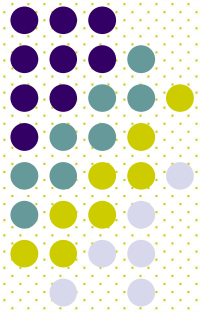
Marketing and Retail Analysis

Vikash Kumar

Agenda & Executive Summary of the data



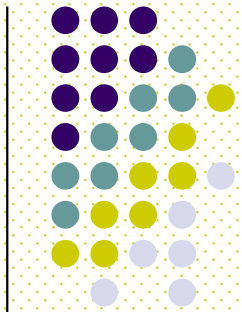
Agenda

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.

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

Contents of the presentation



- Executive Summary of the data
- Exploratory Analysis and Inferences
- Customer Segmentation using RFM analysis
- Inferences from RFM Analysis and identified segments

Executive Summary of the data

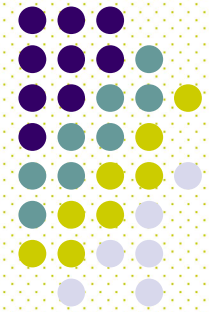
```
Data columns (total 20 columns):
 #   Column                Non-Null Count  Dtype
---  ------                --------------  -----
 0   ORDERNUMBER           2747 non-null   int64
 1   QUANTITYORDERED       2747 non-null   int64
 2   PRICEEACH             2747 non-null   float64
 3   ORDERLINENUMBER       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
```

Data Info, Null, Duplicated, Description summary

- The shape of Sales TS data is: (2747 Rows, 20 columns)
- Duplicate in Sales TS data is: 0
- The Null in Sales TS data is : 0
- Data Found clean

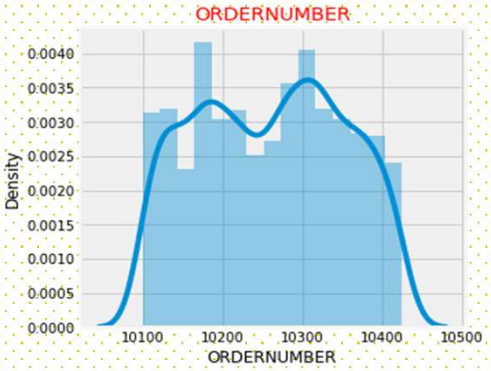
Data Description							
	ORDERNUMBER	QUANTITY ORDERED	PRICEEACH	ORDERLINENUMBER	SALES	DAYS_SINCE_LASTORDER	MSRP
count	2747	2747	2747	2747	2747	2747	2747
mean	10259.76156	35.103021	101.098951	6.491081	3553.047583	1757.085912	100.6917
std	91.877521	9.762135	42.042548	4.230544	1838.953901	819.280576	40.1148
min	10100	6	26.88	1	482.13	42	33
25%	10181	27	68.745	3	2204.35	1077	68
50%	10264	35	95.55	6	3184.8	1761	99
75%	10334.5	43	127.1	9	4503.095	2436.5	124
max	10425	97	252.87	18	14082.8	3562	214

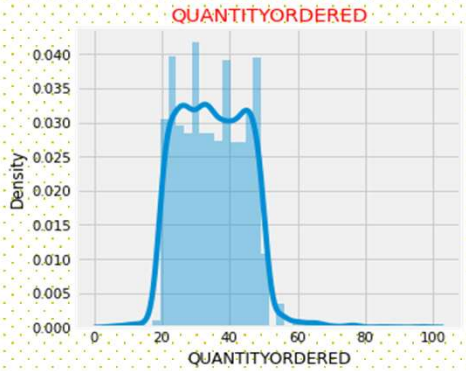
Data Insights:

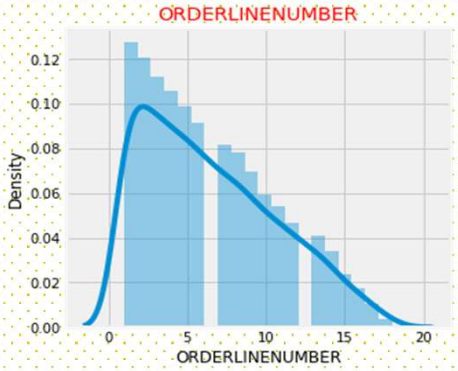


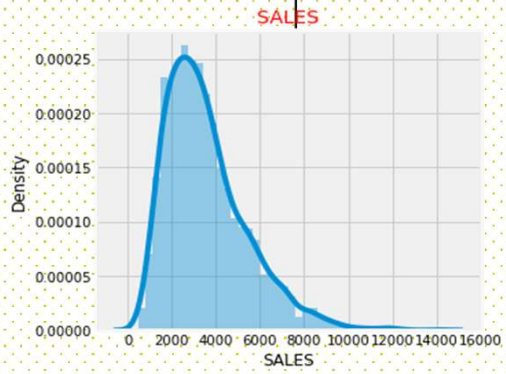
- The average order number is 10259, with a relatively small standard deviation of 92 (Aprox), suggesting that most orders are fairly close in number to each other.
- The average quantity ordered is 35, with a standard deviation of 9.76, indicating that the quantity ordered varies more widely than the order number.
- The average price of each item ordered is 101, with a standard deviation of 42. This suggests that there is a fair amount of variation in the price of items ordered, with some orders containing more expensive items than others.
- The average sales amount per order is 3553, with a standard deviation of 1839. This indicates that there is a wide range of sales amounts, with some orders generating significantly more revenue than others.
- The average time since the last order is 1757 days, with a standard deviation of 819 days. This suggests that there is a fair amount of variability in the time between orders, with some customers ordering frequently and others ordering infrequently.
- The average MSRP (Manufacturer's Suggested Retail Price) of the items ordered is 100, with a standard deviation of 40. This indicates that there is some variability in the price of the items relative to their MSRP, with some items being priced closer to their MSRP than others.
- Overall, these insights provide a basic understanding of the central tendencies and variabilities of the different variables. However, further analysis would be needed to determine any patterns or relationships between the variables and to draw more conclusive insights.

