Infosys Springboard Project Report

DataVista: Sales Data Analysis and Visualization

- Shivani Singh

Project Documentation: Retail Sales Data

Objective: Given the dataset of a retail company, we need to use python libraries to look into the dataset to gain insights of data and identify interesting trends/patterns.

Libraries used:

- import numpy as np: NumPy is used for numerical computing and working with arrays.
- import pandas as pd: Pandas is used for working with datasets for analyzing, cleaning, exploring, and manipulating data.
- import matplotlib.pyplot as plt: Matplotlib is used for data visualization, creating and customizing graphs.
- import math as math: Math provides mathematical functions like square root, trigonometry, etc.
- import seaborn as sns: Seaborn is used to create statistical graphics and visualize data. It is based on matplotlib.

Loading data from CSV

Loading the data and naming it as 'sales'.

- sales = pd.read_csv('/Users/shivanisingh/Documents/retail_sales_dataset.csv')

Knowing about the data

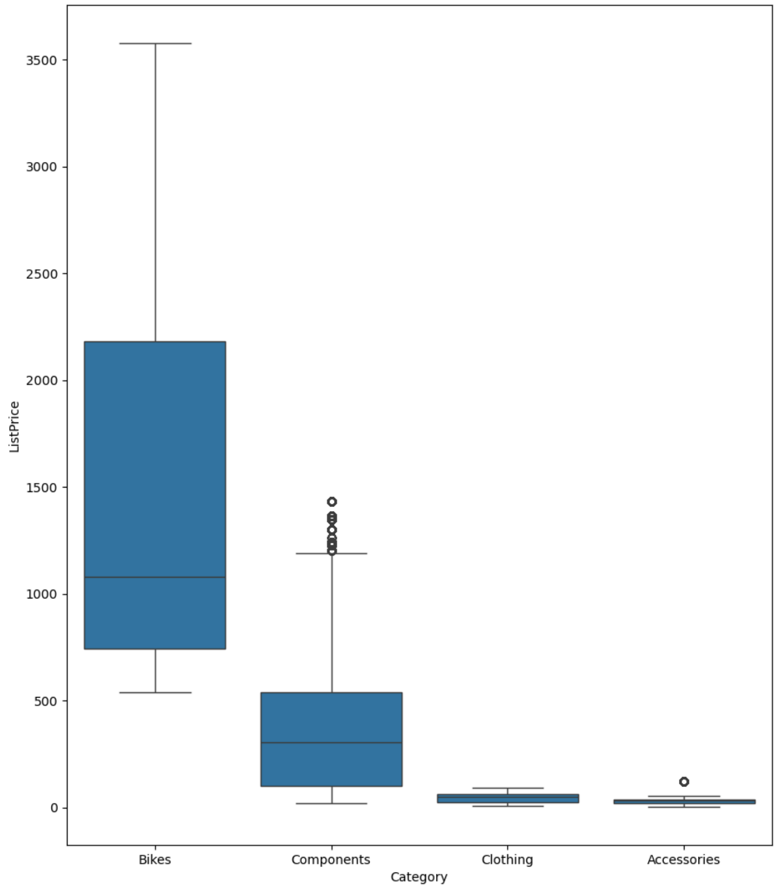
- sales.head(10) : showing first 10 rows
- sales.shape: Check the number of rows and columns that your dataset has.
- sales.dtypes: Check datatype of each attribute
- sales.nunique(): Checking distinct values does each column have
- sales.describe(): describe method to check basic statistical measures of your data like count, mean, min, max, std quantiles etc.

Cleaning data

- data.isnull().sum(): checking null values are in each column.
- Some of the products have Quantity as Null, we need to set their quantity to 1.
- Some of the products have Null List Price, set their price as the mean of that product price in other orders.
- Some orders do not have a sales region assigned, you need to remove such order lines
- Some orders have Due date less than Order date, due date should be set as Order data in such cases.
- Checking and removing the outliers from data.

