SUPERSTORE DATASET

CONTEXT

With growing demands and cut-throat competitions in the market, a Superstore Giant is seeking your knowledge in understanding what works best for them. They would like to understand which products, regions, categories, and customer segments they should target or avoid. You can even take this a step further and try and build a Regression model to predict Sales or Profit.

This analysis aims to uncover trends in sales performance, explore the impact of discounts and shipping methods, and identify opportunities to optimize profitability and customer satisfaction.

1. **Exploratory data analysis**

Descriptive Statistics

	N	Minimum	Maximum	Mean	Std. Deviation
Sales	9994	.4440	22638.4800	229.858001	623.2451005
Quantity	9994	1	14	3.79	2.225
Discount	9994	.00	.80	.1562	.20645
Profit	9994	-6599.9780	8399.9760	28.656896	234.2601077
Valid N (listwise)	9994				

The data is divided into different columns, there are 9994 total items which is our N value. The superstore is located in the United States with different branches in different states and cities. There are different products sold categorized into furniture, technology, and office supplies. Additionally, there are different types of shipping methods, same day shipping, first class, second class and standard shipping. Customers and Products have unique identifying IDs and the sales, quantity, discount and product are also provided.

TRENDS IN THE DATASET

SALES AND PROFITS

This section aims to uncover insights into the sales and profits made by the superstore in terms of region and state. It uncovers the general sales and profit made as well as the best-

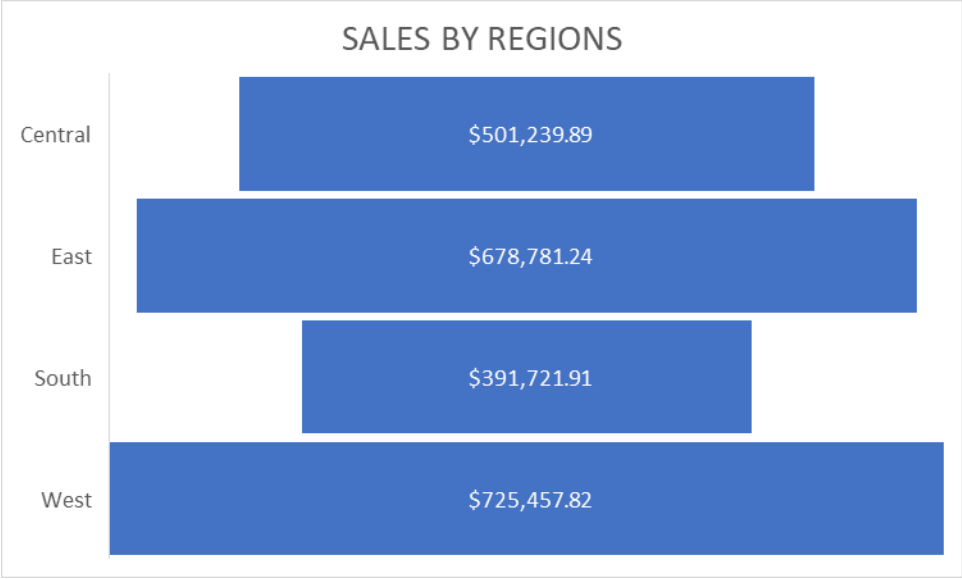
performing regions and states as well as those that aren't performing as well. It recommends solutions for the regions and states that aren't performing as well.

General findings

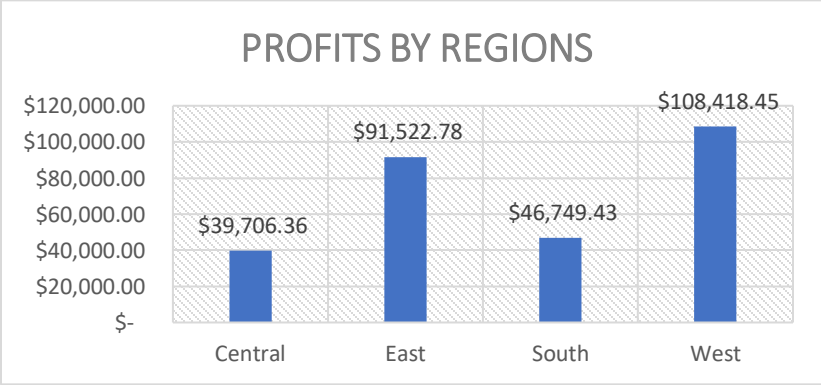
Upon analysis, I found the total amount of sales in the superstore is $ 2,297,200.86 which is approximately two million dollars, with the total amount of profit being $ 286,397.02 for each store, and the average discount is 0.15. I performed an in-depth analysis of the dataset to uncover more insights.