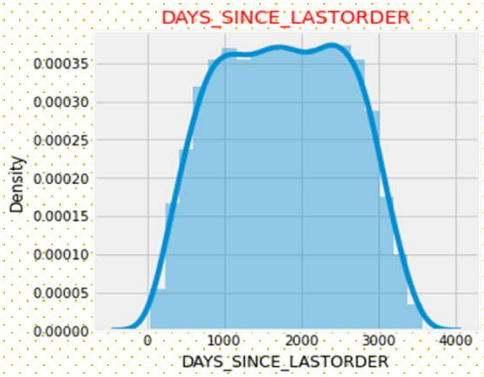
Data Distribution

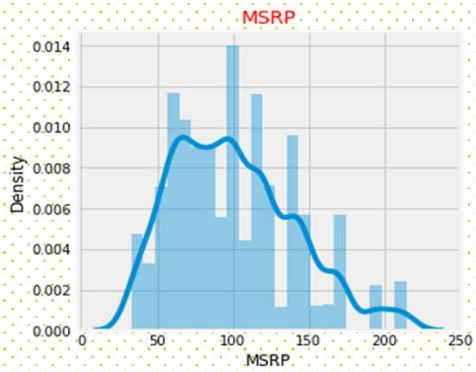












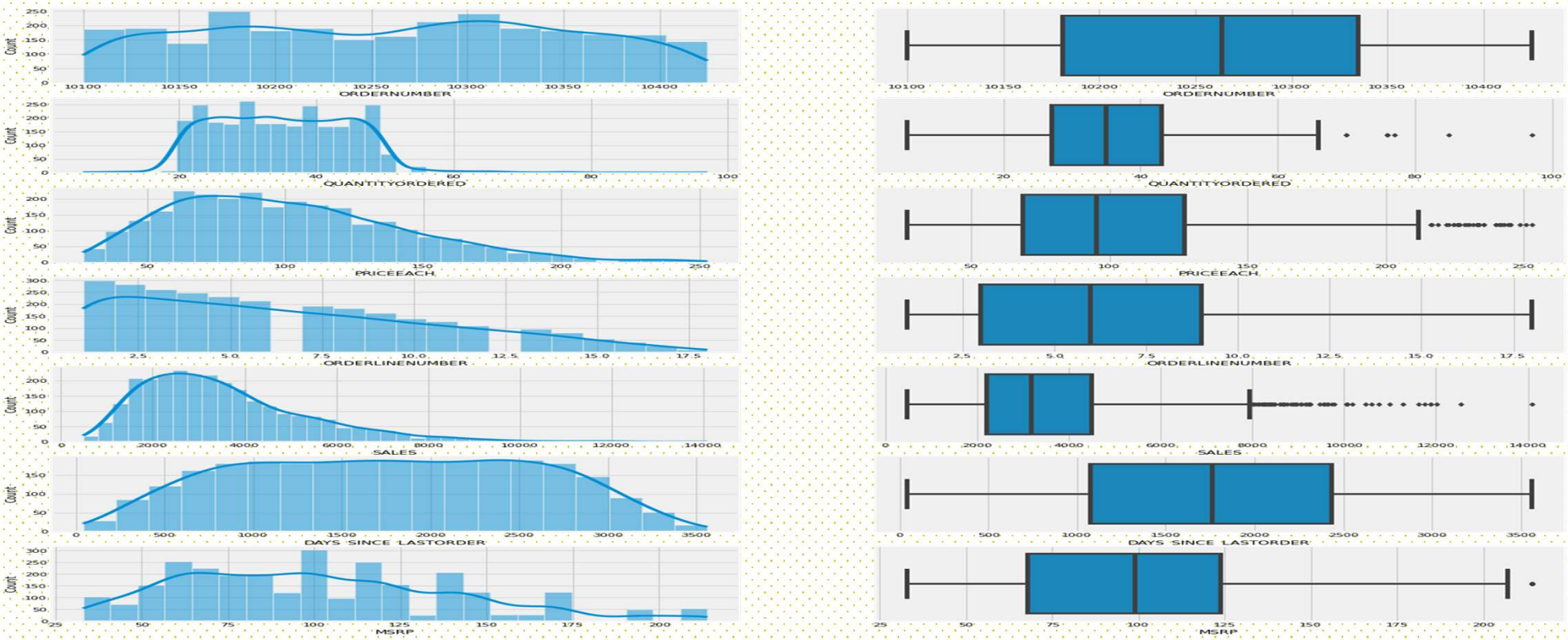
SKEWNESS of DATA SET

ORDERNUMBER	-0.006995
QUANTITYORDERED	0.369286
PRICEEACH	0.697222
ORDERLINENUMBER	0.575327
SALES	1.155940
DAYS_SINCE_LASTORDER	-0.002983
MSRP	0.575646

As summary, the variables QUANTITYORDERED, PRICEEACH, ORDERLINENUMBER, and MSRP all exhibit some degree of right-skewness, while SALES exhibits a highly right-skewed distribution. ORDERNUMBER and DAYS_SINCE_LASTORDER are approximately symmetrical

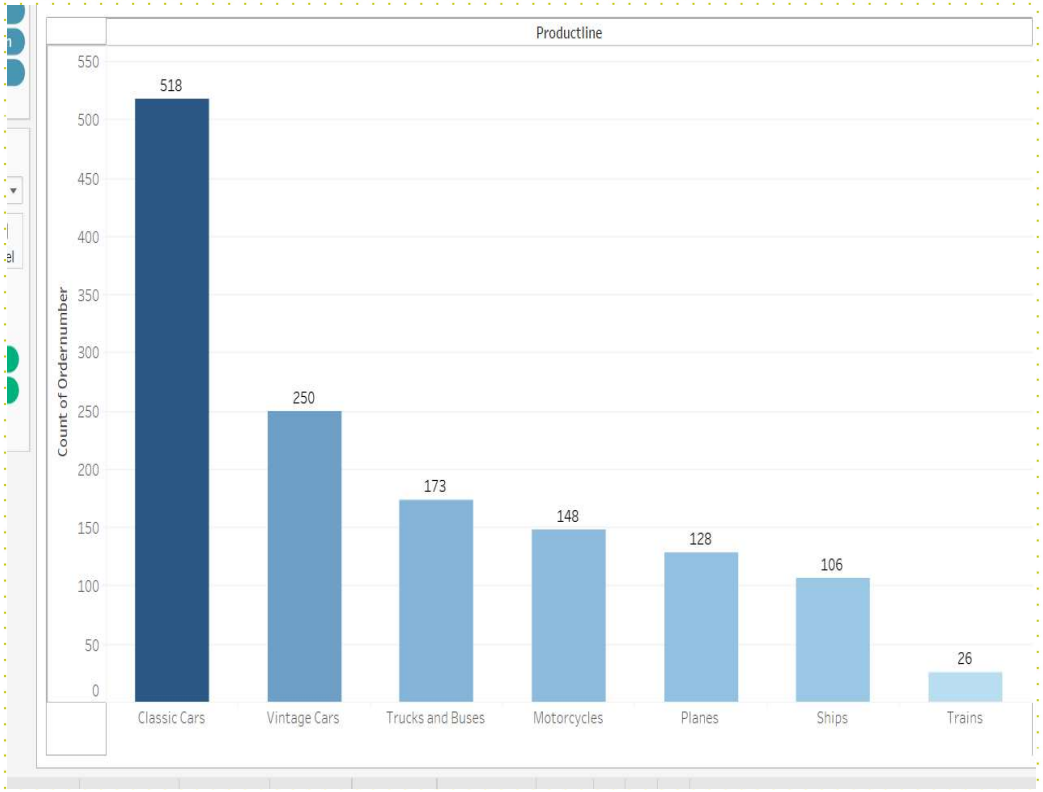
Exploratory Analysis and Inferences

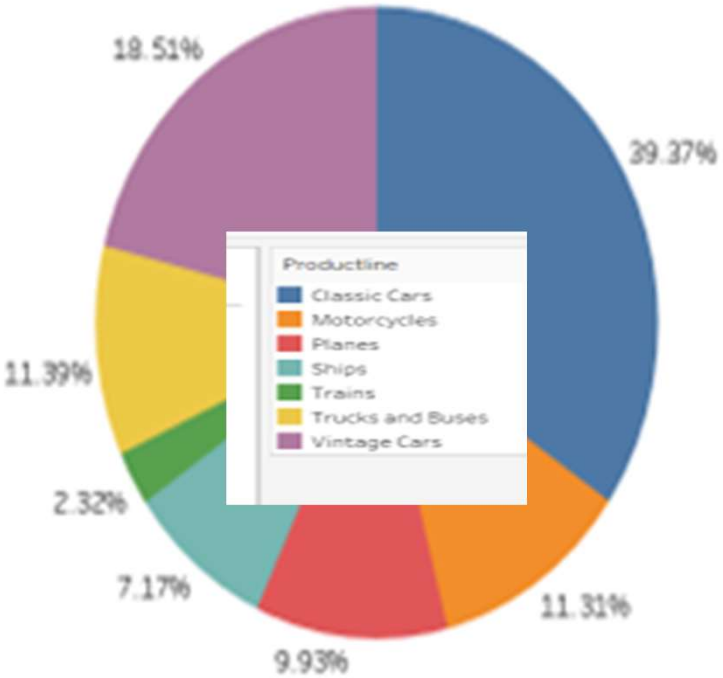
Univariate Analysis



There is presence of severe Outlier in sales, other variables have not so severity except the Order line number, and Day Since last order.

