INFO H – I590 :
Intro to Statistics in R (Project)

Super Store Data Analysis

Index

Purpose of Analysis

Analyzing superstore data serves several purposes, and businesses engage in this practice to gain insights, make informed decisions, and optimize their operations.

- **Improving Sales and Revenue**
- **Enhancing Customer Experience**
- **Competitive Analysis between Regions**
- **Supply Chain Optimization**
- **Market Basket Analysis**
- **Risk Management**

Data Documentation

- Dataset is a US Super Store Data for the years starting from 2014 – 2017.
- It has about 21 columns containing information of each order placed by a customer:

  > Ship Mode - Same day: >=0 days , First class: >=2 days , Second class: >=3 days , Standard class: >= 5 days .

  > Segment - product sector (Consumer, Corporate and Home Office)

  > Country – only 1 i.e. United States (redundant data)

  > City - data records about 531 cities.

  > State - 49 states in this dataset.

  > Postal Code - 631 postal codes present in the dataset

  > Region - 4 regions in the dataset. i.e. South, West, Central and East

  > Category - 3 categories i.e. Furniture, Office Supplies and Technology

  > Sub-Category - 17 sub-categories, they are - Bookcases, Chairs, Labels, Tables, Storage, Furnishings, Art, Phones, Binders, Appliances, Paper, Accessories, Envelopes, Fasteners, Supplies, Machines, Copiers.

  > Order Date - 1237 unique order dates, over the years order have increased

Data Documentation

>Order ID – Unique order identification number

> Ship Date - 1334 unique dates when shipping took place.

>Customer ID – There are 793 customers in the Superstore dataset.

> Customer Name - Count of Customer.ID and Customer. Name match i.e. 793, which states that data is not having issues. Because each Customer ID / Name is unique.

> Product Name

> Product ID :-1862 Unique Products are purchased in the orders placed, this can be figured from the unique id of each.

> Sales - It is the price. Having a range starting from $0.444 to $22638.480

> Profit - It is the profit earned from the product. Can also indicate loss with a negative sign. Having a range -6599.978 to 8399.976

> Discount - Discount given for the product purchased, range starts from 0 to 0.8.

> Quantity - The quantity of products bought, range starts from 1 to 14.

