**Superstore project**

1. **The project description.**

The project includes visualization and analysis of data from an online superstore. Our goal is to gain insights into the performance of the store and identify opportunities for improvement. We need to identify patterns in the data in terms of customer segments, product categories, sales ,etc. We will use various data visualization tools to explore the data and draw meaningful conclusions. Through this project, we hope to provide valuable insights that can help the store optimize its operations and drive growth.

1. **The dataset and variables description.**

* The dataset has 21 columns whose descriptions are given below. Each row represents a product in an order placed by the customer. Don't confuse each row for a separate order. We will treat Orders, Customers, and Products as our main entities and perform our analysis accordingly.
* The defined entities have attributes associated with them like Order ID, Order date, Customer Name, Customer ID, Category(of product), Sub-category(of product) etc.
* Our analysis will mostly revolve around these attributes to get a better understanding of the data
* Dataset has four Shipping modes: 'Second Class' 'Standard Class' 'First Class' 'Same Day'
* It covers data from 531 cities, 49 states of United States divided into 4 regions: 'South' 'West' 'Central' 'East'
* 4 shipping modes are available: 'Second Class' 'Standard Class' 'First Class' 'Same Day'
* Customers have been divided into 3 segments based on the type of account or the place where the order is being delivered. The segments are: **Consumer, Corporate, Home Office**
* The Ship mode distribution for each segment is plotted. We can see that for all ths segments Standard class has the highest occurence followed by second class, first class and same day respectively.
* Number of customers and total sales for a segment have been plotted. We see a similar pattern in both the cases. Consumer Segment has the highest number of customers and sales followed by Corporate and Home office respectively.

Standard class contributes for about 60% of the total orders. First class(19.2%) and Second class(15.7%) also have a bunch of customers. 5.3% of orders were same say orders.

* West and East are the top regions(thanks to states like California and Texas) while the south region that contributes the least sales still makes up 16.4%. This reveals that the sales have uniform geographical distribution.
* 58.2% of the products from the product catalogue belong to office supplies. We also know that technoloogy contributes the highest profits.This suggests that technology products are more profitable compared to office supplies. Furniture has the least amount of products and also generates the least amount of profit
* Products have been distributerd into 3 Categories and 17 sub\_categories. A new attribute profit ratio has been created by dividing profit by sales. It gives us better insights into the profitability of a category. It represents profit generated for every dollar of sale.
* Copiers in general are the highest profit generating products. They have the least variety in offerings and are rarely sold in bulk.

### **Metadata**

* Row ID => Unique ID for each row.
* Order ID => Unique Order ID for each Customer.
* Order Date => Order Date of the product.
* Ship Date => Shipping Date of the Product.
* Ship Mode=> Shipping Mode specified by the Customer.
* Customer ID => Unique ID to identify each Customer.
* Customer Name => Name of the Customer.
* Segment => The segment where the Customer belongs.
* Country => Country of residence of the Customer.
* City => City of residence of Customer.
* State => State of residence of the Customer.
* Postal Code => Postal Code of every Customer.
* Region => Region where the Customer belongs.
* Product ID => Unique ID of the Product.
* Category => Category of the product ordered.
* Sub-Category => Sub-Category of the product ordered.
* Product Name => Name of the Product
* Sales => Sales of the Product.
* Quantity => Quantity of the Product.
* Discount => Discount provided.
* Profit => Profit/Loss incurred.

1. **The problem definition and project objectives.**

With growing demands and cut-throat competition 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.**

**So we need to detect the** top selling products in the superstore , is the sales trend over time (monthly, yearly) , category of products generates the highest revenue and profit , region generates the most sales, the impact of discounts and promotions on sales , average profit margin for each product category and Which sub-category of products has the highest demand.

1. **The data visualization graphics with observations and interpretations of each chart.**

Chart, bar chart

Description automatically generated

As we can see above graph gives information about top 10 states in US where most orders are placed.

**The state of california and New-york city has the biggest total values sales and profit, we can't conclude one commercial activity will have a better profit in New-York City because the data isn't equally reparted but it means the State of California and New-York City have a big commercial activity.**

Chart, pie chart

Description automatically generated

As you can see that most of the customers come from the west region of us.

