MRA Project

Milestone 1

PINAK (DSBA – Aug ' 20)

AGENDA

❑ **Problem Statement**

❑ **About the dataset**

❑ **EDA**

❑ **EDA summary**

❑ **RFM**

❑ **Segmentation**

❑ **Conclusion**

Problem statement

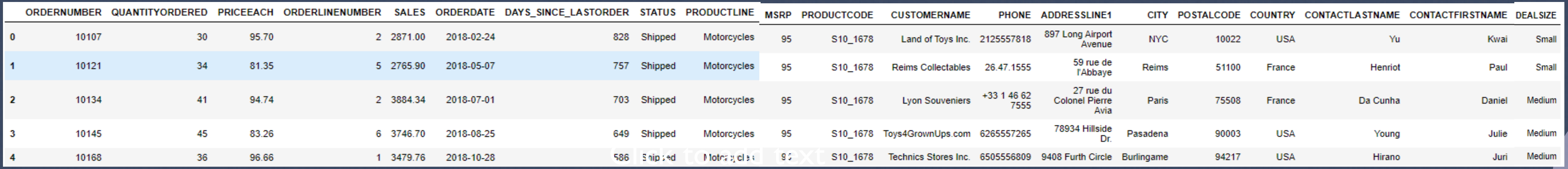
An automobile parts manufacturing company has collected data of transactions for 3 years. They do not have any in-house data science team, thus they have hired you as their consultant. Your job is to use your magical data science skills to provide them with suitable insights about their data and their customers.

Auto Sales Data: **Sales_Data.xlsx**

This project aims 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.

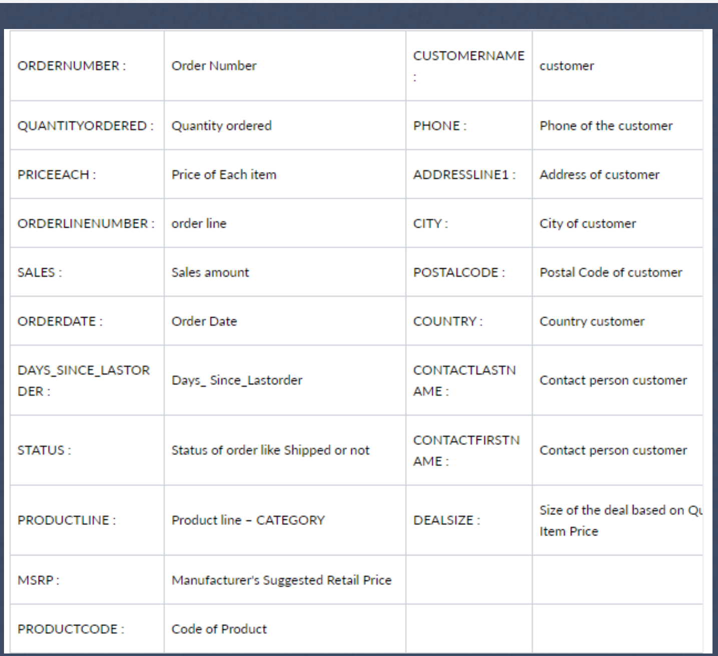
All about the dataset

The dataset has customer data along with sales amount. Other important features included in the dataset are product names along with the date/time information. A small glimpse of the dataset as below,



	ORDERNUMBER	QUANTITYORDERED	PRICEEACH	ORDERLINENUMBER	SALES	ORDERDATE	DAYS_SINCE_LASTORDER	STATUS	PRODUCTLINE	MSRP	PRODUCTCODE	CUSTOMERNAME	PHONE	ADDRESSLINE1	CITY	POSTALCODE	COUNTRY	CONTACTLASTNAME	CONTACTFIRSTNAME	DEALSIZE
0	10107	30	95.70	2	2871.00	2018-02-24	828	Shipped	Motorcycles	95	S10_1678	Land of Toys Inc.	2125557818	897 Long Airport Avenue	NYC	10022	USA	Yu	Kwai	Small
1	10121	34	81.35	5	2765.90	2018-05-07	757	Shipped	Motorcycles	95	S10_1678	Reims Collectables	26.47.1555	59 rue de l'Abbaye	Reims	51100	France	Henriot	Paul	Small
2	10134	41	94.74	2	3884.34	2018-07-01	703	Shipped	Motorcycles	95	S10_1678	Lyon Souveniers	+33 1 46 62 7555	27 rue du Colonel Pierre Avia	Paris	75508	France	Da Cunha	Daniel	Medium
3	10145	45	83.26	6	3746.70	2018-08-25	649	Shipped	Motorcycles	95	S10_1678	Toys4GrownUps.com	6265557265	78934 Hillside Dr.	Pasadena	90003	USA	Young	Julie	Medium
4	10168	36	96.66	1	3479.76	2018-10-28	586	Shipped	Motorcycles	95	S10_1678	Technics Stores Inc.	6505556809	9408 Furth Circle	Burlingame	94217	USA	Hirano	Juri	Medium