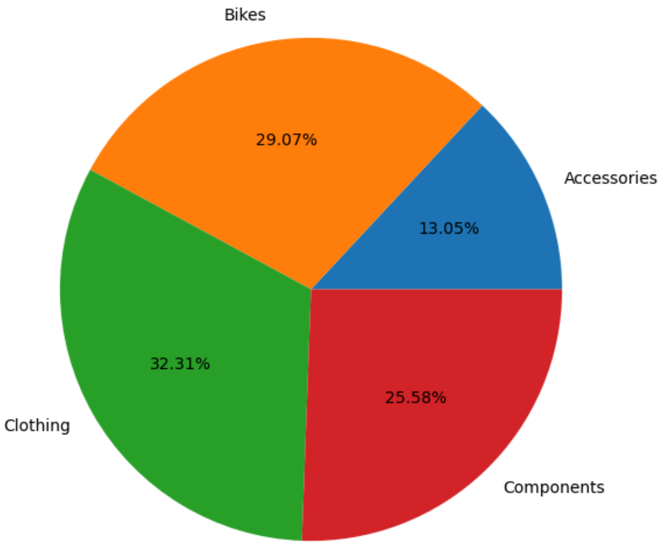
Data Exploration and Visualizations

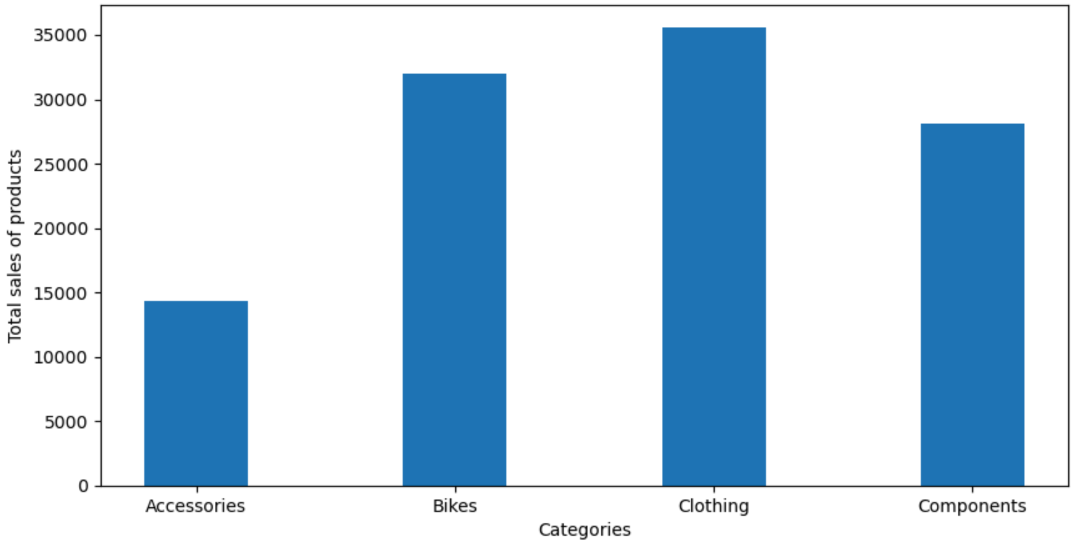
- Checking how many products have been sold under each category.
- **Boxplot** of category vs product list price to compare prices of products across categories.



The box plot shows that List Price of Bikes is very high compared to the rest of the categories. This makes sense because the price of bikes would be a lot more than accessories and clothes. We can also see that the bike List price is positively skewed. This means that most of the bike prices are more than the median price.

