Marketing & Retail Analytics Project

- Prapthi Pandian

Table of Contents

Table of Contents

PART A

1.1 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. As their consultant, use 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.

1.2 ABOUT DATA :

Shape of the dataset:

```
(2747, 20)
```

Duplicate values-

```
Number of duplicate rows = 0
```

Null 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:

```
<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
```

- There are 2747 entries in the dataset, and it appears that there are no missing values.
- There are 12 categorical columns, 7 numeric columns and 1 column of date type.

Summary of numeric variables-

	count	mean	std	min	25%	50%	75%	max
QUANTITYORDERED	2747.0	35.103021	9.762135	6.00	27.000	35.00	43.000	97.00
PRICEEACH	2747.0	101.098951	42.042548	26.88	68.745	95.55	127.100	252.87
SALES	2747.0	3553.047583	1838.953901	482.13	2204.350	3184.80	4503.095	14082.80
DAYS_SINCE_LASTORDER	2747.0	1757.085912	819.280576	42.00	1077.000	1761.00	2436.500	3562.00
MSRP	2747.0	100.691664	40.114802	33.00	68.000	99.00	124.000	214.00

- The mean quantity ordered per transaction is approximately 35 items.
- The mean price of each item sold is approximately $101.10
- The mean sales amount per transaction is approximately $3553.05
- On average, customers place orders roughly every 1757 days
- The mean MSRP is approximately $100.69

Summary of categorical variables-

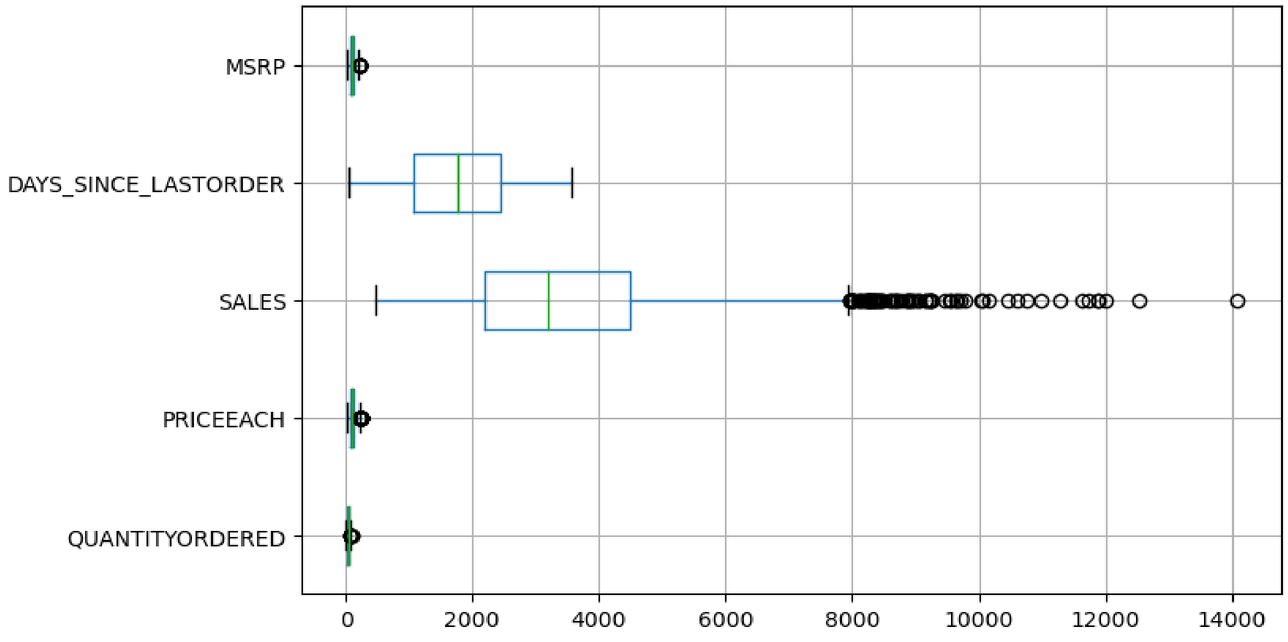
	count	unique	top	freq
STATUS	2747	6	Shipped	2541
PRODUCTLINE	2747	7	Classic Cars	949
COUNTRY	2747	19	USA	928
DEALSIZE	2747	3	Medium	1349

- The dataset contains 2747 records with 6 unique status categories, most of which is in "Shipped“ status with a frequency of 2541.
- There are 7 unique product lines with the most common being the "Classic Cars" with 949 occurrences.
- There exists transactions from 19 different countries. USA being the top with 928 transactions.
- There are 3 deal size categories. “Medium” being the most common deal size with 1349 occurrences.

```
Skewness of variables:
QUANTITYORDERED          0.369286
PRICEEACH                0.697222
SALES                    1.155940
DAYS_SINCE_LASTORDER    -0.002983
MSRP                     0.575646
dtype: float64
```

- For QUANTITYORDERED, PRICEEACH, SALES and MSRP, the positive skewness suggests that there may be some outliers with higher values in these variables.
- DAYS_SINCE_LASTORDER being close to zero indicates that the distribution of days since the last order is nearly symmetrical.

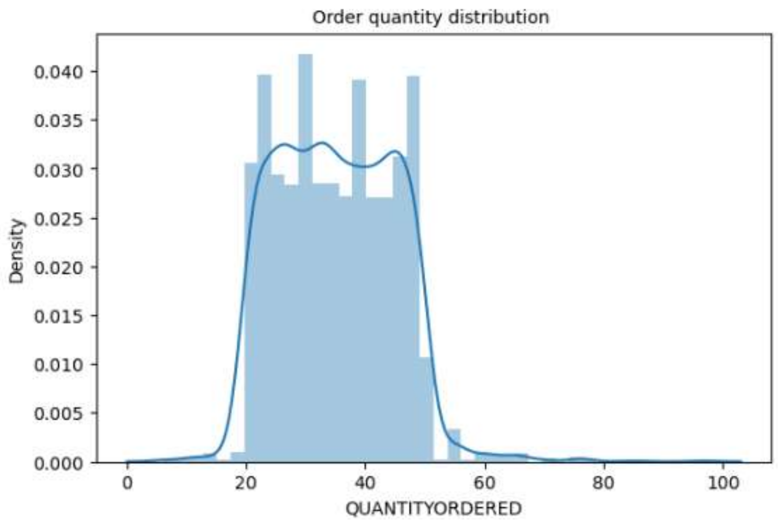
Outliers

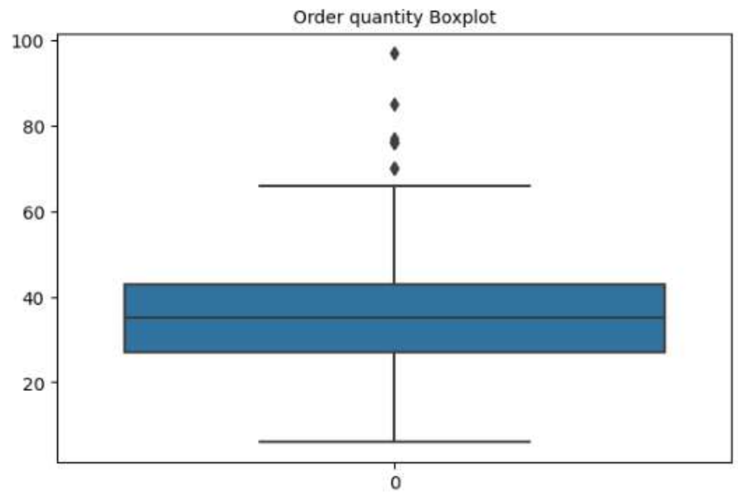


We can notice outliers in the MSRP, Price Each, Quantity Ordered and a significant amount in the Sales. Since, these could be valid, and vary based on the deal size, quantity and other influential factors, we aren’t treating the outliers.

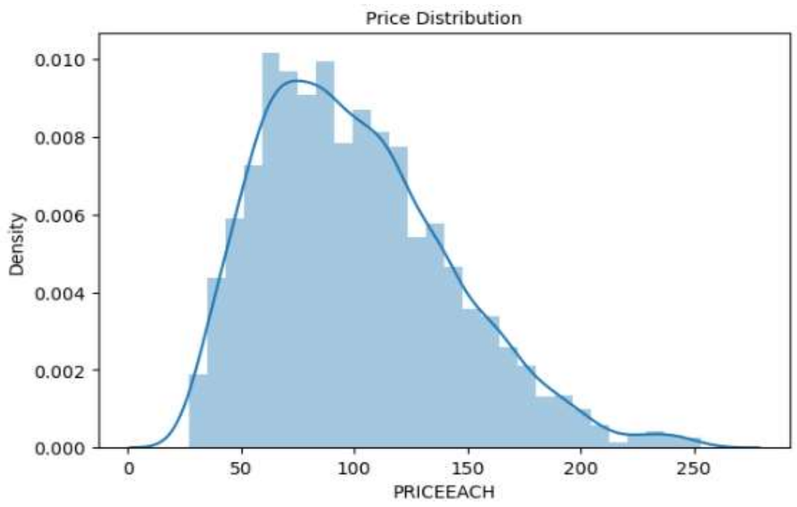
1.3 Exploratory Data Analysis:

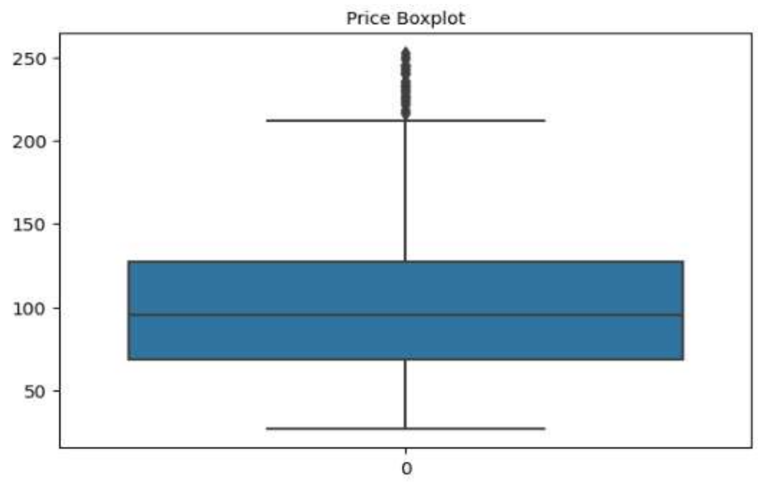
UNIVARIATE ANALYSIS

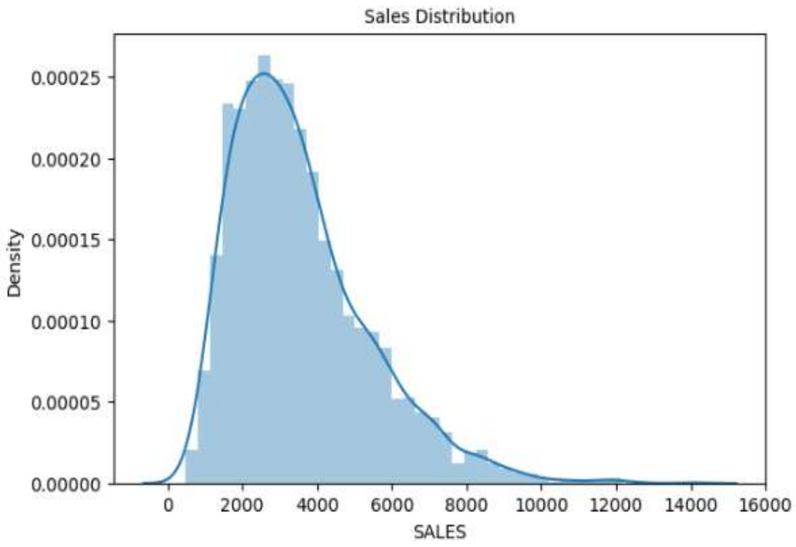


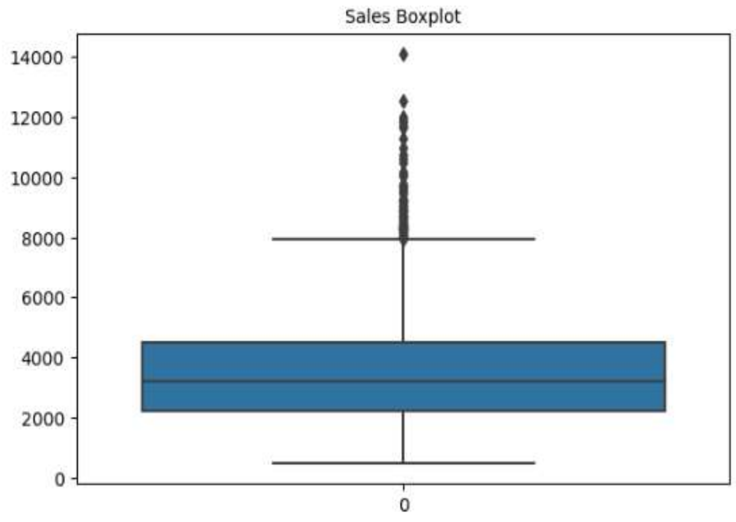


- The order quantities peak at range of 20-45.
- And we can also observe that the mostly ordered items price range from 70$ -130$ each.

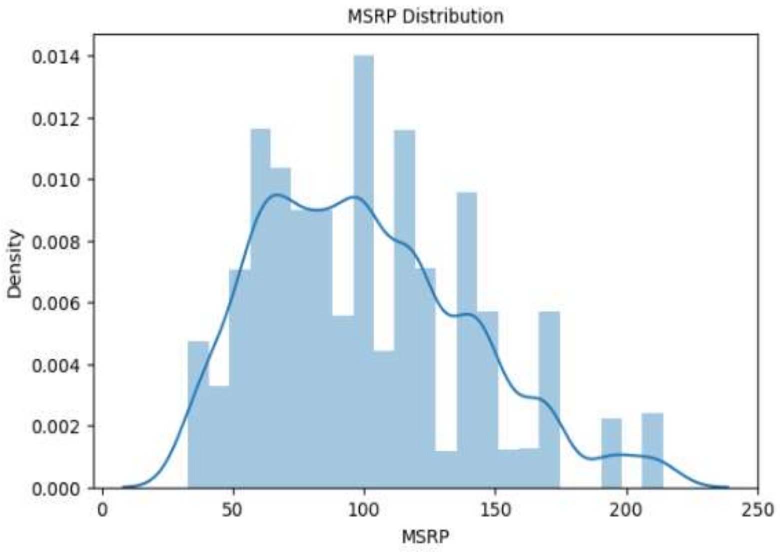


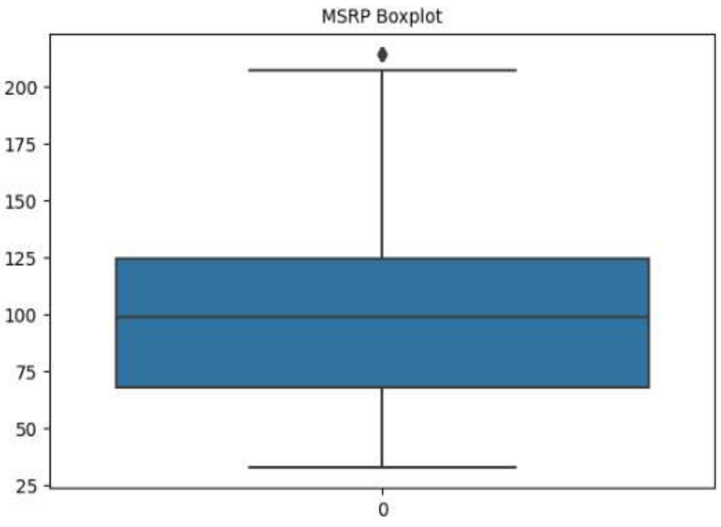




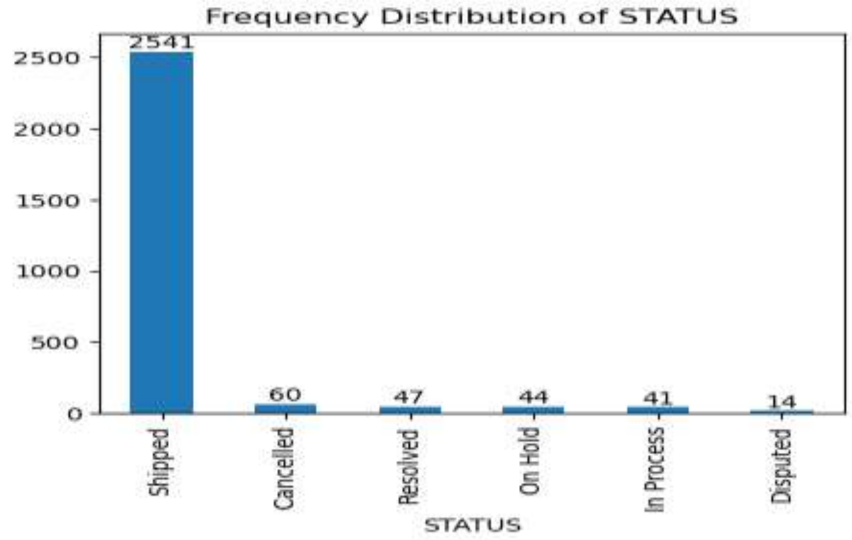


- Highest sales range from $ 2000-$4000.
- And the suggested selling price for each of these items highly vary and is right-skewed.

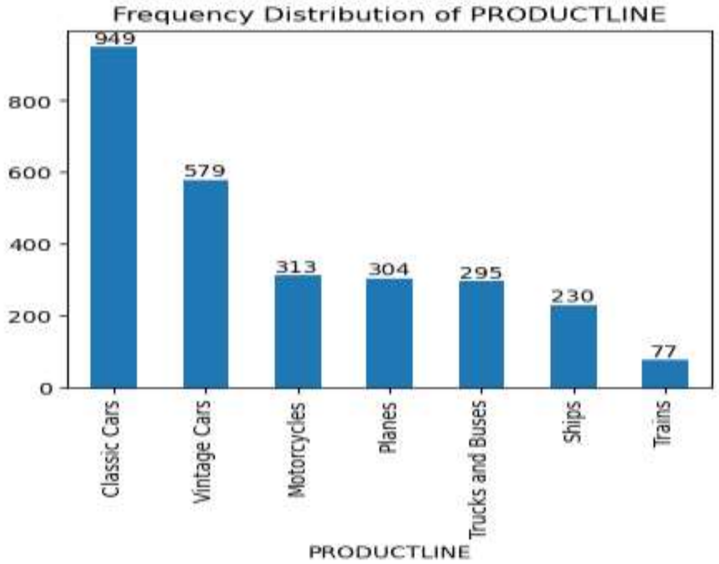




```
Details of STATUS
------------------------------------------------------------
STATUS
Shipped       2541
Cancelled       60
Resolved        47
On Hold         44
In Process      41
Disputed        14
Name: count, dtype: int64
```

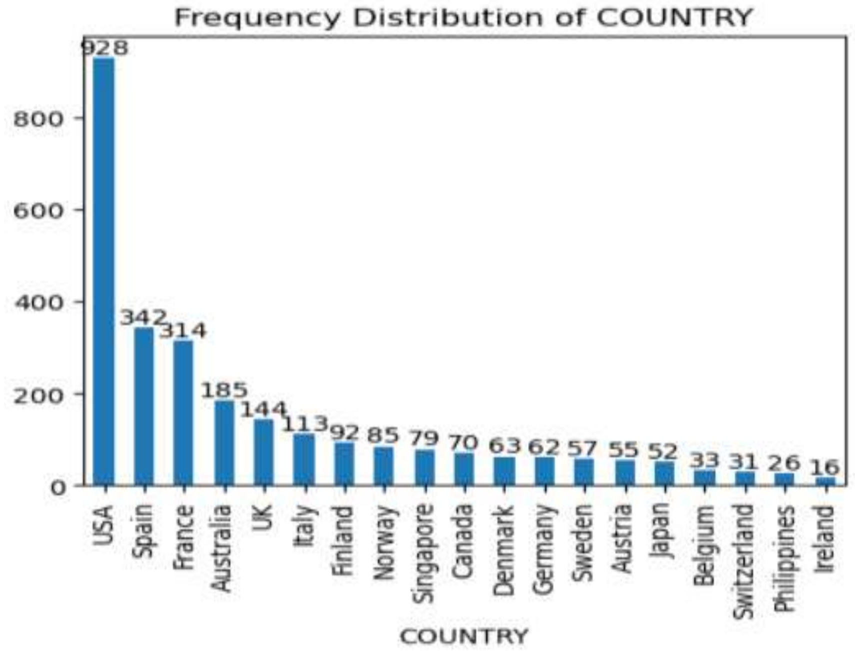



```
Details of PRODUCTLINE
------------------------------------------------------------
PRODUCTLINE
Classic Cars        949
Vintage Cars        579
Motorcycles         313
Planes              304
Trucks and Buses    295
Ships               230
Trains               77
Name: count, dtype: int64
```



