Sardar Patel Institute of Technology

(Autonomous Institute Affiliated to University of Mumbai)
Bhavan's Campus, Munshi Nagar, Andheri (West), Mumbai-400058-India

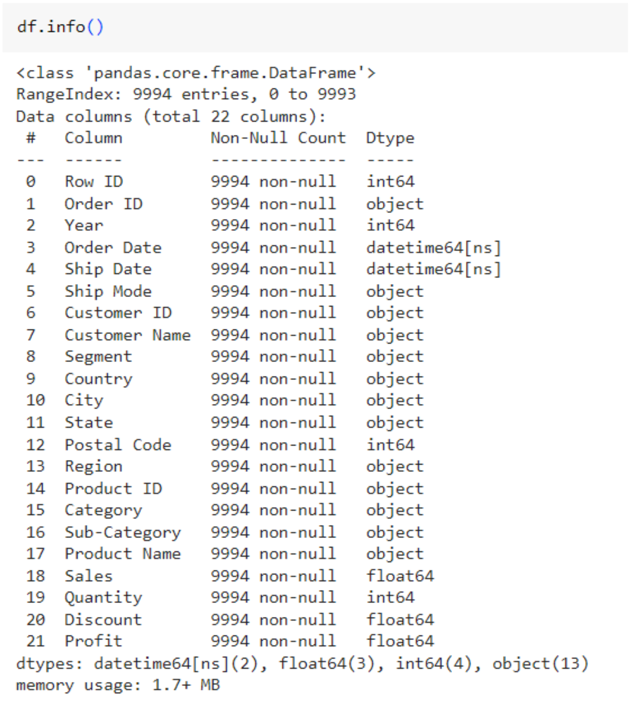
EXPERIMENT 1

NAME:	Sarthak Gharat
ROLL NO:	2021700025
BRANCH:	CSE(DS)
DATE:	23/08/2024

AIM: Create basic charts using Tableau / Power BI / R / Python / D3.js to be performed on the dataset of Ecommerce field

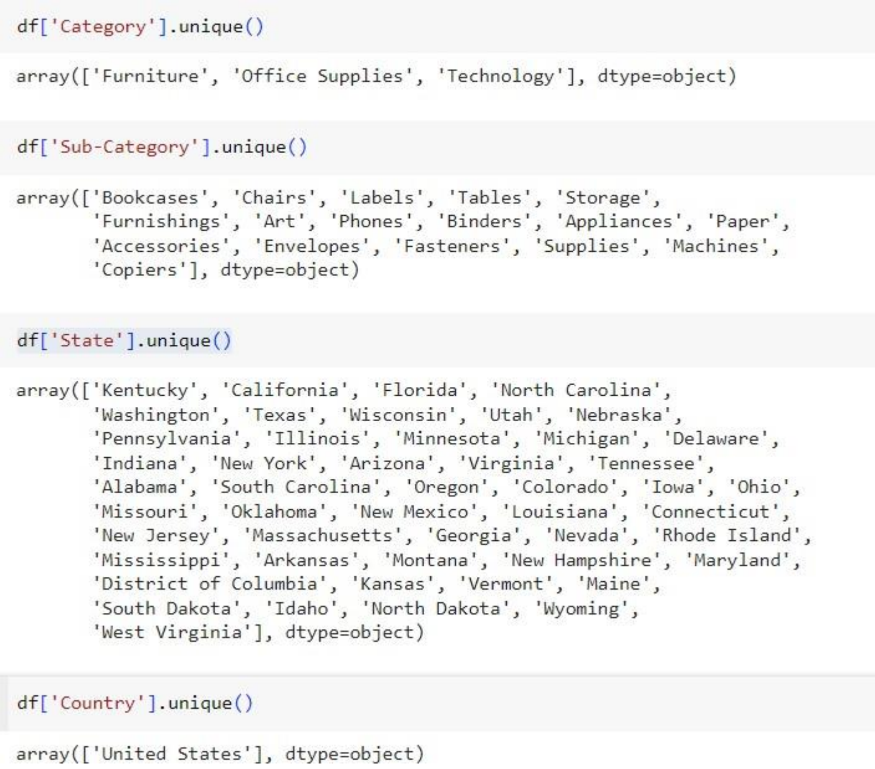
- Basic - Bar chart, Pie chart, Histogram, Timeline chart, Scatter plot, Bubble plot
- Calculate Product wise sales, region wise sales or any other reports
- Write observations from each chart

