Data Dictionary

This report takes reference of the dataset of Superstore sales. The dataset consists of 12 columns representing 12 variables and 250 rows representing 250 customers along with the details of each of its variable.

Ship Mode: The shipping mode used for delivering the order. This dataset consists of 4 shipping modes namely, ’First Class’, ‘Standard’, ‘First Class’ and ‘Second Class’.

Customer Name: The name of the customer who placed the order.

Gender: The gender of the customer.

Segment: The market segment to which the customer belongs. This dataset has 3 market segments namely, Consumer, Corporate, or Home Office.

State: The U.S. state where the customer is located.

Region: The U.S. region where the state is located. Here region is divided into 4 parts namely West, East, Central, or South.

Category: The category of the product ordered. This dataset has 3 categories namely ‘Office Supplies’, ‘Furniture’ and ‘Technology’.

Sub-Category: The sub-category of the product, which is more specific than the main category (e.g., Chairs, Storage, Paper).

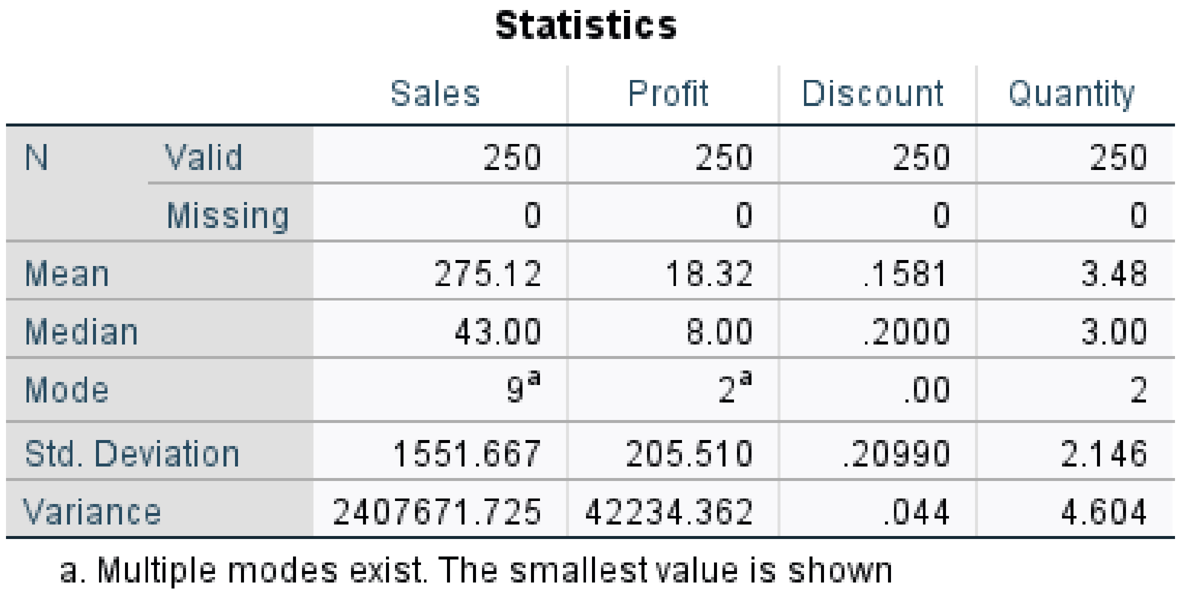
Sales: The total sales amount for the order is in USD.

Quantity: The number of units ordered.

Discount: The discount applied to the order.

Profit: The profit from the order in USD.

Measures of Central Tendecy and Variance



Statistics

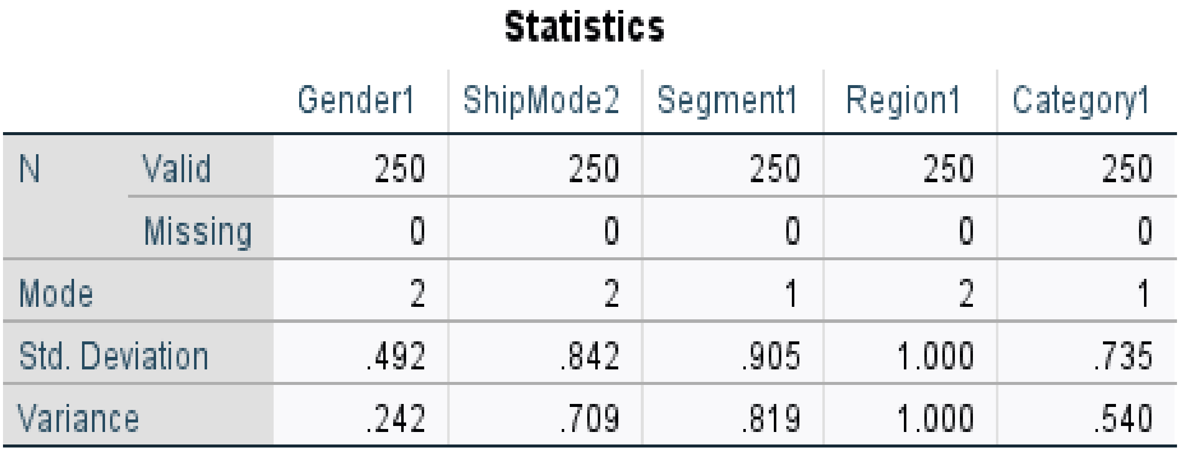
		Sales	Profit	Discount	Quantity
N	Valid	250	250	250	250
	Missing	0	0	0	0
Mean		275.12	18.32	.1581	3.48
Median		43.00	8.00	.2000	3.00
Mode		9[a]	2[a]	.00	2
Std. Deviation		1551.667	205.510	.20990	2.146
Variance		2407671.725	42234.362	.044	4.604