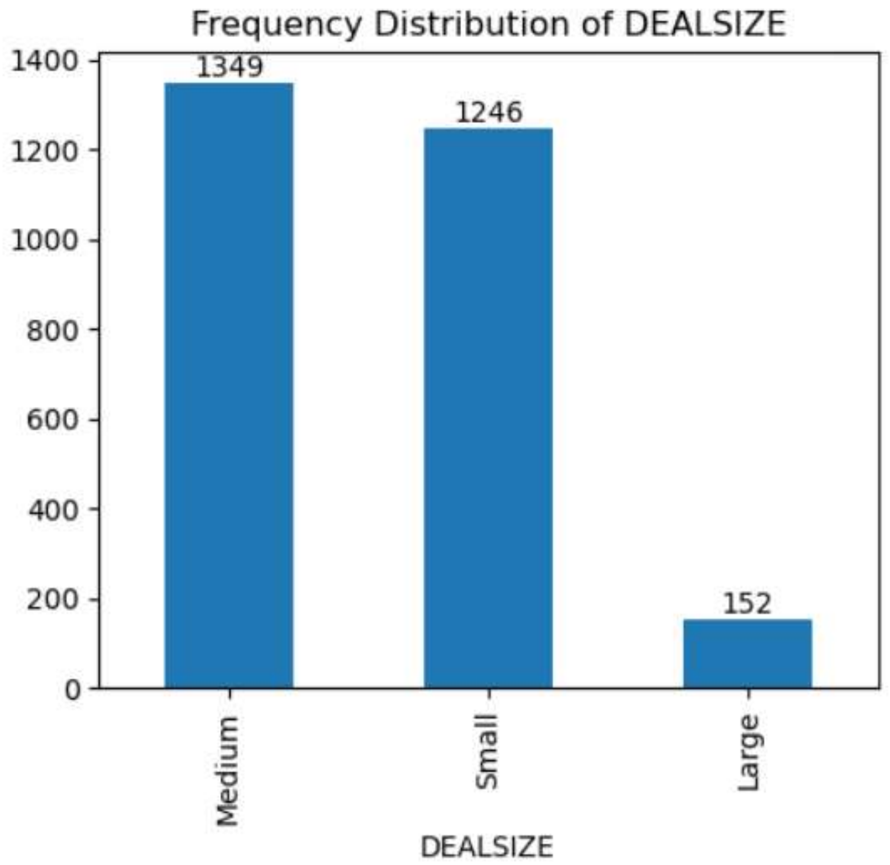
- Although there are quite a few around 60 orders being cancelled, most of the orders have been shipped successfully.
- Few orders indicate some issues where 47 of them have been resolved while 14 orders are still in disputed status.
- Around 85 orders are currently in process and are yet to be shipped.
- Classic cars has been the popular choice with 949 orders followed by vintage cars with 579 orders.
- Trains have been the least preferred with a significant lower order quantity of 77 orders.
- Motorcycles, planes, trucks and buses, ships also have reasonable demand but are less popular than classic and vintage cars.

```
Details of COUNTRY
----------------------------------------------------------------
COUNTRY
USA            928
Spain          342
France         314
Australia      185
UK             144
Italy          113
Finland         92
Norway          85
Singapore       79
Canada          70
Denmark         63
Germany         62
Sweden          57
Austria         55
Japan           52
Belgium         33
Switzerland     31
Philippines     26
Ireland         16
Name: count, dtype: int64
```



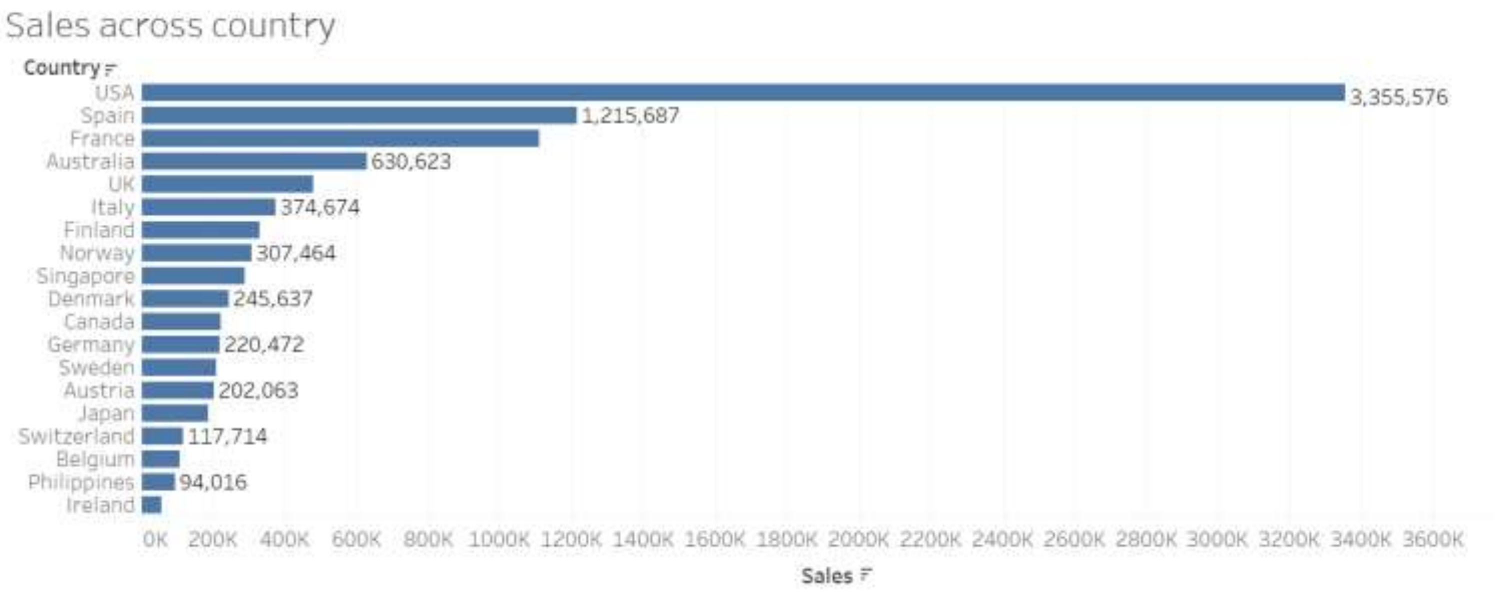
- Majority of order is from USA with 928 orders making it a crucial marketplace.
- It is followed by Spain and France with not much significant difference between them both.
- **Belgium, Switzerland, Philippines, and Ireland** have comparatively lower order counts of less than 50 suggesting lower sales activity in these regions.
- Australia, UK, Italy – although there's not much demand of orders, they do have order amounts of greater than 100.
- **Sales can be boosted in Finland, Norway, Singapore, Canada to achieve greater than 100 order counts.**

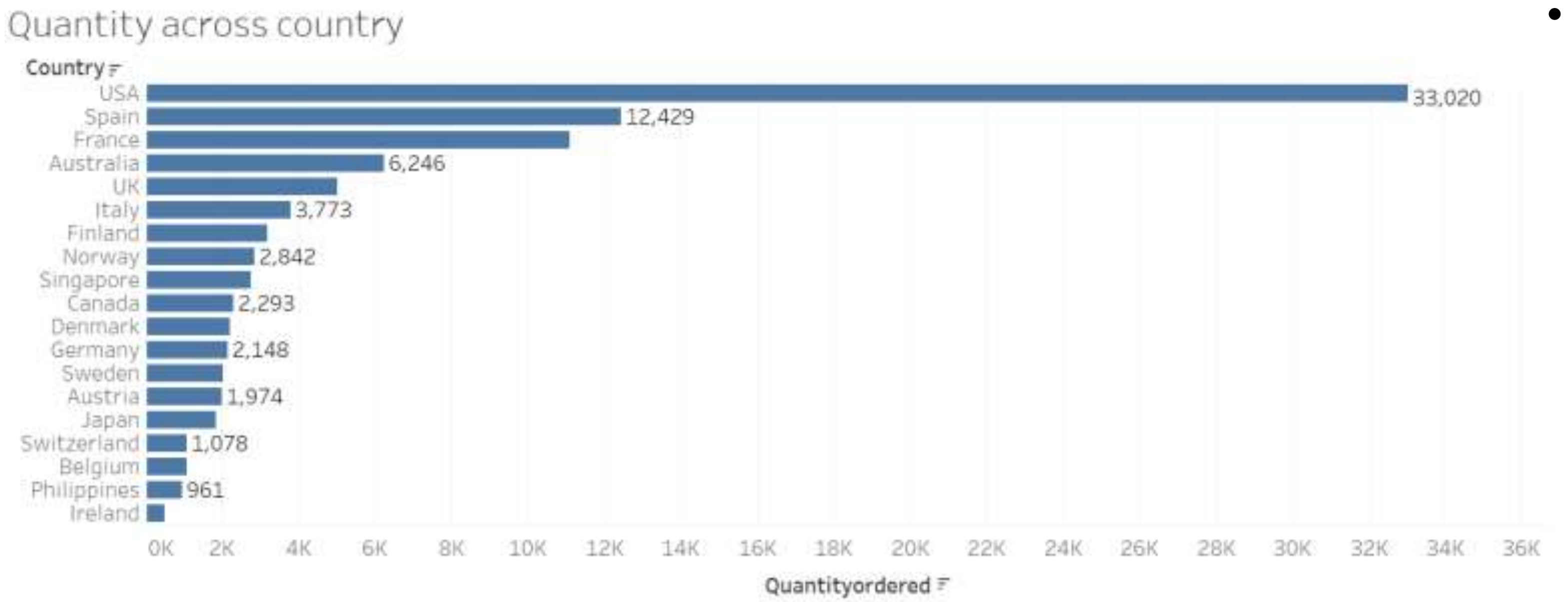
```
Details of DEALSIZE
------------------------------------------------------------
DEALSIZE
Medium    1349
Small     1246
Large      152
Name: count, dtype: int64
```



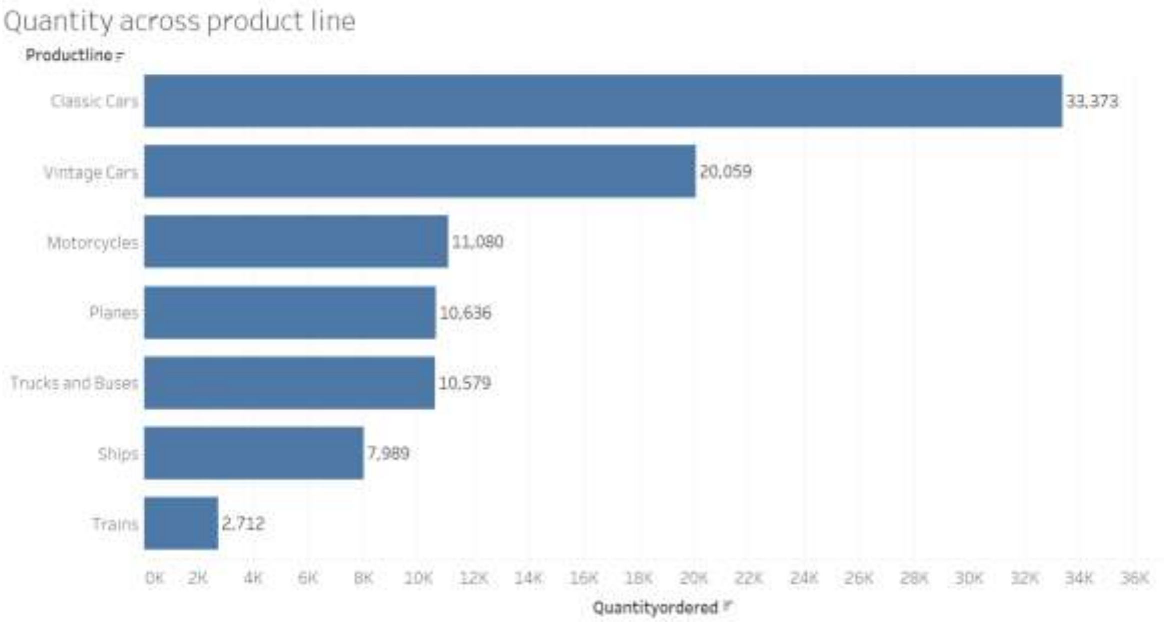
- There are three categories of deal sizes: Medium, Small, and Large.
- Majority of deals fall into the Medium and Small categories, with 1349 and 1246 deals, respectively.
- Deals categorized as Large are less common but they could potentially contribute significantly to overall revenue due to their higher value.

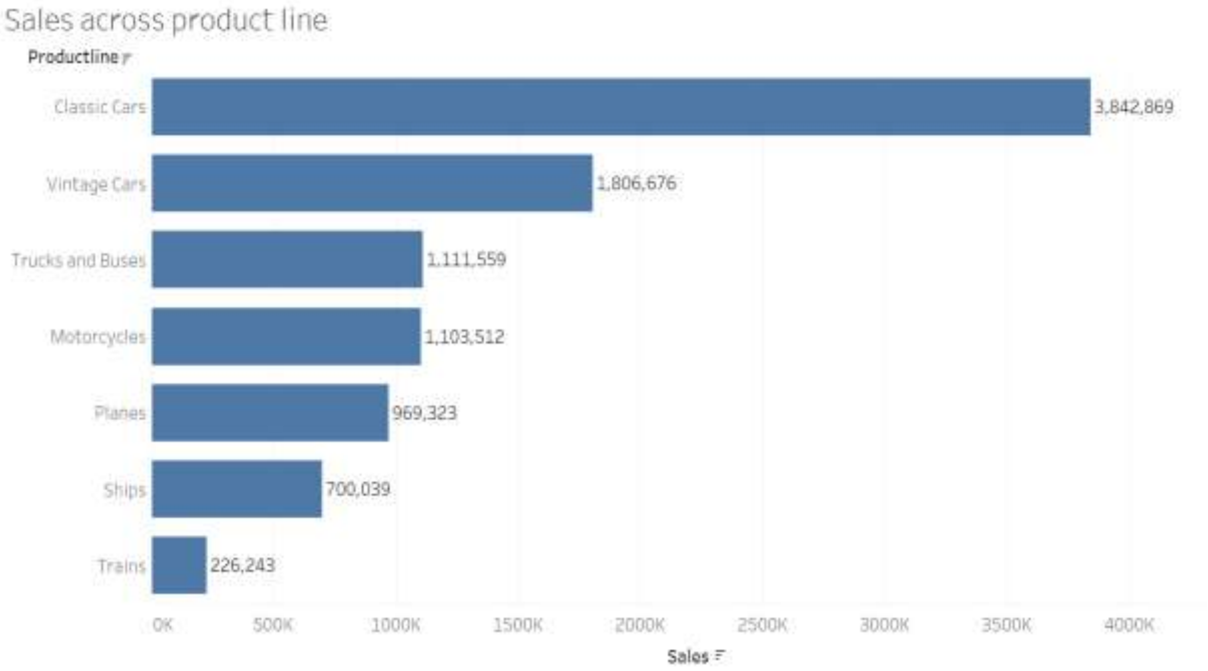
BIVARIATE ANALYSIS



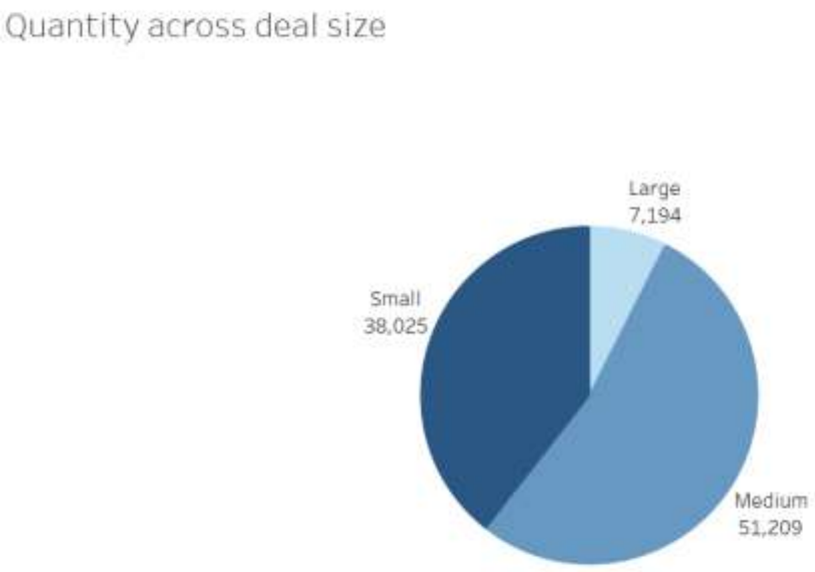


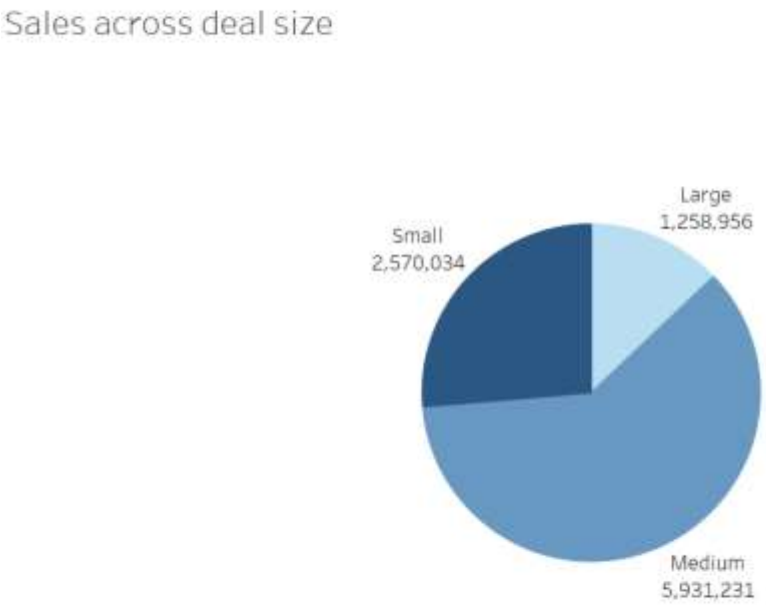
- We can notice highest sales in USA with highest number of quantity ordered making it a critical marketplace.
- It is followed by Spain and France.
- And Ireland being the least contributor towards the sales amount with least quantity of items ordered.
- We can notice the sales and quantity ordered have a relation.
- The more the quantity, the higher the sales.



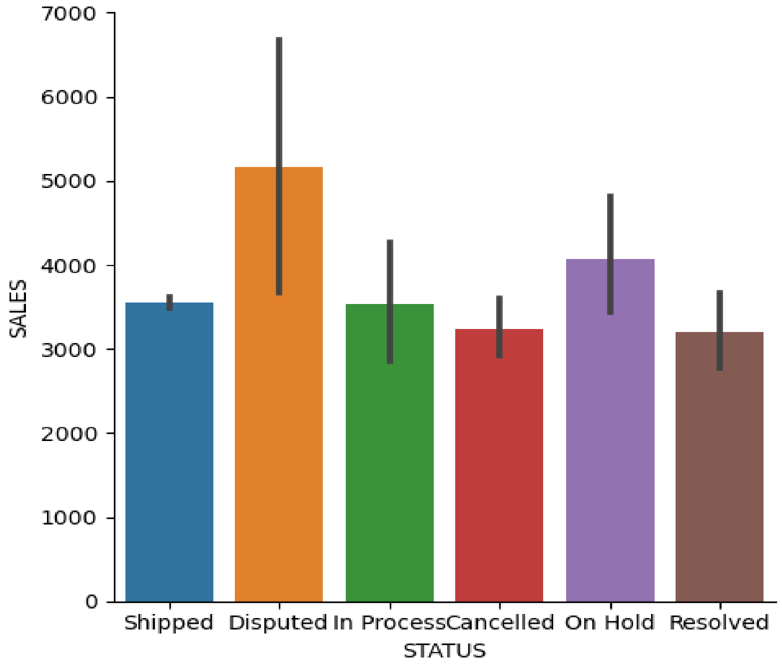


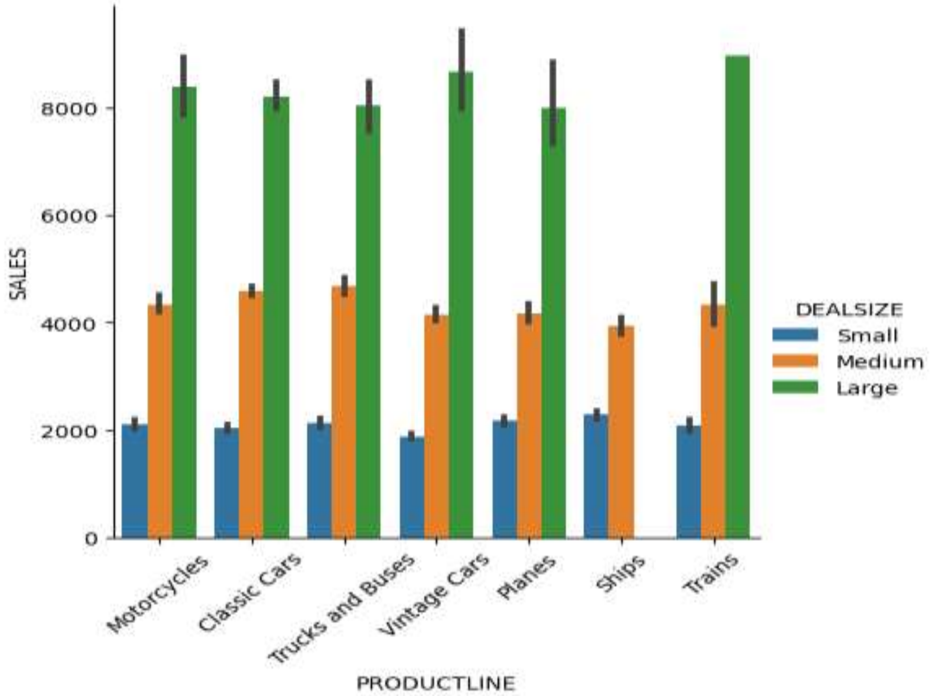
- While comparing the quantity and sales across the product line,
- Classic cars have had the highest number of orders contributing to more sale amount followed by Vintage cars and the Trains being the least preferred order with least sales amount.
- Although Motorcycles and Planes are preferred more than Trains and buses, the sales amount contribution by Trains and buses are more than them.



