Marketing_&_Retail_Analytics

RAGHAVENDRA KUMAR J R

PGP – DSBA Online

Date: 25/02/2024

Summary:

1. Data Analysis
2. Exploratory Analysis:
   1. Univariate, Bivariate and Multivariate analysis
   2. Heat map
3. Use of RFM analysis
4. Customer segmentations – Knime workflow
5. Use of Market Basket Analysis
6. Association Identified
7. A suggestion of possible combos with Lucrative offers.
8. Tableau Link:

MRA Project – Part A – Contents:

PART A: Agenda & Executive Summary of the data -> Contents of the ppt -> Problem statement -> About Data (Info, Shape, Summary Stats, your assumptions about data)

PART A: Exploratory Analysis and Inferences -> Univariate, Bivariate, and multivariate analysis using data visualization (Weekly, Monthly, Quarterly, Yearly Trends in Sales and Sales Across different Categories of different features in the given data) -> Summarise the inferences.

PART A: Customer Segmentation using RFM analysis (4 segments) -> what is RFM? -> What all parameters used and assumptions made? -> Showcase the KNIME workflow image -> what results are there in the output table head?

PART A: Inferences from RFM Analysis and identified segments -> who are your best customers? (Give at least 5) -> Which customers are on the verge of churning? (Give at least 5) -> Who are your lost customers? (Give at least 5) -> Who are your loyal customers? (Give at least 5)

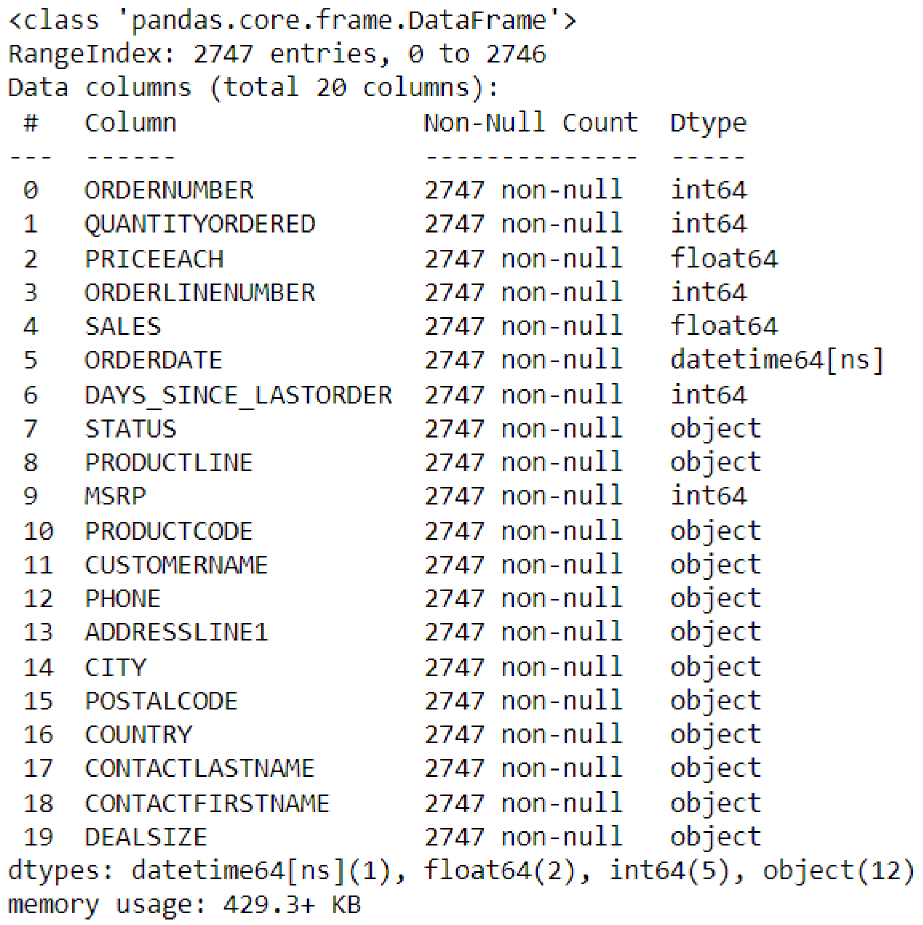
Problem Statement:

An automobile parts manufacturing company has collected data on transactions for 3 years. They do not have any in-house data science team, thus they have hired you as their consultant. Your job is to use your data science skills to find the underlying buying patterns of the customers, provide the company with suitable insights about their customers, and recommend customized marketing strategies for different segments of customers.

PART A: Agenda & Executive Summary of the data -> Contents of the ppt -> Problem statement -> About Data (Info, Shape, Summary Stats, your assumptions about data)

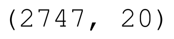
Data analysis - Solution:

DF.info:



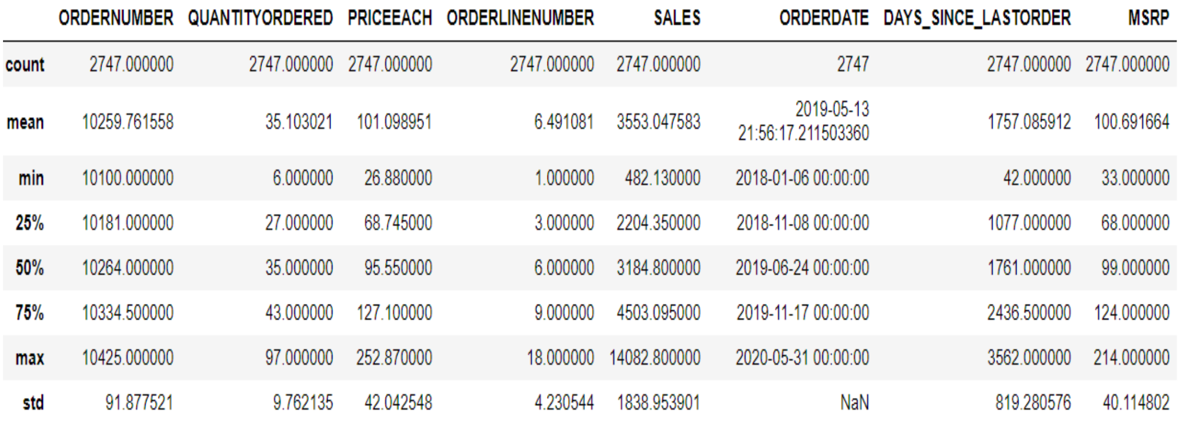
```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 2747 entries, 0 to 2746
Data columns (total 20 columns):
 #   Column                Non-Null Count  Dtype         
---  ------                --------------  -----         
 0   ORDERNUMBER           2747 non-null   int64         
 1   QUANTITYORDERED       2747 non-null   int64         
 2   PRICEEACH             2747 non-null   float64       
 3   ORDERLINENUMBER       2747 non-null   int64         
 4   SALES                 2747 non-null   float64       
 5   ORDERDATE             2747 non-null   datetime64[ns]
 6   DAYS_SINCE_LASTORDER  2747 non-null   int64         
 7   STATUS                2747 non-null   object        
 8   PRODUCTLINE           2747 non-null   object        
 9   MSRP                  2747 non-null   int64         
 10  PRODUCTCODE           2747 non-null   object        
 11  CUSTOMERNAME          2747 non-null   object        
 12  PHONE                 2747 non-null   object        
 13  ADDRESSLINE1          2747 non-null   object        
 14  CITY                  2747 non-null   object        
 15  POSTALCODE            2747 non-null   object        
 16  COUNTRY               2747 non-null   object        
 17  CONTACTLASTNAME       2747 non-null   object        
 18  CONTACTFIRSTNAME      2747 non-null   object        
 19  DEALSIZE              2747 non-null   object        
dtypes: datetime64[ns](1), float64(2), int64(5), object(12)
memory usage: 429.3+ KB
```

Df.shape:



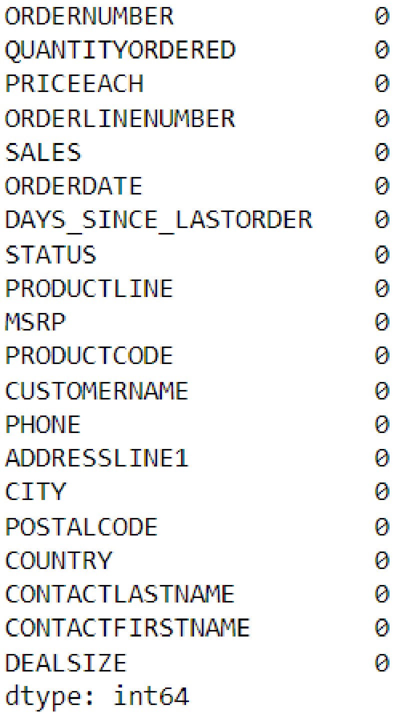
```
(2747, 20)
```

Df.describe:



	ORDERNUMBER	QUANTITYORDERED	PRICEEACH	ORDERLINENUMBER	SALES	ORDERDATE	DAYS_SINCE_LASTORDER	MSRP
count	2747.000000	2747.000000	2747.000000	2747.000000	2747.000000	2747	2747.000000	2747.000000
mean	10259.761558	35.103021	101.098951	6.491081	3553.047583	2019-05-13 21:56:17.211503360	1757.085912	100.691664
min	10100.000000	6.000000	26.880000	1.000000	482.130000	2018-01-06 00:00:00	42.000000	33.000000
25%	10181.000000	27.000000	68.745000	3.000000	2204.350000	2018-11-08 00:00:00	1077.000000	68.000000
50%	10264.000000	35.000000	95.550000	6.000000	3184.800000	2019-06-24 00:00:00	1761.000000	99.000000
75%	10334.500000	43.000000	127.100000	9.000000	4503.095000	2019-11-17 00:00:00	2436.500000	124.000000
max	10425.000000	97.000000	252.870000	18.000000	14082.800000	2020-05-31 00:00:00	3562.000000	214.000000
std	91.877521	9.762135	42.042548	4.230544	1838.953901	NaN	819.280576	40.114802

Missing values:



```
ORDERNUMBER             0
QUANTITYORDERED         0
PRICEEACH               0
ORDERLINENUMBER         0
SALES                   0
ORDERDATE               0
DAYS_SINCE_LASTORDER    0
STATUS                  0
PRODUCTLINE             0
MSRP                    0
PRODUCTCODE             0
CUSTOMERNAME            0
PHONE                   0
ADDRESSLINE1            0
CITY                    0
POSTALCODE              0
COUNTRY                 0
CONTACTLASTNAME         0
CONTACTFIRSTNAME        0
DEALSIZE                0
dtype: int64
```

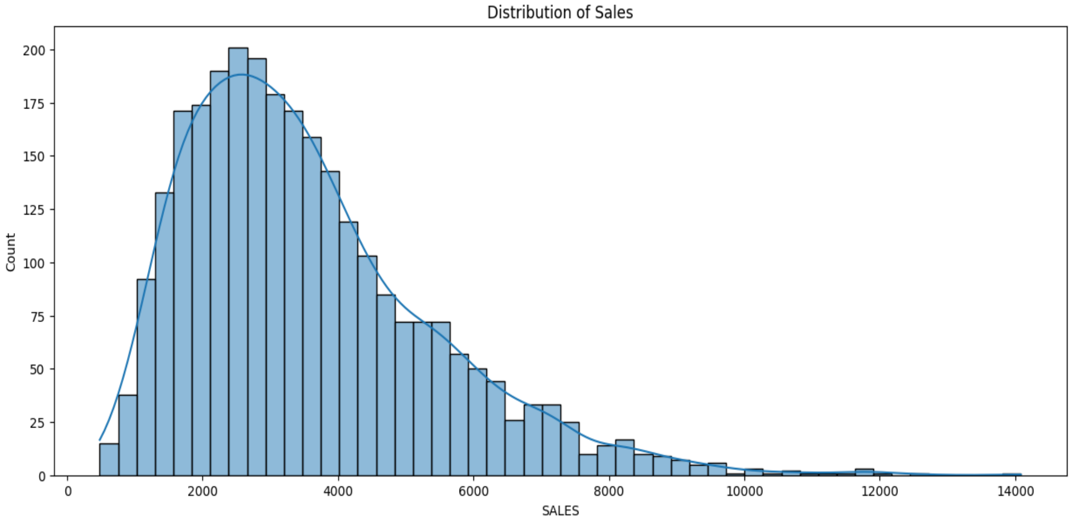
Summary:

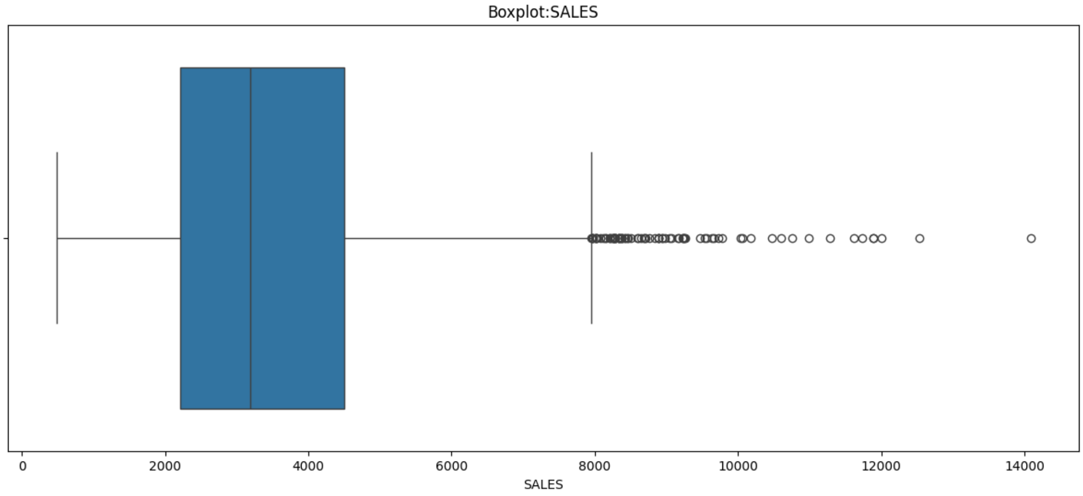
Dataset having 20 variables out of which is 12 categorical, 7 numerical and one is data field and there is no missing values and duplicate values found.

Total number of observations are 2747.

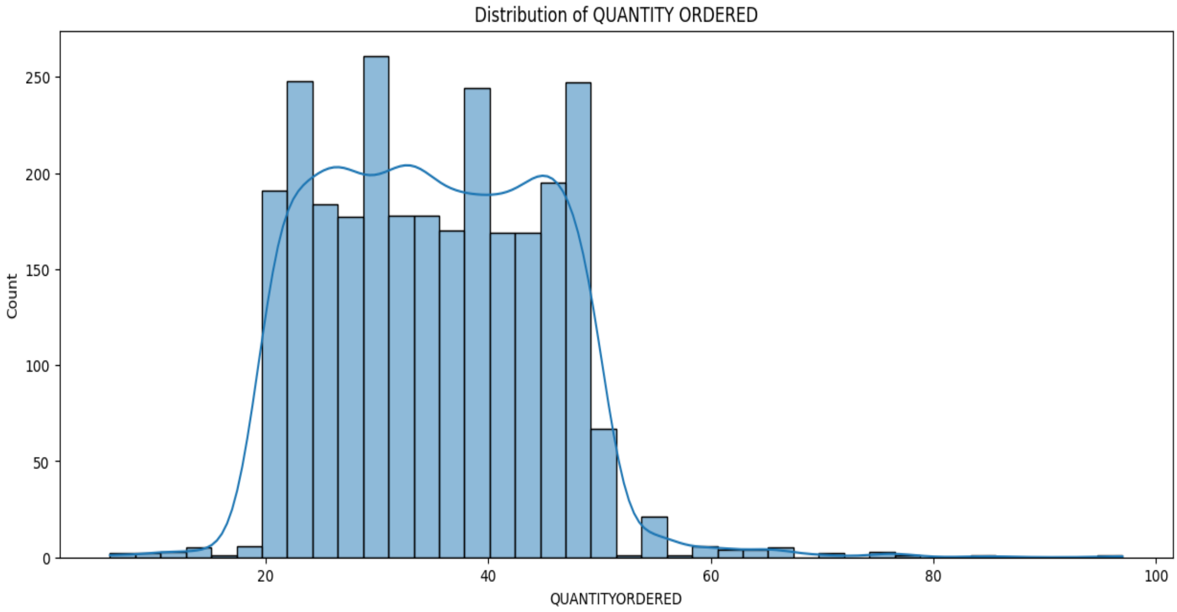
PART A: Exploratory Analysis and Inferences -> Univariate, Bivariate, and multivariate analysis using data visualization (Weekly, Monthly, Quarterly, Yearly Trends in Sales and Sales Across different Categories of different features in the given data) -> Summarise the inferences.

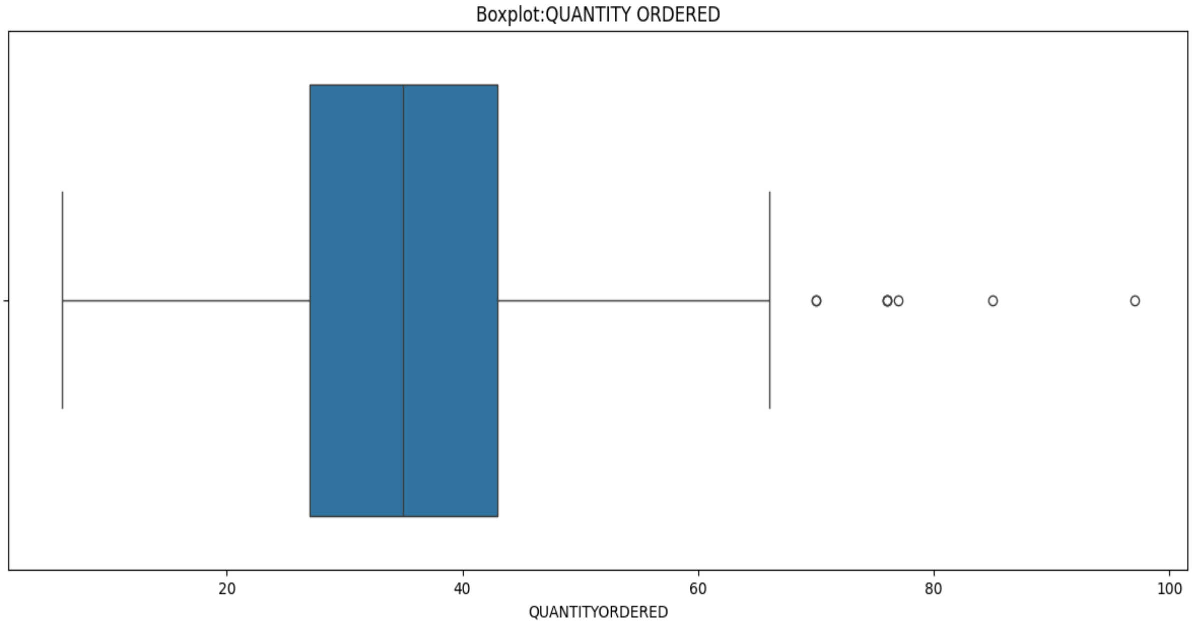
Solution: Univariate Analysis:



