Author: Shashidhar M
Task#3:Exploratory Data Analysis-RETAIL

DATA SCIENCE AND BUSINESS ANALYTICS INTERN AT THE SPARKS FOUNDATION(TSF)

using python

Step0 :Business Understanding

- **Here we are to find a business problem and come up with a solution for a retail shop**
- Ask for the domain person for understanding the data and business problem clearly.

Step 1: Importing libraries and the dataset

- Import the dataset and necessary libraries
- And the dataset shall be imported

Step 2: Data inspection

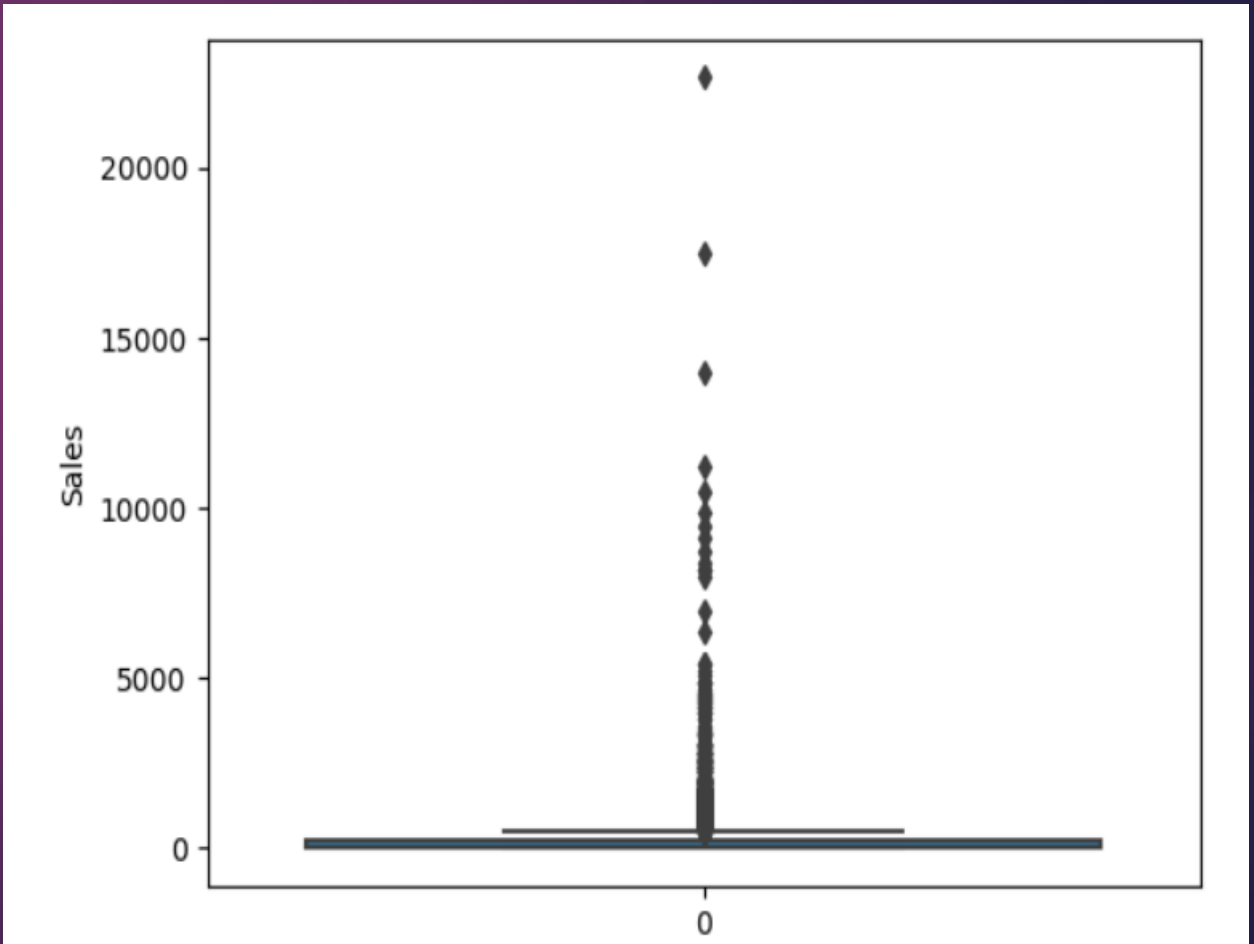
- df.head()
- df.tail()
- df.shape
- df.info()
- df.dtypes
- df.describe()

Step 3: Data manipulation

- **i.) Cleaning the data**
- **check whether there is null values present in the data**
- **There are no null values in the dataset,hence we can proceed with the analysis**
- **ii.) Check if there are any outliers present in the given dataset**
- **applicable for continuous variables**

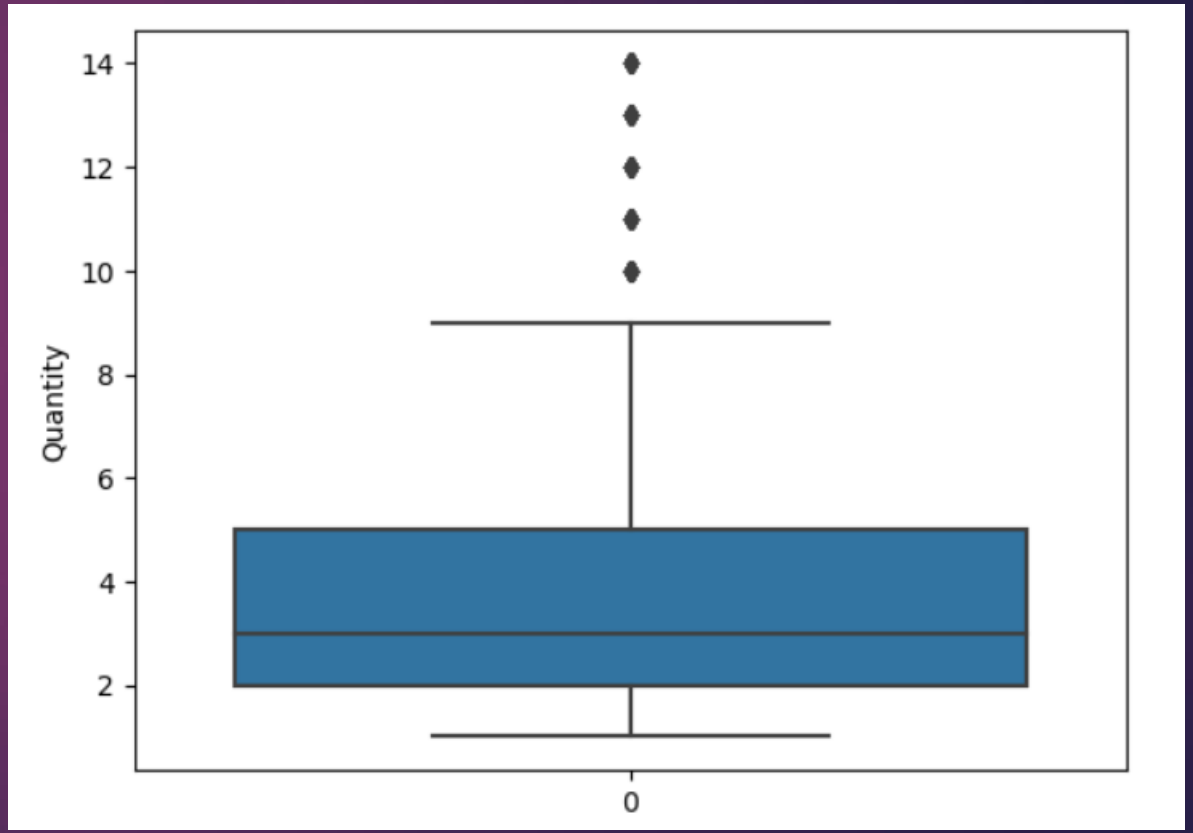
Outliers in sales

- Implies,
- There are outliers in the sales data which is because of some customers purchasing more items.
- There are many of them who have purchased many items.Hence we should conserve the outliers as well for further analysis.
- 11.67% of the customers are having the sales rate high in the entire dataset
- **11.67% of the customers are having the sales rate high in the entire dataset**



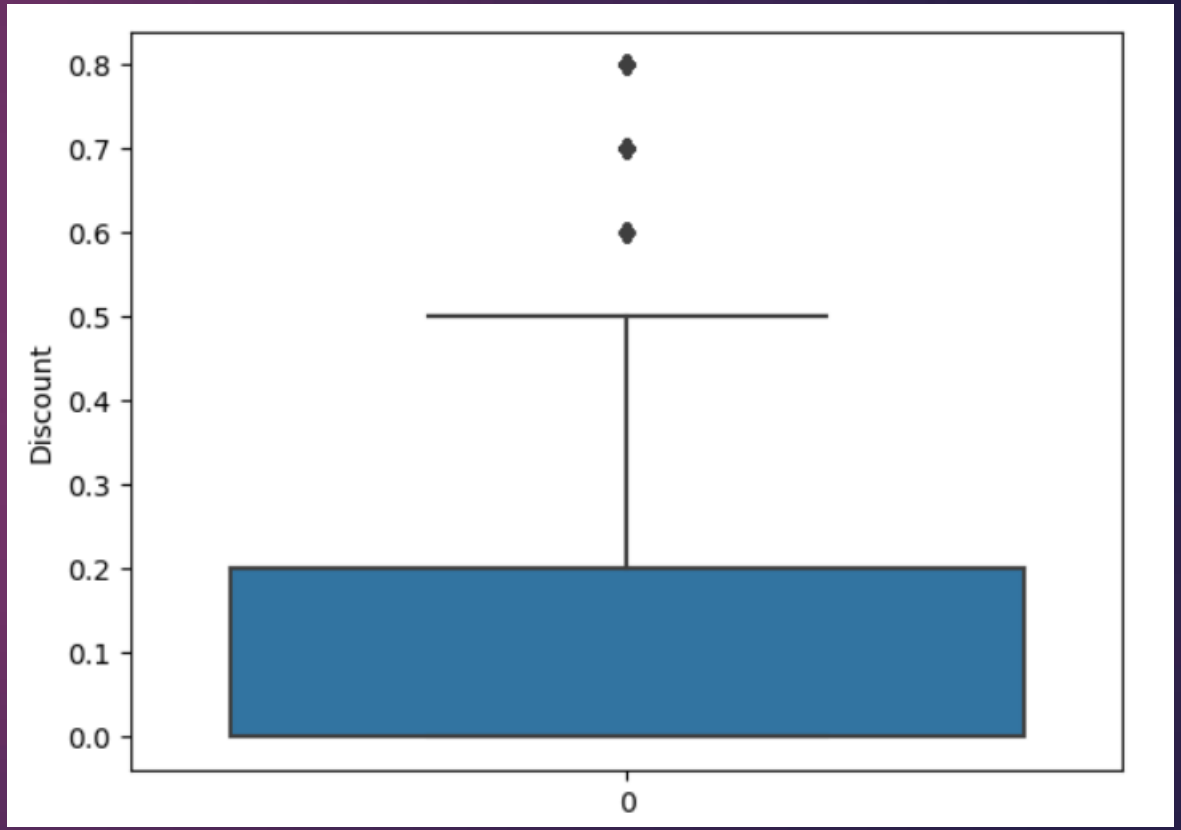
Outlier in quantity

- Maximum Quantities of products purchased are in between 2(minimum) and 5(maximum)
- 5 outliers are seen who have purchased products in quantities of 10,11,12,13,14 respectively
- **upto 4.28% of customers have purchased products in quantity of 9 or more**



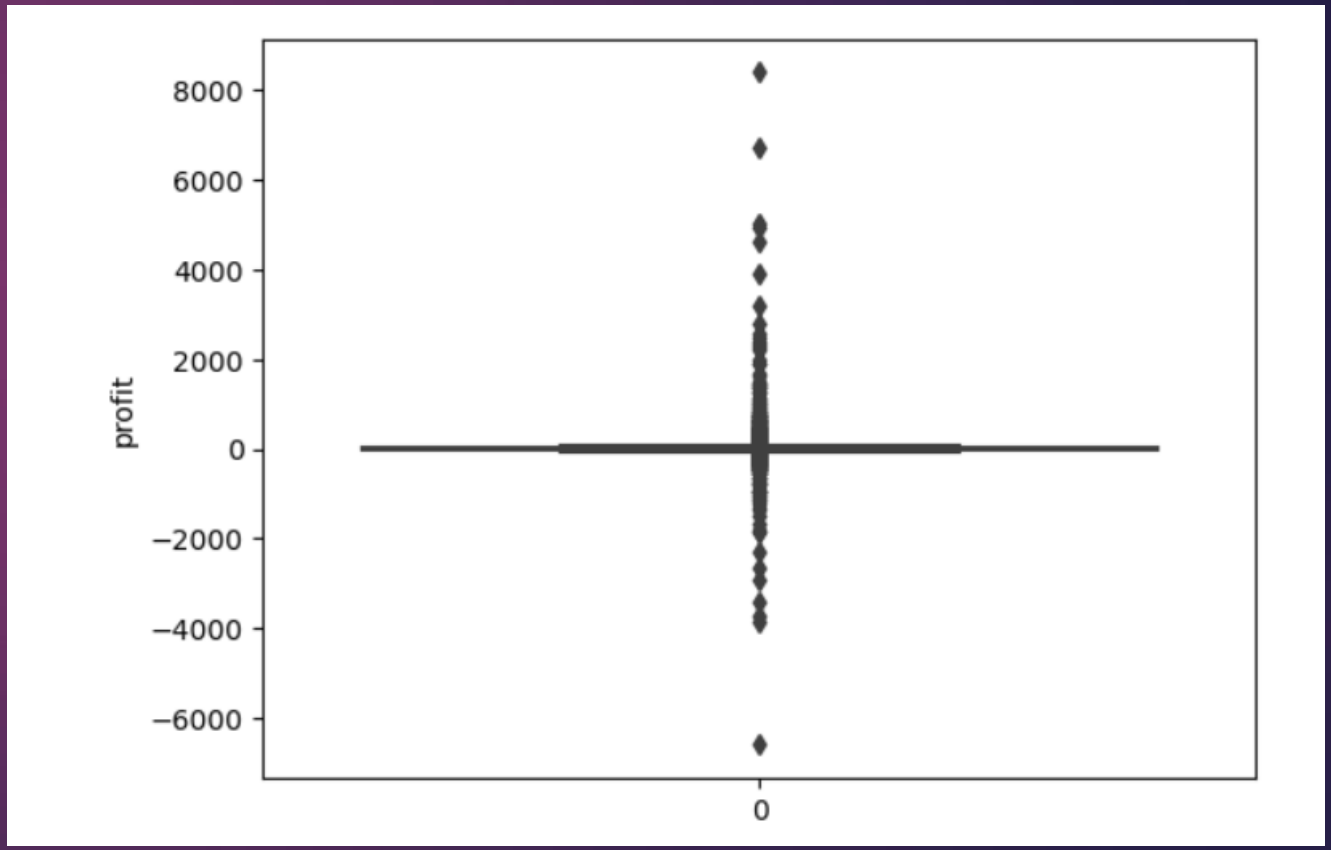
Outliers in Discount

- The discount for majority of customers lie in the range of 0%-20%
- In rare cases the discounts are 60,70 and 80 percents respectively.
- More than 50% of the customers have got the discount more than 20%.=> the discount could be minimized to get even more profit
- 9% of total customers have received discount more than 50%