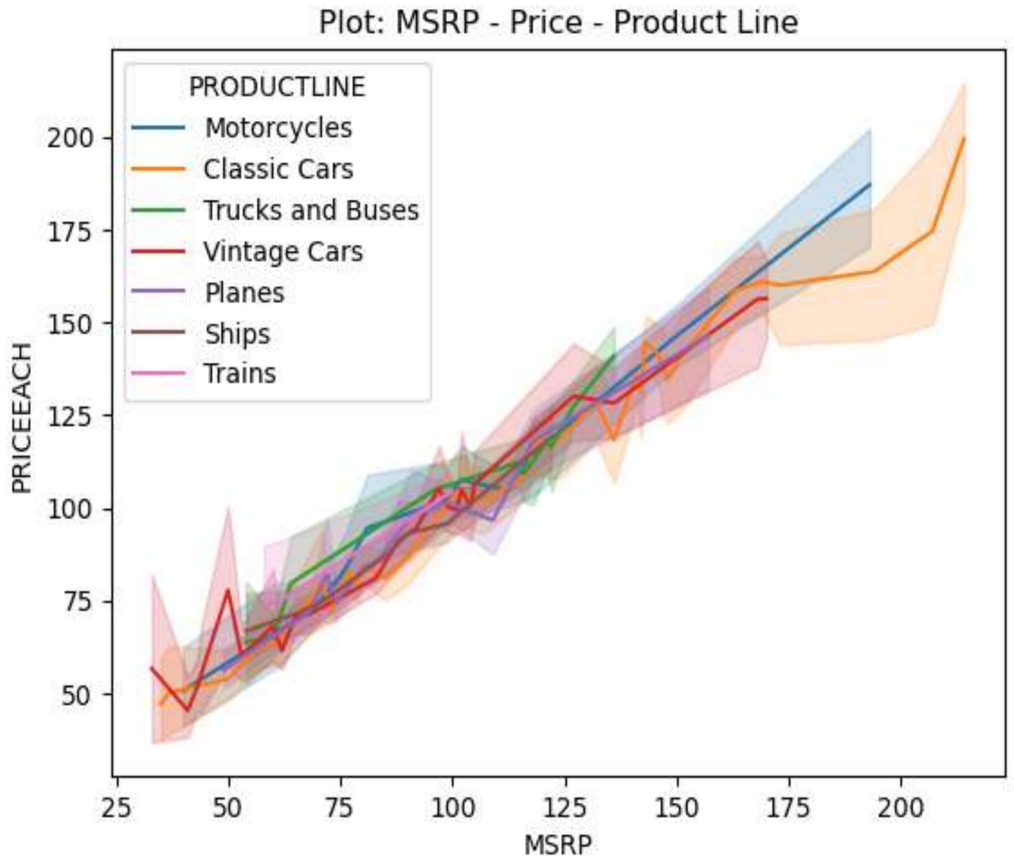


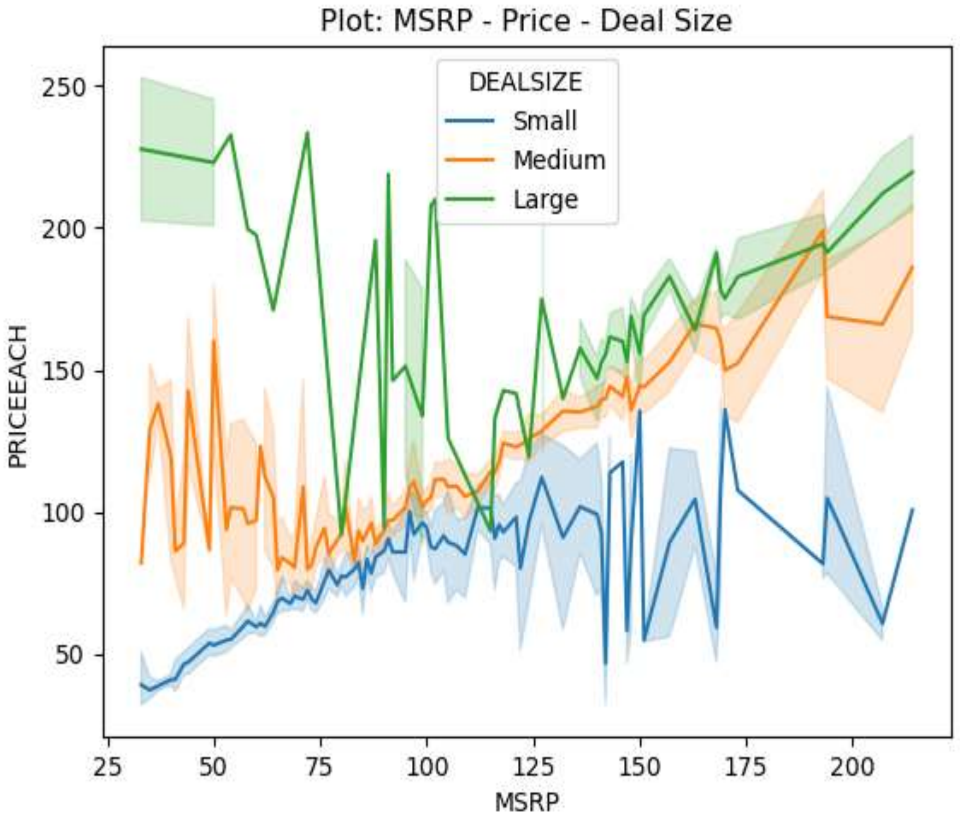
- More quantity of items are ordered in the Medium deal size of 51,209 items contributing to higher sales amount of 5,931,231 $ and the Large deal size being the least with 7,194 items ordered also contributing to good sales amount of 1,258,956 $.
- The small deal size also has a significant share in the sales amount of $ 2,570,034.



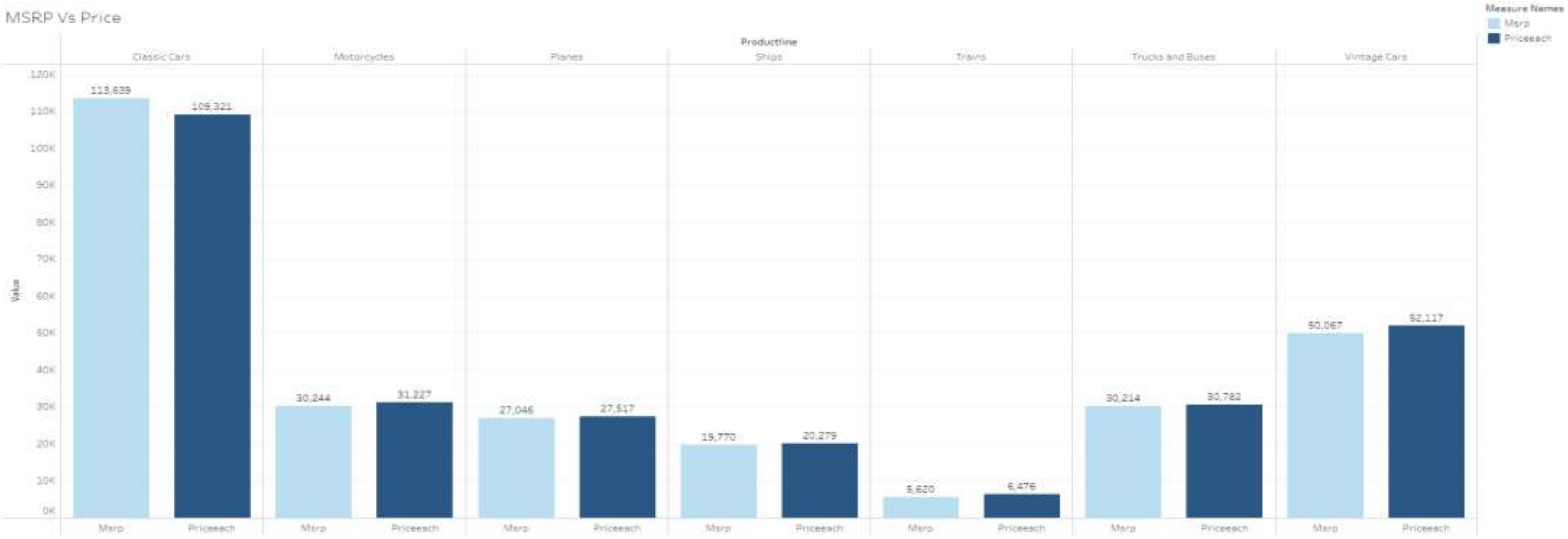


- The sales associated with disputed orders are the highest.
- The sales associated with orders on hold come next. It implies that although these orders are temporarily paused, they still contribute significantly to sales when compared to other statuses.
- Orders that are cancelled or resolved have the lowest sales amounts which is expected as cancelled orders result in no sales and resolved orders might involve adjustments or refunds affecting the overall sales amount.
- The sales amount is highest for large deal size followed by medium and small which corresponds to their value.
- Highest sales for trains amongst the large deal size, trucks and buses in the medium deal size moreover close to the classic cars.
- Although almost all product lines contribute more or less equal sales in the small deal size, trucks and buses are the least contributor and ships, the top.

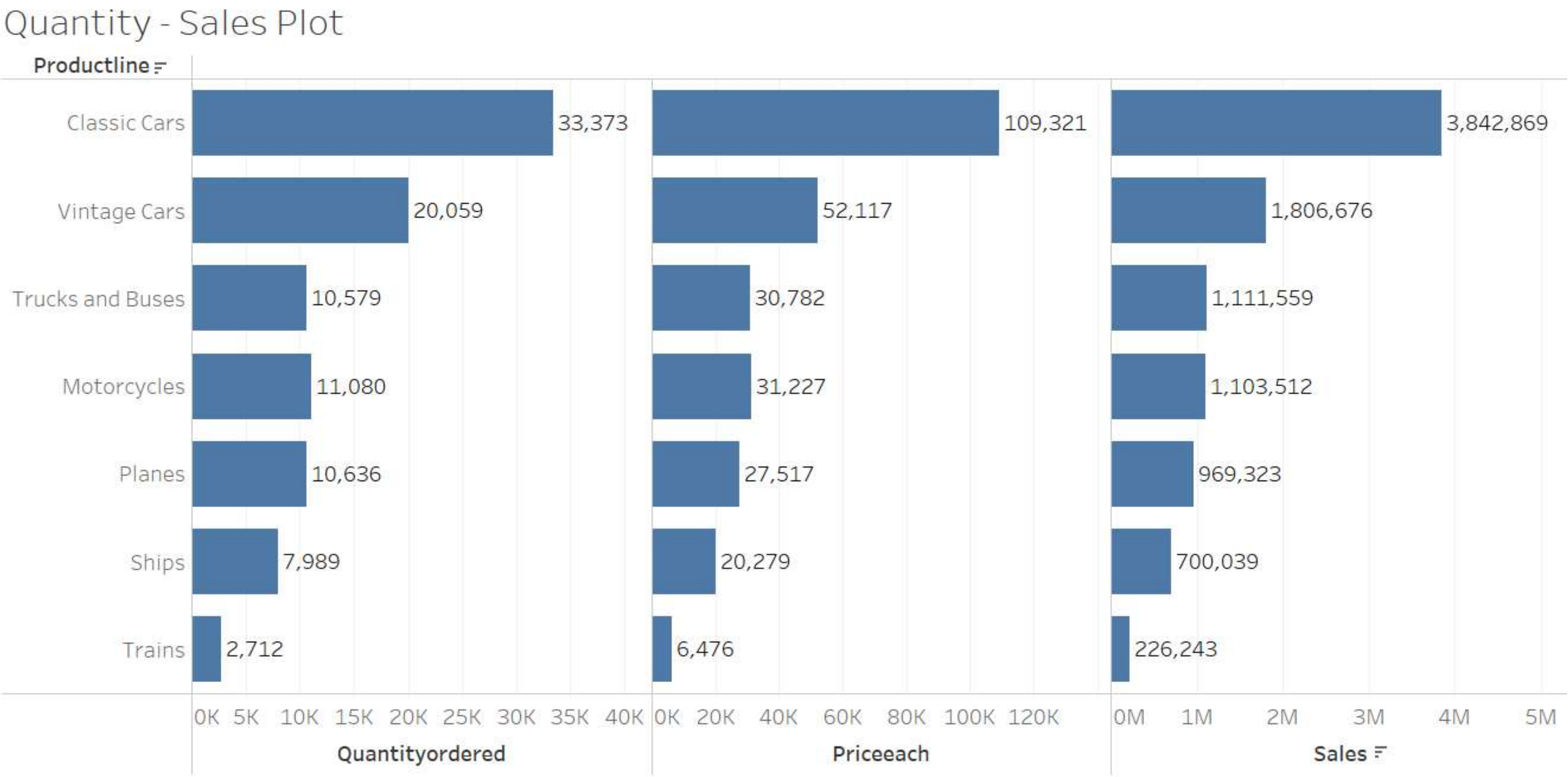




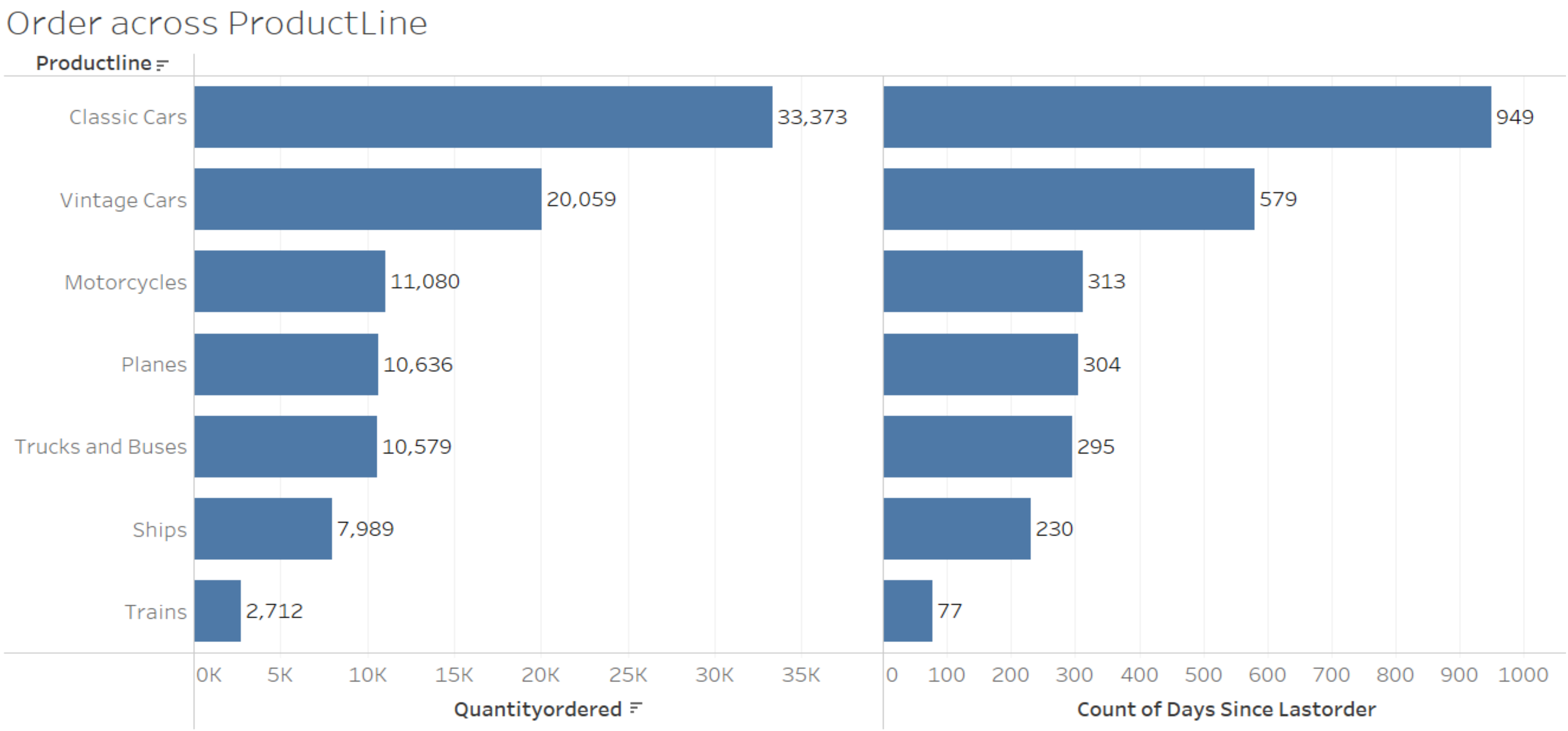
- We can notice wide range in the price for each product line.
- The price of Classic cars and Motorcycles are the highest with 180$ -200$.
- There is a correlation between the price and suggested selling price for the product lines.
- But when its is compared with the deal size, we cannot notice much correlation between them.
- There is high variance among the MSRP and PRICEEACH when considered the deal size.



- Although the Manufacturer’s suggested selling price and the Price of each item do not have a significant difference in them, the price has been higher than the suggested price for all the product lines except for Classic Cars.
- The price is the highest for Classic cars followed by Vintage cars and the least price for Trains.