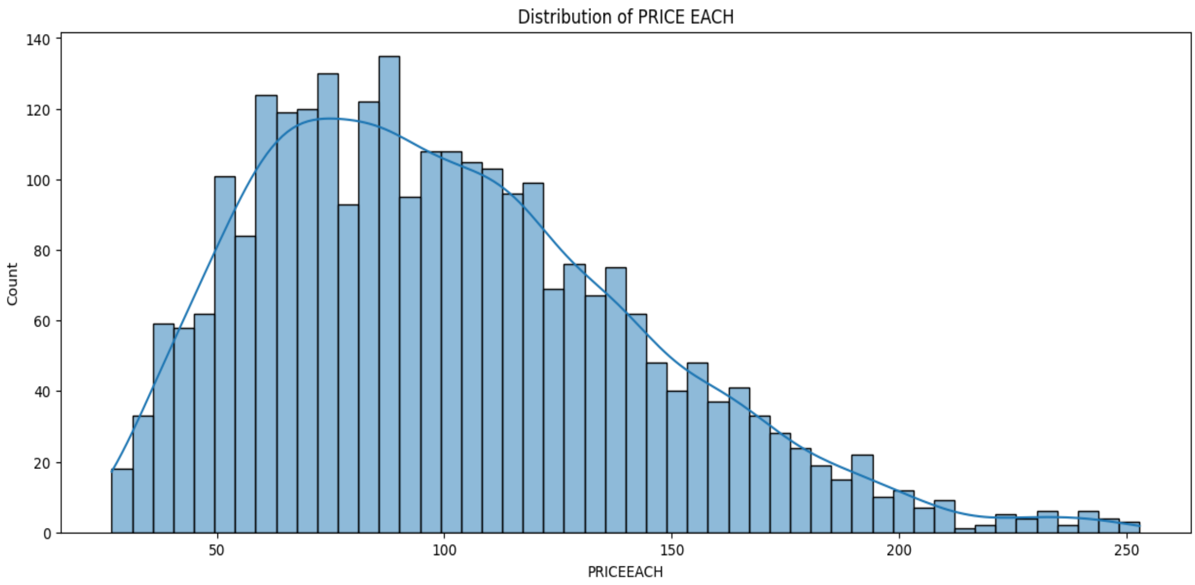


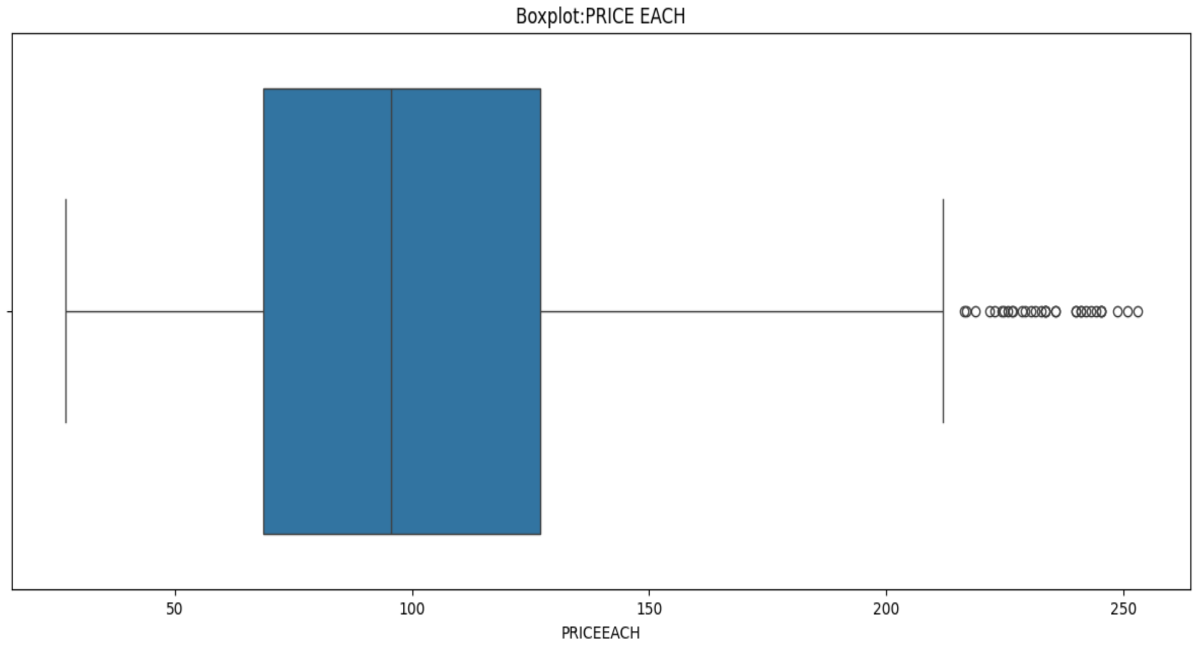
This variable is right skewed with a lot of outliers.



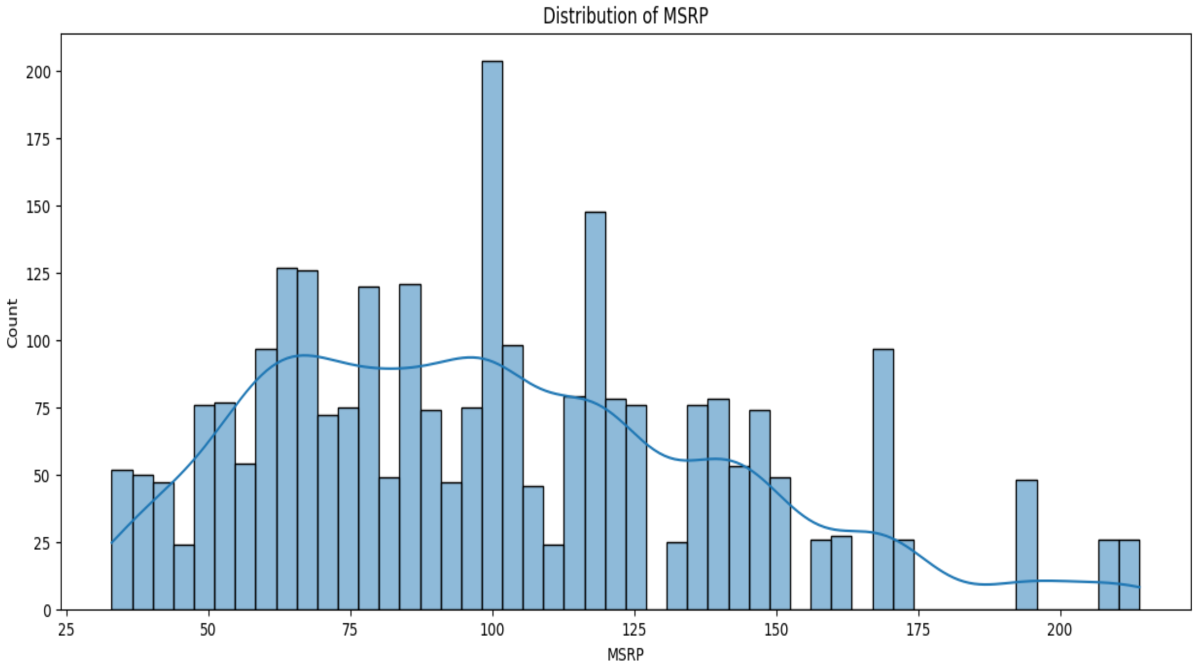


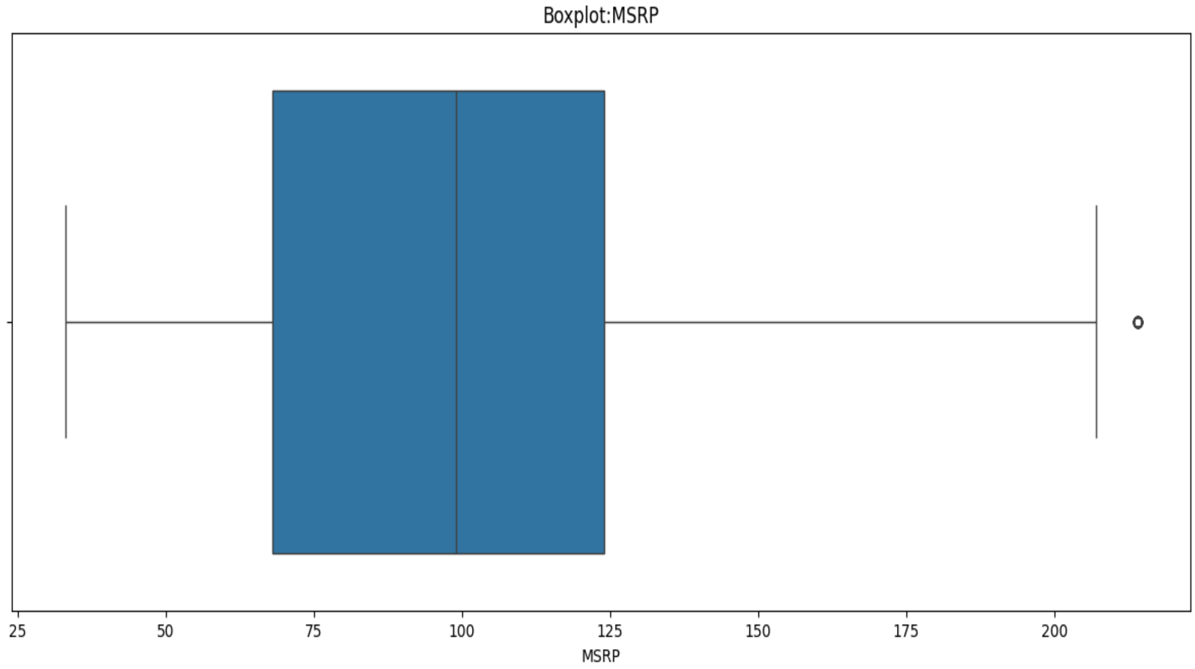
After the exploration of data that there are outliers present in the variable and data is not perfect normally distributed.



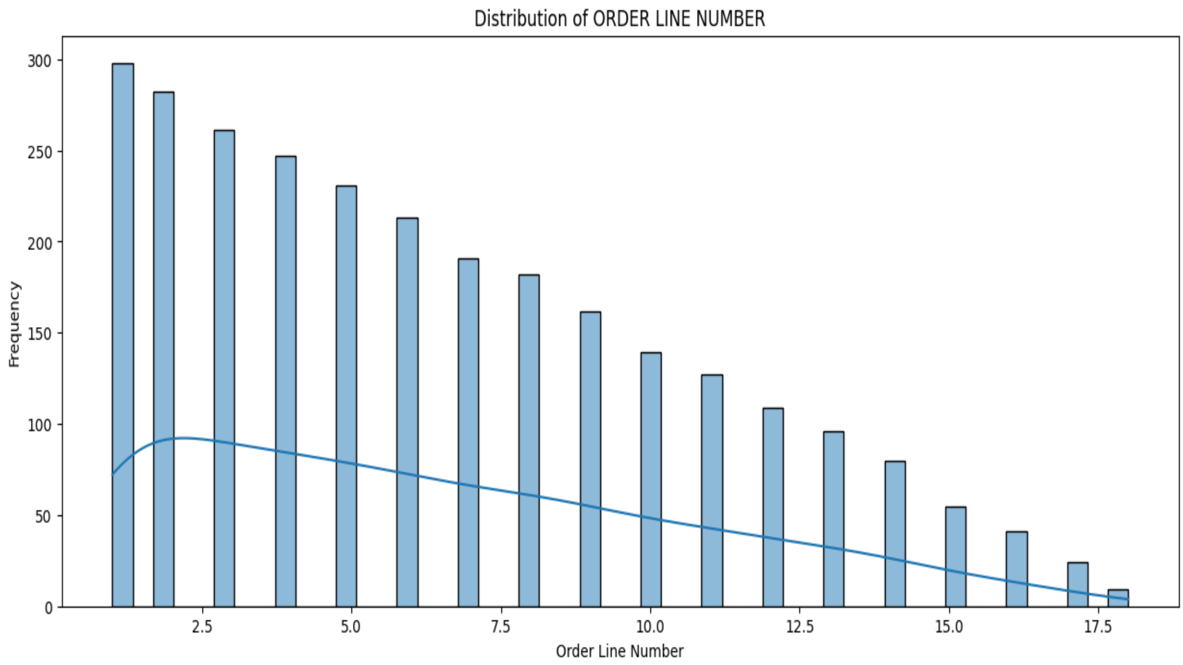


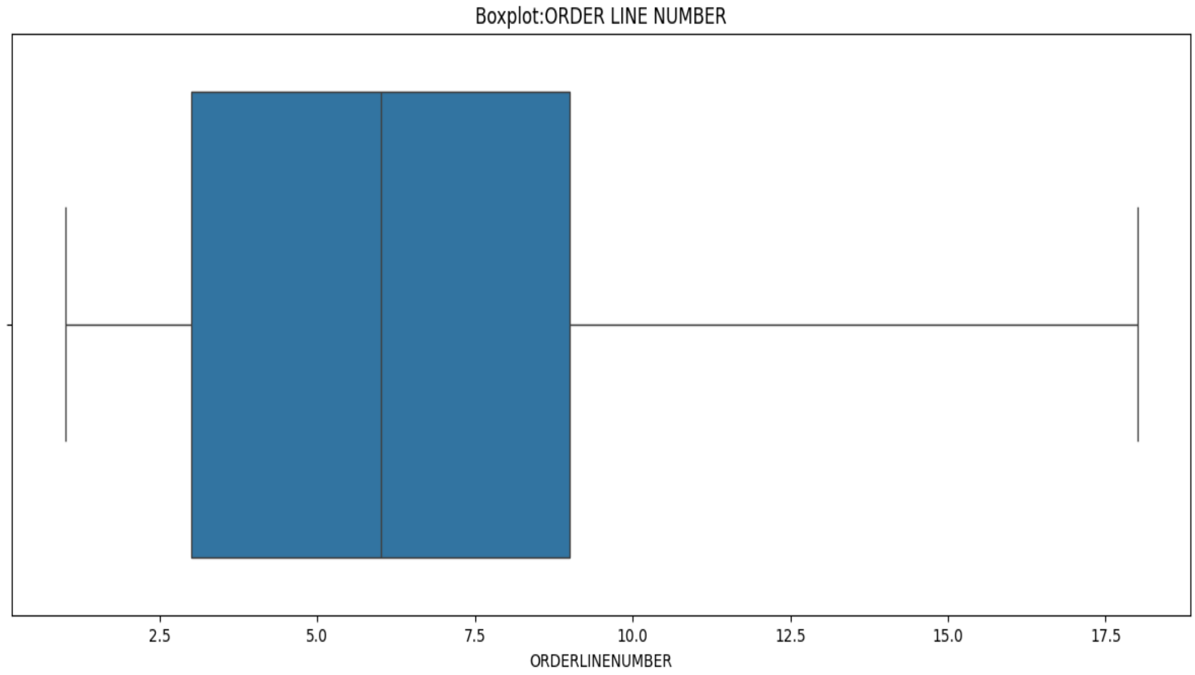
This variable having approx. normal distribution (slight right skewed) but many outliers.





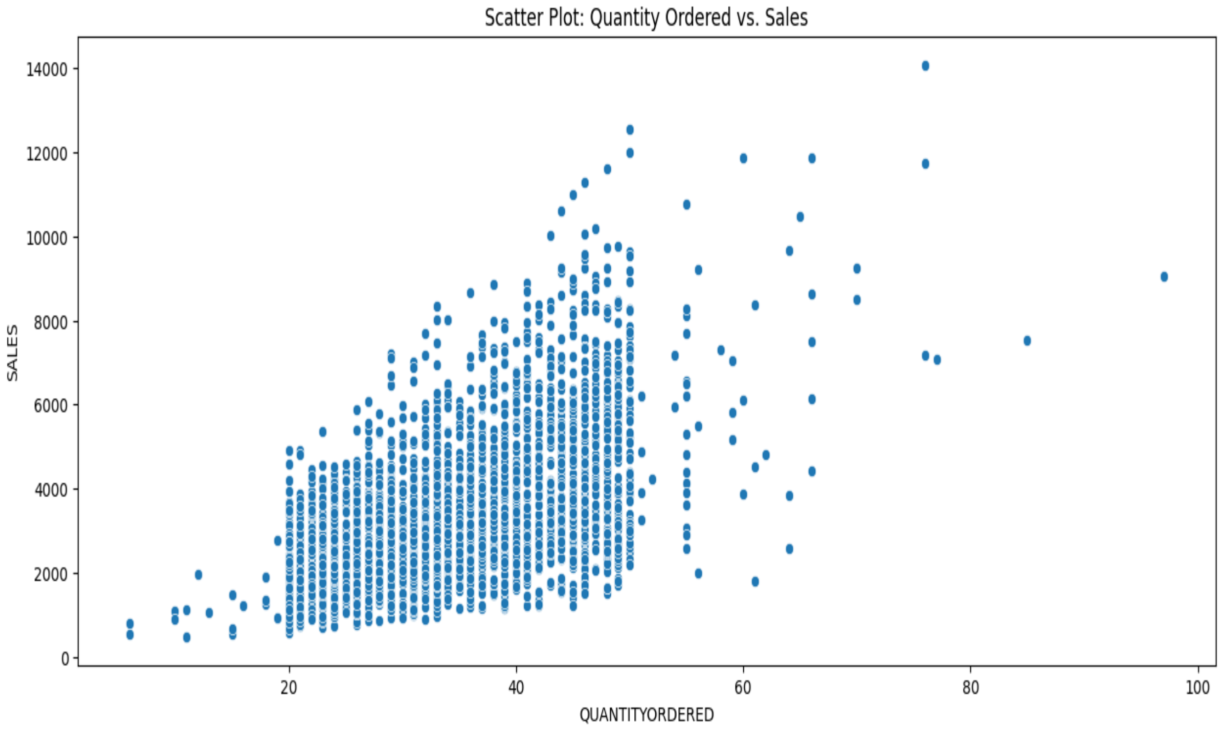
Manufacturer's Suggested Retail Price - This variable having very less outliers.





This variable doesn’t have single outliers.

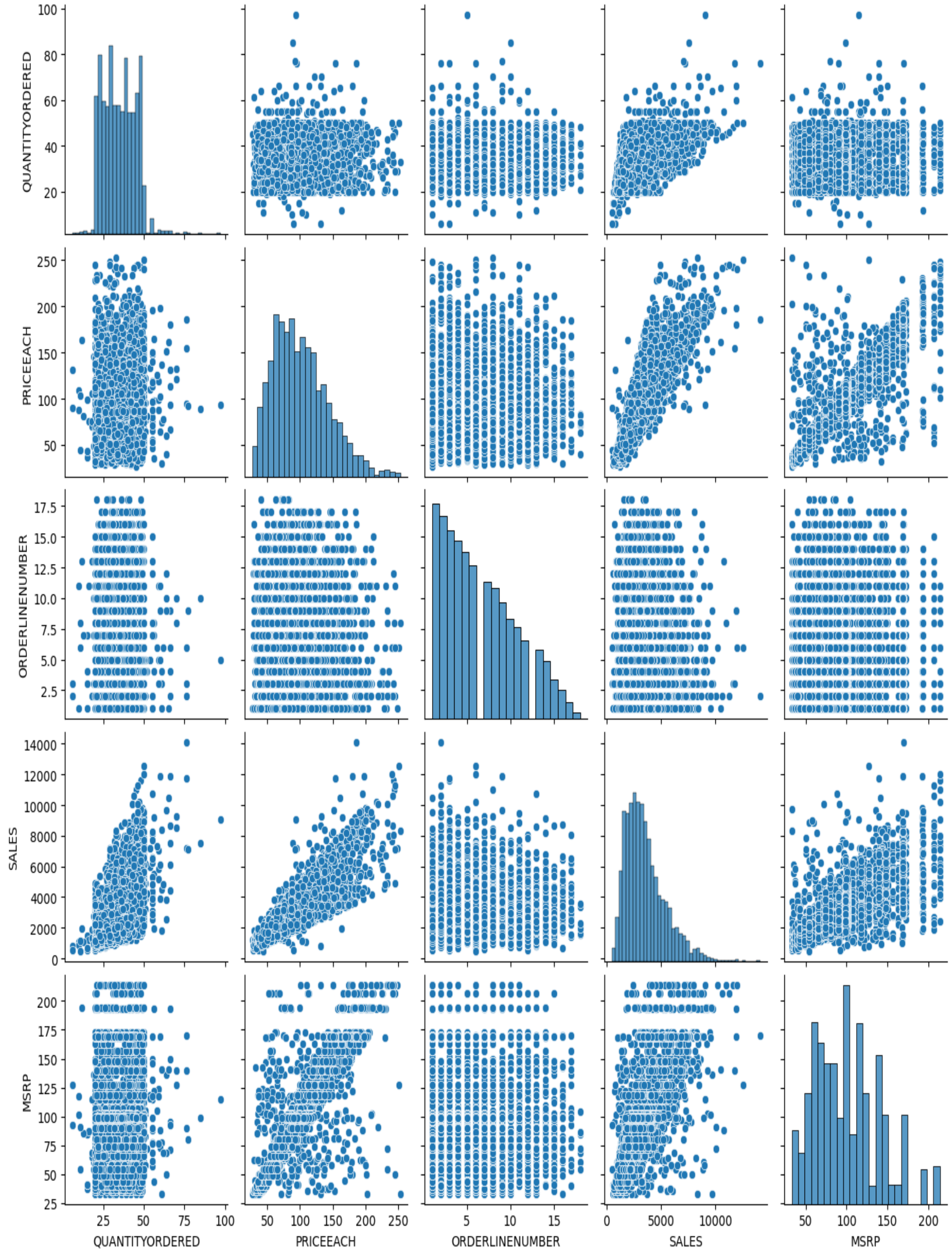
Bivariate Analysis:



- There is a positive correlation between quantity ordered and sales. This means that as the quantity ordered increases, sales also tend to increase. This is likely because customers who order more items are spending more money.