Outliers in Profit variable

- Maximum values of profit appear in between 1.72 and 27.64
- There are somany outliers present
- Upto 12.7% of the customers helps making more profit for retail store
- Therefore, there's upto $1.923lakhs total profit for the retail store

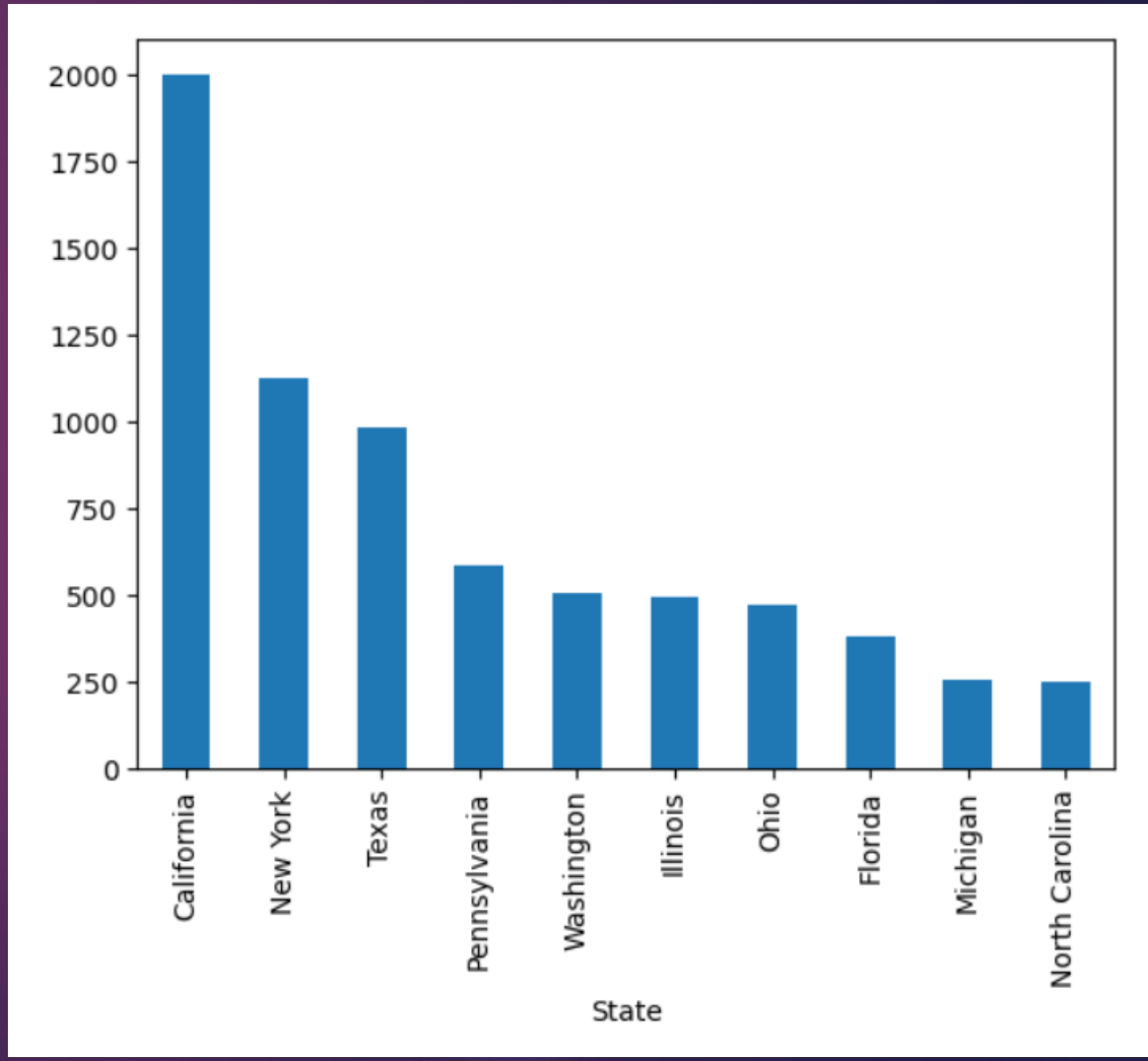


Step4: Exploratory Data Analysis(EDA)

- *-Univariate*
- *-Bivariate*
- *-Multivariate*

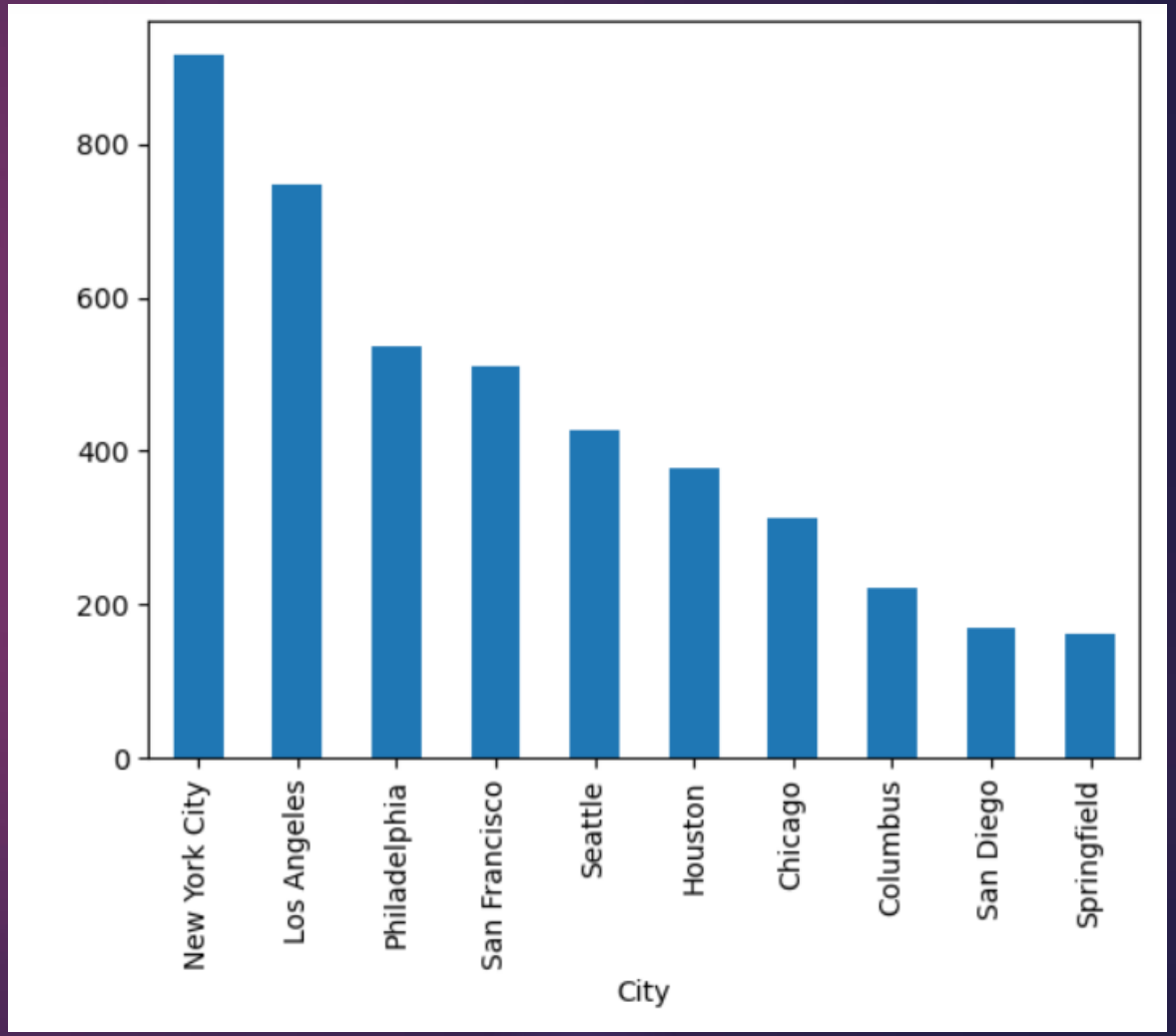
Univariate-bar ,State column

- Bar chart-They map categories to numbers
- -The sales in state California tops with 2000 sales/state. -New york standing 2nd with around 1125 sales/state.
- -Texas stands 3rd with nearly 1000 sales/state.
- -Pennsylvania,Washington,Illinois,Ohio states almost have equal sales of 500 sales/state.
- -Florida, Michigan,North Carolina nearly has around 250 sales/state.
- -The remaining states records very less sales/state.
- "We could do more marketing to attract customers where sales are recorded less,we could provide discounts initially to attract customers to our retail store"



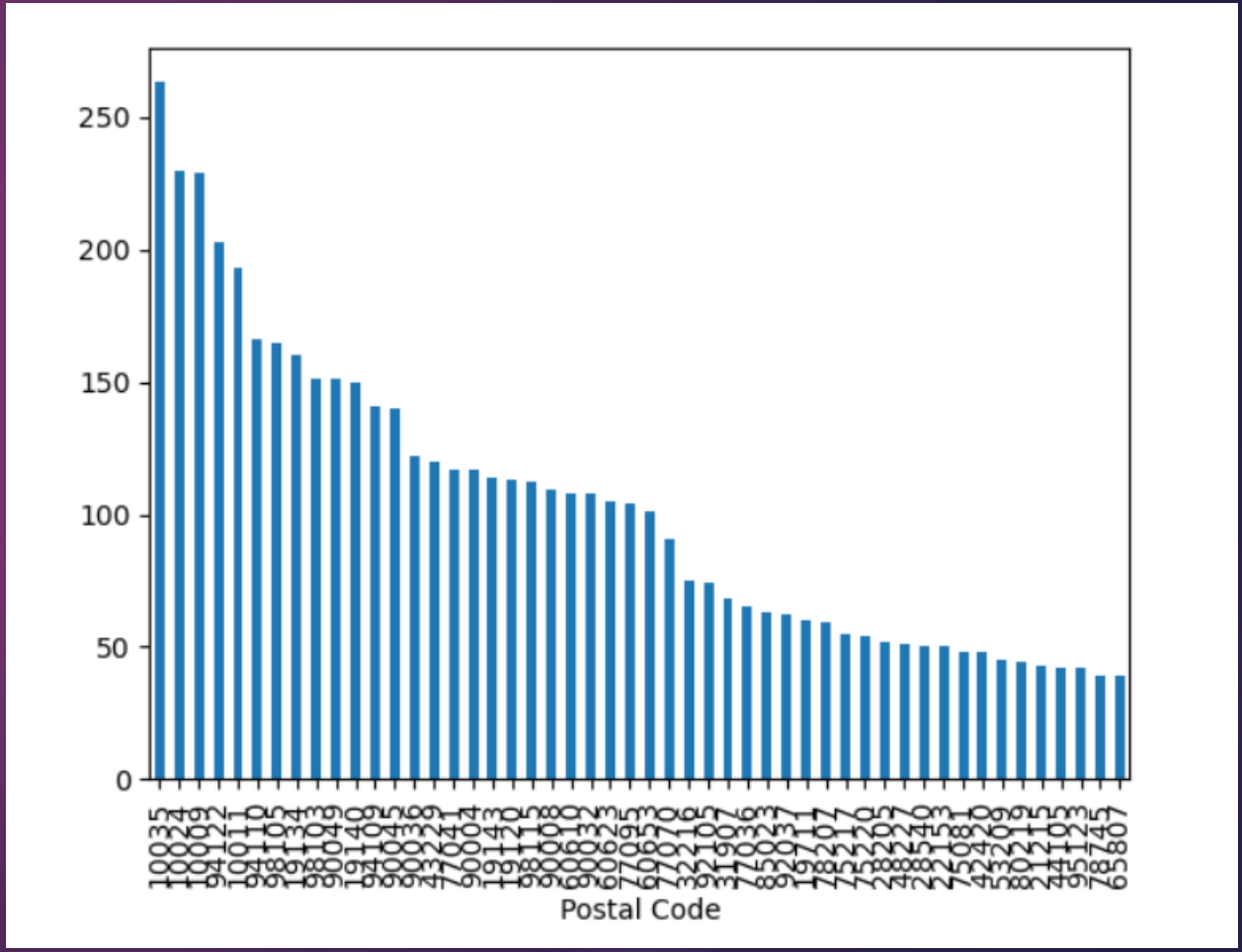
Uni-**City**

- -New york city nearly has 1000 sales/city.
- -Los angeles nearly has 800 sales/city.
- -Philadelphia and San Francisco has nearly 550 sales/city
- . -Seattle,Houston,Chicago has nearly 400 sales/city.
- -The remaining cities have less sales.
- -We could use marketing in cities wherever the sales are less,say we will give discounts,offers to attract customers