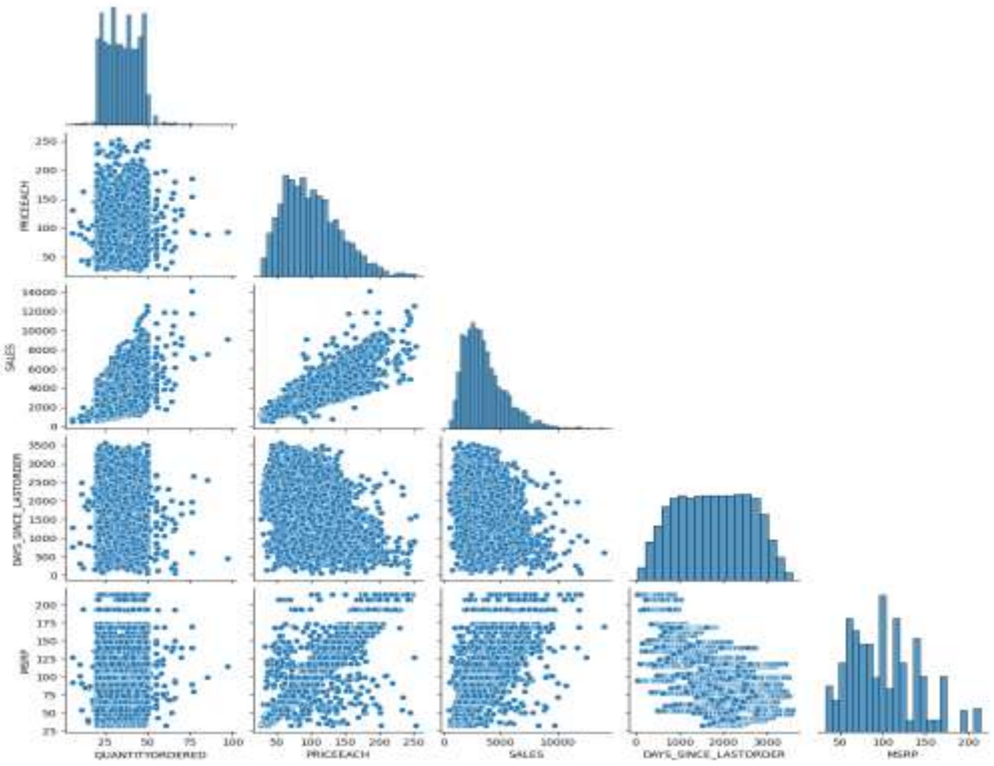


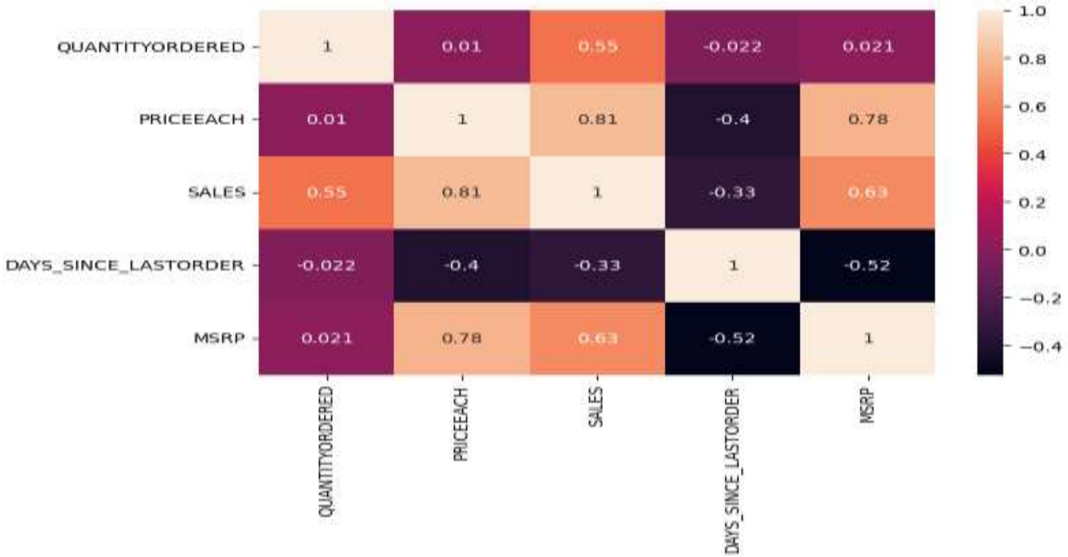
Classic cars are the largest contributor to Sales with highest number of quantity ordered and also with high price. For Trucks and Buses , although the quantity ordered is less than the Motorcycles and have a lesser price, they have contributed more towards the sales.



- Classic cars have the highest quantity ordered, followed by vintage cars, motorcycles, planes, trucks, ships, and trains.
- Classic cars are the most popular among customers, as they have the highest demand compared to other product categories.
- Classic cars also have the highest count of days since the last order. This suggests that there might be a longer interval between orders for classic cars compared to other product categories.
- It could indicate that classic cars are less frequently ordered possibly due to their higher price.

MULTIVARIATE ANALYSIS:

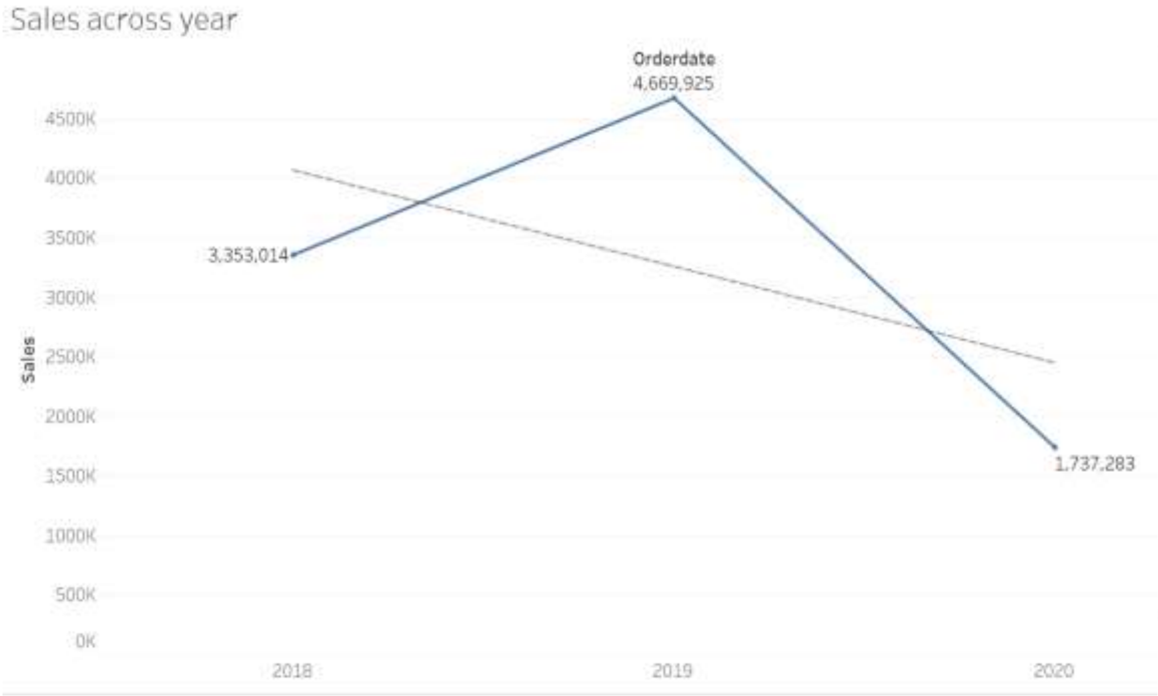


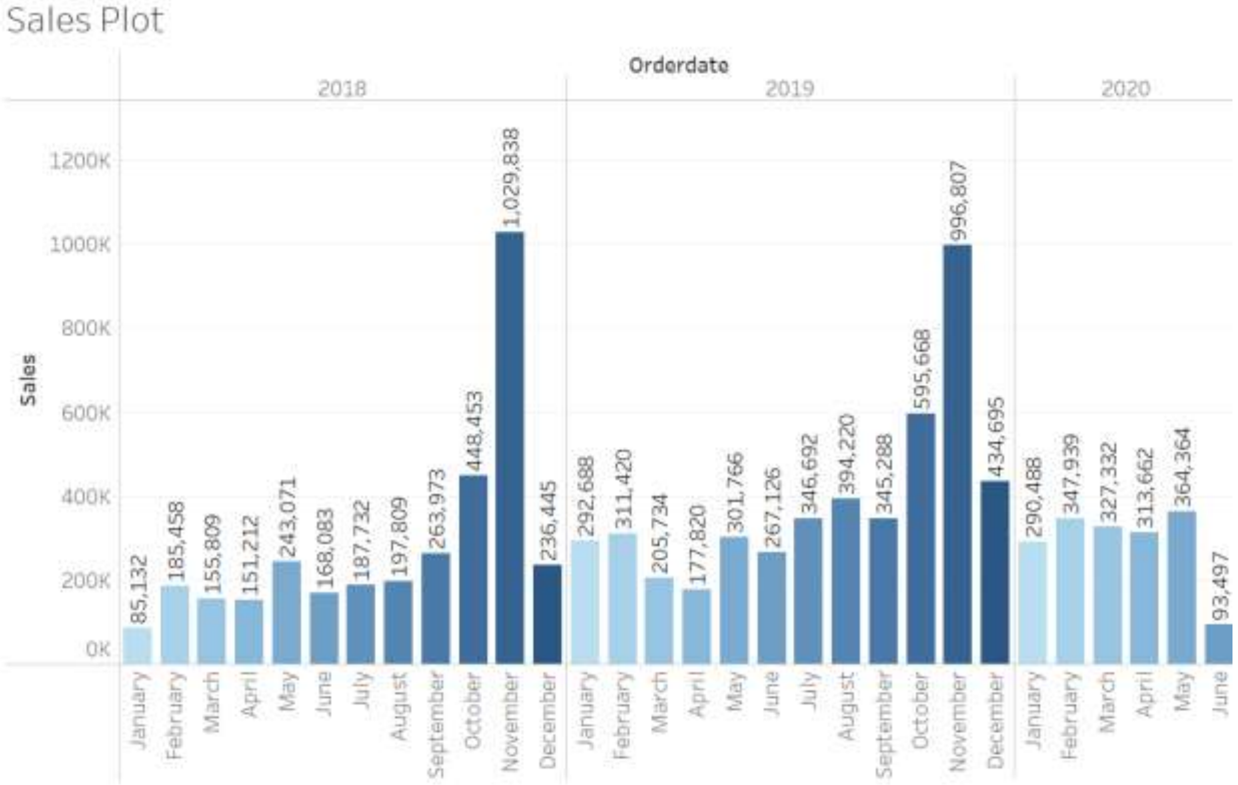


- Quantity ordered and Sales have a strong positive correlation of 0.55 indicating that as the quantity ordered increases, sales also tend to increase.
- Price Each and Sales also have a strong positive correlation of 0.81, suggesting that higher prices per unit are associated with higher sale amount.
- Price Each and MSRP have a strong positive correlation of 0.78 indicating that the price each unit is sold tends to be closely related to the manufacturer's suggested retail price.
- Days since last ordered and Sales have a moderate negative correlation of -0.33. This indicates a potential relationship between frequency of orders and sales volume.
- MSRP and Days since last ordered have a negative correlation coefficient of -0.52. It could indicate a strategy of decreasing prices over time for certain products.
- There's almost no linear relationship between Days since last ordered and Quantity Ordered suggesting that the quantity ordered doesn't significantly change based on the time elapsed since the last order.

1.4 SALES TREND

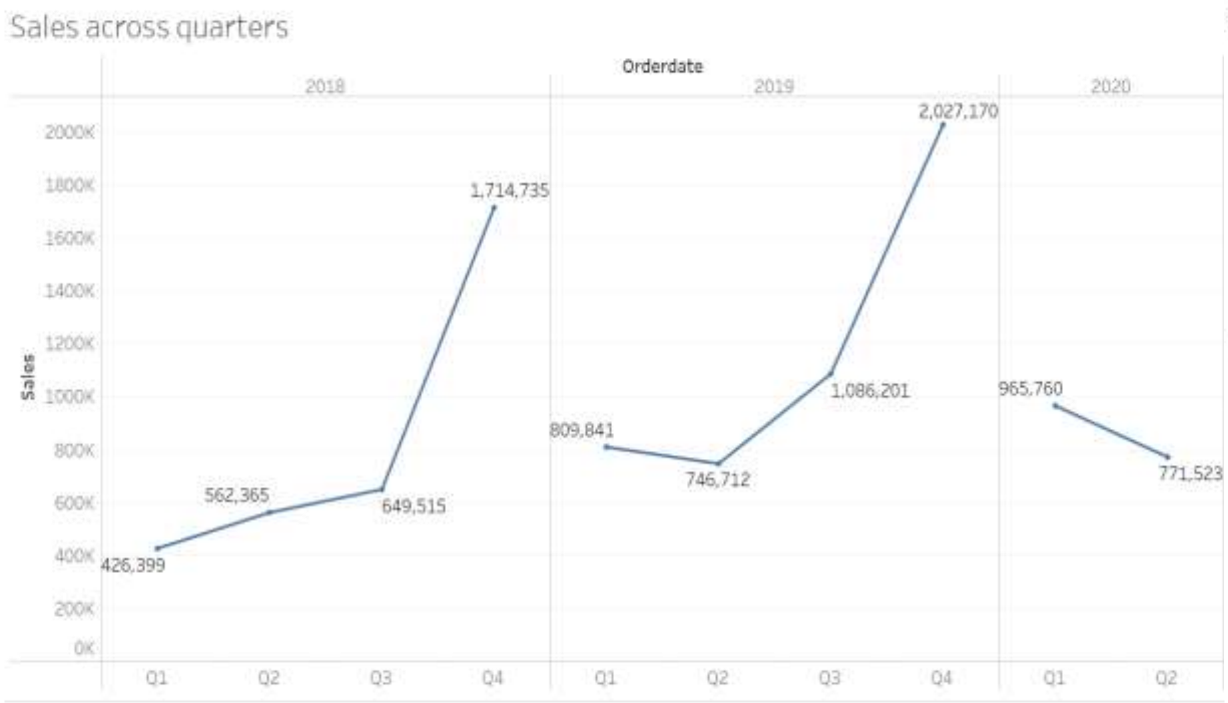
YEARLY TREND IN SALES

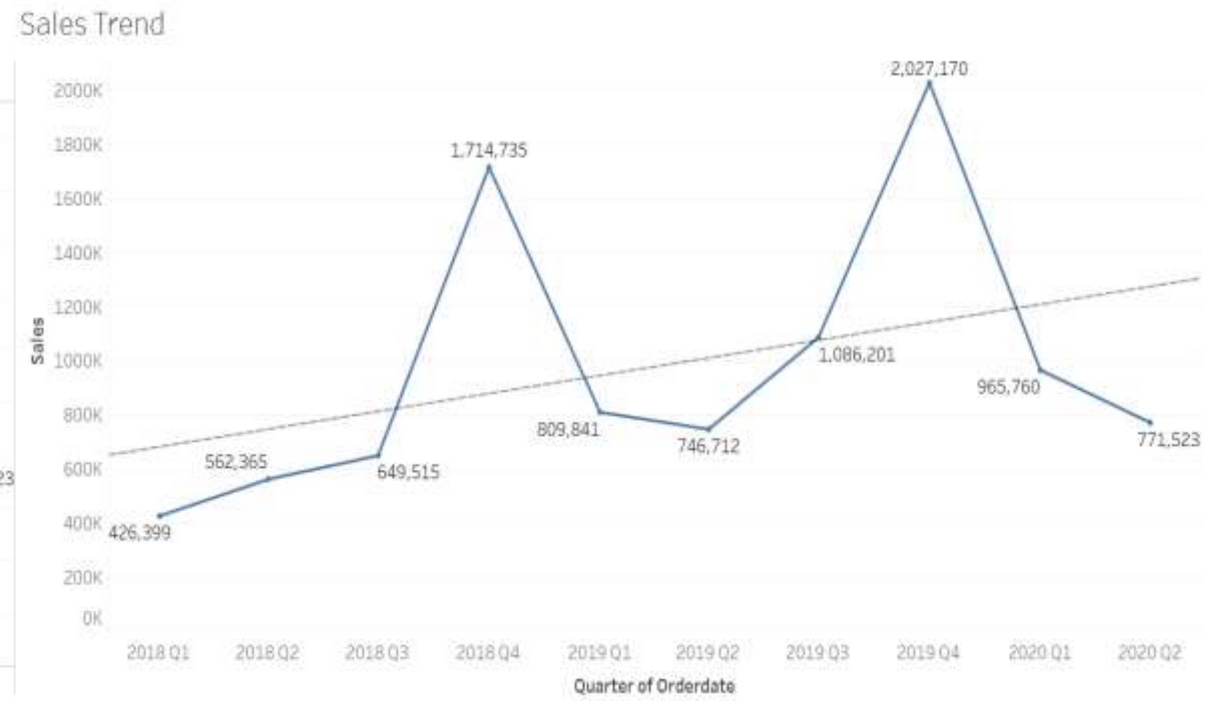




- 2019 has had the highest sales of $ 4,669,925.
- We could see a possible decline in sales trend in 2020 which could even improve towards the last two quarters.
- November is the peak sales month with close to $1,000,000 followed by October with nearly half of Nov month sale.
- The sales are likely to improve in Q3 and Q4 of each year.

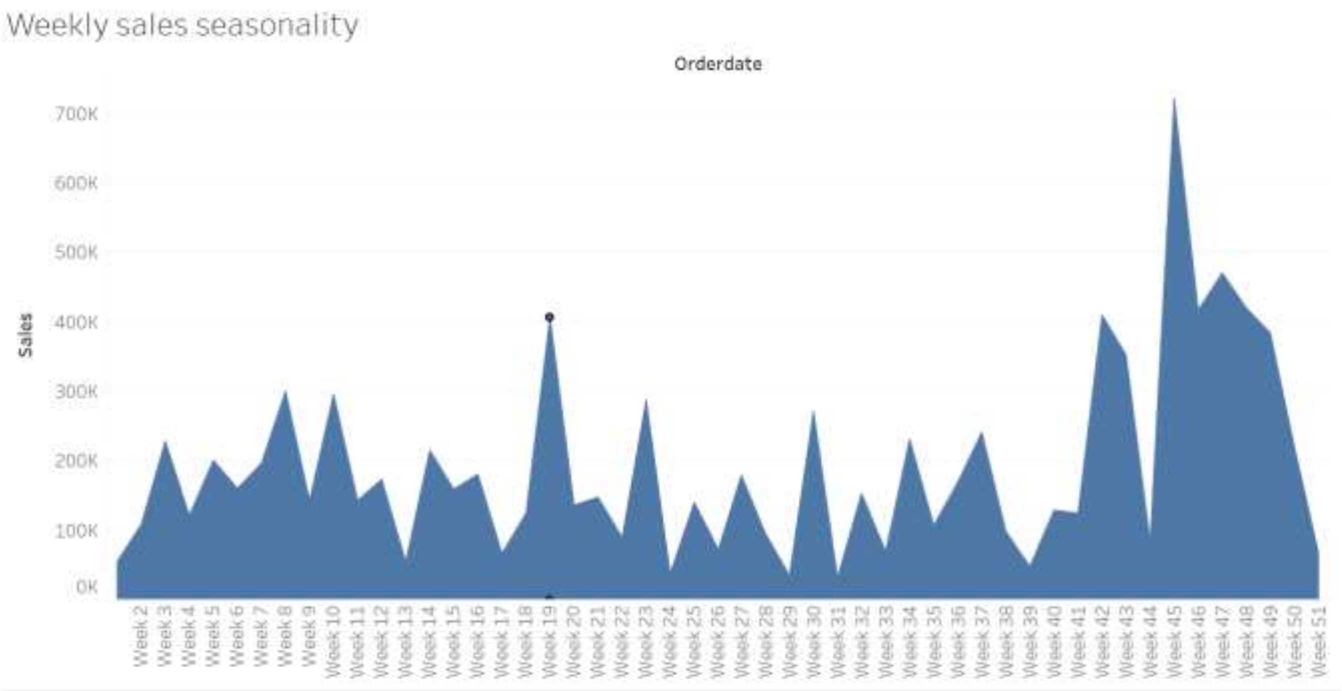
QUARTERLY TREND IN SALES

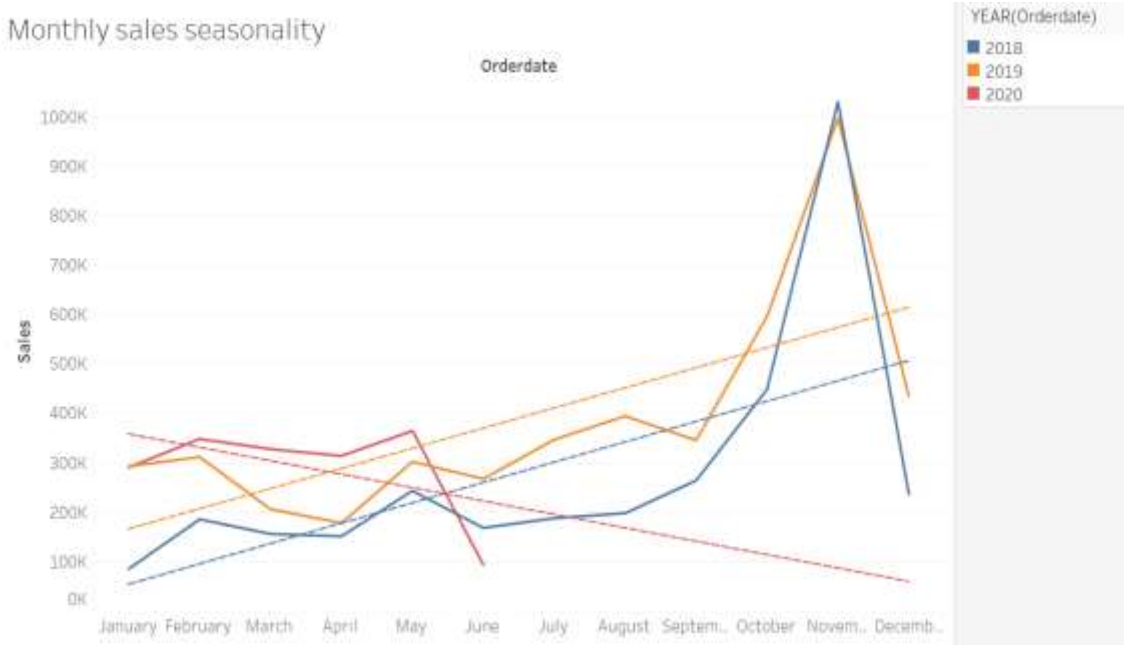




- We could notice an increase in sales trend over the years.
- Sales tend to pickup at Q3 and reach a peach towards Q4.
- 2019 Q4 has had the highest sales amount of $2,027,170.

