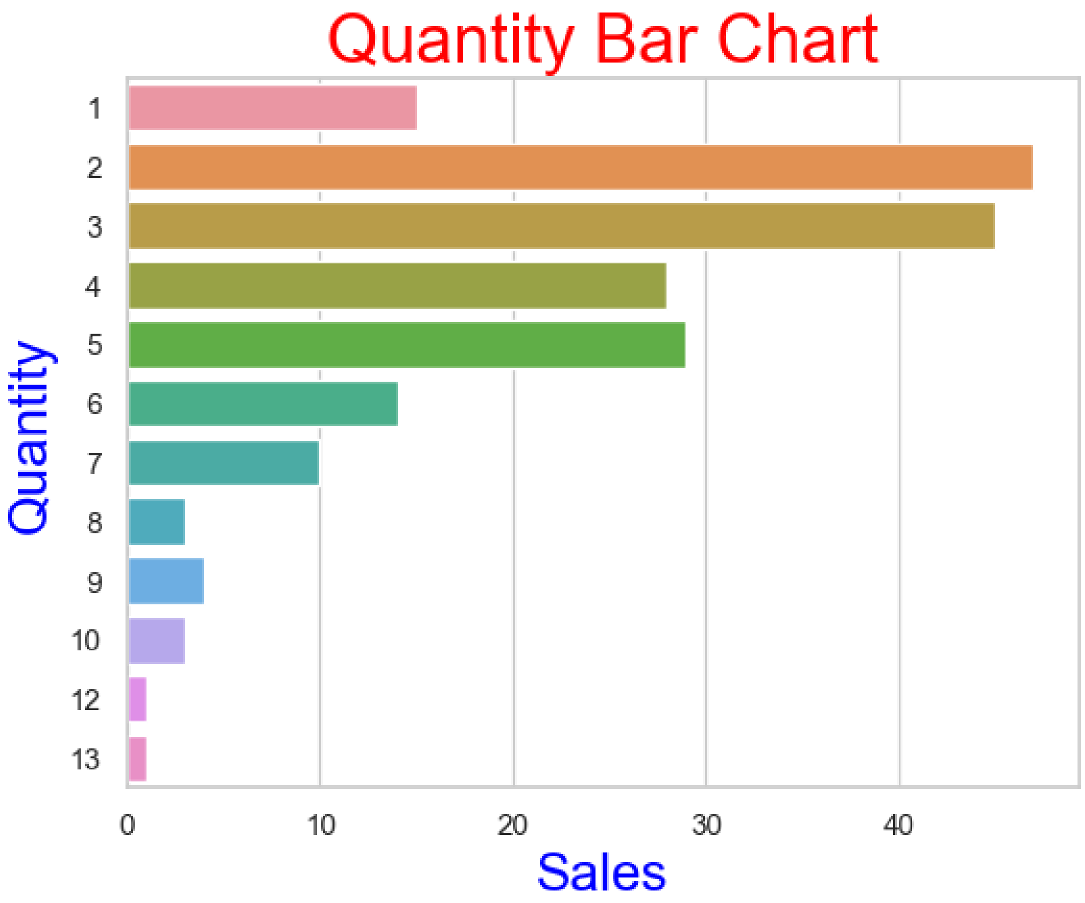


Python Code Fig 2

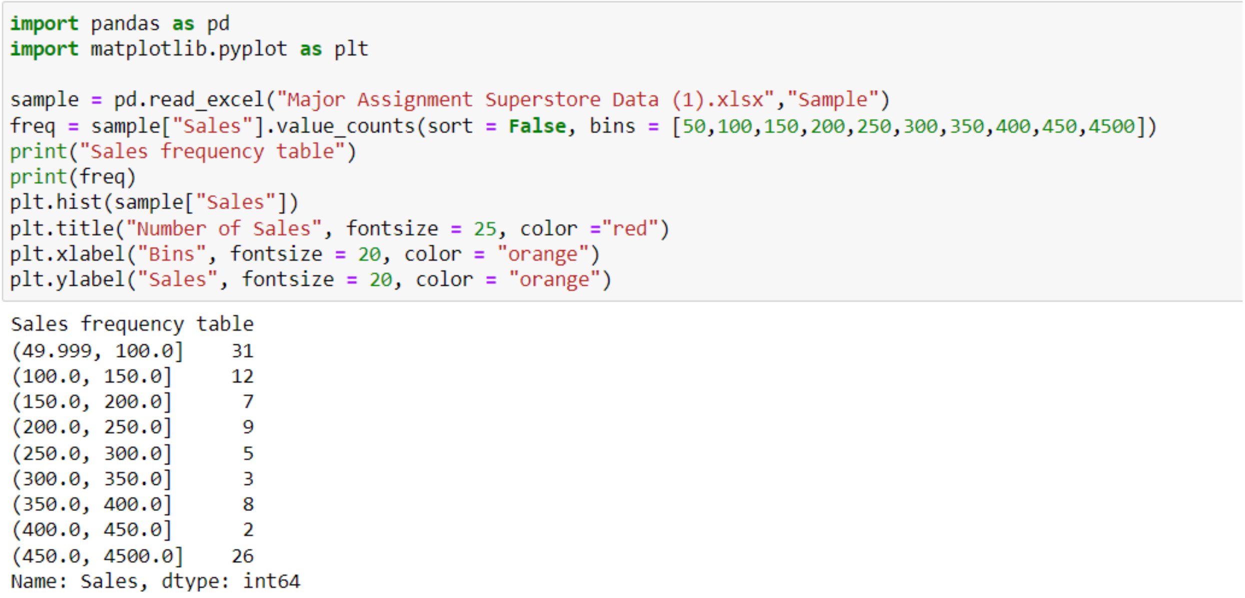


```
import pandas as pd
import matplotlib.pyplot as plt
import seaborn as sns
sample = pd.read_excel("Major Assignment Superstore Data (1).xlsx","Sample")
freq = sample["Quantity"].value_counts(sort = False)
print("Quantity frequency table")
print(freq)

sns.countplot(y="Quantity", data = sample)
sns.set(style = 'whitegrid', color_codes = True)
plt.title("Quantity Bar Chart", color= "red", fontsize = 25)
plt.ylabel("Quantity", color = "blue", fontsize = 20)
plt.xlabel("Sales", color = "blue", fontsize = 20)
```

```
Quantity frequency table
4     28
3     45
2     47
5     29
8      3
1     15
6     14
9      4
7     10
13     1
12     1
10     3
Name: Quantity, dtype: int64
```