Data Dictionary



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

The following steps were followed while analysing the data,

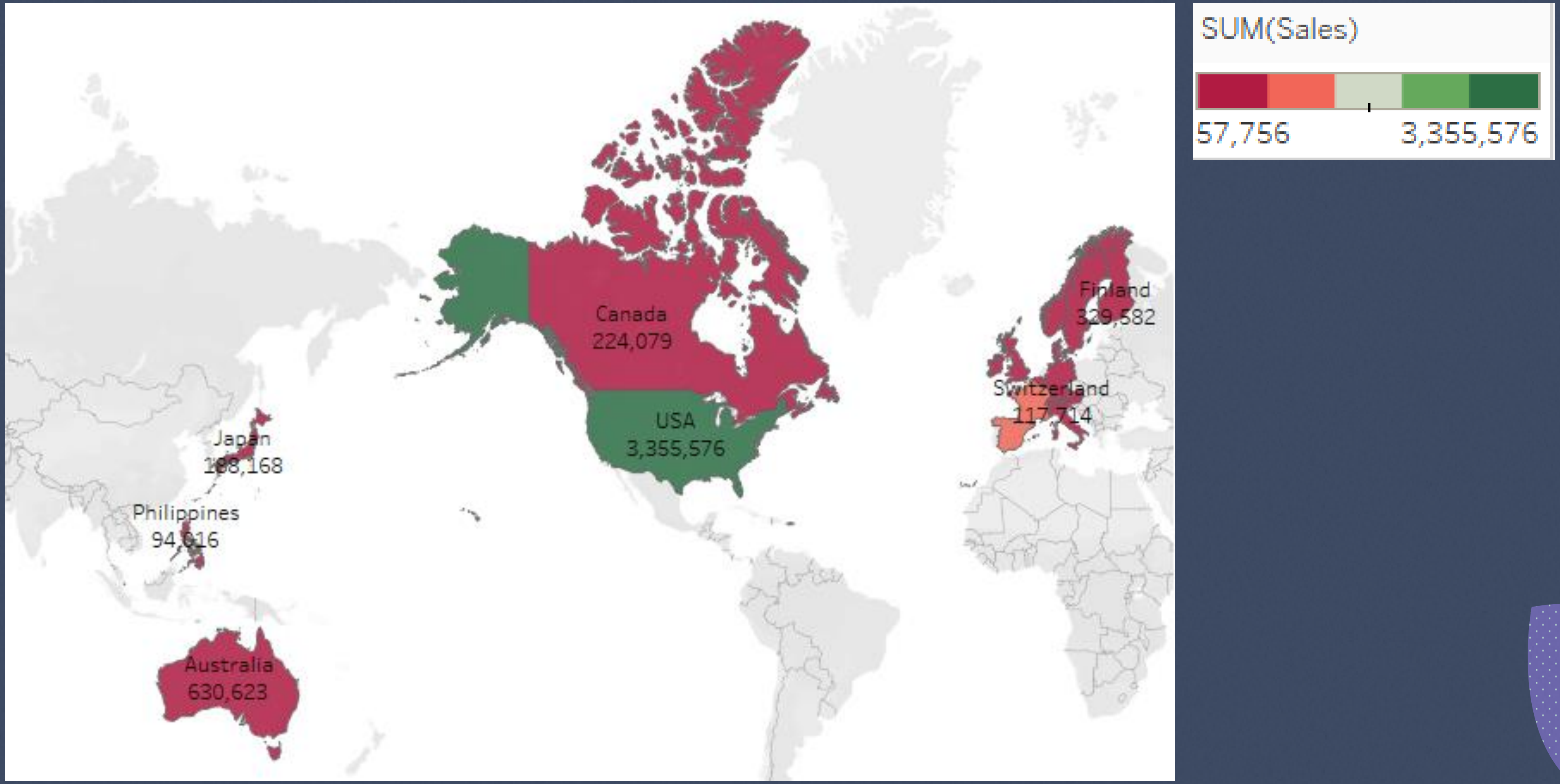
1. The raw dataset was first checked using Python 3 with Pandas and Numpy libraries. The basic stats of the dataset was derived using the above two libraries.
2. Tableau Public was used for data visualization and inferences for EDA.
3. Finally KNIME and Excel was used for RFM analysis.