a. Multiple modes exist. The smallest value is shown

I formed this table through SPSS. From this table things that we should really focus on is *Mean, Standard Deviation* and *Variance* of different variables. Mean of total sales is 275.12$ and mean of total profit is 18.32$. The *standard deviation* shows the degree of dispersion of the values around the mean. A larger standard deviation indicates a *wider* range of values.
The standard deviation squared represents *variance*, which shows how widely distributed the data collection is.

My interpretation from the data:

- There is considerable variation in the sales numbers, as seen by the large standard deviation and variance of sales relative to the mean.
- Even though the profit is smaller especially when compared to sales, its standard deviation and fluctuation still reveal a significant spread.
- The low mean, median, and fairly small variance and standard deviation of the quantity indicate that most orders are for a limited number of items.



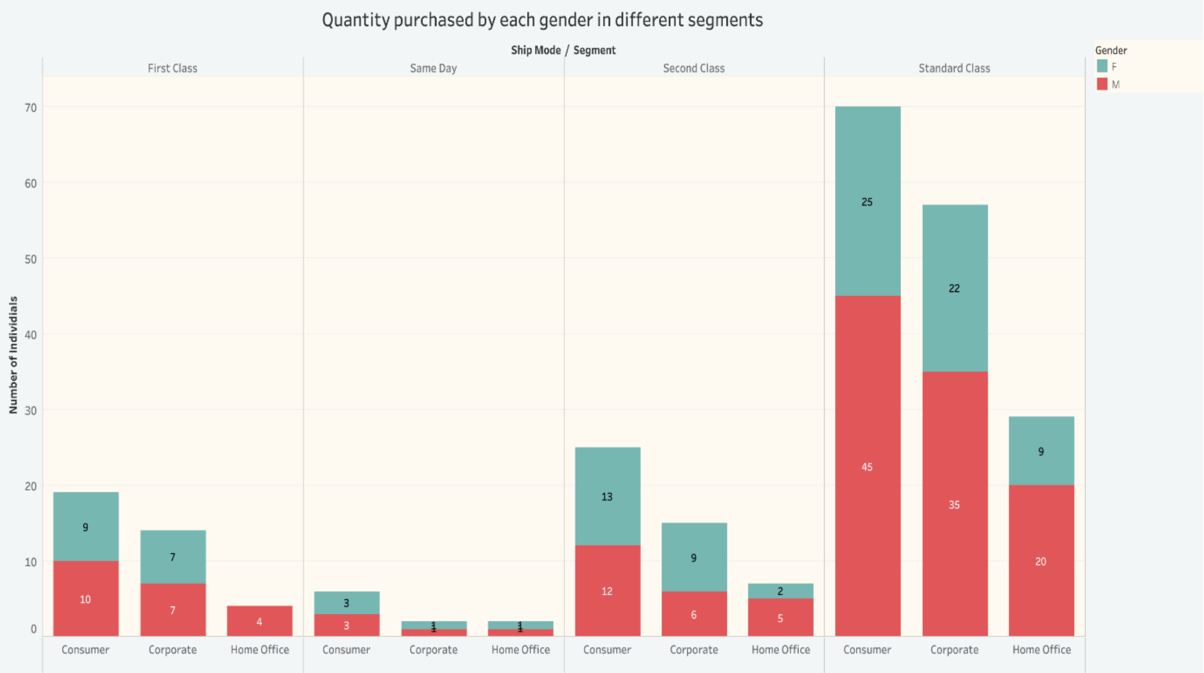
Statistics

		Gender1	ShipMode2	Segment1	Region1	Category1
N	Valid	250	250	250	250	250
	Missing	0	0	0	0	0
Mode		2	2	1	2	1
Std. Deviation		.492	.842	.905	1.000	.735
Variance		.242	.709	.819	1.000	.540

Mode is a suitabale indicator to examine these kind of variables. For example, group *'2'* in Gender represents Male. Mode of *'2'* in Gender1 represents frequency of 'Males' in total number of Transactions is more than 'Females'. Simililarly people prefer group *'2'* of shipment mode more than other shipment modes. Here group *'2'* represents 'Standard Class'.