- There is a lot of scatter in the data. This means that there is a lot of variability in the relationship between quantity ordered and sales. In other words, not all customers who order a lot of items also have high sales. There are many other factors that can affect sales, such as the price of the items, the customer's budget, and the customer's needs.

- There are a few outliers in the data. These are the data points that fall far away from the main trend of the data. Outliers can be caused by errors in the data or by unusual events. It is important to investigate outliers to see if they are legitimate or if they should be removed from the data analysis.


QUANTITYORDERED
PRICEEACH
ORDERLINENUMBER
SALES
MSRP
100
80
60
40
20
250
200
150
100
50
17.5
15.0
12.5
10.0
7.5
5.0
2.5
14000
12000
10000
8000
6000
4000
2000
0
200
175
150
125
100
75
50
25
25 50 75 100
50 100 150 200 250
5 10 15
0 5000 10000
50 100 150 200

Variables:

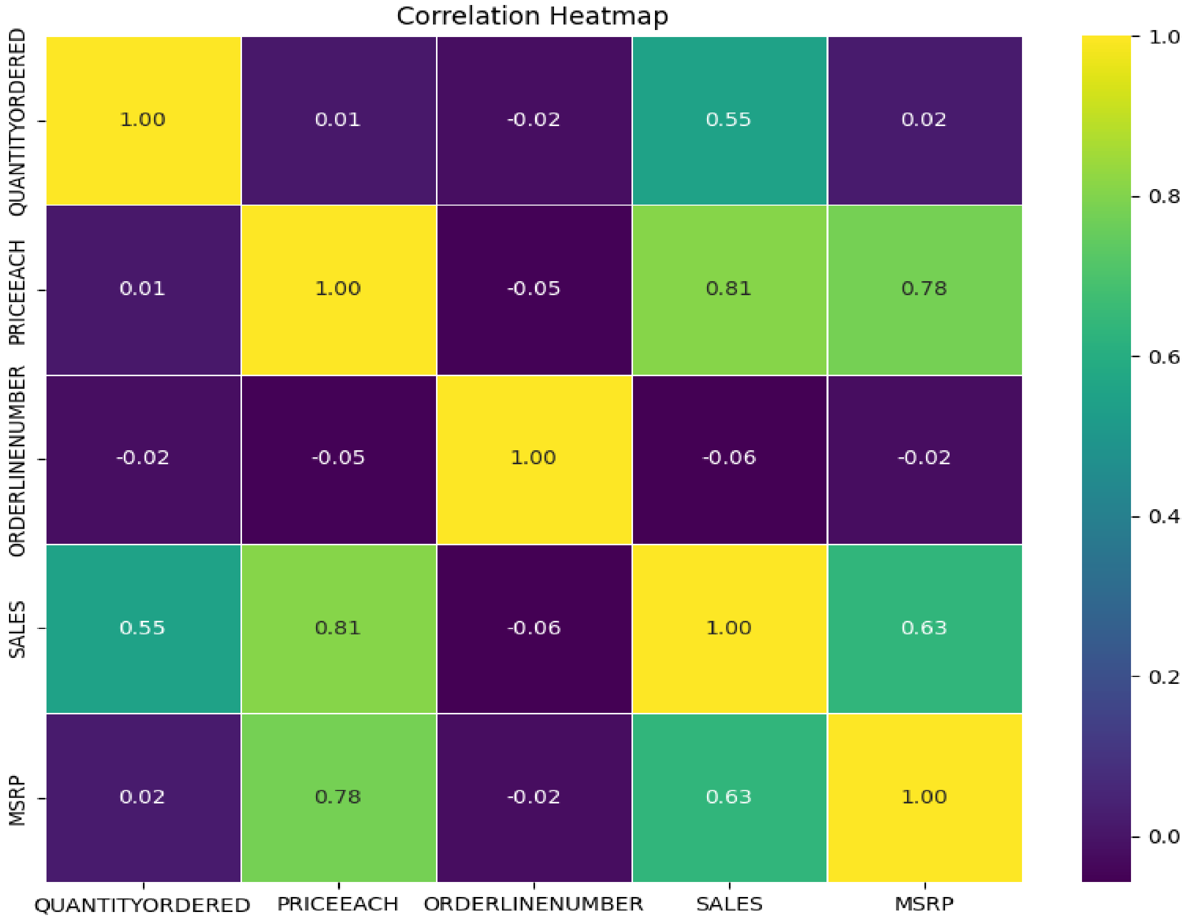
-
- The x-axis variables include "QUANTITYORDERED", "PRICEEACH", and "ORDERLINENUMBER", which seem to be numerical.
- The y-axis variables include "MSRP", "SALES", and again "ORDERLINENUMBER" which appears twice. This might be a mistake, or it could be intentional to show the distribution of order numbers on the diagonal plot.

Relationships:

- There seems to be a positive correlation between "QUANTITYORDERED" and "SALES". This means that as the number of items ordered increases, the total sales also tend to increase.
- There is a weaker positive correlation between "PRICEEACH" and "SALES". This means that as the price per item increases, the total sales also tend to increase, but not as strongly as with the quantity ordered.
- There is no clear correlation between "ORDERLINENUMBER" and either "SALES" or "MSRP". This means that the order number itself doesn't seem to have a predictive relationship with the sales or the MSRP.

Distributions:

- The diagonal plots (histograms) show that "QUANTITYORDERED" and "PRICEEACH" have right-skewed distributions, meaning there are more orders and prices on the lower end with a few outliers on the higher end.
- "ORDERLINENUMBER" also has a right-skewed distribution, but it's difficult to say more without knowing the context of the data (e.g., is it a continuous increasing number or does it restart periodically?).
- "SALES" and "MSRP" also seem to have right-skewed distributions, but it's hard to be certain without more data points.
- Overall, this pair plot suggests that there is a positive relationship between the quantity of items ordered and the total sales, with a weaker relationship between price per item and sales. The order number itself doesn't seem to be a good predictor of sales or MSRP.



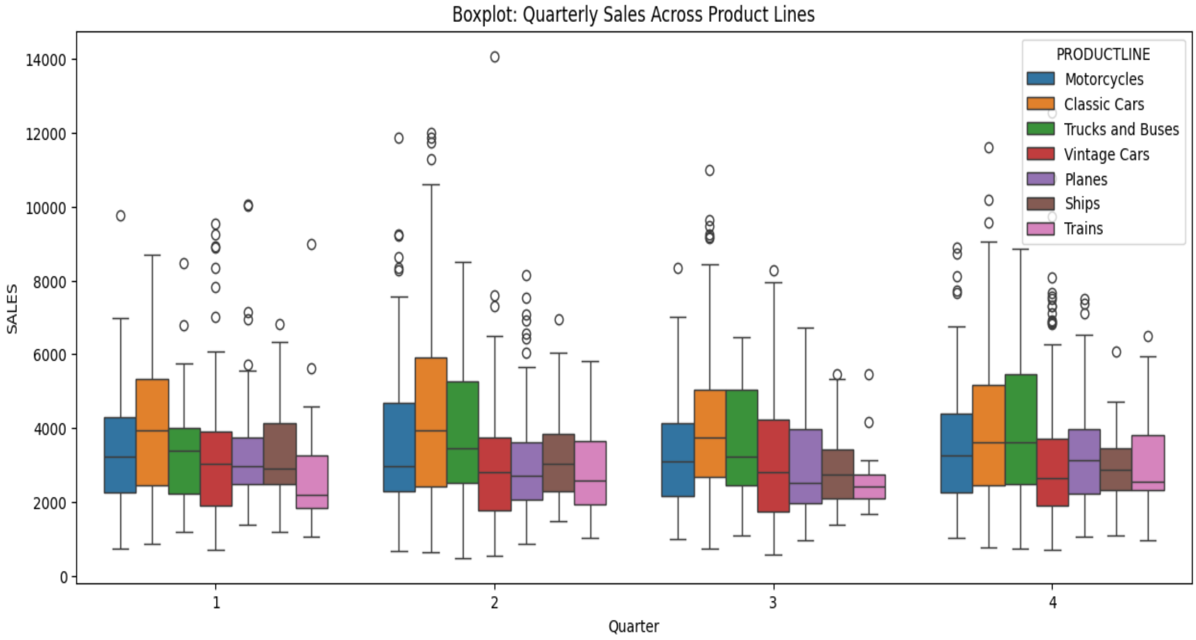
- There is a positive correlation between the quantity of orders and the number of orders received. This means that as the quantity ordered increases, the number of orders received also tends to increase. This is likely because customers who order more items are more likely to place multiple orders.

- The purple boxes represent the highest orders received, the green boxes represent the lowest orders received, and the yellow boxes represent the middle values. This means that the heat map is showing the distribution of the number of orders received for different quantities ordered.

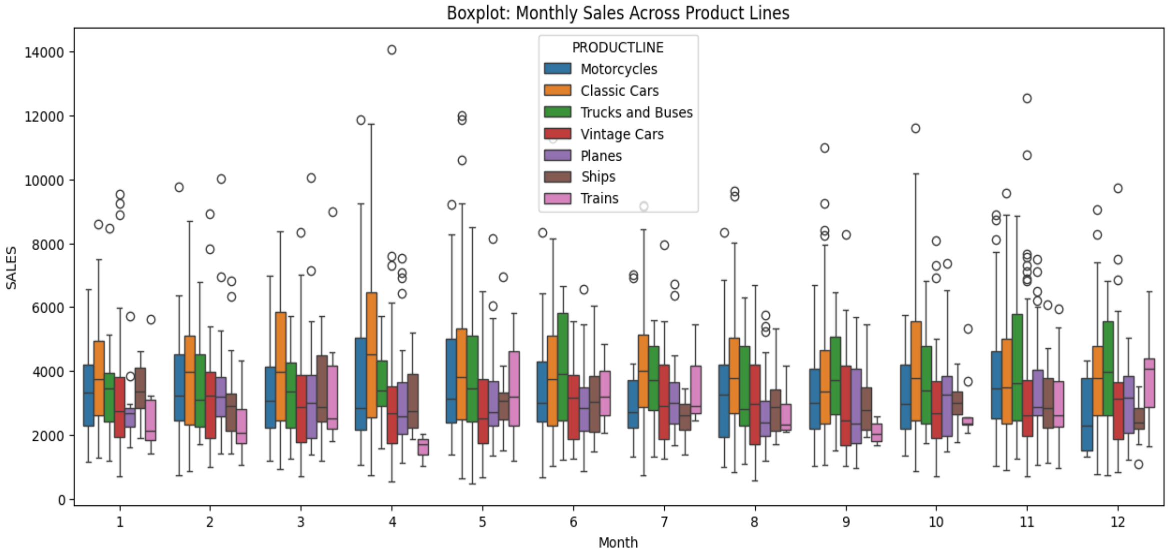
- There is a lot of variability in the data. This means that there is a lot of variation in the number of orders received for a given quantity ordered. In other words, not all customers who order a lot of items also receive a lot of orders. There are many other factors that can affect the number of orders received, such as the price of the items, the customer's budget, and the customer's needs.

- Overall, the heat map shows that there is a positive relationship between quantity ordered and the number of orders received, but there is also a lot of variability in this relationship. Other factors besides quantity ordered can also affect the number of orders received.

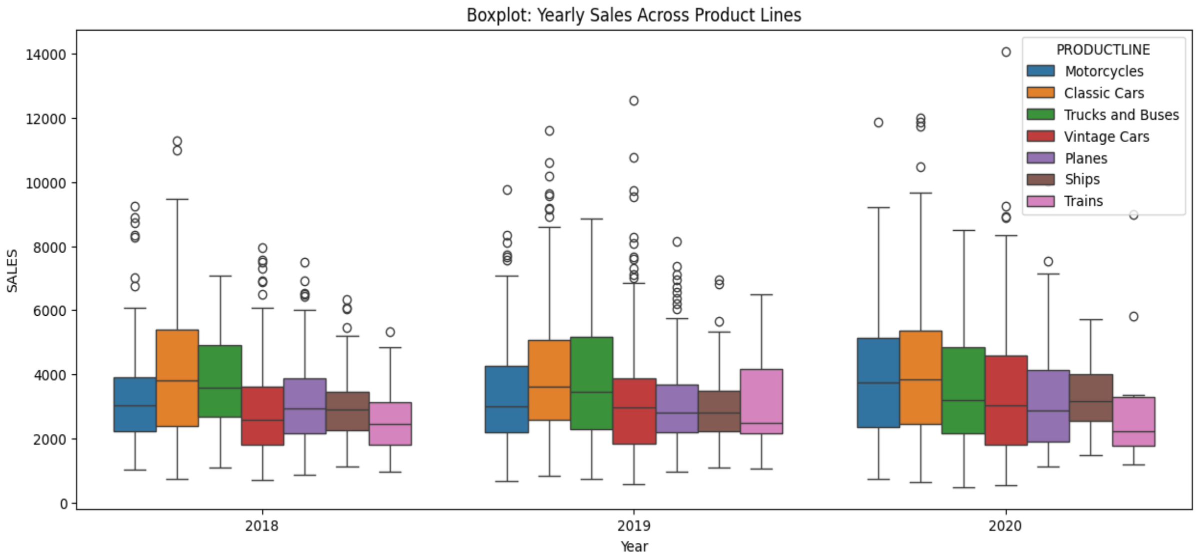
Multivariate Analysis:



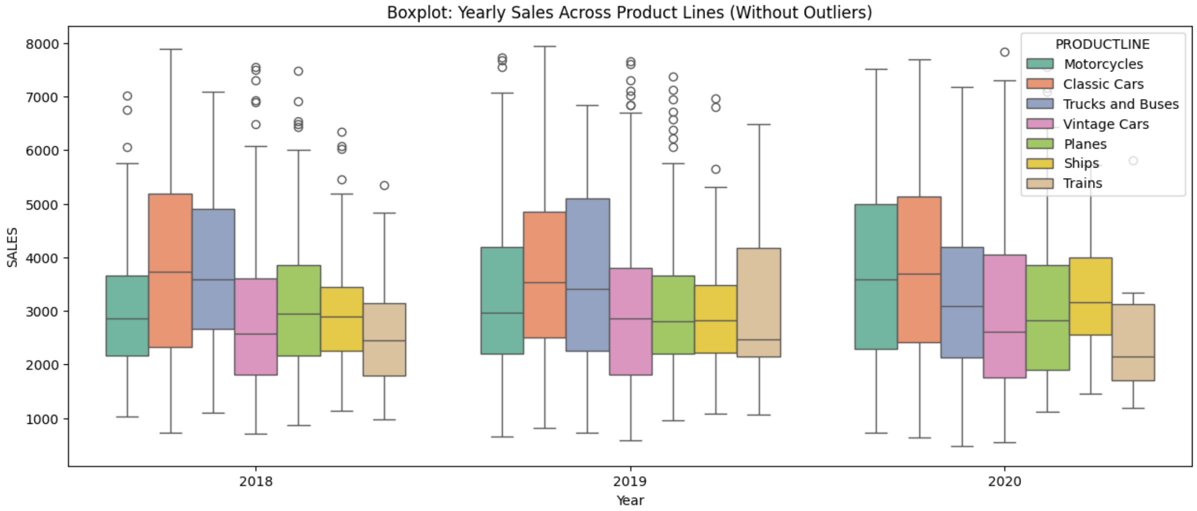
The boxplot shows that there are some differences in sales across product lines, with motorcycles and classic cars having higher sales on average. However, there is also a lot of variability within each product line, and there are some outliers in the data.



The boxplot suggests that Motorcycles and Classic Cars tend to have higher sales than the other product lines, but there is also more variability in their sales figures. Trucks and Buses have the lowest median sales, and their sales are also less variable sales patterns could change over time.



- Overall, sales appear to be higher for motorcycles and classic cars compared to other product lines. The medians for motorcycles and classic cars are both around 12,000, while the medians for the other product lines are all below 10,000.
- Trucks and buses seem to have the lowest median sales, around 6,000.
- There is a lot of variability in sales within each product line. The boxplots show that the interquartile ranges (IQRs) are all quite large, which means that there is a lot of variation in sales even within the same product line.

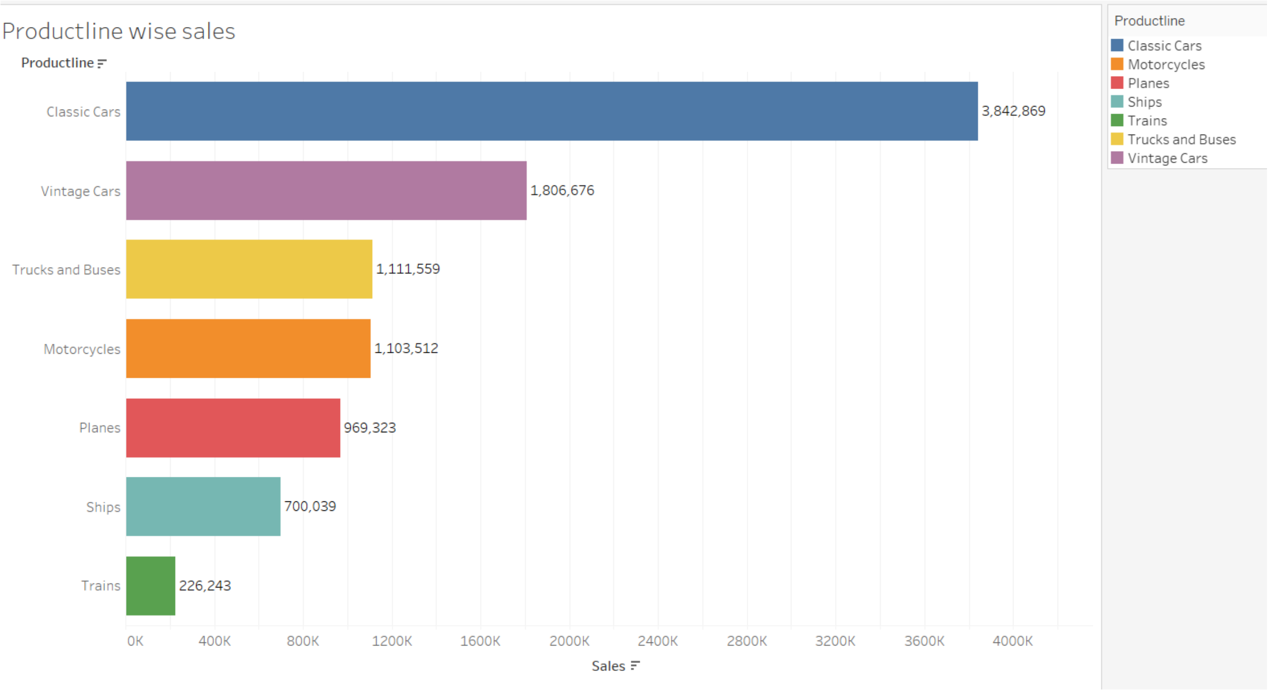


- The boxplot shows the distribution of sales across different product lines for a given year. It appears to be based on 5 product lines: Classic Cars, Motorcycles, Planes, Ships, and Trains.
- Motorcycles have the highest median sales, followed by Classic Cars. Their boxes are also shorter, indicating less variability in sales compared to other product lines.
- Trucks and Buses and Ships have the lowest median sales. Trucks and Buses also have a wider range of sales and a few outliers, suggesting more variability.
- Trains have the fewest sales overall, with a median sale close to 0. There are also no outliers for trains.