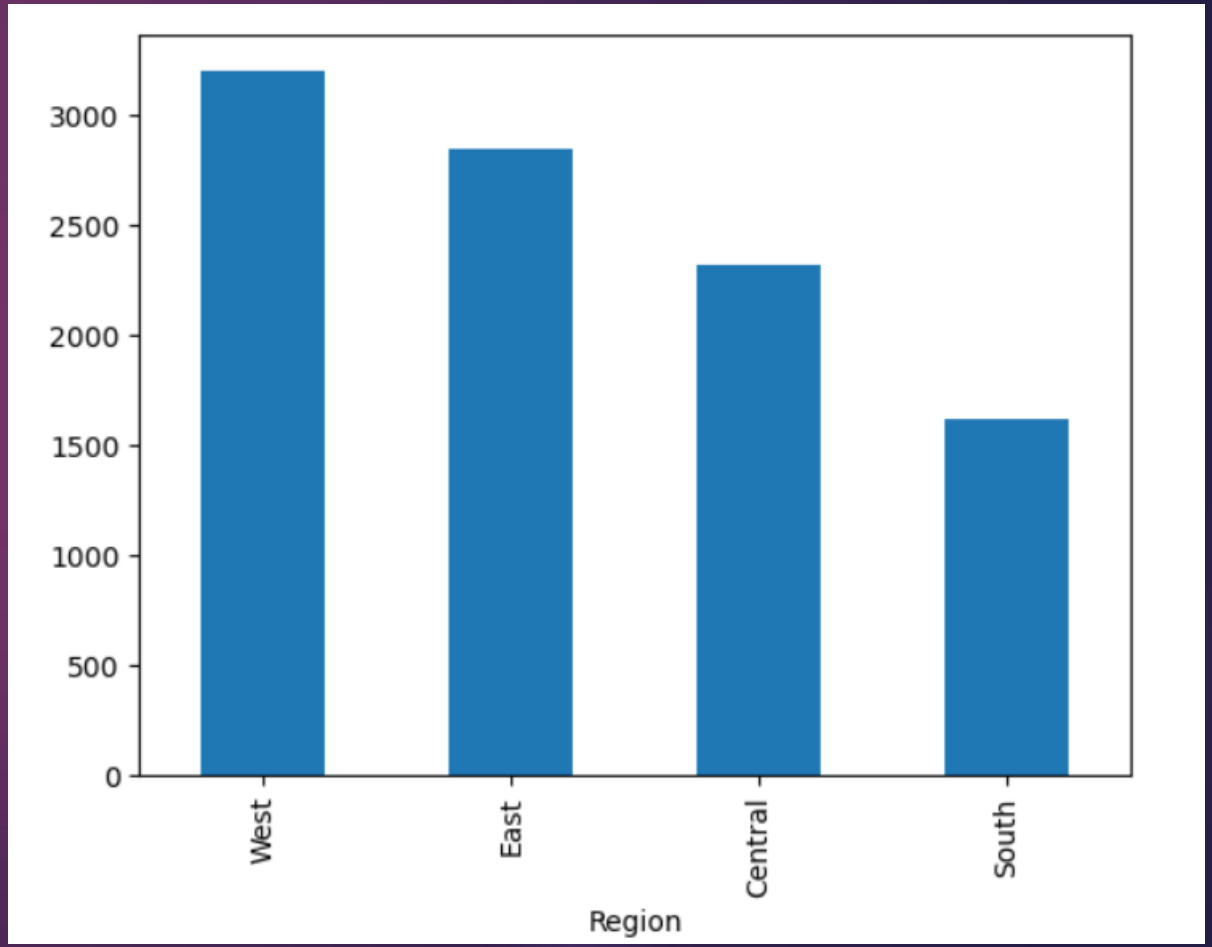
Uni-Postal code

- Distribution of sales according to various postal code of customers are seen
  It seems to have a exponential decrease of sales



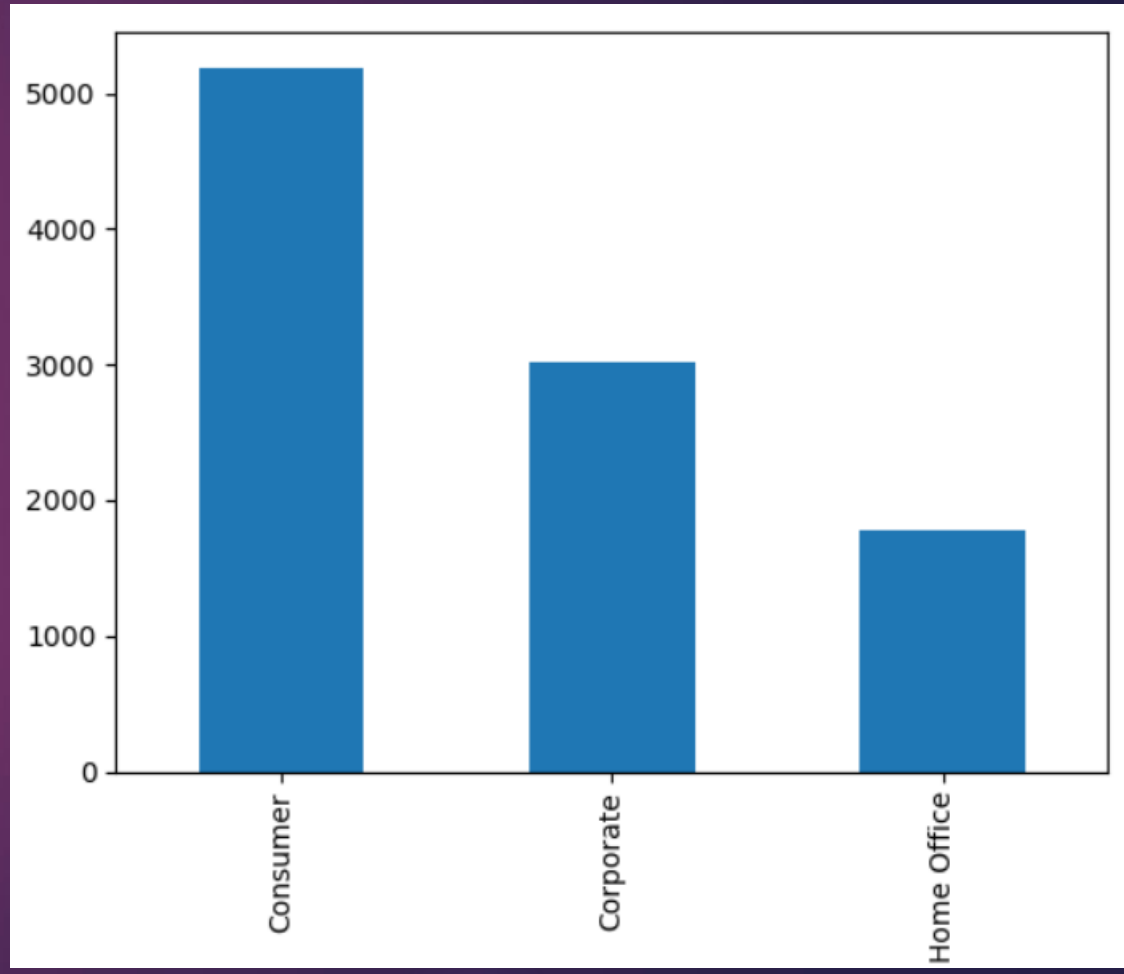
Uni-**Region**

- We can see the West USA has more sales than any other part of USA. Gradually sales decreasing from West to East,East to Central, and from central to south



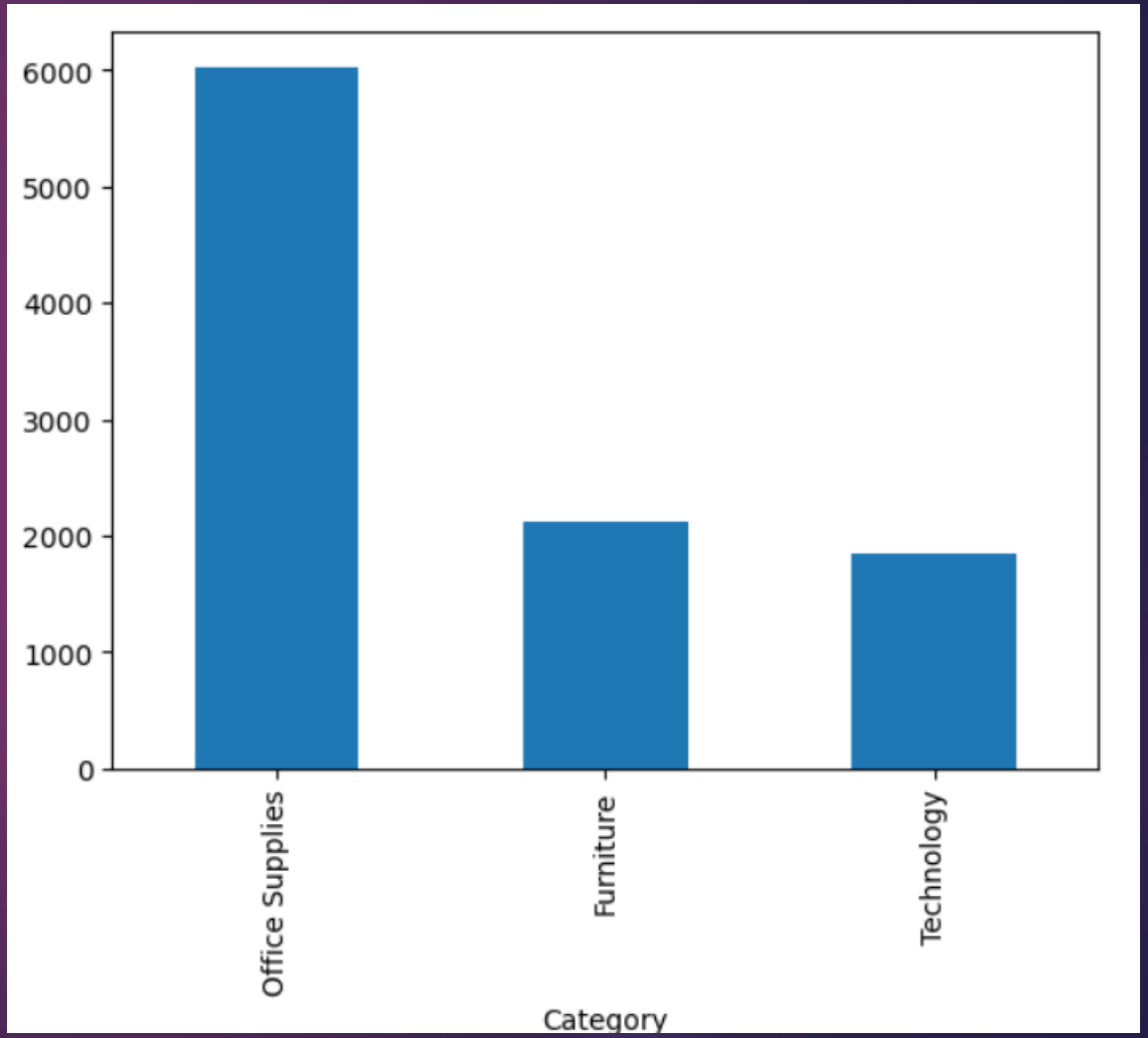
Segment

- -We can see that the sales are more under the segment Consumer which has more than 5000sales/segment.
- -Corporate has 3000sales/segment.
- -And the least being HomeOffice 1500sales/segment -we can do good marketing on "Home office" segment to increase sales from this segment



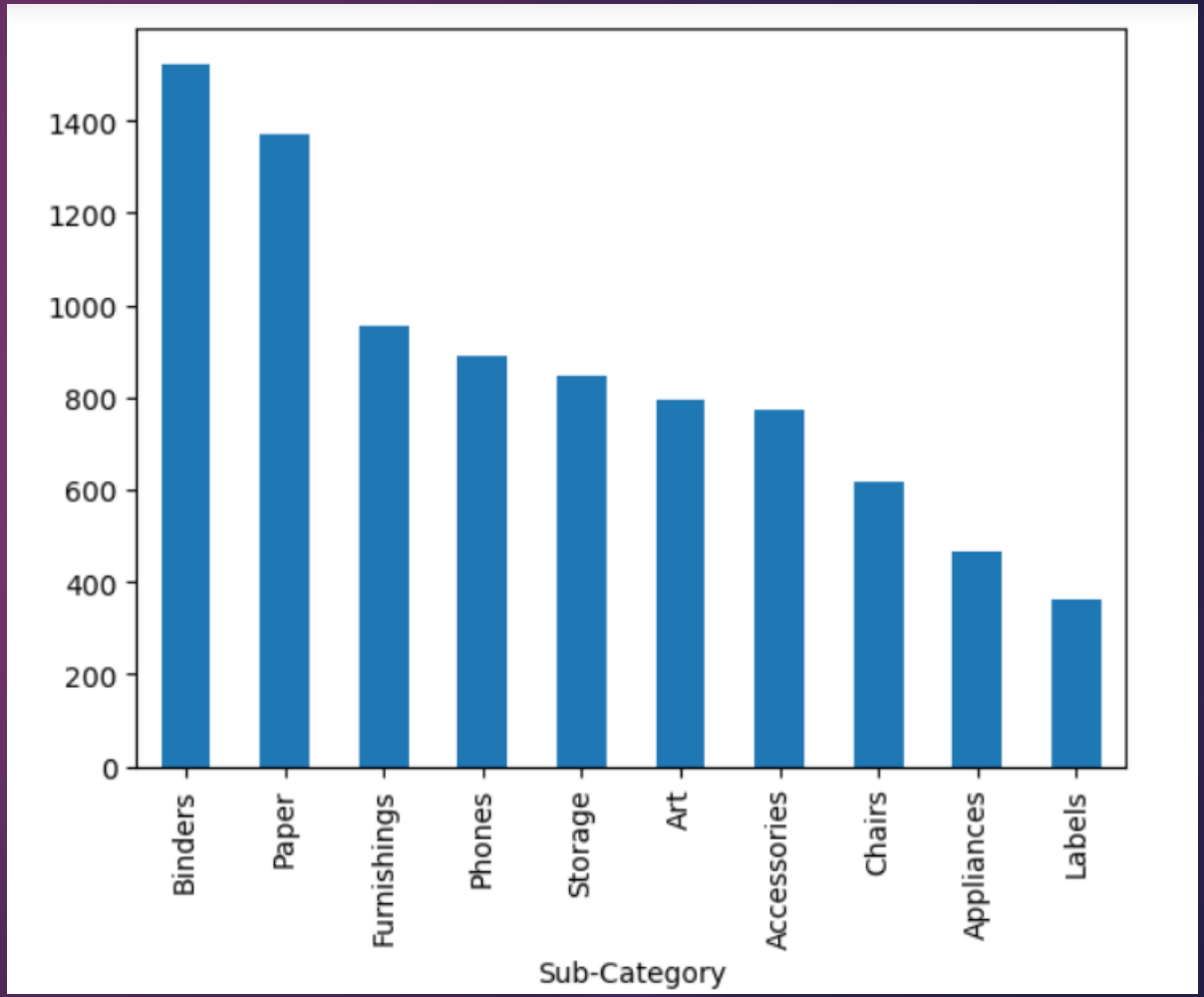
Category

- Office supplies category has got most sales/category. Furniture and technology has got very less. we can give some discounts to increase sales/category in other categories also.