DATASET:
https://docs.google.com/spreadsheets/d/1L6aBX0uNlzKiJb7JHdkNUile18s9CI4r/edit?gid=1589100670#gid=1589100670



```
df.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 9994 entries, 0 to 9993
Data columns (total 22 columns):
 #   Column         Non-Null Count  Dtype
---  ------         --------------  -----
 0   Row ID         9994 non-null   int64
 1   Order ID       9994 non-null   object
 2   Year           9994 non-null   int64
 3   Order Date     9994 non-null   datetime64[ns]
 4   Ship Date      9994 non-null   datetime64[ns]
 5   Ship Mode      9994 non-null   object
 6   Customer ID    9994 non-null   object
 7   Customer Name  9994 non-null   object
 8   Segment        9994 non-null   object
 9   Country        9994 non-null   object
 10  City           9994 non-null   object
 11  State          9994 non-null   object
 12  Postal Code    9994 non-null   int64
 13  Region         9994 non-null   object
 14  Product ID     9994 non-null   object
 15  Category       9994 non-null   object
 16  Sub-Category   9994 non-null   object
 17  Product Name   9994 non-null   object
 18  Sales          9994 non-null   float64
 19  Quantity       9994 non-null   int64
 20  Discount       9994 non-null   float64
 21  Profit         9994 non-null   float64
dtypes: datetime64[ns](2), float64(3), int64(4), object(13)
memory usage: 1.7+ MB
```



```
df['Category'].unique()
```

```
array(['Furniture', 'Office Supplies', 'Technology'], dtype=object)
```

```
df['Sub-Category'].unique()
```

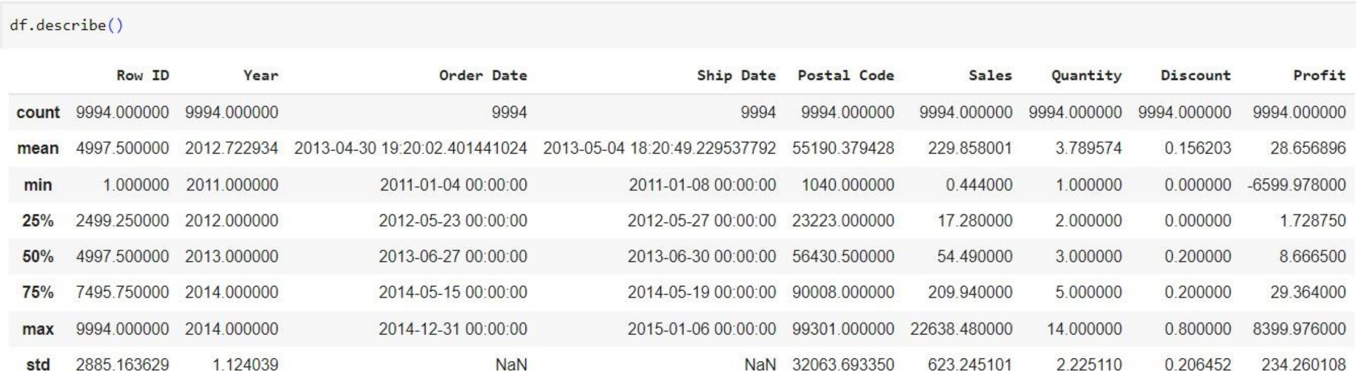
```
array(['Bookcases', 'Chairs', 'Labels', 'Tables', 'Storage',
       'Furnishings', 'Art', 'Phones', 'Binders', 'Appliances', 'Paper',
       'Accessories', 'Envelopes', 'Fasteners', 'Supplies', 'Machines',
       'Copiers'], dtype=object)
```

```
df['State'].unique()
```

```
array(['Kentucky', 'California', 'Florida', 'North Carolina',
       'Washington', 'Texas', 'Wisconsin', 'Utah', 'Nebraska',
       'Pennsylvania', 'Illinois', 'Minnesota', 'Michigan', 'Delaware',
       'Indiana', 'New York', 'Arizona', 'Virginia', 'Tennessee',
       'Alabama', 'South Carolina', 'Oregon', 'Colorado', 'Iowa', 'Ohio',
       'Missouri', 'Oklahoma', 'New Mexico', 'Louisiana', 'Connecticut',
       'New Jersey', 'Massachusetts', 'Georgia', 'Nevada', 'Rhode Island',
       'Mississippi', 'Arkansas', 'Montana', 'New Hampshire', 'Maryland',
       'District of Columbia', 'Kansas', 'Vermont', 'Maine',
       'South Dakota', 'Idaho', 'North Dakota', 'Wyoming',
       'West Virginia'], dtype=object)
```

```
df['Country'].unique()
```

```
array(['United States'], dtype=object)
```



```
df.describe()
```