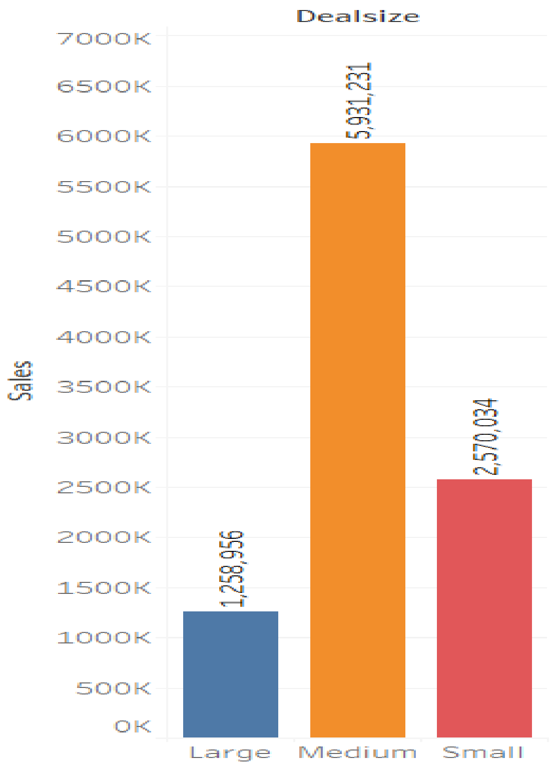
EDA analysis:-

- Univariate analysis
- Bivariate analysis
- Multi-variate analysis

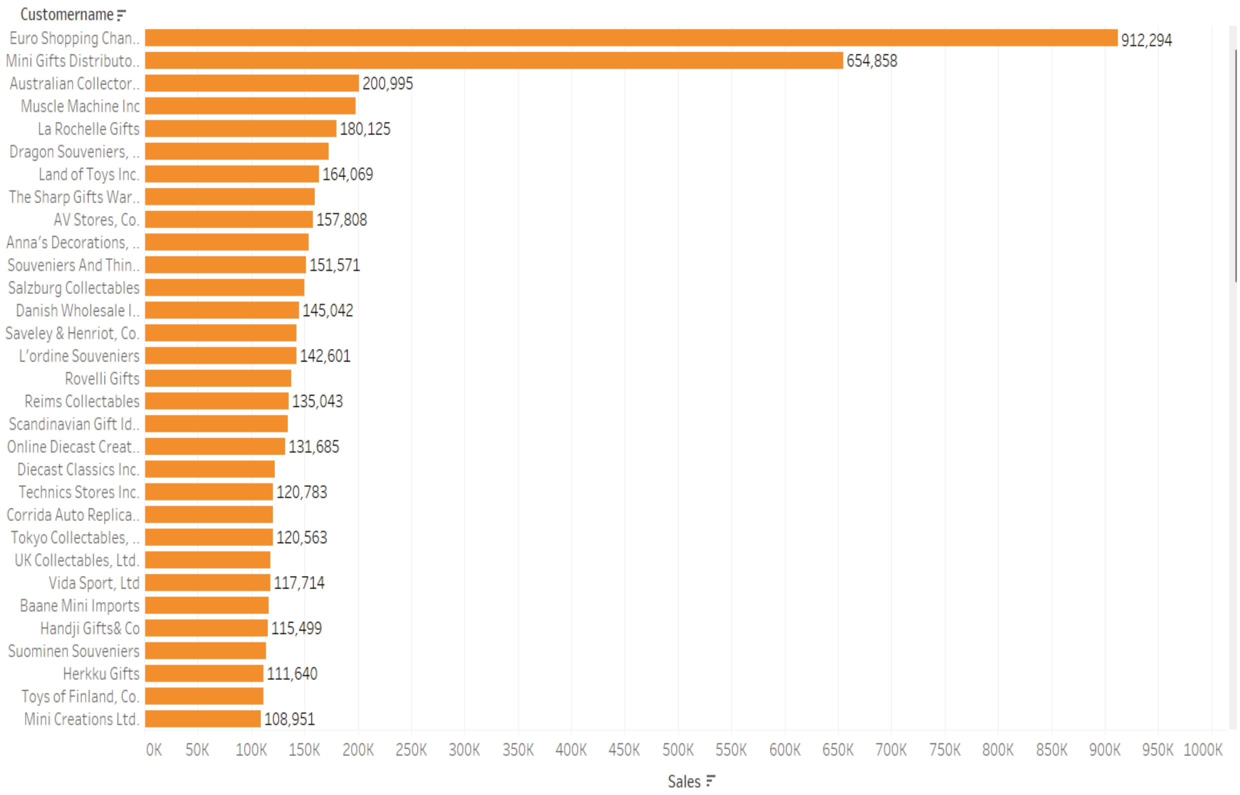
Tableau link - https://public.tableau.com/app/profile/raghavendra.kumar1327