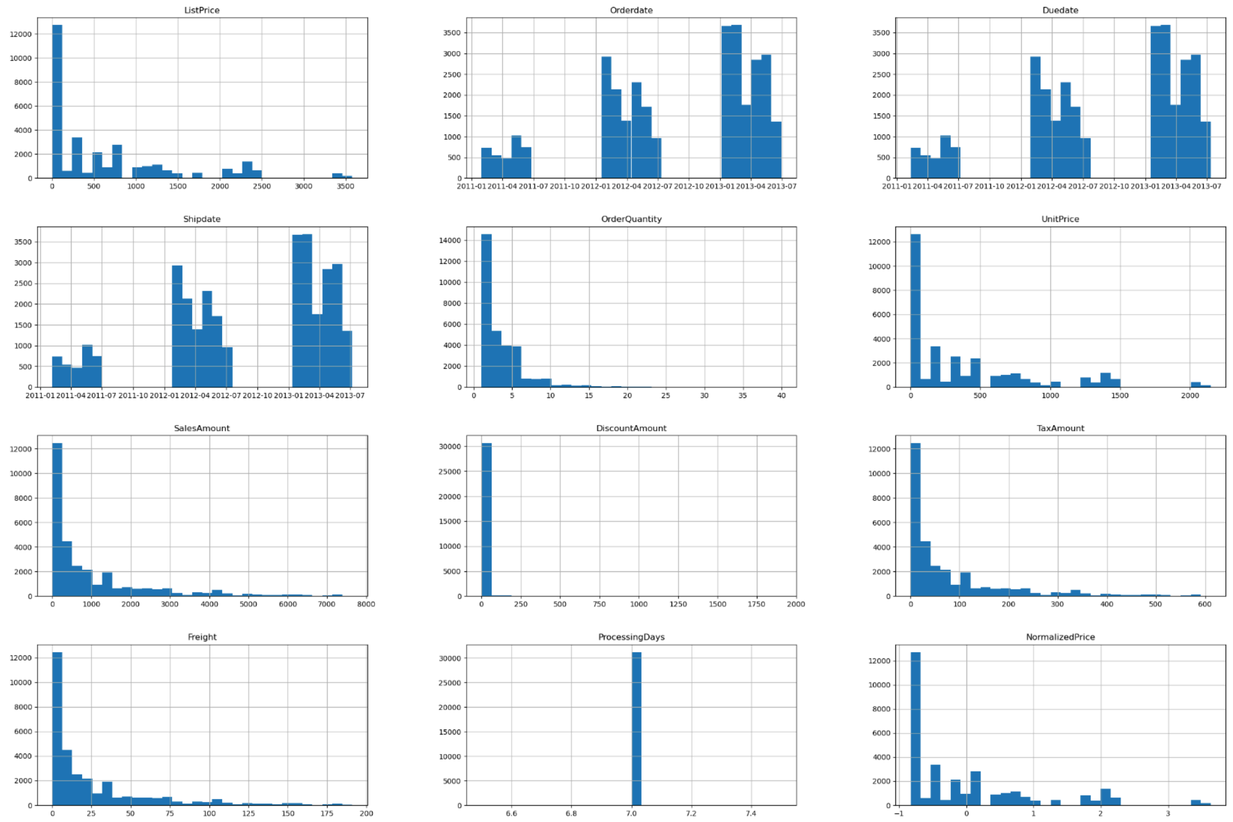
- **Bar chart** and **pie chart** to see total sales of products in each category.





As can be observed from the charts, the highest number of products sold come from Clothing category followed by Bikes then components and then Accessories with the least number of products sold.

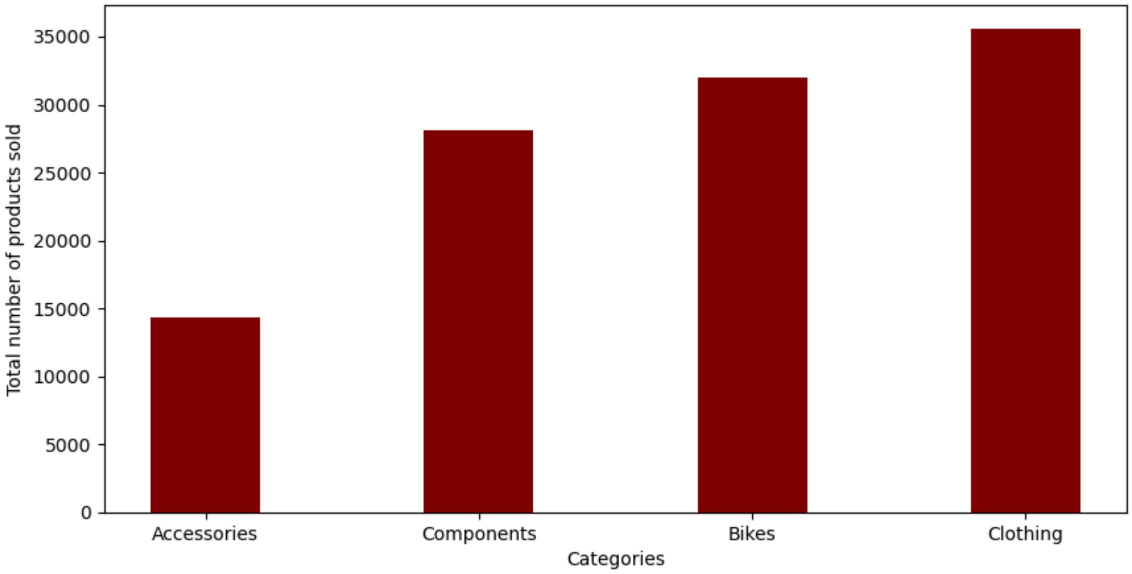
- **Histogram** of all numeric attributes to see their distribution



