Name : Gembali Rakesh

Reg No : 12009254

Course code : INT-353

My Data set : Sample superstore

My GitHub link :

https://github.com/rakeshgem/Sample_super_store/blob/main/Sample_super_store.ipynb

My dataset link :

https://www.kaggle.com/datasets/bravehart101/sample-supermarket-dataset

Project introduction: -

I have chosen a data set from Kaggle it is about a superstore in the United States, it contains 13 columns namely ship mode, segment, country, city, state, postal code, region, categories of Office supplies, furniture, and technology, sub category`s like Binder's paper, phones, art, furnishings, envelopes, bookcases, chairs, phones, storage, labels, accessories, fasteners, tables, supplies, machines, appliances, sales column, quantity, discount and finally profit.

I mainly chose this data set to analyze how a store work under different situations and how they tackle profit or loss, while they are in profit how they got profit and how to increase it further, and if they are in loss how they tend to increase the sales and background works like how they analyze to make profit and improvement in sales. Not only for one store huge number of stores near 10,000

Domain: -

Superstore, a large retail store operated on a self-service basis, selling groceries, fresh produce, bakery, and dairy products, and sometimes an assortment of non-food goods.

About my data set: -

It is all about sample superstores in the United States, it contains 13 columns

are as follows: -

- Ship mode contains the class of shipping modes: standard, second, and first.
- Segment contains segment categories: consumer, corporate, and home office.
- Country contains the country of the super store in this data set all the super stores are taken from the United States only.
- City it contains the city of the superstore.
- State it contains the state of the superstore.
- Postal code.
- Region, it contains the region that is from west or east or central or south or north.
- Category contains the type of products, in its three categories, are there office supplies, furniture, and technology?
- Sub-category contains different products depending on the category of the superstore.

- Sales contains the number of times sold depending on the sub-category.
- Quantity contains the number of products sold at a time depending on the sub-category.
- Discount contains how much discount got on an average of sales, and quantity.
- Profit.

Analysis I will do: -

- I will analyze the sales of certain products and the profit and discount that the superstore provides.
- I will analyze categories and sub-categories depending on sales and profit so that I will get the relation between sales and profit, and in which categories the sales are more and profit also more, or sales less and profit more, or sales less and profit also less.
- I will analyze the profit or loss by giving discounts.
- I will analyze sub-categories that which products are more in the sale and which products getting more profits, and how the super stores give the discount depending on the product and category.
- I will analyze profit and sales depending on shipping mode.
- I will analyze the segment that is in which segment category and how the sales and profits going on.
- I will analyze which region is attractive by seeing sales and in which region the profits are more.

Top 5 rows in my data set:

	Ship Mode	Segment	Country	City	State	Postal Code	Region	Category	Sub-Category	Sales	Quantity	Discount	Profit
0	Second Class	Consumer	United States	Henderson	Kentucky	42420	South	Furniture	Bookcases	261.9600	2	0.00	41.9136
1	Second Class	Consumer	United States	Henderson	Kentucky	42420	South	Furniture	Chairs	731.9400	3	0.00	219.5820
2	Second Class	Corporate	United States	Los Angeles	California	90036	West	Office Supplies	Labels	14.6200	2	0.00	6.8714
3	Standard Class	Consumer	United States	Fort Lauderdale	Florida	33311	South	Furniture	Tables	957.5775	5	0.45	-383.0310
4	Standard Class	Consumer	United States	Fort Lauderdale	Florida	33311	South	Office Supplies	Storage	22.3680	2	0.20	2.5164

Data cleaning:

Checking information:

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 9994 entries, 0 to 9993
Data columns (total 13 columns):
 #   Column        Non-Null Count  Dtype
---  ------        --------------  -----
 0   Ship Mode     9994 non-null   object
 1   Segment       9994 non-null   object
 2   Country       9994 non-null   object
 3   City          9994 non-null   object
 4   State         9994 non-null   object
 5   Postal Code   9994 non-null   int64
 6   Region        9994 non-null   object
 7   Category      9994 non-null   object
 8   Sub-Category  9994 non-null   object
 9   Sales         9994 non-null   float64
 10  Quantity      9994 non-null   int64
 11  Discount      9994 non-null   float64
 12  Profit        9994 non-null   float64
dtypes: float64(3), int64(2), object(8)
memory usage: 1015.1+ KB
```

No missing values and all data types are okay

From the information of my data set 8 categorical and 5 numerical columns are there and no null values are there. If null values are there then two ways to handle them:

1. Deleting the missing values:

   With this method, we delete the row if some random null values are there whether a column contains more than 30% null values we can drop the column but the main problem with this method we are deleting the data but it also cost some money to gather the data.

2. Imputing null values:

   There are different ways of replacing the missing values:

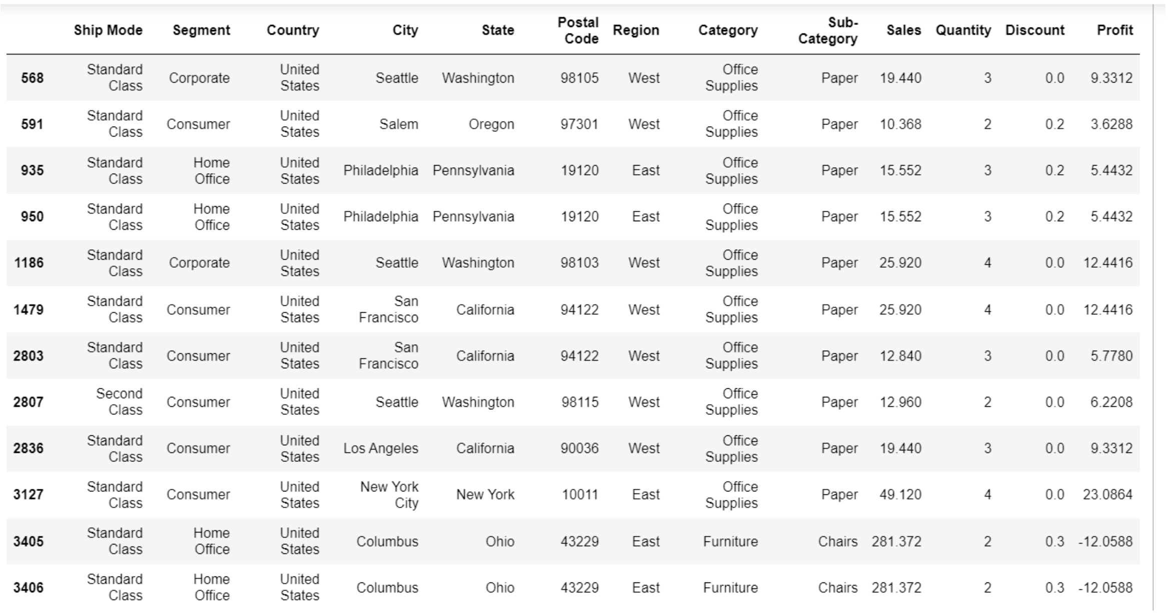
   a. Replacing With Arbitrary Value

   b. Replacing with mean

   c. Replacing with mode

   d. Replacing with median etc.

Checking for duplicate values:

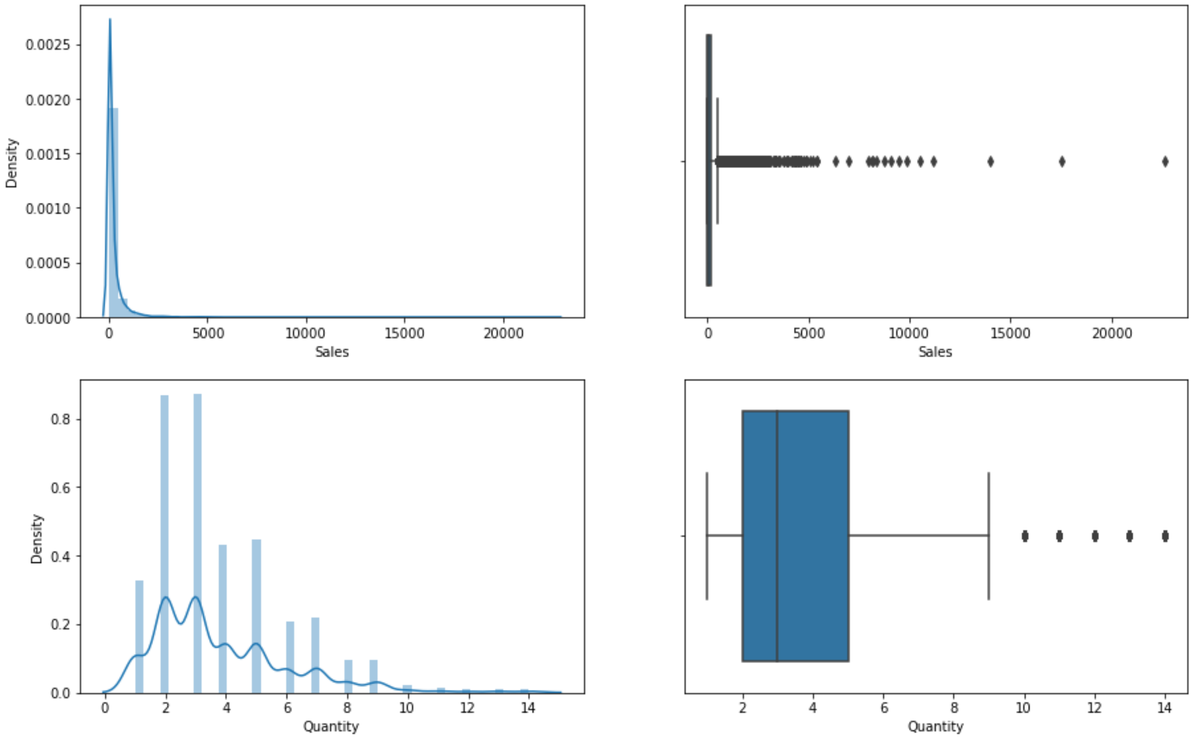


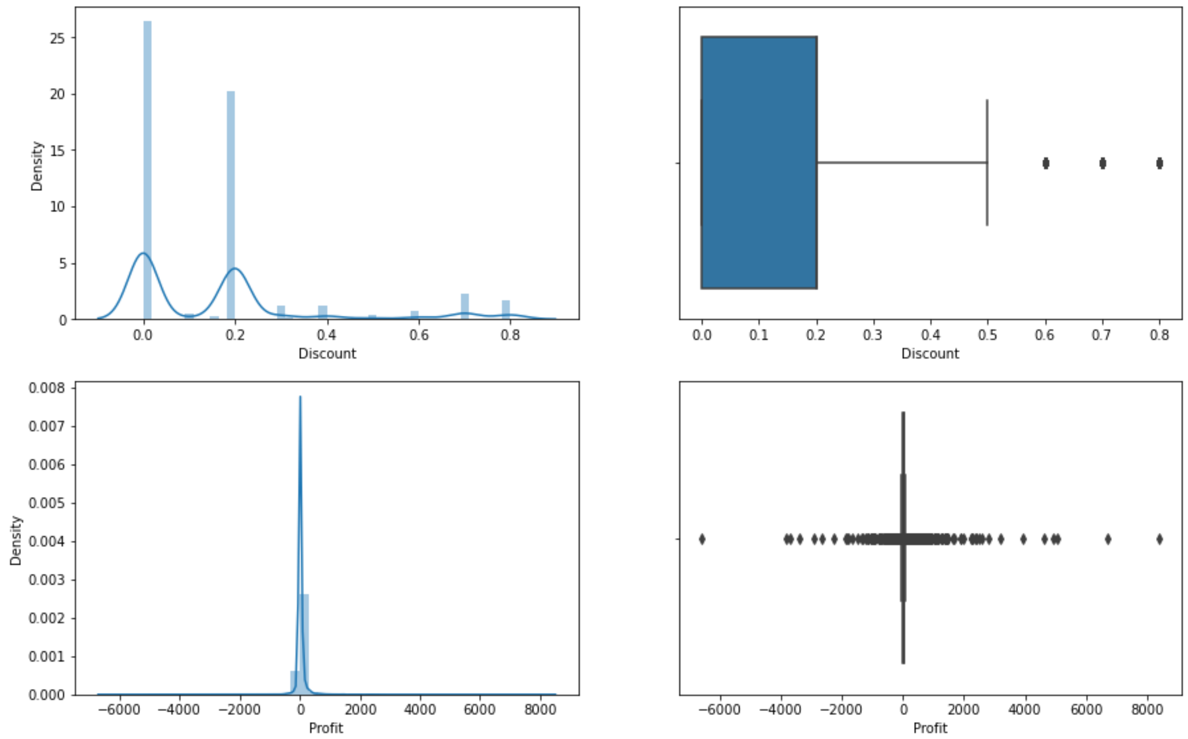
	Ship Mode	Segment	Country	City	State	Postal Code	Region	Category	Sub-Category	Sales	Quantity	Discount	Profit
568	Standard Class	Corporate	United States	Seattle	Washington	98105	West	Office Supplies	Paper	19.440	3	0.0	9.3312
591	Standard Class	Consumer	United States	Salem	Oregon	97301	West	Office Supplies	Paper	10.368	2	0.2	3.6288
935	Standard Class	Home Office	United States	Philadelphia	Pennsylvania	19120	East	Office Supplies	Paper	15.552	3	0.2	5.4432
950	Standard Class	Home Office	United States	Philadelphia	Pennsylvania	19120	East	Office Supplies	Paper	15.552	3	0.2	5.4432
1186	Standard Class	Corporate	United States	Seattle	Washington	98103	West	Office Supplies	Paper	25.920	4	0.0	12.4416
1479	Standard Class	Consumer	United States	San Francisco	California	94122	West	Office Supplies	Paper	25.920	4	0.0	12.4416