All about the dataset

1) The dataset has 2747 rows and 20 columns
2) The dataset has 7 numerical, 1 datetime and 12 object variables
3) There is no null value in the dataset
4) There is no duplicates of any of the entries
5) There are outliers in most of the numerical variables

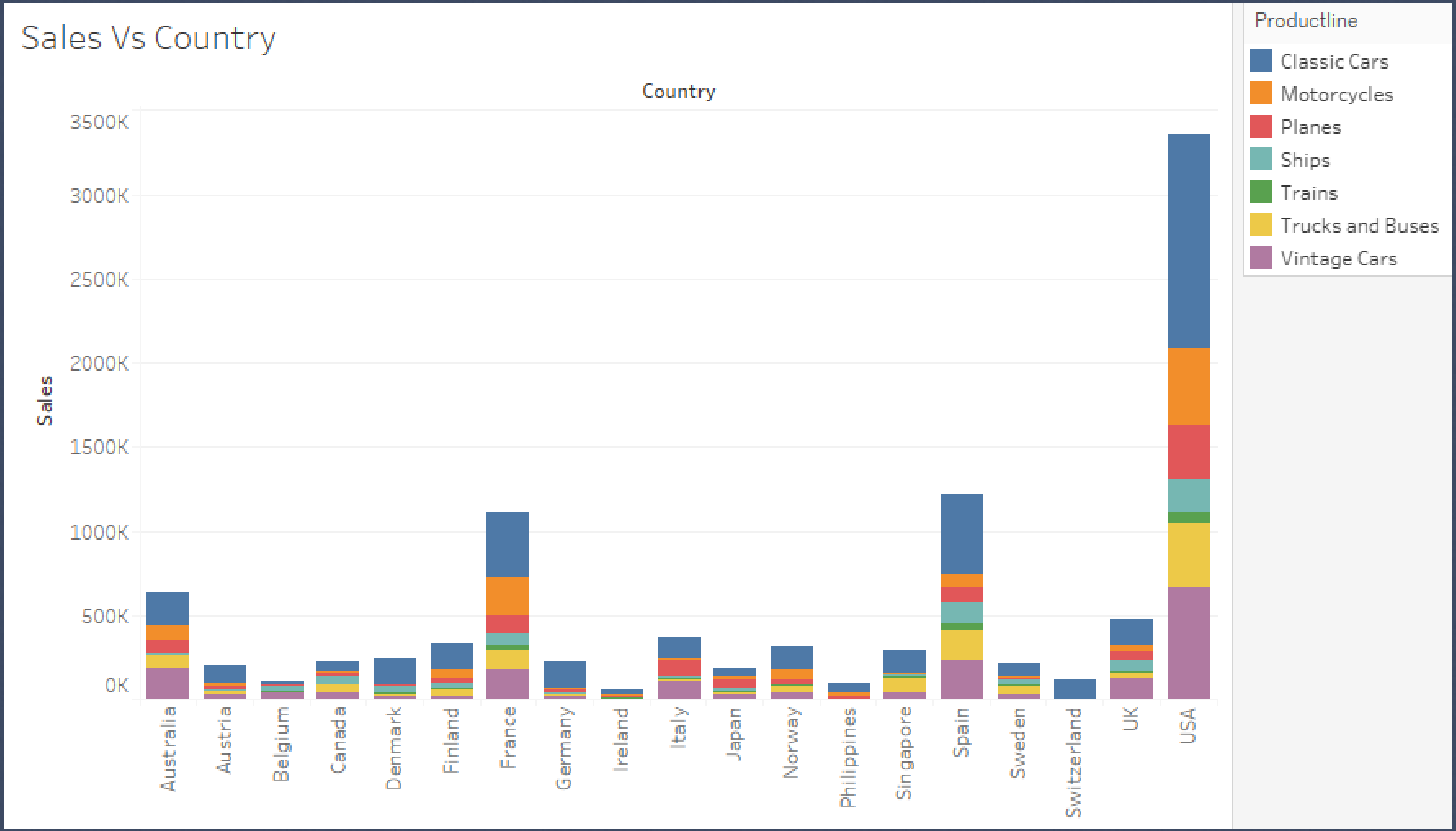
A detailed EDA (Univariate/bi-variate/Multivariate) on the dataset has been performed which is shown in subsequent slides. Tableau public has been used for data visualization.