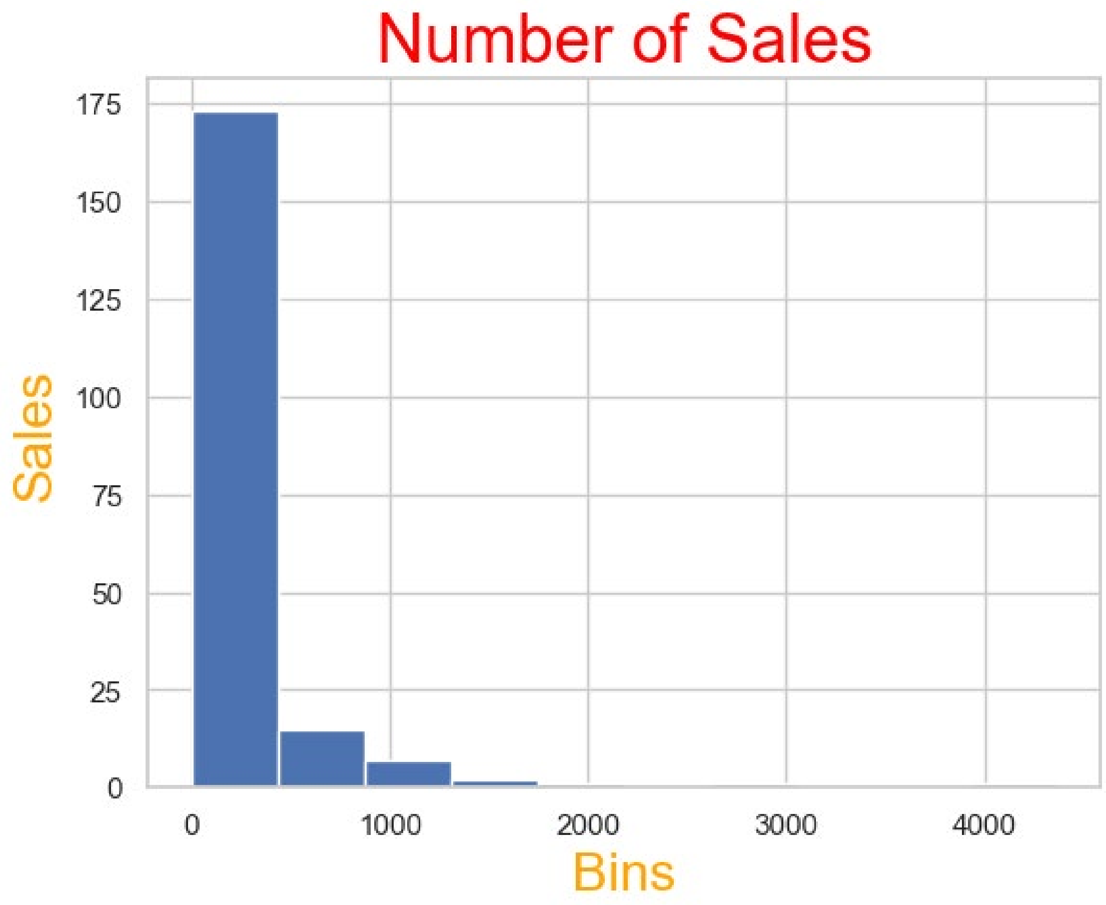


Python Code Fig 3

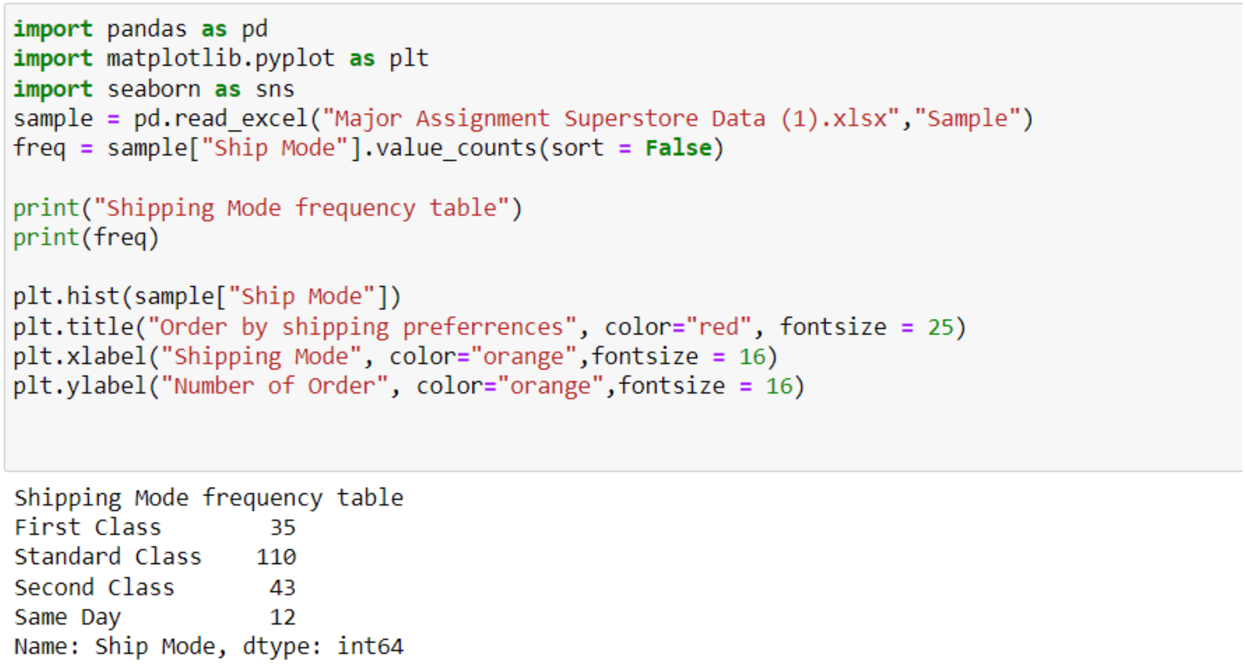


```
import pandas as pd
import matplotlib.pyplot as plt

sample = pd.read_excel("Major Assignment Superstore Data (1).xlsx","Sample")
freq = sample["Sales"].value_counts(sort = False, bins = [50,100,150,200,250,300,350,400,450,4500])
print("Sales frequency table")
print(freq)
plt.hist(sample["Sales"])
plt.title("Number of Sales", fontsize = 25, color ="red")
plt.xlabel("Bins", fontsize = 20, color = "orange")
plt.ylabel("Sales", fontsize = 20, color = "orange")
```

```
Sales frequency table
(49.999, 100.0]    31
(100.0, 150.0]     12
(150.0, 200.0]      7
(200.0, 250.0]      9
(250.0, 300.0]      5
(300.0, 350.0]      3
(350.0, 400.0]      8
(400.0, 450.0]      2
(450.0, 4500.0]    26
Name: Sales, dtype: int64
```