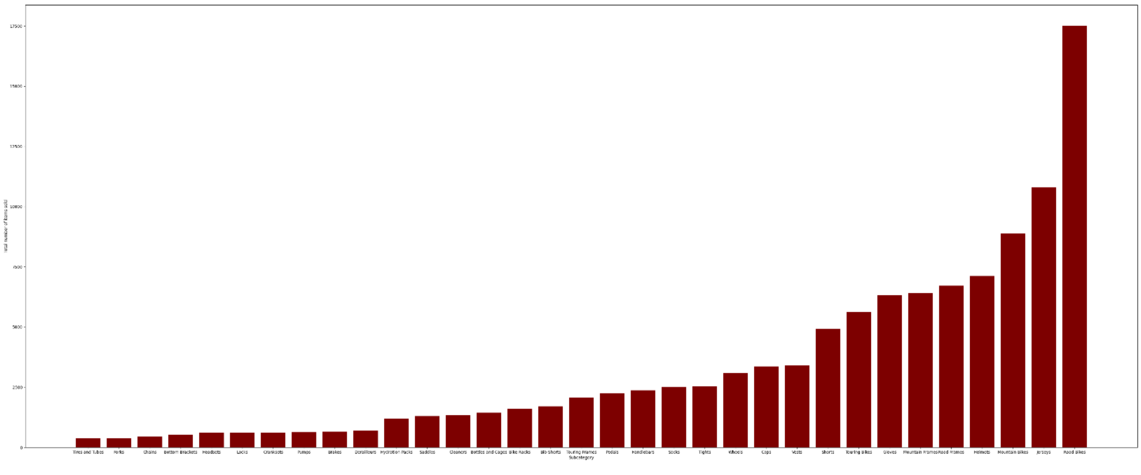
From the above histograms we can make some conclusions:

1. Most of the Products have been ordered without any discount i.e. at full price.
2. Most of the orders have a very low freight amount charged.
3. The list Price and Unit Price are following the same pattern. The prices are also very similar in both histograms.
4. Most of the orders have 1-2 products of each kind. Many orders also have up to 7 same products but more than that is quite rare.
5. Sales amount, tax amount and Freight are following the same trends and most of the orders have comparatively low sales amount (less than $1000) and hence low tax amount and freight applied on them.

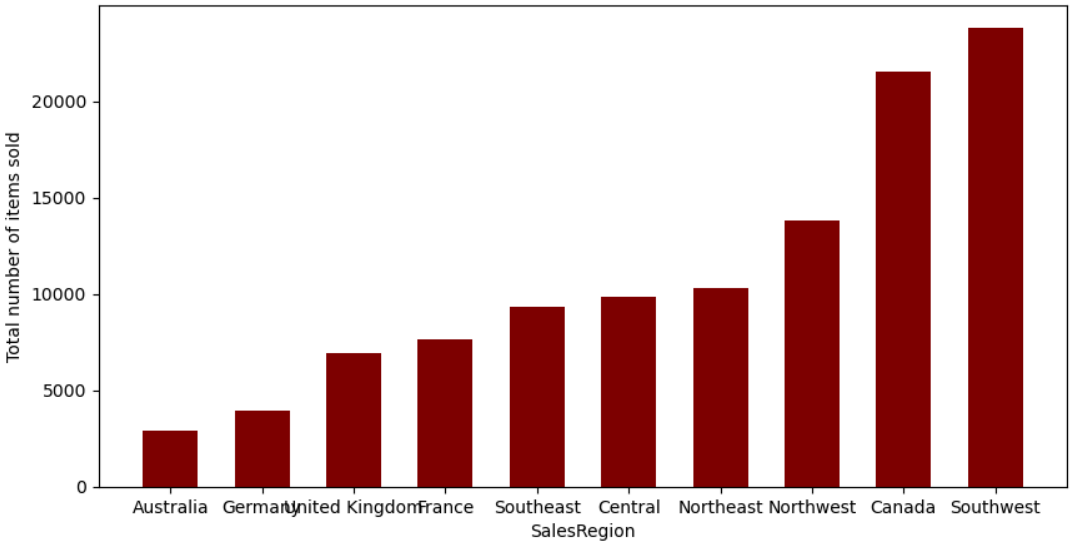
- **Bar chart** for categorical attributes Category, SubCategory, Promotion and Region.