	Row ID	Year	Order Date	Ship Date	Postal Code	Sales	Quantity	Discount	Profit
count	9994.000000	9994.000000	9994	9994	9994.000000	9994.000000	9994.000000	9994.000000	9994.000000
mean	4997.500000	2012.722934	2013-04-30 19:20:02.401441024	2013-05-04 18:20:49.229537792	55190.379428	229.858001	3.789574	0.156203	28.656896
min	1.000000	2011.000000	2011-01-04 00:00:00	2011-01-08 00:00:00	1040.000000	0.444000	1.000000	0.000000	-6599.978000
25%	2499.250000	2012.000000	2012-05-23 00:00:00	2012-05-27 00:00:00	23223.000000	17.280000	2.000000	0.000000	1.728750
50%	4997.500000	2013.000000	2013-06-27 00:00:00	2013-06-30 00:00:00	56430.500000	54.490000	3.000000	0.200000	8.666500
75%	7495.750000	2014.000000	2014-05-15 00:00:00	2014-05-19 00:00:00	90008.000000	209.940000	5.000000	0.200000	29.364000
max	9994.000000	2014.000000	2014-12-31 00:00:00	2015-01-06 00:00:00	99301.000000	22638.480000	14.000000	0.800000	8399.976000
std	2885.163629	1.124039	NaN	NaN	32063.693350	623.245101	2.225110	0.206452	234.260108

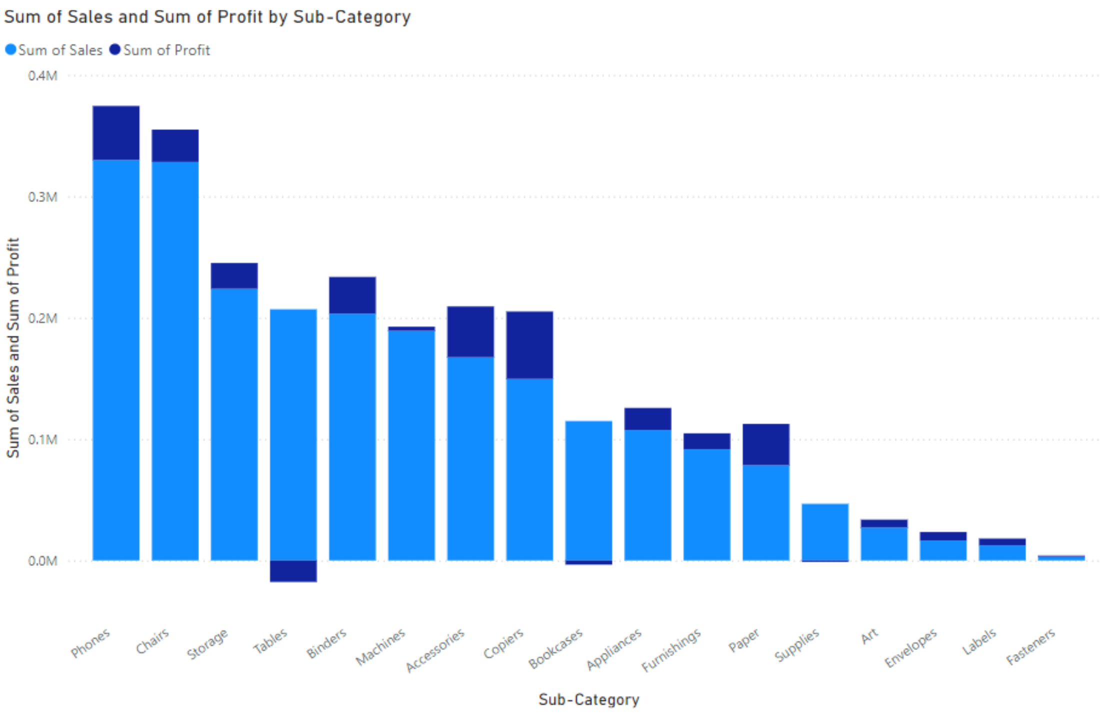
Dataset was cleaned i.e. no null values were found and the data types are correctly assigned.

