MRA Project - Milestone 1

Name: G. Umamakeshwari
PGPDSBA - April' B21

Agenda

Objective of this project is to find the underlying buying patterns of the customers of an automobile part manufacturer based on the past 3 years of the Company's transaction data and hence recommend customized marketing strategies for different segments of customers.

- ## Executive Summary of the data

We have received the 3 years data of automobile parts manufacturing company. It containing 2747 transactions with 20 variables regarding the orders of the product and customer information.

Contents of the presentation

- Problem Statement.
- About Data
  - Sample data
  - Shape of data
  - Info of data
  - Data Summary
- Exploratory Analysis and Inferences
  - Univariate Analysis – Numerical variable
  - Univariate Analysis – Categorical Variable
  - Bivariate Analysis
  - Multivariate Analysis
- Customer Segmentation using RFM analysis.
  - Workflow image of KNIME
  - Output table head
- Inferences from RFM Analysis and identified segments
  - Who are your best customers?
  - Which customers are on the verge of churning?
  - Who are your lost customers?
  - Who are your loyal customers?
  - Summary

- # Problem statement

An automobile parts manufacturing company has collected data of 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 magical data science skills to provide them with suitable insights about their data and their customers.

Data Dictionary:			
ORDERNUMBER :	Order Number	**PRODUCTCODE :**	Code of Product
QUANTITYORDERED :	Quantity ordered	**CUSTOMERNAME :**	customer
PRICEEACH :	Price of Each item	**PHONE :**	Phone of the customer
ORDERLINENUMBER :	order line	**ADDRESSLINE1 :**	Address of customer
SALES :	Sales amount	**CITY :**	City of customer
ORDERDATE :	Order Date	**POSTALCODE :**	Postal Code of customer
DAYS_SINCE_LASTORDER :	Days_ Since_Lastorder	**COUNTRY :**	Country customer
STATUS :	Status of order like Shipped or not	**CONTACTLASTNAME :**	Contact person customer
PRODUCTLINE :	Product line – CATEGORY	**CONTACTFIRSTNAME :**	Contact person customer
MSRP :	Manufacturer's Suggested Retail Price	**DEALSIZE :**	Size of the deal based on Quantity and Item Price

- **About Data** (Info, Shape, Summary Stats, your assumptions about data)

Sample data

	ORDERNUMBER	QUANTITYORDERED	PRICEEACH	ORDERLINENUMBER	SALES	ORDERDATE	DAYS_SINCE_LASTORDER	STATUS	PRODUCTLINE	MSRP
0	10107	30	95.70	2	2871.00	2018-02-24	828	Shipped	Motorcycles	95
1	10121	34	81.35	5	2765.90	2018-05-07	757	Shipped	Motorcycles	95
2	10134	41	94.74	2	3884.34	2018-07-01	703	Shipped	Motorcycles	95
3	10145	45	83.26	6	3746.70	2018-08-25	649	Shipped	Motorcycles	95
4	10168	36	96.66	1	3479.76	2018-10-28	586	Shipped	Motorcycles	95

Shape of data

- There are 20 variables available regarding the orders of the product and customer information with 2747 records.

Info of data

- The data has 1 datetime, 2 float, 5 int, and 12 Object data types variables. There is no missing values present in the data set.

```
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
```

Data Summary

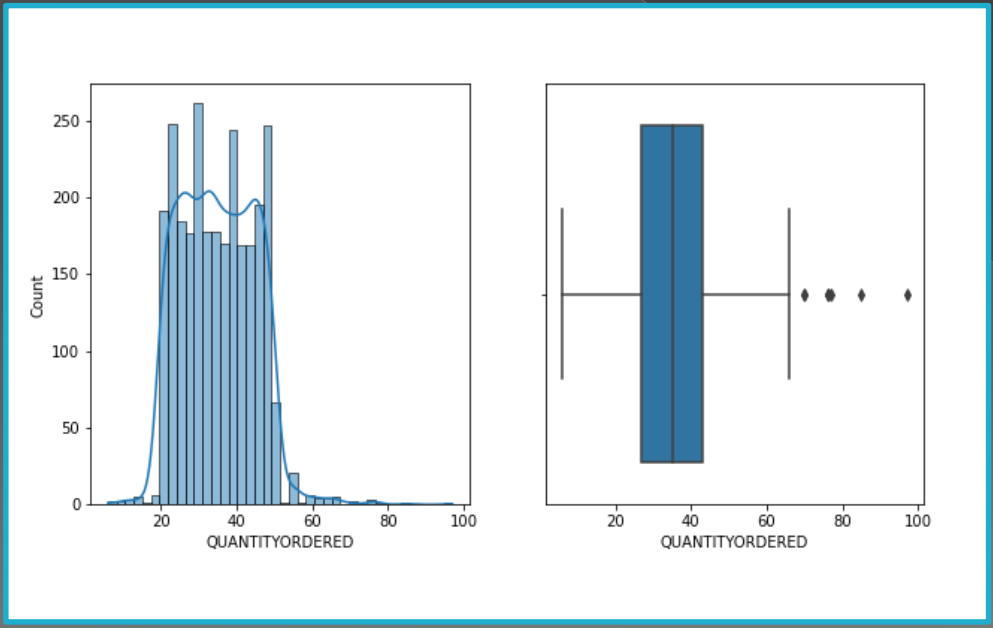
	QUANTITYORDERED	PRICEEACH	SALES	DAYS_SINCE_LASTORDER	MSRP
count	2747.000000	2747.000000	2747.000000	2747.000000	2747.000000
mean	35.103021	101.098951	3553.047583	1757.085912	100.691664
std	9.762135	42.042548	1838.953901	819.280576	40.114802
min	6.000000	26.880000	482.130000	42.000000	33.000000
25%	27.000000	68.745000	2204.350000	1077.000000	68.000000
50%	35.000000	95.550000	3184.800000	1761.000000	99.000000
75%	43.000000	127.100000	4503.095000	2436.500000	124.000000
max	97.000000	252.870000	14082.800000	3562.000000	214.000000

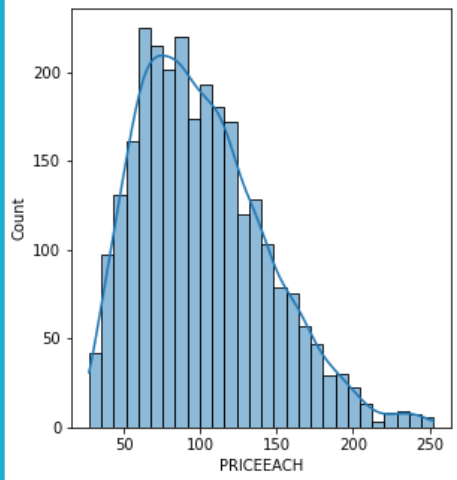
- The data has 1 datetime, 2 float, 5 int, and 12 Object data types variables. There is no missing values present in the data set. Here 2747 records available without any missing values with 20 variables.
- The company is into automobile part manufacture, and they have different product line like Classic car , Motorcycle, plane, train, ship, Bus truck, vintage cars etc
- Here 5 numerical variables are described with count, mean, std, min, max and percentile details. Other variables are categorical or numerical but not include for describe function.
- The data maintained each transactions entry as order number and for each order number maintained all required information like customer identity details , and product details like price , quantity , product code, and sales for each customer.
- We noticed that one order number has many different entries with different product codes.