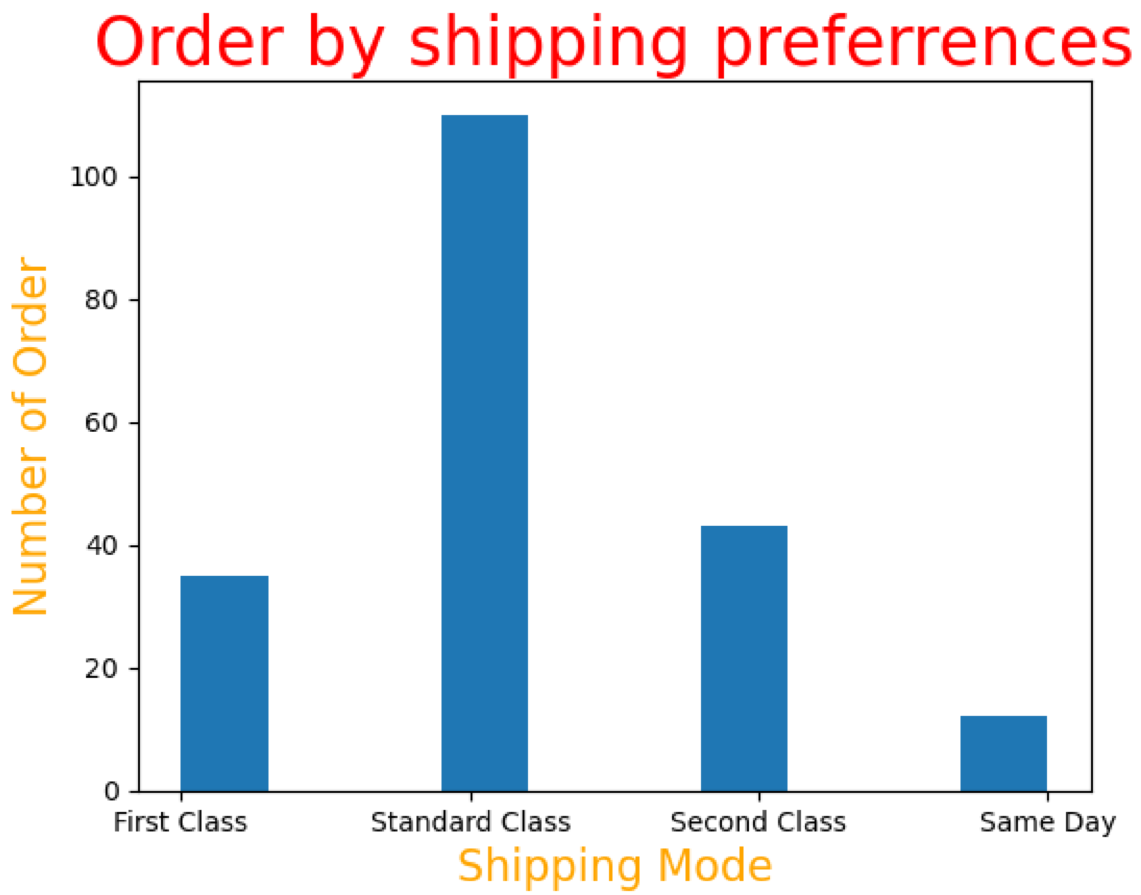


Python code Fig 4

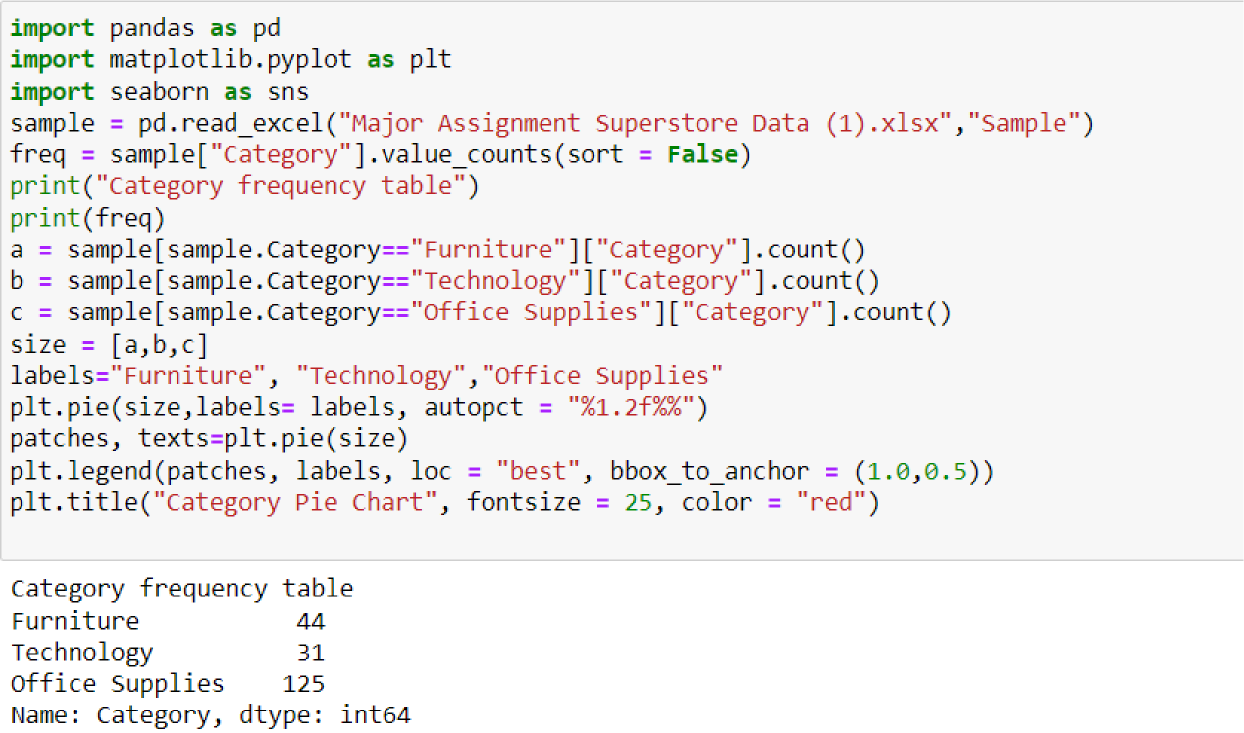


```
import pandas as pd
import matplotlib.pyplot as plt
import seaborn as sns
sample = pd.read_excel("Major Assignment Superstore Data (1).xlsx","Sample")
freq = sample["Ship Mode"].value_counts(sort = False)

print("Shipping Mode frequency table")
print(freq)

plt.hist(sample["Ship Mode"])
plt.title("Order by shipping preferrences", color="red", fontsize = 25)
plt.xlabel("Shipping Mode", color="orange",fontsize = 16)
plt.ylabel("Number of Order", color="orange",fontsize = 16)
```

```
Shipping Mode frequency table
First Class        35
Standard Class    110
Second Class       43
Same Day           12
Name: Ship Mode, dtype: int64
```

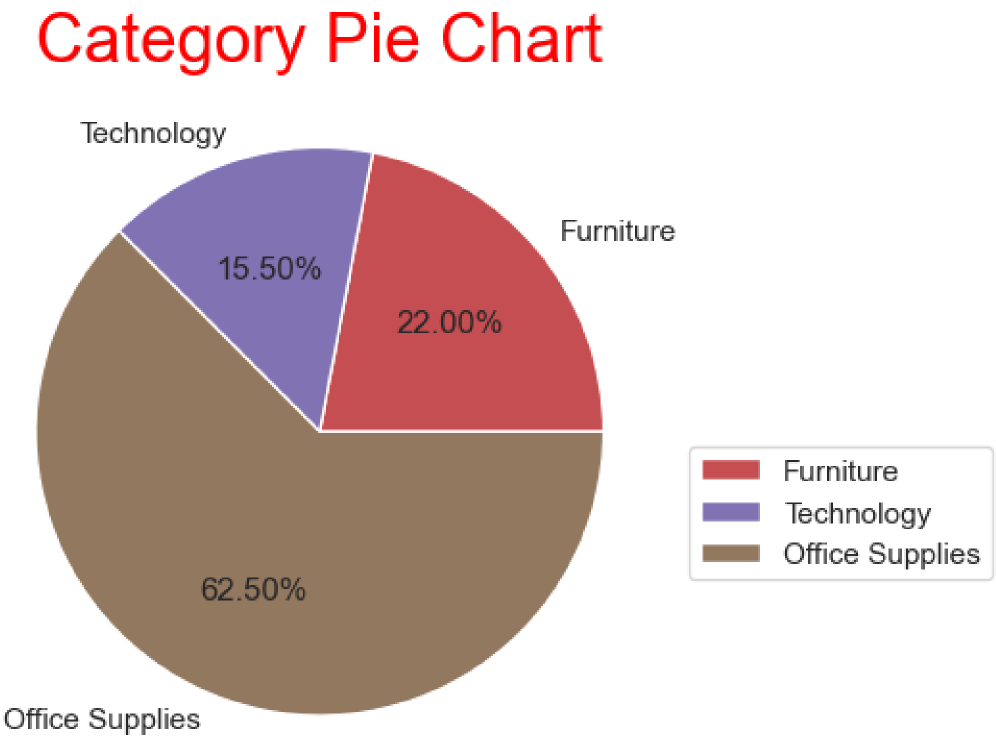


Python code Fig 5



```python
import pandas as pd
import matplotlib.pyplot as plt
import seaborn as sns
sample = pd.read_excel("Major Assignment Superstore Data (1).xlsx","Sample")
freq = sample["Category"].value_counts(sort = False)
print("Category frequency table")
print(freq)
a = sample[sample.Category=="Furniture"]["Category"].count()
b = sample[sample.Category=="Technology"]["Category"].count()
c = sample[sample.Category=="Office Supplies"]["Category"].count()
size = [a,b,c]
labels="Furniture", "Technology","Office Supplies"
plt.pie(size,labels= labels, autopct = "%1.2f%%")
patches, texts=plt.pie(size)
plt.legend(patches, labels, loc = "best", bbox_to_anchor = (1.0,0.5))
plt.title("Category Pie Chart", fontsize = 25, color = "red")
```