PROCEDURE DESCRIPTION, OUTPUT & ANALYSIS:

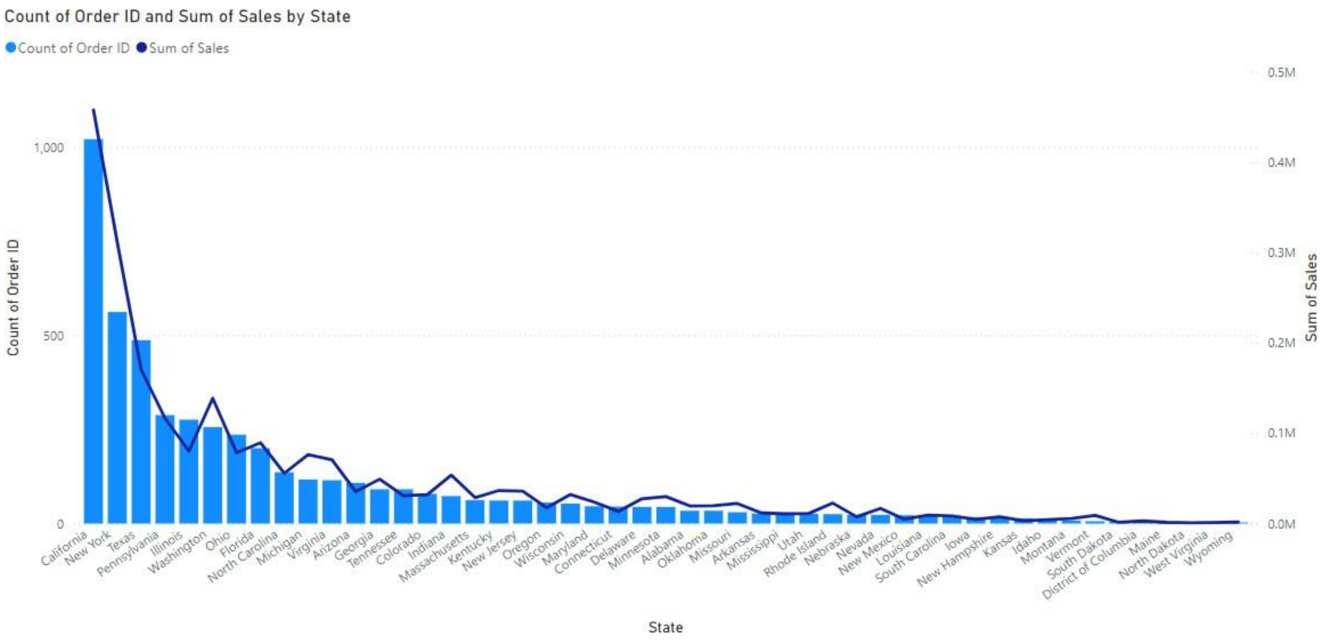
STEP 1. Data is prepared, loaded into Power BI desktop and proceeding further with basic plots and analysis

STEP 2. Bar plots

- Bar plots can be used for a wide range of data types, including nominal (categorical), ordinal, and even interval data.
- It can display single variables (e.g., sales by sub-category) or multiple variables (e.g., sales and profit by sub-category) within the same chart.