2803	Standard Class	Consumer	United States	San Francisco	California	94122	West	Office Supplies	Paper	12.840	3	0.0	5.7780
2807	Second Class	Consumer	United States	Seattle	Washington	98115	West	Office Supplies	Paper	12.960	2	0.0	6.2208
2836	Standard Class	Consumer	United States	Los Angeles	California	90036	West	Office Supplies	Paper	19.440	3	0.0	9.3312
3127	Standard Class	Consumer	United States	New York City	New York	10011	East	Office Supplies	Paper	49.120	4	0.0	23.0864
3405	Standard Class	Home Office	United States	Columbus	Ohio	43229	East	Furniture	Chairs	281.372	2	0.3	-12.0588
3406	Standard Class	Home Office	United States	Columbus	Ohio	43229	East	Furniture	Chairs	281.372	2	0.3	-12.0588

etc

By checking my data set some duplicate values are there so I will remove those.

Finding outliers:





For Normal distributions: we need to use empirical relations of Normal distribution.

– The data points which fall below mean-3*(sigma) or above mean+3*(sigma) are outliers.

where mean and sigma are a particular column's average value and standard deviation.

For Skewed distributions: Use Inter-Quartile Range (IQR) proximity rule.

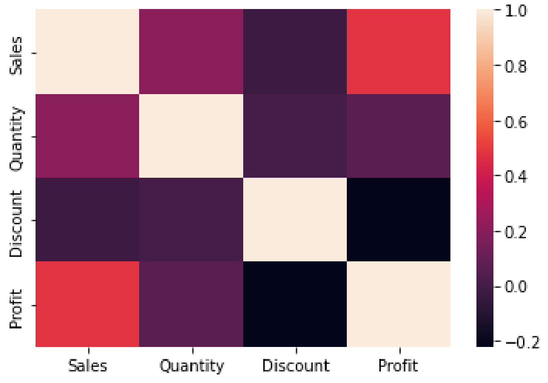