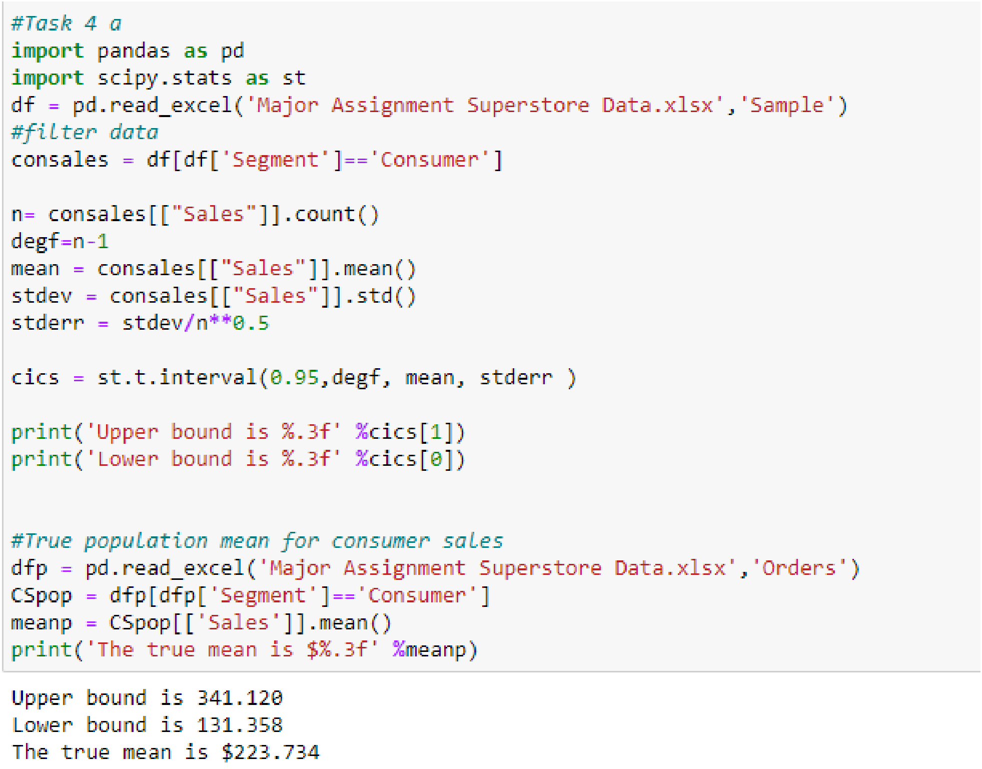
```
Category frequency table
Furniture           44
Technology          31
Office Supplies    125
Name: Category, dtype: int64
```



Appendix-D

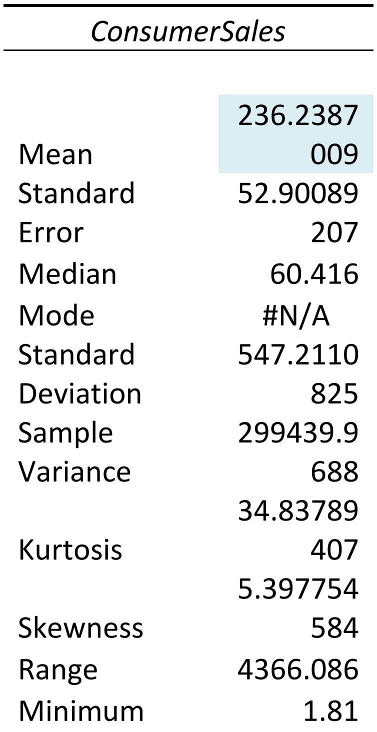
Confidence interval calculation

4a

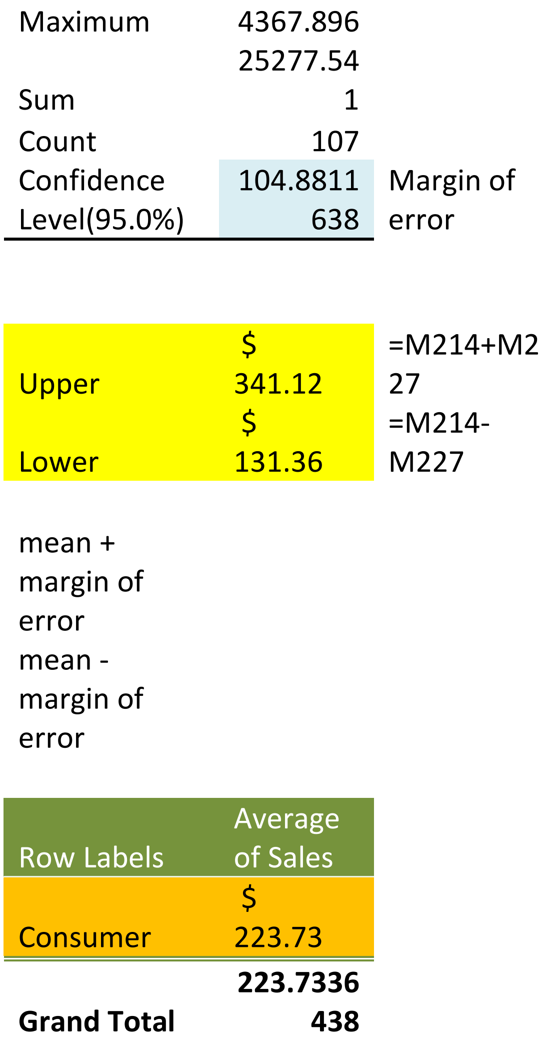


```
#Task 4 a
import pandas as pd
import scipy.stats as st
df = pd.read_excel('Major Assignment Superstore Data.xlsx','Sample')
#filter data
consales = df[df['Segment']=='Consumer']

n= consales[["Sales"]].count()
degf=n-1
mean = consales[["Sales"]].mean()
stdev = consales[["Sales"]].std()
stderr = stdev/n**0.5

cics = st.t.interval(0.95,degf, mean, stderr )

print('Upper bound is %.3f' %cics[1])
print('Lower bound is %.3f' %cics[0])


#True population mean for consumer sales
dfp = pd.read_excel('Major Assignment Superstore Data.xlsx','Orders')
CSpop = dfp[dfp['Segment']=='Consumer']
meanp = CSpop[['Sales']].mean()
print('The true mean is $%.3f' %meanp)
```

```
Upper bound is 341.120
Lower bound is 131.358
The true mean is $223.734
```