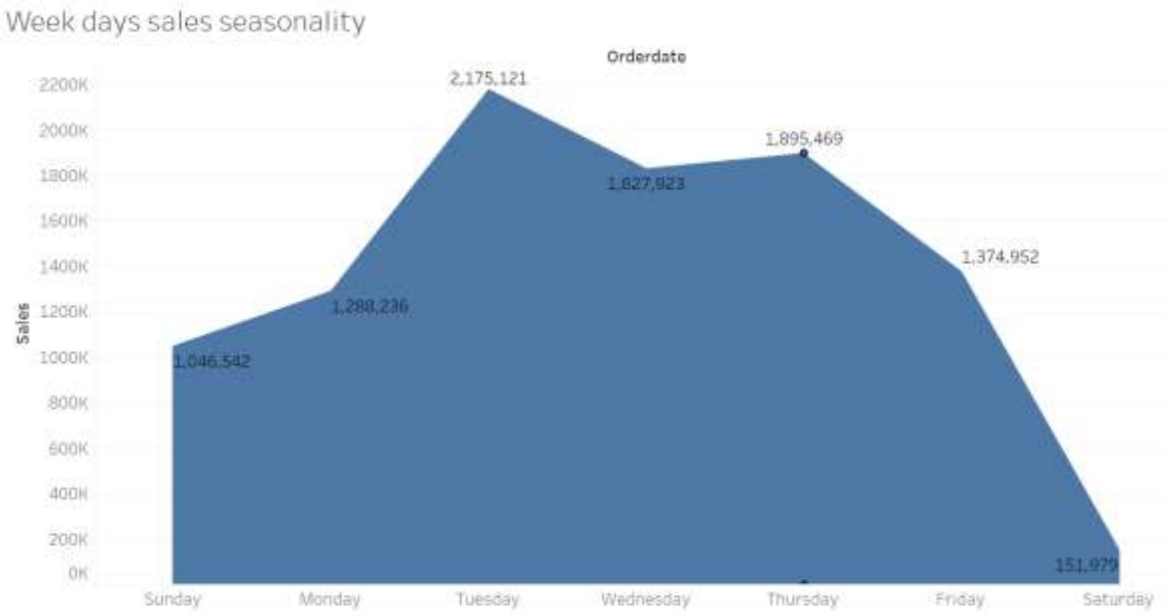
MONTHLY TREND IN SALES

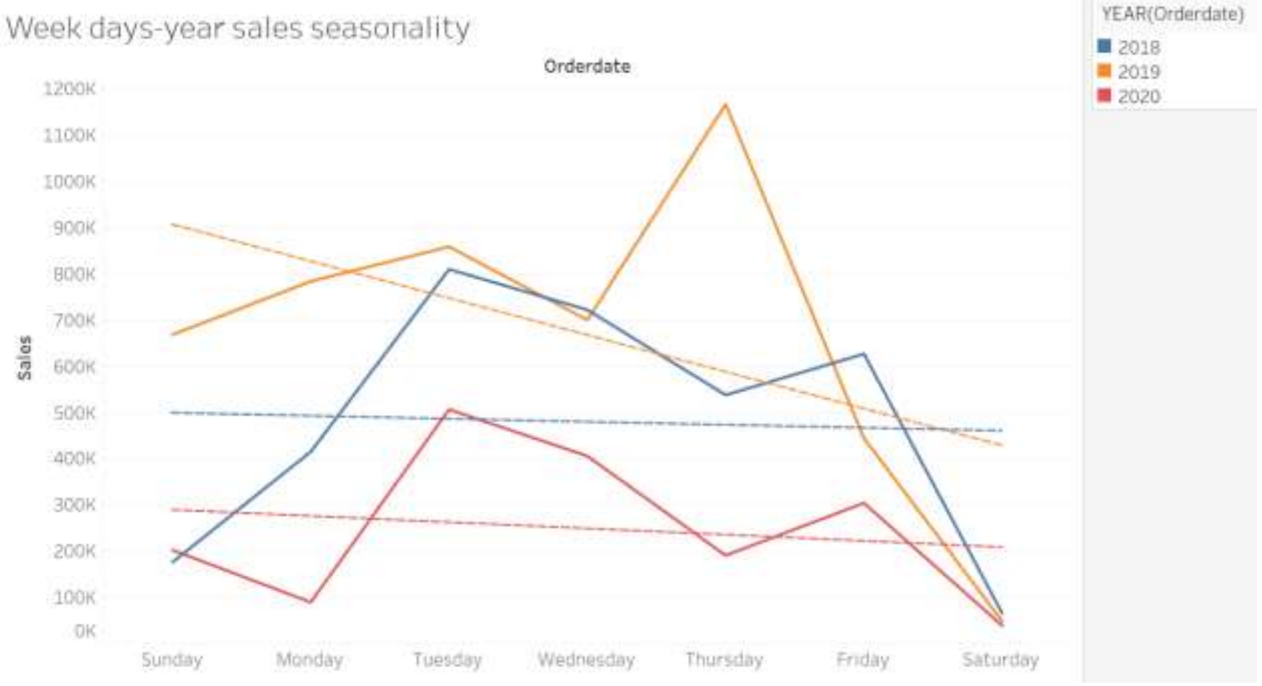




- 2020 has been doing better in regard to sales in the first two quarters but there is a noticeable decline in the month of June
- Sales likely to have a rise in October, November.
- 2018 Nov has had highest sale amount of $1,029,838.
- The sales have taken a rise at week45, 46 and week 47 but has led to decline over the next few weeks.

WEEKLY TREND IN SALES





- We can observe that Tuesdays seem to be the best day for sales. And Saturday is likely to have a decline in sales.
- But the strength of the relationship between sales and weekdays vary across different years.
- While there is a significant relationship in 2019, with weekdays explaining a moderate amount of variability in sales, this relationship is not observed in 2018. Additionally in 2020, the relationship is not statistically significant and explains only a minimal amount of variability in sales.

1.5 RFM

- Segmentation technique that gives you a measure of the loyalty of the customer
- KNIME tool is used for the purpose.
- **RFM Metrics-**
    - Recency- Freshness of customer activity- purchases/ visits (Time since last order/ last engaged)
    - Frequency- Freq of customer transaction/ visits (Total no. of transactions or average time between transactions/ engaged visits)
    - Monetary- Intention of customer to spend or purchasing power of customer (Total or average transaction value)
- **Parameters-** CUSTOMERNAME, ORDERDATE, ORDERNUMBER, SALES
- **Assumptions-**
    - For Recency, we are considering the difference from order date to the current date
    - Frequency- Total no of order by each customer by taking the count of Order Number
    - Monetary- Considering Sales column, we are calculating total sales by each customer
    - And have grouped data by Customer name.

- We have created 3 bins and provided the appropriate RFM Label based on the Recency, Frequency and Monetary Value.

	S Key	S Recency...	S Freq_RFM	S Monetar...
Row0	Bin 1	High	Low	Low
Row1	Bin 2	Medium	Medium	Medium
Row2	Bin 3	Low	High	High
Row3				
Row4				
Row5				
Row6				

- Based on the RFM label, we have segmented the customers into Active, At-risk and Inactive customers.

		Monetary_RFM			
Recency_RFM	**Freq_RFM**	**High**	**Medium**	**Low**	
High	**High**	9	1		
	Medium	1	8	1	
	Low		2	1	**Active**
Medium	**High**	10	1		
	Medium	1	21	1	
	Low		2	8	**At risk**
Low	**High**	1			
	Medium		7		
	Low		2	12	**Inactive**

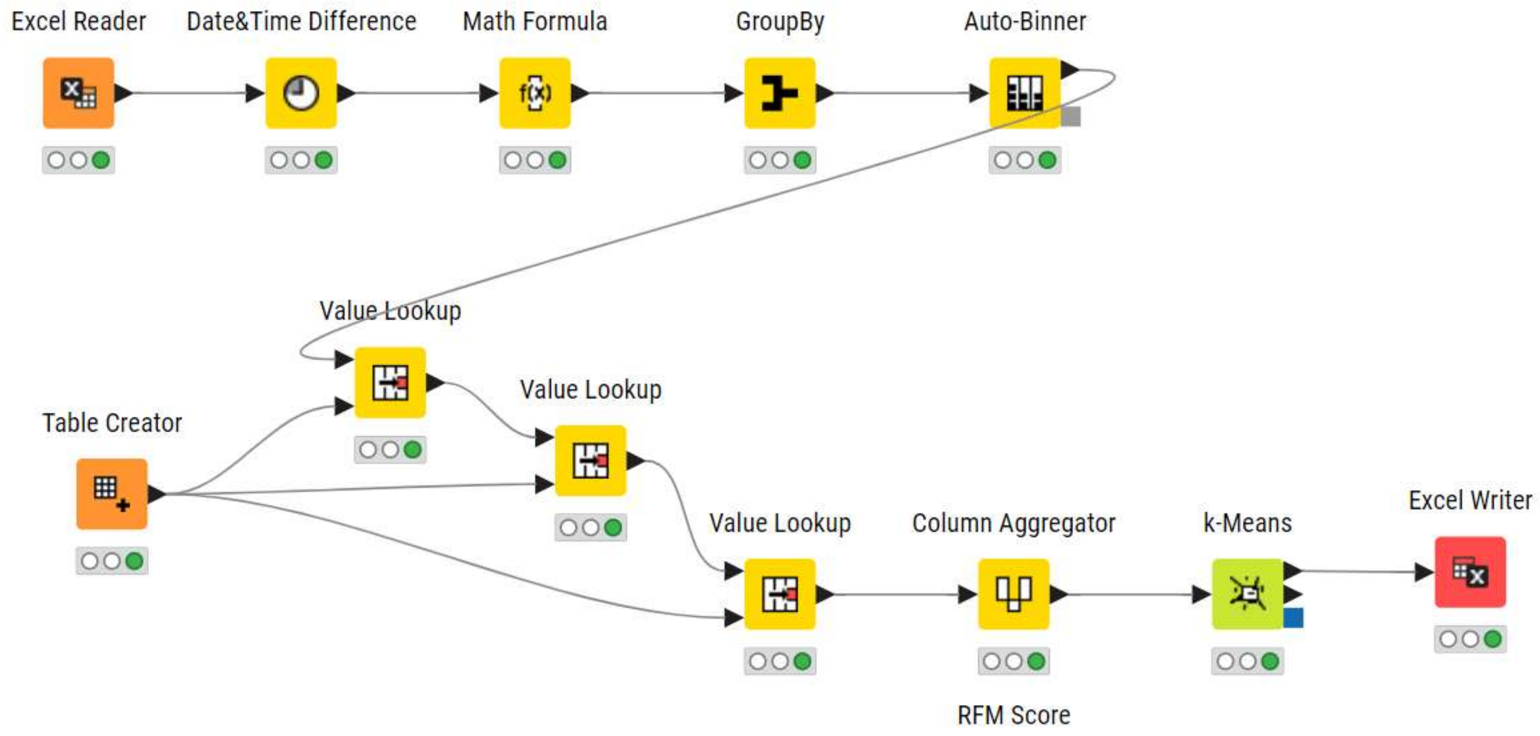
1.6 Segmentation

- Using RFM analysis, we have different segments of customers
- Cluster 0 - Customers in this cluster are highly active, making frequent purchases and they contribute significantly to the company's revenue. They have made purchases relatively recently, indicating ongoing engagement with the company.
- Cluster 1- Customers in this cluster are moderately active, making fewer purchases compared to Cluster 0. They haven't made purchases as recently as Cluster 0 but still contribute a significant amount to the company's revenue
- Cluster 2- Customers in this cluster are less active, making fewer purchases compared to other clusters. They haven't made purchases as recently as customers in Cluster 0 and Cluster 1 and they contribute less to the company's revenue.
- Cluster 3- Customers in this cluster are least active, making the fewest purchases and haven't made purchases recently. They contribute the least to the company's revenue compared to other clusters.

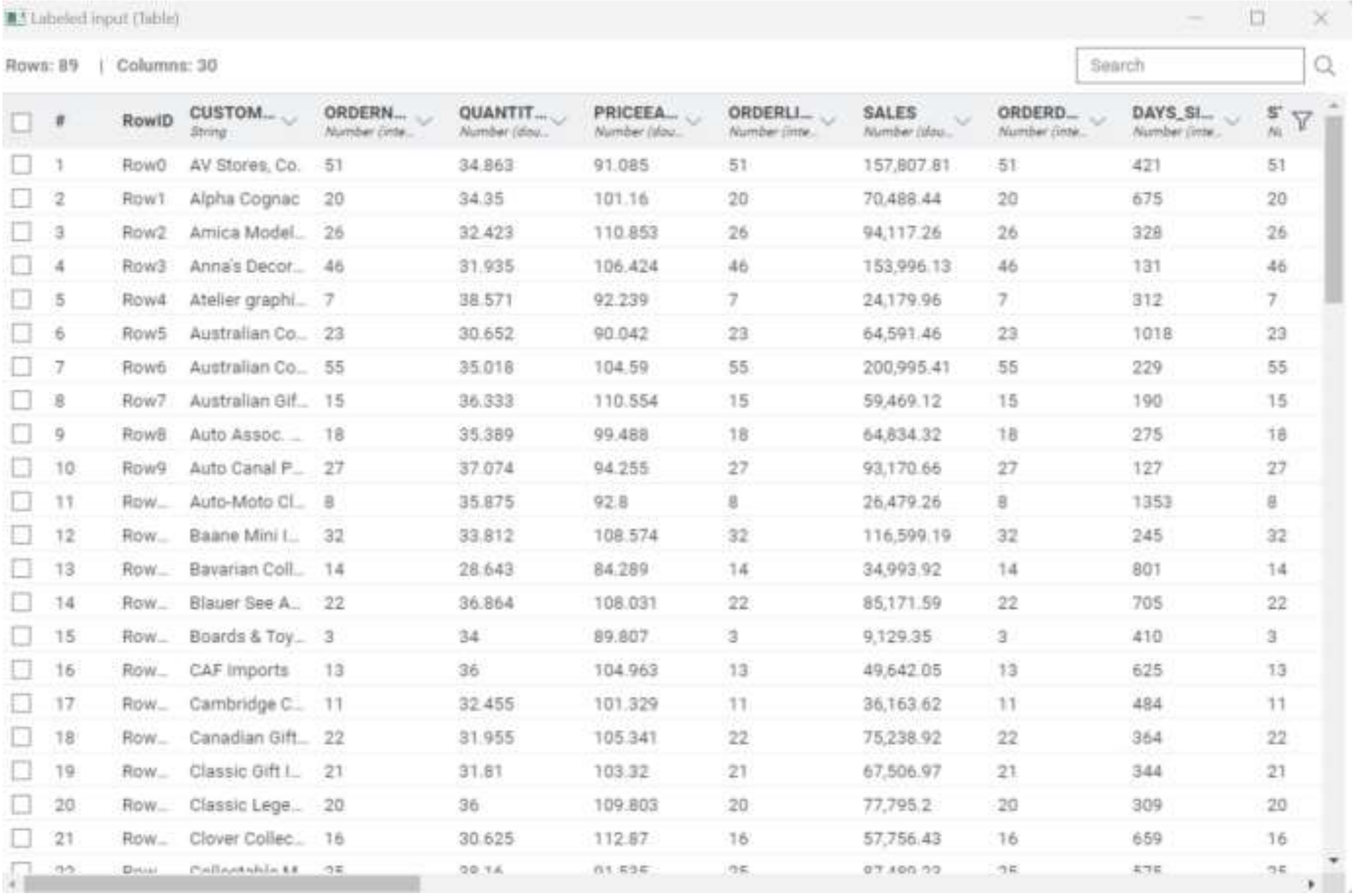
#	RowID	ORDERNUMBER Number (double)	Recency Number (double)	Monetary Number (double)
1	cluster_0	219.5	1,388	783,576.085
2	cluster_1	44	1,508.882	156,385.921
3	cluster_2	26.326	1,551.413	94,377.6
4	cluster_3	14.542	1,651.5	49,714.137

Cluster	RFM Score	Average of SALES	Average of ORDERNUMBER	Average of QUANTITYORDERED
cluster_0	High High High	$ 783,576.09	219.5	35.68912484
cluster_0 Total		**$ 783,576.09**	**219.5**	**35.68912484**
cluster_1	High High High	$ 152,027.24	42.14285714	35.87427081
	Low High High	$ 142,874.25	41	34.82926829
	Medium High High	$ 161,277.31	45.77777778	34.92961013
cluster_1 Total		**$ 156,385.92**	**44**	**35.3126856**
cluster_2	High High Medium	$ 115,498.73	36	34.33333333
	High Low Medium	$ 79,504.41	19.5	37.6
	High Medium High	$ 122,138.14	31	35.83870968
	High Medium Medium	$ 93,656.48	25.875	36.08236515
	Low Low Medium	$ 79,472.07	17	37.41176471
	Low Medium Medium	$ 95,414.26	26.85714286	34.48939036
	Medium High High	$ 120,783.07	34	34.67647059
	Medium High Medium	$ 108,951.13	35	32.57142857
	Medium Low Medium	$ 76,775.04	19.5	35.02631579
	Medium Medium High	$ 120,615.28	32	36.34375
	Medium Medium Medium	$ 92,580.98	26.33333333	34.46843155
cluster_2 Total		**$ 94,377.60**	**26.32608696**	**35.00760445**
cluster_3	High Low Low	$ 70,488.44	20	34.35
	High Medium Low	$ 64,591.46	23	30.65217391
	Low Low Low	$ 52,024.30	15.16666667	34.44880421
	Low Low Medium	$ 70,859.78	18	38.61111111
	Medium Low Low	$ 36,925.13	10.625	36.01863252
	Medium Medium Low	$ 67,506.97	21	31.80952381
cluster_3 Total		**$ 49,714.14**	**14.54166667**	**34.87322998**
Grand Total		**$ 109,665.41**	**30.86516854**	**35.04495763**

1.7 KNIME WORKFLOW:



1.8 OUTPUT TABLE:


Labeled input (Table)

Rows: 89 | Columns: 30

#	RowID	CUSTOM... String	ORDERN... Number (inte...	QUANTIT... Number (dou...	PRICEEA... Number (dou...	ORDERLI... Number (inte...	SALES Number (dou...	ORDERD... Number (inte...	DAYS_SI... Number (inte...	S' N...
1	Row0	AV Stores, Co.	51	34.863	91.085	51	157,807.81	51	421	51
2	Row1	Alpha Cognac	20	34.35	101.16	20	70,488.44	20	675	20
3	Row2	Amica Model...	26	32.423	110.853	26	94,117.26	26	328	26
4	Row3	Anna's Decor...	46	31.935	106.424	46	153,996.13	46	131	46
5	Row4	Atelier graphi...	7	38.571	92.239	7	24,179.96	7	312	7
6	Row5	Australian Co...	23	30.652	90.042	23	64,591.46	23	1018	23
7	Row6	Australian Co...	55	35.018	104.59	55	200,995.41	55	229	55
8	Row7	Australian Gif...	15	36.333	110.554	15	59,469.12	15	190	15
9	Row8	Auto Assoc. ...	18	35.389	99.488	18	64,834.32	18	275	18
10	Row9	Auto Canal P...	27	37.074	94.255	27	93,170.66	27	127	27
11	Row...	Auto-Moto Cl...	8	35.875	92.8	8	26,479.26	8	1353	8
12	Row...	Baane Mini I...	32	33.812	108.574	32	116,599.19	32	245	32
13	Row...	Bavarian Coll...	14	28.643	84.289	14	34,993.92	14	801	14
14	Row...	Blauer See A...	22	36.864	108.031	22	85,171.59	22	705	22
15	Row...	Boards & Toy...	3	34	89.807	3	9,129.35	3	410	3
16	Row...	CAF Imports	13	36	104.963	13	49,642.05	13	625	13
17	Row...	Cambridge C...	11	32.455	101.329	11	36,163.62	11	484	11
18	Row...	Canadian Gift...	22	31.955	105.341	22	75,238.92	22	364	22
19	Row...	Classic Gift I...	21	31.81	103.32	21	67,506.97	21	344	21
20	Row...	Classic Lege...	20	36	109.803	20	77,795.2	20	309	20
21	Row...	Clover Collec...	16	30.625	112.87	16	57,756.43	16	659	16

1.9 Best customers:

CUSTOMERNAME
AV Stores, Co.
Anna's Decorations, Ltd
Australian Collectors, Co.
Corrida Auto Replicas, Ltd
Danish Wholesale Imports
Diecast Classics Inc.
Dragon Souveniers, Ltd.
Euro Shopping Channel
L'ordine Souveniers
La Rochelle Gifts
Land of Toys Inc.
Mini Gifts Distributors Ltd.
Muscle Machine Inc
Online Diecast Creations Co.
Reims Collectables
Rovelli Gifts
Salzburg Collectables
Scandinavian Gift Ideas
Souveniers And Things Co.
Technics Stores Inc.
The Sharp Gifts Warehouse

- The "best" customers are typically those with the highest monetary value, high/ medium freq and high/ medium recency
- These customers contribute significantly to the company's revenue and profitability making them highly valuable from a financial perspective.
- They engage with the company regularly, make repeat purchases over time and are more likely to make future purchases.