Data Transformation

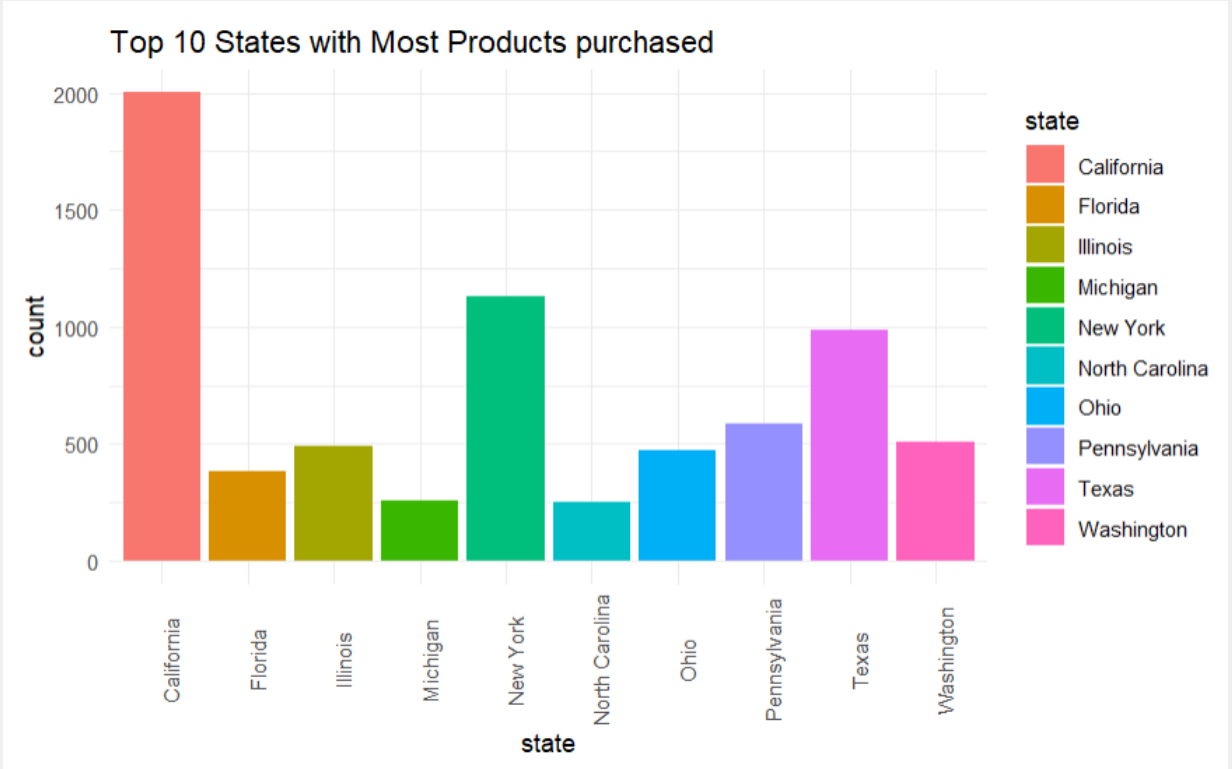
- The data didn’t require much cleaning. As there were no duplicate rows.
- Only few things that were required to be done –
  1. To understand the unique value of each column, pre-processing was done. Calculated the distinct count and unique values.
  2. Further, data type of certain columns like Order Date and Ship Date were transformed from <char> to <date>.
- To understand Data better, we can ask few questions.

Exploratory Data Analysis

which are the top 10 states that have purchased the products from the Superstore?

state <chr>	count <int>
California	2001
New York	1128
Texas	985
Pennsylvania	587
Washington	506
Illinois	492
Ohio	469
Florida	383
Michigan	255
North Carolina	249