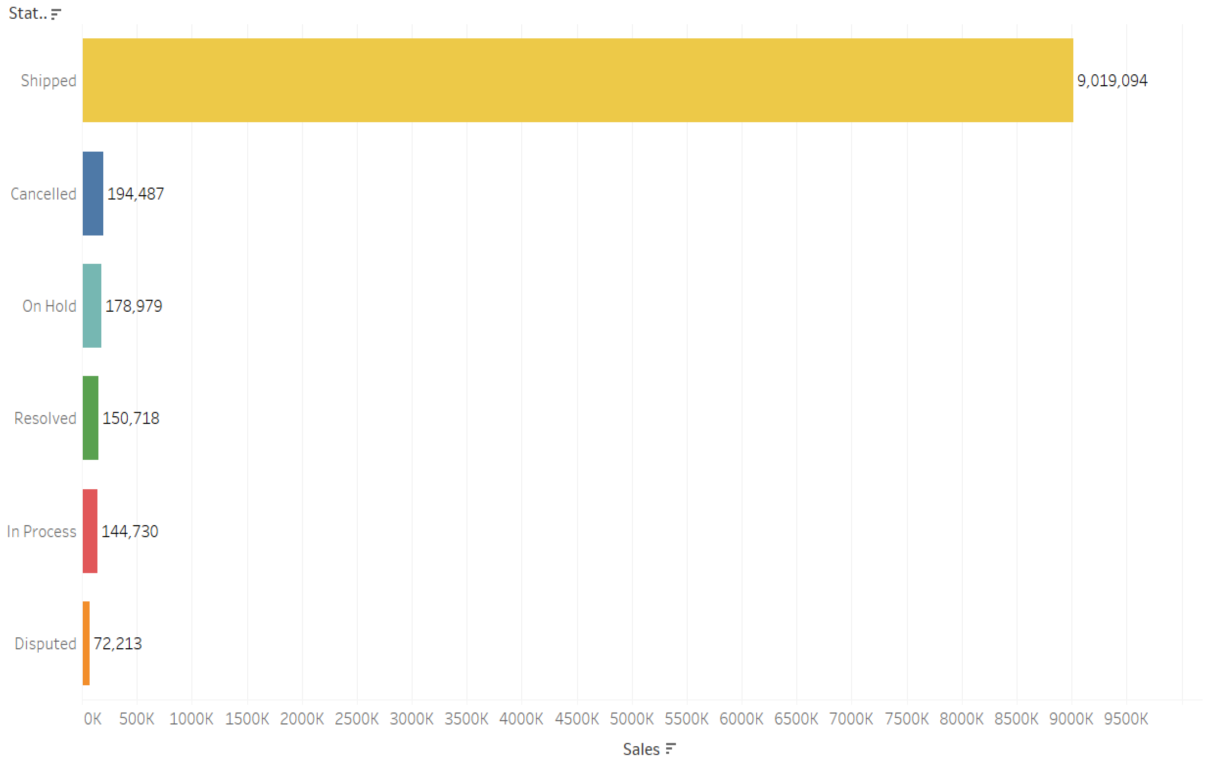
Deal vs Sales



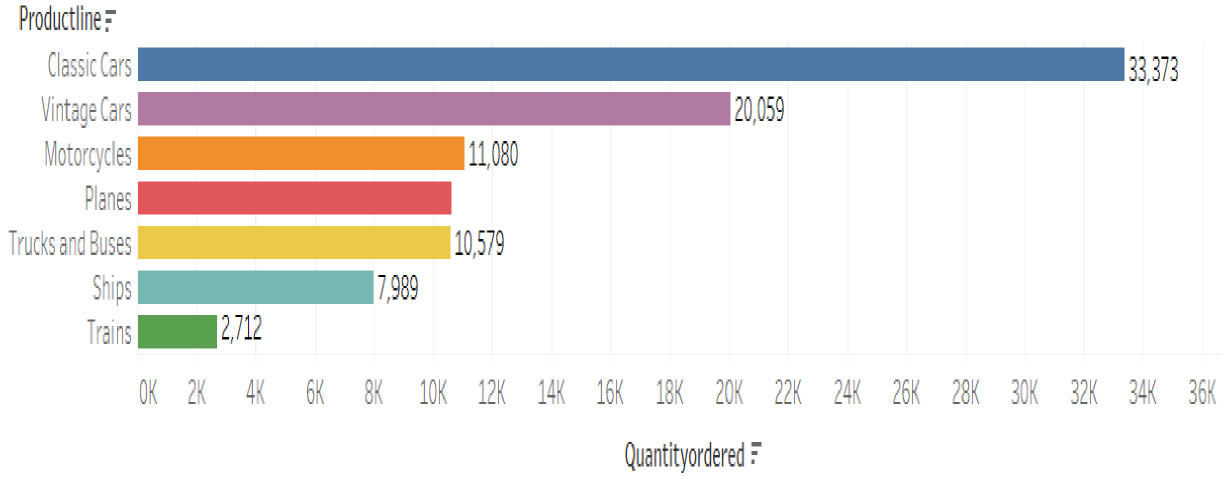
Customername vs Sales



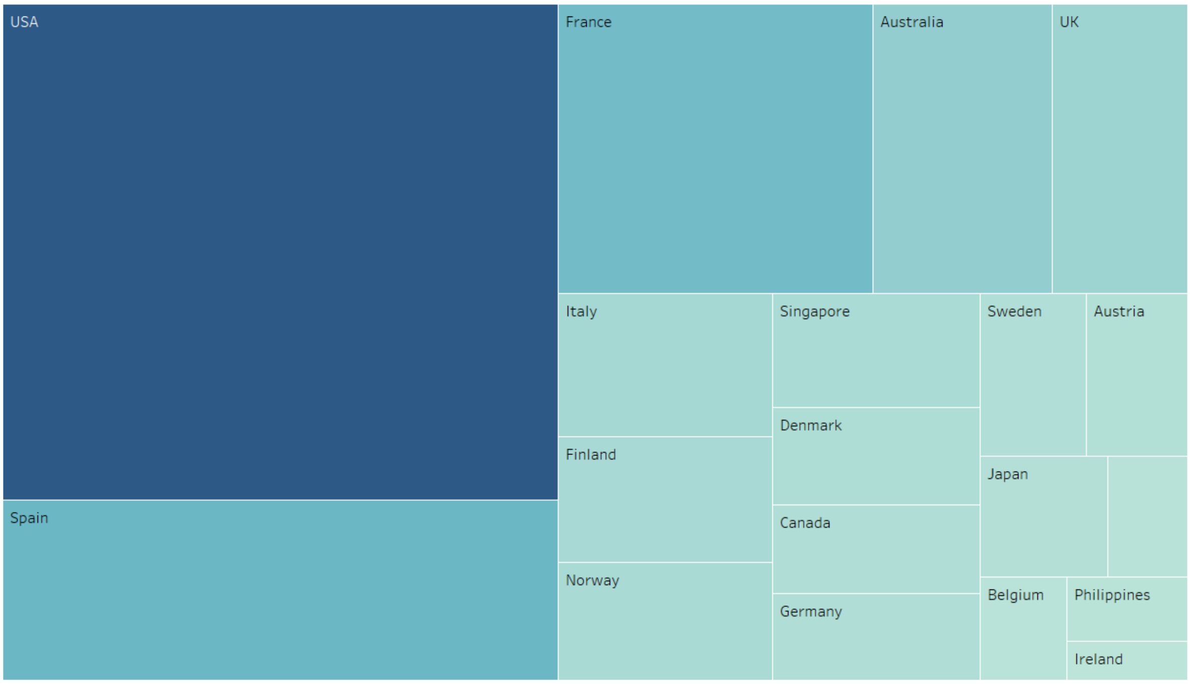
Sales across Status



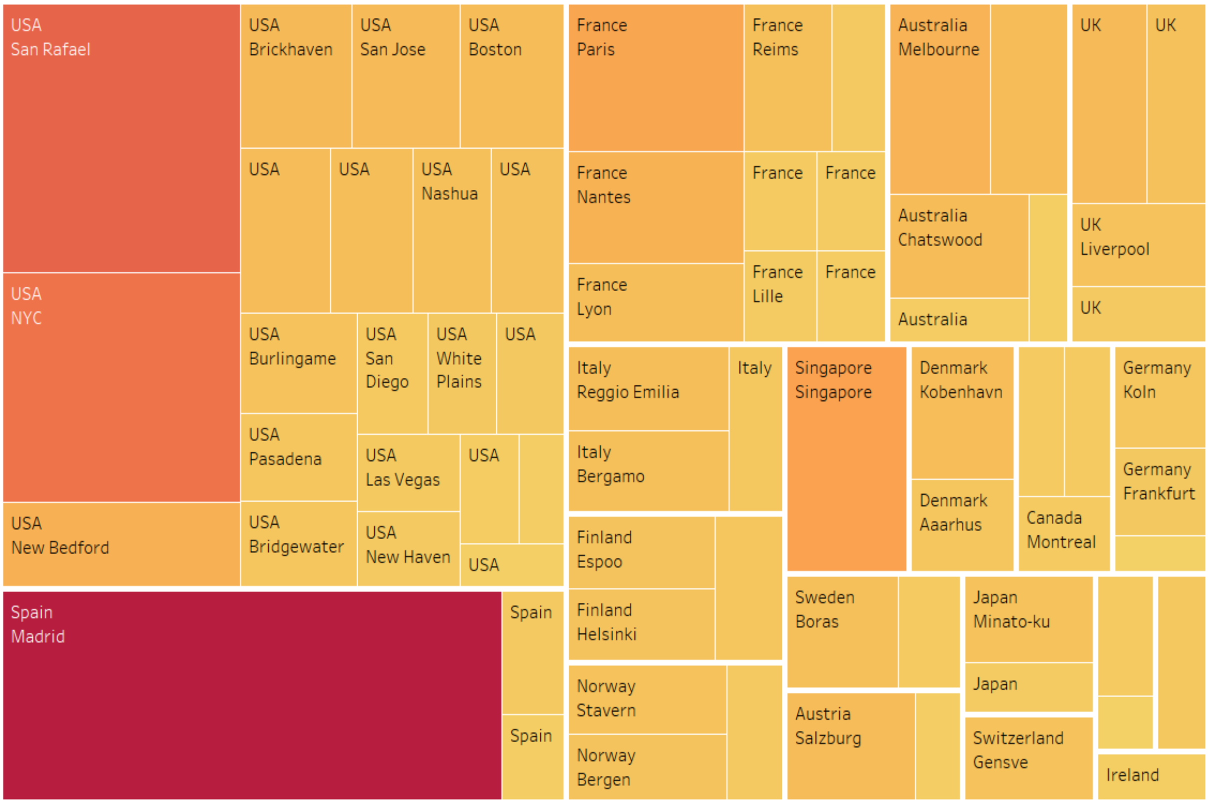
Quantity ordered on each product

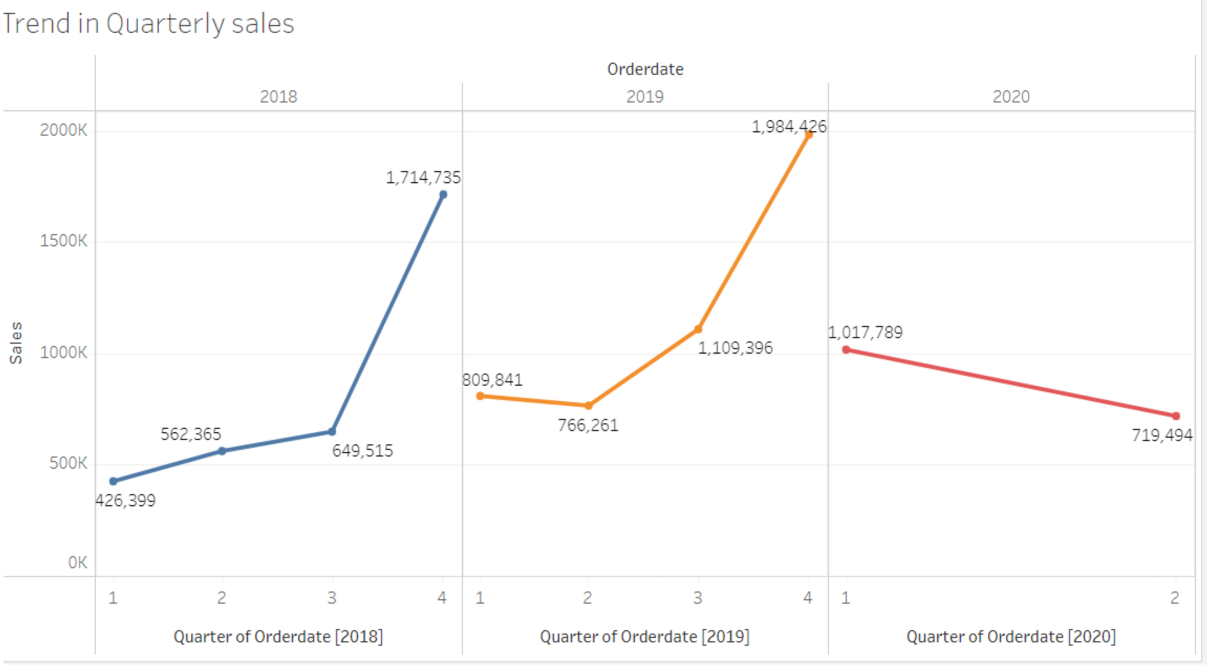


Countrywise Sales

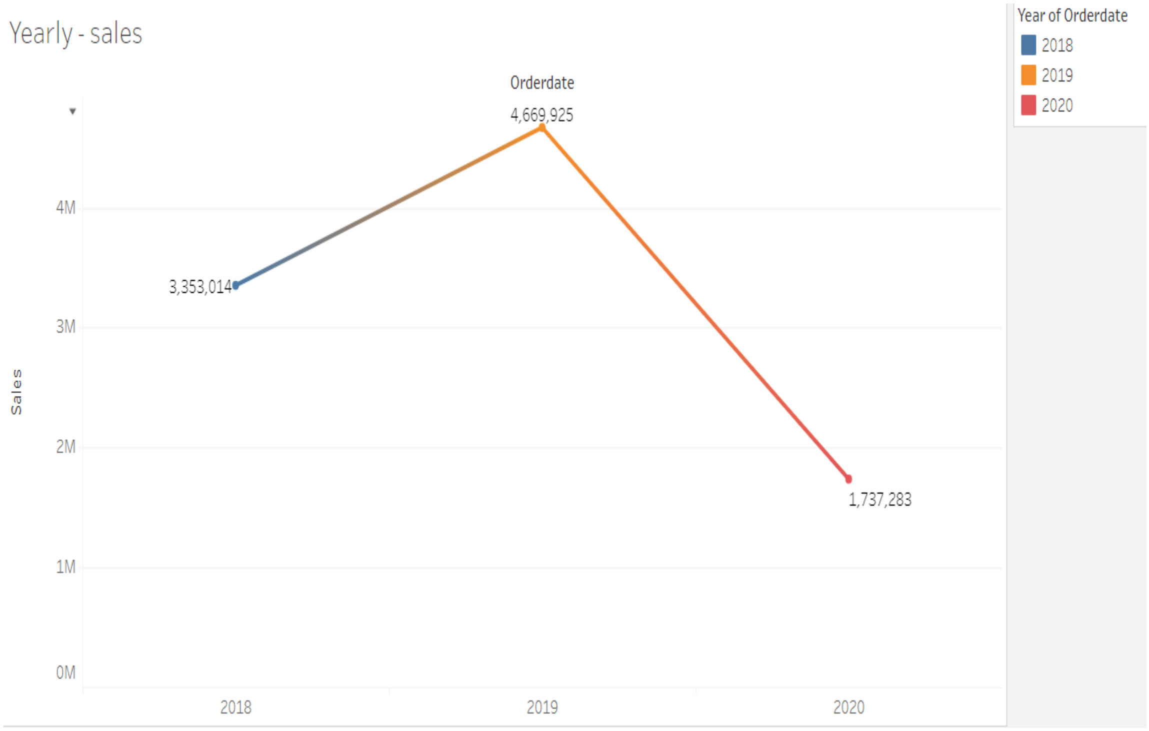


Country and Citywise Sales

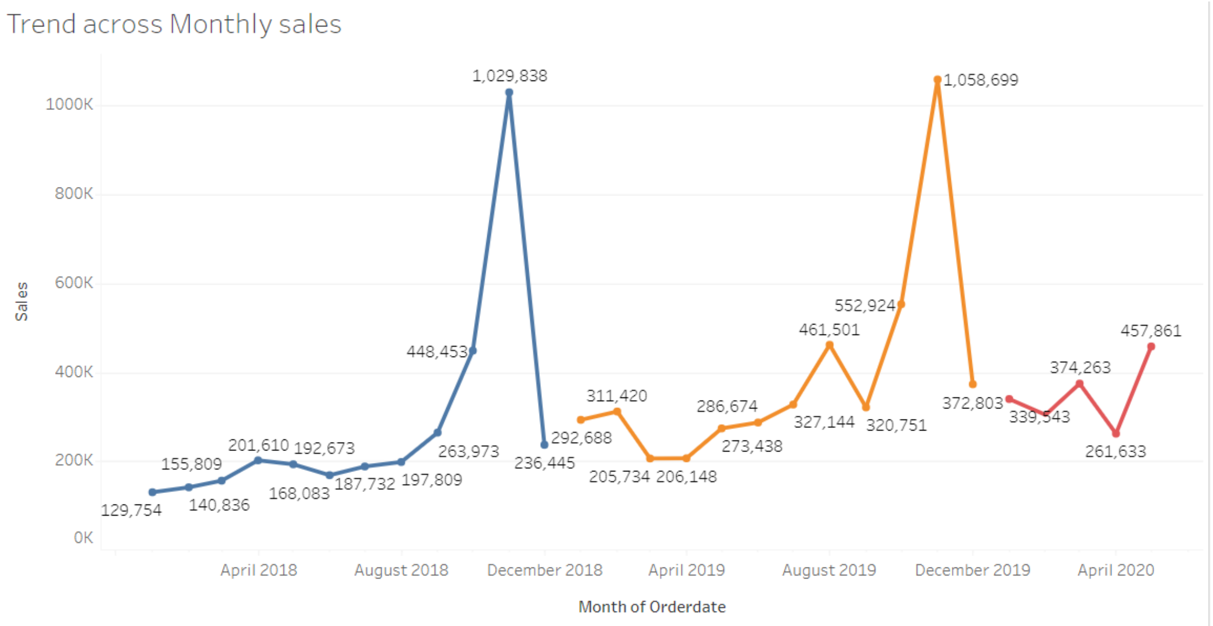




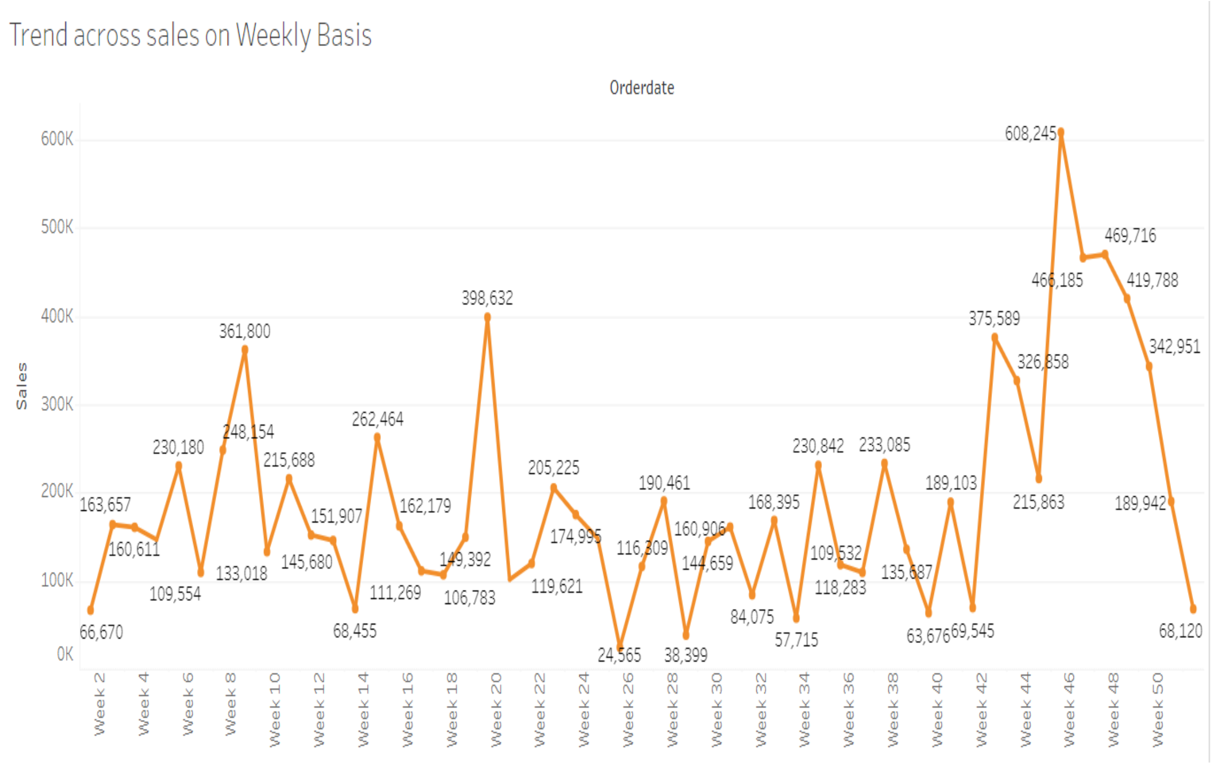
Trend in Quarterly sales



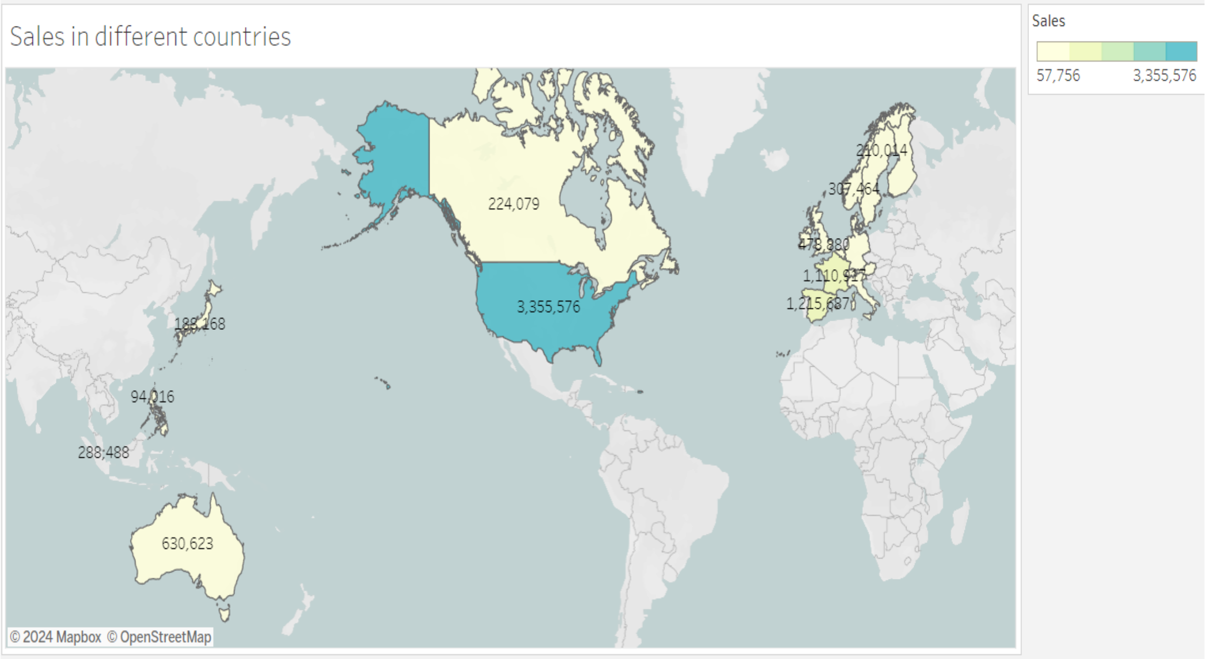
Yearly - sales

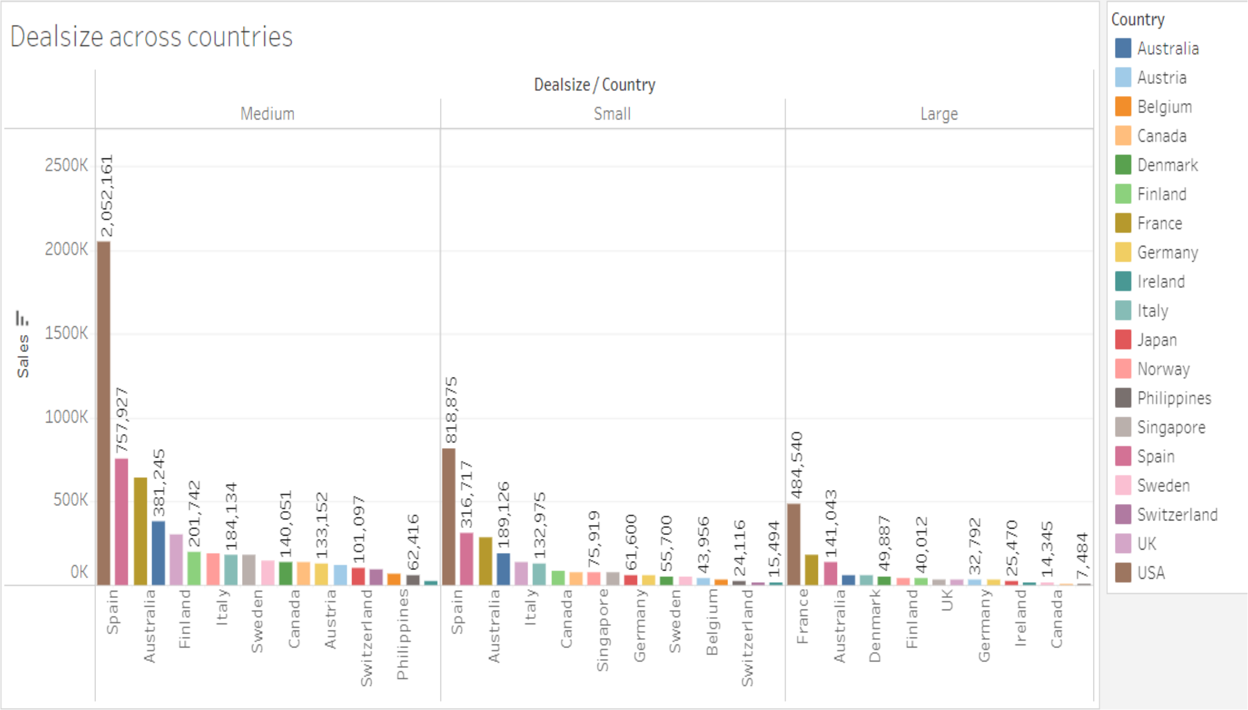


Trend across Monthly sales



Trend across sales on Weekly Basis

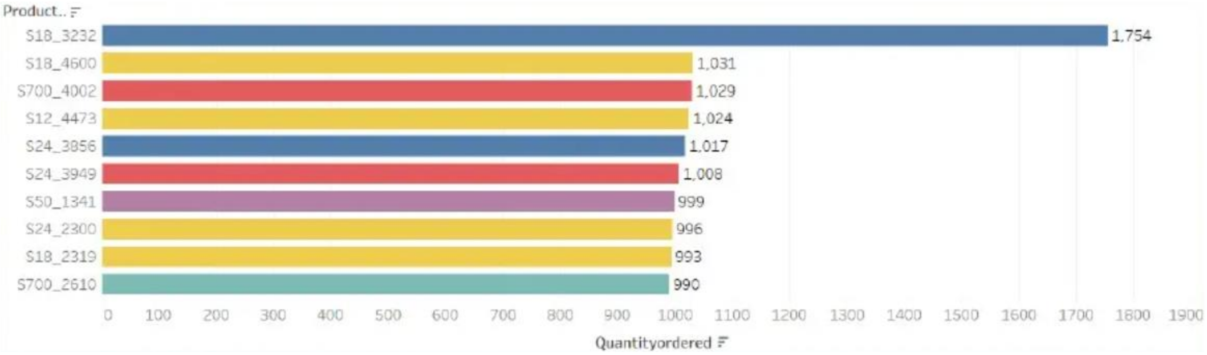




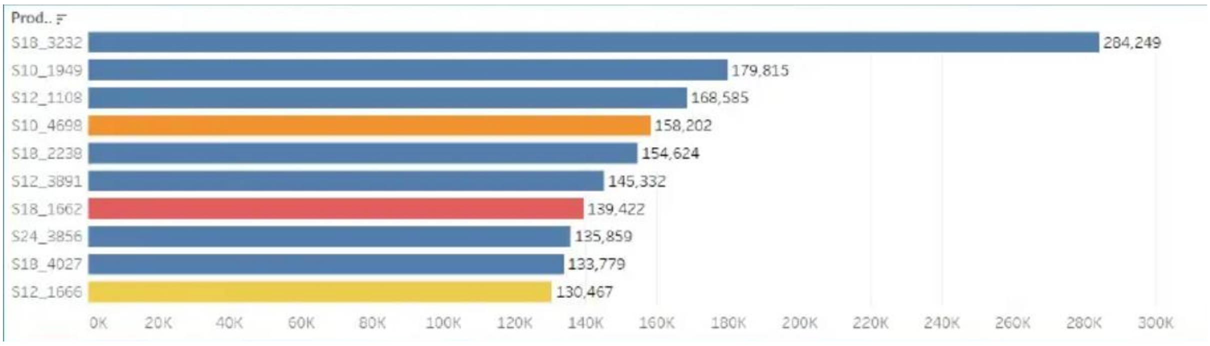
Trends in sales summary:

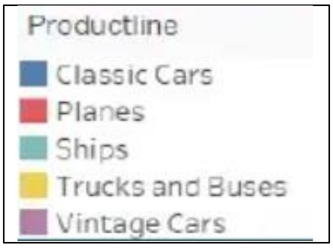
- It shows that week 45 seems to have the highest sales as compared to other weeks.
- The November month gives out the highest sales & week 45 falls in November, hence, it also validates the first point.
- Sales are decreasing from Q1 to Q3 and took a huge rise in Q4.
- 2019 seems to deliver the highest sales.

MOST PURCHASED TOP 10 PRODUCTS



TOP 10 MOST SELLING PRODUCTS





Sales trends across different weeks and months:

- Weeks: The x-axis represents weeks, ranging from 1 to 52.
- Months: The y-axis represents months, likely from January to December.
- Sales Values: Each cell in the heat map is coloured according to the sales value for that week and month. Darker colours indicate higher sales, while lighter colours indicate lower sales.
- Highest Sales: Week 45 appears to have the highest sales value, followed by weeks 44 and 46.
- Highest Sales Month: November seems to be the month with the highest overall sales.
- Sales Trend: Sales seem to be decreasing from the first quarter (Q1) to the third quarter (Q3), before experiencing a significant rise in the fourth quarter (Q4).
- Year: Based on the sales trend and November being the highest sales month, it is likely that the data represents the year 2019.

PART A: Customer Segmentation using RFM analysis (4 segments) -> what is RFM? -> What all parameters used and assumptions made? -> Showcase the KNIME workflow image -> what results are there in the output table head?

- RFM analysis is a method used for customer segmentation based on three key metrics.
- RFM analysis is a marketing technique used to segment customers based on their purchasing patterns Recency, Frequency, and Monetary Value.

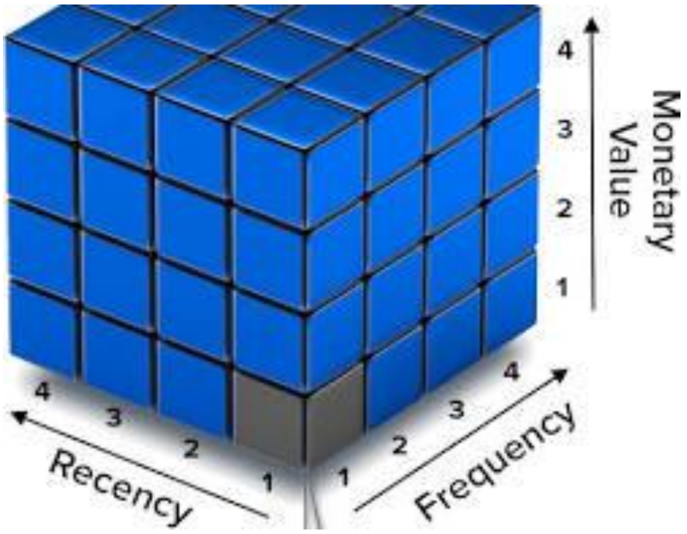
1. Recency (R): It measures how recently a customer has made a purchase. Customers who made a purchase more recently are considered more valuable.
2. Frequency (F): It measures how often a customer makes a purchase. Customers who make more frequent purchases are considered more valuable.
3. Monetary Value (M): It measures the total monetary value of a customer's purchases. Customers who have spent more money are considered more valuable.

Each customer is assigned a score for each of these metrics, and these scores are used to categorize customers into segments.

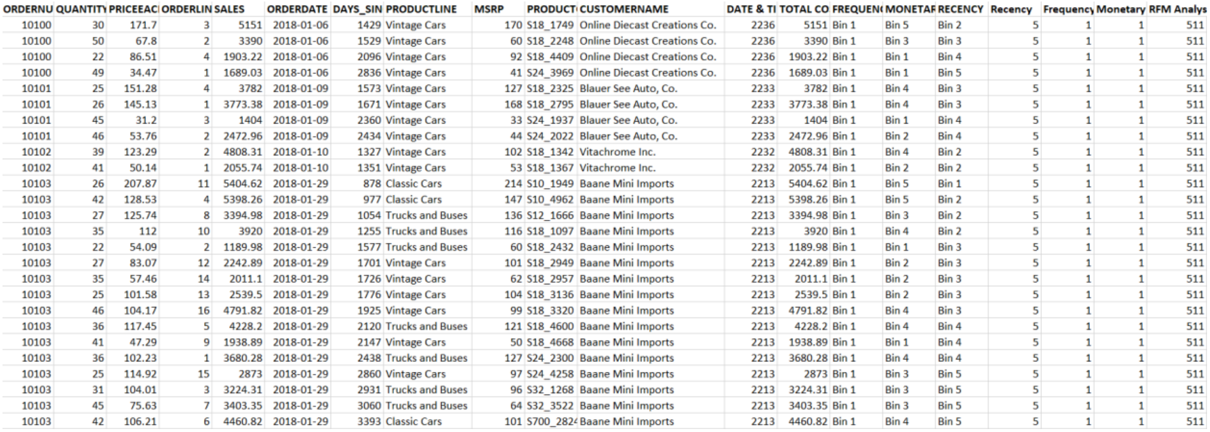
Parameters and Assumptions:

- The RFM analysis involves assigning numerical scores to each customer based on the Recency, frequency, and monetary value parameters.
- Typically, customers are segmented into four categories (quartiles or percentiles) for each parameter, resulting in a total of 64 segments (4^3).
- The specific scoring system and segmentation thresholds may vary based on business needs and domain knowledge.
- Commonly, a higher score indicates a more valuable customer in each parameter.
- The combination of these scores creates segments that can be used for targeted marketing and personalized communication.