– The data points which fall below Q1 – 1.5 IQR or above Q3 + 1.5 IQR are outliers.

where Q1 and Q3 are the 25th and 75th percentile of the dataset respectively, and IQR represents the inter-quartile range and is given by Q3 – Q1.

By using these distributions I removed the outliers.

Exploratory Data Analysis:

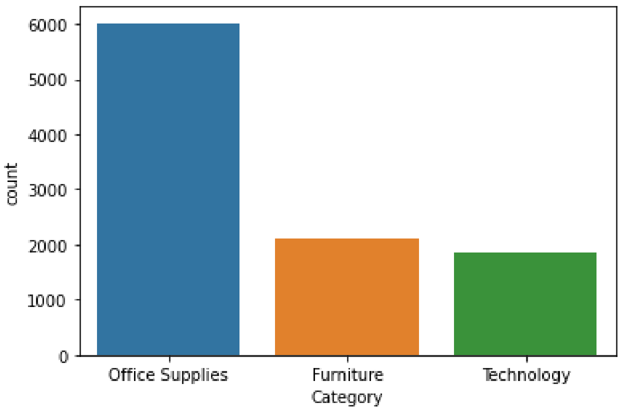
First, I will plot a heat map to find any correlation is there between the columns:



By the heat map, we notice there's a correlation between (profit, discount) and (profit, sales) and (quantity, sales)

Univariant analysis:

- Office Supplies has the highest number

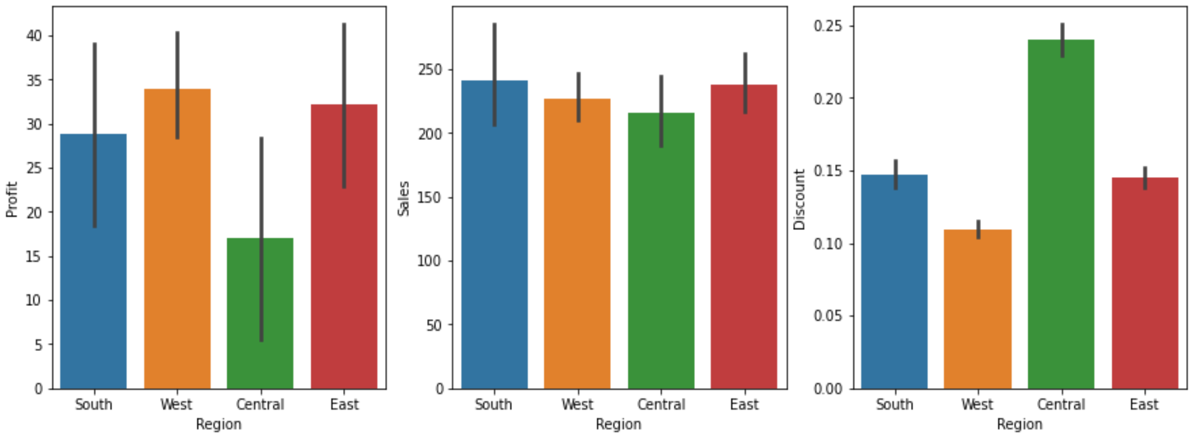


Bivariate analysis:

Categorical to numerical:

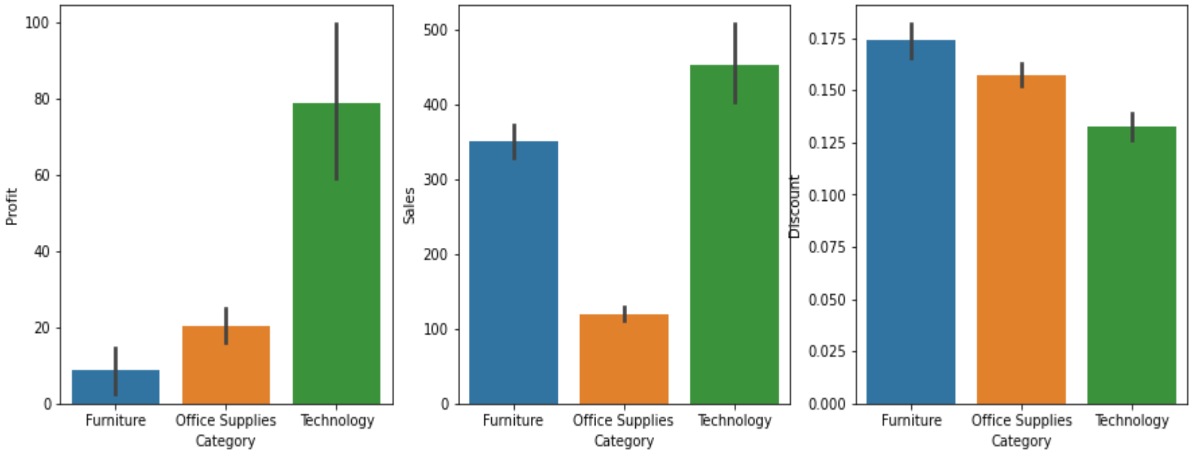
I will check profit, sales, and discounts based on the categorical columns:

For each region:



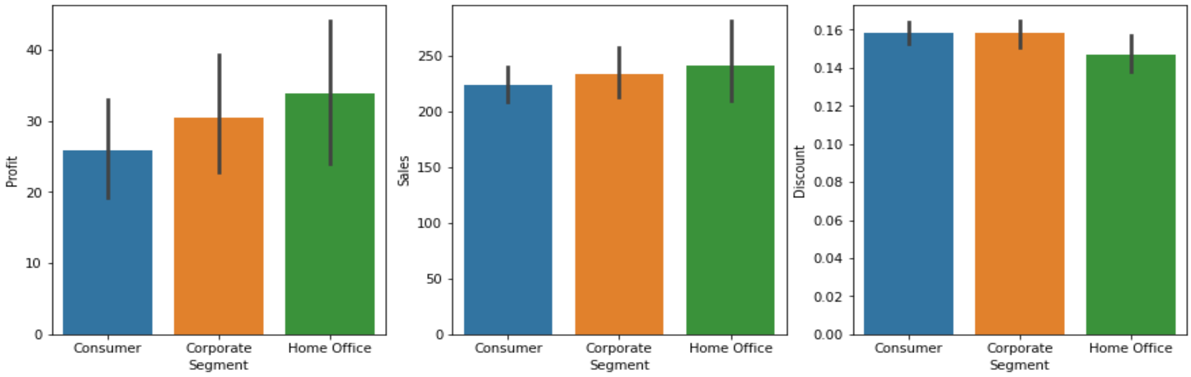
- West region has the highest profit and finds the same region has the lowest discount
- South has the highest sales
- Central has the lowest profit and the highest discount, maybe that is why.

For each category:



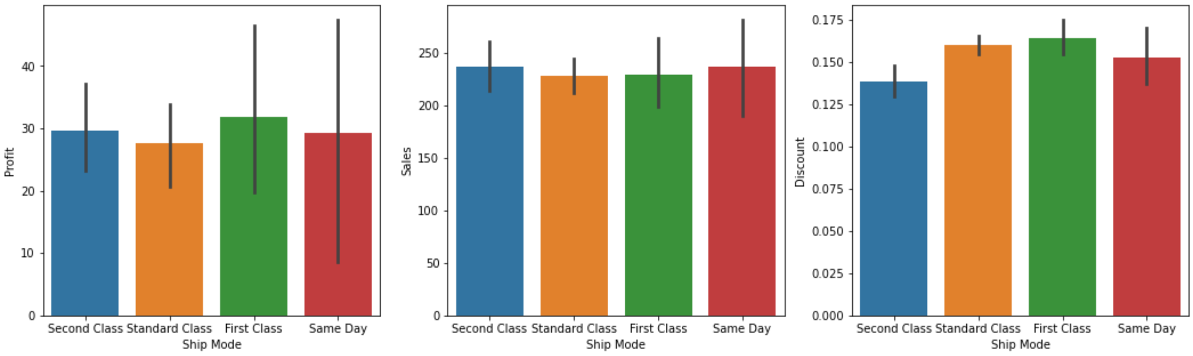
- Furniture sales are high but have very low profit maybe a high discount is the reason.