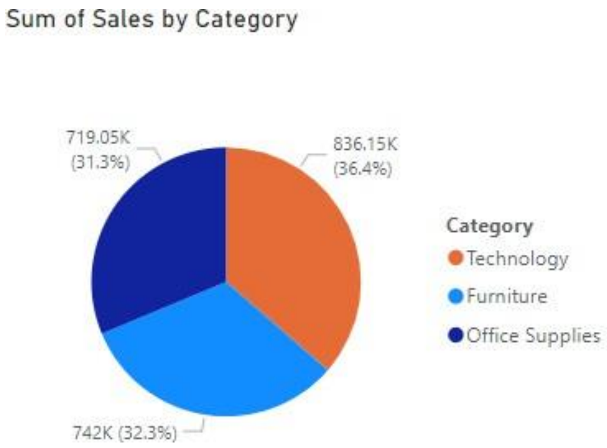
- Sub-categories like accessories, copiers, paper, despite having lower sales generate high profit , these are potentially high margin items where we can focus more.
- Sub-categories like Tables, Storage, Chairs have low profit despite having very high sales, this might need cost optimization and control.
- The bottom 5 ones have low sales and profit as well.

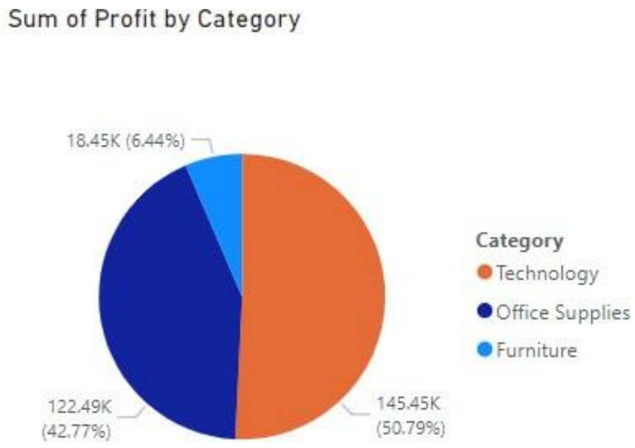


- Above dual axis chart shows the number of orders (distinct) placed and total sales (in Million $) in each state.
- California, New York, Texas, Pennsylvania and Illinois are the top 5 selling states.

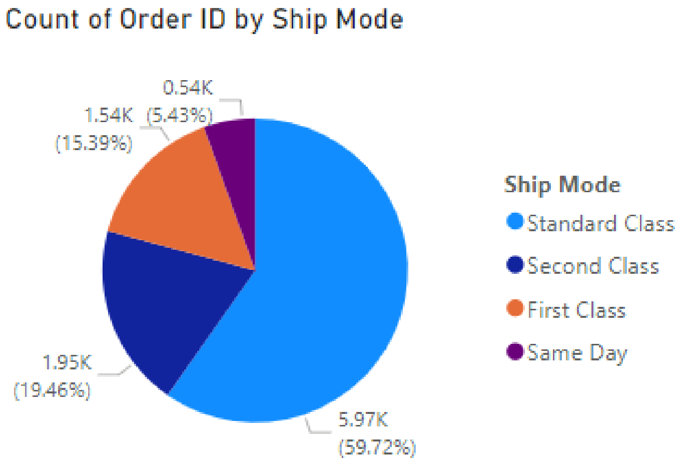
STEP 3. Pie Chart

- Pie charts can be used for comparison of different categorical variables by percentage of total.
- It can work well only in case of limited categories to prevent cluttered visualization, if there are more than we can go ahead with bar charts.