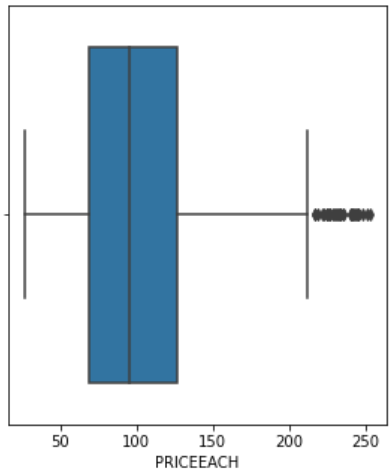
Exploratory Analysis and Inferences

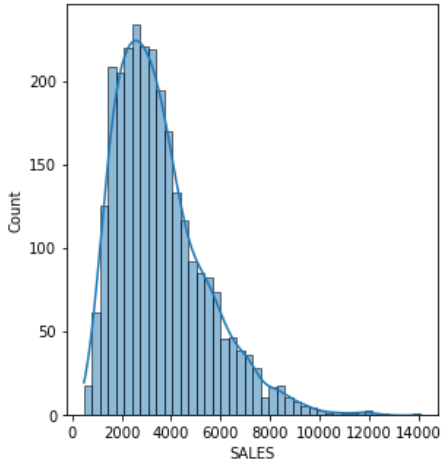
- Univariate, Bivariate, and multivariate analysis using data visualization
  - Weekly, Monthly, Quarterly, Yearly Trends in Sales
  - Sales Across different Categories of different features in the given data

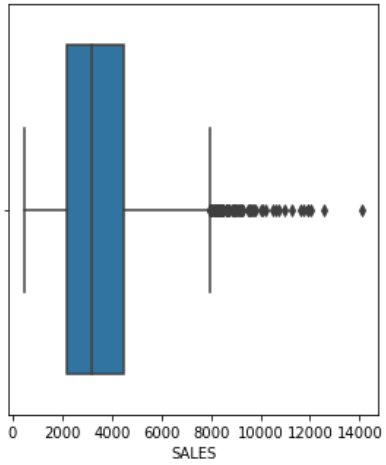
Univariate Analysis – Numerical variable