As can be observed from the charts, the highest number of products sold come from Clothing category followed by bikes then components and then Accessories with the least number of products sold.

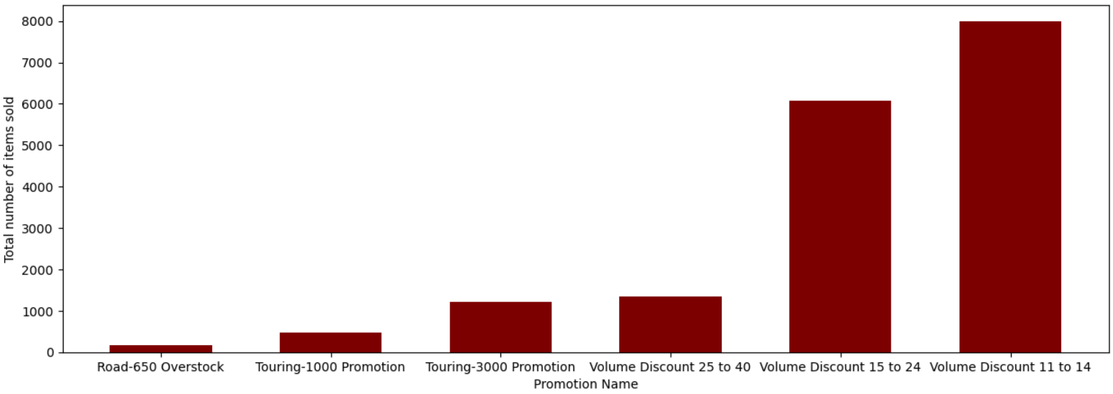


If we zoom in we can observe that the sub category with highest number of products sold is Road Bikes followed by jerseys and then mountain bikes. The sub categories with the least number of products sold are tire tubes and forks.



Here we see the highest number of products sold are in Southwest, Canada and then Northwest whereas the least number of products sold are in Australia and European countries like Germany and UK.

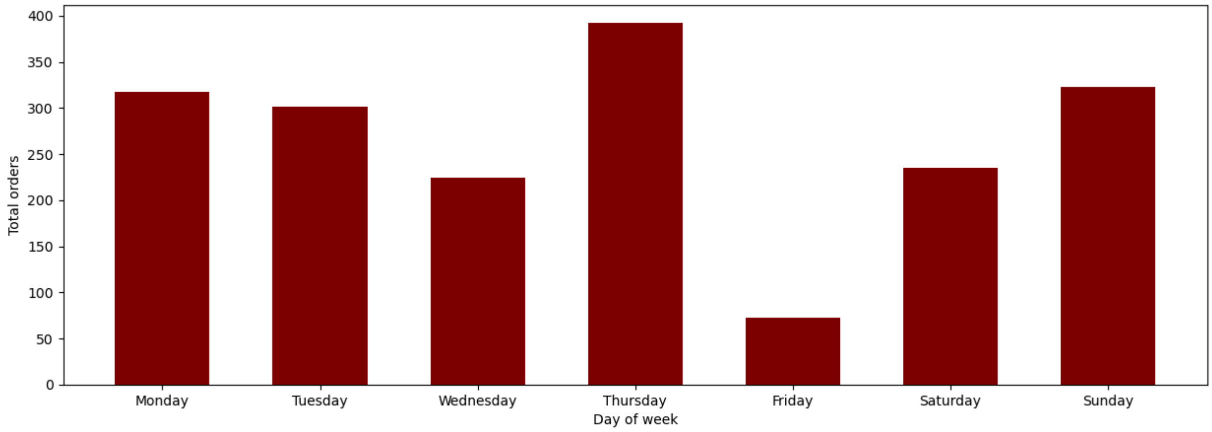
- **Bar chart** for promotions



Highest number of products sold are under the discount, 'Volume Discount 11 to 14' followed closely by 'Volume Discount 15 to 24' after which there is a stark decrease in products sold under discounts with least number of products sold under the discount 'Road-650 Overstock'.