1. Sales and profits by region

I. Sales



II. Profits



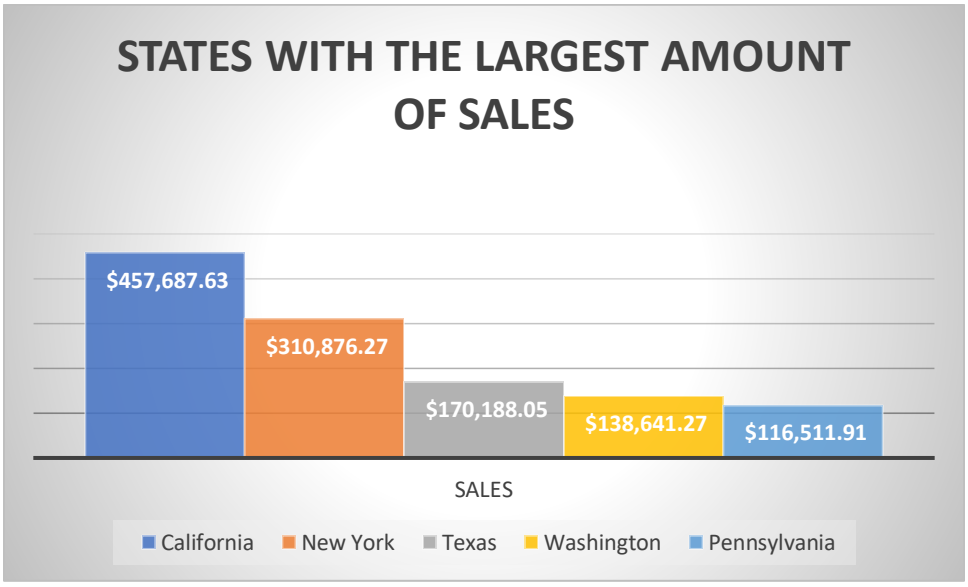
- The region with the highest amount of sales and profits is the Western region of the United States which comprises of states such as Alaska, California, and Washington. The region with the least amount of sales and profits is the southern region which is composed of states such as Alabama and Kentucky.

Key insights

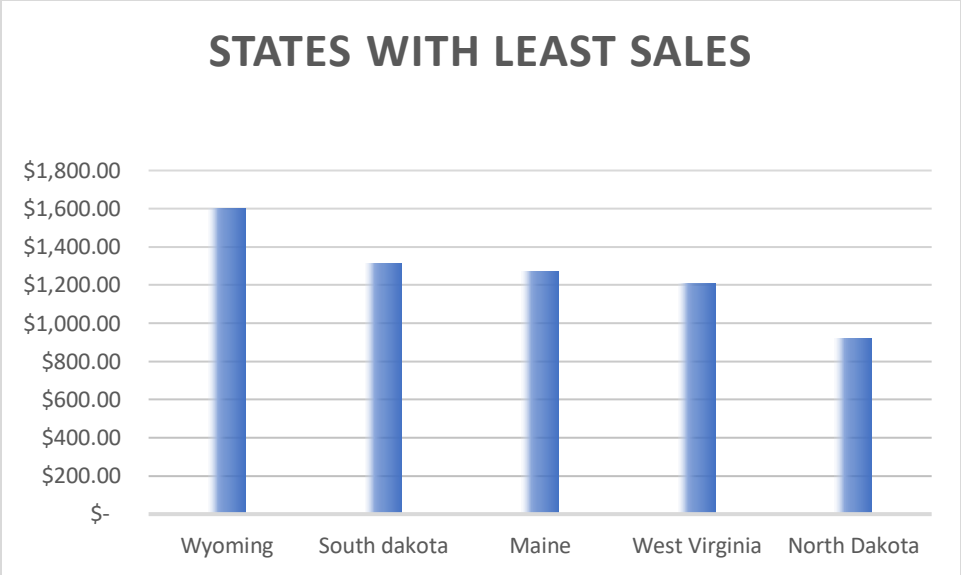
- The western region is a key driver of the stores success with the highest profits and sales margins.

2. Sales and profits by state

I. Sales



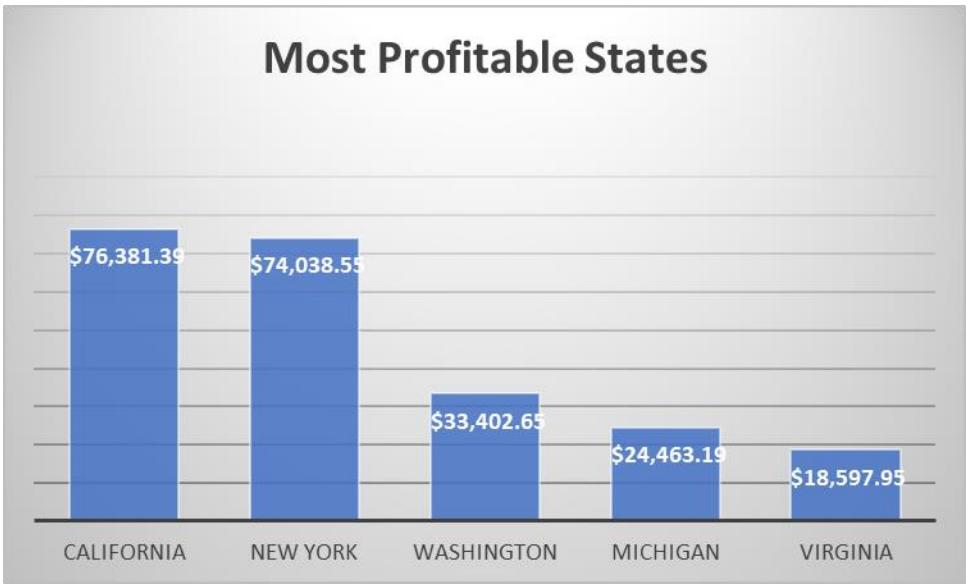
- The state of California had the largest number of sales at $457, 687 dollars. It is followed by New York, Texas, Washington ,and Pennsylvania.