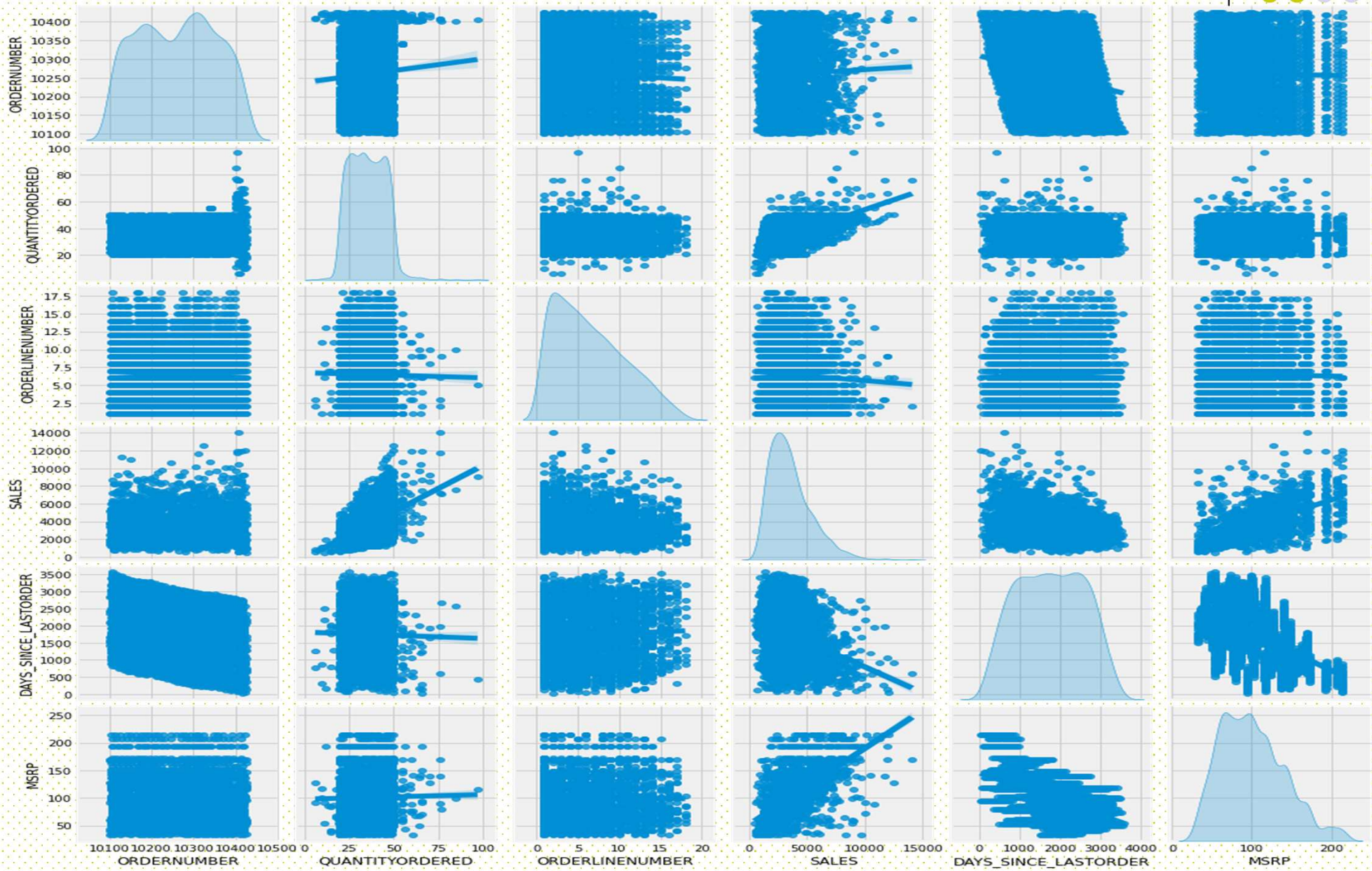
Sales analysis and total percentage



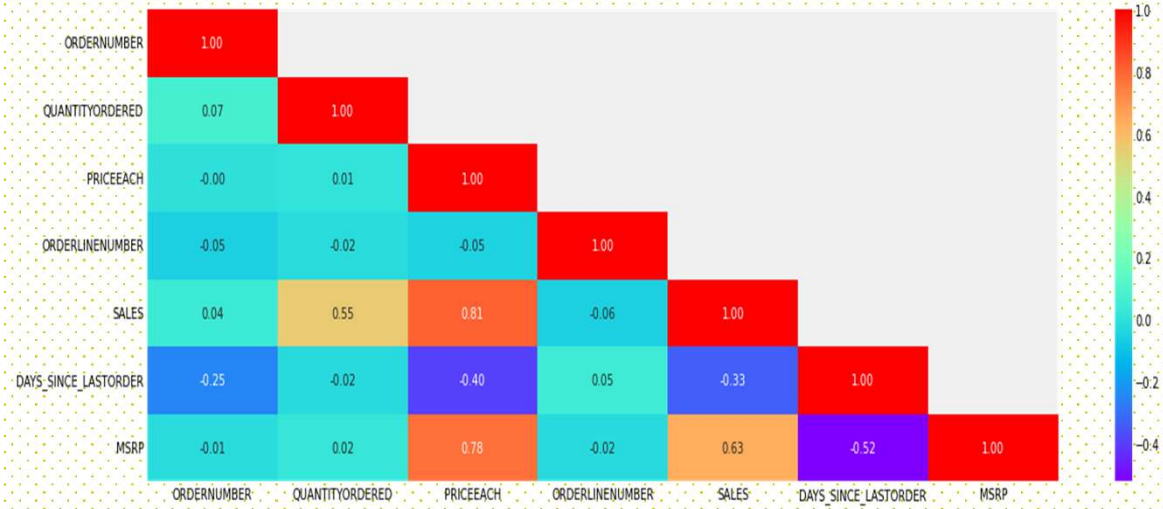


[Project ARM Visual | Tableau Public]