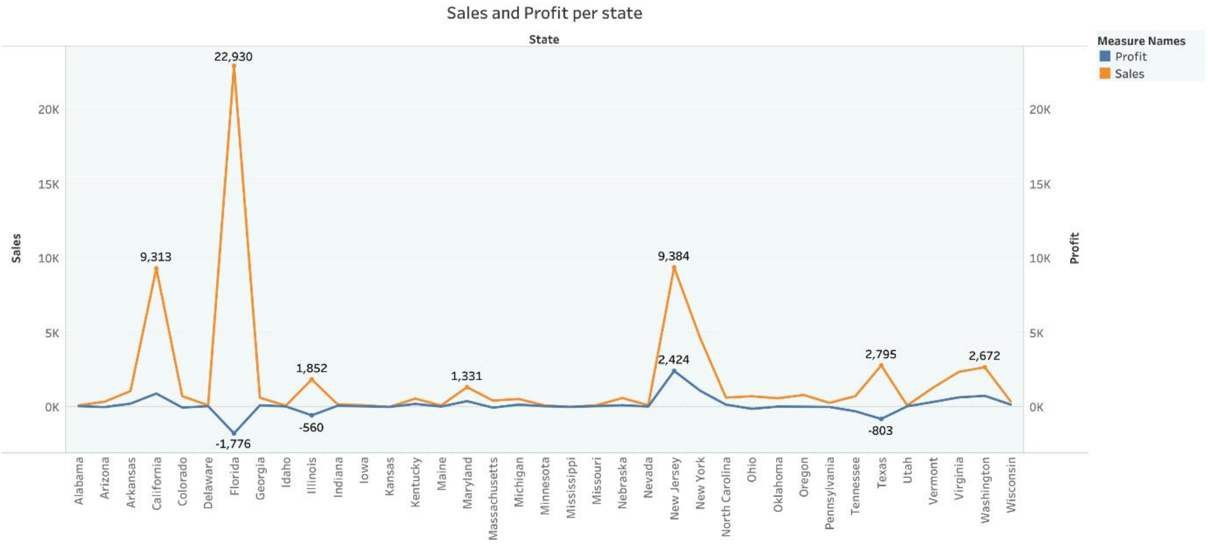
Data Visualization

1) Quantity purchased by each Gender in different segments



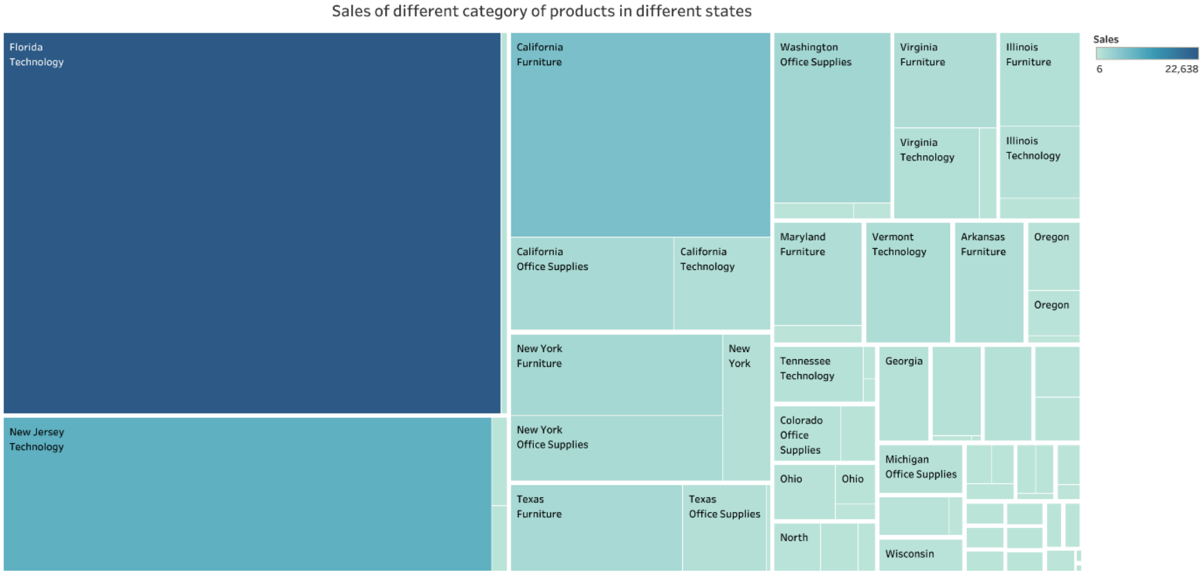
This graph made in tableau software, shows total quantity of products in different *Segment*, *Ship mode* and *Gender*. According to the visual, *'Standard Class'* is the most common shipping option, and men typically make larger purchases than women, particularly in the *'Consumer'* and *'Corporate'* groups. The least common shipping option is *'Same Day'*. In addition to pointing to possible marketing opportunities—especially with regard to targeting male consumers—this also shows price sensitivity, as seen by the preference for the presumably less expensive ‘Standard Class’ shipping.

2) Sales and Profit per state



Uneven sales and profit distribution amongst states are shown in the graph. New York and California have high sales compared to California's poor profit, which shows the influence of high cost of operation. States like Vermont and Wyoming exhibit small sale results , while Texas and Pennsylvania makes loss despite having large sales. Profits are not always directly correlated with sales.