EDA



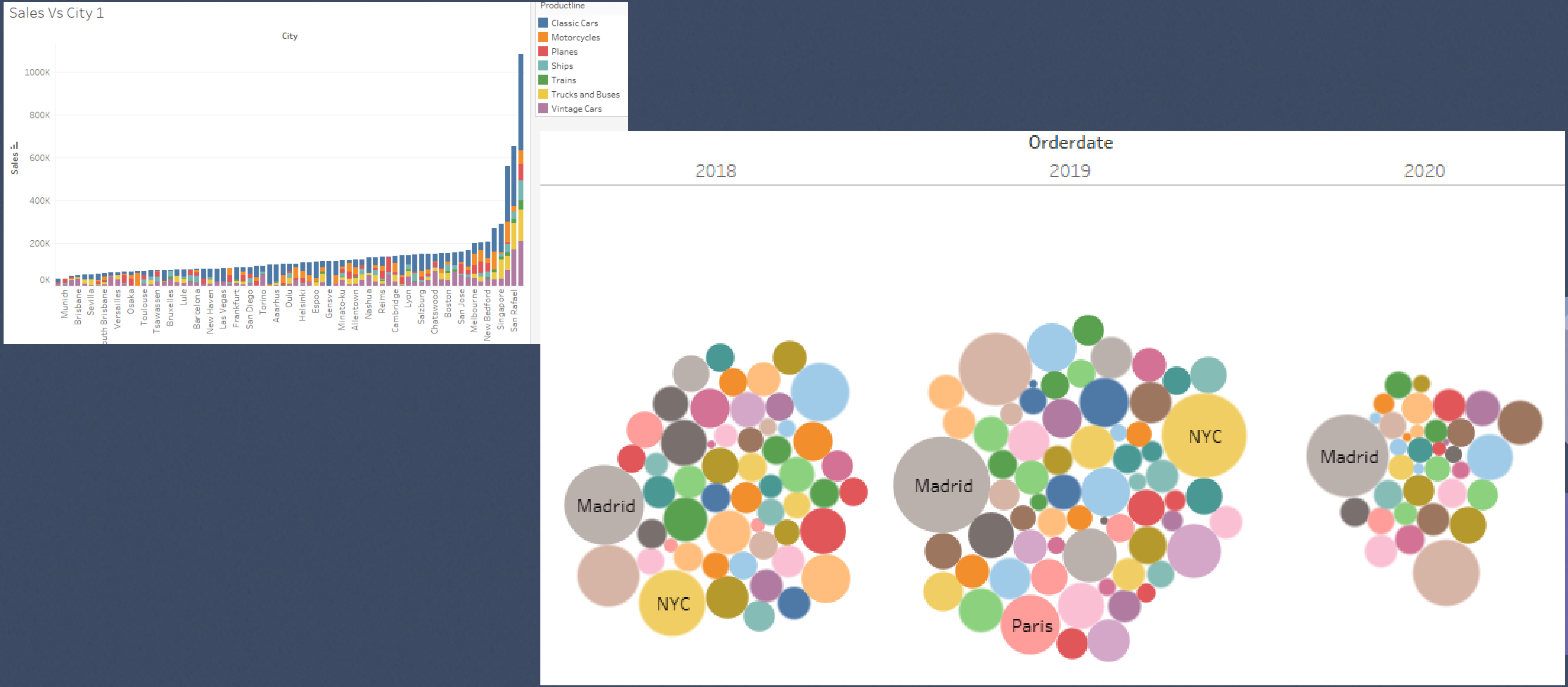
1. Most of the countries have less sales in terms of sum totals.
2. US has highest sales followed by Spain and France

EDA



1. Classics cars are most saleable cars among countries with highest sales such as USA, Spain, France.
2. Vintage car has significant sales in USA