Bivariate Analysis



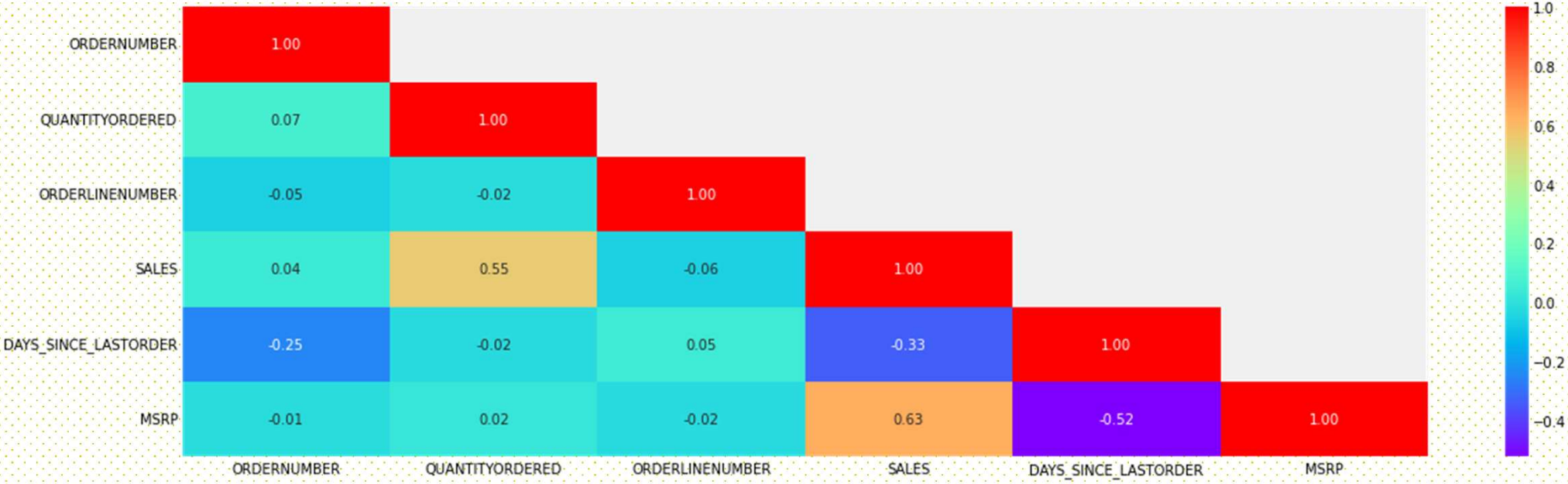
Multivariate Analysis



Correlation

PRICEEACH	SALES	0.808287
MSRP	PRICEEACH	0.778393
SALES	MSRP	0.634849
QUANTITYORDERED	SALES	0.553359
MSRP	DAYS_SINCE_LASTORDE	0.524285

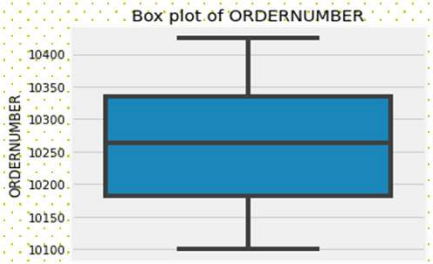
After dropping of Correlation

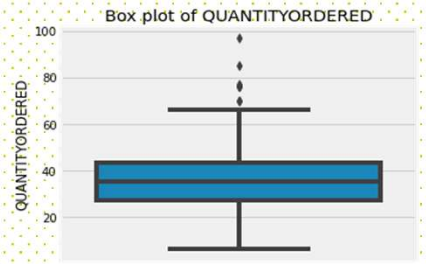


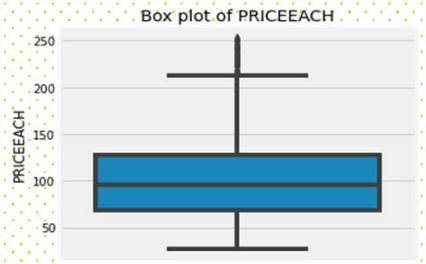
Variables Insights

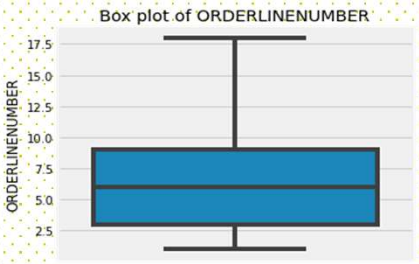
- There is presence of severe Outlier in sales, other variables have not so severity except the Order line number, and Day Since last order.
- Overall Sales are Highest with Category classic cars while lowest in Trains
- Small Deal size are in Classic cars and Vintage cars while large and medium deal size are in Classic cars category.
- There is linear relationship between sales and Price
- Some meaningful relation among Sales and MSRP with correlation of .63.
- Some positive relation among Sales and Quantity ordered with correlation of 0.55

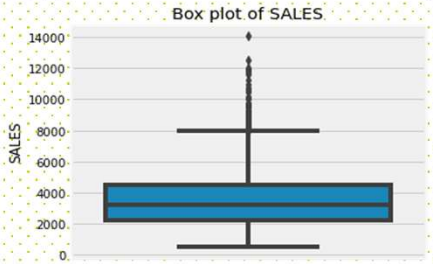
Outlier Detection & Treatment

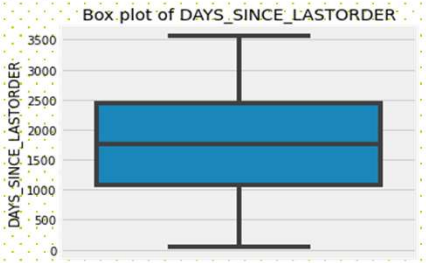


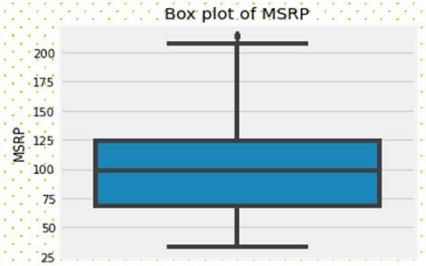




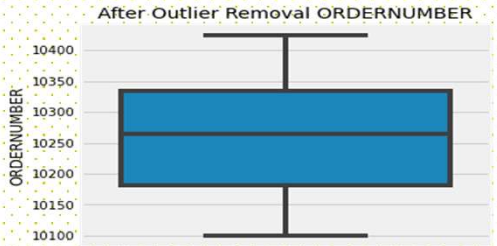


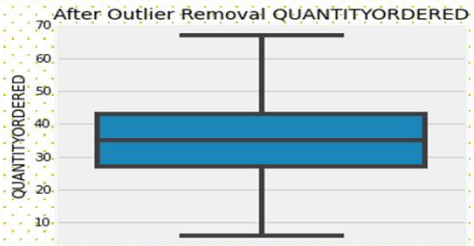


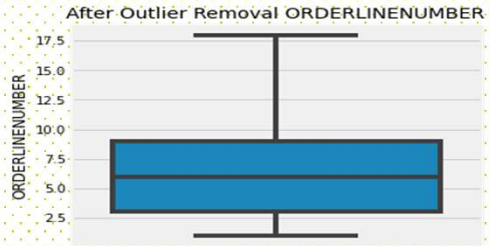


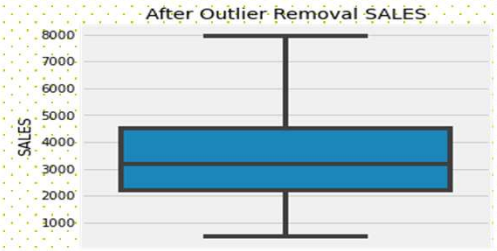


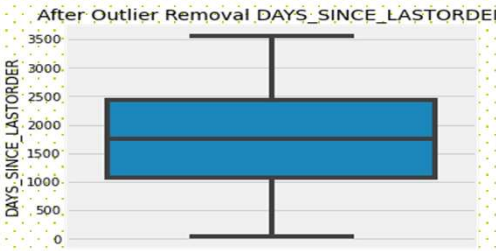
Before and after the treatment of Outlier