1-10 of 10 rows

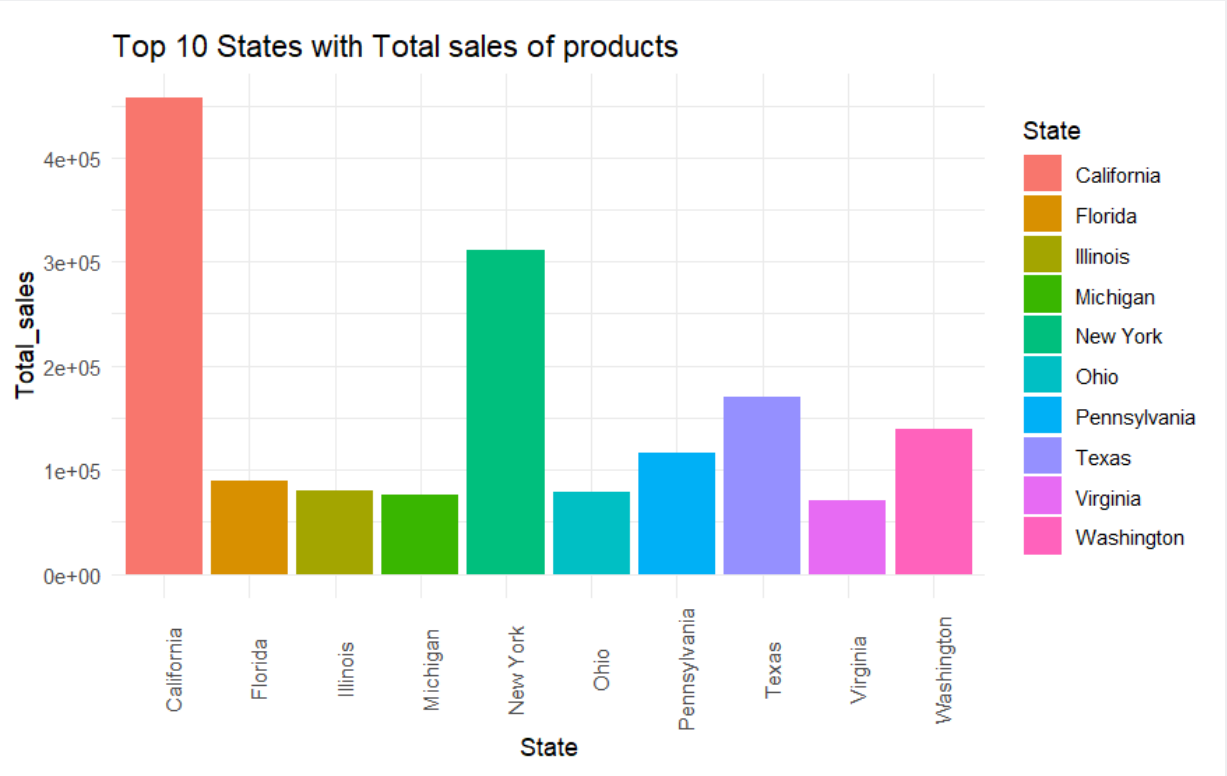


Exploratory Data Analysis

which are the top 10 states that have highest total sales over the entire time period?

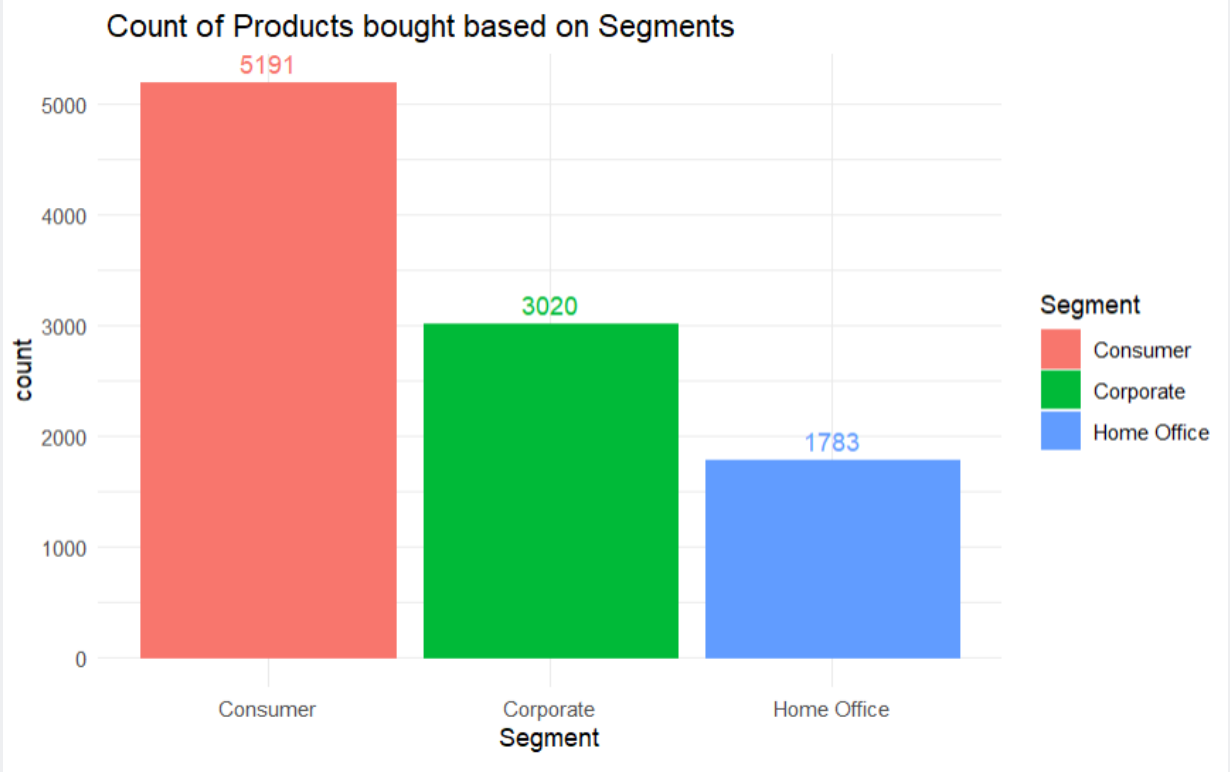
State <chr>	Total_sales <dbl>
California	457687.63
New York	310876.27
Texas	170188.05
Washington	138641.27
Pennsylvania	116511.91
Florida	89473.71
Illinois	80166.10
Ohio	78258.14
Michigan	76269.61
Virginia	70636.72