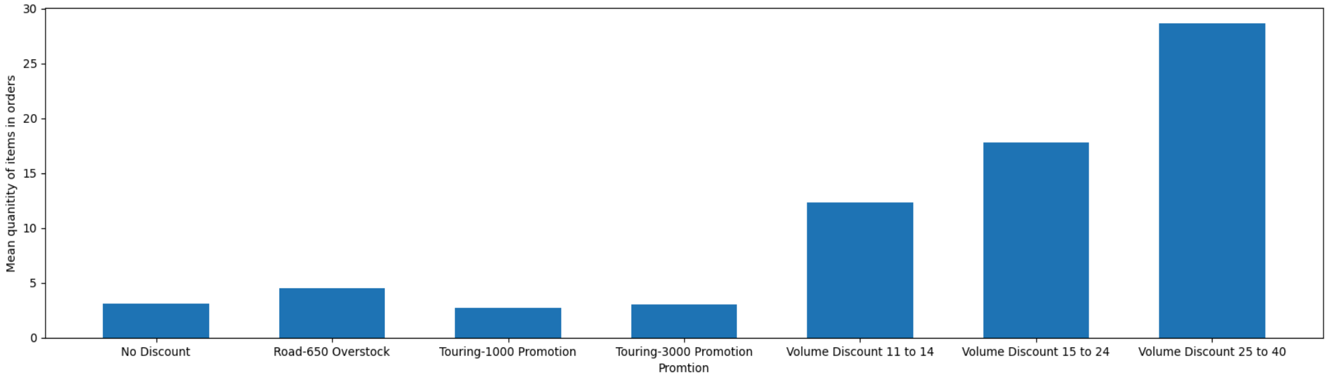
Finding Trends

- Bar plot for number of orders placed on each day of week. Are there any interesting trends?



There are more orders placed on weekdays than weekends with the highest number of orders being placed on Thursdays. The trend drops to the least number of orders placed on Friday after which it picks up on Saturday and Sunday.

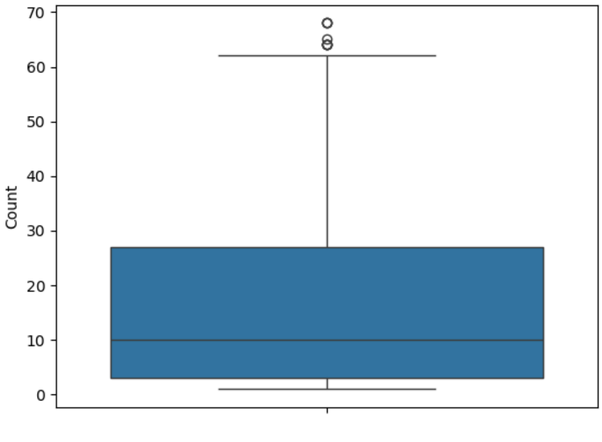
- Is there any impact of promotion on overall product sales?



As we can see, the number of average quantities in an order increases overall when there is a promotion as compared to without any discount. We can also see discounts 25 to 40 having most order quantities followed by vol 15 to 20 and then vol 11 to 14. Hence, promotions have a positive impact on product sales

.

- Performing some analysis on the number of products in an order. Finding their average, min, max etc.

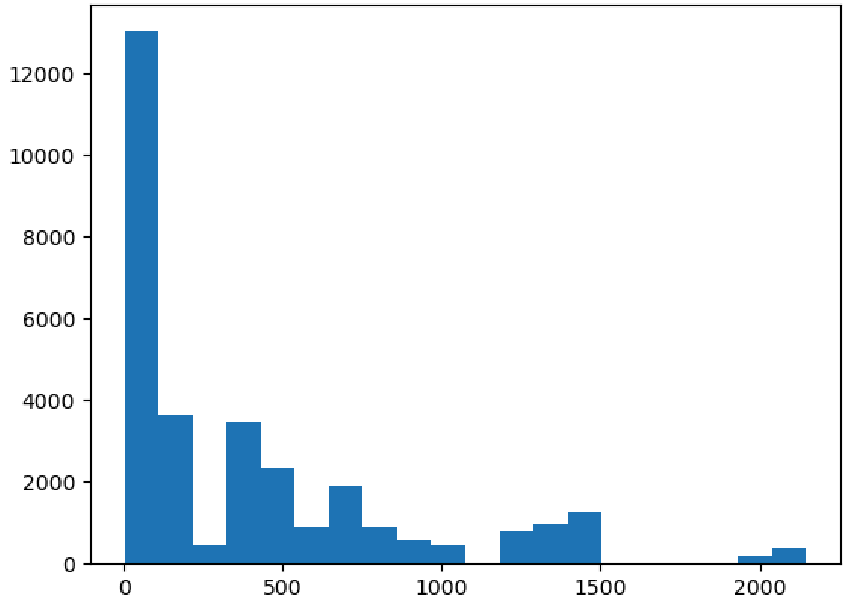


The above box plot shows that the median of the number of products in an order is 10 and the mean is 16.7 therefore we see the box plot positively skewed. The maximum number of products in an order are 68 and this is considered as an outlier.