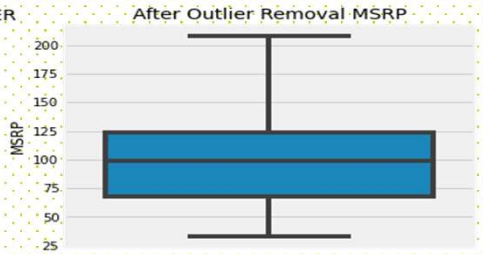




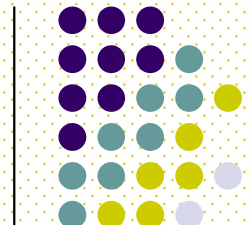


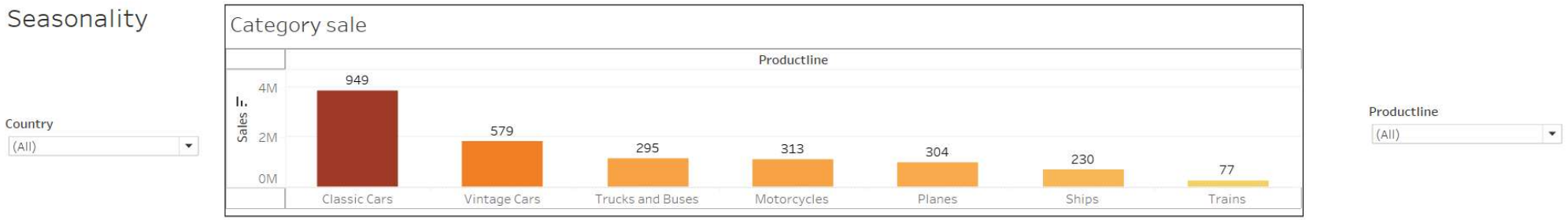


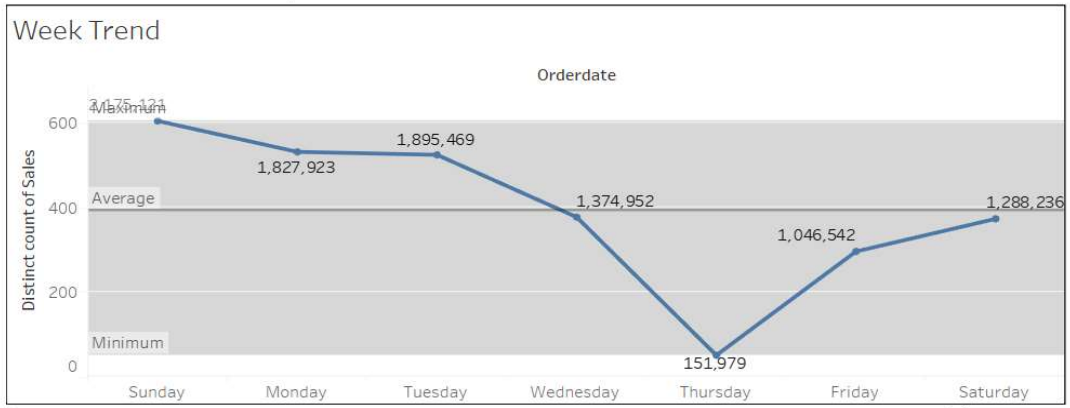


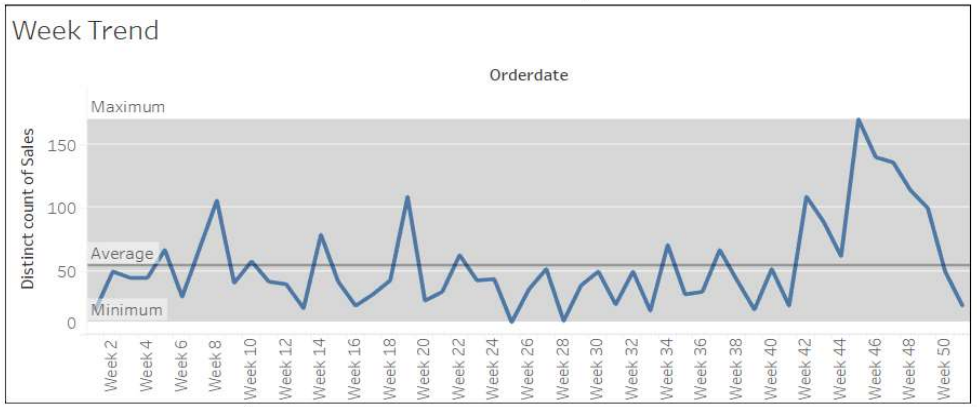


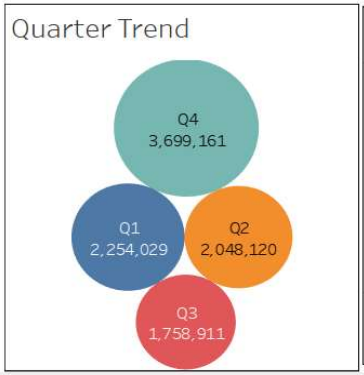
Trend Analysis