For each Segment:



- Home Office has the lowest discount but has also the highest profit as sales

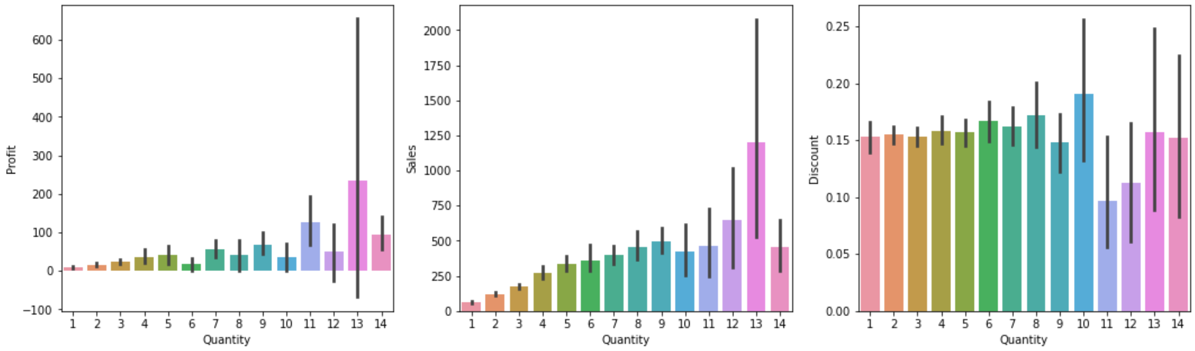
For each ship mode:



- Same-day shipping has the highest sales
- First class has the highest discount but also the highest profit

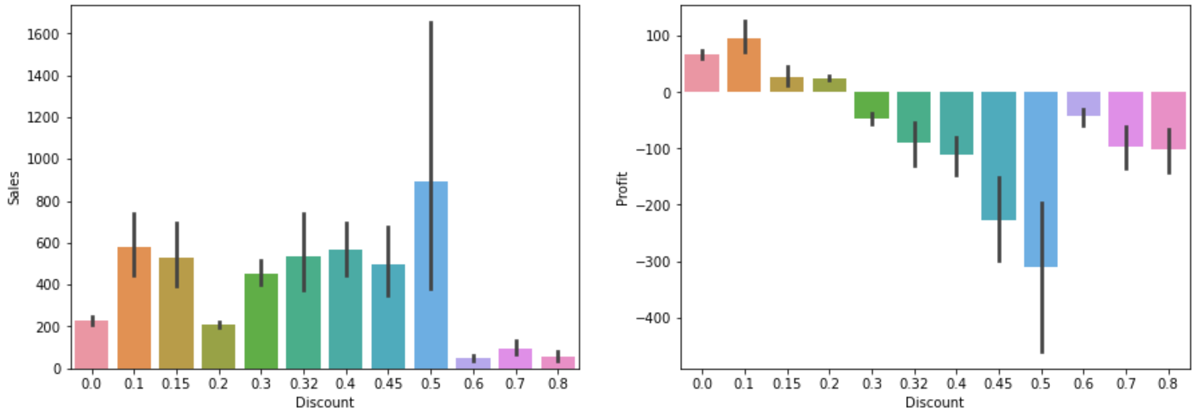
Numerical to numerical:

For quantity:



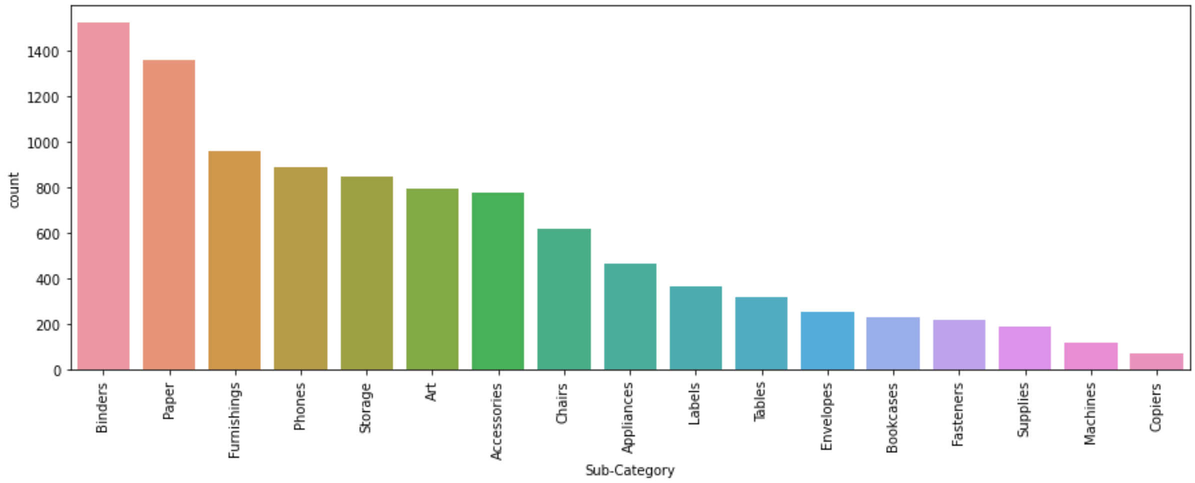
- quantity 13 is the highest in sales and profit but number 10 is the highest in discount and gains lower profit
- we need to balance between quantity and discount

For Discount:



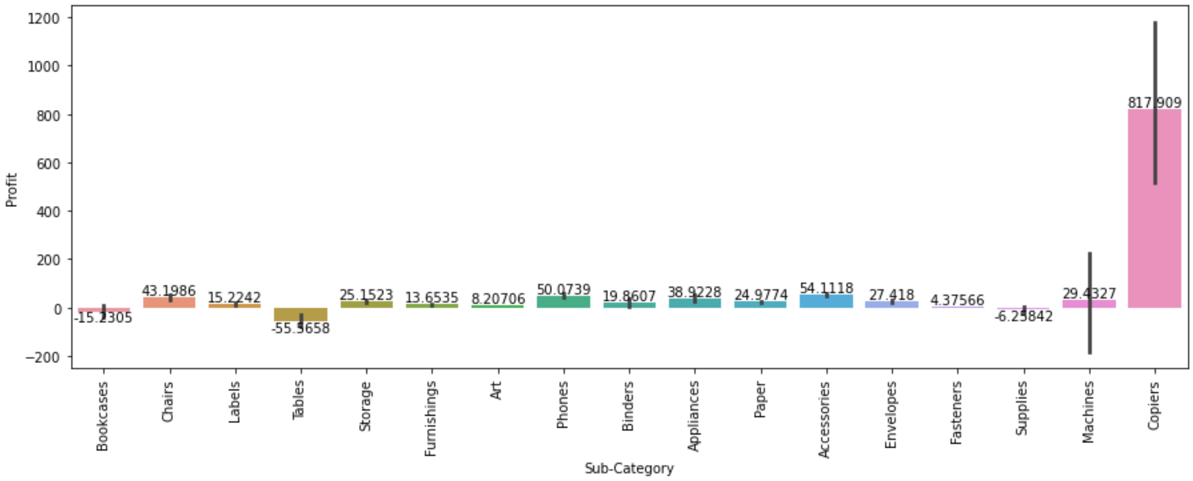
- We note that when the discount is high, the profit is negative

check the number of each sub-category:



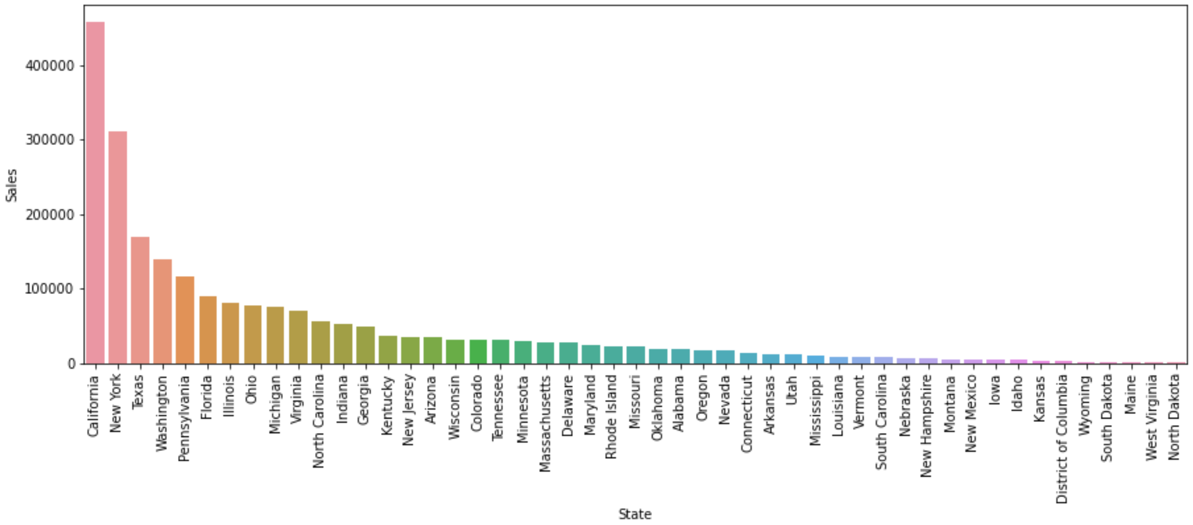
- Sales of Binders as Paper are the highest.

check Sub-category with profit:



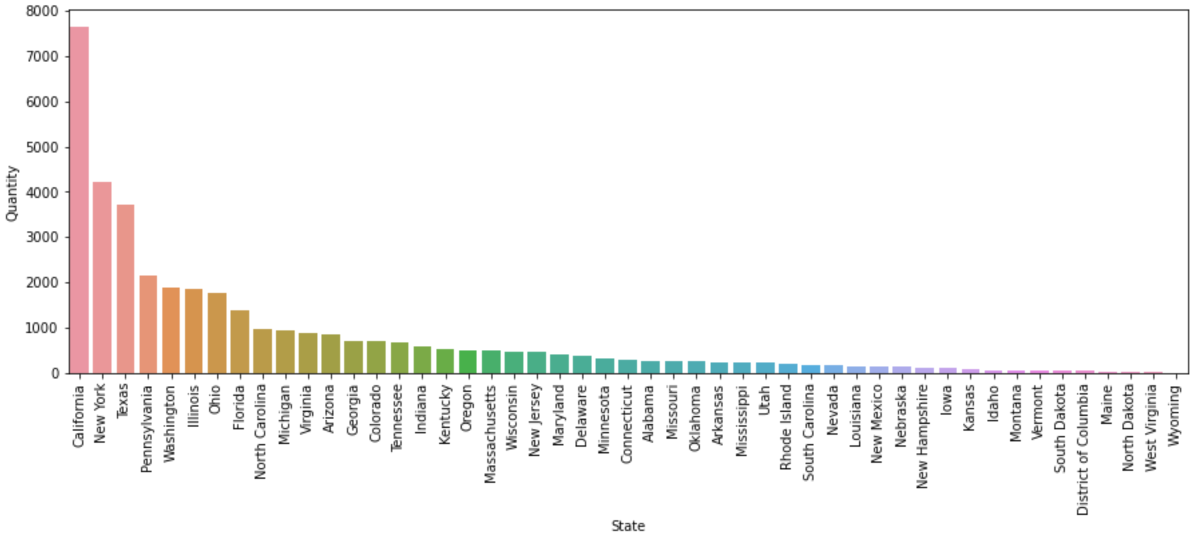
- Tables Supplies and Bookcases have negative profit

check sales:



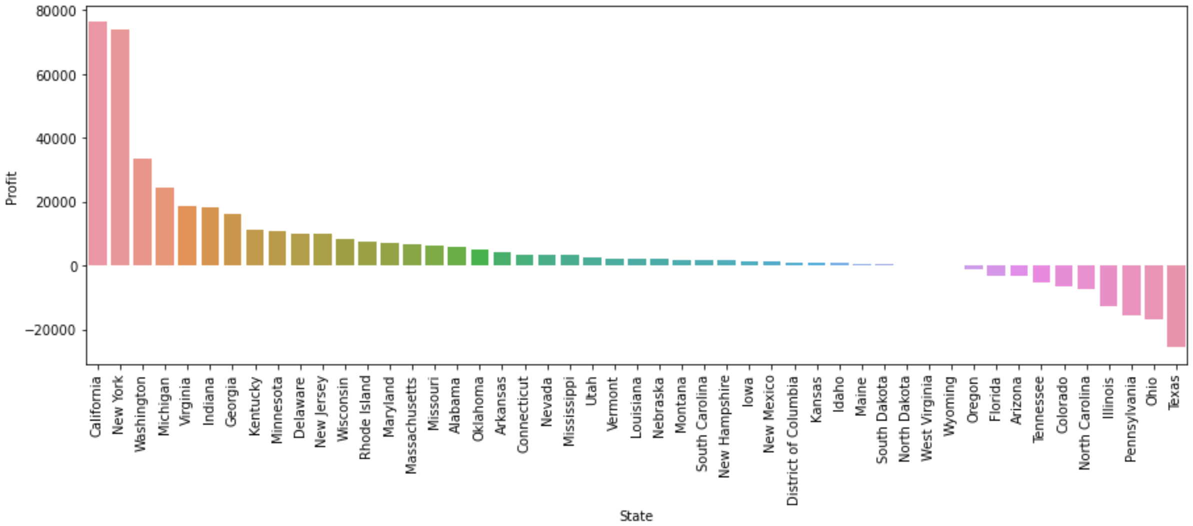
- California and New York have the maximum sales
- Many states have lower sales

Check Quantity:



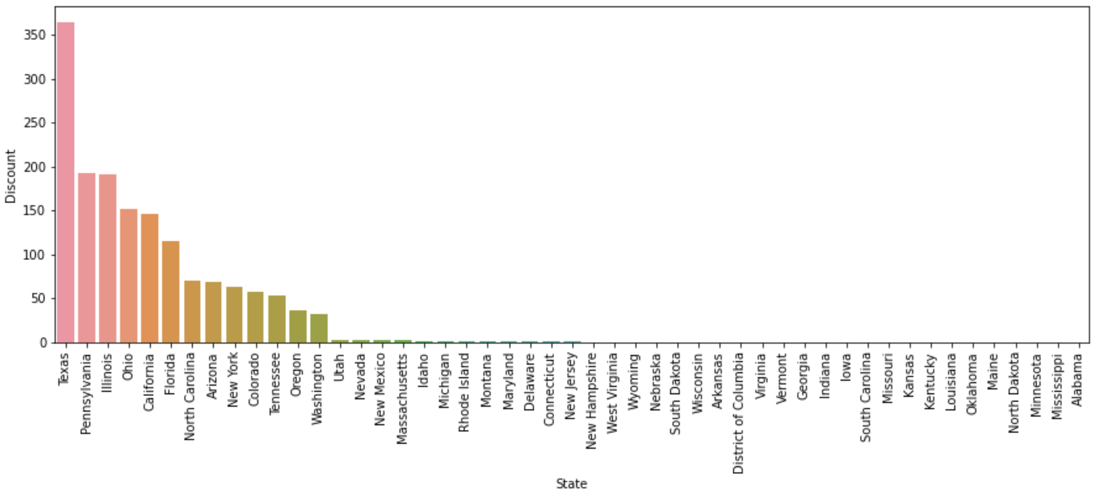
- In California and New York people like to take more quantity of products at a time.