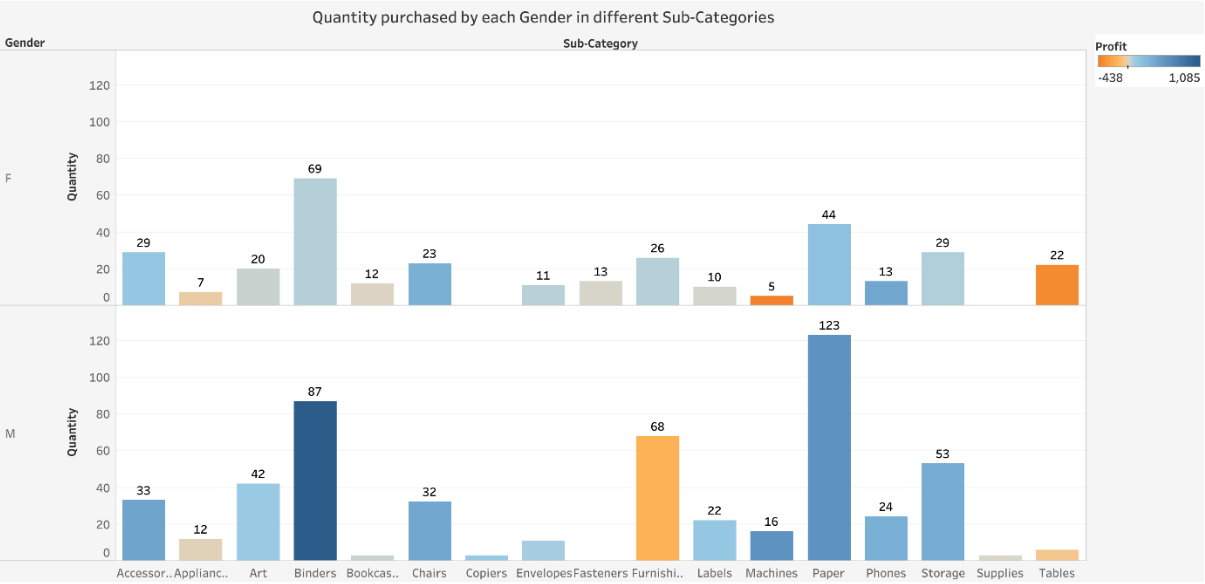
3) Sales of different category of products in different states



The *Treemap* displays sales of Technology, Furniture, and Office Supplies across various states. Greater sales are shown by larger rectangles, across all categories, California and New

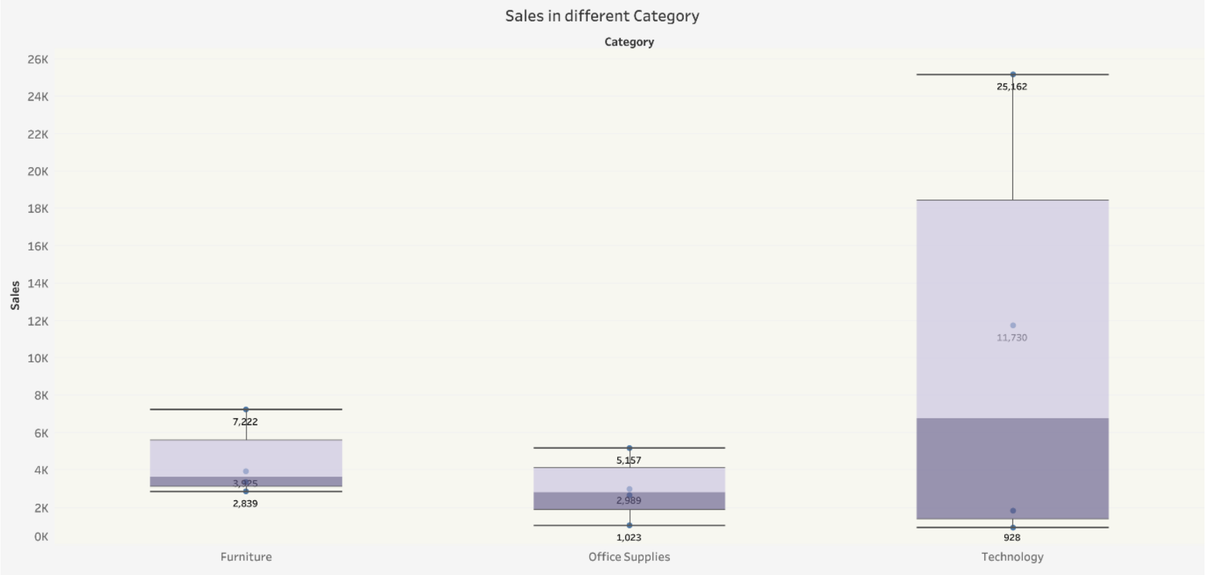
York have notable volumes. The chart indicates market strengths by state and category, which is helpful in determining important growth areas and focusing sales strategy as per different areas.

4) Sub-category wise sales as per Gender



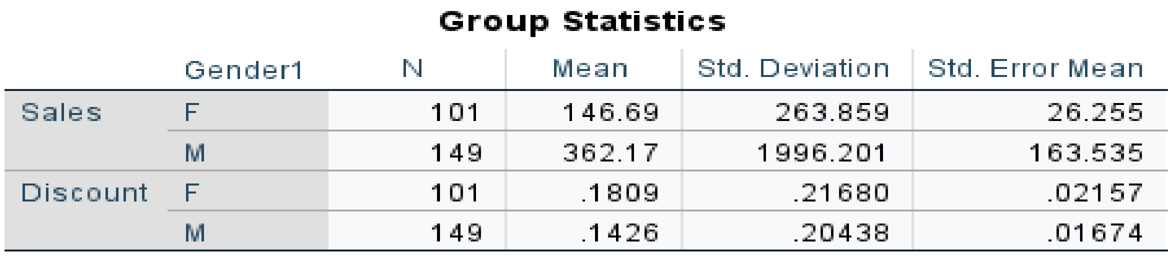
When comparing purchases by *Gender* across *subcategories*, the chart shows that men buy a lot of paper, which is quite profitable, while women buy more binders. The color grading from orange to blue indicates the profit margin in different sub-categories. *Men* buy more products overall in most categories. The information could direct gender- and product-specific focused sales and marketing campaigns.