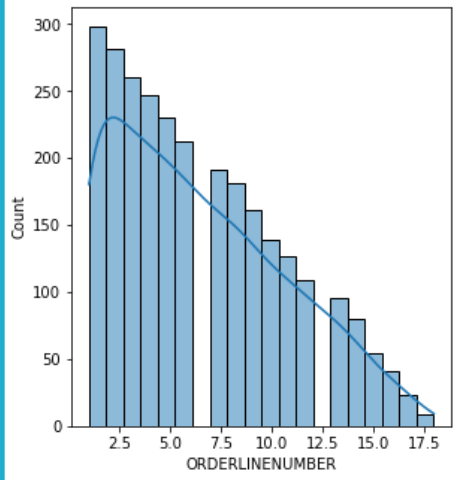


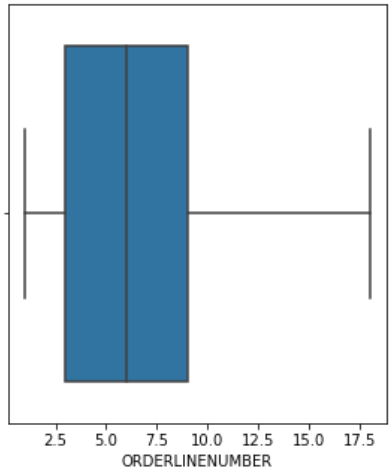


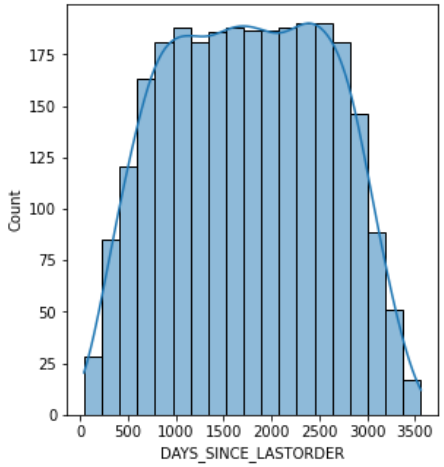


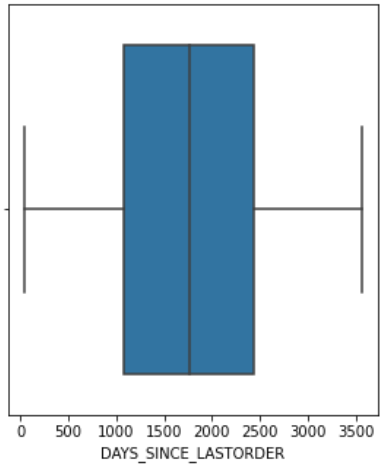








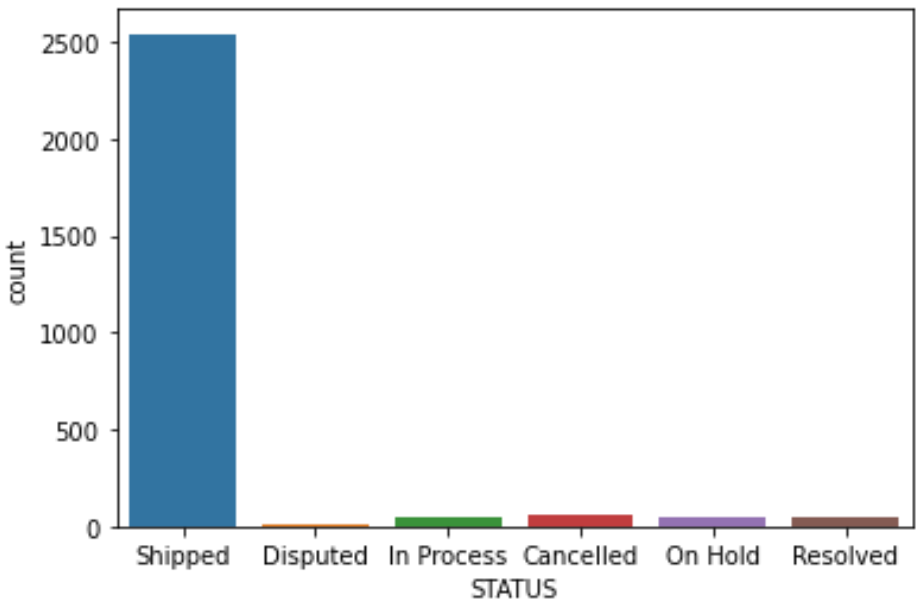


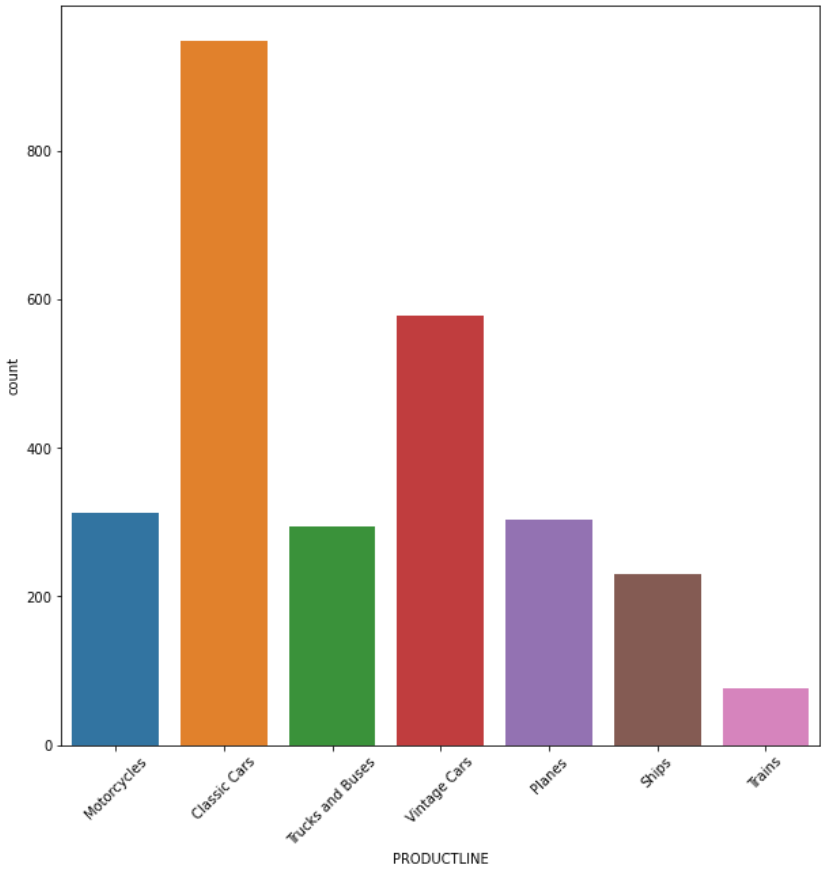


Inference:

- Quantity, Price and Sales have outliers in dataset. All fields are have different data pattern waves.

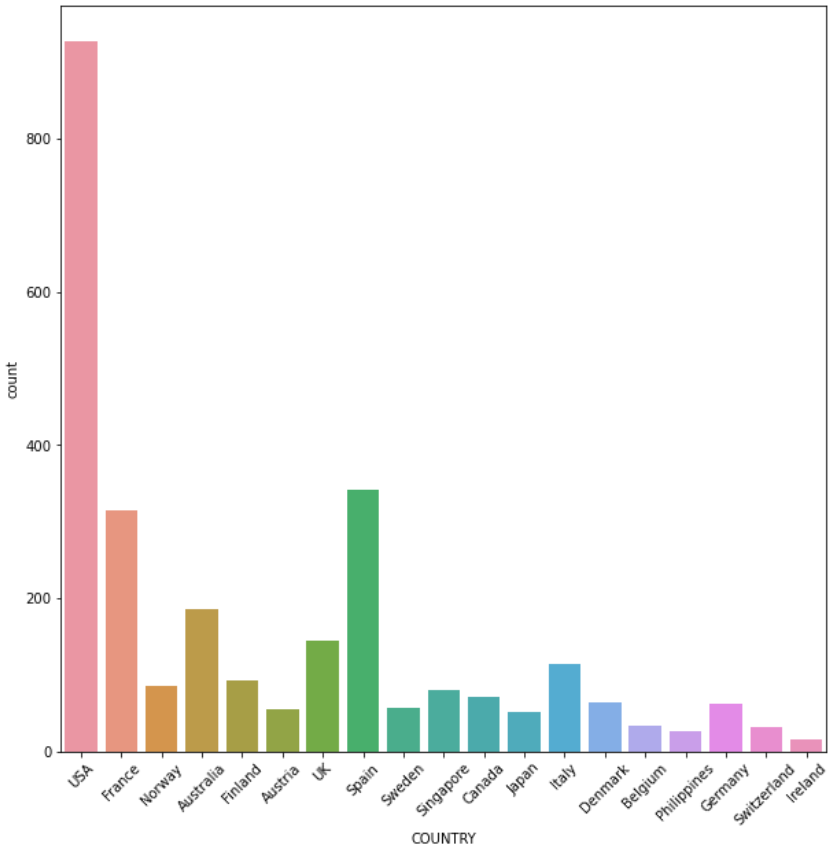
Univariate Analysis – Categorical Variable

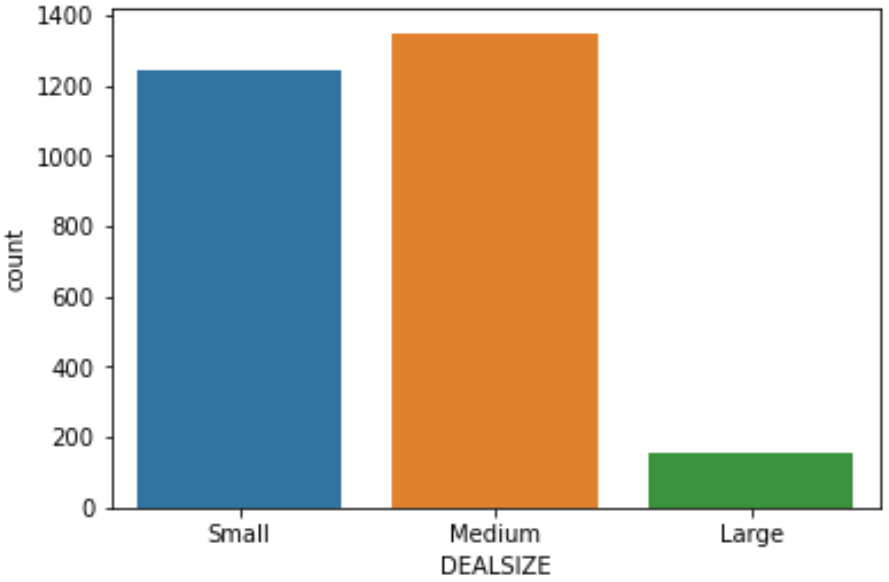




Inference:

- As per the plot ‘Shipped’ status have high records and related to product ‘Classic cars’ after that ‘Vintage cars’ have high.