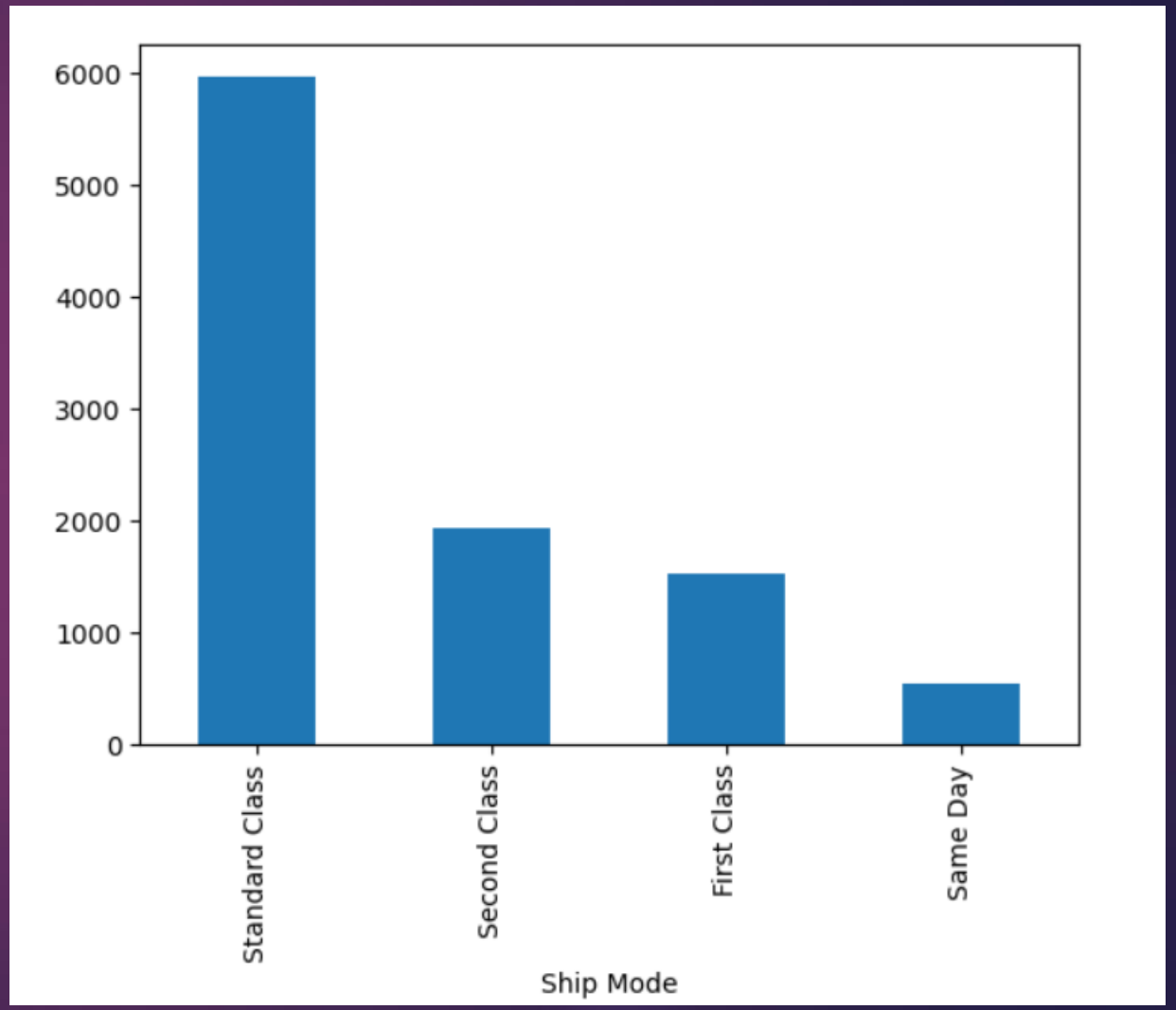
Sub Category

- -We can see the top10 subcategories sold in a retail store
- -Top sales/subcategory is seen in Binders,Paper.
- -sales/subcategory is almost same in Furnishings,phones,storage,Art,Acce ssories
- -sales/subcategory is very less in Appliances,Labels,and so on..we can market these and give attractive discounts under these sub-categories



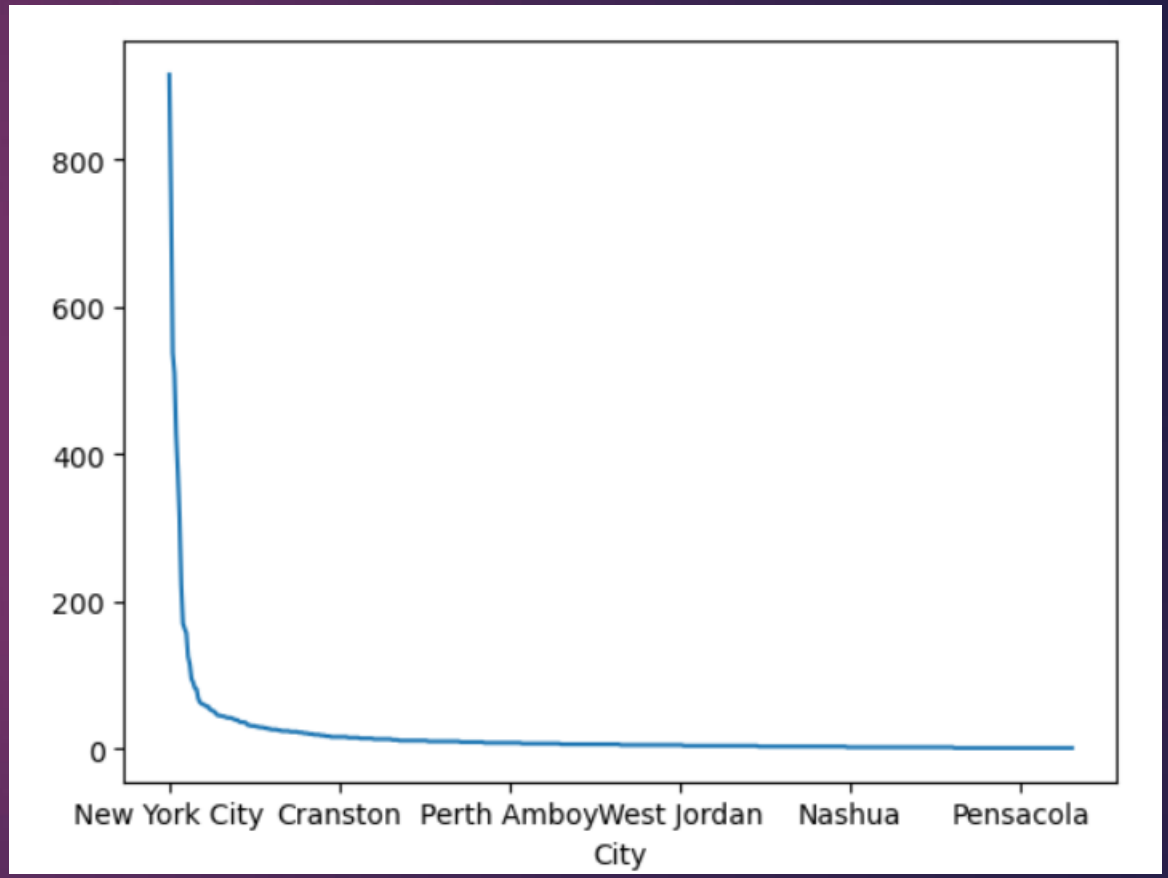
Ship Mode

- The category ship mode where ship mode-"standard class" has more sales The category other than "Standard class" have less sales WE can improve the sales by providing "same day" ship mode, can attract customers who are in a hurry to buy products from a retail store



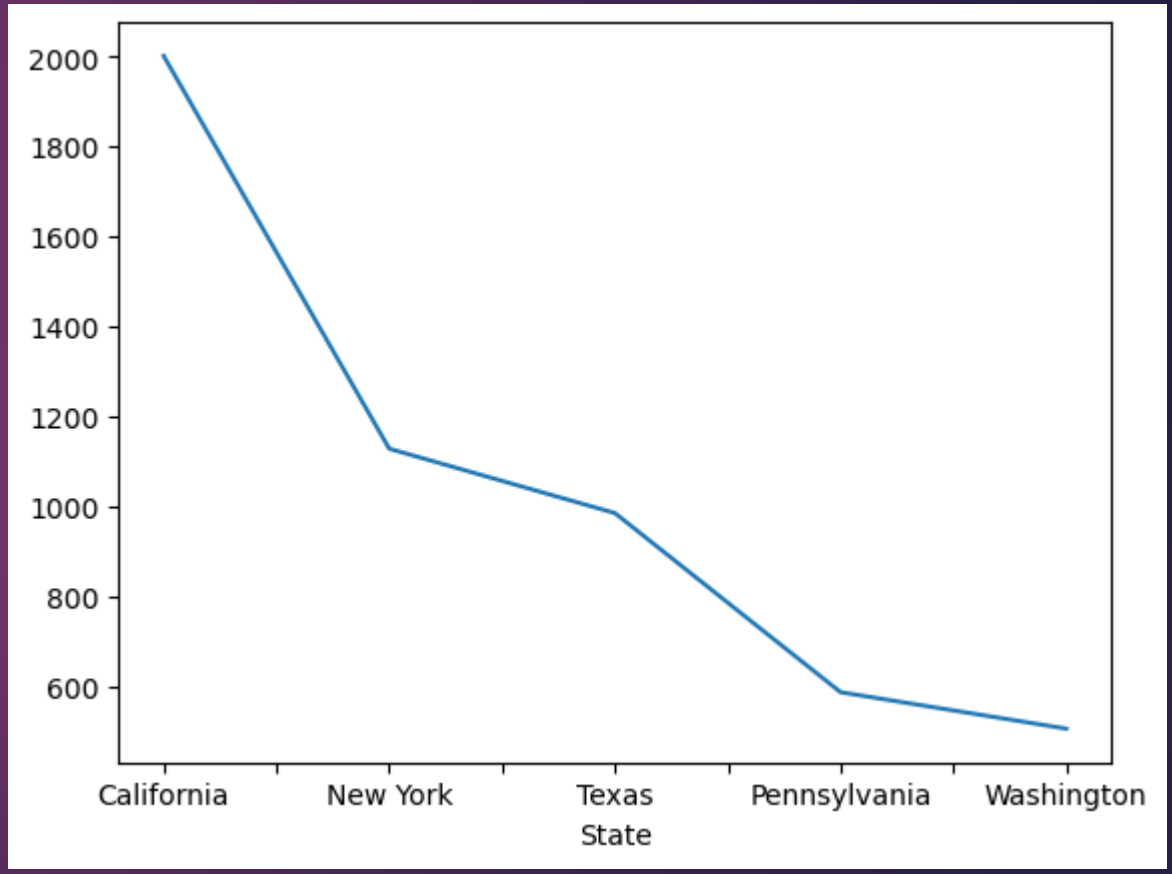
Line chart, City

- new york tops the sales/city category and all others less



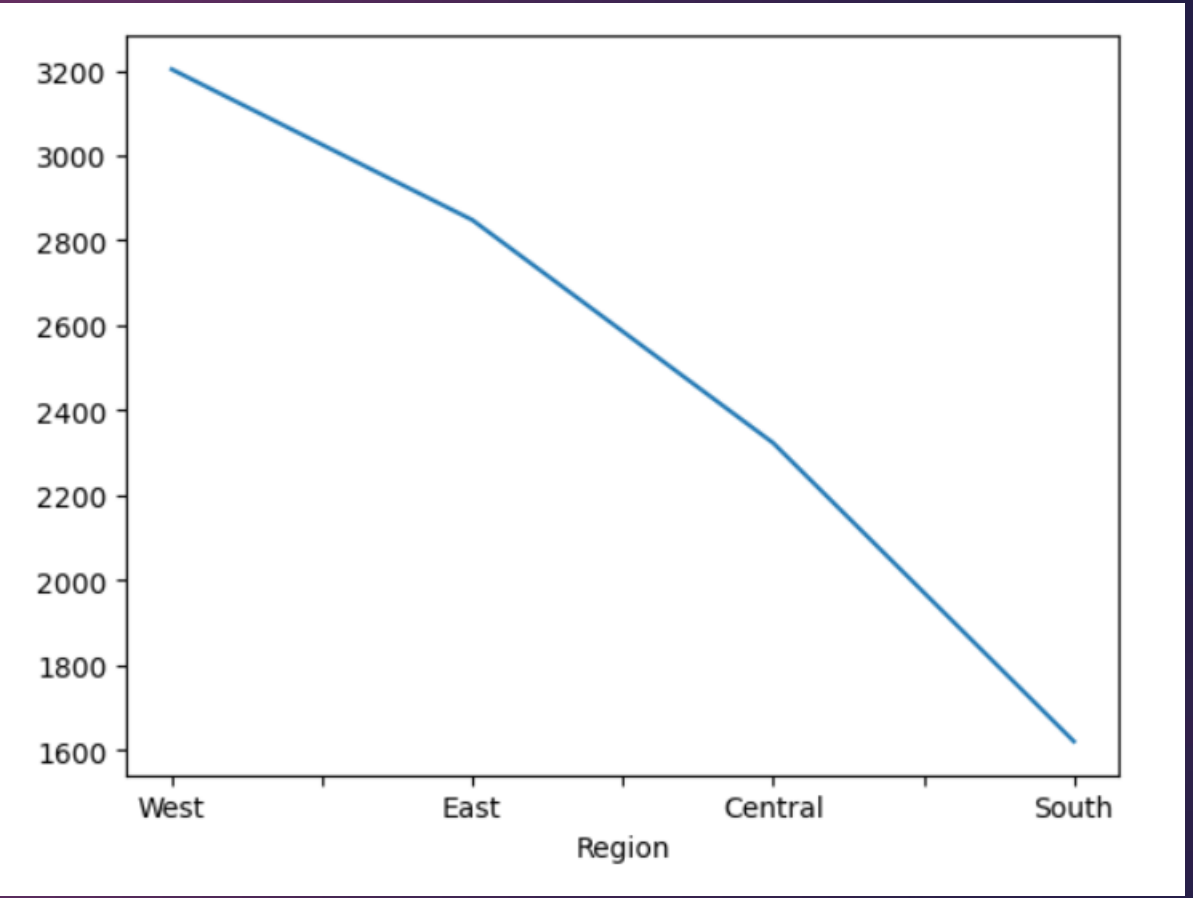
State

- The sales/state is high in California gradually decreasing from newyork till washington and all other cities we should try to attract customers in states other than California to increase the profit of retail store