5) Sales in different categories



The Box plot displays the revenue distributions across categories, with *Technology* having the largest revenue lead, followed by Office Supplies, Furniture, and then Furniture. While the range demonstrates variety, the *medians* represent common sales. The data indicates that the *Technology* market is performing well, which makes it a potentially profitable area of concentration for business plans.

T-Test



Group Statistics

	Gender1	N	Mean	Std. Deviation	Std. Error Mean
Sales	F	101	146.69	263.859	26.255
	M	149	362.17	1996.201	163.535
Discount	F	101	.1809	.21680	.02157
	M	149	.1426	.20438	.01674

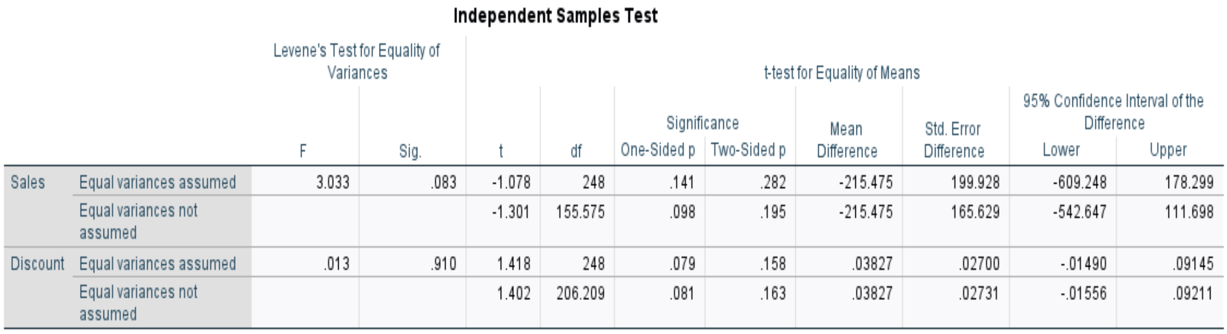
With regard to two variables, Sales and Discount, the figure displays a set of group data for two gender groups, labelled **'F'** (Female) and **'M'** (Male). For each gender group for both variables, the table shows the number of observations **(N)**, the mean, the standard deviation, and the standard error of the mean for the two variables *'Sales'* and *'Discount'*.

Sales

- The average amount of sales for men is more than that of women.
- Males had a significantly higher standard deviation than females, suggesting that there is more variation in sales, made by different male customers.
- Males have significantly higher mean sales values than females do, but the male standard deviation is also significantly larger, suggesting that the male mean may be impacted by extreme or outlier values.

Discount

- Females have a higher average discount received than males.
- The standard deviations of the discount for the two genders are almost equal.
- The mean discount shows a slight gender difference, with females receiving a higher average discount. This explain marketing tactics of company where they try to target women customers.



Independent Samples Test

		Levene's Test for Equality of Variances		t-test for Equality of Means							
						Significance		Mean Difference	Std. Error Difference	95% Confidence Interval of the Difference	
		F	Sig.	t	df	One-Sided p	Two-Sided p			Lower	Upper
Sales	Equal variances assumed	3.033	.083	-1.078	248	.141	.282	-215.475	199.928	-609.248	178.299
	Equal variances not assumed			-1.301	155.575	.098	.195	-215.475	165.629	-542.647	111.698
Discount	Equal variances assumed	.013	.910	1.418	248	.079	.158	.03827	.02700	-.01490	.09145
	Equal variances not assumed			1.402	206.209	.081	.163	.03827	.02731	-.01556	.09211

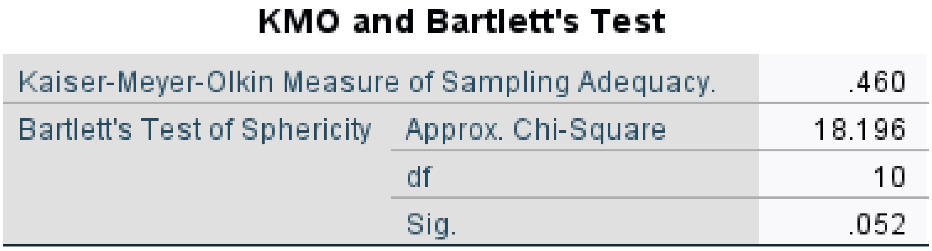
- F value of **3.033** with a significance of **0.83** determines that the two groups(Male and Female) are not statistically different with a alpha level of **0.05**. Even for discount the two groups are not much different.

- With p-value of **0.282** for Sales and **0.158** for discounts which is both greather than **0.05**, therefore there is no significant statistical difference in the sales created by males and the sales created by females for ‘Equal Variances assumed’ section.

- With insignificant p-value, we can observe that any differences in the sample means could be due to chance, rather than a true difference in the population count of both males and females.