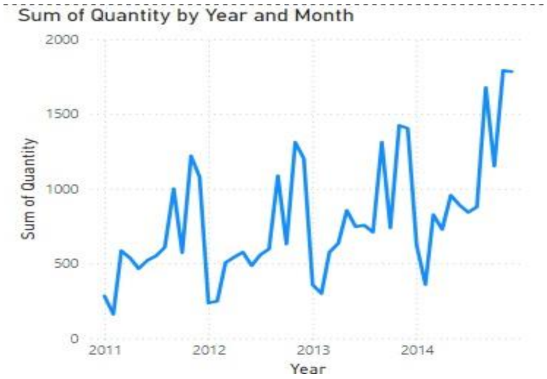


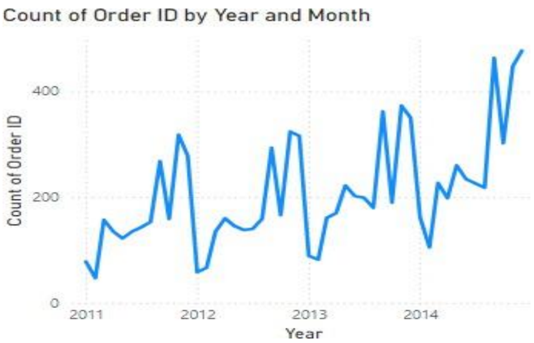
- Pie charts above will show the percentage of total sales contributed by each category, percentage of total profit by category and percentage of total number of orders by shipping mode.
- This helps in understanding which category dominates the most, and clearly above we can see that Technology has both highest sales and profit. We can ensure that these are well-stocked to meet demand.
- From marketing perspective, we need to promote the low selling category office supplies (though it lags by a very less percentage from furniture, and the sales of all categories are almost similar).
- Despite having decent sales for furniture, it has comparatively very low profit. This may indicate inefficiency, required change in pricing strategy and cost management.



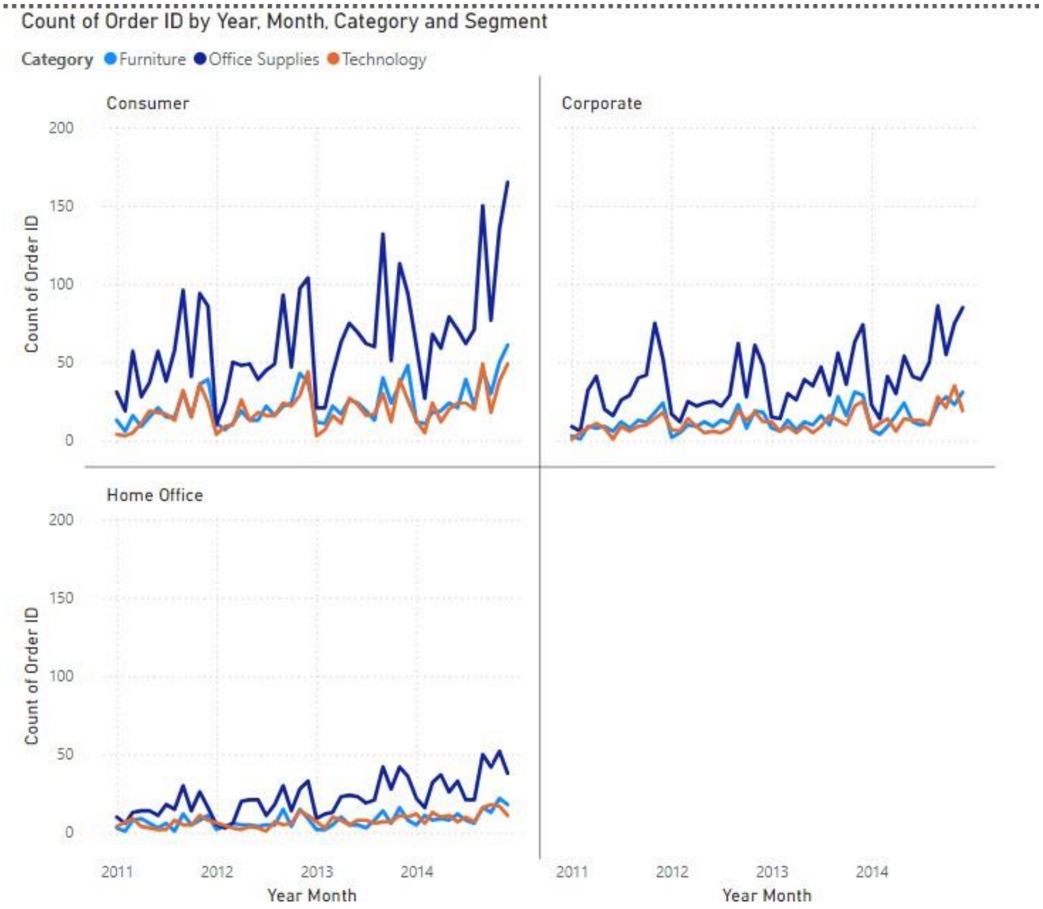
- The most preferred shipping mode is Standard class (almost about 60% of all). This could be due to efficiency (both in terms of cost and time) and better customer service.
- Same day services are used the least.