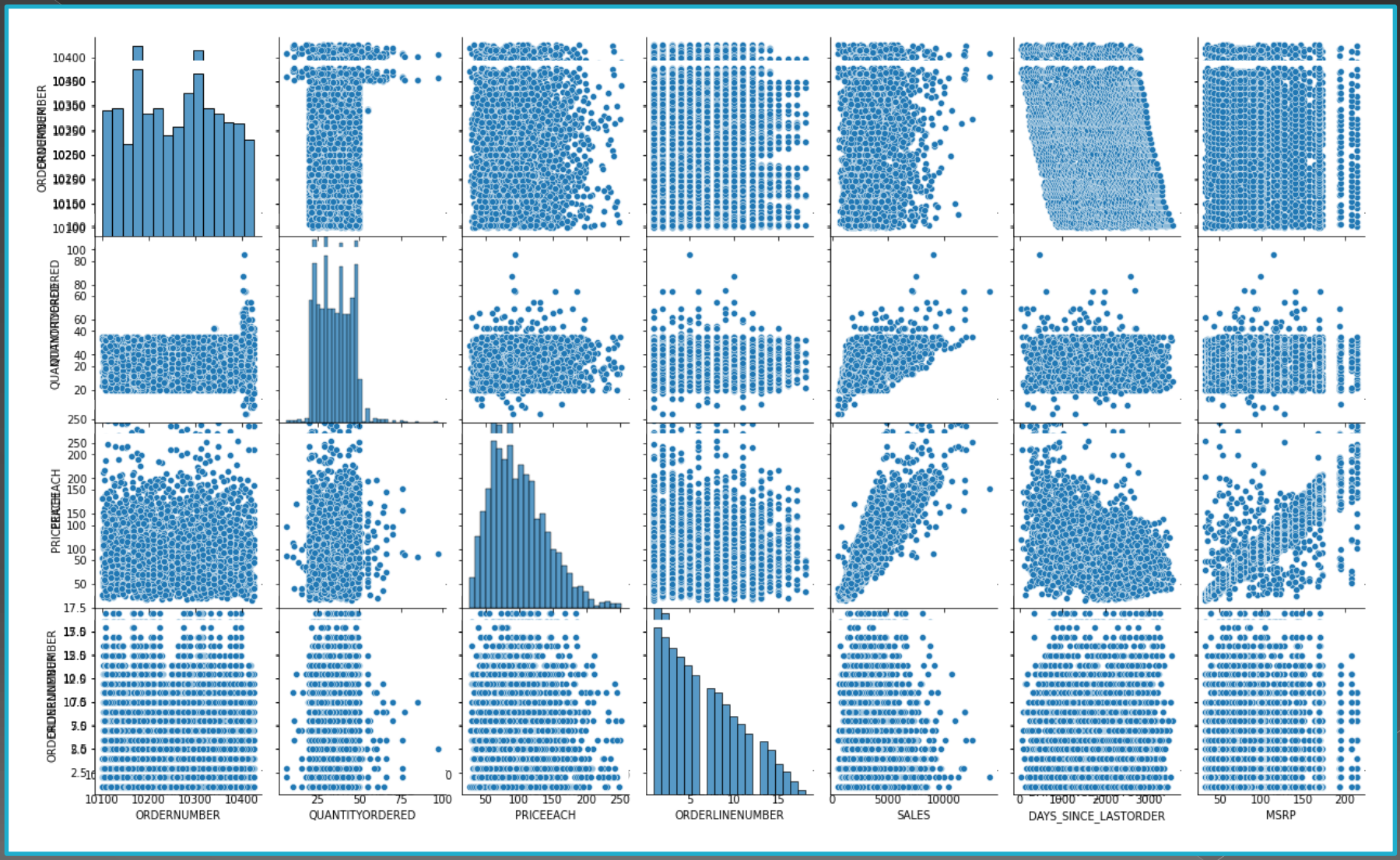




Inference:

- As per the plot ‘USA’ country have high records and related to Deal size ‘Medium’ have high.

Bivariate Analysis



Inference:

- As per the plot all variables are
- Quantity, Price and Sales have outliers in dataset. All fields are have different data pattern waves.
- As per scatter plot all variables are in fully scatter related to each other variables

Multivariate Analysis

	ORDERNUMBER	QUANTITYORDERED	PRICEEACH	ORDERLINENUMBER	SALES	DAYS_SINCE_LASTORDER	MSRP
ORDERNUMBER	1.000000	0.067110	-0.003369	-0.054300	0.037289	-0.251476	-0.013910
QUANTITYORDERED	0.067110	1.000000	0.010161	-0.016295	0.553359	-0.021923	0.020551
PRICEEACH	-0.003369	0.010161	1.000000	-0.052646	0.808287	-0.397092	0.778393
ORDERLINENUMBER	-0.054300	-0.016295	-0.052646	1.000000	-0.057414	0.046615	-0.020956
SALES	0.037289	0.553359	0.808287	-0.057414	1.000000	-0.334274	0.634849
DAYS_SINCE_LASTORDER	-0.251476	-0.021923	-0.397092	0.046615	-0.334274	1.000000	-0.524285
MSRP	-0.013910	0.020551	0.778393	-0.020956	0.634849	-0.524285	1.000000