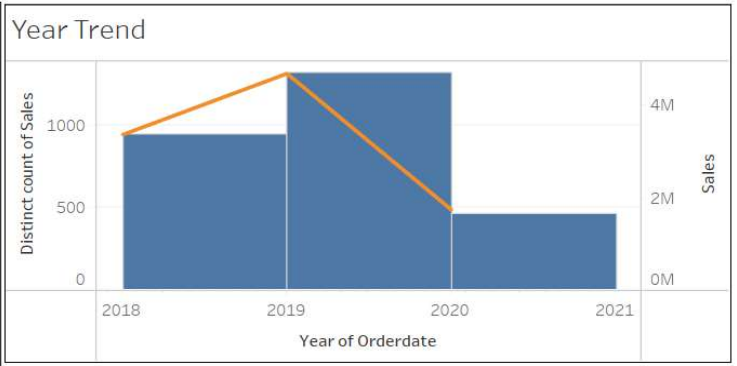






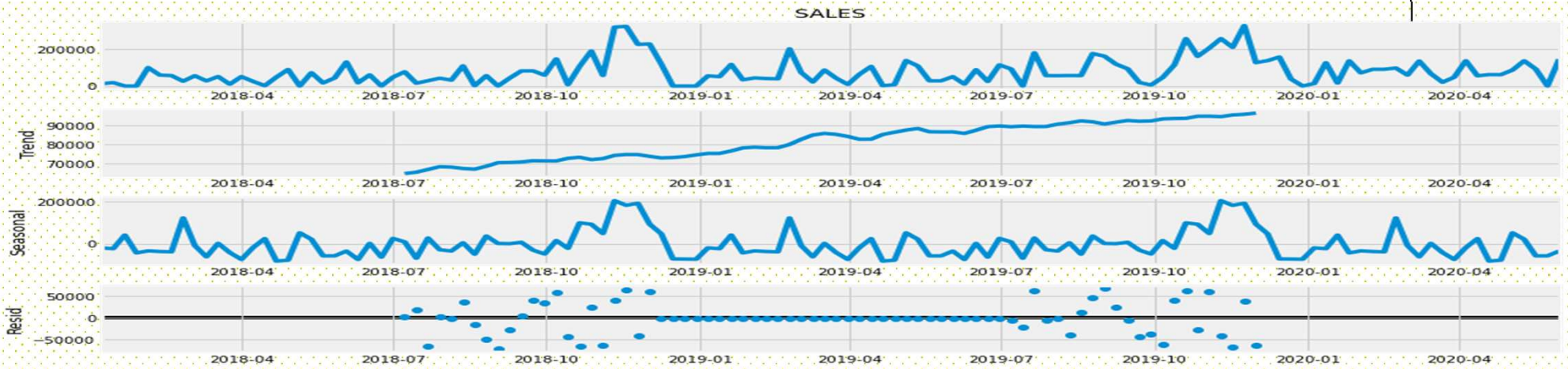


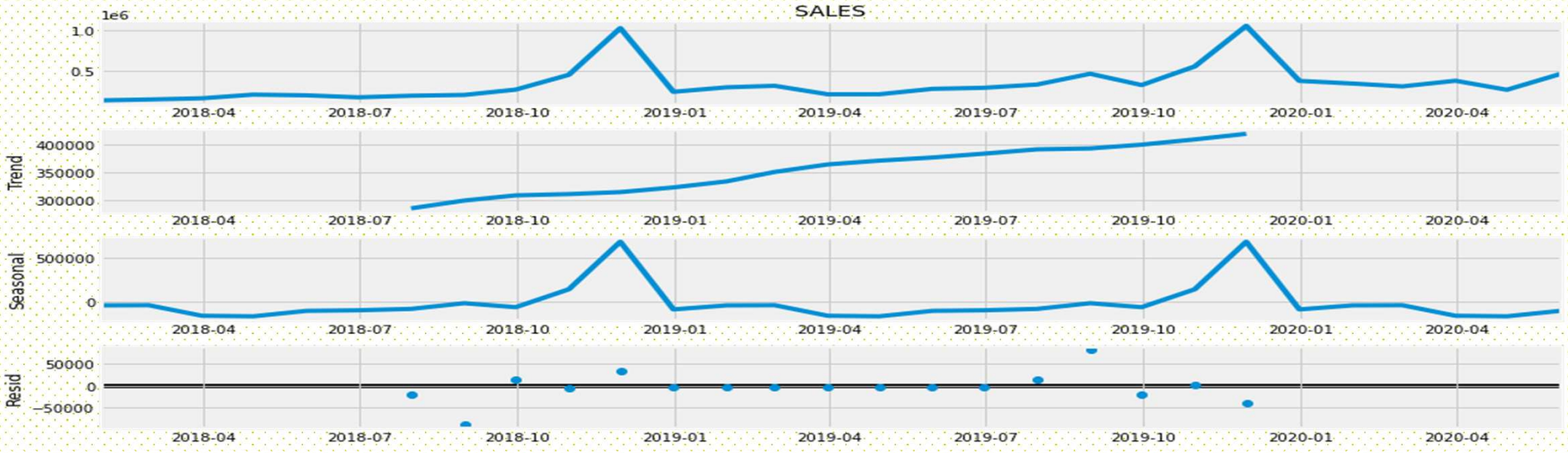




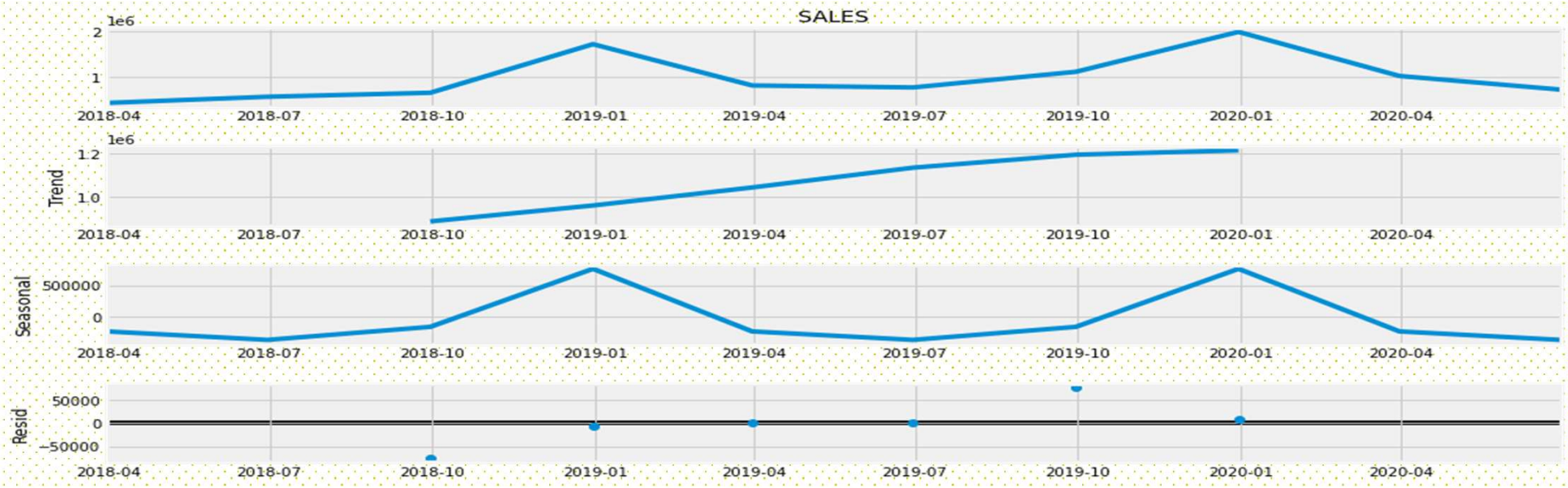
Category sale | Year Trend | Quarter Trend | Month Trend | Week Trend | Week Trend | Seasonality