Factor Analysis

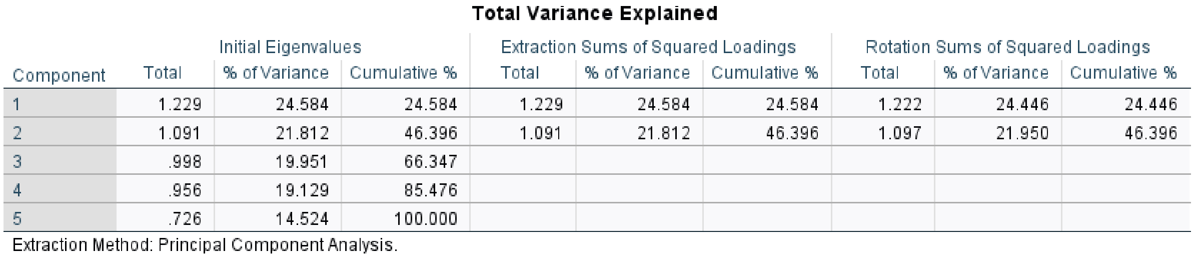
For factor analysis, this report analyzes 5 variables namely *Sales(1), Category(2), Gender(3), Discount(4) and Segment(5)* to analyze the correlation between them. Particularly, we want to identify how different factors affect the sales of the supermarket.



KMO and Bartlett's Test

Kaiser-Meyer-Olkin Measure of Sampling Adequacy.		.460
Bartlett's Test of Sphericity	Approx. Chi-Square	18.196
	df	10
	Sig.	.052

A moderately significant Bartlett's Test and an inadequate KMO score imply that factor analysis *may not be* the best technique for evaluating this dataset. It's possible that the data don't have enough significant correlations between variables or common variance to support the application of factor analysis.



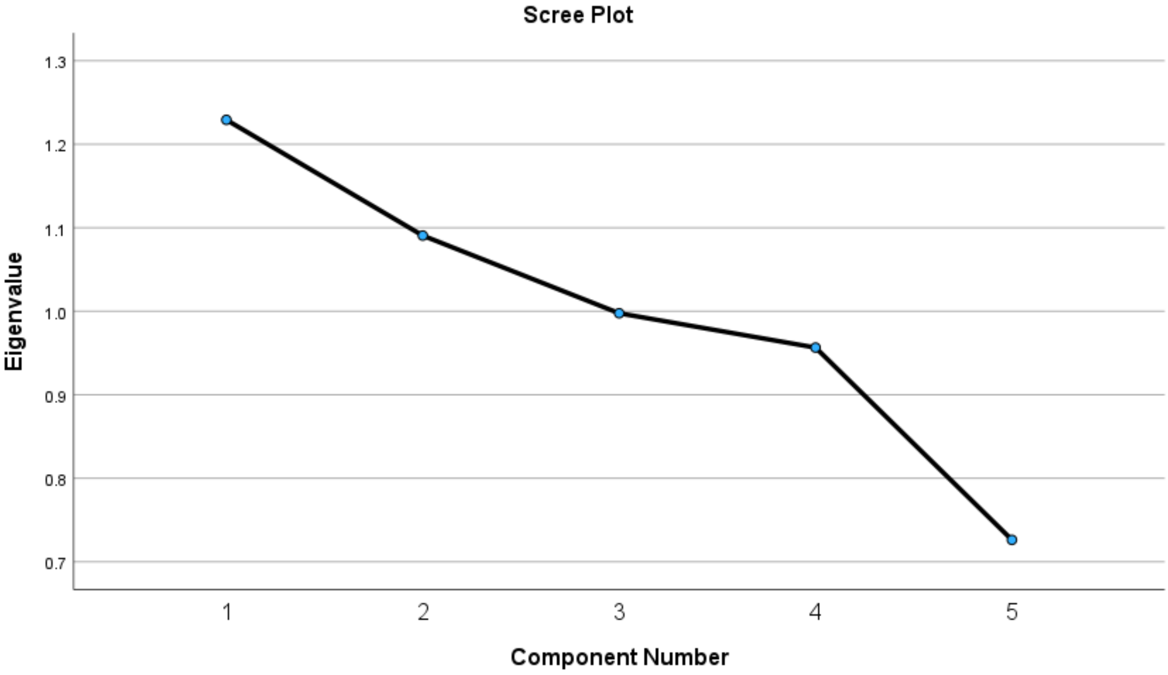
Total Variance Explained

	Initial Eigenvalues			Extraction Sums of Squared Loadings			Rotation Sums of Squared Loadings		
Component	Total	% of Variance	Cumulative %	Total	% of Variance	Cumulative %	Total	% of Variance	Cumulative %
1	1.229	24.584	24.584	1.229	24.584	24.584	1.222	24.446	24.446
2	1.091	21.812	46.396	1.091	21.812	46.396	1.097	21.950	46.396
3	.998	19.951	66.347						
4	.956	19.129	85.476						
5	.726	14.524	100.000						

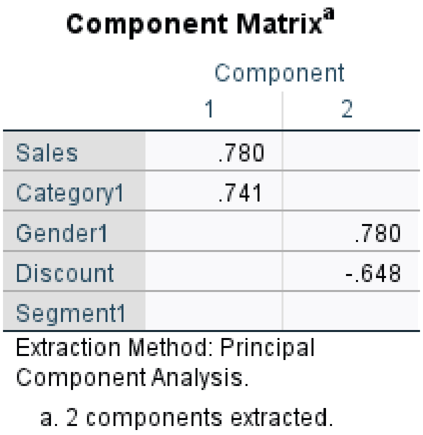
Extraction Method: Principal Component Analysis.