- The state with the least amount of total sales is North Dakota, followed by West Virginia, Maine, South Dakota and Wyoming.

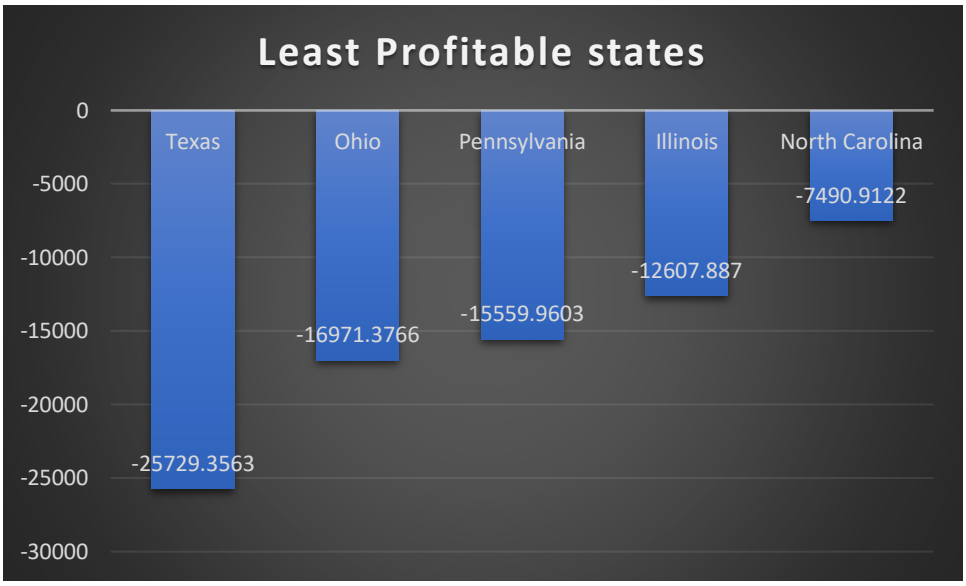
II. Profits

I used excel to identify the top 5 states in terms of profit.



- California is the best-performing state, with a margin profit of $76,381, followed by New York, Washington, Michigan, and Virginia.

The states with the least amount of sales on the other hand are:



- These states had a negative profit, meaning these stores generate losses instead of profits. The state with the greatest amount of loss is Texas with a loss of $ 25, 729 dollars. It is followed by Ohio, Pennslyvania, Illinois, and North Carolina.
- It is interesting to note that Texas and Pennsylvania are ranked among the top 5 states with most sales but these states are not making a profit and instead are making a loss. However, California, New York, and Washington are consistent in terms of profit and the amount of sales made.

<u>Recommendations</u>

- It is important for the superstore to continue investing more in the states with greater profit and sales margin i.e California, New York and Washington, these states could be key drivers of growth and continued investment could yield further returns.
- It is also important to do a root cause analysis and find out what the issue is with the states that are making a loss. It is also necessary to find out why Texas and Pennsylvania have a lot of sales but are making losses. These could be because of various reasons such as customer behavior, shipping costs and product mixes.
- Texas is also making the most amount of losses and it should be prioritized for investigation. This could be done by market research.

<u>Investigating why Texas and Pennysylvania are making a loss.</u>

Comparing discounts with the national average.

State	Discount level	National level
Pennsylvania	0.328620102	-0.178620102
Texas	0.370192893	-0.220192893

From this, we can see that these two states offer very high discount rates, these are in fact the highest discount rates in all the states, this explains why they generate a lot of sales but they are making a loss. Texas specifically, which makes the highest losses offers the highest discounts, it is more than double of the national average which is 0.15.

Calculating the profit margin to understand further.

We dive deeper to find out what exactly causes the unprofitability of texas and Pennsylvania, the discounts already show us where the problem might by lying.

The profit margin shows how much profit or loss you earn from each sale:

Calculated by profit/ sales * 100

For texas generally: It has a profit margin of - 15.12 (0.15) which means that it loses $0.15 for every dollar that it makes. The lowest profit margin is in furniture which is 0.41 which means that it loses 0.40 dollars for every dollar worth of office supplies that it sells in texas.