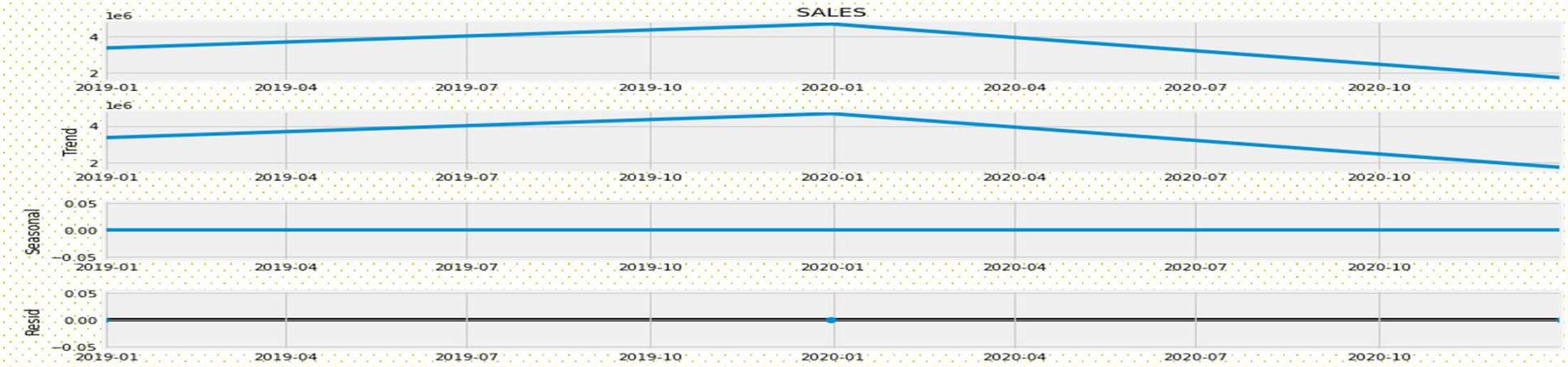
Trend Analysis in Sales Weekly and monthly





Trend Analysis in Sales Quarterly and Yearly





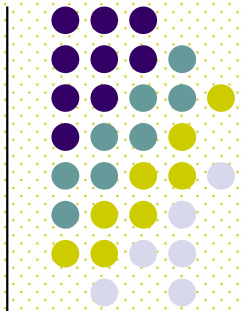
Trend Analysis in Sales

Plot Insights

- In Yearly Trend the year 2019 was good for all categories while registered a steep fall in very next year.
- In quarterly sale analysis the factor seasonality also been there because the Q4 has registered the maximum sales compared to other quarters.
- The granularity of seasonality can be observed better in month wise plot of data. The month of November has registered the maximum number of sale while December a steep downward trend.
- Further magnification of trend on week basis the Sunday, Monday, and Tuesday are days of good business.

Customer Segmentation using RFM analysis

RFM



RFM is a data-driven customer segmentation technique that allows marketers to take tactical decisions. It empowers marketers to quickly identify and segment users into homogeneous groups and target them with differentiated and personalized marketing strategies. This in turn improves user engagement and retention

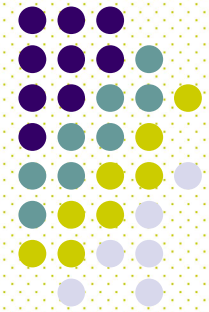
RFM stands for Recency, Frequency, and Monetary value, each corresponding to some key customer trait. These RFM metrics are important indicators of a customer's behavior because frequency and monetary value affects a customer's lifetime value, and recency affects retention, a measure of engagement.

- Recency: How recently a customer has made a purchase
- Frequency: How often a customer makes a purchase
- Monetary value: How much money a customer spends on purchases

RFM analysis allows a comparison between potential contributors and clients. It gives organizations a sense of how much revenue comes from repeat customers (vs. new customers), and which levers they can pull to try to make customers happier so they become repeat purchasers

Customer Segmentation using RFM analysis

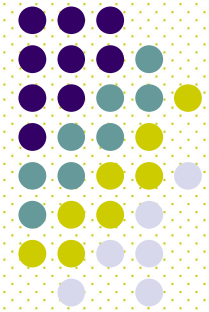
The parameters that used for analysis with the variables



- SALES: This variable represents the total sales for an order. It can be used to calculate revenue and track sales performance.
- ORDERDATE: This variable represents the date on which an order was placed. It can be used to analyze sales trends over time.
- DAYS_SINCE_LASTORDER: This variable represents the number of days since the customer's last order. It can be used to identify loyal customers and track customer behavior.
- ORDERLINENUMBER: This variable represents the order line number for a product. It can be used to identify unique products and track inventory levels.
- PRICEEACH: This variable represents the price for each unit of a product. It can be used to calculate revenue, profit margins, and customer value.
- PHONE: This variable is treated as Unique ID for customers.

Customer Segmentation using RFM analysis

Assumptions



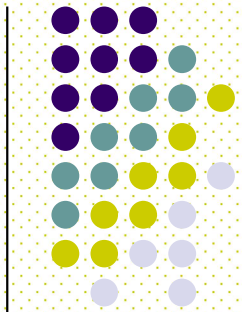
- Customers who have made recent purchases are more likely to make more purchases in the future.
- Customers who have made more purchases in the past are more likely to make more purchases in the future.
- Customers who have spent more money on their purchases are more valuable to the company
- All customers are treated equally, regardless of their demographics or purchase history.
- The time period for calculating RFM scores is typically set to the past given three year periods.
- The scores for each parameter (Recency, Frequency, Monetary) are typically divided into four quintiles (LOW, LM,UM, HIGH), with a score of LOW indicating the lowest value and a score of HIGH indicating the highest value, While reverse in case of recency.

- The final RFM score is usually calculated by concatenating the scores for each parameter.

RFM Output Table

RFM Analysis						
		MONETORY				
RECENCY	FREQUENCY	HIGH	UM	LM	LOW	Status
HIGH	HIGH	9	1	0	0	Active
	UM	1	3	0	0	
	LM	0	1	5	1	
	LOW	0	0	1	0	
UM	HIGH	6	1	0	0	Loyal
	UM	1	4	1	0	
	LM	0	2	2	0	
	LOW	0	0	0	6	
LM	HIGH	3	0	0	0	At Risk
	UM	1	3	0	0	
	LM	0	4	7	0	
	LOW	0	0	1	2	
LOW	HIGH	1	0	0	0	Lost
	UM	0	2	0	0	
	LM	0	1	3	3	
	LOW	0	0	2	10	

RFM Table Output Summary



- Customers who have a high recency, high frequency, and high monetary value are the most valuable customers for the company, as they have made recent purchases, have made purchases frequently, and have spent a lot of money on their purchases. There are 9 such customers in the dataset.

- Customers who have a high recency, high frequency, and low monetary value are still valuable, but not as much as the first group, as they have made recent and frequent purchases, but have spent less money. There is only 1 such customer in the dataset.