STEP 4. Timeline chart



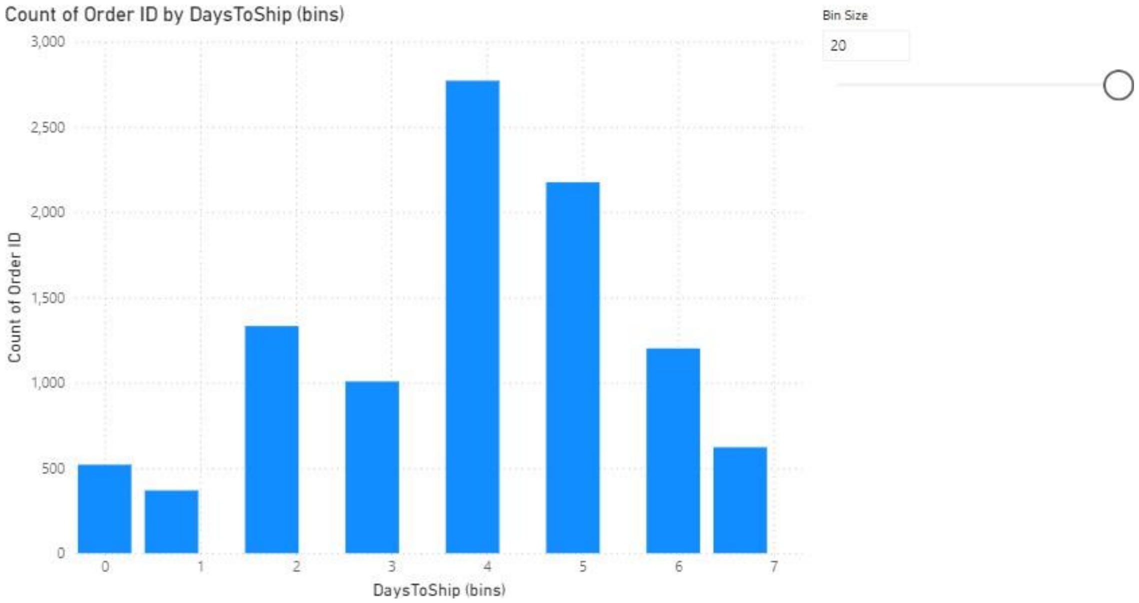


- We see that there is seasonality both in number of orders and quantity in them. There are peaks in March, September and November, which possibly shows marketing strategies like sales or new product launch and consumer behavior.
- Products inventory should be stocked up during these times to maximize the sales hence profit.



- The number of orders are maximum in office supplies category in all the 3 segments, while for furniture and technology products are almost similar.
- Overall the amount is maximum in Consumer segment and minimum in home office.
- Orders are increasing year on year basis, maximum in December 2014.