State	**Sum of Sales**	**Sum of Profit**	**Profit margin**
Texas	**$ 170,188.05**	**$ (25,729.36)**	**-15.12**
Furniture	$ 60,593.29	$ (10,436.14)	-17.22
Office Supplies	$ 44,490.53	$ (18,584.64)	-41.77
Technology	$ 65,104.22	$ 3,291.43	5.06

From these two investigations, it seems that texas and Pennslyvania are largely unprofitable because they offer a large discount generally and particularly on office supplies and this is what causes the irregularities in the profits and sales.

Product analysis

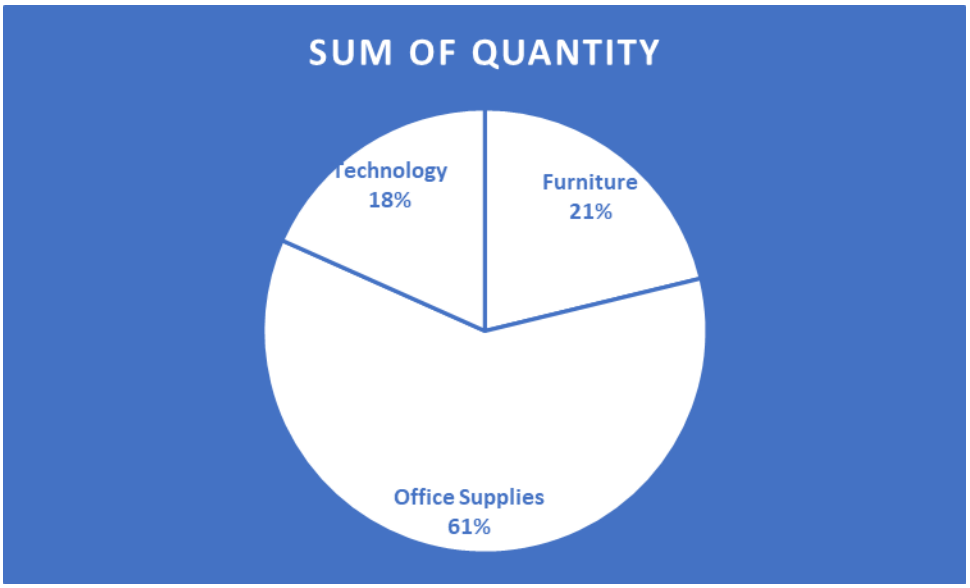
This section digs deep into different product categories to discover which products are purchased more and which products generate a lot of revenue. It also uncovers insights on different products purchased by different categories.

The products are divided into three different categories:

1. Office supplies
2. Technology
3. Furniture

The analysis below is to find out the behavior of different product categories.

1. Product categories and quantities

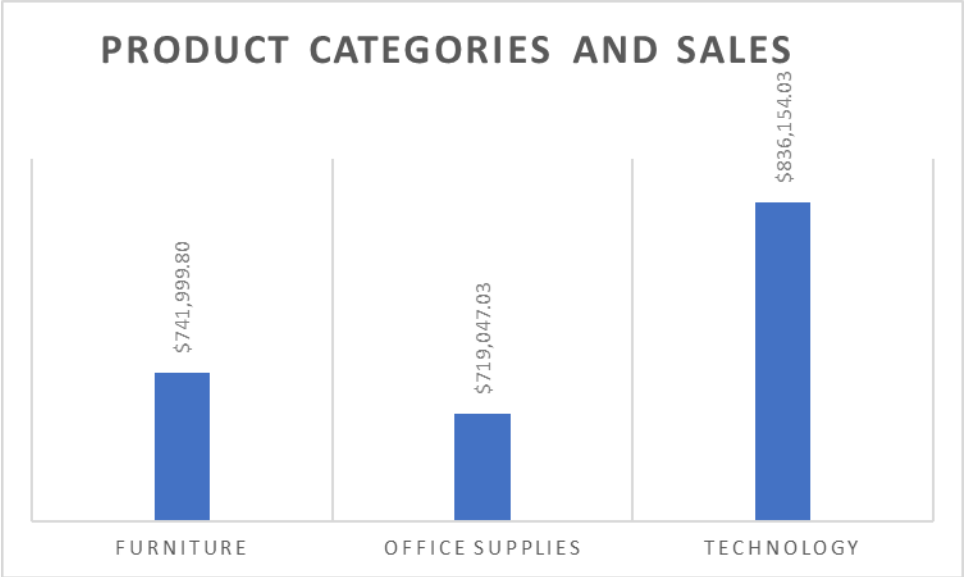


Office supplies are sold in the highest percentage at 61%. Technology, on the other hand, has the least number of quantity of products sold at 18% of the total quantity sold.