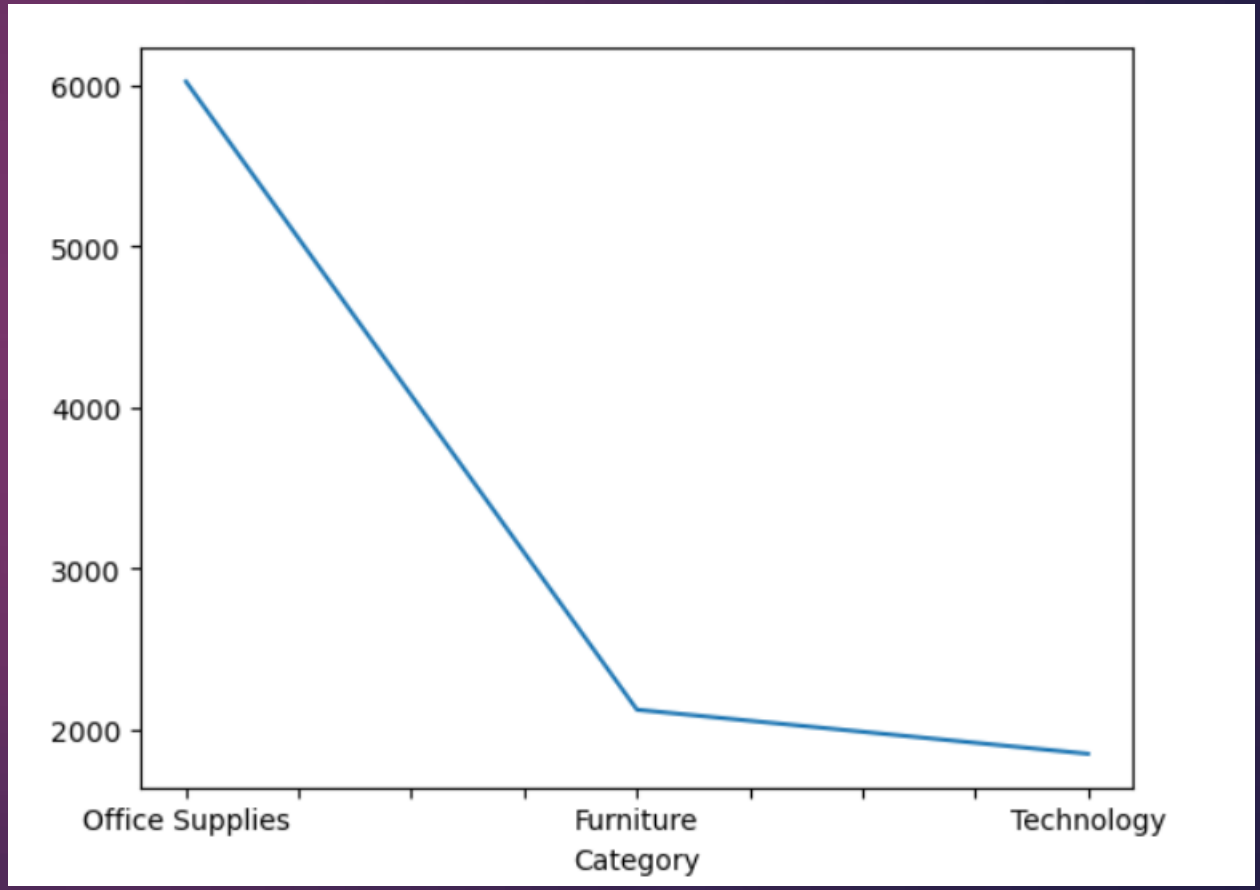
Region

- The sales trend across different regions of USA is as above



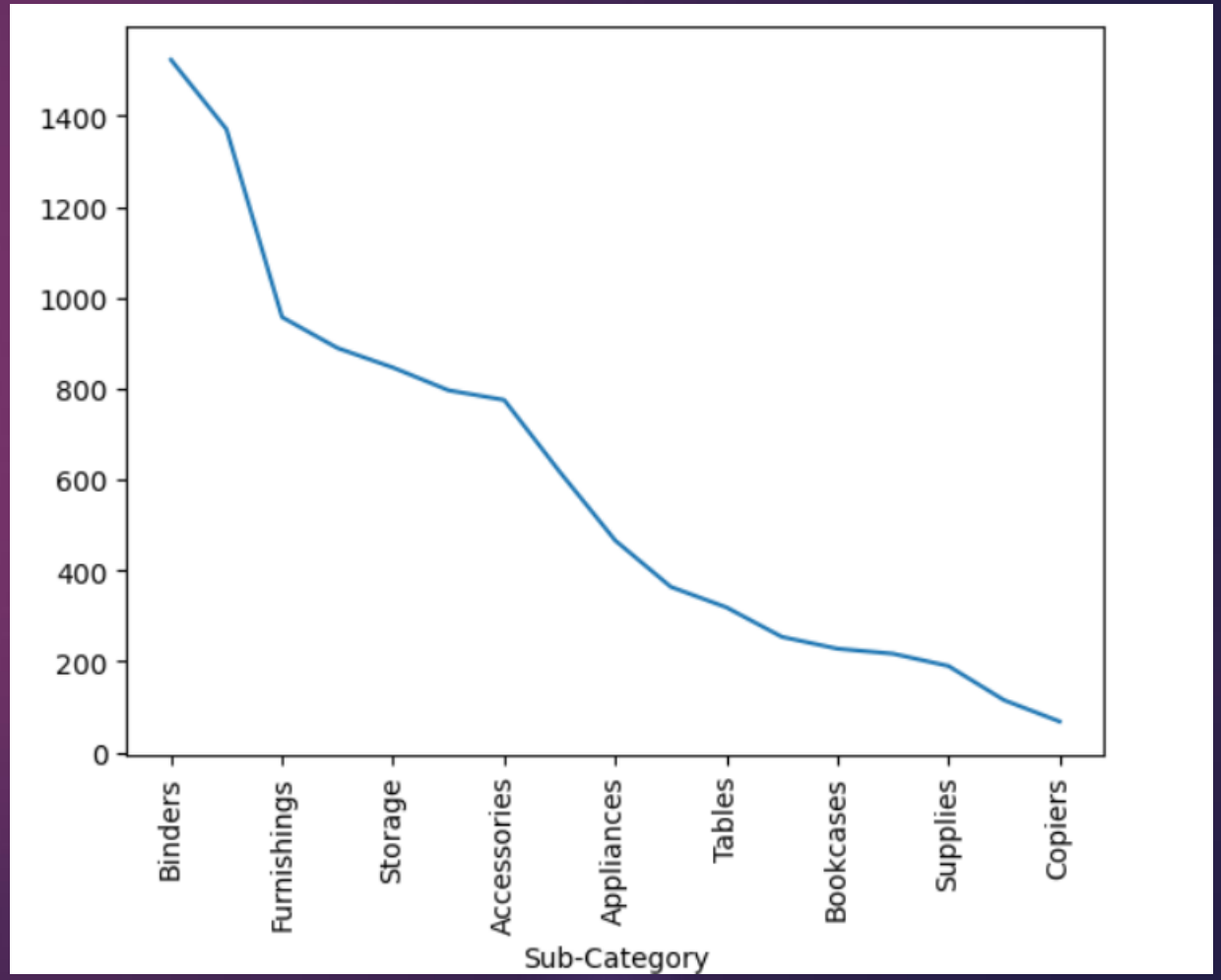
Category

- Trend in sales of different categories are seen,office supplies dominating



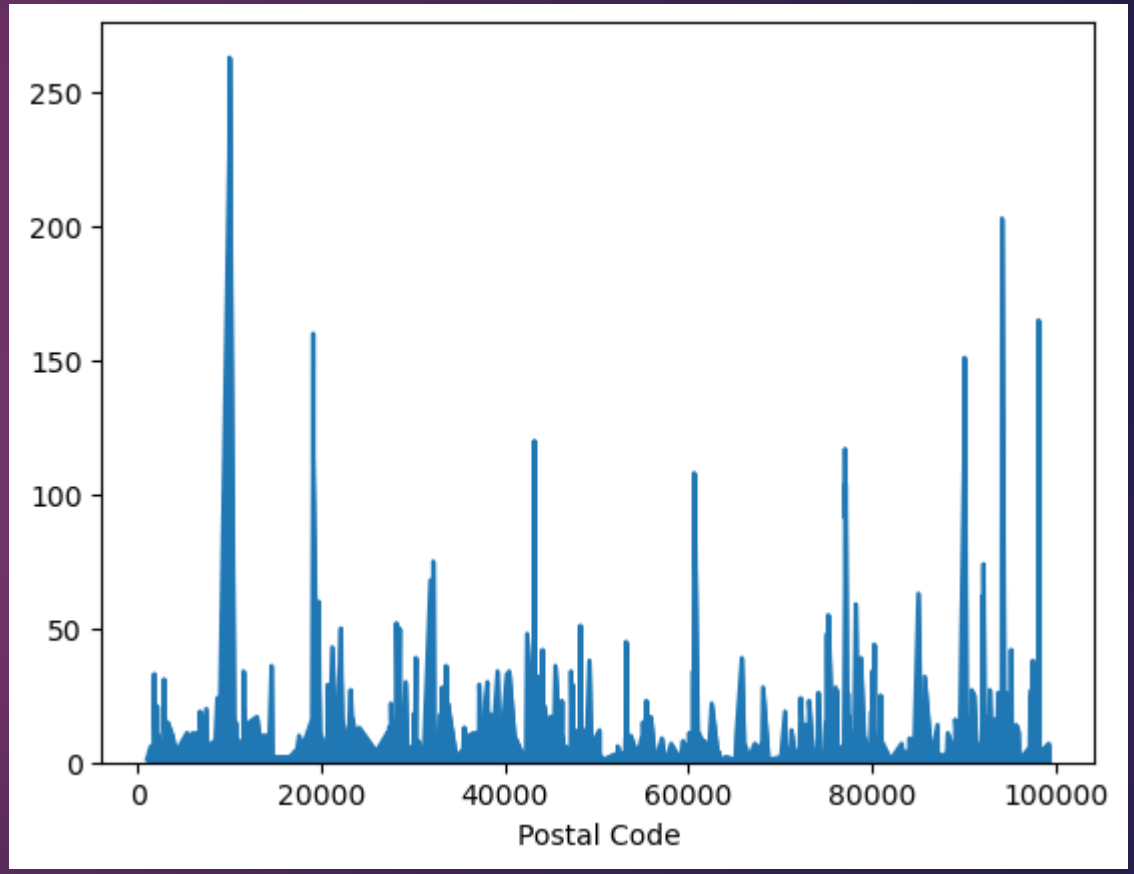
Sub-Category

- Sub-categories trend is seen above,Binders are most sold,followed by furnishings,storage and others decreasing exponentially



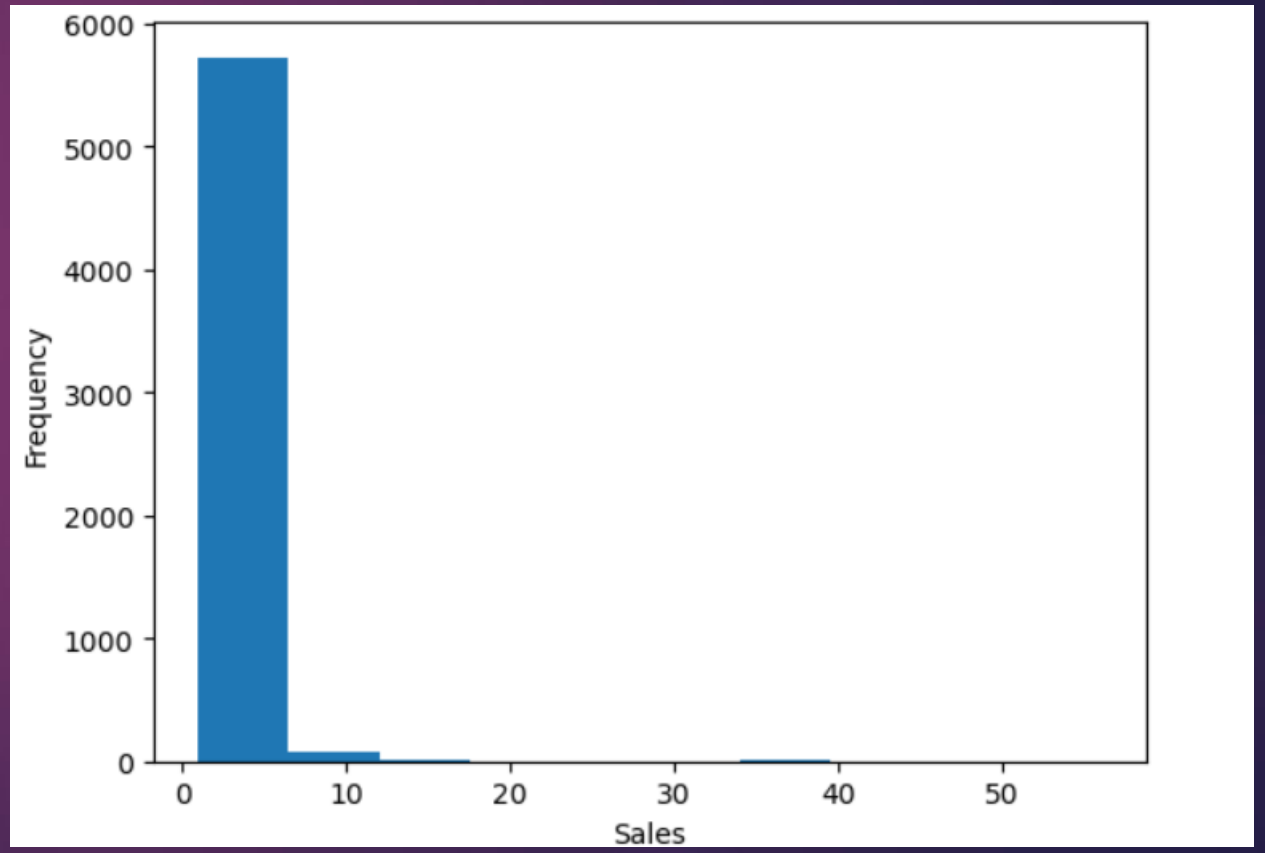
Area chart -- nominal categorical

- We can see the areas of postal codes and their corresponding, approximate postal codes. however, for this dataset We can't conclude much using Area Chart