1-10 of 10 rows



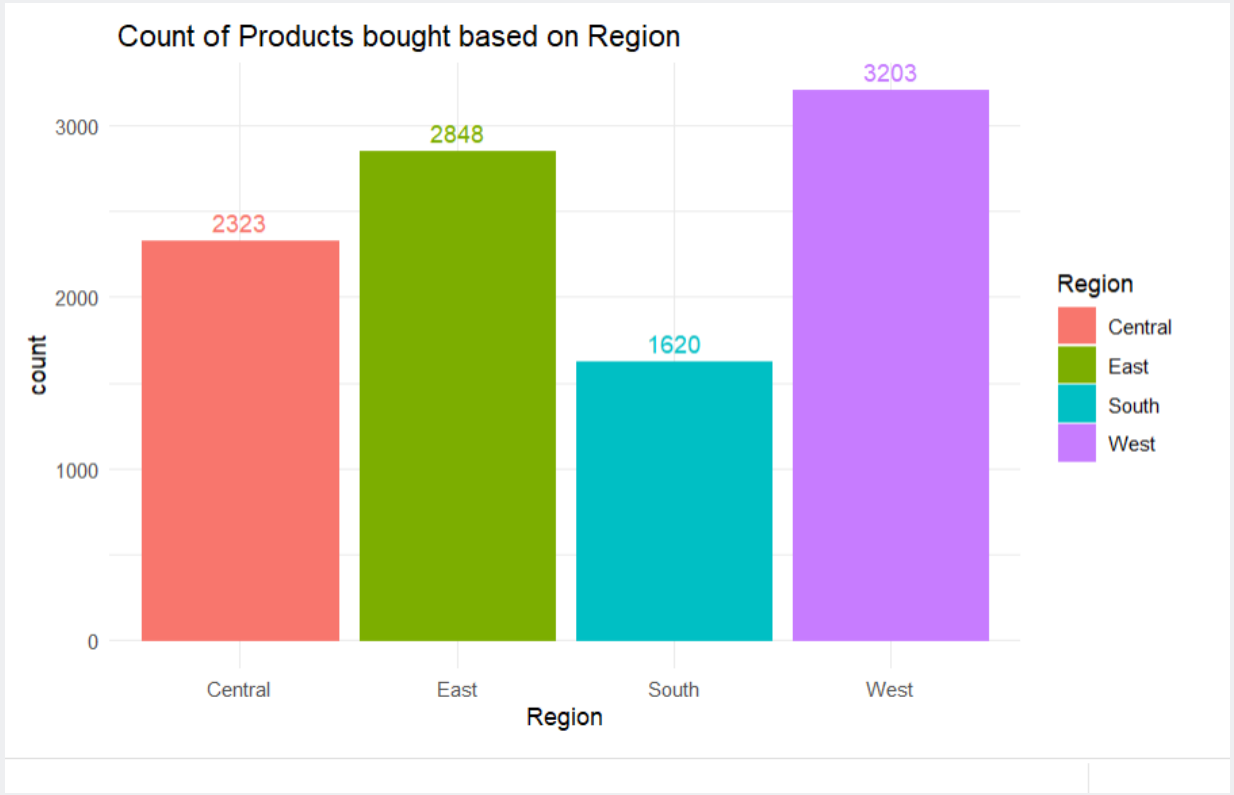
Exploratory Data Analysis

which Segment saw most purchases in products?