STEP 5. Histogram

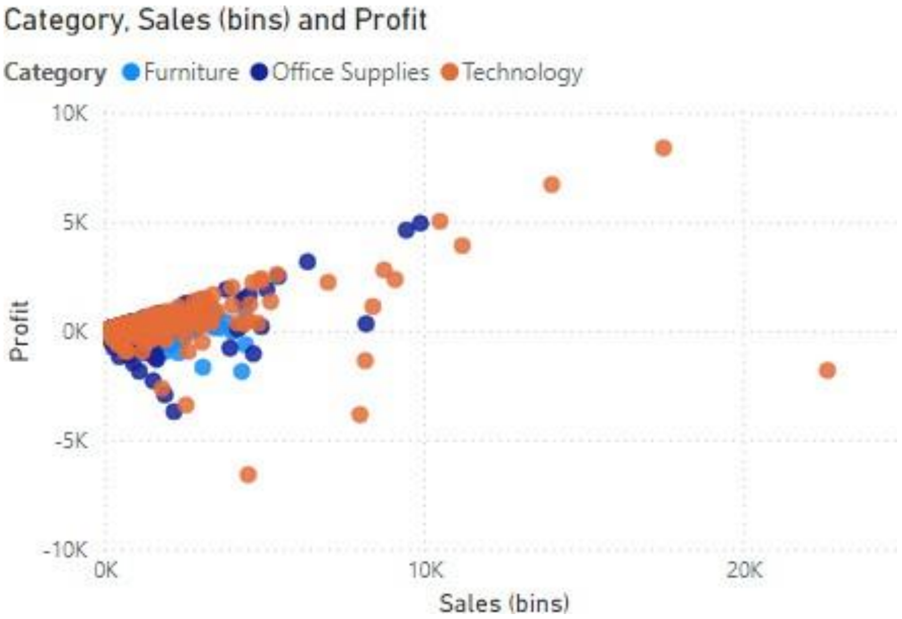


Bin Size

20

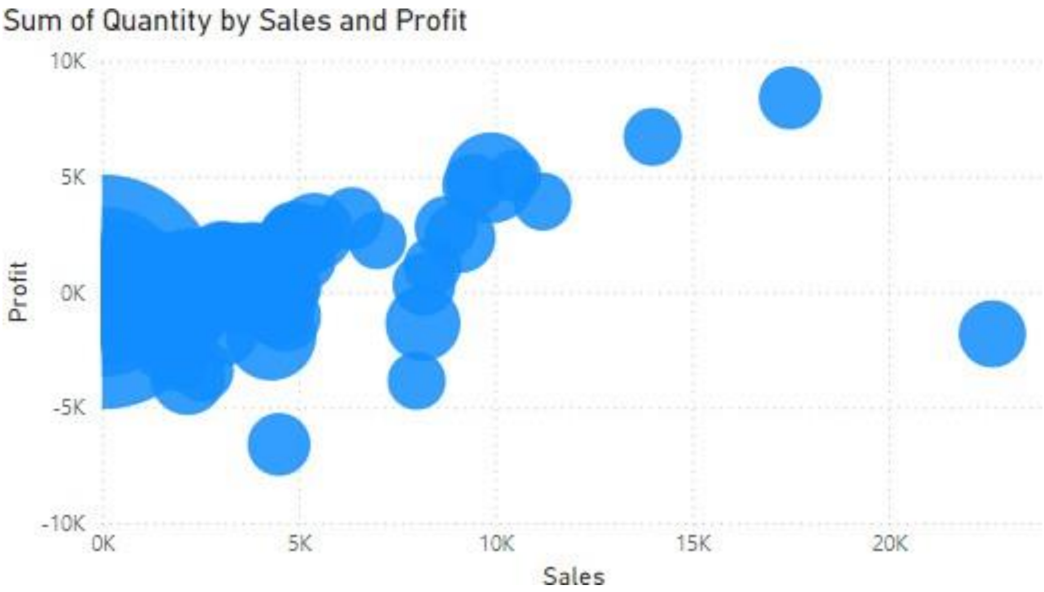
- The DaysToShip is the difference between order date and shipping date. And the histogram depicts the number of orders delivered in the time duration.
- The average number of days required to ship most orders are around 4-5 days.

STEP 6. Scatter plot



- There is a positive correlation between sales and profit (majority), specifically for Technology category and some outliers as well (very high sales but less profit).
- For office supplies, there is some negative correlation and losses.
- For furniture, there is no prominent correlation, average sales give less profit here.

STEP 7. Bubble plot



- The only difference between scatter and bubble plot is, that scatter plot is differentiated on basis of categories (color), while bubble is done on sum of quantity (size). Greater the quantity, greater the bubble size.

CONCLUSION:

- In this experiment, I was able to understand the basics of visualization tool - Power BI
- Also about various charts, its use and analysis on the chosen ecommerce dataset.