**Sales and Profit for different segments**

A picture containing text, screenshot, colorfulness, diagram

Description automatically generated

As we see that the most profitable segment is the consumer

Chart, pie chart

Description automatically generated

as we see that the consumer is the most participant segment

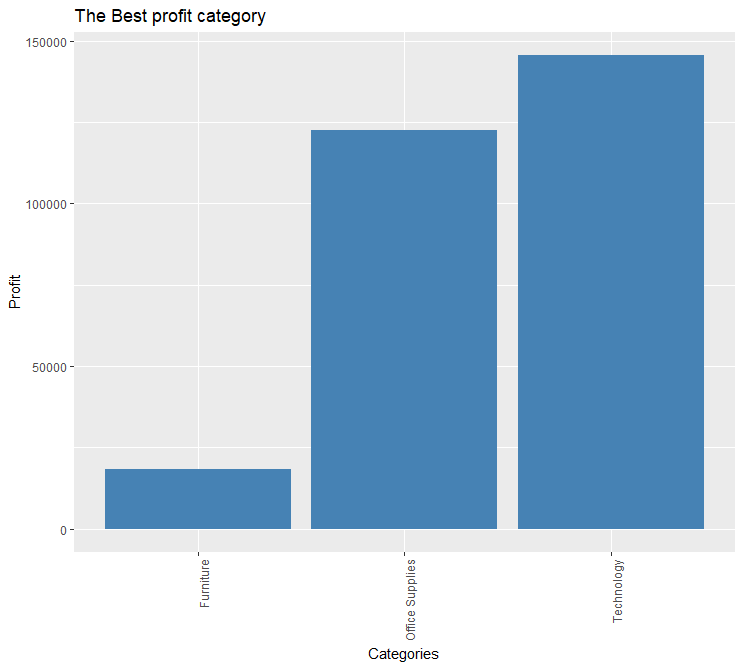
Chart, scatter chart

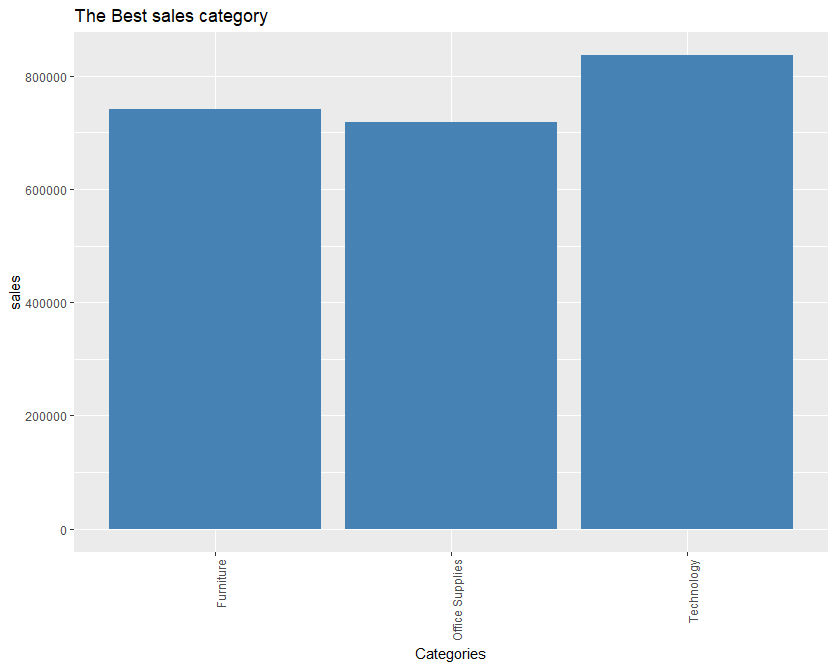
Description automatically generated

1. Profit tends to go up when sales is higher
2. Discount really affect the profit. The bigger discount gives minus profit to the store
3. Discount is not directly proportional to the sales. As you can see there are still many sales

even though there are no discount applied.