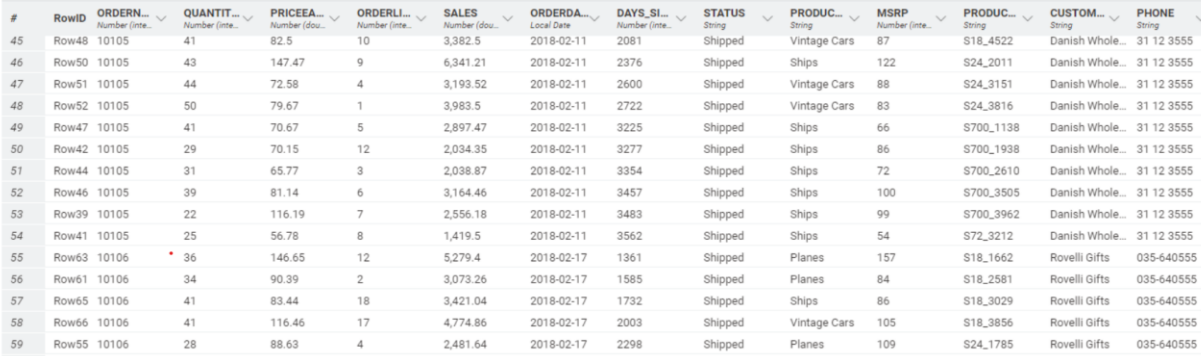
RFM Example :



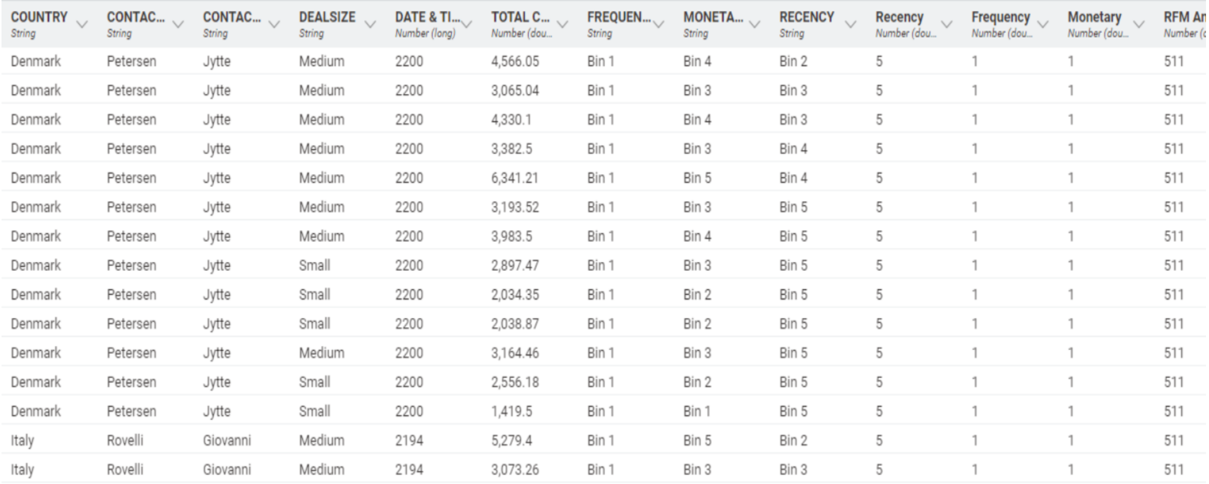
OUTPUT TABLE HEAD:



ORDERNU	QUANTITY	PRICEEAC	ORDERLIN	SALES	ORDERDATE	DAYS_SIN	PRODUCTLINE	MSRP	PRODUCT	CUSTOMERNAME	DATE & TI	TOTAL CO	FREQUEN	MONETAF	RECENCY	Recency	Frequency	Monetary	RFM Analys
10100	30	171.7	3	5151	2018-01-06	1429	Vintage Cars	170	S18_1749	Online Diecast Creations Co.	2236	5151	Bin 1	Bin 5	Bin 2	5	1	1	511
10100	50	67.8	2	3390	2018-01-06	1529	Vintage Cars	60	S18_2248	Online Diecast Creations Co.	2236	3390	Bin 1	Bin 3	Bin 3	5	1	1	511
10100	22	86.51	4	1903.22	2018-01-06	2096	Vintage Cars	92	S18_4409	Online Diecast Creations Co.	2236	1903.22	Bin 1	Bin 1	Bin 4	5	1	1	511
10100	49	34.47	1	1689.03	2018-01-06	2836	Vintage Cars	41	S24_3969	Online Diecast Creations Co.	2236	1689.03	Bin 1	Bin 1	Bin 5	5	1	1	511
10101	25	151.28	4	3782	2018-01-09	1573	Vintage Cars	127	S18_2325	Blauer See Auto, Co.	2233	3782	Bin 1	Bin 4	Bin 3	5	1	1	511
10101	26	145.13	1	3773.38	2018-01-09	1671	Vintage Cars	168	S18_2795	Blauer See Auto, Co.	2233	3773.38	Bin 1	Bin 4	Bin 3	5	1	1	511
10101	45	31.2	3	1404	2018-01-09	2360	Vintage Cars	33	S24_1937	Blauer See Auto, Co.	2233	1404	Bin 1	Bin 1	Bin 4	5	1	1	511
10101	46	53.76	2	2472.96	2018-01-09	2434	Vintage Cars	44	S24_2022	Blauer See Auto, Co.	2233	2472.96	Bin 1	Bin 2	Bin 4	5	1	1	511
10102	39	123.29	2	4808.31	2018-01-10	1327	Vintage Cars	102	S18_1342	Vitachrome Inc.	2232	4808.31	Bin 1	Bin 4	Bin 2	5	1	1	511
10102	41	50.14	1	2055.74	2018-01-10	1351	Vintage Cars	53	S18_1367	Vitachrome Inc.	2232	2055.74	Bin 1	Bin 2	Bin 2	5	1	1	511
10103	26	207.87	11	5404.62	2018-01-29	878	Classic Cars	214	S10_1949	Baane Mini Imports	2213	5404.62	Bin 1	Bin 5	Bin 1	5	1	1	511
10103	42	128.53	4	5398.26	2018-01-29	977	Classic Cars	147	S10_4962	Baane Mini Imports	2213	5398.26	Bin 1	Bin 5	Bin 2	5	1	1	511
10103	27	125.74	8	3394.98	2018-01-29	1054	Trucks and Buses	136	S12_1666	Baane Mini Imports	2213	3394.98	Bin 1	Bin 3	Bin 2	5	1	1	511
10103	35	112	10	3920	2018-01-29	1255	Trucks and Buses	116	S18_1097	Baane Mini Imports	2213	3920	Bin 1	Bin 4	Bin 2	5	1	1	511
10103	22	54.09	2	1189.98	2018-01-29	1577	Trucks and Buses	60	S18_2432	Baane Mini Imports	2213	1189.98	Bin 1	Bin 1	Bin 3	5	1	1	511
10103	27	83.07	12	2242.89	2018-01-29	1701	Vintage Cars	101	S18_2949	Baane Mini Imports	2213	2242.89	Bin 1	Bin 2	Bin 3	5	1	1	511
10103	35	57.46	14	2011.1	2018-01-29	1726	Vintage Cars	62	S18_2957	Baane Mini Imports	2213	2011.1	Bin 1	Bin 2	Bin 3	5	1	1	511
10103	25	101.58	13	2539.5	2018-01-29	1776	Vintage Cars	104	S18_3136	Baane Mini Imports	2213	2539.5	Bin 1	Bin 2	Bin 3	5	1	1	511
10103	46	104.17	16	4791.82	2018-01-29	1925	Vintage Cars	99	S18_3320	Baane Mini Imports	2213	4791.82	Bin 1	Bin 4	Bin 3	5	1	1	511
10103	36	117.45	5	4228.2	2018-01-29	2120	Trucks and Buses	121	S18_4600	Baane Mini Imports	2213	4228.2	Bin 1	Bin 4	Bin 4	5	1	1	511
10103	41	47.29	9	1938.89	2018-01-29	2147	Vintage Cars	50	S18_4668	Baane Mini Imports	2213	1938.89	Bin 1	Bin 1	Bin 4	5	1	1	511
10103	36	102.23	1	3680.28	2018-01-29	2438	Trucks and Buses	127	S24_2300	Baane Mini Imports	2213	3680.28	Bin 1	Bin 4	Bin 4	5	1	1	511
10103	25	114.92	15	2873	2018-01-29	2860	Vintage Cars	97	S24_4258	Baane Mini Imports	2213	2873	Bin 1	Bin 3	Bin 5	5	1	1	511
10103	31	104.01	3	3224.31	2018-01-29	2931	Trucks and Buses	96	S32_1268	Baane Mini Imports	2213	3224.31	Bin 1	Bin 3	Bin 5	5	1	1	511
10103	45	75.63	7	3403.35	2018-01-29	3060	Trucks and Buses	64	S32_3522	Baane Mini Imports	2213	3403.35	Bin 1	Bin 3	Bin 5	5	1	1	511
10103	42	106.21	6	4460.82	2018-01-29	3393	Classic Cars	101	S700_2824	Baane Mini Imports	2213	4460.82	Bin 1	Bin 4	Bin 5	5	1	1	511

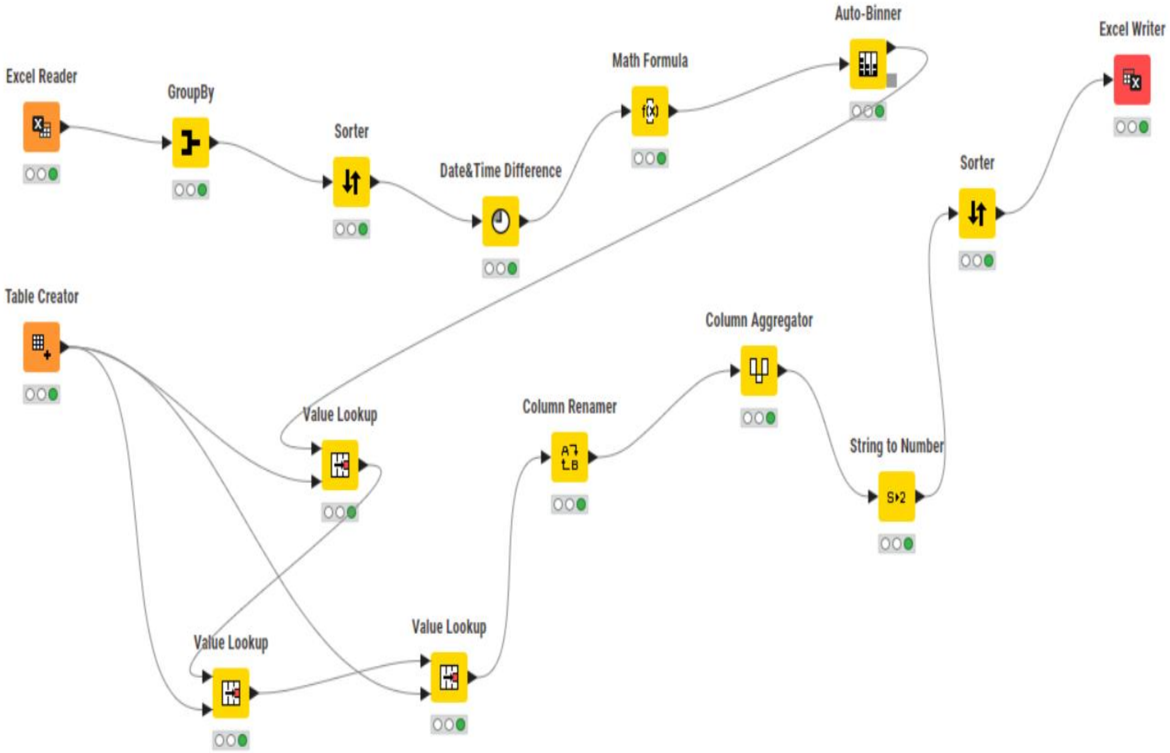


#	RowID	ORDERN...	QUANTIT...	PRICEEA...	ORDERLI...	SALES	ORDERDA...	DAYS_SI...	STATUS	PRODUC...	MSRP	PRODUC...	CUSTOM...	PHONE
		Number (inte...	Number (inte...	Number (dou...	Number (inte...	Number (dou...	Local Date	Number (inte...	String	String	Number (inte...	String	String	String
45	Row48	10105	41	82.5	10	3,382.5	2018-02-11	2081	Shipped	Vintage Cars	87	S18_4522	Danish Whole...	31 12 3555
46	Row50	10105	43	147.47	9	6,341.21	2018-02-11	2376	Shipped	Ships	122	S24_2011	Danish Whole...	31 12 3555
47	Row51	10105	44	72.58	4	3,193.52	2018-02-11	2600	Shipped	Vintage Cars	88	S24_3151	Danish Whole...	31 12 3555
48	Row52	10105	50	79.67	1	3,983.5	2018-02-11	2722	Shipped	Vintage Cars	83	S24_3816	Danish Whole...	31 12 3555
49	Row47	10105	41	70.67	5	2,897.47	2018-02-11	3225	Shipped	Ships	66	S700_1138	Danish Whole...	31 12 3555
50	Row42	10105	29	70.15	12	2,034.35	2018-02-11	3277	Shipped	Ships	86	S700_1938	Danish Whole...	31 12 3555
51	Row44	10105	31	65.77	3	2,038.87	2018-02-11	3354	Shipped	Ships	72	S700_2610	Danish Whole...	31 12 3555
52	Row46	10105	39	81.14	6	3,164.46	2018-02-11	3457	Shipped	Ships	100	S700_3505	Danish Whole...	31 12 3555
53	Row39	10105	22	116.19	7	2,556.18	2018-02-11	3483	Shipped	Ships	99	S700_3962	Danish Whole...	31 12 3555
54	Row41	10105	25	56.78	8	1,419.5	2018-02-11	3562	Shipped	Ships	54	S72_3212	Danish Whole...	31 12 3555
55	Row63	10106	36	146.65	12	5,279.4	2018-02-17	1361	Shipped	Planes	157	S18_1662	Rovelli Gifts	035-640555
56	Row61	10106	34	90.39	2	3,073.26	2018-02-17	1585	Shipped	Planes	84	S18_2581	Rovelli Gifts	035-640555
57	Row65	10106	41	83.44	18	3,421.04	2018-02-17	1732	Shipped	Ships	86	S18_3029	Rovelli Gifts	035-640555
58	Row66	10106	41	116.46	17	4,774.86	2018-02-17	2003	Shipped	Vintage Cars	105	S18_3856	Rovelli Gifts	035-640555
59	Row55	10106	28	88.63	4	2,481.64	2018-02-17	2298	Shipped	Planes	109	S24_1785	Rovelli Gifts	035-640555



COUNTRY	CONTAC...	CONTAC...	DEALSIZE	DATE & TI...	TOTAL C...	FREQUEN...	MONETA...	RECENCY	Recency	Frequency	Monetary	RFM An
String	String	String	String	Number (long)	Number (dou...	String	String	String	Number (dou...	Number (dou...	Number (dou...	Number (
Denmark	Petersen	Jytte	Medium	2200	4,566.05	Bin 1	Bin 4	Bin 2	5	1	1	511
Denmark	Petersen	Jytte	Medium	2200	3,065.04	Bin 1	Bin 3	Bin 3	5	1	1	511
Denmark	Petersen	Jytte	Medium	2200	4,330.1	Bin 1	Bin 4	Bin 3	5	1	1	511
Denmark	Petersen	Jytte	Medium	2200	3,382.5	Bin 1	Bin 3	Bin 4	5	1	1	511
Denmark	Petersen	Jytte	Medium	2200	6,341.21	Bin 1	Bin 5	Bin 4	5	1	1	511
Denmark	Petersen	Jytte	Medium	2200	3,193.52	Bin 1	Bin 3	Bin 5	5	1	1	511
Denmark	Petersen	Jytte	Medium	2200	3,983.5	Bin 1	Bin 4	Bin 5	5	1	1	511
Denmark	Petersen	Jytte	Small	2200	2,897.47	Bin 1	Bin 3	Bin 5	5	1	1	511
Denmark	Petersen	Jytte	Small	2200	2,034.35	Bin 1	Bin 2	Bin 5	5	1	1	511
Denmark	Petersen	Jytte	Small	2200	2,038.87	Bin 1	Bin 2	Bin 5	5	1	1	511
Denmark	Petersen	Jytte	Medium	2200	3,164.46	Bin 1	Bin 3	Bin 5	5	1	1	511
Denmark	Petersen	Jytte	Small	2200	2,556.18	Bin 1	Bin 2	Bin 5	5	1	1	511
Denmark	Petersen	Jytte	Small	2200	1,419.5	Bin 1	Bin 1	Bin 5	5	1	1	511
Italy	Rovelli	Giovanni	Medium	2194	5,279.4	Bin 1	Bin 5	Bin 2	5	1	1	511
Italy	Rovelli	Giovanni	Medium	2194	3,073.26	Bin 1	Bin 3	Bin 3	5	1	1	511

KNIME WORKFLOW:



PART A: Inferences from RFM Analysis and identified segments -> who are your best customers? (Give at least 5) -> Which customers are on the verge of churning? (Give at least 5) -> Who are your lost customers? (Give at least 5) -> Who are your loyal customers? (Give at least 5)

Best Customers:

1. Euro Shopping Channel with an RFM score of 511
2. Anna's Decorations, Ltd with an RFM score of 511
3. Online Diecast Creations Co. with an RFM score of 511
4. Souvenirs and Things Co. with an RFM score of 511
5. Salzburg Collectibles with an RFM score of 511

Customers on the verge of churning

1. Dhanish Wholesale Imports with an RFM score of 144
2. Reims Collectibles with an RFM score of 144
3. Dragon Souvenirs, Ltd. with an RFM score of 144
4. Muscle Machine Inc. with an RFM score of 144
5. Land of Toys Inc. with an RFM score of 144

Lost Customers:

1. Alpha Cognac with RFM score of 555
2. Mini Auto Werke with RFM score of 555
3. Australian Gift Network, Co with RFM score of 555
4. Gift Ideas Corp. with RFM score of 555
5. Auto-Moto Classics Inc. with RFM score of 555

Loyal Customers:

1. Euro Shopping Channel with RFM score of 511
2. Anna's Decorations, Ltd with RFM score of 511
3. Online Diecast Creations Co. with RFM score of 511
4. UK Collectables, Ltd. with RFM score of 511
5. Oulu Toy Supplies, Inc. with RFM score of 511

MRA – PART B – Contents:

PART B: Exploratory Analysis --> Exploratory Analysis of data & an executive summary (in PPT) of your top findings, supported by graphs. --> Are there trends across months/years/quarters/days etc. that you are able to notice?

PART B: Use of Market Basket Analysis (Association Rules) -->Write Something about the association rules and its relevance in this case -->Add KNIME workflow image -->Write about threshold values of Support and Confidence

PART B: Associations Identified --> Put the associations in a tabular manner --> Explain about support, confidence, & lift values that are calculated

PART B: Suggestion of Possible Combos with Lucrative Offers --> Write recommendations --> Make discount offers or combos (or buy two get one free) based on the associations and your experience

Problem Statement:

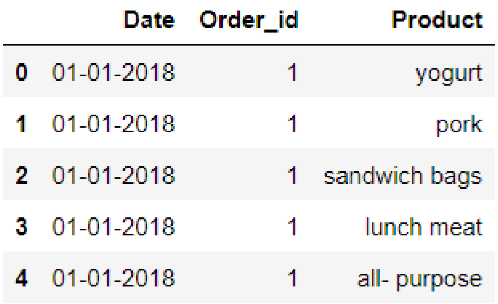
A grocery store shared the transactional data with you. Your job is to conduct a thorough analysis of Point of Sale (POS) data, identify the most commonly occurring sets of items in the customer orders, and provide recommendations through which a grocery store can increase its revenue by popular combo offers & discounts for customers.