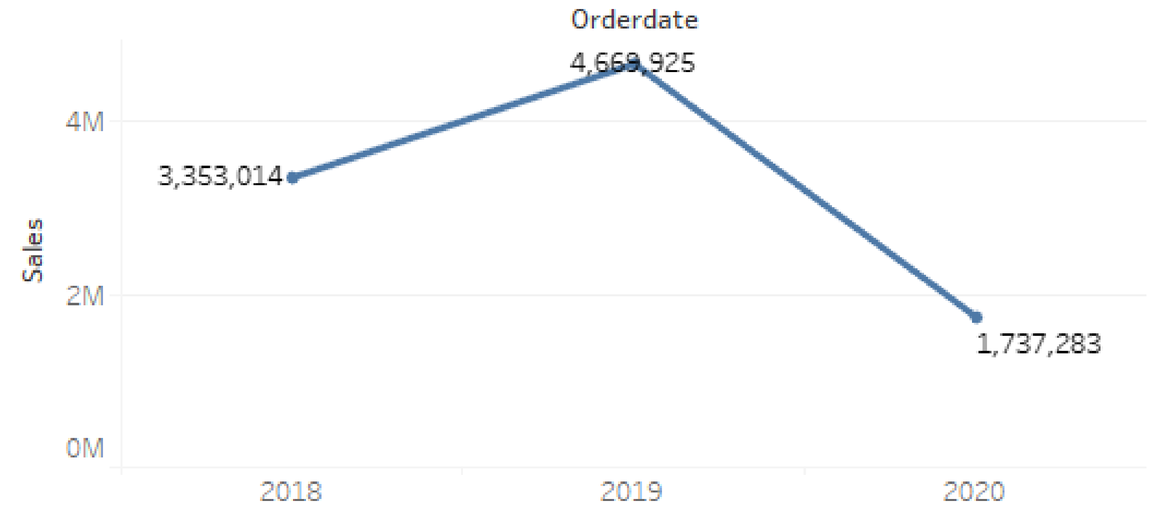
EDA



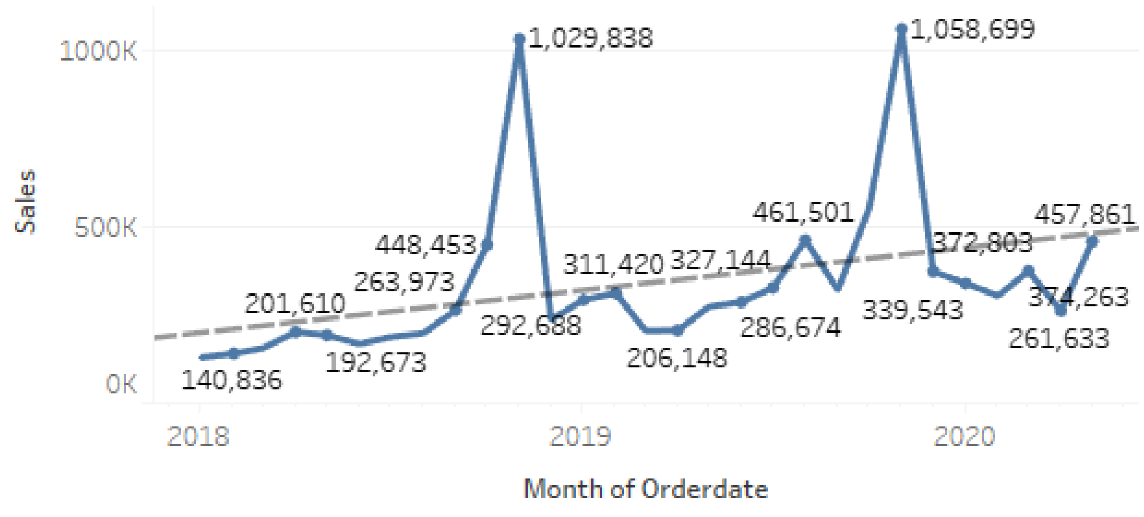
1. Madrid has the highest sales city wise in last 3 years
2. Classics cars followed by Vintage car have highest sales in city such as Madrid, San Rafael, NYC etc.