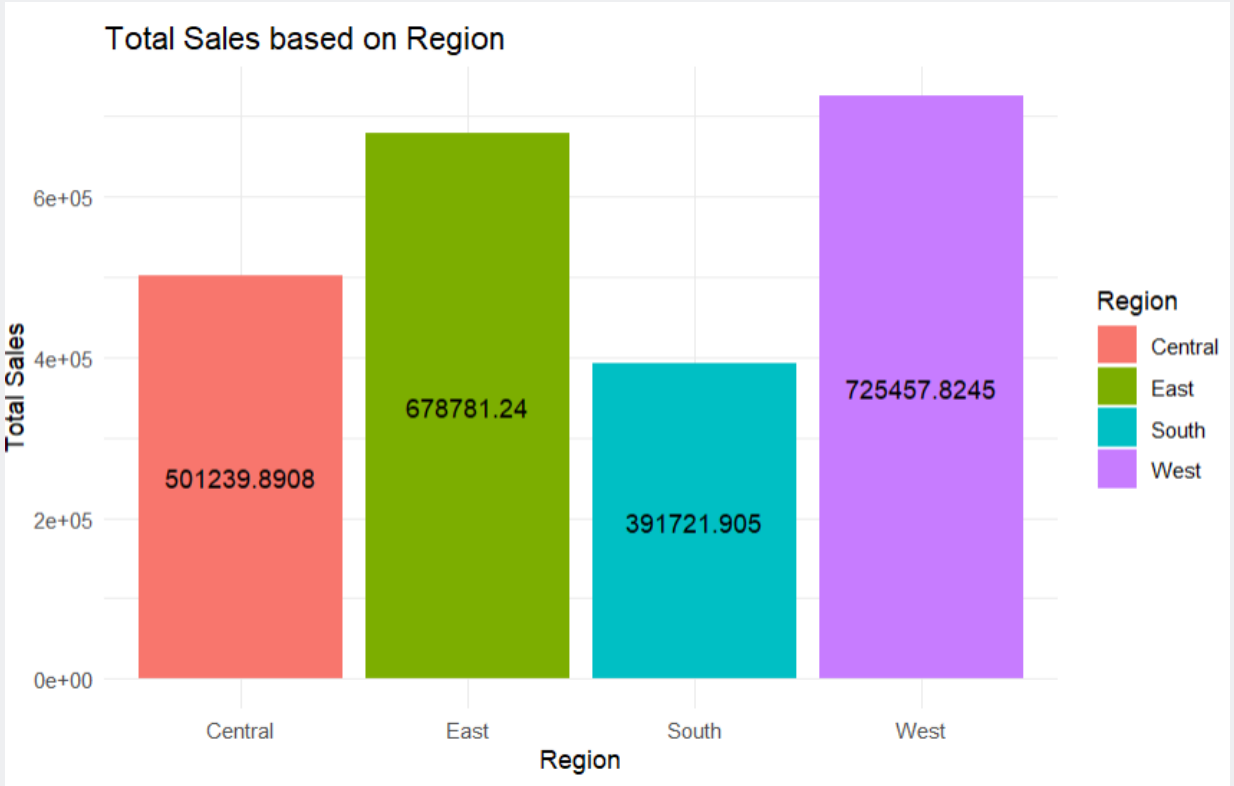
Exploratory Data Analysis

which region has had a lot of Orders placed?