PART B: Exploratory Analysis --> Exploratory Analysis of data & an executive summary (in PPT) of your top findings, supported by graphs. --> Are there trends across months/years/quarters/days etc. that you are able to notice?

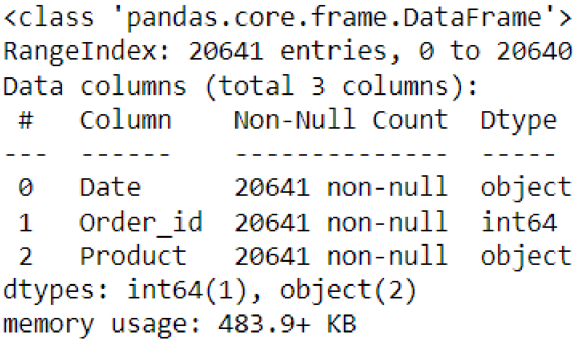
Shape of the data set 20641 rows, 3 Columns

Number of variables: 1 numerical columns, 1 – date time, 1- categorical variables.

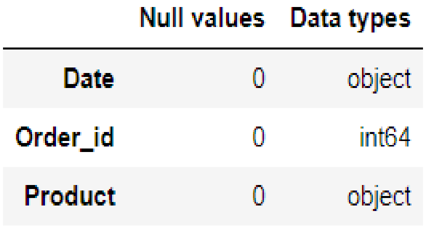
Zero ‘0’ null values in the data.



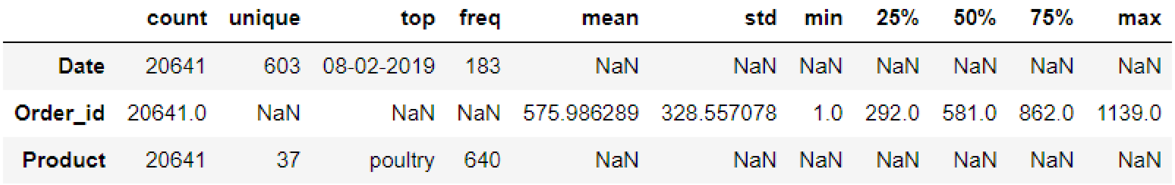
	Date	Order_id	Product
0	01-01-2018	1	yogurt
1	01-01-2018	1	pork
2	01-01-2018	1	sandwich bags
3	01-01-2018	1	lunch meat
4	01-01-2018	1	all- purpose



```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 20641 entries, 0 to 20640
Data columns (total 3 columns):
 #   Column    Non-Null Count  Dtype 
---  ------    --------------  ----- 
 0   Date      20641 non-null  object
 1   Order_id  20641 non-null  int64 
 2   Product   20641 non-null  object
dtypes: int64(1), object(2)
memory usage: 483.9+ KB
```



	Null values	Data types
Date	0	object
Order_id	0	int64
Product	0	object

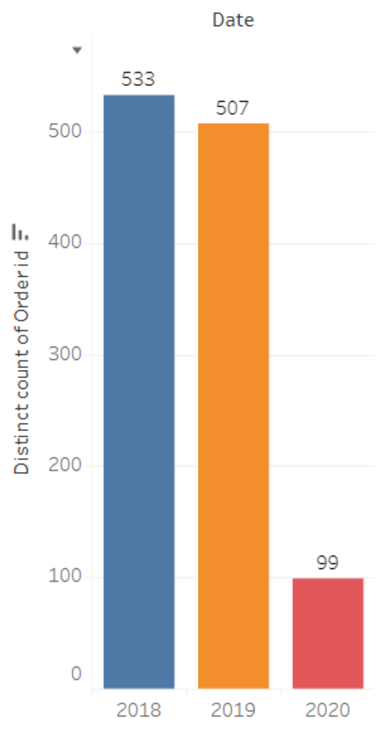


	count	unique	top	freq	mean	std	min	25%	50%	75%	max
Date	20641	603	08-02-2019	183	NaN	NaN	NaN	NaN	NaN	NaN	NaN
Order_id	20641.0	NaN	NaN	NaN	575.986289	328.557078	1.0	292.0	581.0	862.0	1139.0
Product	20641	37	poultry	640	NaN	NaN	NaN	NaN	NaN	NaN	NaN

The EDA analysis done in Tableau tool – with the workflow published in tableau public:

Link - https://public.tableau.com/app/profile/raghavendra.kumar1327

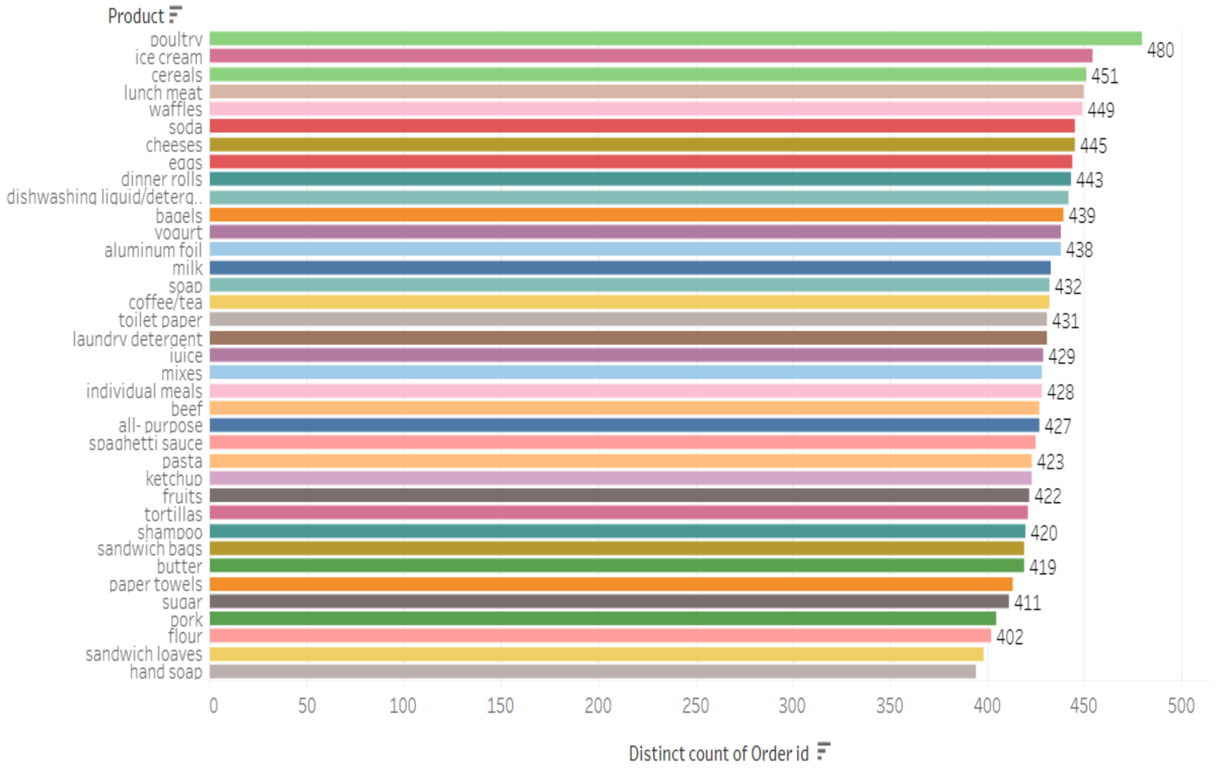
Yearly Sales Data

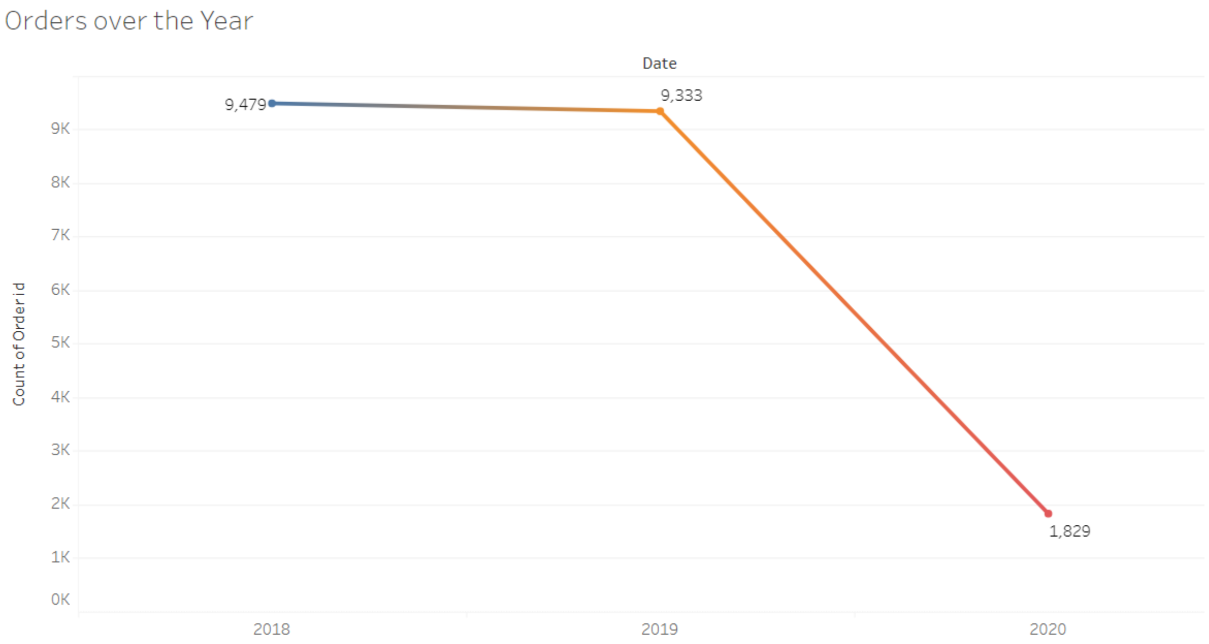


Orders over the years:

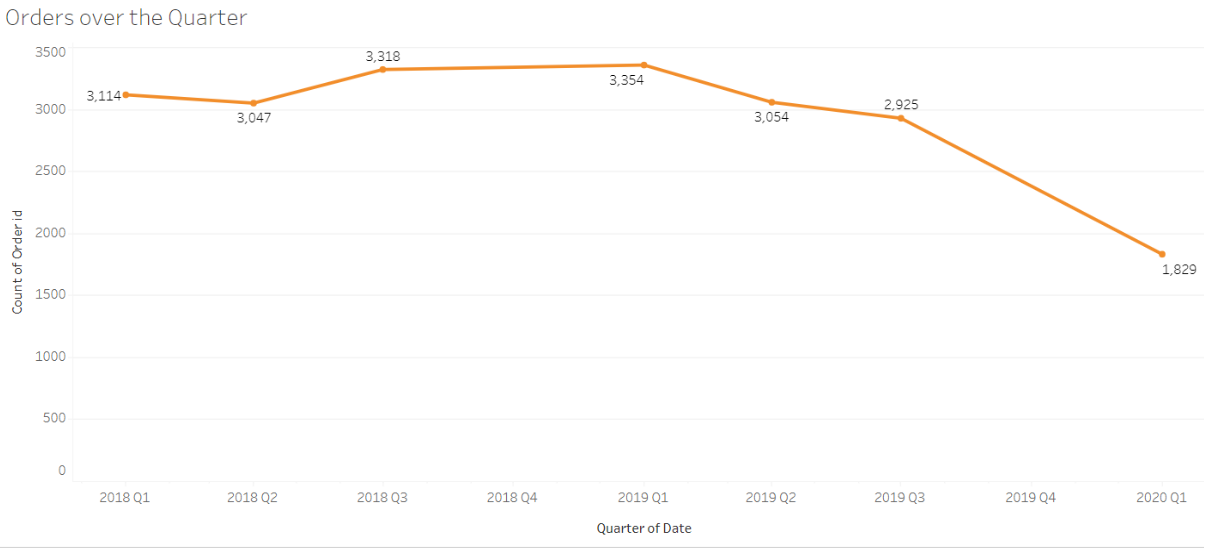
In 2018 the order number is highest amongst the consolidated data and 2020 is the lowest numbers of orders but having only two months of data.

Products and Orders





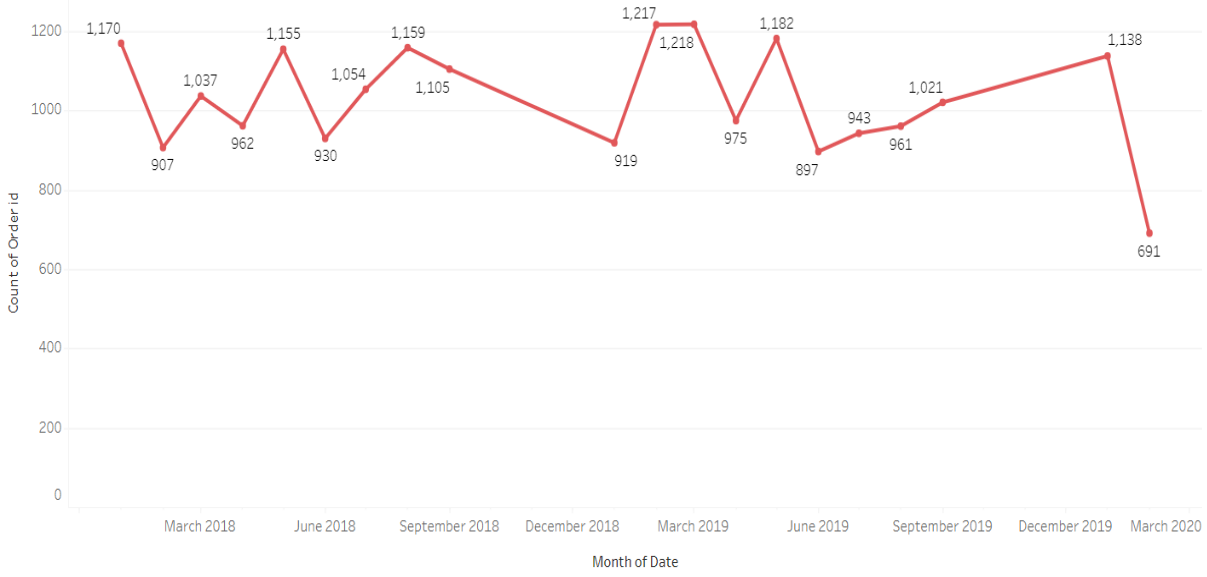
The order trend for the given data decreasing over the years with 2018 having highest orders and then followed closely by 2019 and then 2020 with the lowest number.



The order trend for the given data over the quarters indicate 2019 with highest number of orders flowed by Q3 2018.

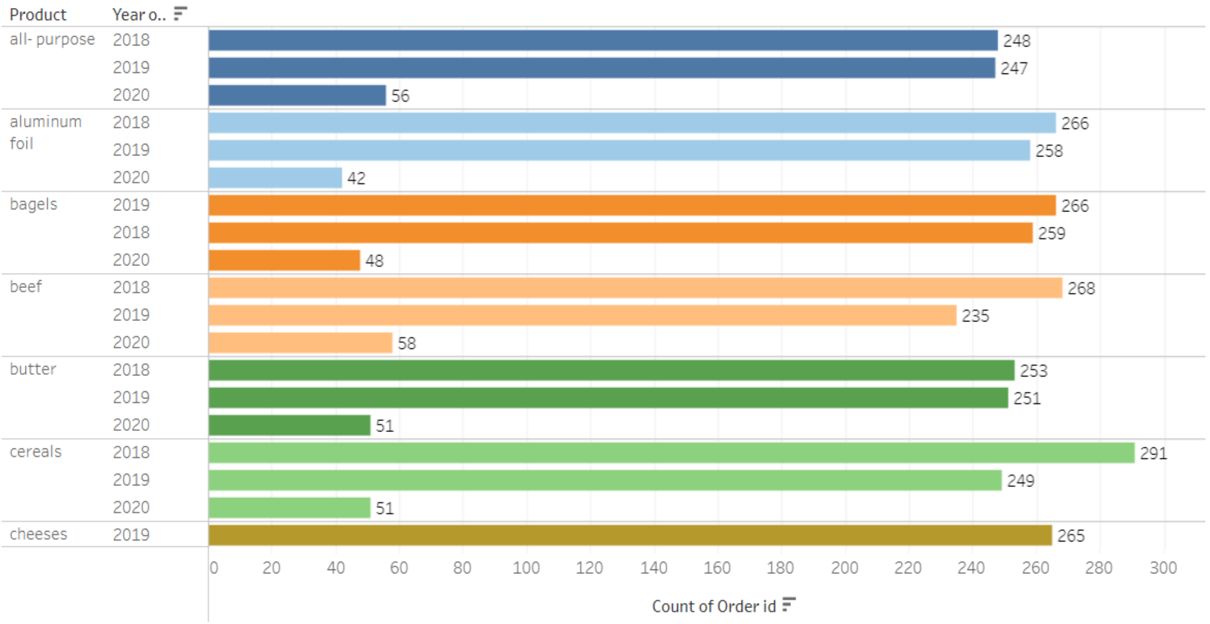
It could mean a trend of high orders during Q3 - Q4 – Q1 of every year – but then the data is limited so this hypothesis could not be proved.

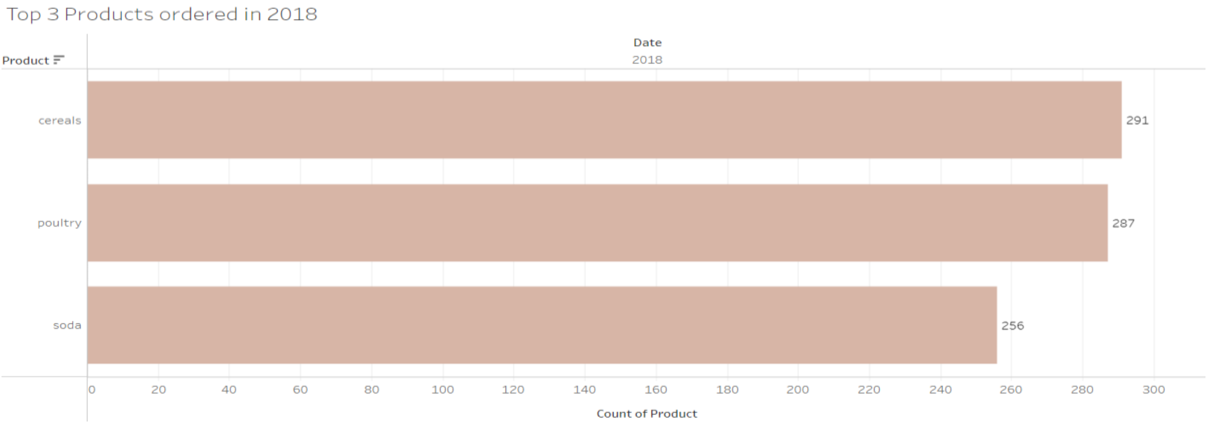
Oders over the Month



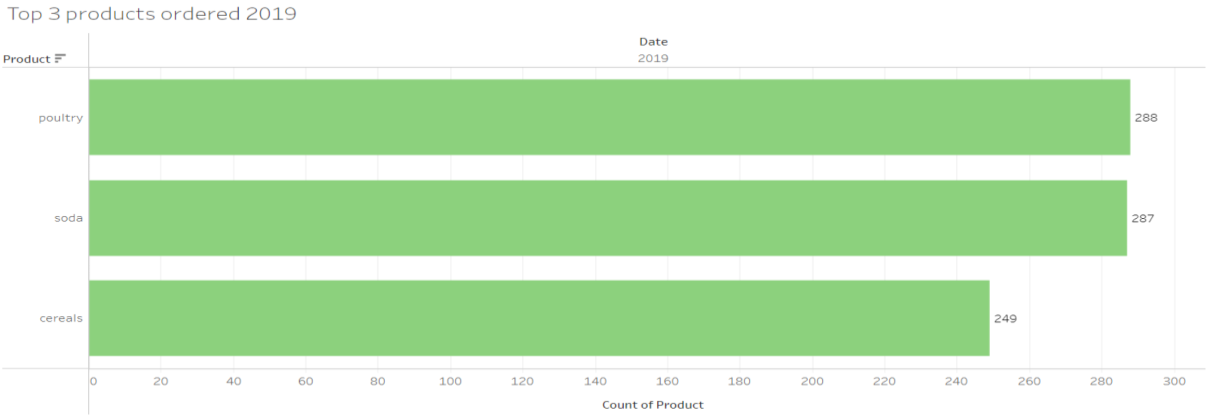
The orders trend for the given data over the months, however indicate Jan and Feb being months where high orders are placed.