EDA

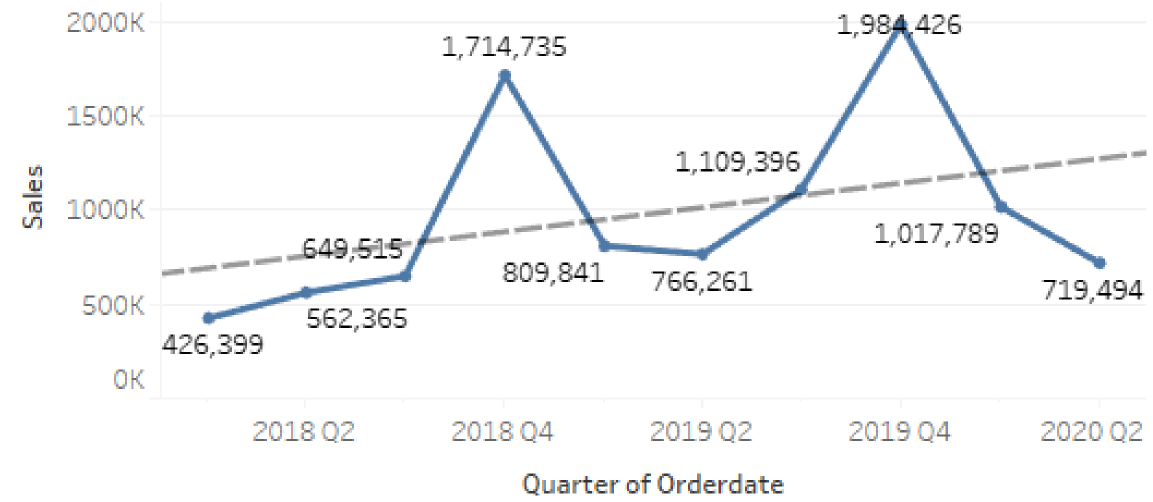
YEARLY



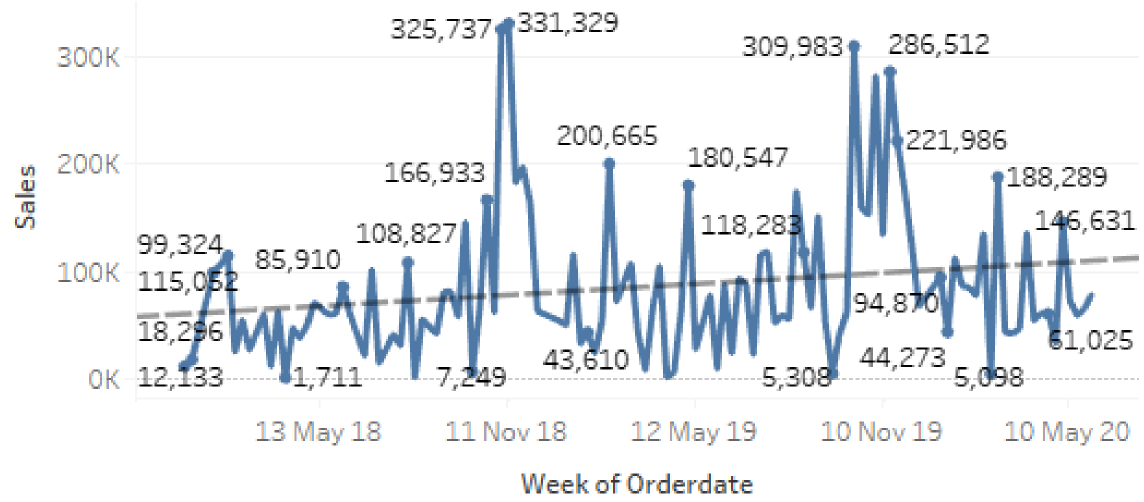
MONTHLY



QUARTERLY



WEEKLY



EDA

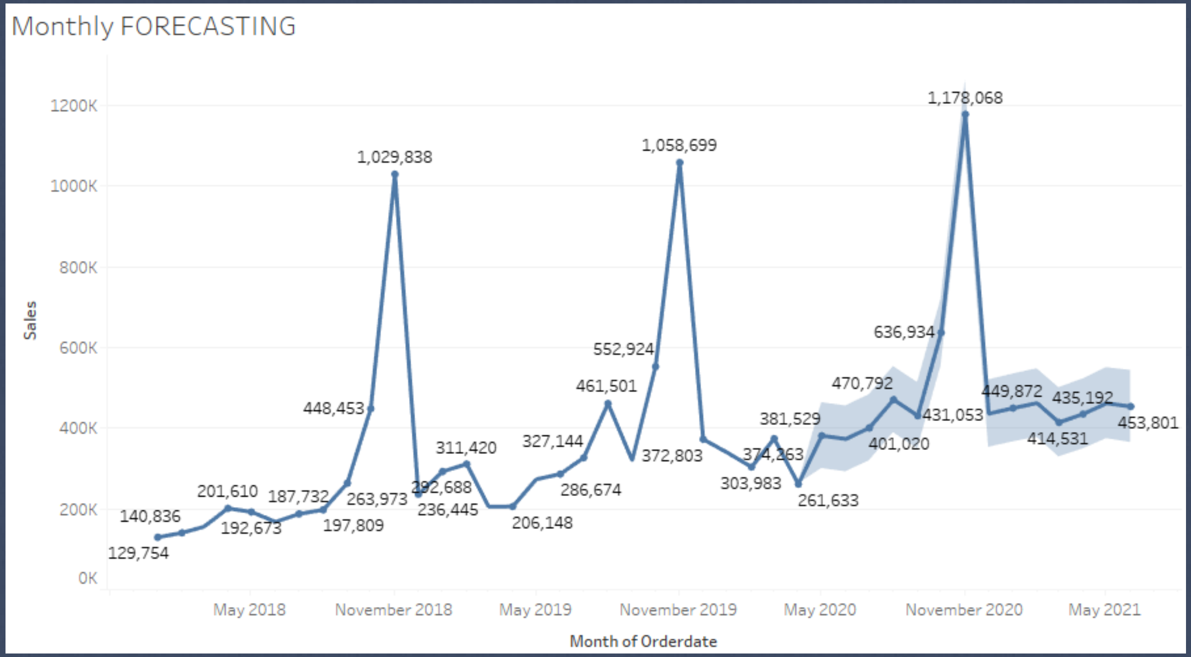
1. Yearly sales was highest in year 2019, has declined in 2020.