1.10 Customers are on the verge of churning:

CUSTOMERNAME
Marseille Mini Autos
Canadian Gift Exchange Network
giftsbymail.co.uk
Enaco Distributors
Collectables For Less Inc.
Signal Gift Stores
Motor Mint Distributors Inc.
Blauer See Auto, Co.
Mini Classics

- Customers who fall into the "Medium Recency" category with "High/ medium Freq" but with lower/ medium monetary value are considered at risk of churning.
- They still purchase frequently but may be spending less compared to before, indicating a potential decline in engagement or satisfaction.

1.11 Lost customers:

CUSTOMERNAME
Auto Assoc. & Cie.
Bavarian Collectables Imports, Co.
CAF Imports
Cambridge Collectables Co.
Clover Collections, Co.
Daedalus Designs Imports
Double Decker Gift Stores, Ltd
Iberia Gift Imports, Corp.
Online Mini Collectables
Osaka Souveniers Co.
Signal Collectibles Ltd.
West Coast Collectables Co.

- Customers who are labeled as "Low Recency" with "Low Freq" and "Low Monetary" are considered lost or inactive customers.
- They haven't made recent purchases and have low engagement levels and have spent the least amount of money indicating they may have already churned or are at high risk of doing so.

1.12 Loyal customers:

CUSTOMERNAME
The Sharp Gifts Warehouse
Souveniers And Things Co.
Salzburg Collectables
Reims Collectables
Mini Gifts Distributors Ltd.
La Rochelle Gifts
L'ordine Souveniers
Euro Shopping Channel
Danish Wholesale Imports

- High Recency, High Frequency, High Monetary: Customers in this segment are actively purchasing frequently, contribute significantly to revenue and have recently made purchases.

1.13 Recommendations:

- For Cluster 0- High-value customers,
  - Offer exclusive loyalty rewards or VIP programs for continued purchases.
  - Provide product recommendations based on their past purchase history.
  - Send targeted promotional offers or discounts on complementary products to encourage them.
  - Excel in customer service and prioritize their needs to enhance their overall experience.
- For Cluster 1- Medium-value customers-
  - Offer limited-time promotions or discounts to boost purchases.
  - Implement a referral program to encourage them to refer friends or family members to the company.
  - Provide programs/ content related to products they've purchased to enhance their product knowledge and usage.
- For Cluster 2 & 3- Low-value customers,
  - Set campaigns, provide special offers or discounts to promote them to make a purchase.
  - Conduct customer surveys or feedback polls to understand their needs and preferences better.
  - Provide personalized recommendations based on their past purchases to reignite their interest.
  - Offer incentives for customers to update their profiles or preferences to receive more relevant communications in the future.

1.14 Marketing Strategies:

- Tailor marketing campaigns to highlight popular product lines such as "Classic Cars" and "Vintage Cars" based on their high demand.
- Showcase unique features or benefits of less preferred product lines like "Trains" or "Ships" to increase interest and sales.
- Customize marketing promotions based on regional and cultural preferences.
- Offer promotions or discounts to cater to the specific needs and preferences of customers in different countries or regions.
- Adjust pricing strategies or promotional offers based on deal size preferences to maximize sales and profitability.
- Offer flexible pricing options or financing plans for customers interested in larger deal sizes to facilitate larger transactions.
- By implementing these recommendations and customized marketing strategies, the automobile parts manufacturing company can effectively leverage insights about their customers' buying patterns to drive engagement, increase sales, and foster long-term loyalty.

1.15 Dataset: Sales_Data.xlsx

Data Dictionary:

Column Name	Description
ORDERNUMBER	This column represents the unique identification number assigned to each order.
QUANTITYORDERED	It indicates the number of items ordered in each order.
PRICEEACH	This column specifies the price of each item in the order.
ORDERLINENUMBER	It represents the line number of each item within an order.
SALES	This column denotes the total sales amount for each order, which is calculated by multiplying the quantity ordered by the price of each item.
ORDERDATE	It denotes the date on which the order was placed.
DAYS_SINCE_LASTORDER	This column represents the number of days that have passed since the last order for each customer. It can be used to analyze customer purchasing patterns.
STATUS	It indicates the status of the order, such as "Shipped," "In Process," "Cancelled," "Disputed," "On Hold," or "Resolved"
PRODUCTLINE	This column specifies the product line categories to which each item belongs.
MSRP	It stands for Manufacturer's Suggested Retail Price and represents the suggested selling price for each item.
PRODUCTCODE	This column represents the unique code assigned to each product.
CUSTOMERNAME	It denotes the name of the customer who placed the order.
PHONE	This column contains the contact phone number for the customer.
ADDRESSLINE1	It represents the first line of the customer's address.
CITY	This column specifies the city where the customer is located.
POSTALCODE	It denotes the postal code or ZIP code associated with the customer's address.
COUNTRY	This column indicates the country where the customer is located.
CONTACTLASTNAME	It represents the last name of the contact person associated with the customer.
CONTACTFIRSTNAME	This column denotes the first name of the contact person associated with the customer.
DEALSIZE	It indicates the size of the deal or order, which are the categories "Small," "Medium," or "Large."

PART B

2.1 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.

2.2 ABOUT DATA :

Shape of the dataset: (20641, 3)

Head of the dataset:

	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

Null values:

```
Date        0
Order_id    0
Product     0
dtype: int64
```

- The dataset contains 20,641 entries and 3 columns.
- The columns are 'Date', 'Order_id', and 'Product’.
- There are no null values present in the dataset.

SUMMARY:

- 'Date' and 'Product' are of object data type, while 'Order_id' is of integer data type.
- We have converted the ‘date’ to Datetime data type.

```
<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
```

```
<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  datetime64[ns]
 1   Order_id  20641 non-null  int64
 2   Product   20641 non-null  object
dtypes: datetime64[ns](1), int64(1), object(1)
memory usage: 483.9+ KB
```

Summary of Numeric Variable:

	Order_id
count	15911.000000
mean	574.150462
std	328.537425
min	1.000000
25%	289.500000
50%	579.000000
75%	859.000000
max	1139.000000

- There are 15,911 unique order IDs in the dataset.
- The minimum order ID is 1, which is expected as order IDs typically start from 1.
- The maximum order ID is 1139, indicating the highest order number in the dataset.
- 25% of the order IDs are below 289.5.
- 50% of the order IDs are below 579.
- 75% of the order IDs are below 859.

Summary of Categorical Variable:

	Product
count	20641
unique	37
top	poultry
freq	640

- There are 20,641 entries in the 'Product' column, indicating the total number of transactions recorded in the dataset.
- There are 37 unique products in the dataset.
- The most frequently occurring product in the dataset is 'poultry’.
- The product 'poultry' appears 640 times in the dataset, indicating it's the most commonly purchased item among all the transactions.

Duplicate values:

```
Number of duplicate rows = 4730
```

	Date	Order_id	Product
10	2018-01-01	1	all- purpose
13	2018-01-01	1	all- purpose
18	2018-01-01	1	dinner rolls
29	2018-01-01	2	waffles
31	2018-01-01	2	hand soap
...	...	...	...
20616	2020-02-24	1137	paper towels
20632	2020-02-25	1138	sandwich bags
20633	2020-02-25	1138	toilet paper
20635	2020-02-25	1138	soda
20636	2020-02-25	1138	soda

4730 rows × 3 columns

```
Number of duplicate rows = 0
```

- Initially, there were 4,730 duplicate rows in the DataFrame.
- Since they had the same combination of 'Date', 'Order_id', and 'Product', we removed the duplicate records.
- After dropping the duplicate rows, the DataFrame now contains no duplicate entries.

Unique values:

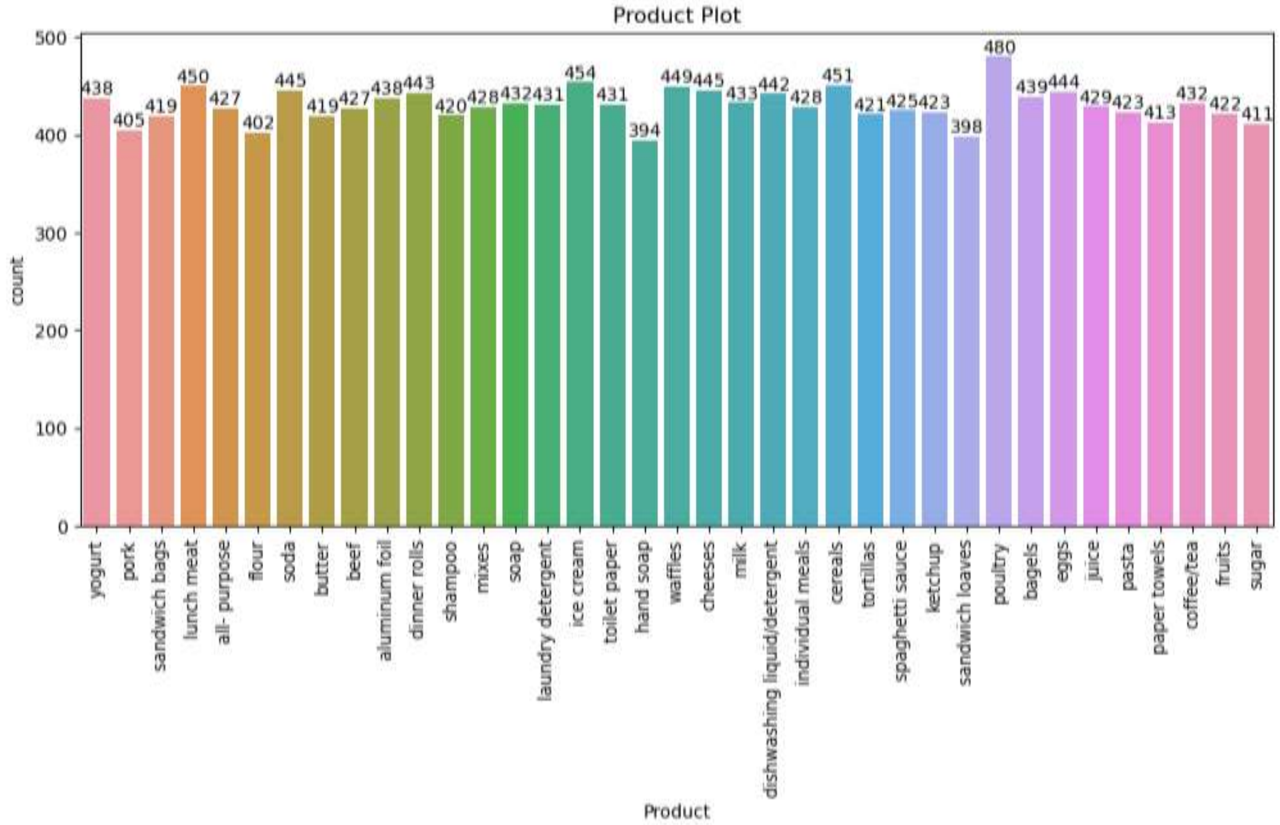
```
PRODUCT :  37
Product
hand soap                        394
sandwich loaves                  398
flour                            402
pork                             405
sugar                            411
paper towels                     413
butter                           419
sandwich bags                    419
shampoo                          420
tortillas                        421
fruits                           422
ketchup                          423
pasta                            423
spaghetti sauce                  425
beef                             427
all- purpose                     427
mixes                            428
individual meals                 428
juice                            429
laundry detergent                431
toilet paper                     431
soap                             432
coffee/tea                       432
milk                             433
aluminum foil                    438
yogurt                           438
bagels                           439
dishwashing liquid/detergent     442
dinner rolls                     443
eggs                             444
cheeses                          445
soda                             445
waffles                          449
lunch meat                       450
cereals                          451
ice cream                        454
poultry                          480
Name: count, dtype: int64
```

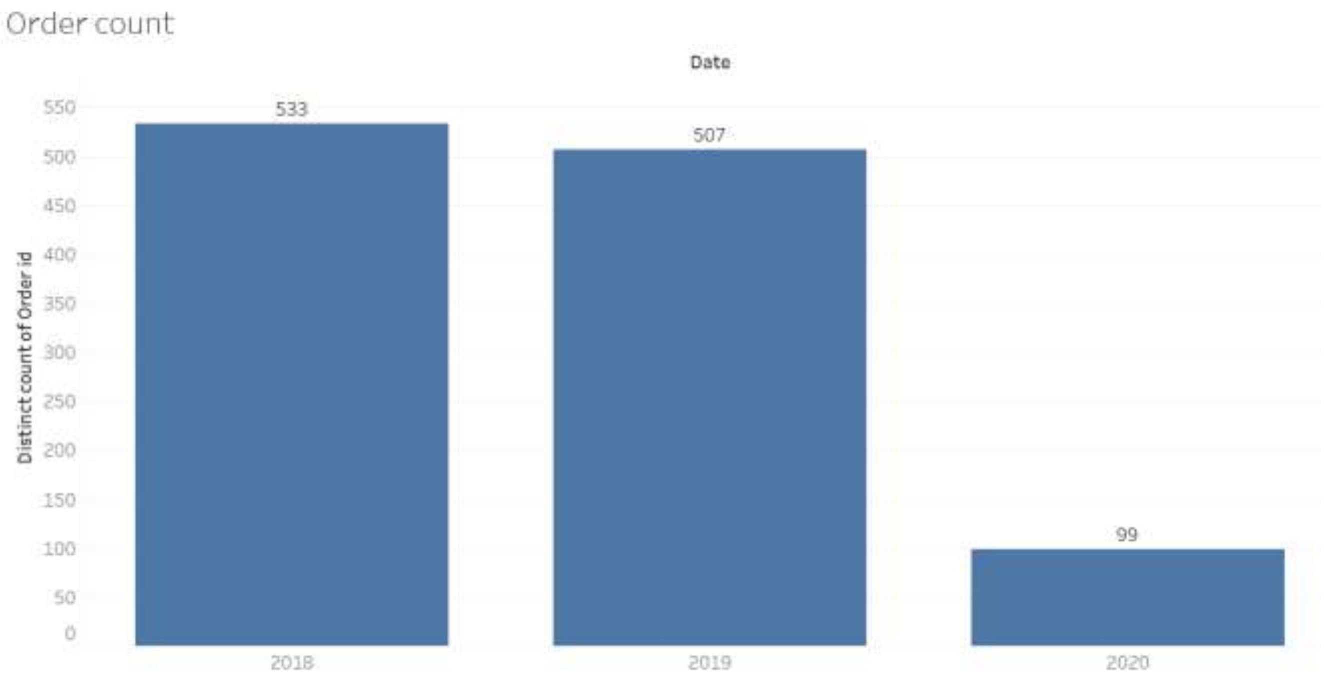
- There are 37 unique products in the dataset, indicating a diverse range of items sold by the grocery store.
- 'poultry' appears to be the most popular product, with a count of 480, indicating that it was purchased the most frequently among all the products.
- Other popular products include 'ice cream' (454), 'cereals' (451), 'lunch meat' (450), 'waffles' (449), 'soda' (445), 'cheeses' (445), 'eggs' (444), etc.
- There's variability in the popularity of products, with some items being purchased more frequently than others.
- Hand soap, Sandwich loaves, Flour, Pork are some of the least preferred products comparatively.

2.3 Exploratory Data Analysis:

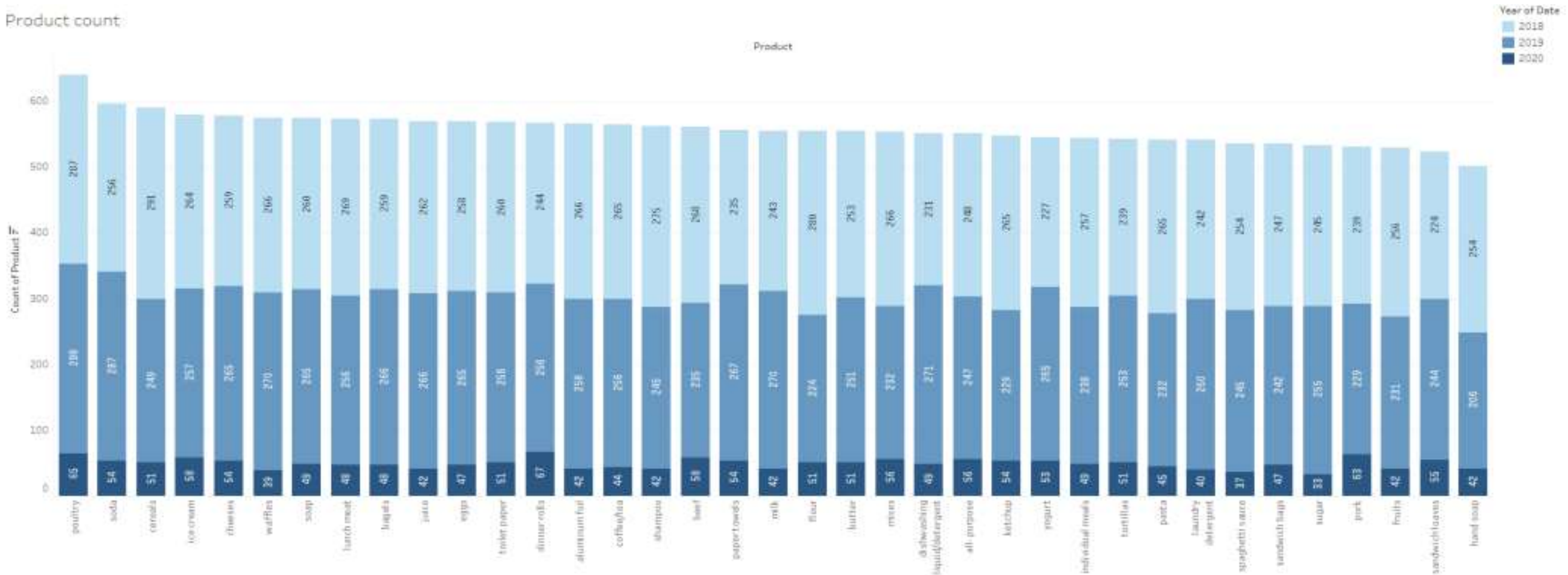
Univariate Analysis:

- There are 37 unique products in the dataset, indicating a diverse range of items sold by the grocery store.
- 'poultry' appears to be the most popular product, with a count of 480, indicating that it was purchased the most frequently among all the products.
- Other popular products include 'ice cream' (454), 'cereals' (451), 'lunch meat' (450), 'waffles' (449), 'soda' (445), 'cheeses' (445), 'eggs' (444), etc.
- There's variability in the popularity of products, with some items being purchased more frequently than others.
- Hand soap, Sandwich loaves, Flour, Pork are some of the least preferred products comparatively.