Orders products over the Years

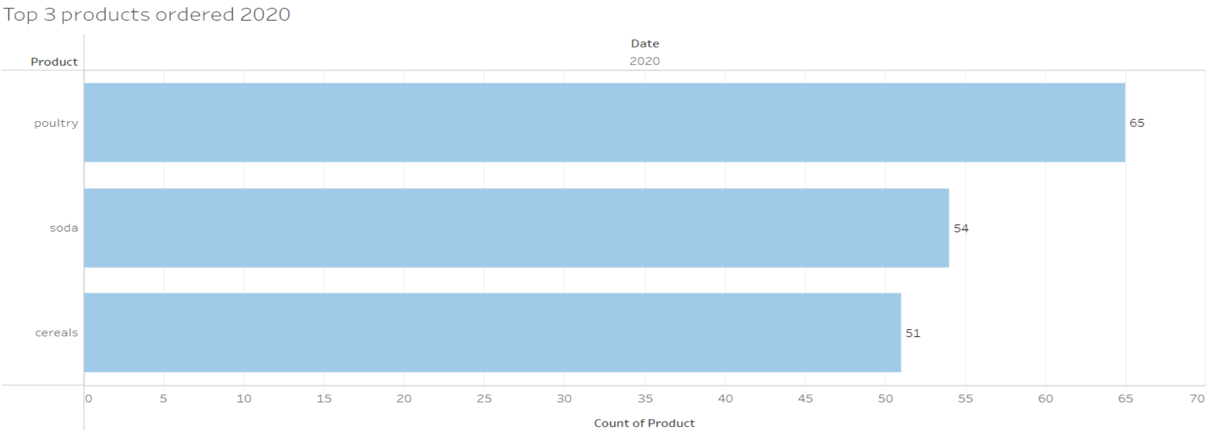




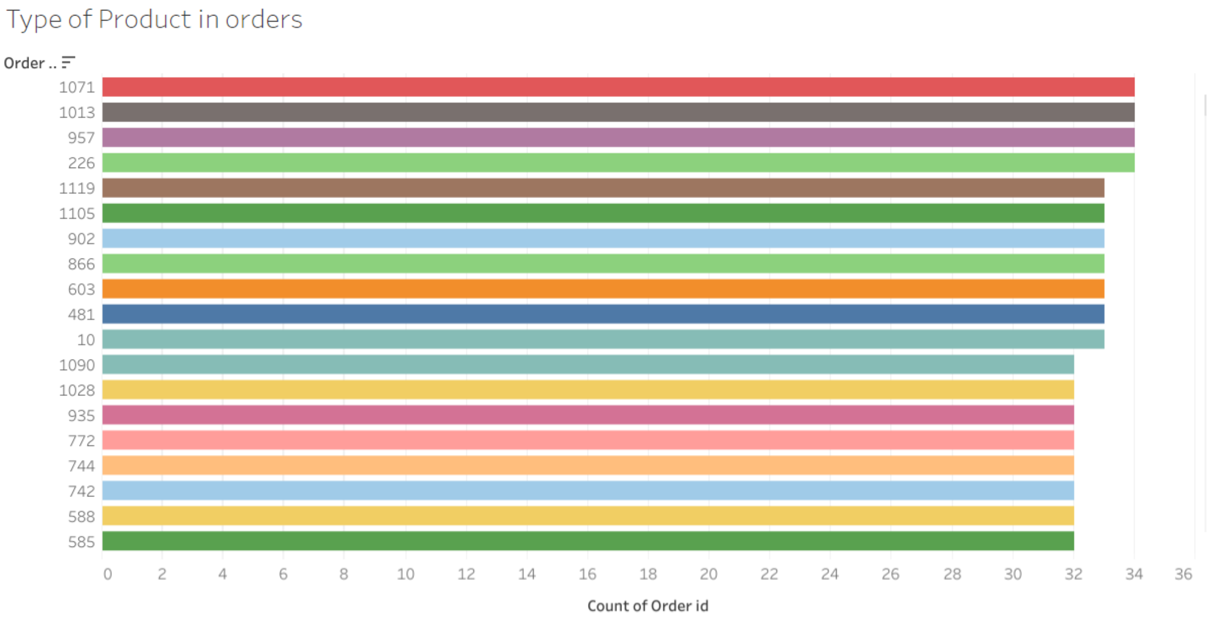
Top three products ordered in 2018 are cereals, poultry and soda with the mentioned no. of product orders.



Top three products ordered in 2019 are poultry, soda and cereals with the mentioned no. of product orders.



Top three products ordered in 2020 are poultry, soda and cereals with the mentioned no. of product orders.



Comprehensive view of the ordered products over the year.

PART B: Use of Market Basket Analysis (Association Rules) -->Write Something about the association rules and its relevance in this case -->Add KNIME workflow image -->Write about threshold values of Support and Confidence.

Market basket analysis (MBA) is a data mining technique used by retailers to uncover hidden relationships between products that customers frequently purchase together.

Understanding customer behaviour: Its primary goal is to understand customer purchasing patterns to make better decisions about:

Product Placement: Stores can place frequently purchased items together.

Promotions: Design targeted offers and cross-selling strategies.

Inventory Management: Forecast demand and optimize inventory levels.

Customer Recommendations: Build recommendation systems ("Customers who bought this also bought...")

How Market Basket Analysis Works:

Data Collection: MBA starts with collecting large amounts of customer transaction data. This data shows which items have been purchased together in each transaction.

Association Rule Mining: Algorithms analyse the data and identify frequent item sets (groups of products that appear together often). From these item sets, association rules are generated with the form "IF item X is purchased, THEN item Y is also likely to be purchased."

Benefits of Market Basket Analysis:

Improved customer understanding: Provides deep insights into how customers shop and what they like to buy in combination.

Increased sales: Helps businesses boost sales through cross-selling and up-selling.

Optimized marketing: Targeted promotions and recommendations become more effective.

Better inventory management: Prevents stock outs of popular items and avoids overstocking slow-moving items.

Association rules are a powerful technique used to uncover hidden relationships between items purchased together by customers. They are often expressed in the form of "IF item X is purchased, THEN item Y is also likely to be purchased."

These rules help businesses understand customer behaviour and make data-driven decisions about product placement, promotions, and recommendations.

Knime workflow for market basket analysis look:

1. Data Preparation:

Load your transaction data.

Clean and pre-process the data (e.g., handle missing values, convert data into the appropriate format).

2. Association Rule Mining:

Use the "Association Rule Learner" node in KNIME to generate association rules. You'll need to experiment with minimum support and confidence settings (explained below).

3. Rule Filtering and Visualization:

Filter the rules based on the desired support, confidence, and lift thresholds.

Visualize the rules using nodes like "Association Rules Viewer" or "Network Viewer"

Threshold Values: Support, Confidence, and Lift:

Support: A measure of how frequently a particular set of items occurs together in transactions. Higher support indicates a more common pattern.

Confidence: A measure of how likely an item Y is purchased when item X is purchased. This indicates the strength of the association rule.

Lift: Measures how much more likely item Y is purchased when item X is purchased, compared to how likely item Y would be purchased anyway. Lift values greater than 1 indicate a positive correlation between items.

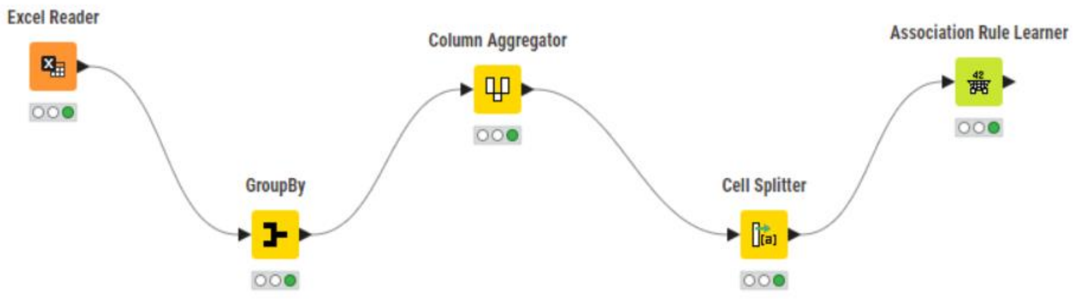
Important Considerations:

1 The choice of support and confidence thresholds is key. Lower thresholds will result in more rules but may include less significant patterns.

2 Higher thresholds ensure stronger rules but may miss out on some valuable insights.

3 Consider exploring the 'lift' metric for more robust rule selection.

PART B: Associations Identified --> Put the associations in a tabular manner --> Explain about support, confidence, & lift values that are calculated.



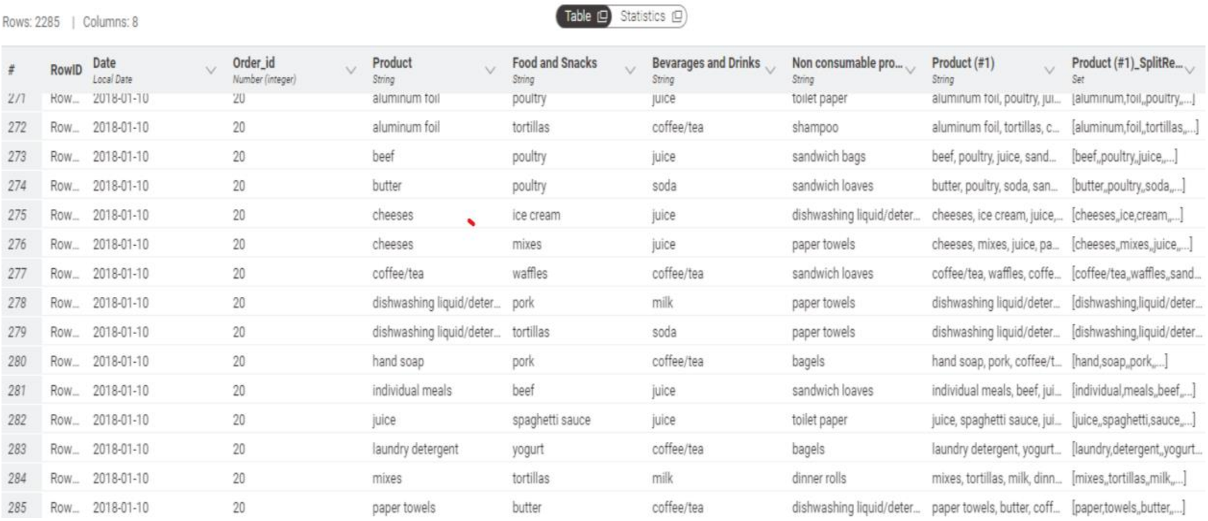
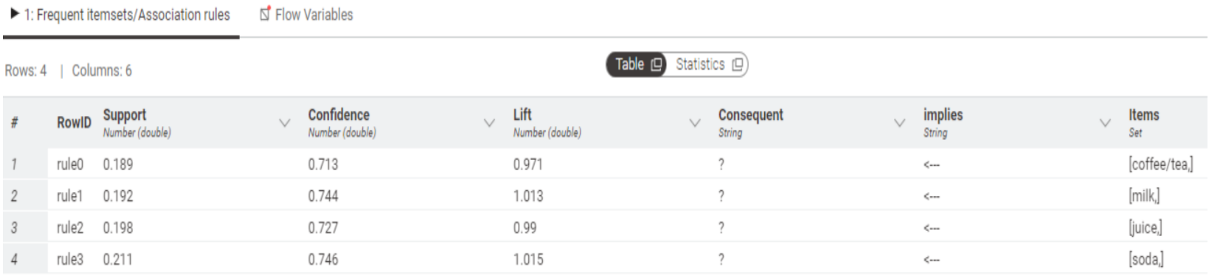

Rows: 2285 | Columns: 8

Table Statistics

#	RowID	Date Local Date	Order_id Number (integer)	Product String	Food and Snacks String	Bevarages and Drinks String	Non consumable pro... String	Product (#1) String	Product (#1)_SplitRe... Set
271	Row...	2018-01-10	20	aluminum foil	poultry	juice	toilet paper	aluminum foil, poultry, jui...	[aluminum,foil,,poultry,,...]
272	Row...	2018-01-10	20	aluminum foil	tortillas	coffee/tea	shampoo	aluminum foil, tortillas, c...	[aluminum,foil,,tortillas,,...]
273	Row...	2018-01-10	20	beef	poultry	juice	sandwich bags	beef, poultry, juice, sand...	[beef,,poultry,,juice,,...]
274	Row...	2018-01-10	20	butter	poultry	soda	sandwich loaves	butter, poultry, soda, san...	[butter,,poultry,,soda,,...]
275	Row...	2018-01-10	20	cheeses	ice cream	juice	dishwashing liquid/deter...	cheeses, ice cream, juice,...	[cheeses,,ice,cream,,...]
276	Row...	2018-01-10	20	cheeses	mixes	juice	paper towels	cheeses, mixes, juice, pa...	[cheeses,,mixes,,juice,,...]
277	Row...	2018-01-10	20	coffee/tea	waffles	coffee/tea	sandwich loaves	coffee/tea, waffles, coffe...	[coffee/tea,,waffles,,sand...
278	Row...	2018-01-10	20	dishwashing liquid/deter...	pork	milk	paper towels	dishwashing liquid/deter...	[dishwashing,liquid/deter...
279	Row...	2018-01-10	20	dishwashing liquid/deter...	tortillas	soda	paper towels	dishwashing liquid/deter...	[dishwashing,liquid/deter...
280	Row...	2018-01-10	20	hand soap	pork	coffee/tea	bagels	hand soap, pork, coffee/t...	[hand,soap,,pork,,...]
281	Row...	2018-01-10	20	individual meals	beef	juice	sandwich loaves	individual meals, beef, jui...	[individual,meals,,beef,,...]
282	Row...	2018-01-10	20	juice	spaghetti sauce	juice	toilet paper	juice, spaghetti sauce, jui...	[juice,,spaghetti,sauce,,...]
283	Row...	2018-01-10	20	laundry detergent	yogurt	coffee/tea	bagels	laundry detergent, yogurt...	[laundry,detergent,,yogurt...
284	Row...	2018-01-10	20	mixes	tortillas	milk	dinner rolls	mixes, tortillas, milk, dinn...	[mixes,,tortillas,,milk,,...]
285	Row...	2018-01-10	20	paper towels	butter	coffee/tea	dishwashing liquid/deter...	paper towels, butter, coff...	[paper,towels,,butter,,...]


▸ 1: Frequent itemsets/Association rules   Flow Variables

Rows: 4 | Columns: 6

Table Statistics

#	RowID	Support Number (double)	Confidence Number (double)	Lift Number (double)	Consequent String	implies String	Items Set
1	rule0	0.189	0.713	0.971	?	<---	[coffee/tea,]
2	rule1	0.192	0.744	1.013	?	<---	[milk,]
3	rule2	0.198	0.727	0.99	?	<---	[juice,]
4	rule3	0.211	0.746	1.015	?	<---	[soda,]

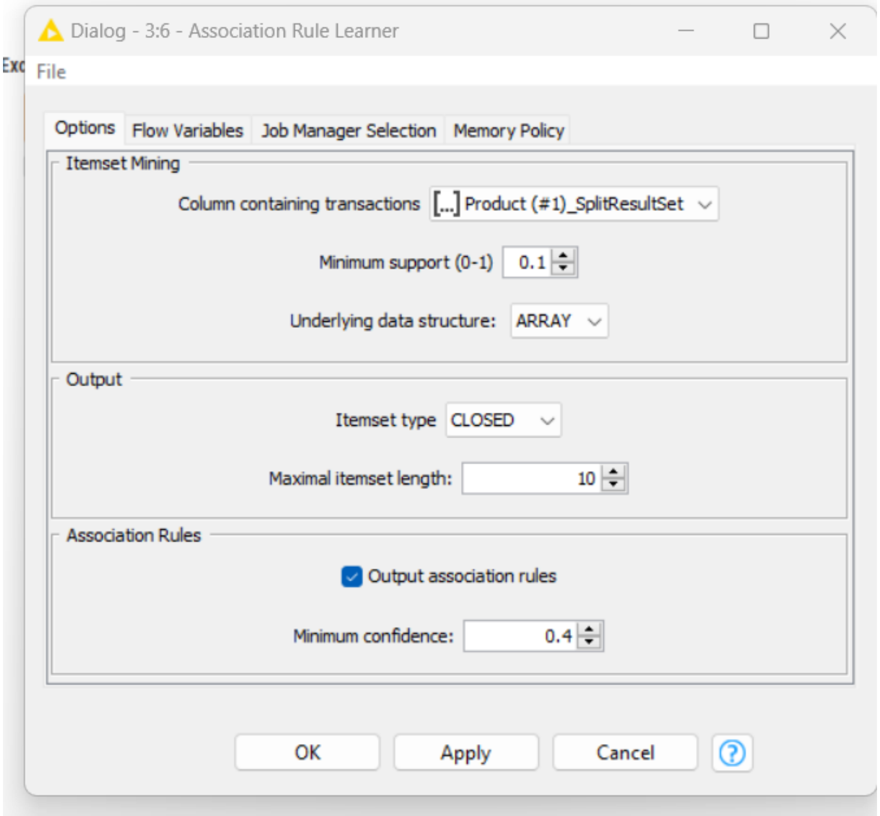
MRA Association rules Parameters:

Threshold values are found out by various regressions and shown here:

Support of minimum: 0.01

Maximum item set length: 10

Minimum confidence level: 0.04



Itemset Mining

Column containing transactions [...] Product (#1)_SplitResultSet

Minimum support (0-1) 0.1

Underlying data structure: ARRAY

Output

Itemset type CLOSED

Maximal itemset length: 10

Association Rules

☑ Output association rules

Minimum confidence: 0.4

PART B: Suggestion of Possible Combos with Lucrative Offers --> Write recommendations --> Make discount offers or combos (or buy two get one free) based on the associations and your experience.

MRA suggestions and Recommendations:

- Top Combos with good confidence
- Soda could be another item which can be offered in a combo.

MRA Suggestions and Recommendations:

- We can have easy hit-up counter of the top combinations near sales counter or billing counter to increase the sale of the offering combinations.
- Since Poultry and Soda are the most sold items and Hand soap and Sandwich loaves are the least sold items, a combo offer of these would eventually increase a sale of Hand soap and Sandwich loaves as well.
- We can have frequent sale offer on the least sold products to increase its sales.
- We can offer special discount coupon on the least sold products purchased, on the next shopping on all the product to increase the sale of the least product and increase the frequency of the customers.

Combo Suggestions:

- "Meal Deal": Combine Poultry (a high-selling item) with a side like Sandwich Loaves (low-selling), and offer a small discount on the bundle. This encourages customers to try the lesser-known item.
- "Cleaning Combo": Pair Soda (a popular item) with Hand Soap (low-selling) as a cleaning supply bundle, offering it at a slightly lower price than buying them individually.
- "Themed Bundles": If you find other associations through further analysis, bundle products based on themes. For example, a "Game Night" bundle with snacks and beverages, or a "Picnic Pack" with relevant items.

Lucrative Offers:

- Buy One, Get One at a Discount: On Hand Soap (low-selling), offer "Buy one, get the second at 50% off." This clears stock while still making a profit.
- Free Samples: Offer a small free sample of a low-selling item (like Sandwich Loaves) with the purchase of a popular item (Poultry or Soda). This can introduce customers to new products.
- Loyalty Programs: Reward customers who purchase low-selling products with bonus points or exclusive discounts, encouraging future purchases.

Additional Considerations

- Placement: Place combo items near each other in the store for higher visibility.
- Clear Signage: Promote combo offers and discounts with clear signs in-store and on your website or marketing materials.
- Limited-Time Offers: Create a sense of urgency to drive customers to take advantage of special offers.

Before launching these offers, make sure to calculate your profit margins to ensure they remain financially viable for your business.