	ConsumerSales
Mean	236.2387009
Standard Error	52.90089207
Median	60.416
Mode	#N/A
Standard Deviation	547.2110825
Sample Variance	299439.9688
Kurtosis	34.83789407
Skewness	5.397754584
Range	4366.086
Minimum	1.81



Maximum	4367.896	
Sum	25277.541	
Count	107	
Confidence Level(95.0%)	104.8811638	Margin of error

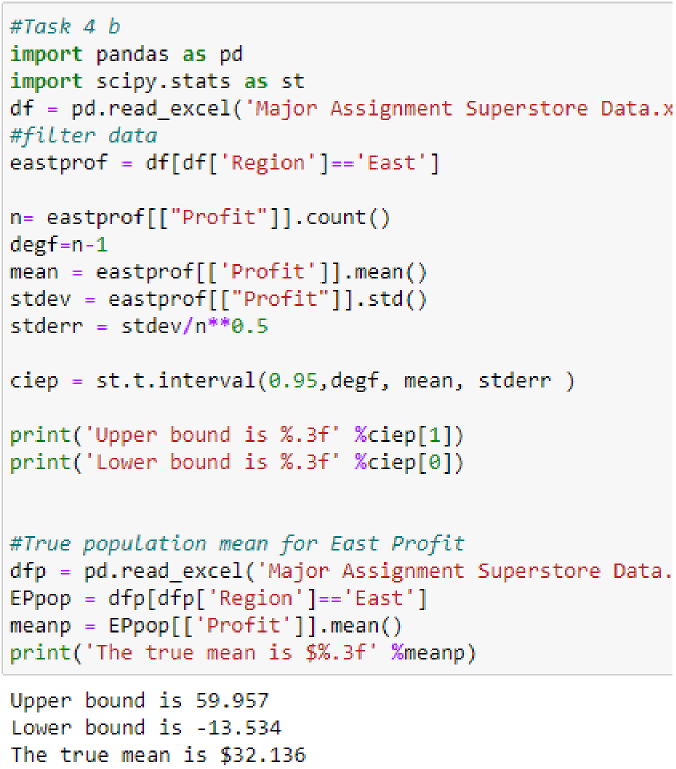
Upper	$ 341.12	=M214+M227
Lower	$ 131.36	=M214-M227

mean + margin of error

mean - margin of error

Row Labels	Average of Sales
Consumer	$ 223.73
Grand Total	**223.7336438**

4b



```
#Task 4 b
import pandas as pd
import scipy.stats as st
df = pd.read_excel('Major Assignment Superstore Data.x
#filter data
eastprof = df[df['Region']=='East']

n= eastprof[["Profit"]].count()
degf=n-1
mean = eastprof[['Profit']].mean()
stdev = eastprof[["Profit"]].std()
stderr = stdev/n**0.5

ciep = st.t.interval(0.95,degf, mean, stderr )

print('Upper bound is %.3f' %ciep[1])
print('Lower bound is %.3f' %ciep[0])


#True population mean for East Profit
dfp = pd.read_excel('Major Assignment Superstore Data.
EPpop = dfp[dfp['Region']=='East']
meanp = EPpop[['Profit']].mean()
print('The true mean is $%.3f' %meanp)
```

```
Upper bound is 59.957
Lower bound is -13.534
The true mean is $32.136
```

Profit	
Mean	23.21160385
Standard Error	18.30317714
Median	8.6828
Mode	#N/A
Standard Deviation	131.9860873
Sample Variance	17420.32725
Kurtosis	7.861164025
Skewness	0.771632218
Range	990.208
Minimum	-407.976
Maximum	582.232
Sum	1207.0034
Count	52

Confidence Level(95.0%)	36.74516136	Margin of Error

Upper	$ 59.96
Lower	-$ 13.53

Row Labels	Average of Profit
East	$ 32.14
Grand Total	**$ 32.14**

Hypothesis Testing 1

5A

Ho	µHome - µCrop ≤ 0
Ha	µCrop - µHome > 0

1 tail

t-Test: Two-Sample Assuming Equal Variances

	CorpQuantity	*HomeQuantity*
Mean	3.728813559	4.117647059
Variance	4.787258913	5.561497326
Observations	59	34
Pooled Variance	5.068026689	
Hypothesized Mean Difference	0	
df	91	
t Stat	-0.802173327	
P(T<=t) one-tail	0.212271312	
t Critical one-tail	1.661771155	
P(T<=t) two-tail	0.424542624	
t Critical two-tail	1.986377154	