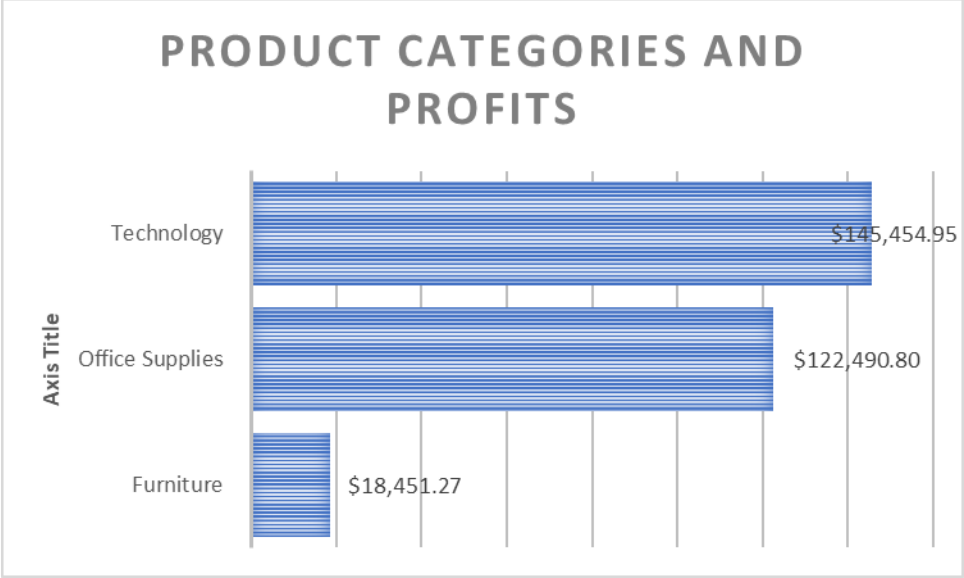
2. Product categories and sales/ profits

Product categories	Sum of Sales	Sum of Profit
Furniture	$ 741,999.80	$ 18,451.27
Office Supplies	$ 719,047.03	$ 122,490.80

Technology	$ 836,154.03	$ 145,454.95



- Technology, despite having the least quantity sold, has the highest amount of sales at $836,154 while office supplies have the highest quantity sold but the lowest amount in terms of sale at $719,047.



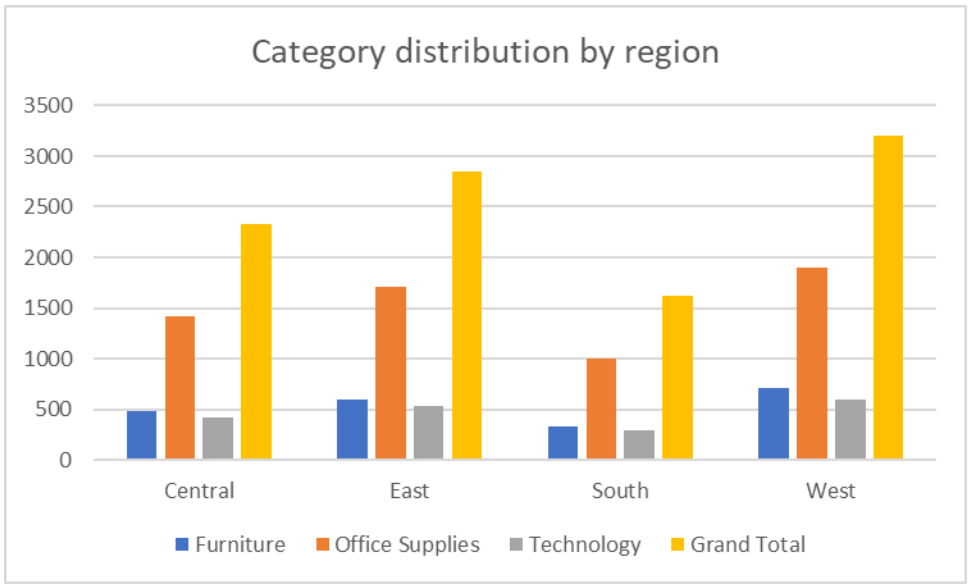
- Technological products make the highest profit at $ 145, 454 while Furniture makes a significantly lower profit of $18,451.

Key insights

- From the above, it is likely that office supplies consist of lower-value items that are sold in bulk and this explains the price vs quantity difference. It also suggests that Technology on the other hand consists of high-value items that are not sold in bulk.