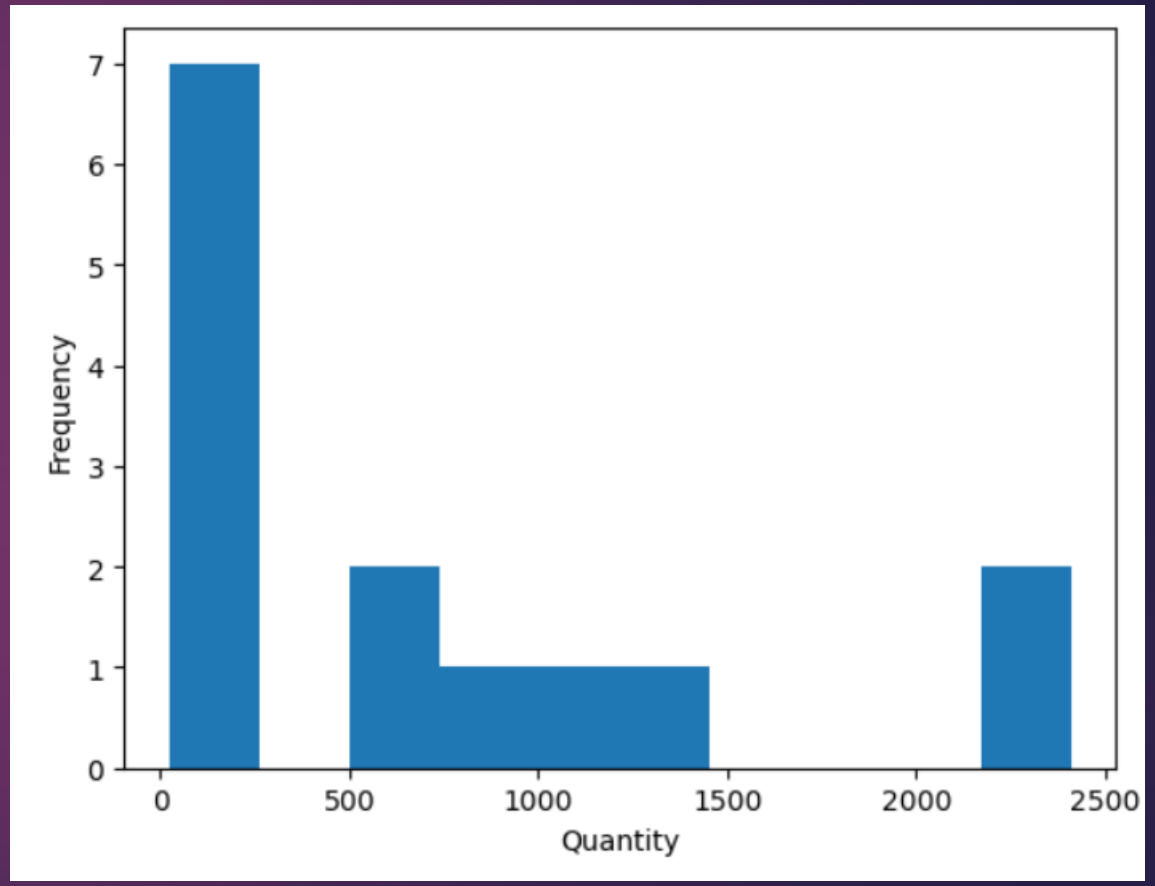
Histogram- plotted for nominal interval scales-Sales

- We can clearly see that the sales is more in between 0 and 10



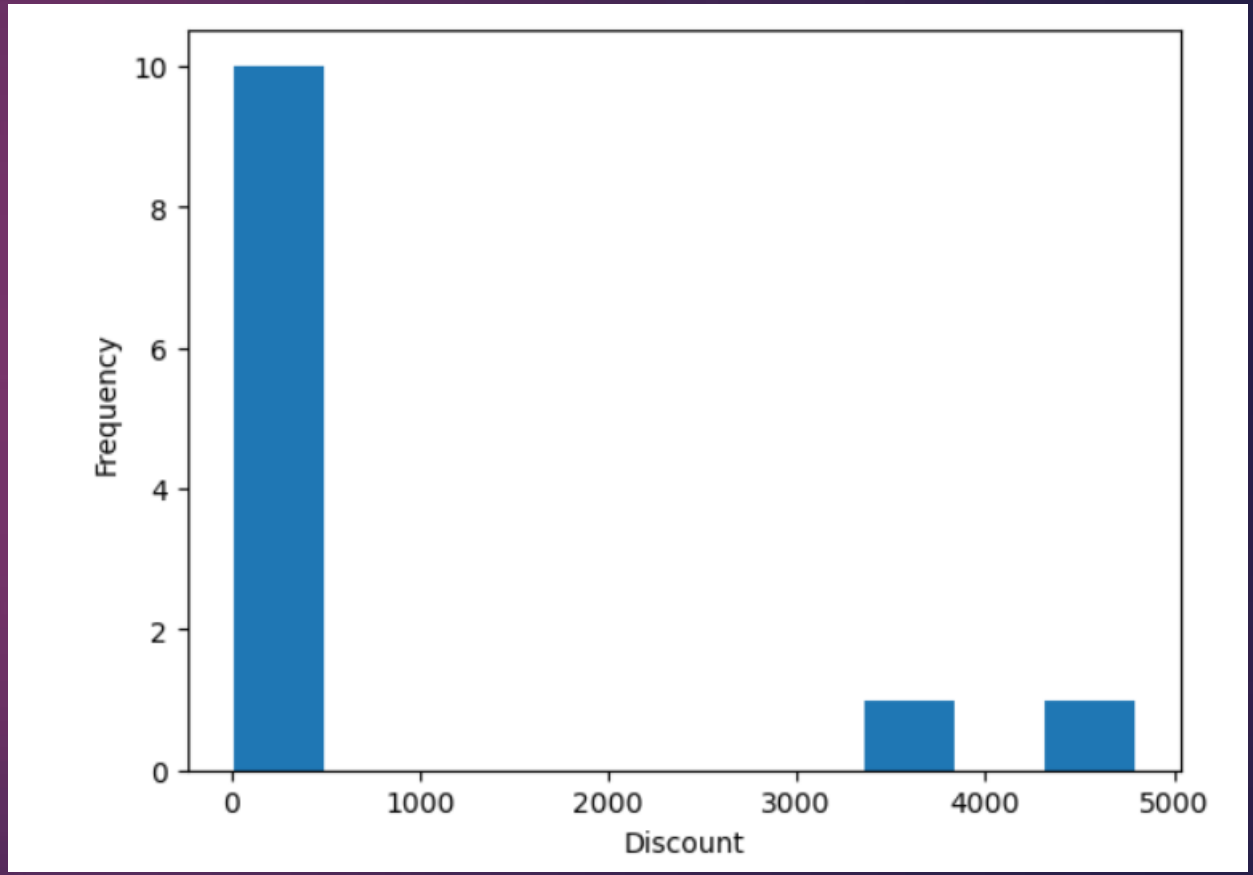
Quantity

- more frequently bought quantities are between '0-200' less frequently bought quantities are between '500-1500'



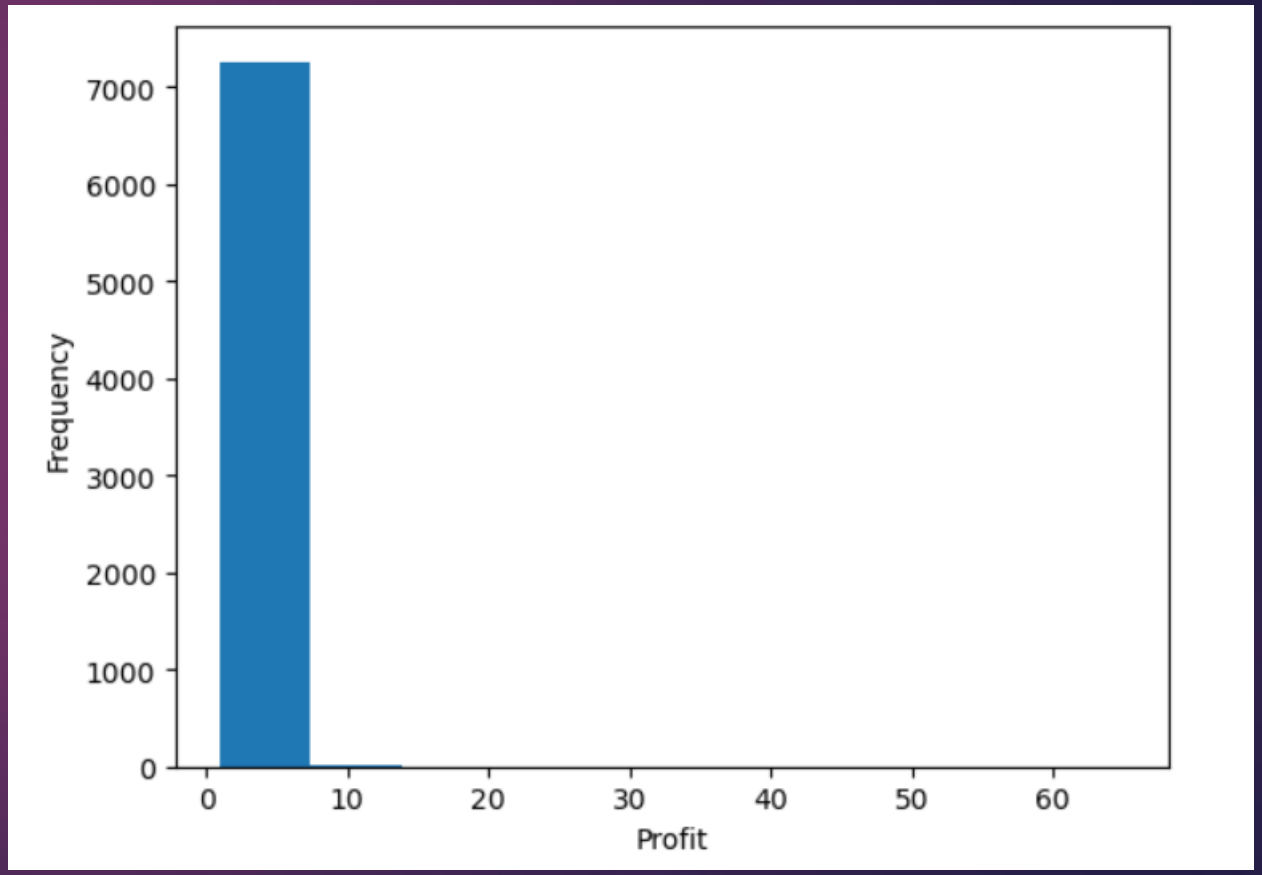
Discount

- The discount offered is high in the range of 0 to 500
- and very less frequently the discounts of 3000, 4000,And 5000 are obtained by customers



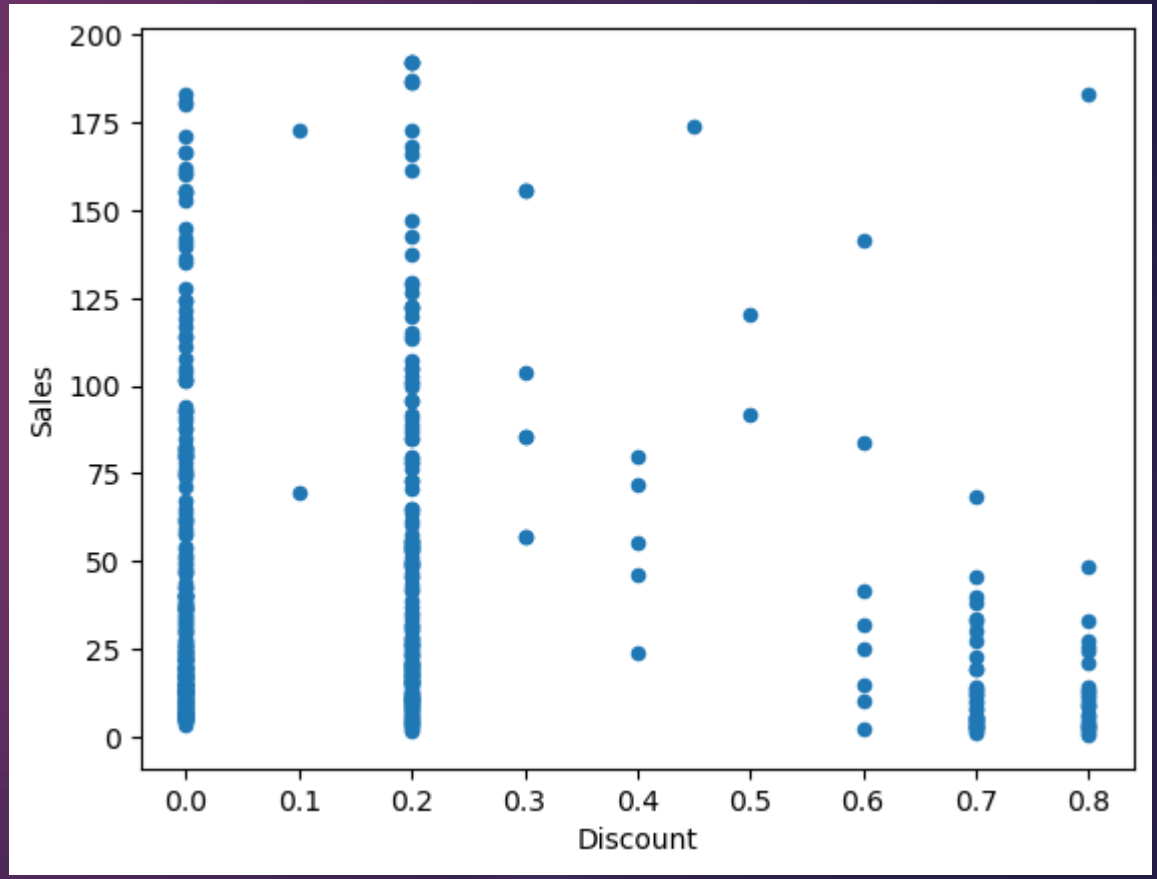
Profit

- the profit margin of all products are in the range of '0 - 10' we could get more profit margin out of each customers so that we get more profit