Do not reject the Ho

pvalue	0.212271312
alpha	0.05

pvalue > alpha thus do not reject the Ho

Hypothesis testing 2

Ho: µFurn - µTech = 0

Ha: µFurn - µTech ≠ 0

two tail

t-Test: Two-Sample Assuming Equal Variances

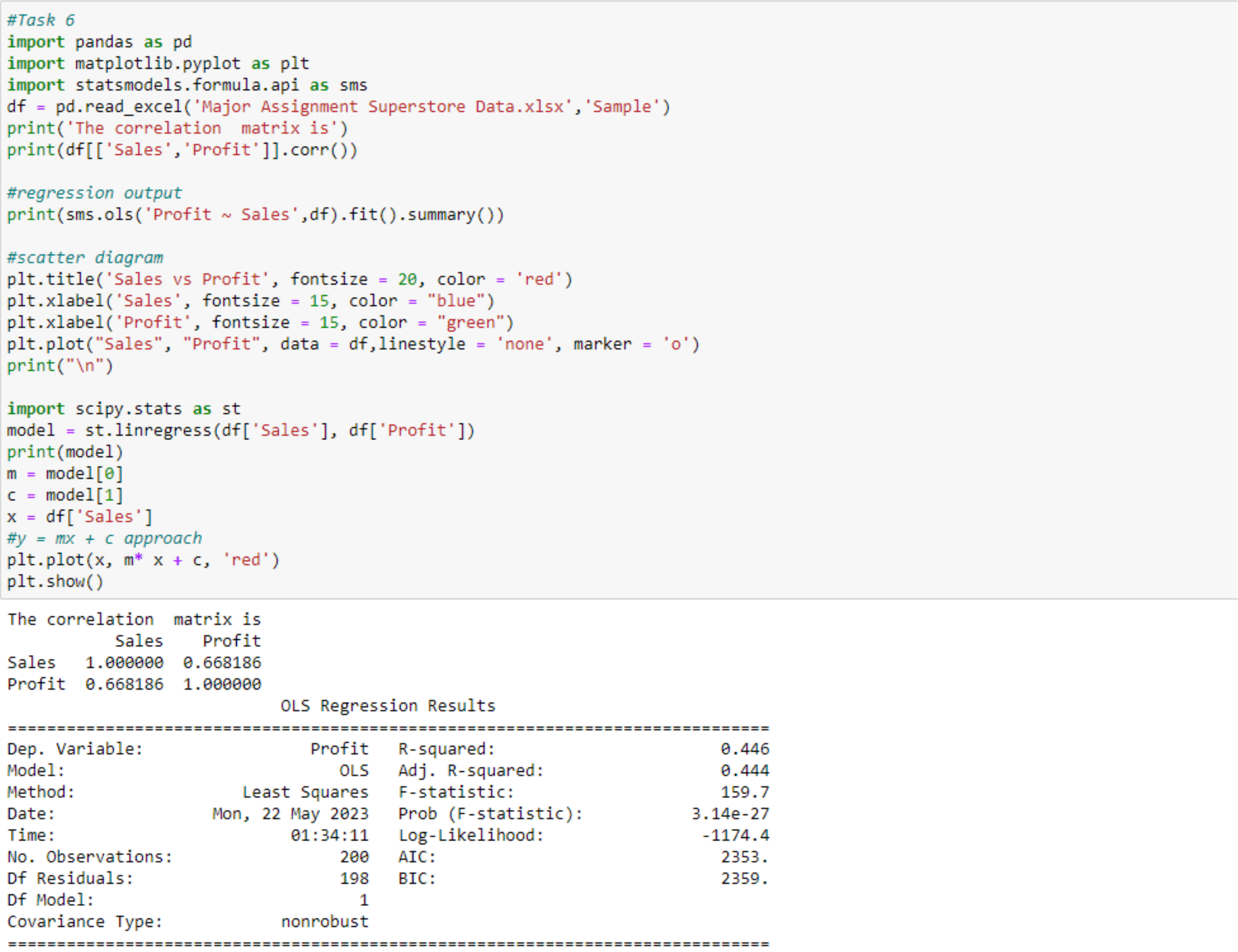
	FunProfit	*TechProfit*
Mean	11.39762955	83.39895161
Variance	12655.56046	35451.04712
Observations	44	31
Pooled Variance	22023.56868	
Hypothesized Mean Difference	0	
df	73	
t Stat	-2.069060443	
P(T<=t) one-tail	0.021040742	
t Critical one-tail	1.665996224	
P(T<=t) two-tail	0.042081483	
t Critical two-tail	±1.99299712588986	

alpha	0.05
pvalue	0.042081483

pvalue is less thatn alpha so reject.