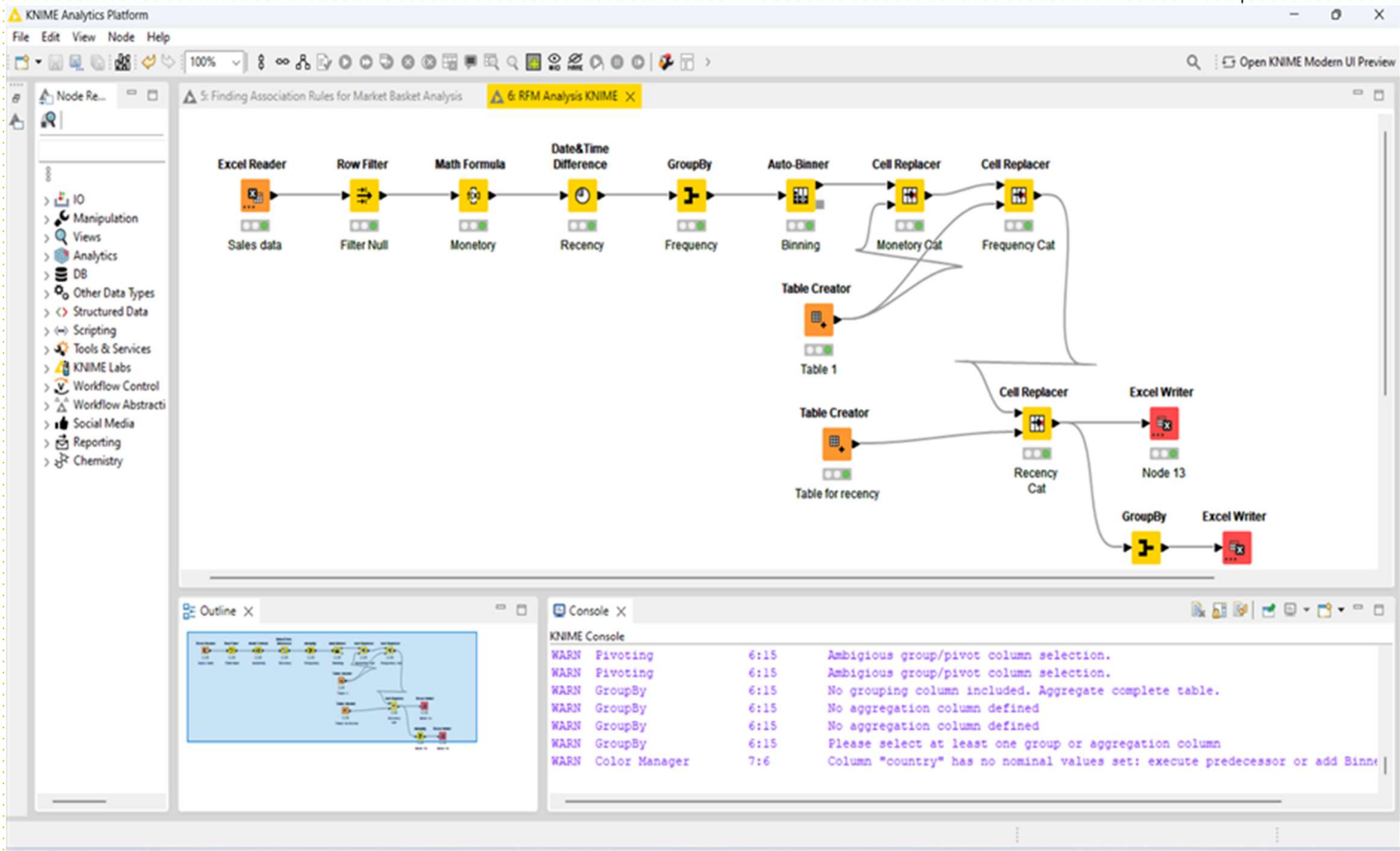
- Customers who have a high recency, medium frequency, and medium monetary value are still valuable, but again, not as much as the first group. They have made recent purchases, but not as frequently as the first group, and have spent less money. There is only 1 such customer in the dataset.

- Customers who have a high recency, low frequency, and high monetary value are valuable, but have not made recent purchases. They have only made a few purchases, but have spent a lot of money. There is only 1 such customer in the dataset.

- Customers who have a medium recency, high frequency, and high monetary value are also valuable, as they have made frequent purchases and spent a lot of money, but it has been some time since their last purchase. There is only 1 such customer in the dataset.

Customer Segmentation using RFM analysis

KNIME Work Flow



Best Customers:

- Customers who have a high recency, high frequency, and high monetary value are the most valuable customers for the company,
- As they have made recent purchases, have made purchases frequently, and have spent a lot of money on their purchases.
- There are 9 such customers in the dataset.

Best Five Customers

Index	PHONE	PRODUCTLINE	CONTACTFIRSTNAME	DEALSIZE	Monetory	Recency	MONETORY Cat	FREQUENCY Cat	RECENCY Cat
13	(91) 555 94 44	259	Diego	Large	912294.11	1074	HIGH	HIGH	HIGH
24	+61 2 9495 8555	46	Adrian	Small	151570.98	1076	HIGH	HIGH	HIGH
35	0522-556555	39	Maurizio	Large	142601.33	1095	HIGH	HIGH	HIGH
53	26.47.1555	41	Paul	Small	135042.94	1136	HIGH	HIGH	HIGH
55	31 12 3555	36	Jytte	Large	145041.6	1120	HIGH	HIGH	HIGH

Loyal customers

- Loyal customers are those who have high frequency and monetary values, and UM recency value.
- These customers have made frequent purchases with a high monetary value in the recent past and are likely to continue to do so in the future.
- They are considered to be the most valuable customers as they generate a significant portion of the company's revenue.

Five Loyal Customers

Index	PHONE	QUANTITYORDERED	CONTACTFIRSTNAME	DEALSIZE	Monetory	Recency	MONETORY Cat	FREQUENCY Cat	RECENCY Cat
26	+65 221 7555	43	Eric	Large	172989.68	1164	HIGH	HIGH	UM
31	02 9936 8555	46	Anna	Small	153996.13	1157	HIGH	HIGH	UM
33	03 9520 4555	55	Peter	Medium	200995.41	1258	HIGH	HIGH	UM
36	0695-34 6555	38	Maria	Large	134259.33	1163	HIGH	HIGH	UM
47	2125557413	48	Jeff	Large	197736.94	1256	HIGH	HIGH	UM

Customer at Risk

Customers who have the LM recency, LM frequency, and LM monetary value segments are considered to be at the highest risk of churning

There are 7 such customers in the dataset

More holistic approach that includes additional data sources and analyses might be necessary for a comprehensive customer retention strategy.

Five At Risk Customers

Index	PHONE	ORDERNUMBER	CONTACTFIRSTNAME	DEALSIZE	Monetory	Recency	MONETORY Cat	FREQUENCY Cat	RECENCY Cat
8	(198) 555-8888	26	Helen	Medium	78240.84	1286	LM	LM	LM
10	(604) 555-3392	22	Yoshi	Large	75238.92	1296	LM	LM	LM
14	(93) 203 4555	23	Eduardo	Medium	78411.86	1263	LM	LM	LM
22	+49 69 66 90 2555	22	Roland	Medium	85171.59	1282	LM	LM	LM
49	2125558493	20	Maria	Medium	77795.2	1266	LM	LM	LM

Lost Customers

- Customers who have a Low recency, low frequency, and low monetary value are the least valuable customers in the dataset,
- As they have made purchases some how in past, but not frequently, and have spent relatively little money.
- There are 10 such customers in the dataset.

Lost Five Customers

Index	PHONE	ORDERNUMBER	CONTACTFIRSTNAME	DEALSIZE	Monetory	Recency	MONETORY Cat	FREQUENCY Cat	RECENCY Cat
7	(171) 555-7555	12	Thomas	Medium	36019.04	1569	LOW	LOW	LOW
15	(95) 555 82 82	15	Jose Pedro	Medium	54723.62	1312	LOW	LOW	LOW
17	+34 913 728 555	13	Jesus	Large	49642.05	1513	LOW	LOW	LOW
18	+353 1862 1555	16	Dean	Large	57756.43	1332	LOW	LOW	LOW
23	+49 89 61 08 9555	14	Michael	Medium	34993.92	1333	LOW	LOW	LOW

Tools Used

- Python for basic EDA and visualisation (ipynb attached)
- Tableau for Visualisation (Project ARM Visual | Tableau Public)
- KNIME for RFM analysis (Work flow mentioned)


THANK
YOU