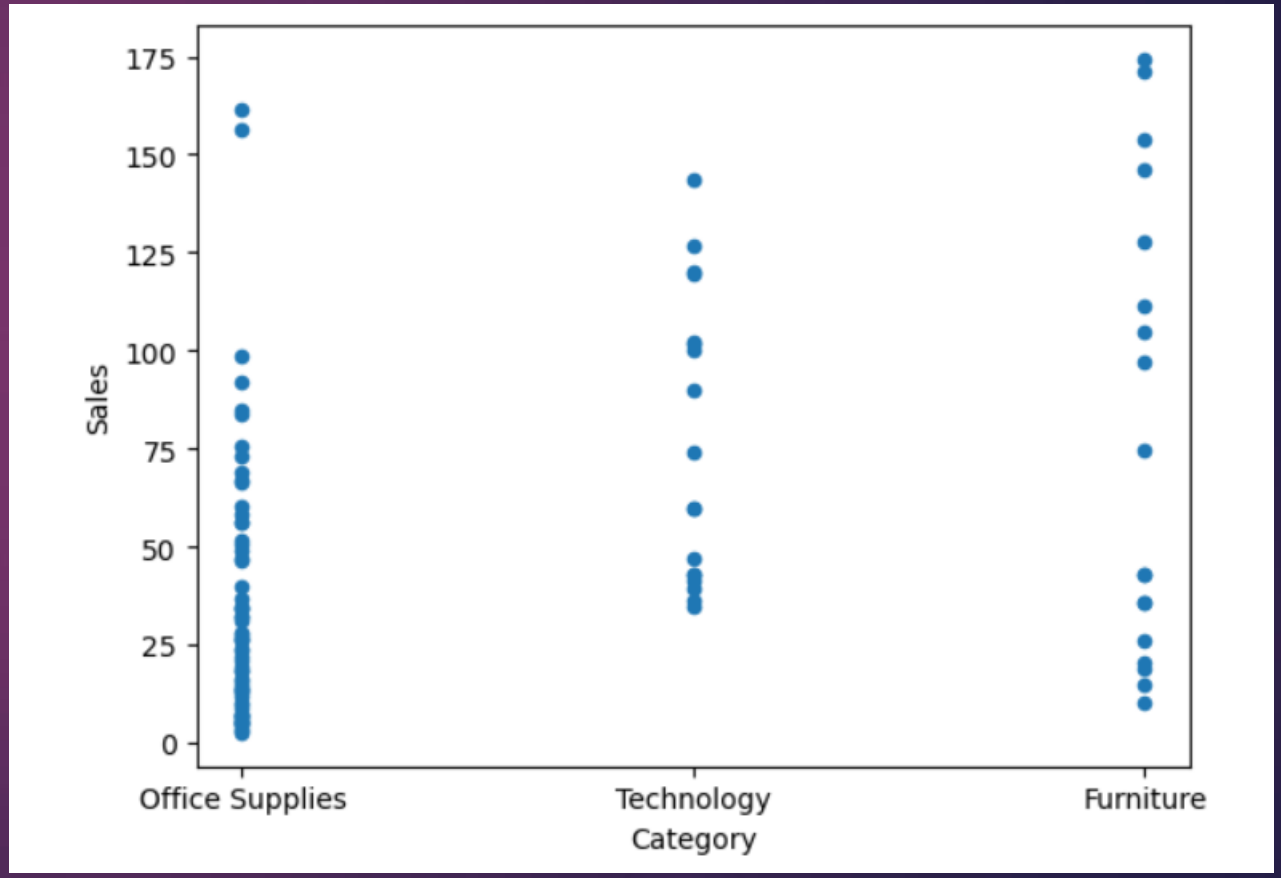
Bivariate Analysis 1st Scatter plot sales,discount

- Clearly graph shows non-linearity between sales(dependent variable) and independent variable



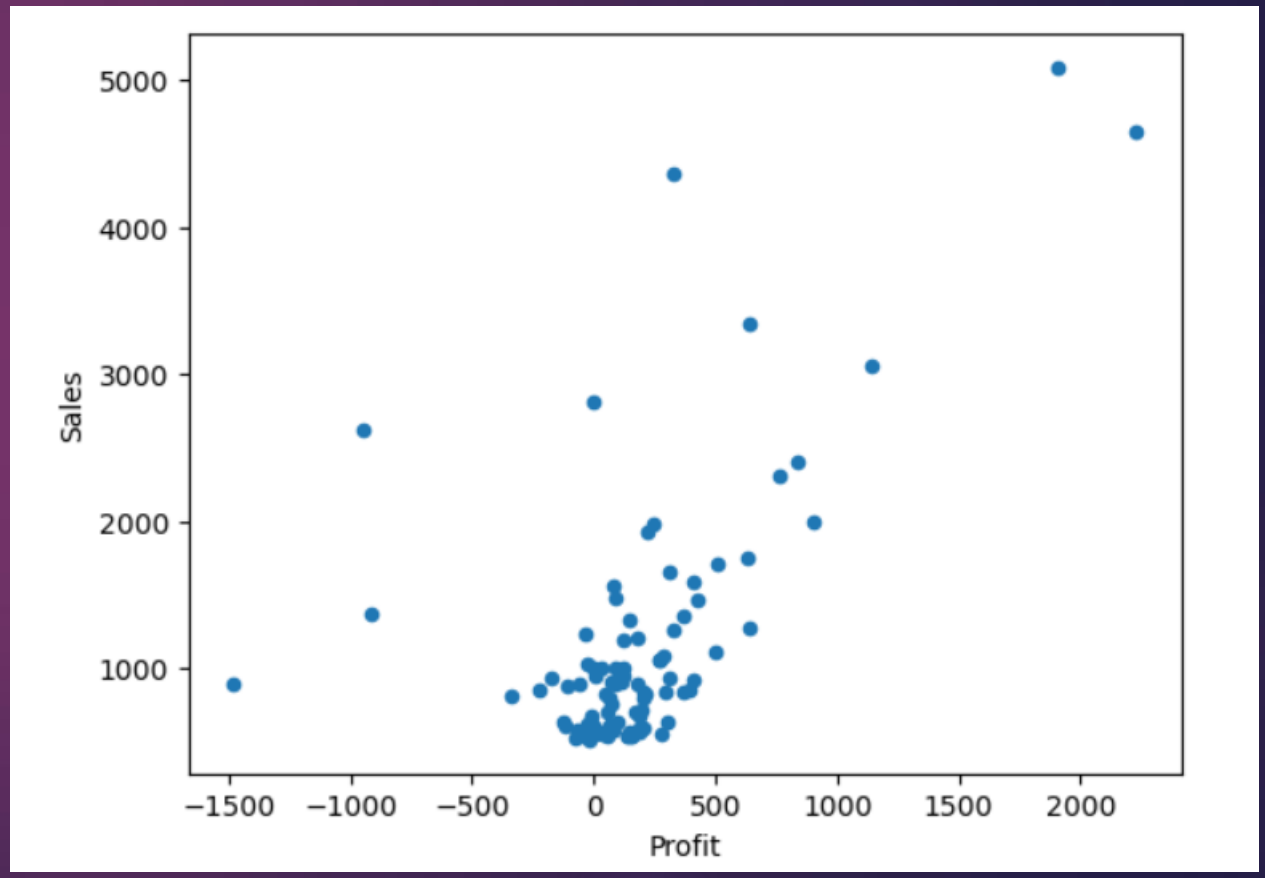
Sales,category

- Clearly graph shows non-linearity between sales(dependent variable) and independent variable category



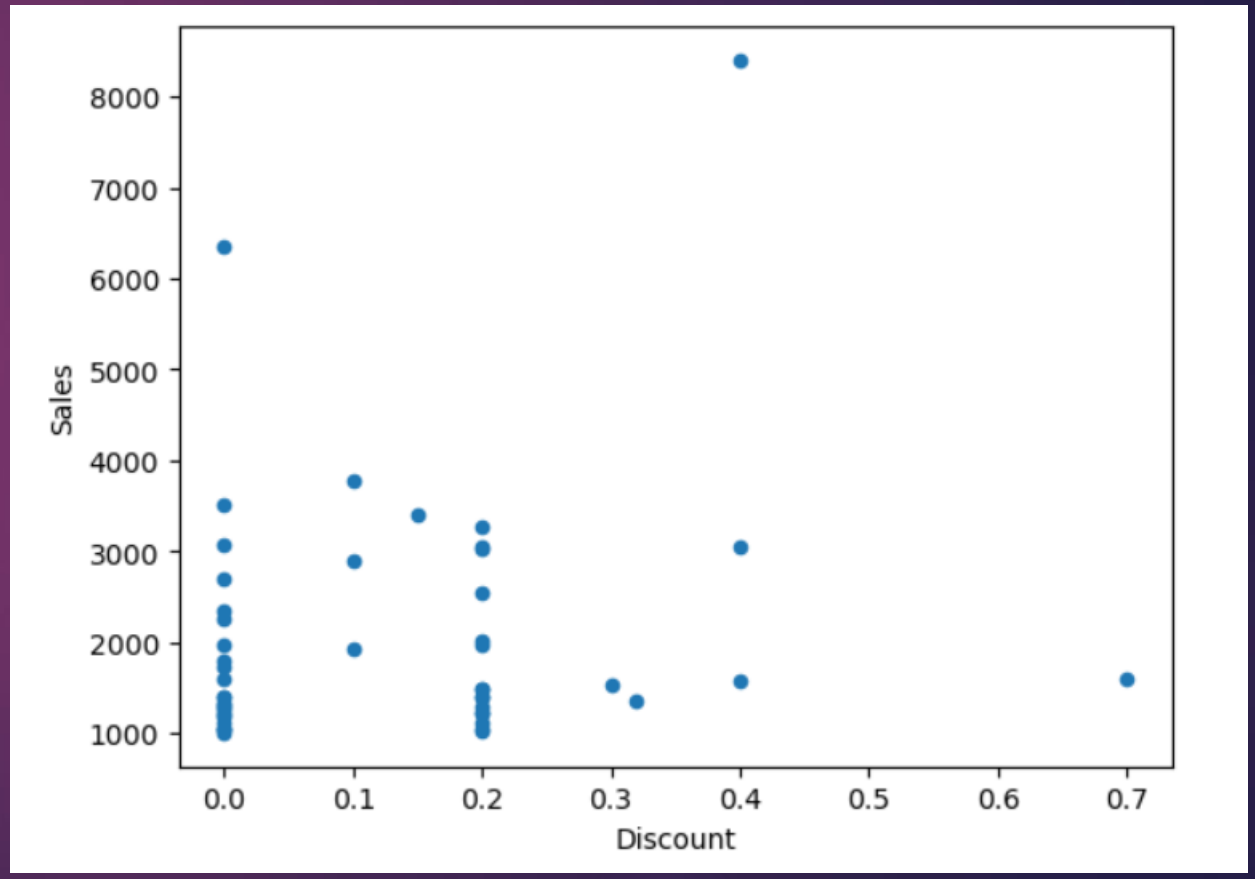
sales,profit-key findings

- eliminate negative profits,consider it as loss
- sales and profit are linearly dependent as sales(dependent variable) increases profit also increases



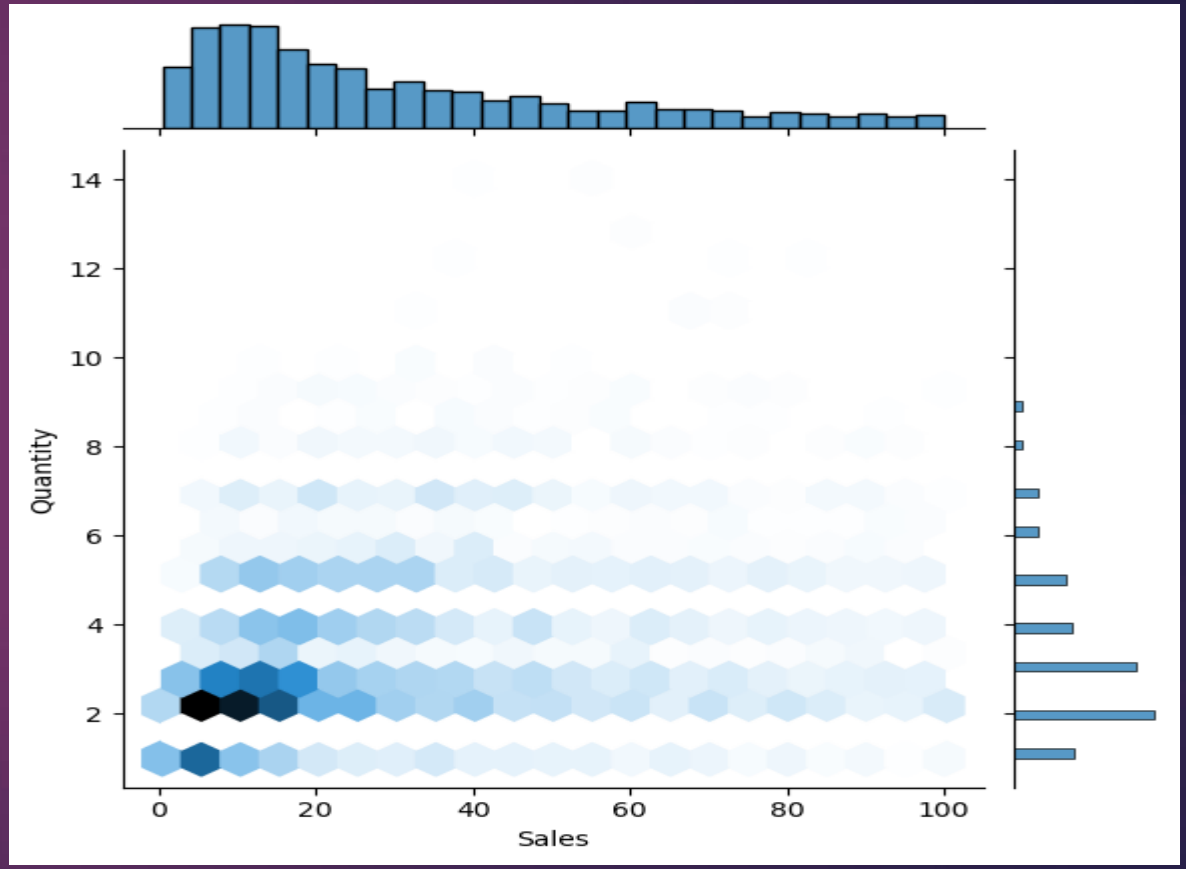
sales,discount(sales-dependent,discount-independent)