Normalization

- Creating a new column 'NormalizedPrice' that contains normalized ListPrice of products. Z-Score normalization will be performed.

- Is product price following a normal distribution? Is it skewed? Plot histogram to check the same.



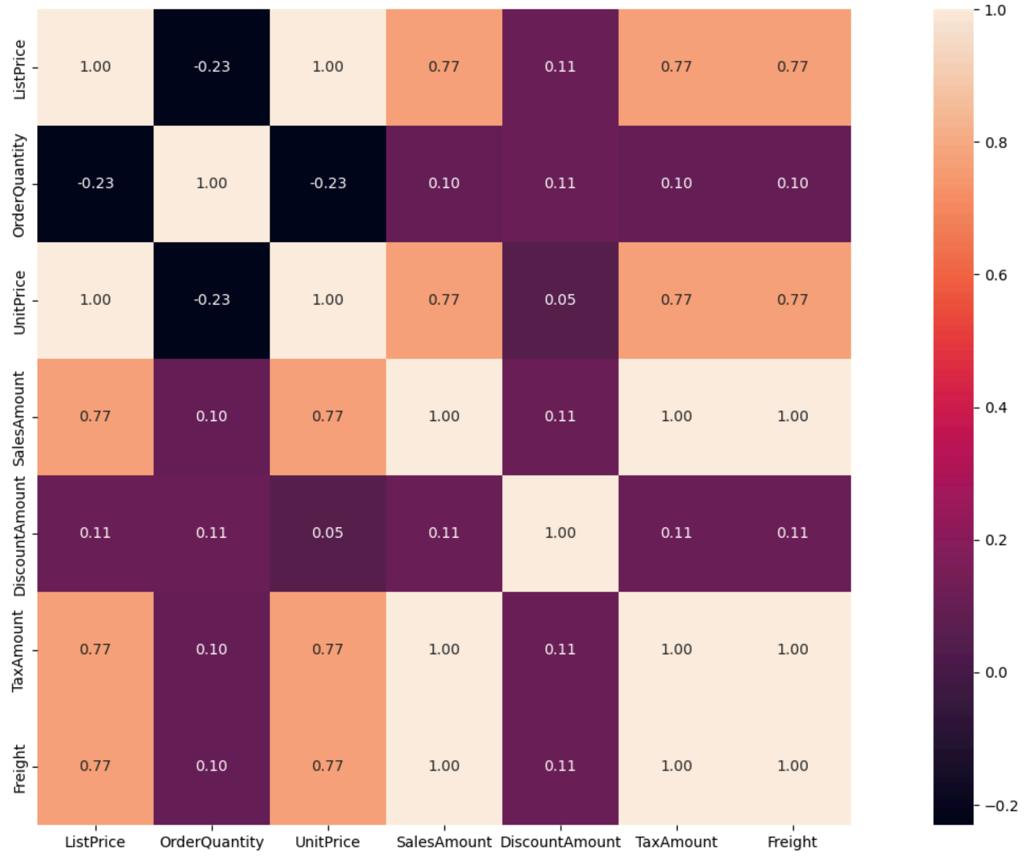
The histogram is not following a normal distribution, rather it is right-skewed.

- Finding k means and it does discretization best as can be seen it divides the data set into most expensive, medium expensive, less expensive and then reasonably priced and cheap categories.

Correlation

Studying correlation between attributes via:

- correlation coefficients
- scatter plot matrix
- plotting heatmap



List Price and Unit Price are positively correlated. If we analyze our data, we can see that the unit price is always approximately 0.6x the List Price.
Sales amount, tax amount and freight charges are also positively correlated with one another. If we analyze our data, we can see that the tax amount is always 8% of the Sales amount and the freight is 2.5% of the Sales amount.

The sales amount also has a positive correlation with Unit Price and ListPrice as the higher the prices of individual products, the more the total sales amount.
Unit Price and Order Quantity have a negative correlation although it is not very strong. The negative correlation would make sense as the higher the price, the lower the quantity of products you would buy.