check Profit:



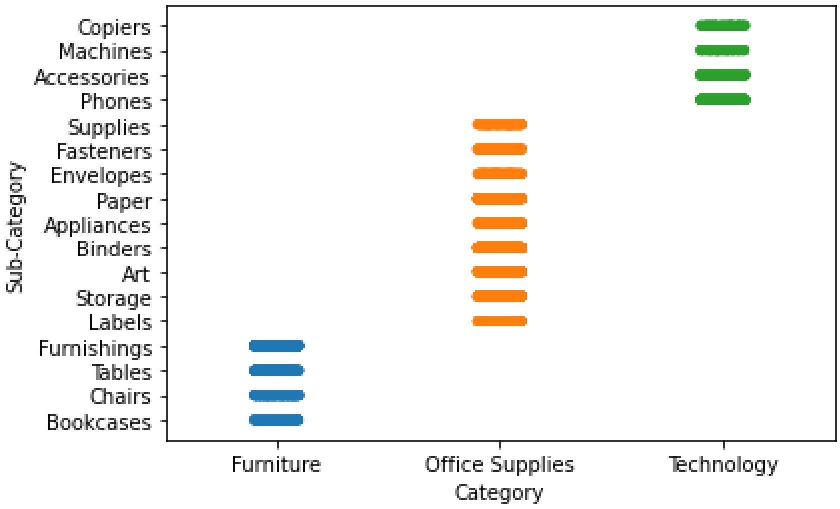
- California and New York have the highest profit
- Some states have good sales but profits in negative

Check Discount for the state:



- high discounts sometimes cause to loss profit and we can see that

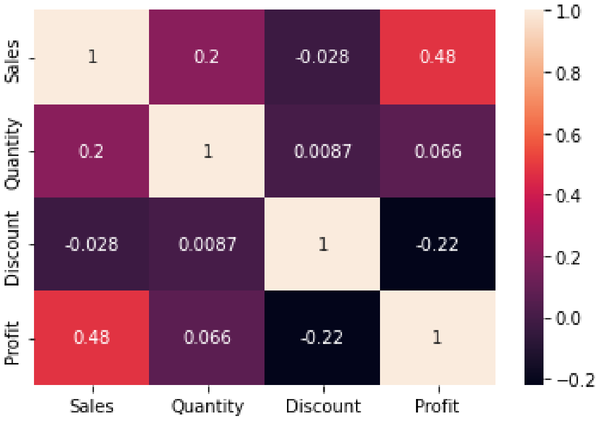
Category to category:

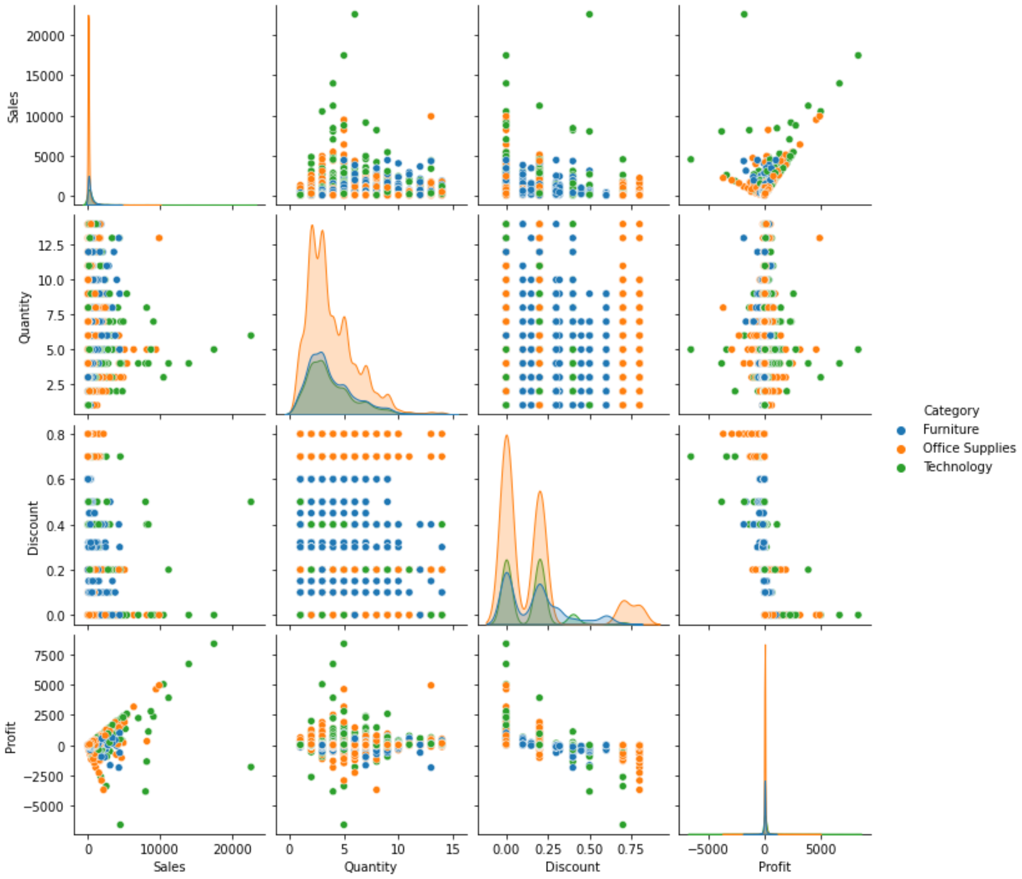


- From this graph we can understand that different sub-category of categories

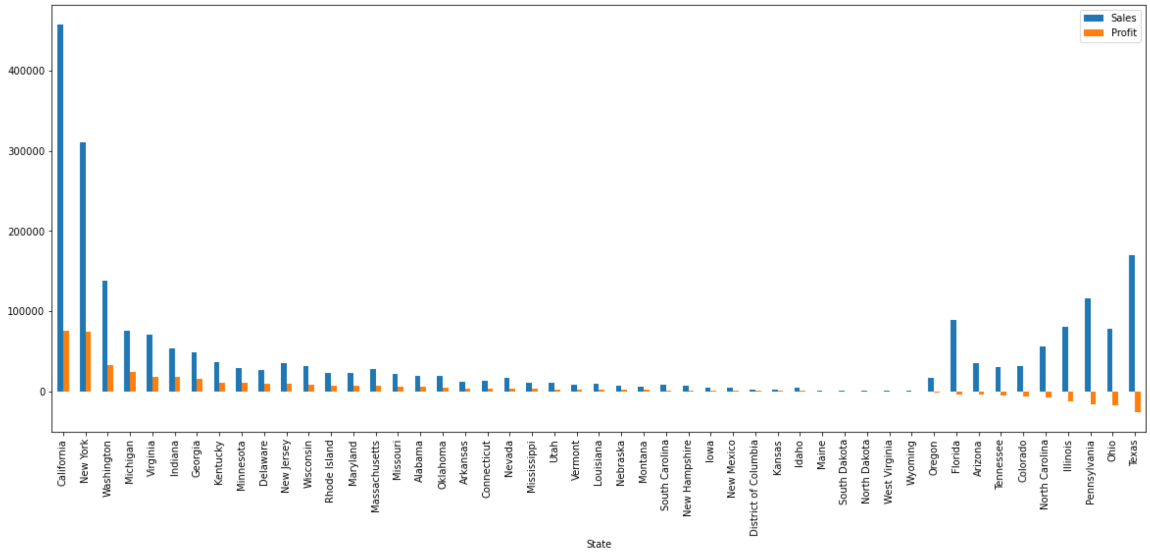
Multivariant analysis:

- From this heat map profit and sales have more correlation and then sales and quantity have a good correlation.





- The Technology category is high in sales, low in discounts, and high in profit.
- The Furniture category is low in profit, high in discount, and compared to office supplies high in sales.
- The Office supplies category is low in sales, compared to furniture high in profit, and compared to technology high discount.



- Not all good sales make a good profit

Conclusion:

- For each region:
  - West region has the highest profit and finds the same region has the lowest discount
  - South has the highest sales
  - Central has the lowest profit and the highest discount, maybe that is why.
- For each category:
  - Furniture sales are high but have very low profit maybe a high discount is the reason.
- For each segment:
  - Home Office has the lowest discount but has also the highest profit as sales
- For each ship mode:
  - Same-day shipping has the highest sales
  - First class has the highest discount but also the highest profit

- Sales and Profit are not linear for most States
- Central region needs more attention
- Furniture and Office Supplies have high loss profit with high discount
- Office Supplies has maximum loss at 80% and 0% discount
- Furniture and Technology have a maximum loss between 30% to 50%
- Technology also has maximum loss profit at 70%
- Tables and Supplies and bookcases in the sub-category have negative profit

Overall analysis and predictions:

- Technology earns more profit compared to furniture
- Same-day shipping earns high sales
- West region has the highest profit
- Vermont state has a good profit with low sales
- Discount with 50% and less gain more profit
- Need to give more discounts on Furniture to attract customers
- In segment need to give more discounts in Consumer and in Home-Office promote more for higher the profit