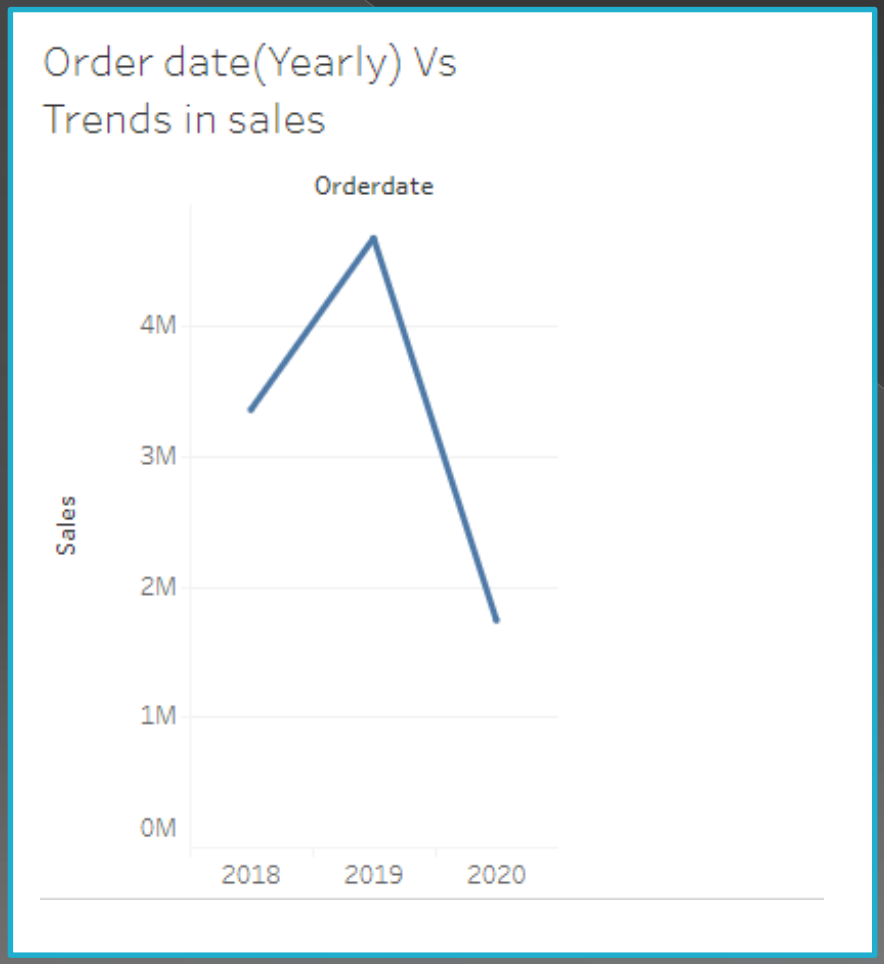
Inference:

- As per the heat map, multi collinearity present between 3 variables in the dataset. High correlation between Price each and sales, price each and MSRP, Sales and MSRP, Quantity ordered and sales.


ORDERNUMBER
QUANTITYORDERED
PRICEEACH
ORDERLINENUMBER
SALES
DAYS_SINCE_LASTORDER
MSRP
1.00 0.07 -0.00 -0.05 0.04 -0.25 -0.01
0.07 1.00 0.01 -0.02 0.55 -0.02 0.02
-0.00 0.01 1.00 -0.05 0.81 -0.40 0.78
-0.05 -0.02 -0.05 1.00 -0.06 0.05 -0.02
0.04 0.55 0.81 -0.06 1.00 -0.33 0.63
-0.25 -0.02 -0.40 0.05 -0.33 1.00 -0.52
-0.01 0.02 0.78 -0.02 0.63 -0.52 1.00
ORDERNUMBER
QUANTITYORDERED
PRICEEACH
ORDERLINENUMBER
SALES
DAYS_SINCE_LASTORDER
MSRP
1.0
0.8
0.6
0.4
0.2
0.0
−0.2
−0.4

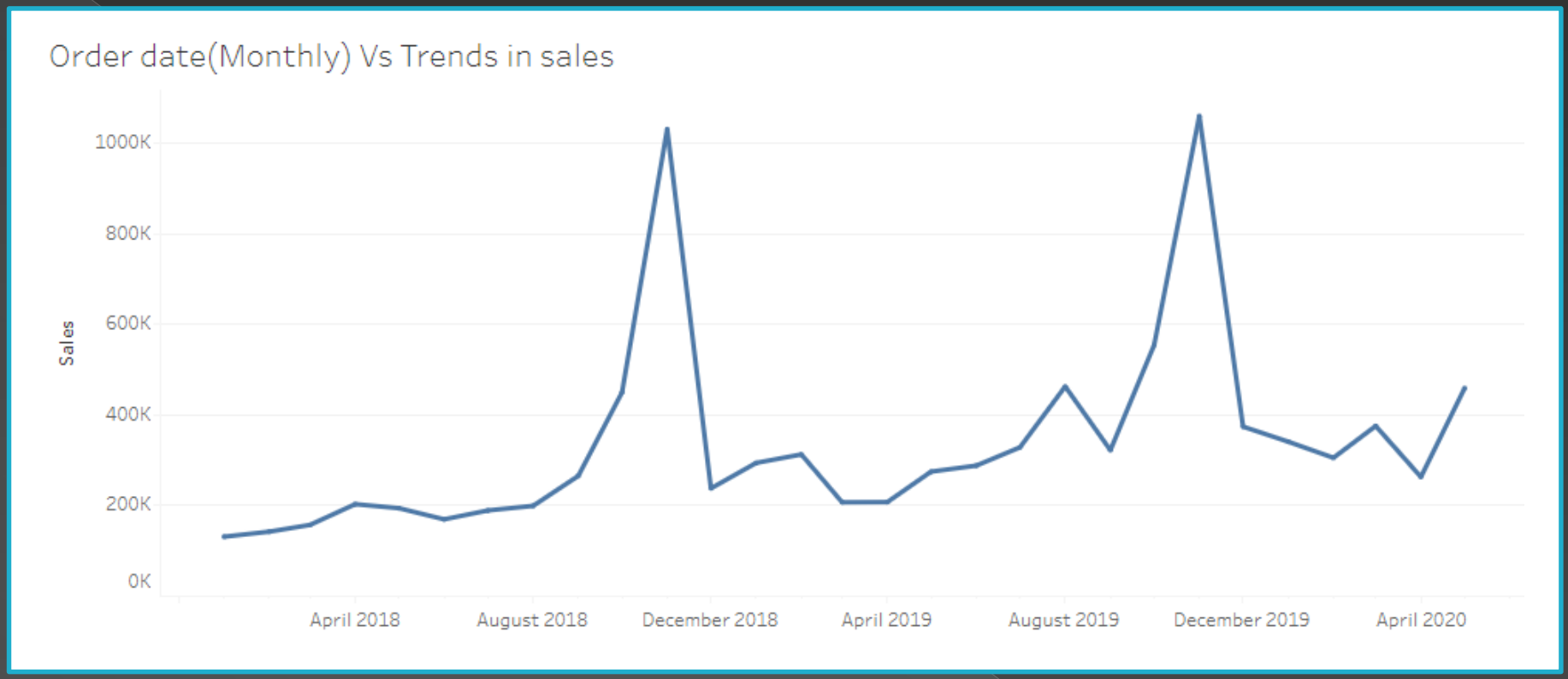
Time series & Trends in Sales

- Weekly, Monthly, Quarterly, Yearly Trends in Sales



Inference:

- As per Yearly order date Vs Sales, we only have 3 years data. Here 2018 to 2019 sales has increase but 2019 to 2020 sales has decreased.