2. In monthly as well as quarterly data, seasonality is observed in month of September to November.

3. In weekly data, the variation in sales data is more.

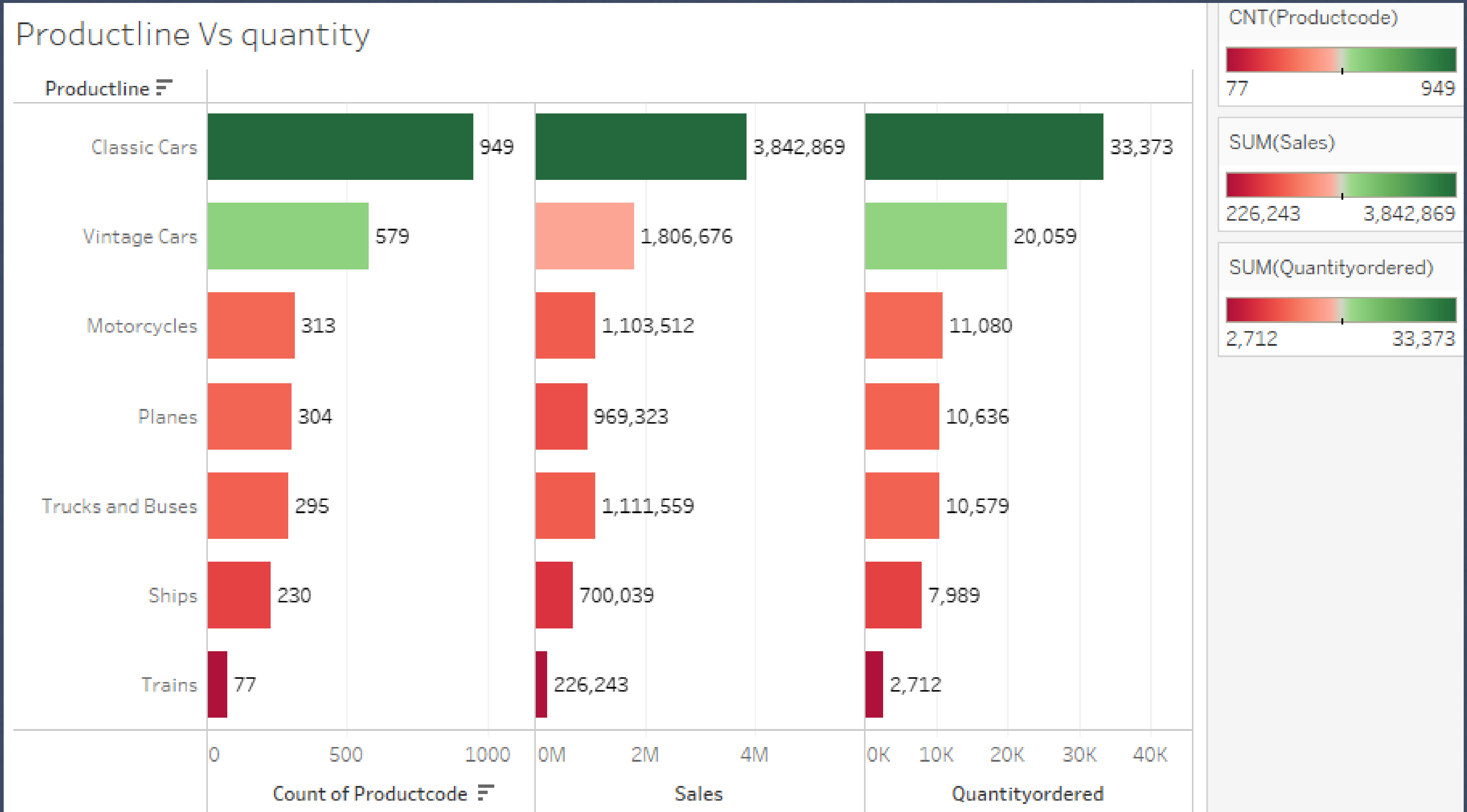
4. Overall the sales has been in a uptrend on monthly, quarterly and weekly data.