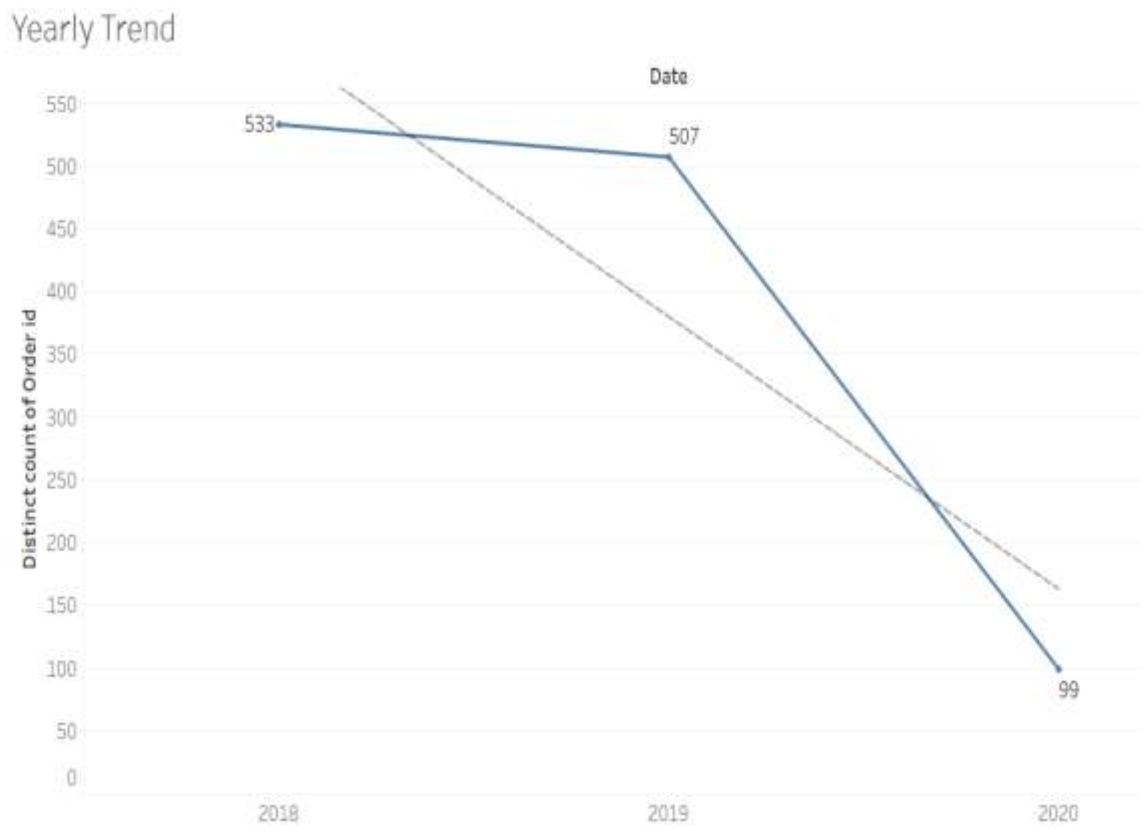


- 2018 has had the highest order of 533 which has reduced in 2019 with an order count of 507.
- Considering the Q1 of 2020, the order count is 99.

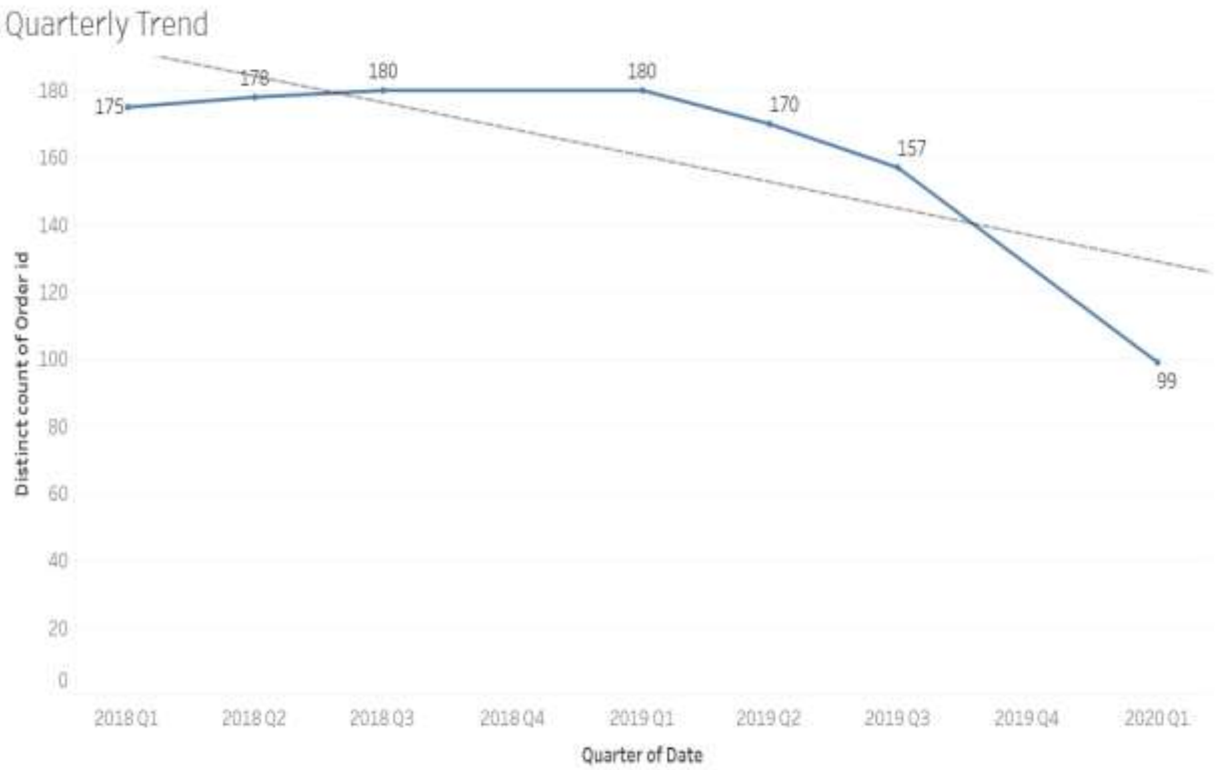


- In 2018, cereals were the most purchased product with a count of 291, followed closely by poultry with a count of 287, and flour with a count of 280. These products were among the top choices for customers.
- Sandwich loaves were the least purchased product in 2018 with a count of 224, indicating relatively lower demand compared to other items.
- The rankings of some products changed between 2018 and 2019. For example, soda, which was the most purchased product in 2018 with a count of 256, became the second most purchased product in 2019 with a count of 287.
- Dishwashing liquid saw an increase in demand from 231 in 2018 to 271 in 2019, indicating a rise in popularity or possibly changes in consumer behavior or preferences.
- In the first quarter of 2020, dinner rolls emerged as the most purchased product with a count of 67, followed closely by poultry with a count of 65, and pork with a count of 63. This shows a shift in consumer preferences compared to the previous years.
- Sugar was the least purchased product in the first quarter of 2020, with a count of 33, indicating a decline in demand for this item.
- The fluctuations in product counts across the years and quarters highlight the variability in consumer preferences.

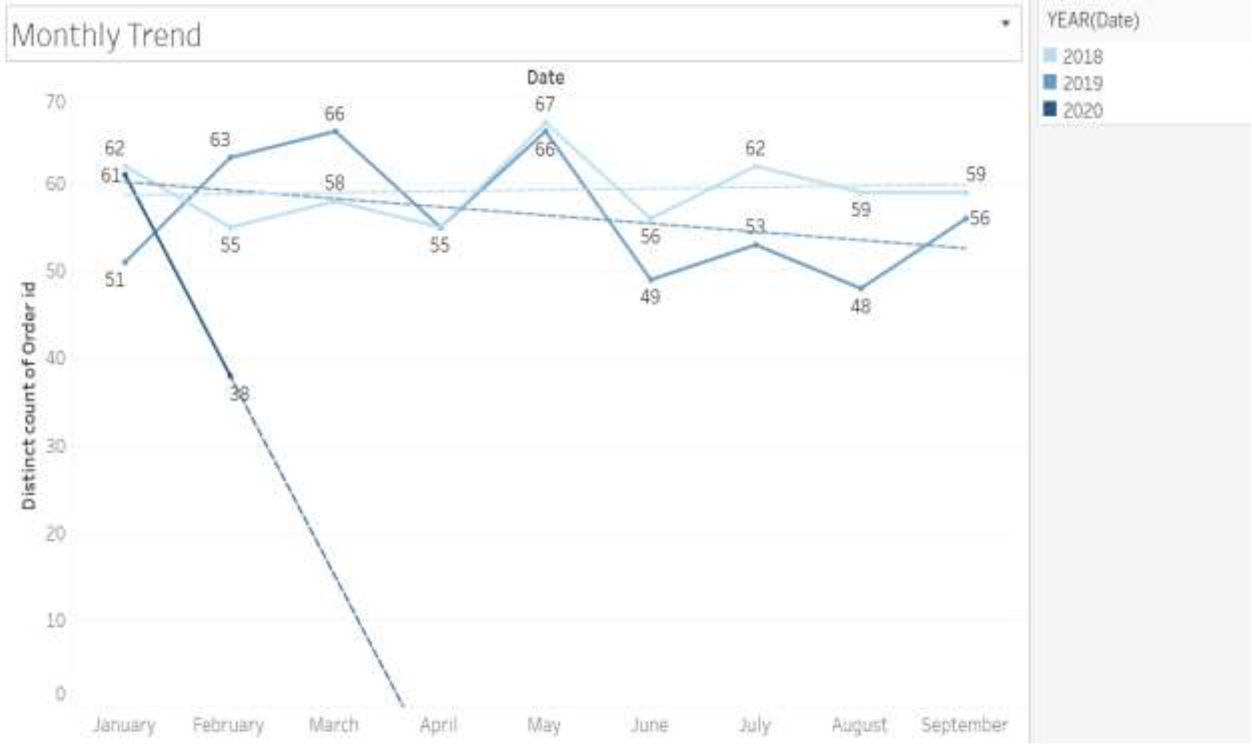
2.4 SALES TREND

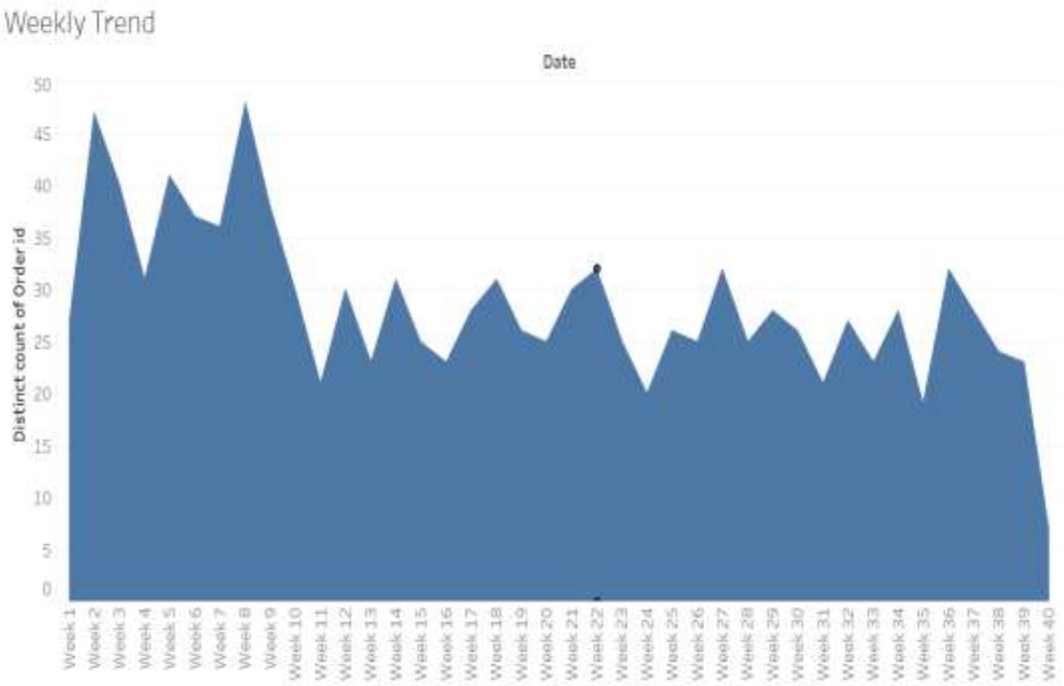


- In 2018, there were 533 distinct order IDs.
- In 2019, the number decreased slightly to 507 distinct order IDs.
- The gradual decrease suggests a potential decline in the number of unique transactions or customers during this period.
- In the first quarter of 2020, there were 99 distinct order IDs.
- The R-squared value of 0.79476 suggests a strong correlation between the year and the distinct count of order IDs, indicating that the trend is fairly predictable.

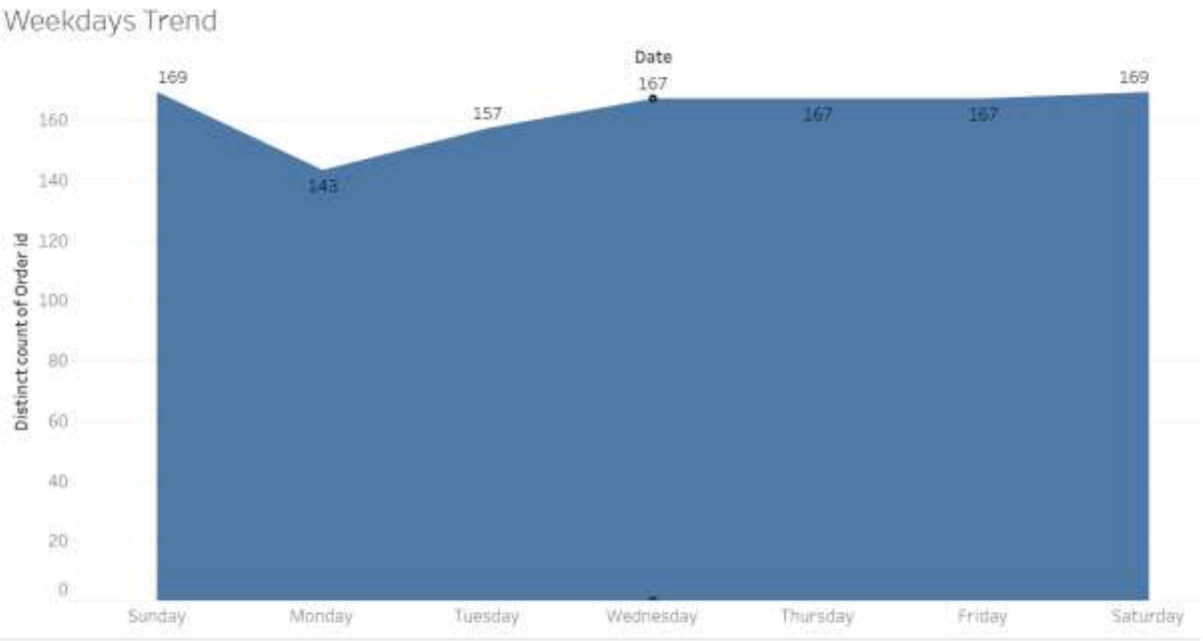


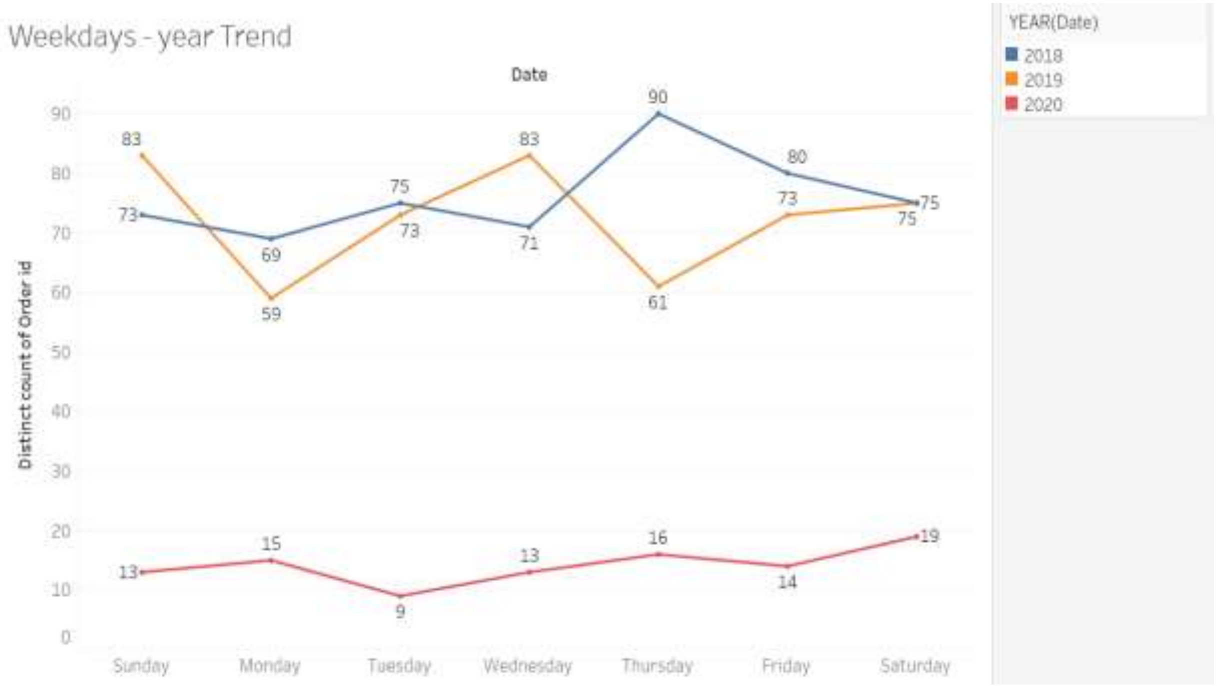
- There is a trend of fluctuating distinct counts of order IDs over the quarters from 2018 Q1 to 2020 Q1.
- In 2018, there is a slight increase from Q1 to Q3, with Q3 having the highest count of 180.
- However, in 2019, there is a decline in the distinct count of order IDs over the quarters, with Q3 having the lowest count of 157.
- In 2020 Q1, there is a significant decrease in the distinct count to 99, which is notably lower compared to the counts in the previous quarters.
- The trend line analysis indicates an R-squared value of 0.597032, suggesting that approximately 59.7% of the variability in the distinct count of order IDs can be explained by the trend line.





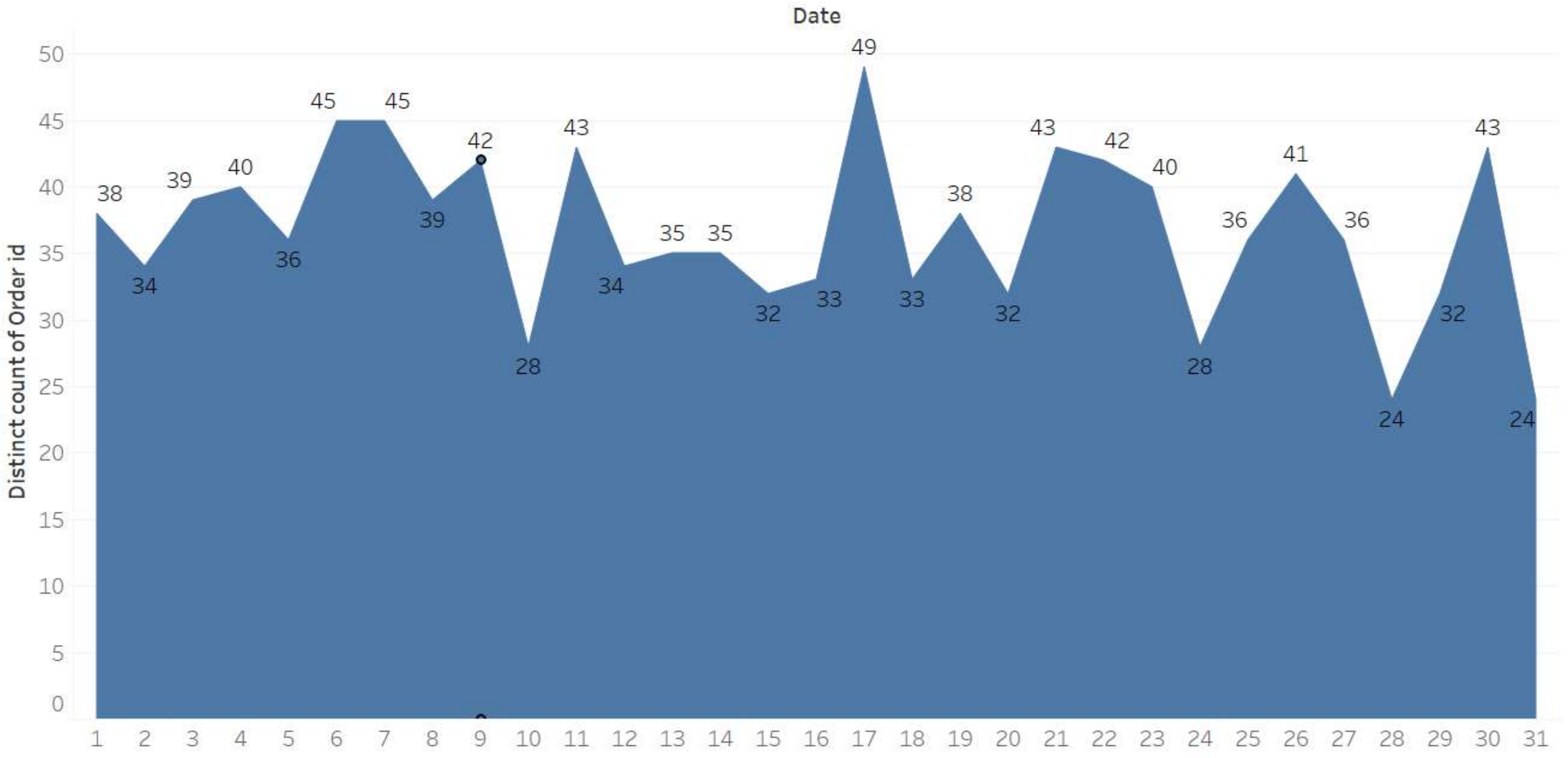
- The transaction counts in 2018 exhibit some variability from month to month, ranging from 55 to 67 transactions per month.
- May had the highest transaction count in 2018 with 67 transactions, while February had the lowest with 55 transactions.
- In 2019, there is also variability in the transaction counts across the months, ranging from 48 to 66 transactions per month.
- March had the highest transaction count in 2019 with 66 transactions, while August had the lowest with 48 transactions.
- Transaction counts dropped significantly in February 2020 compared to January, with January having 61 transactions and February having only 38 transactions.
- While there are fluctuations in transaction counts from month to month, there doesn't seem to be a consistent pattern of growth or decline over the years.
- Overall, week 8 has had highest order of 48 followed by week 2 with order of 47 and week 40 the least with 7 orders.