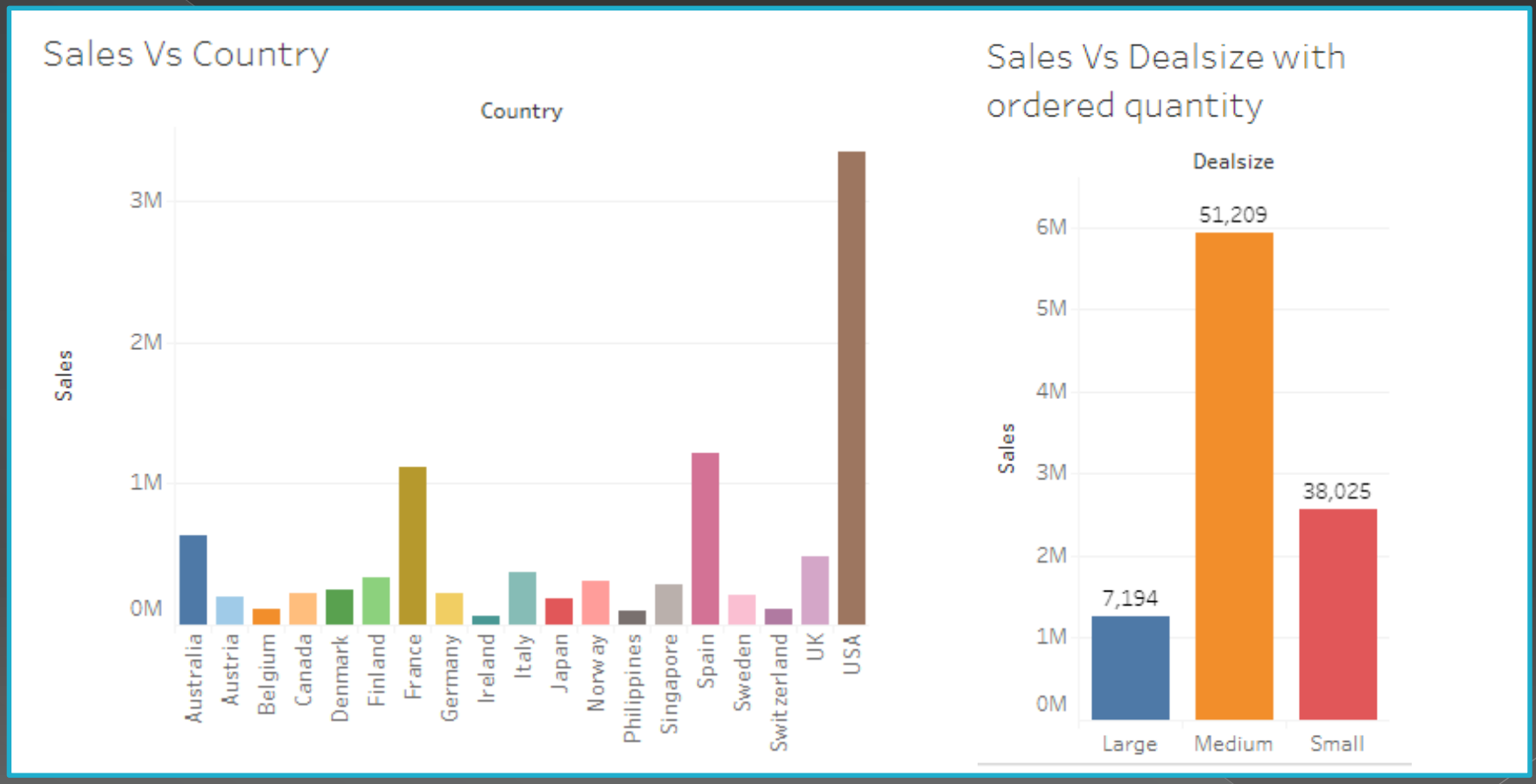
Inference:

- As per monthly order date Vs Sales, we can see there is a small increasing trend with 2 seasonality patterns from April to next year April.

Inference:

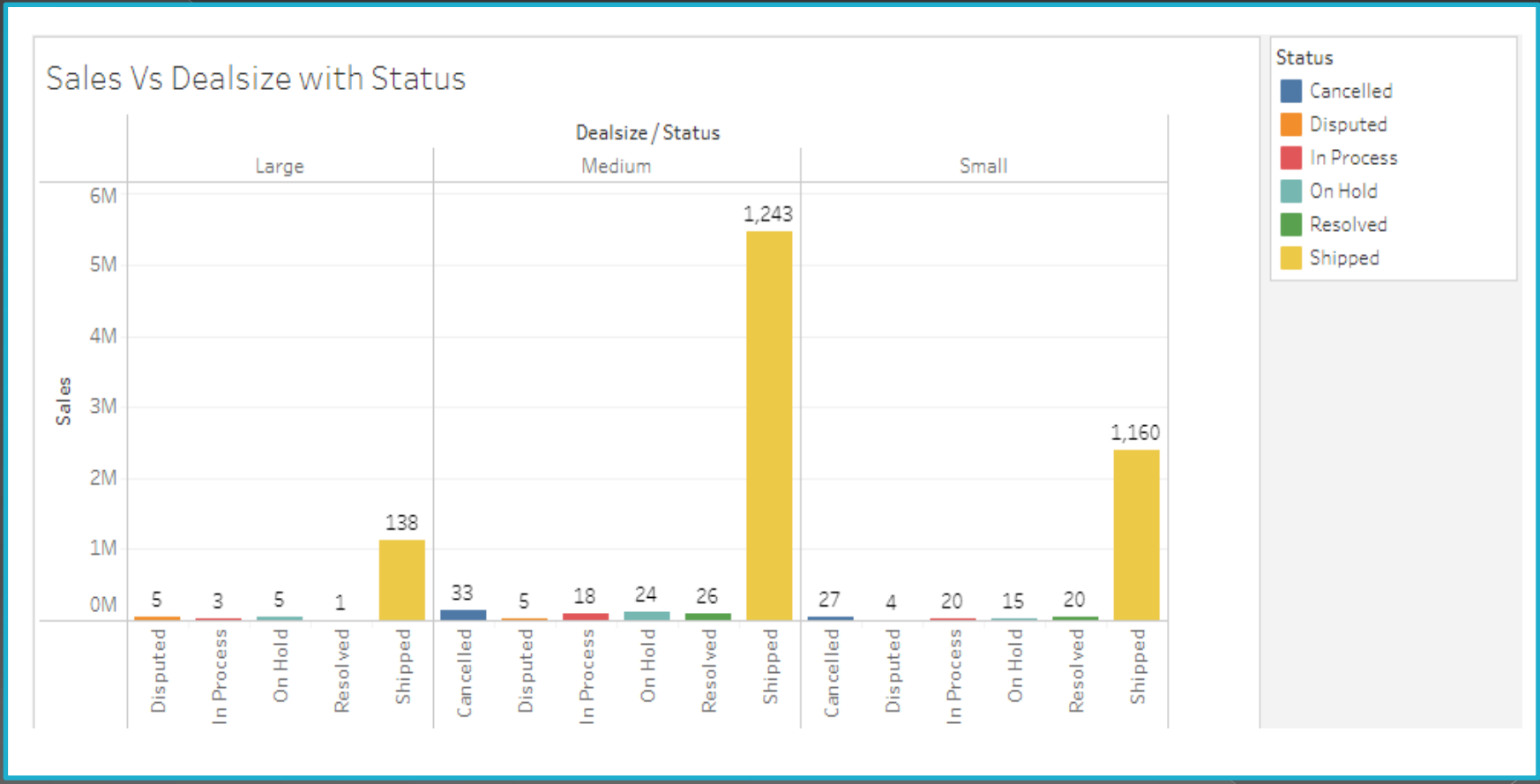
- As per quarterly order date Vs Sales, we can see there is a small increasing trend with 2 seasonality patterns from Q1 to next Q1.