Components with eigenvalues *greater than 1* are often considered *significant*. Together, the first two components explain roughly *46.396%* of the dataset's variation, indicating that they don't fully capture all of the data. Since all *five* components jointly account for all of the variance in the dataset, the cumulative proportion of variance explained by all five components is 100%. This is always the case with *PCA(Principal Component Analysis)*.

These findings suggest that no single factor is predominating when it comes to explaining a significant amount of the variance. *This means all the five factors that has been taken, affects the sales of the supermarket store though to different intensities.* Depending on the complexity of the data and the purpose of the research, it is preferable to reduce dimensionality by using the first two or three components while using PCA by choosing components that cumulatively explain a significant percentage of the variation (typical cutoffs are 70-80%).



Components with eigenvalues *greater than 1* are considered significant according to the *Kaiser* criterion. Between the third and fourth components is the "elbow"—the point at which the plot begins to flatten out. Given that the first three components *(Sales, Category and Gender)* explain the majority of the change in the data before the *eigenvalues* approach 1.0 suggesting that, indicating that *4th(Discount)* and *5th(Segment)* components offer less information, it would seem sensible to analyze the first three components.



Component Matrix[a]

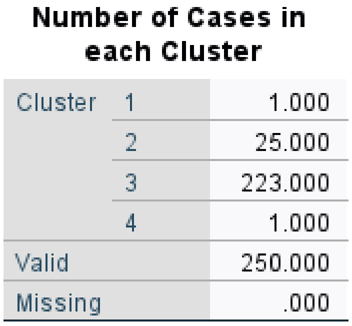
	Component	
	1	2
Sales	.780	
Category1	.741	
Gender1		.780
Discount		-.648
Segment1		

Extraction Method: Principal Component Analysis.

a. 2 components extracted.

The matrix facilitates interpretation of the variables that are the primary drivers of variation. Component 1 is more closely associated to *sales* and *category*, whereas Component 2 is more closely related to *gender*, according to the component matrix, which implies that these components are capturing distinct patterns of variance in the data set.
Since, Segment1 is not present in the matrix, it is possible that its link with the first two primary components is not very strong.

Clustering

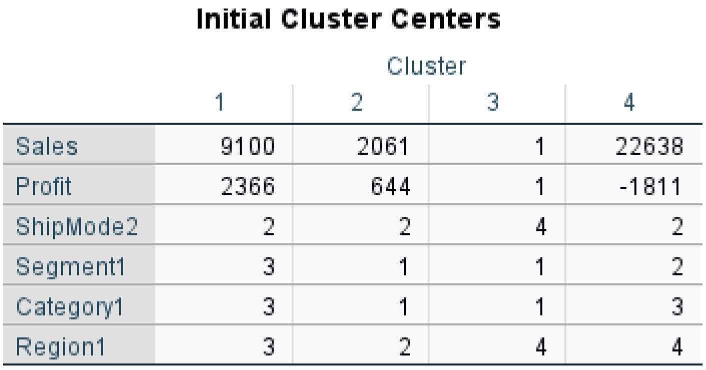


Number of Cases in each Cluster

Cluster	1	1.000
	2	25.000
	3	223.000
	4	1.000
Valid		250.000
Missing		.000

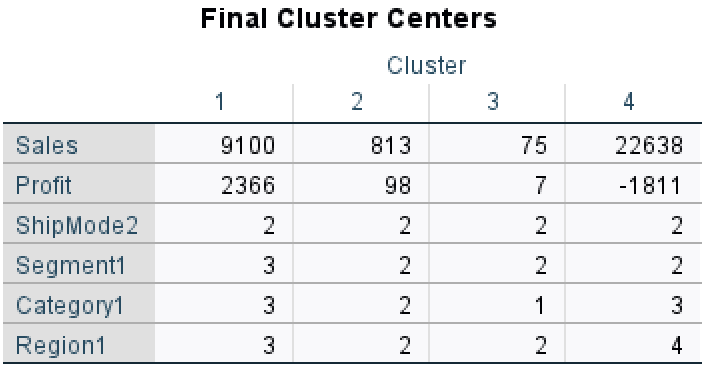
This table shows total nuber of cases in each cluster. Cluster 1 and 4 has just 1 case each while majority of cases quoting *223* fall into cluster *3*. Cluster 2 has *25* casees. There is a significant disparity in the distribution of cases among the clusters; *223* out of *250* cases fall into Cluster 3. There is just one case in each of Clusters 1 and 4, which may point to outliers or singular observations that did not fit in with the other observations, or it may point to a problem with the clustering procedure. There are a fair amount of cases in Cluster 2. This *unequal* distribution could potentially reduce the comparative effectiveness between clusters 1, 2, and 4 compared to cluster 3, which could impact the *validity* and *reliability* of any research conducted on the clusters.

From the given dataset, this report shows 4 clusters. 6 variables from the dataset namely Sales, Profit, ShipMode2, Segmen1, Category1 and Region1 are been used to analyze these clusters. Now we will try to analyze and interpret these 4 clusters.