We Conclude That Discounts More Than 20% Results a Negative Profit , Therefore Our Discounts Shouldn’t Be More Than 20% In Order to Maintain Our Sales & Profit .





As we see that the best profitable and the highest level of sales is the technology category but furniture has highest sales than profit. So we can conclude that not all high profitable categories achieve highest sales.

Chart, bar chart

Description automatically generated

The highest product of sales is staple envelope

Chart

Description automatically generated

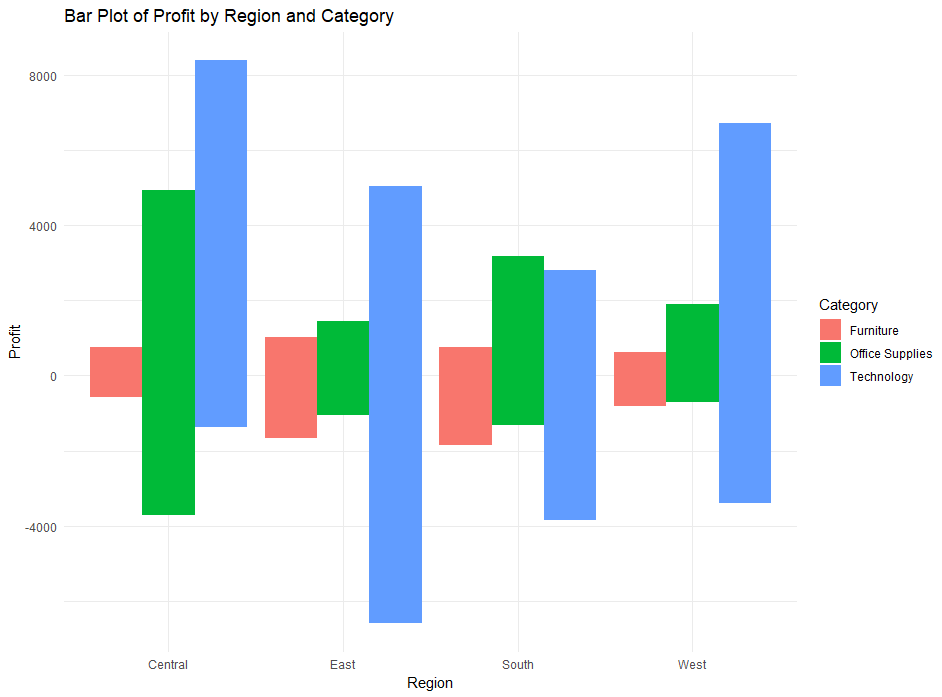
We Can Apply Discounts On These Products In Order to Increase Its Sales & Obtain A Satisfactory Profit From Them. The least product of sales is newil 342.

**Products distribution in the regions**

A picture containing text, screenshot, diagram, colorfulness

Description automatically generated

We can conclude that the count of ordered categories are very high at west region and of office supplies segment. And the type of category ordered differs according to the region.



The central region has the highest profitable result and also technology category is the highest profitable category in all regions.

A picture containing text, diagram, black and white

Description automatically generated

A picture containing text, screenshot, diagram, font

Description automatically generated

A picture containing text, screenshot, font

Description automatically generated

A picture containing text, line, screenshot, diagram

Description automatically generated

1. **Any applied data cleaning or transformation methods (if applied).**

* **Identify Missing Values**
* **Identify data types :**

**so in case of different data so we can detect it,**

As we can see there are 3 types of datatypes.

1. object: data is in categorical format like Order ID, Product Name, State, etc.
2. int64: data is in numerical format like Postal code, Quantity, row ID.
3. float64: data is in decimal format like Sales, Discount, Profit.

* filling gaps by average :

as by calculation of mean and average of profit, discounts and sales columns and fill each gap in each column with this value.

* convert Ship\_Date column to the date:

We can see that Ship Date and Order Date are in object format so we have to convert it into datetime format.

* summarizing data : it reviews the maximum, minimum, median ,mean and length of each column.

1. **Any applied Hypothesis and the test results and interpretations (if applied).**

Determine the correlation between the two variables -0.2194875

indicate the relation between dependent and independent values.

This number indicates a strong negative linear regression.