Sales Across different Categories of different features in the given data



Inference:

- USA have more sales value compared with other countries. Related to deal size, medium size deals have high sales value with highest ordered quantity.



Inference:

- Shipped status order sales are high in all type of deal size. In this Shipped status in medium size deals have more sales value compared with large and small size of deals.

Customer Segmentation using RFM analysis

- **Which tool used?**

  I have used KNIME Tool for RFM analysis

- **What all parameters used and assumptions made?**

  1. I have assumed "01-06-2020" as Max(order date) to find the **Recency**.

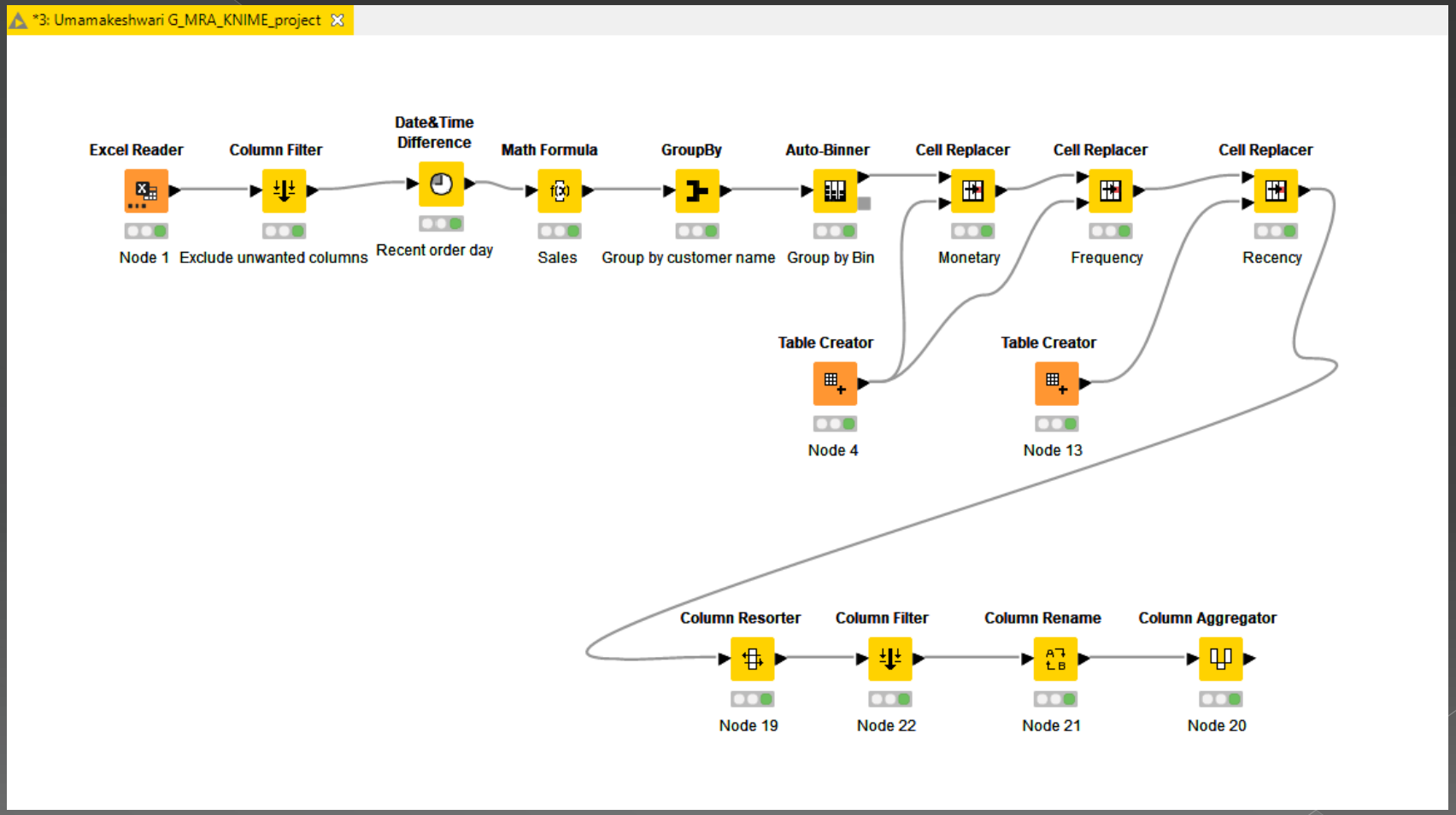
     Recency days= [Max(order date) - order date)] and take a **minimum of Recency days** for each customer as **Recency**.

  2. I have found that same customer have multiple order with multiple products. I have calculated the **Frequency** based on unique customer name with **no.of unique order numbers** for each customer.

  3. Sales column is already available for each products based on orders. I have calculated **Monetary** by **sum of sales for each customers**.

  Then created four different bin for each Recency, Frequency & Monetary using percentile range(0, 0.25, 0.50, 0.75, 100). Based on above 4 bin assumption I have considered 4 segments.