3. Categories vs Regions

Region	Furniture	Office Supplies	Technology	Grand Total
Central	481	1422	420	2323
East	601	1712	535	2848
South	332	995	293	1620
West	707	1897	599	3203

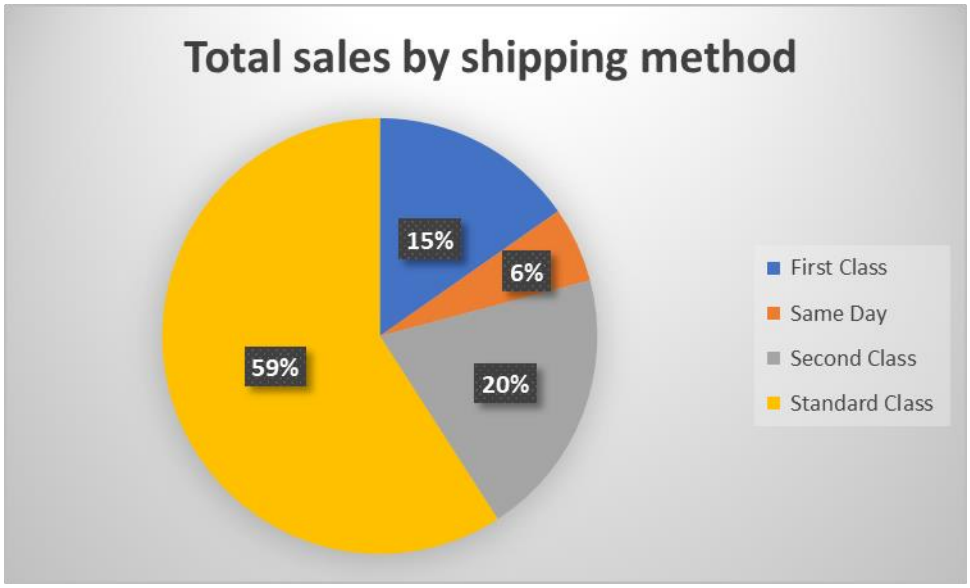


- The western region has the highest number of office supplies, furniture, and technology quantities sold – (the highest generating income product) which corresponds to the fact that it has the highest number of sales and profits, and in equal measure, the southern region also has the lowest amount of quantities sold.
- This would imply that the store should continue investing in the western region states as they are the highest profit areas.

Shipping methods

This section aims to check the effectiveness of various shipping methods, i.e Same day, first class, second class, and the standard class. It also checks the profitability and impact on sales.

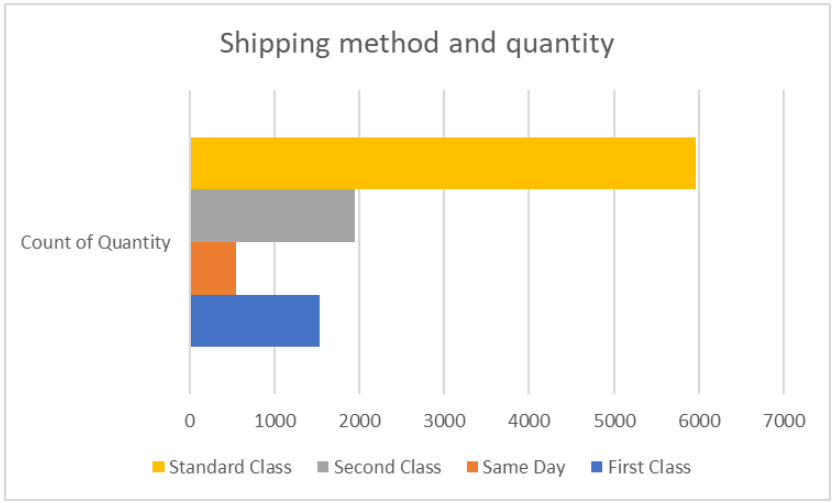
1. Sales comparison by shipping method



The shipping mode with the largest number of sales is the standard class, which suggests that most people are not comfortable or willing to pay extra prices for shipping.

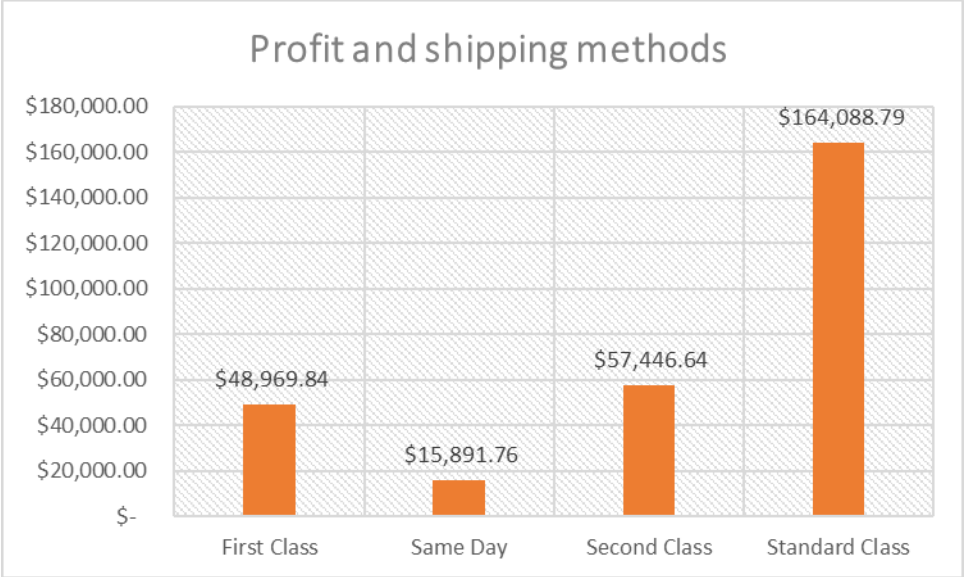
The shipping mode with the lowest number of sales is the same-day shipping method.