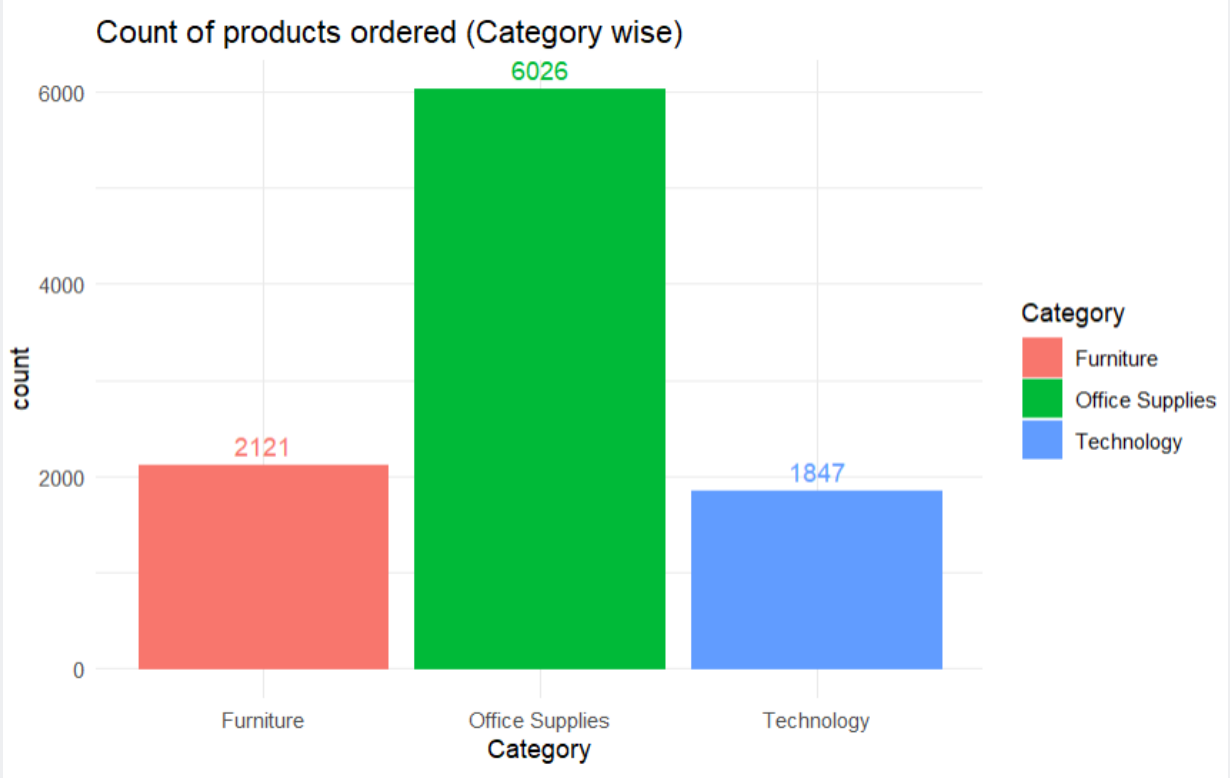
Exploratory Data Analysis

For each region, what has been the Total Sales for the Orders placed?