- # Workflow image of KNIME



- **Output table head**

File Edit Hilite Navigation View

Table "default" - Rows: 89 | Spec - Columns: 8 | Properties | Flow Variables

Row ID	S CUSTOMERNAME	I ORDERFREQUENCY	L RECENTORDERDAYS	D TOTALORDERVALUE	S MONETARY	S FREQUENCY	S RECENCY	S Concatenate
Row0	AV Stores, Co.	3	197	157,807.81	1	1	2	112
Row1	Alpha Cognac	3	65	70,488.44	1	1	3	113
Row2	Amica Models & Co.	2	266	94,117.26	1	1	2	112
Row3	Anna's Decorations, Ltd	4	84	153,996.13	1	1	3	113
Row4	Atelier graphique	3	189	24,179.96	1	1	2	112
Row5	Australian Collectables, Ltd	3	23	64,591.46	1	1	3	113
Row6	Australian Collectors, Co.	5	185	200,995.41	2	1	2	212
Row7	Australian Gift Network, Co	3	120	59,469.12	1	1	3	113
Row8	Auto Assoc. & Cie.	2	234	64,834.32	1	1	2	112
Row9	Auto Canal Petit	3	55	93,170.66	1	1	3	113
Row10	Auto-Moto Classics Inc.	3	181	26,479.26	1	1	2	112

Inferences from RFM Analysis and identified segments

- **Who are your best customers?**

CUTOMER NAME	ORDER FREQUENCY	RECENT ORDER DAYS	TOTAL ORDER VALUE	RECENCY	FREQUENCY	MONETARY	CONCATENATE
Danish Wholesale Imports	5	47	145041.6	4	4	4	444
Diecast Classics Inc.	4	2	122138.14	4	4	4	444
Euro Shopping Channel	26	1	912294.11	4	4	4	444
La Rochelle Gifts	4	1	180124.9	4	4	4	444
Mini Gifts Distributors Ltd.	17	3	654858.06	4	4	4	444

On basis on Recency, frequency & monetary I have grouped best customers. In according to RFM model the most importance is given to Recency. So, that I have kept it as first parameter for selecting best customers and then frequency order after that Monetary order.

For instance, Customer name - Danish Wholesale Imports, have recently made a purchase and have high frequency with a high monetary.

Best customers are,

1. **Danish Wholesale Imports**
2. **Diecast Classics Inc.**
3. **Euro Shopping Channel**
4. **La Rochelle Gifts**
5. **Mini Gifts Distributors Ltd.**

Which customers are on the verge of churning?

CUTOMER NAME	ORDER FREQUENCY	RECENT ORDER DAYS	TOTAL ORDER VALUE	RECENCY	FREQUENCY	MONETARY	CONCATENATE
Saveley & Henriot, Co.	3	457	142874.25	1	2	4	124
Herkku Gifts	3	272	111640.28	1	2	3	123
Amica Models & Co.	2	266	94117.26	1	1	3	113
Marta's Replicas Co.	2	232	103080.38	1	1	3	113
Vida Sport, Ltd	2	276	117713.56	1	1	3	113

On basis on Recency, frequency & monetary I have grouped customers who are on verge of churning. In according to RFM model the most importance is given to Recency. So, that I have kept it as first parameter and then frequency order after that Monetary order.

In this case customer on verge of churning is customer's have high recency days(Low recency) but high in Frequency and Monetary. We should definitely focus on this group before we lose them and try to convert them into regular customers by giving some offers.

For instance, Customer name - Saveley & Henriot, Co., have high frequency with a high monetary but recency is low.

Customers on verge of churning are,

1. **Saveley & Henriot, Co.**
2. **Herkku Gifts**
3. **Amica Models & Co.**
4. **Marta's Replicas Co.**
5. **Vida Sport, Ltd**

Who are your lost customers?

CUTOMER NAME	ORDER FREQUENCY	RECENT ORDER DAYS	TOTAL ORDER VALUE	RECENCY	FREQUENCY	MONETARY	CONCATENATE
Auto Assoc. & Cie.	2	234	64834.32	1	1	1	111
Bavarian Collectables Imports, Co.	1	260	34993.92	1	1	1	111
CAF Imports	2	440	49642.05	1	1	1	111
Cambridge Collectables Co.	2	390	36163.62	1	1	1	111
Clover Collections, Co.	2	259	57756.43	1	1	1	111

On basis on Recency, frequency & monetary I have grouped Lost customers. In according to RFM model the most importance is given to Recency. So, that I have kept it as first parameter and then frequency order after that Monetary order in low to high order.

In this case customer have low Recency, low Frequency and low Monetary. We can collect a reviews and comments based on recent orders and make a best possible offers to them to bring them back as our customers.

For instance, Customer name - Auto Assoc. & Cie., have low Recency, Frequency and Low Monetary.

Lost customers are,

1. **Auto Assoc. & Cie.**
2. **Bavarian Collectables Imports, Co.**
3. **CAF Imports**
4. **Cambridge Collectables Co.**
5. **Clover Collections, Co.**

Who are your loyal customers?

CUTOMER NAME	ORDER FREQUENCY	RECENT ORDER DAYS	TOTAL ORDER VALUE	RECENCY	FREQUENCY	MONETARY	CONCATENATE
Danish Wholesale Imports	5	47	145041.6	4	4	4	444
Diecast Classics Inc.	4	2	122138.14	4	4	4	444
Euro Shopping Channel	26	1	912294.11	4	4	4	444
Handji Gifts& Co	4	39	115498.73	4	4	3	443
La Rochelle Gifts	4	1	180124.9	4	4	4	444

On basis on Recency, frequency & monetary I have grouped loyal customers. In according to RFM model the most importance is given to Frequency to find loyal customer. So, that I have kept it as first parameter and then Recency order. But, Monetary is not so much important to know the loyal customer.

In this case customer have high Frequency and then high Recency and not consider Monetary. We can give some offers to these customer to get high Monetary value.

For instance, Customer name - Danish Wholesale Imports., have high frequency with a high Recency.

Customers on verge of churning are,

1. **Danish Wholesale Imports**
2. **Diecast Classics Inc.**
3. **Euro Shopping Channel**
4. **Handji Gifts& Co**
5. **La Rochelle Gifts**

Summary

- Recency, Frequency and Monetary are the parameters mostly used for Marketing Retail Analysis.
- Using Recency, frequency & monetary parameters we have grouped our Best, Loyal, on the verge of churning and Lost customers. Customers with good RFM(High Recency, Frequency, Monetary) are the Best customers and customers who have low RFM are Lost customer in the list.
- Customer on verge of churning is customer's have high recency days(Low recency) but high in Frequency and Monetary. We should definitely focus on this group before we lose them and try to convert them into regular customers by giving some offers.
- For lost customer, they have low Recency, low Frequency and low Monetary. We can collect a reviews and comments based on recent orders and make a best possible offers to them to bring them back as our customers.
- Loyal customers high Frequency and then high Recency without Monetary consideration. We can give some offers to these customer to get high Monetary value.