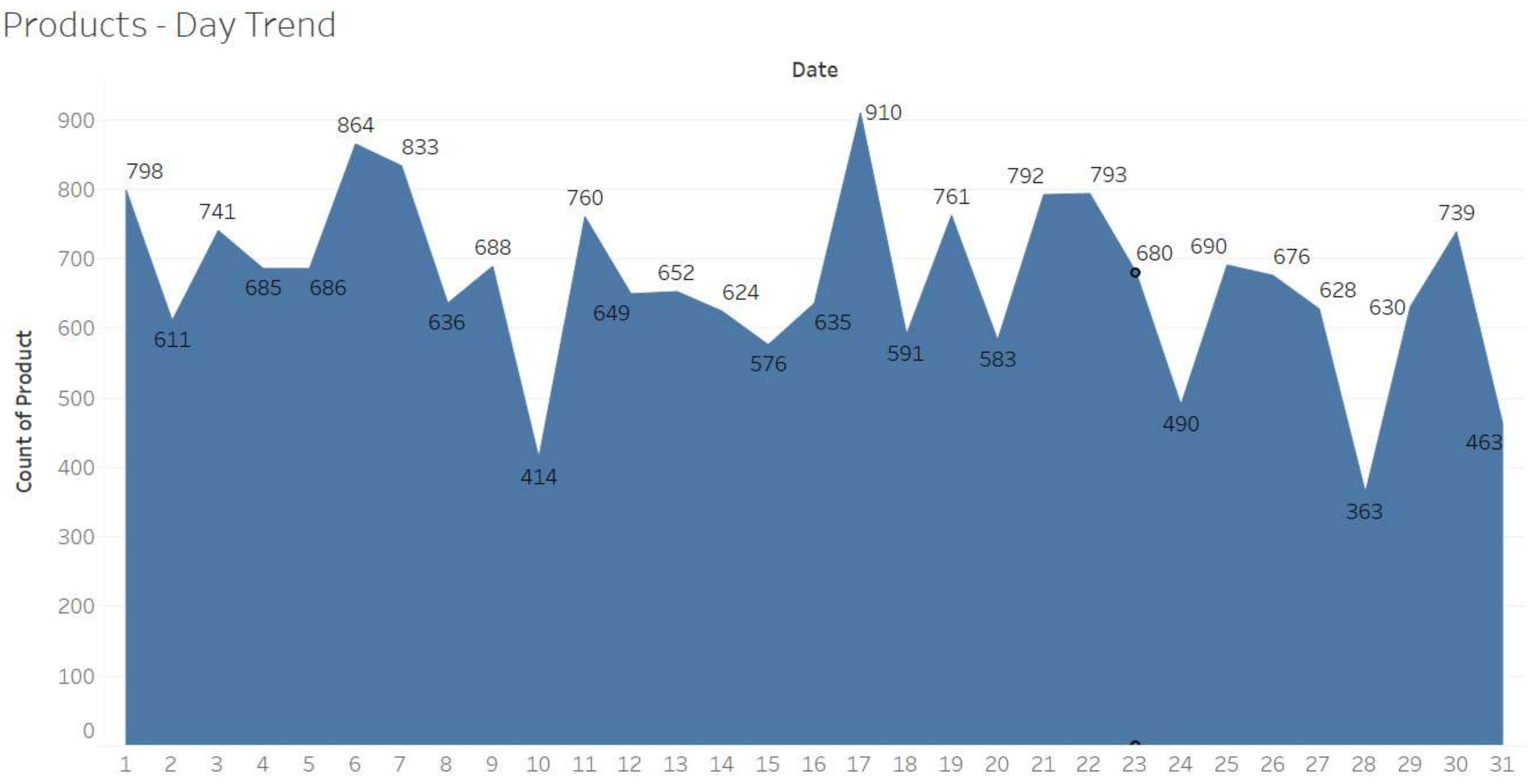


- Sunday and Saturday consistently have the highest number of orders, with 169 orders each. This suggests that weekends are generally busier days for the grocery store.
- Monday, Wednesday, Thursday, and Friday have relatively similar order counts, ranging from 143 to 167 orders, indicating consistent demand throughout the weekdays.
- Comparing the same days across the years, there are fluctuations in order counts. For example, Sunday order counts increased from 73 in 2018 to 83 in 2019 and then decreased to 13 in the first quarter of 2020.
- There is variability in order counts across all days of the week in each year, suggesting that consumer behavior may be influenced by various factors beyond just the day of the week.

Orders - Day Trend



- The daily order counts vary throughout the month, ranging from 24 to 49 orders per day.
- There doesn't seem to be a consistent pattern of increase or decrease in order counts over the month.
- The highest order count observed is 49 orders, while the lowest is 24 orders.
- Days with relatively higher order counts (above 40) are days 6, 7, 10, 17, 21, 22, and 30.
- Days with relatively lower order counts (below 30) are days 10, 11, 15, 19, 24, 27, and 29.



- There is variability in the counts of the products bought across the 31 days of the month. The counts range from 363 to 910, indicating fluctuations in demand or sales of the product throughout the month.

2.5 Market Basket Analysis

- Enticing customers based on buying pattern.
- Market basket analysis (MBA) is a data mining technique used in retail and e-commerce to discover relationships between products that are frequently purchased together.
- The goal of market basket analysis is to identify patterns, associations and correlations among items that co-occur in transactions.

2.6 Association Rule-

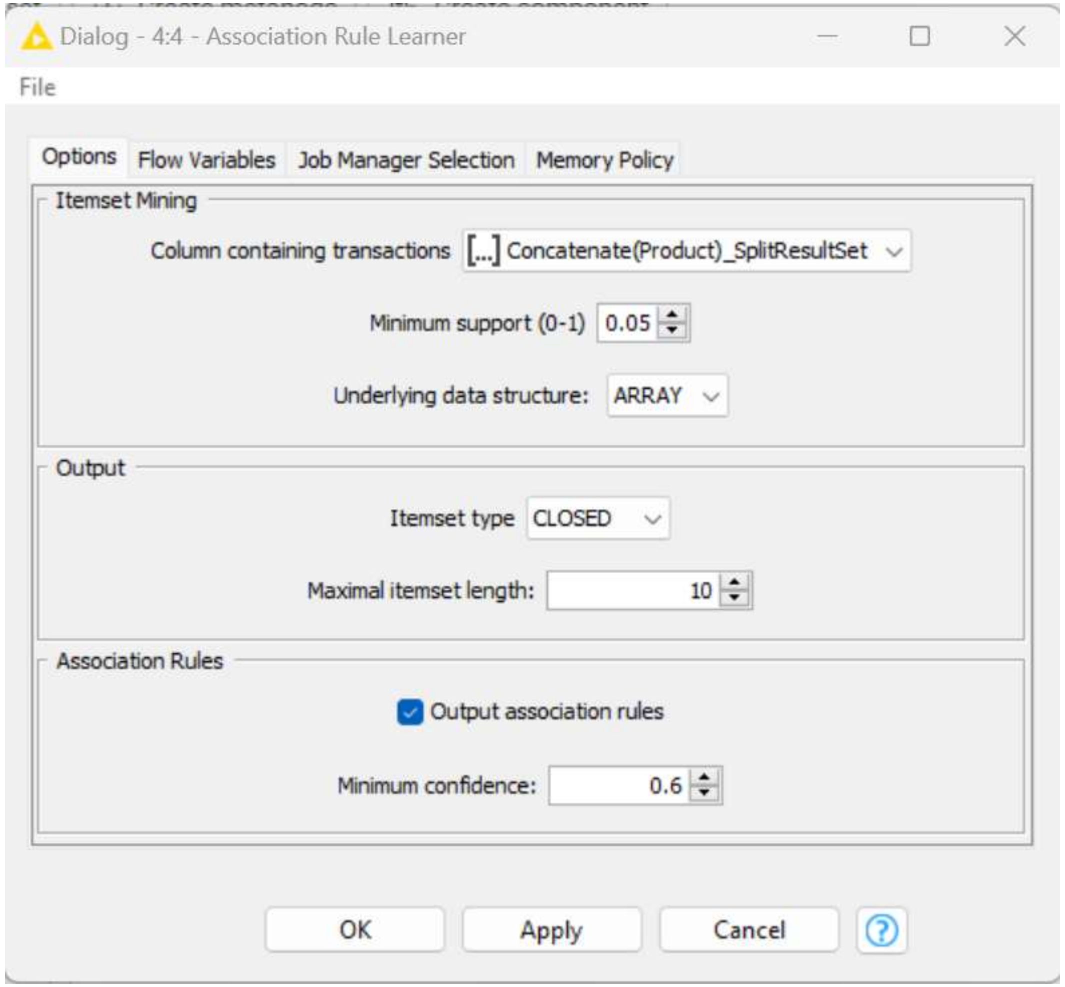
- Set of rules where likelihood of buying a product is great
- They describe relationships between items in a dataset of transactions
- They help businesses understand patterns in customer purchasing behavior, identify cross-selling and upselling opportunities and make data-driven decisions to improve marketing strategies and customer satisfaction.
- **Metrics used in market basket analysis:**

  Support:

  - It indicates how frequently a particular combination of items appears together in transactions.
  - It is calculated as the proportion of transactions that contain all the items in the rule

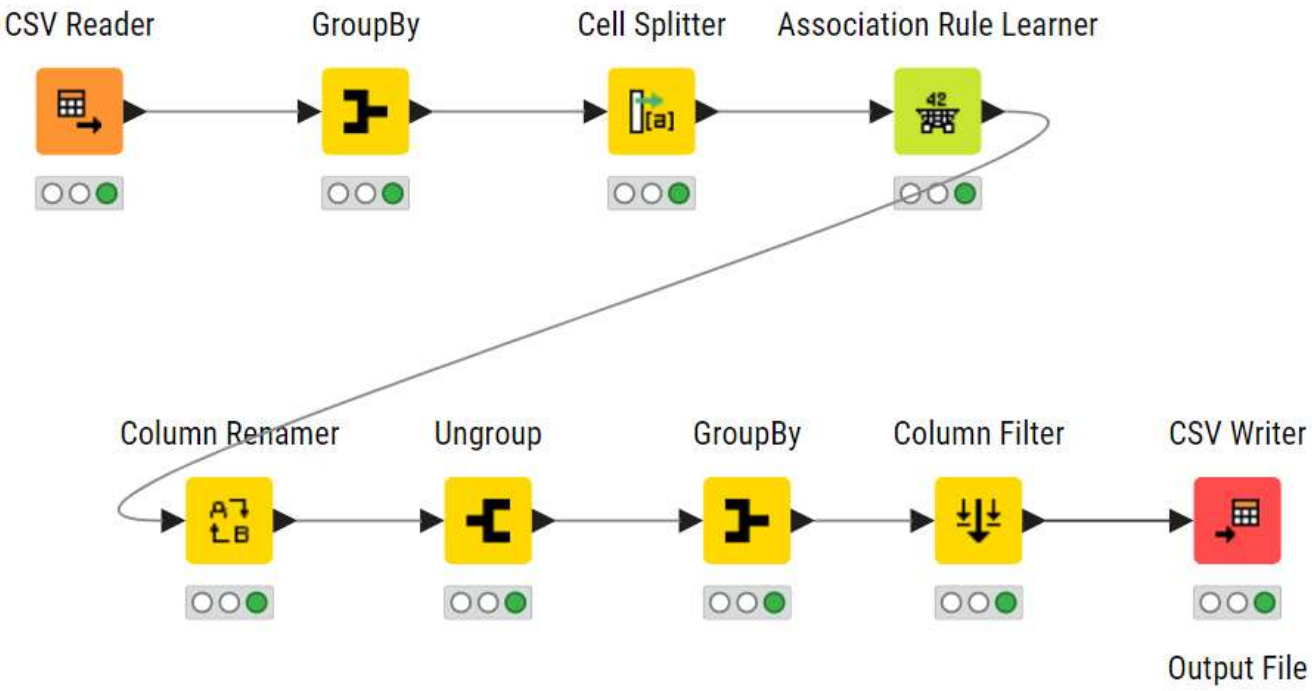
  Confidence:

  - This measures the reliability or strength of the rule.
  - The confidence of an association rule measures the likelihood that the presence of one item in a transaction implies the presence of another item.



- Here, after some trials. we have set a minimum support threshold of **0.05**, it means that an item set must appear in at least 5% of all transactions to be considered frequent.
- And, a minimum confidence of **0.6**, it means that in 60% of cases where the antecedent items are present, the consequent items are also observed in the same transaction.
- Setting a higher minimum support threshold results in fewer item sets being considered frequent, as only those with a high level of support are included in the analysis.
- Setting a higher minimum confidence threshold results in only the most reliable association rules being considered, as rules with lower confidence are filtered out.
- By setting appropriate minimum support and minimum confidence thresholds, we can control the number and quality of association rules generated from the data.
- These thresholds help ensure that the discovered rules are both frequent and reliable, allowing for more meaningful insights into customer behavior and purchasing patterns.

2.7 KNIME Workflow:



2.8 Output Table:

Rows: 24 | Columns: 5

#	RowID	Recommended I... (String)	Support (Number (double))	Confidence (Number (double))	Lift ↓ (Number (double))	Items (#1) (String)
16	Row...	paper towels	0.055	0.649	1.791	eggs, ice cream, pasta
18	Row...	pasta	0.055	0.643	1.731	paper towels, eggs, ice cream
2	Row1	cheeses	0.051	0.674	1.726	bagels, cereals, sandwich bags
13	Row...	juice	0.05	0.64	1.7	yogurt, toilet paper, aluminum foil
15	Row...	mixes	0.051	0.63	1.678	yogurt, poultry, aluminum foil
24	Row...	sandwich bags	0.051	0.611	1.66	cheeses, bagels, cereals
7	Row6	dinner rolls	0.054	0.642	1.651	spaghetti sauce, poultry, laundry detergent
5	Row4	dinner rolls	0.052	0.641	1.649	spaghetti sauce, poultry, ice cream
12	Row...	juice	0.05	0.62	1.645	yogurt, poultry, aluminum foil
20	Row...	poultry	0.052	0.686	1.628	dinner rolls, spaghetti sauce, ice cream
9	Row8	eggs	0.052	0.634	1.627	paper towels, dinner rolls, pasta
17	Row...	pasta	0.052	0.602	1.621	paper towels, eggs, dinner rolls
4	Row3	dinner rolls	0.051	0.63	1.621	spaghetti sauce, poultry, cereals
10	Row9	eggs	0.055	0.63	1.616	paper towels, ice cream, pasta
3	Row2	coffee/tea	0.05	0.613	1.616	yogurt, cheeses, cereals
6	Row5	dinner rolls	0.052	0.628	1.614	spaghetti sauce, poultry, juice
8	Row7	eggs	0.052	0.628	1.61	dinner rolls, poultry, soda
14	Row...	milk	0.051	0.604	1.589	poultry, laundry detergent, cereals
11	Row...	ice cream	0.055	0.624	1.565	paper towels, eggs, pasta
1	Row0	cereals	0.051	0.617	1.558	cheeses, bagels, sandwich bags
22	Row...	poultry	0.054	0.656	1.556	dinner rolls, spaghetti sauce, laundry detergent
19	Row...	poultry	0.051	0.637	1.512	dinner rolls, spaghetti sauce, cereals
21	Row...	poultry	0.052	0.602	1.429	dinner rolls, spaghetti sauce, juice
23	Row...	poultry	0.05	0.6	1.424	dishwashing liquid/detergent, laundry detergent, mix...

2.9 Insights:

- 24 Association rules have been derived from the dataset.
- Lift is a metric in association rule mining that measures the strength of association between items in a rule, helping identify meaningful patterns and relationships in transactional data.
- Higher lift values indicate stronger associations.
- The support values range from approximately 0.05 to 0.055, indicating that the item sets occur in around 5% to 5.5% of transactions.
- The confidence values of the association rules are above 0.6, indicating strong relationships between the antecedent and consequent items.
- The highest lift value of 1.791 suggests that the Paper towels are frequently purchased together with eggs, ice cream, and pasta.
  - Rule: eggs, ice cream, pasta -> paper towels
  - Support: 0.0553 (5.53% of transactions contain eggs, ice cream, pasta, and paper towels)
  - Confidence: 0.649 (64.9% of transactions with eggs, ice cream, and pasta also contain paper towels)
  - Lift: 1.791 (Transactions with eggs, ice cream, and pasta are approximately 1.791 times more likely to contain paper towels compared to the expected likelihood of the items bought independent)

2.10 Recommendation:

- Popular combo offers:

  For example, the association rule "paper towels -> eggs, ice cream, pasta" has a high lift value of 1.791, indicating a strong association between paper towels and the combination of eggs, ice cream, and pasta. The store can create a combo offer or discount for customers purchasing paper towels along with eggs, ice cream, and pasta, to encourage additional purchases and increase revenue.

- Bundle related products together and offer them at a discounted price to encourage customers to purchase multiple items.

  For instance, based on the association rule "cheeses -> bagels, cereals, sandwich bags", the store can create a breakfast bundle including cheeses, bagels, cereals, and sandwich bags, and offer it at a discounted price to attract customers looking for convenient breakfast options.

- Cross-selling opportunity:

  Use association rules to identify cross-selling opportunities where one product purchase can lead to the sale of complementary items. For example, based on the association rule "juice -> yogurt, toilet paper, aluminum foil", the store can promote juice purchases by offering discounts on yogurt, toilet paper, and aluminum foil when purchased together with juice.

- Targeted Promotions:

  Tailor promotions and marketing campaigns based on the insights derived from association rules.

  For instance, if the association rule "dinner rolls -> spaghetti sauce, poultry, laundry detergent" suggests a strong association between dinner rolls and a combination of spaghetti sauce, poultry, and laundry detergent, the store can run targeted promotions on dinner rolls along with these related items to encourage customers to buy them together.

- Place high-demand items and frequently purchased combinations in prominent locations like near the counter to increase visibility and encourage impulse purchases.

- Utilize customer transaction data to provide personalized recommendations and offers based on individual purchase history and preferences. This can enhance customer satisfaction and increase repeat purchases.

- Monitor the trends in product preferences over time to anticipate changes in consumer behavior and adjust inventory and marketing strategies accordingly.

- Regularly review and update promotions, discounts, and product offerings to keep them relevant and appealing to customers.

- Tailor promotions and discounts based on seasonal trends and quarterly sales patterns. For example, offer promotions during the summer months or during festive holidays.

- Capitalize on the higher order counts observed on weekends by offering special weekend promotions or deals to attract more customers.

- Implement weekday-specific promotions or discounts to drive traffic during slower periods and increase sales on weekdays.

- Continuously monitor sales data, customer feedback, and market trends to identify opportunities for improvement and adapt strategies accordingly.

- Continuously monitor the performance of combo offers and discounts to assess their effectiveness in driving sales and revenue.

- Regularly update and rotate combo offers to keep customers interested and coming back to explore new deals and promotions.

- Analyze customer feedback and purchasing patterns to refine and adjust combo offers over time for maximum impact.

- By implementing these strategies , the grocery store can effectively increase its revenue, enhance customer satisfaction, and drive business growth.

2.11 DATASET : dataset_group.csv

Data Dictionary:

Column Name	Description
Order_id	This column represents the unique identification number assigned to each order.
Date	It denotes the date on which the order was placed.
Product	This column represents the product purchased.

3. OUTPUT FILES:

- PART A: Sales Output.xlsx
- PART B: Product output.csv

THANK YOU