2. Quantity shipped and the method used



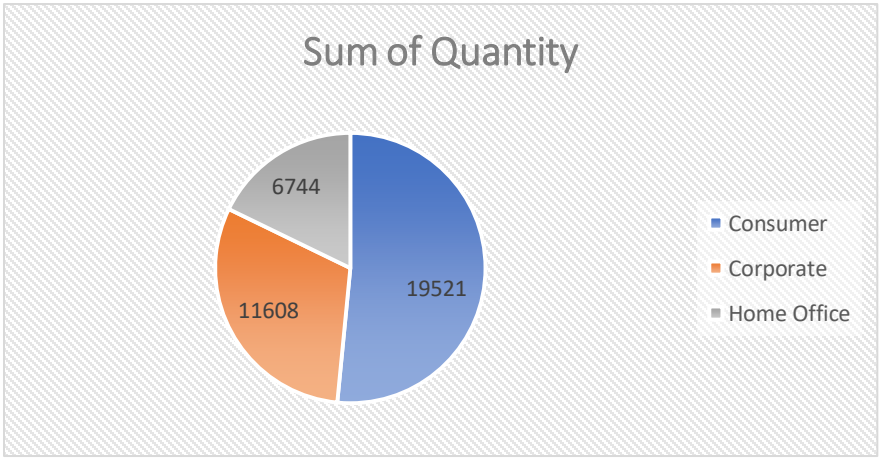
The standard class is also the shipping mode with the highest quantity of products delivered.

3. Shipping methods and profitability



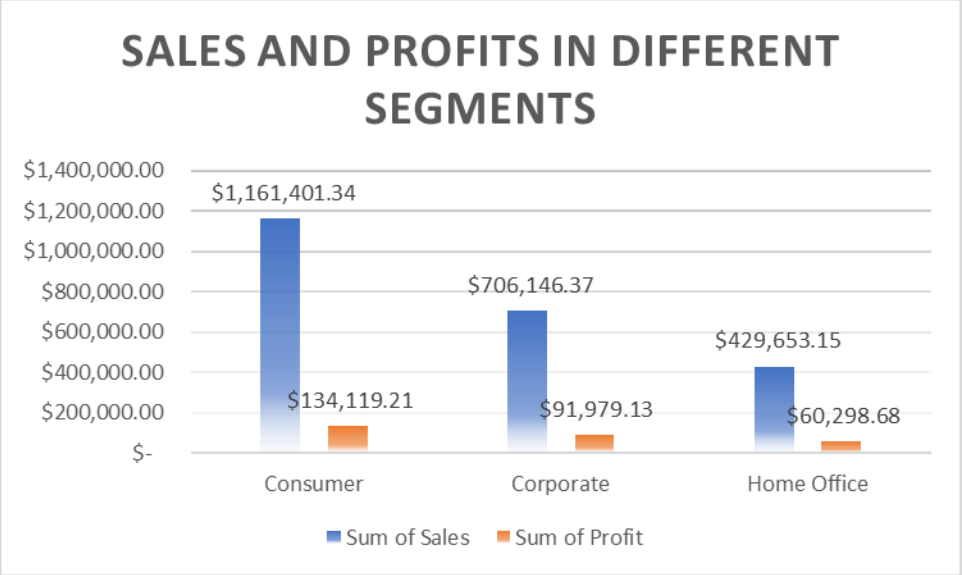
The data is consistent also in profits as it shows that the standard class generates the most amount of profit and that the same-day deliveries generate the least amount of profit.

CUSTOMER SEGMENTS, PROFITS AND SALES



- The consumer segment has the largest number in terms of quantities sold and the home office segment has the least amount in terms of quantities sold.

Customer segment	Sum of Sales	Sum of Profit	profit margin
Consumer	$ 1,161,401.34	$ 134,119.21	11.5480501
Corporate	$ 706,146.37	$ 91,979.13	13.02550552
Home Office	$ 429,653.15	$ 60,298.68	14.03426897



- From this chat, consumers have the highest sales and profits but the lowest profit margin out of the three groups. This means that consumers are a key revenue driver and should not be ignored despite the lower profit margin. If the business wishes to maximize its earnings, it should focus on the consumer segment, and it should improve its profit margin

- The home office segment despite having the least amount in sales and profits has the highest profit margin. This means that if the business aims to optimize efficiency and reduce the risk of low-margin sales, it should focus on the home office segment.

Regression model

1. To predict sales

We predict sales as the dependent variables using shipping modes, regions and product categories as the independent variables.

I categorized shipping modes, regions, and product categories to make the regression analysis easier. For the shipping method, ordinal ranking was ideal as the shipping methods are categorized according to speed.