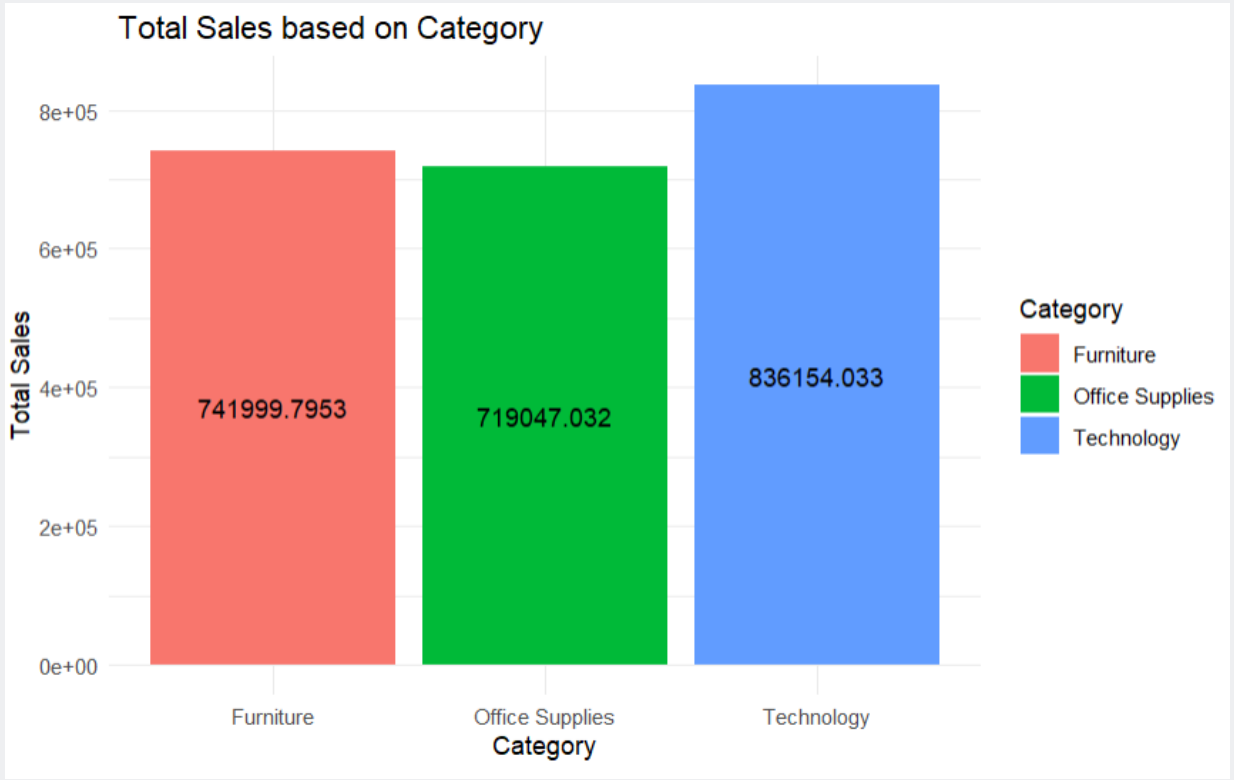
Exploratory Data Analysis

which category of products were bought the most?



Exploratory Data Analysis

which category of products had most sales ?



- In order for any business, to expand their store in various regions. They need to understand which category of products are being purchased by customer? And how the sales are being affected by it?
- For the same purpose, from the previous slide of EDA. We can say that overall Total sales for each category is somewhat similar.
- Hence a Hypothesis Test can be done over the 2 of the Categories say Technology and Office Supplies.
- The reason for picking Technology and Office Supplies are -
  * Technology is widely being used by everyone all around the world (digital world).
  * Furniture as a lot of people are investing to buy homes/ start Offices.
- Demand for those categories wouldn’t go away in long run.

Hypothesis :

Average Total Sales of Technology = Average Total Sales of Office Supplies.

- Based on calculations done, by setting rejection factor (alpha) with least probability of 0.05 and power level i.e. True Negative Rate to 0.8, such that right predictions of the Test are obtained. We did Testing using Neyman-Pearson's method.
- From that we figured that the stated Hypothesis must be rejected as the criteria set by the method was not being met for accepting our assumption.
- Hence, we can't say based on the exploration done that Average sales of Technology related products purchased by customers is same as the Average sales of Office Supplies.

Conclusion

- Each Category has its own Total Sales, we can’t really generalize them or even equalize them to find overall Sales that the Store might receive.
- Each category must be treated with great consideration while calculating Sales as it would allow the business to expand better.

Time Series Analysis

- Also, to understand how the business is doing over time, plotted the Trend of Orders placed with time (2014 - 2017)
- Conclusion – Sales Increasing Over Time, irrespective of Category