- no relationship can be seen in above plot



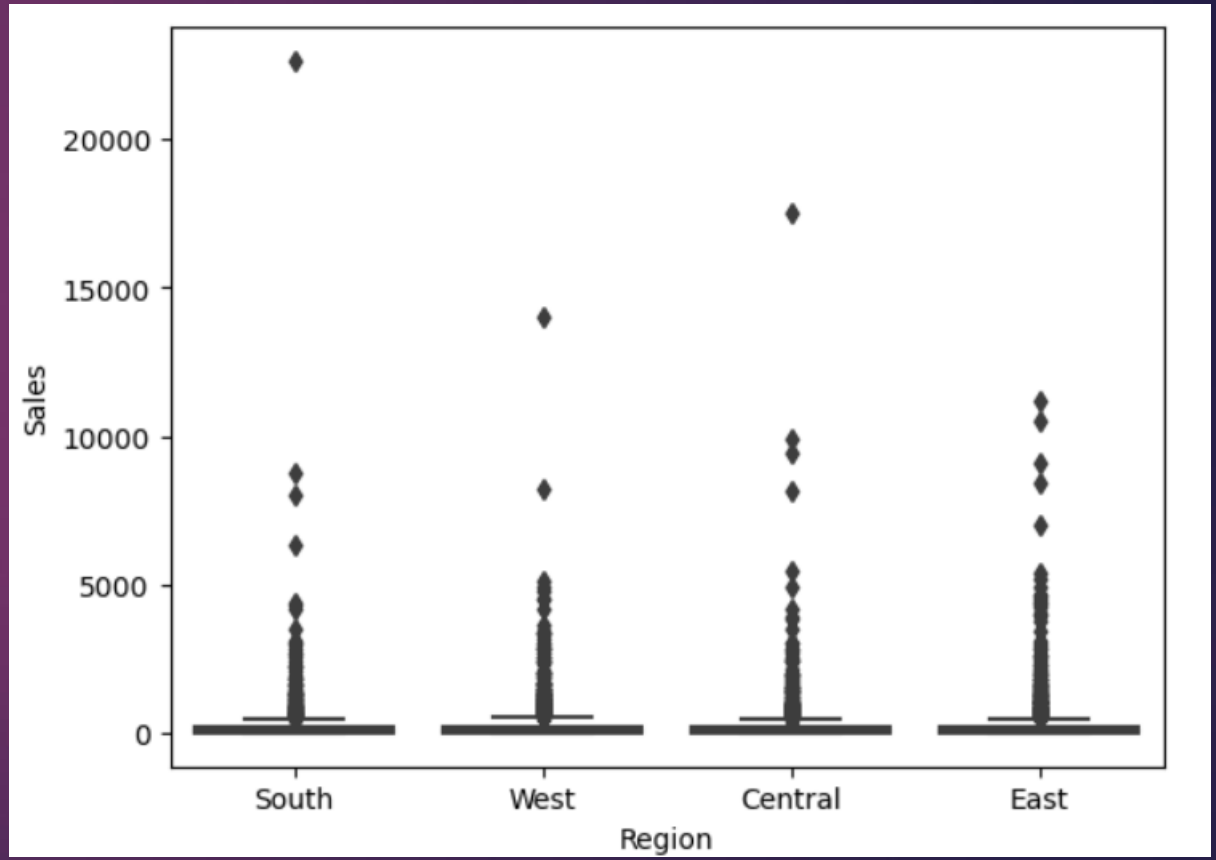
Plotting using Seaborn,jointplot,hexbin

- The color of each hexagon represents the number of data points within that bin. A darker color hexbin means that there are more observations, or more density, within that region.



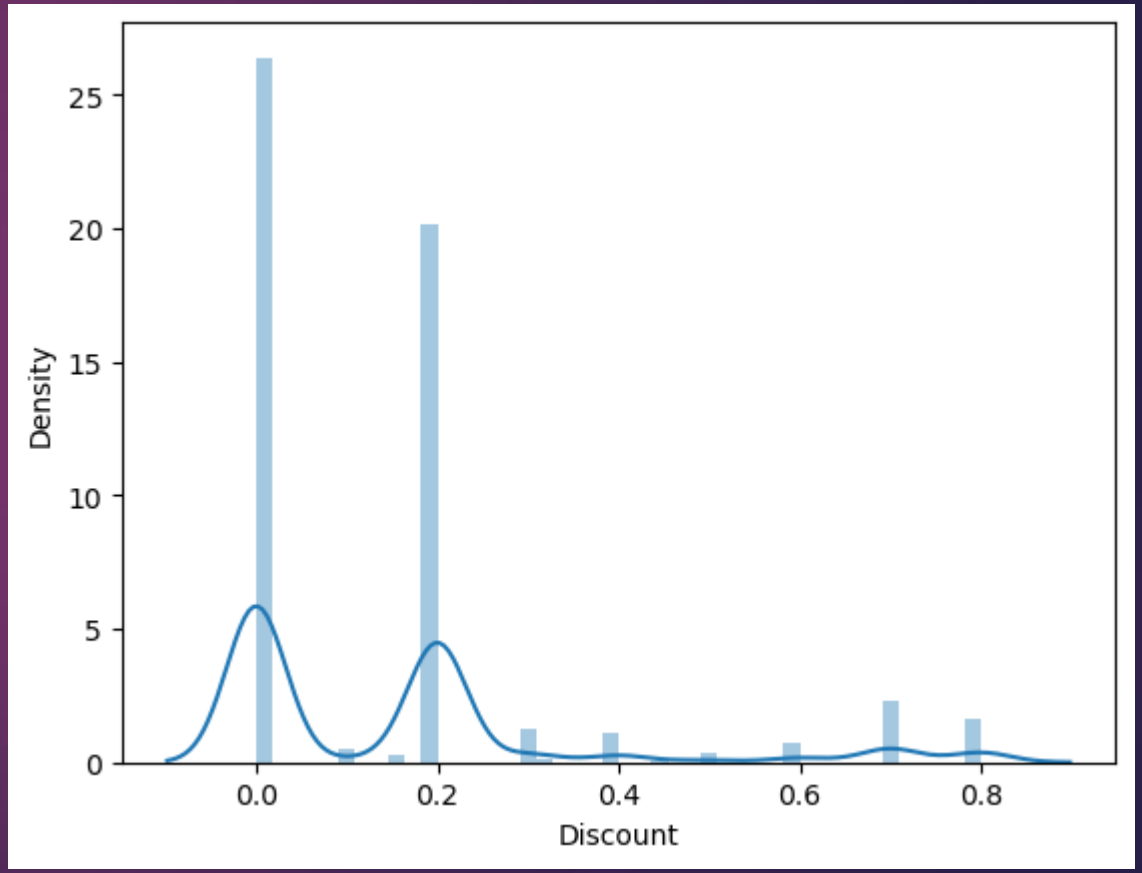
combined Boxplot,sales vs region

- Combined Boxplot of sales in various regions



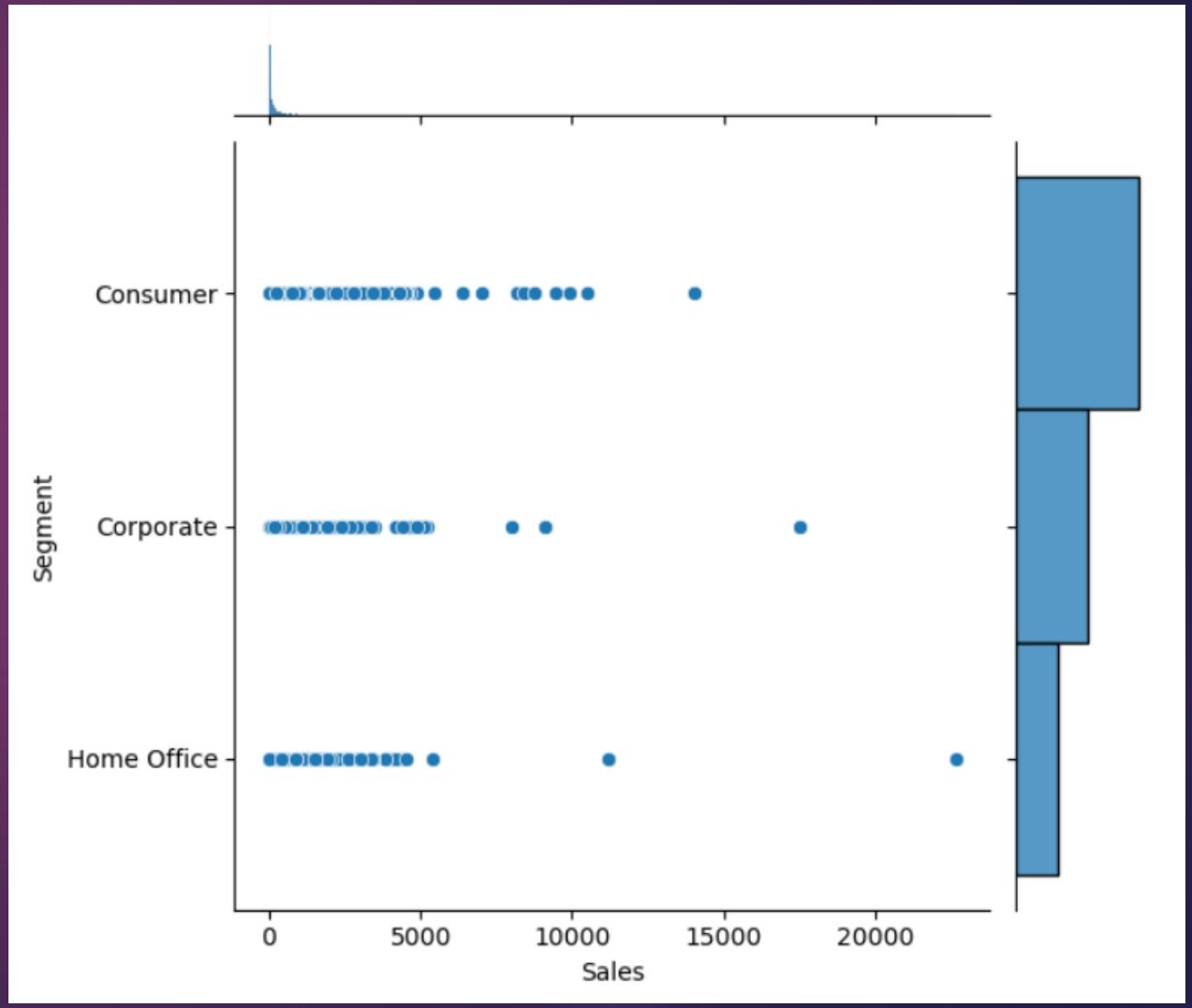
distplot

- we could see the discounts given more is in the range of 0-20% and rarely the discounts given are 40%,60%,80%



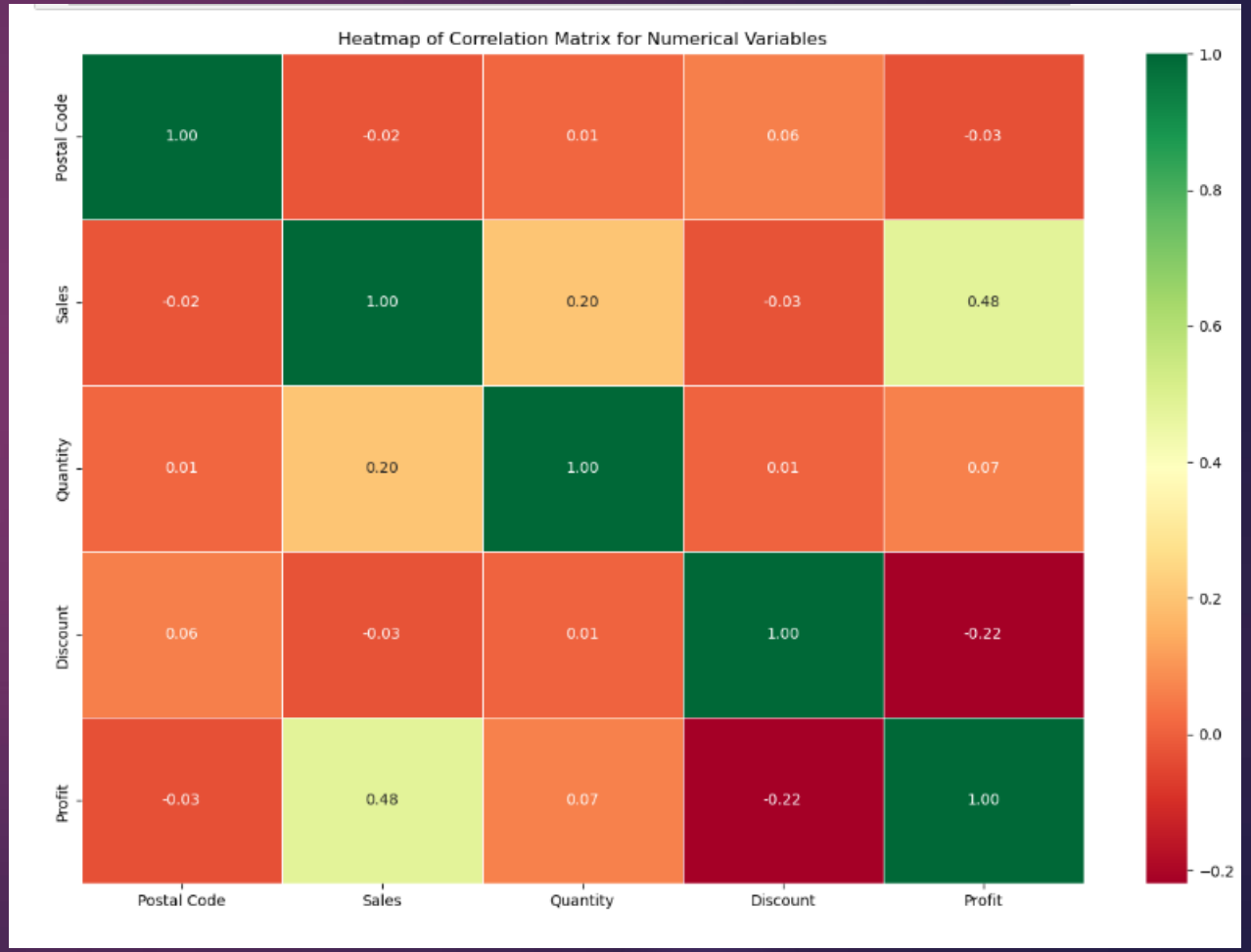
Jointplot, segment vs sales

- Consumer segment has more sales,than corporate and Home office



Multivariate analysis-Heatmap

- Profit and discount is least correlated => less correlated
- sales and profit have moderate correlation
- quantity and sales are correlated



Thank you