However, for product categories and regions, since there isn't any inherent order, one hot encoding is more effective. To avoid multicollinearity, I drop the furniture category and the central region category.

ANOVA[a]

Model		Sum of Squares	df	Mean Square	F	Sig.
1	Regression	356130125.5	7	50875732.22	144.106	<.001[b]
	Residual	3525495386	9986	353043.800		
	Total	3881625512	9993			

a. Dependent Variable: Sales

b. Predictors: (Constant), Quantity, Office supplies, South , Shipping, East, Technology, West

Our ANOVA model has a significant value that is less than 0.001 which means that these results are statistically significant, and since our F- Value is also quite large at 144.156 it indicates that there is a strong relationship between the independent variables and the dependent variable.

Coefficients[a]

Model		Unstandardized Coefficients		Standardized Coefficients		
		B	Std. Error	Beta	t	Sig.
1	(Constant)	135.626	29.098		4.661	<.001
	Shipping	-3.410	6.429	-.005	-.530	.596
	East	21.918	16.621	.016	1.319	.187
	West	2.375	16.202	.002	.147	.883
	South	23.372	19.235	.014	1.215	.224
	Technology	104.286	18.911	.065	5.515	<.001
	Office supplies	-231.526	15.004	-.182	-15.431	<.001
	Quantity	56.770	2.673	.203	21.241	<.001

a. Dependent Variable: Sales

The unstandardized coefficients represent the actual change in the dependent variable for a one-unit change in the corresponding independent variable. For instance, a unit increase in Technology causes a 104.286 increase in sales, and for office sales one unit increase in Office Supplies causes a -231.526 decrease in sales.

For the t- values, a larger t value indicates a stronger relationship between the dependent and independent variables, and for instance quantity has a t-value of 21 which means that it statistically impacts sales.

The p-values show whether these results are statistically significant. Shipping values and the regions have a p-value that is greater than 0.05 which means that these values are not statistically significant.

GENERALLY:

- **(Constant)**: 135.626 - This is the expected sales value when all the other predictors are held constant.
- **Shipping**: -3.410 — A one-unit increase in Shipping (e.g., moving from one shipping category to another) is associated with a decrease of 3.410 in Sales. However, the high p-value (0.596) suggests that this result is not statistically significant, so we cannot confidently conclude that Shipping has a meaningful effect on sales.
- **East**: 21.918 — Being in the East region (compared to the reference region) is associated with an increase of 21.918 in Sales, but the p-value (0.187) is high, indicating this may not be a statistically significant result.

- **West**: 2.375 — Being in the West region (compared to the reference region) is associated with a small increase in sales (2.375), but again, the high p-value (0.883) means this is not statistically significant.
- **South**: 23.372 — Being in the South region (compared to the reference region) is associated with an increase of 23.372 in Sales, but the p-value (0.224) suggests this result is not statistically significant either.
- **Technology**: 104.286 — The Technology category (compared to the reference category, possibly "Office supplies" or "Standard") is associated with a significant increase of 104.286 in Sales. This is statistically significant because the p-value is 0.000, indicating a meaningful relationship between Technology and Sales.
- Office supplies: -231.526 — The Office supplies category is associated with a decrease of 231.526 in Sales compared to the reference category, and this is statistically significant (p-value = 0.000).
- Quantity: 56.770 — A one-unit increase in Quantity is associated with an increase of 56.770 in Sales, and this is statistically significant (p-value = 0.000).

Conclusion for Predicting Sales:

Technology and quantity are the key predictors of sales while office supplies have a negative impact on sales and the regional factors are not very significant.

Therefore, the regression model is

$Sales=\beta0+\beta1(Quantity)+\beta2(Technology)+\beta3(Office\ supplies)+\epsilon$

Coefficients[a]

Model		Unstandardized Coefficients		Standardized Coefficients		
		B	Std. Error	Beta	t	Sig.
1	(Constant)	135.237	16.391		8.251	<.001
	Office supplies	-231.429	15.001	-.182	-15.428	<.001
	Quantity	56.697	2.671	.202	21.225	<.001
	Technology	104.468	18.910	.065	5.525	<.001

a. Dependent Variable: Sales

ADDITIONAL QUESTIONS

Is there a correlation between profit and discount?

Null hypothesis: There is no correlation between profit and discount

Alternative hypothesis: There is a correlation between profit and discount

Correlations

		Discount	Profit
Discount	Pearson Correlation	1	-.219**
	Sig. (2-tailed)		<.001
	N	9994	9994
Profit	Pearson Correlation	-.219**	1
	Sig. (2-tailed)	<.001	
	N	9994	9994

**. Correlation is significant at the 0.01 level (2-tailed).

We obtain a significant value that's less than 0.001 which is less than the p-value which means that these results are significant at 0.05.

The Pearson correlation coefficient is -.219 (21.9%) which means that as profits increase the discounts reduce and vice versa.

This concludes the analysis