1. **The dataset preparation in terms of machine learning (training set, learning set).**

The training data is used for building a model, while the testing data is used for making predictions. This means after fitting a model on the training data set, finding of the errors, and minimizing those errors, the model is used for making predictions on the unseen data which is the test data.

Use80% of the dataset as a training set and 20% as a test set

The split method splits the data into train and test datasets with a ratio of 0.8 This means 80% of our dataset is passed in the training dataset and 20% in the testing dataset.

The train dataset gets all the data points after split which are 'TRUE' and similarly the test dataset gets all the data points which are 'FALSE'.

Train set is 7614 of 21 column

And the test set is 2380 of 21 column

1. **The used data analytics techniques (e.g. ID3, Regression, Apriori, Kmeans, etc.), with justification for your choice.**

Linear regression analysis is used to predict the value of a variable based on the value of another variable. The variable you want to predict is called the dependent variable. The variable you are using to predict the other variable's value is called the independent variable.

Lm() function in R provides us the linear regression equation which helps us to predict the data. It is one of the most important functions which is widely used in statistics and mathematics. The only limitation with the lm() function is that we require historical data set to predict the value in this function. But we can’t treat this as any limitation because historical data is a must if we have to predict anything. Historical data shows us the trend and with the help of a trend, we can predict the data.

summary gives the summary result of training model , the performance metrics r2 and rmse obtained helps us to check how well our metrics is performing

Regression equation

**Y = β1 + β2X + ϵ**

Where β1 is the intercept of the regression equation and β2 is the slope of the regression equation. β1 & β2 are also known as regression coefficients. ϵ is the error term.

* **β1:**Intercept of The Regression Equation
* **β2:** Slope of The Regression Equation
* **Y:** Dependent Variable
* **X**: Independent Variable

Call:

lm(formula = Profit ~ Discount, data = train)

Residuals:

Min 1Q Median 3Q Max

-6493.9 -54.4 -16.0 9.4 8332.5

Coefficients:

Estimate Std. Error t value Pr(>|t|)

(Intercept) 67.431 3.179 21.21 <2e-16 \*\*\*

Discount -247.865 12.347 -20.07 <2e-16 \*\*\*

---

Signif. codes: 0 ‘\*\*\*’ 0.001 ‘\*\*’ 0.01 ‘\*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 220.9 on 7612 degrees of freedom

Multiple R-squared: 0.05028, Adjusted R-squared: 0.05016

F-statistic: 403 on 1 and 7612 DF, p-value: < 2.2e-16

1. **The performance measures used and the evaluation of the analytical technique.**

Linear Regression is a supervised learning algorithm used for continuous variables. The simple Linear Regression describes the relation between 2 variables, an independent variable (x) and a dependent variable (y). The equation for simple linear regression is \*\*y = mx+ c\*\*, where m is the slope and c is the intercept. In Linear regression, a scatter plot is plotted between the x and y initially and a best fit line is drawn over it. The model is then trained and predictions are made over the test dataset(y\_pred) and a line between x and y\_pred is fitted over. The difference between the actual values and the fitted values is known as residual values or errors / RESIDUAL SUM OF SQUARES (RSS) and this must be as low as possible. To keep RSS minimal, there are two methods used i.e - OLS (ordinary least square) - Gradient descent method. The accuracy of this model is checked using the \*\*performance metrics\*\* R squared and RMSE -root mean squared error. R squared ranges between 0-1 and must be as high as possible as it represents the proportion of information in the data that can be explained by the model. RMSE (28.26363)

determines how far the predicted data points are from the actual data points on the best fit line. In this recipe, a dataset where the relation between the profit and discount, is to be determined using simple linear regression.

1. **Discussion/Quantification for relevant project findings for your project.**

Conclusion

* **Discounts of more than 20% lead to loss of profits, so they must be avoided in order to preserve our sales and profits.**
* The best-selling does not necessarily mean that it is the most profitable.
* Office Supplies are the most popular among customers.
* While all categories achieve good high sales and profits, furniture products did not achieve high profits.
* 18.7% of the total number of orders achieved negative profits (1871 orders out of a total of 9994 orders).