Initial Cluster Centers

	Cluster			
	1	2	3	4
Sales	9100	2061	1	22638
Profit	2366	644	1	-1811
ShipMode2	2	2	4	2
Segment1	3	1	1	2
Category1	3	1	1	3
Region1	3	2	4	4



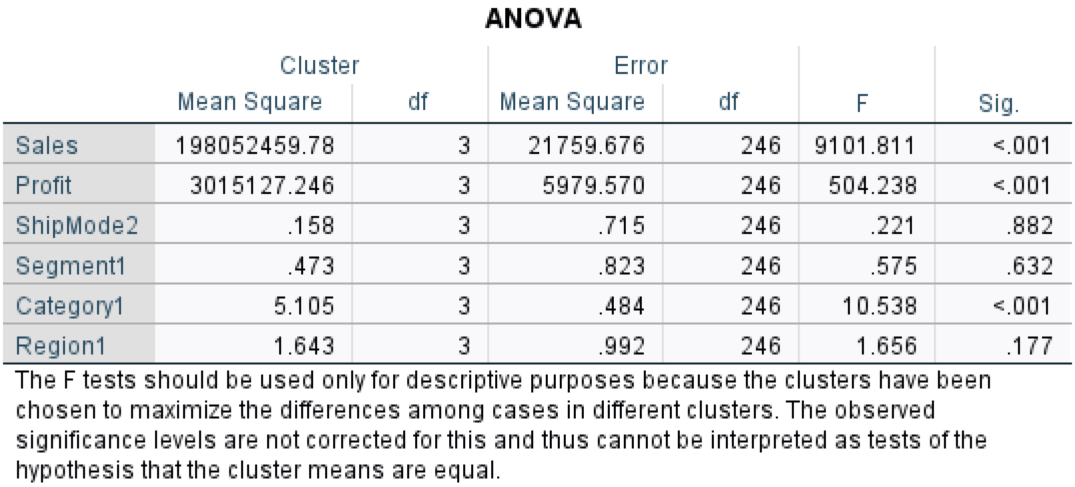
Final Cluster Centers

	Cluster			
	1	2	3	4
Sales	9100	813	75	22638
Profit	2366	98	7	-1811
ShipMode2	2	2	2	2
Segment1	3	2	2	2
Category1	3	2	1	3
Region1	3	2	2	4

The clustering appears to have identified a range of transaction types within the data, from highly profitable to loss-making, with distinct characteristics in terms of sales and profit. The *Sales* and *Profit* variables, which are probably the most important in deciding cluster membership, show the biggest differences between the initial and final cluster centers.

In terms of sales and profit, Clusters 1 and 4 stand out from the others, with Cluster 1 being the most profitable and Cluster 4 being loss-making. With Cluster 2 being average and Cluster 3 being *smaller-scale*, Clusters 2 and 3 may represent more common transactions.

Various business strategies can be formed by the clustering, such as fixing the losses in Cluster 4 or concentrating on growth in Cluster 1. When analyzing these clusters and putting any actionable recommendations into practice, it is crucial to take business knowledge and other data characteristics into account.



ANOVA

	Cluster		Error			
	Mean Square	df	Mean Square	df	F	Sig.
Sales	198052459.78	3	21759.676	246	9101.811	<.001
Profit	3015127.246	3	5979.570	246	504.238	<.001
ShipMode2	.158	3	.715	246	.221	.882
Segment1	.473	3	.823	246	.575	.632
Category1	5.105	3	.484	246	10.538	<.001
Region1	1.643	3	.992	246	1.656	.177

The F tests should be used only for descriptive purposes because the clusters have been chosen to maximize the differences among cases in different clusters. The observed significance levels are not corrected for this and thus cannot be interpreted as tests of the hypothesis that the cluster means are equal.

Mean Square (Cluster): For the variable under consideration, this is the variance (mean of the squares of the departures from the mean) within each group or cluster.

F: The variance obtained for the groups (Mean Square Cluster) divided by the variance within the groups (Mean Square Error) is the F-statistic. It's employed in p-value determination.

***Sig.**:* The F-statistic's significance level, which indicates the likelihood that the observed results are an outcome of chance. A typical significance criterion is set at 0.05.

The analysis reveals significant variations among clusters in *Sales* and *Profit*, supported by *high* F-statistics (9101.811 and 504.238, respectively) and p-values < .001. Conversely, ShipMode2 and Segment1 exhibit no significant differences between clusters, indicated by *low* F-statistics (.221 and .575, respectively) and high p-values. Category1 displays a significant discrepancy (F-statistic: 10.538; p < .001), while Region1 shows *no* significant variance (F-statistic: 1.656; p = .177).
Higher F-values in an ANOVA study suggest that the *mean differences* are probably real. The F-value analyzes variances between groups to evaluate mean differences. High *F-values* for sales and profit indicate significant variations between clusters. *ShipMode2* and *Segment 1* show negligible differences, with *F-values* at or below 1. *Category 1* is in the midst. Higher *p-values* for ShipMode2, Segment1, and Region1 suggest randomness, but low *p-values* for Sales, Profit, and Category1 support substantial differences. This aligns with their low F-values, indicating insignificant clustering effects.