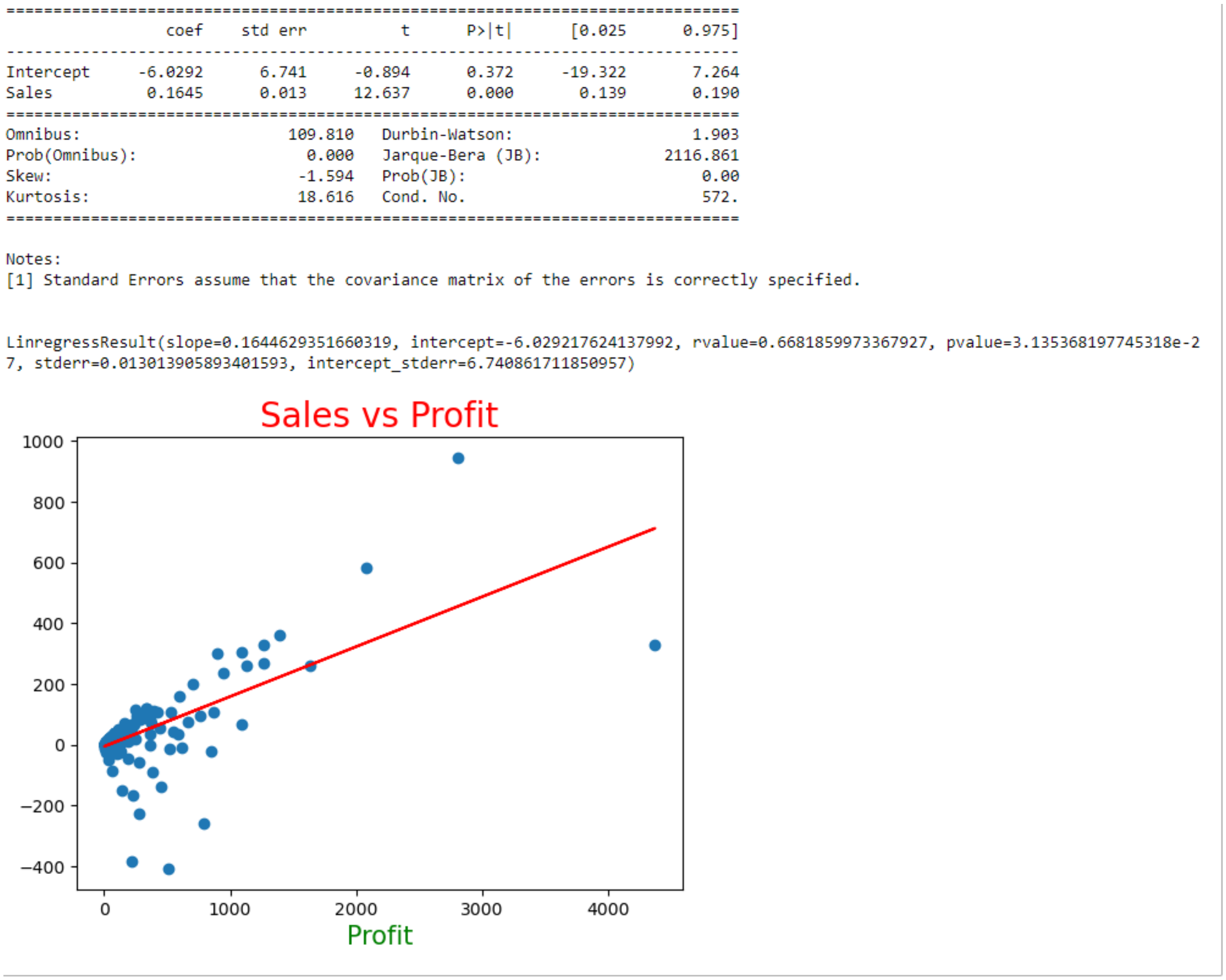
Reject Ho

Coefficient and Correlation

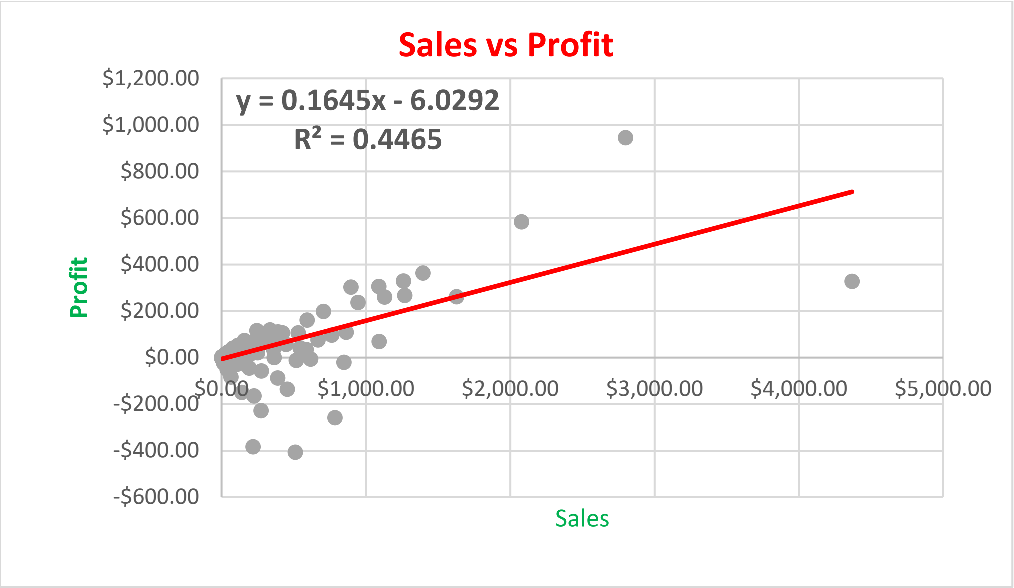


```
#Task 6
import pandas as pd
import matplotlib.pyplot as plt
import statsmodels.formula.api as sms
df = pd.read_excel('Major Assignment Superstore Data.xlsx','Sample')
print('The correlation  matrix is')
print(df[['Sales','Profit']].corr())

#regression output
print(sms.ols('Profit ~ Sales',df).fit().summary())

#scatter diagram
plt.title('Sales vs Profit', fontsize = 20, color = 'red')
plt.xlabel('Sales', fontsize = 15, color = "blue")
plt.xlabel('Profit', fontsize = 15, color = "green")
plt.plot("Sales", "Profit", data = df,linestyle = 'none', marker = 'o')
print("\n")

import scipy.stats as st
model = st.linregress(df['Sales'], df['Profit'])
print(model)
m = model[0]
c = model[1]
x = df['Sales']
#y = mx + c approach
plt.plot(x, m* x + c, 'red')
plt.show()
```

```
The correlation  matrix is
           Sales    Profit
Sales   1.000000  0.668186
Profit  0.668186  1.000000
                            OLS Regression Results
==============================================================================
Dep. Variable:                 Profit   R-squared:                       0.446
Model:                            OLS   Adj. R-squared:                  0.444
Method:                 Least Squares   F-statistic:                     159.7
Date:                Mon, 22 May 2023   Prob (F-statistic):           3.14e-27
Time:                        01:34:11   Log-Likelihood:                -1174.4
No. Observations:                 200   AIC:                             2353.
Df Residuals:                     198   BIC:                             2359.
Df Model:                           1
Covariance Type:            nonrobust
==============================================================================
```



```
==============================================================================
                 coef    std err          t      P>|t|      [0.025      0.975]
------------------------------------------------------------------------------
Intercept     -6.0292      6.741     -0.894      0.372     -19.322       7.264
Sales          0.1645      0.013     12.637      0.000       0.139       0.190
==============================================================================
Omnibus:                      109.810   Durbin-Watson:                   1.903
Prob(Omnibus):                  0.000   Jarque-Bera (JB):             2116.861
Skew:                          -1.594   Prob(JB):                         0.00
Kurtosis:                      18.616   Cond. No.                         572.
==============================================================================

Notes:
[1] Standard Errors assume that the covariance matrix of the errors is correctly specified.


LinregressResult(slope=0.1644629351660319, intercept=-6.029217624137992, rvalue=0.6681859973367927, pvalue=3.135368197745318e-2
7, stderr=0.013013905893401593, intercept_stderr=6.740861711850957)
```



SUMMARY OUTPUT

Regression Statistics		
Multiple R	0.668186	coefficient of correlation
R Square	0.446473	coefficient of determination
Adjusted R Square	0.443677	
Standard Error	86.32758	
Observations	200	

	Coefficients	*Standard Error*	*t Stat*	*P-value*	*Lower 95%*	*Upper 95%*
Intercept	-6.02922	6.740862	-0.89443	0.372178714	-19.3223	7.263879
Sales	0.164463	0.013014	12.63748	3.14E-27	0.138799	0.190127

Profit = 0.164463 Sales - 6.02922

t critical 1.972017

df = n - 2 198

Lower 95.0%
-19.322
0.13879

Ho β1 = 0

Ha β1 ≠ 0

Reject Ho There is evidence of a linear relationship between sales and profit

pvalue	3.14E-27
alpha	0.05

p alpha thus reject Ho