5. Below is a forecasting on monthly data for next 14 months with 95% CI.



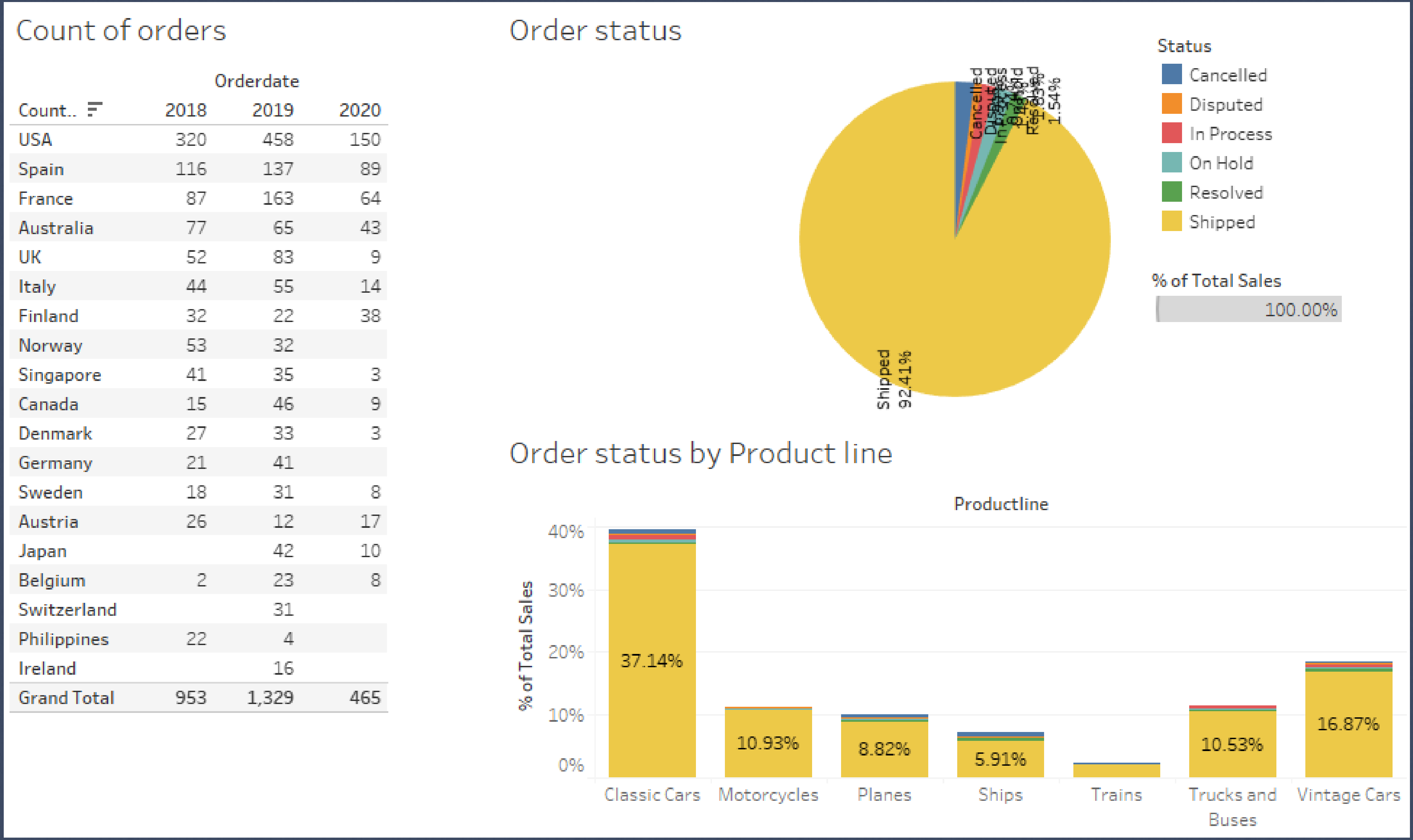
Monthly FORECASTING

EDA



1. Classic cars and Vintage cars are leaders in sales as well as quantity ordered.
2. Planes has significant sales , an interesting fact.

EDA



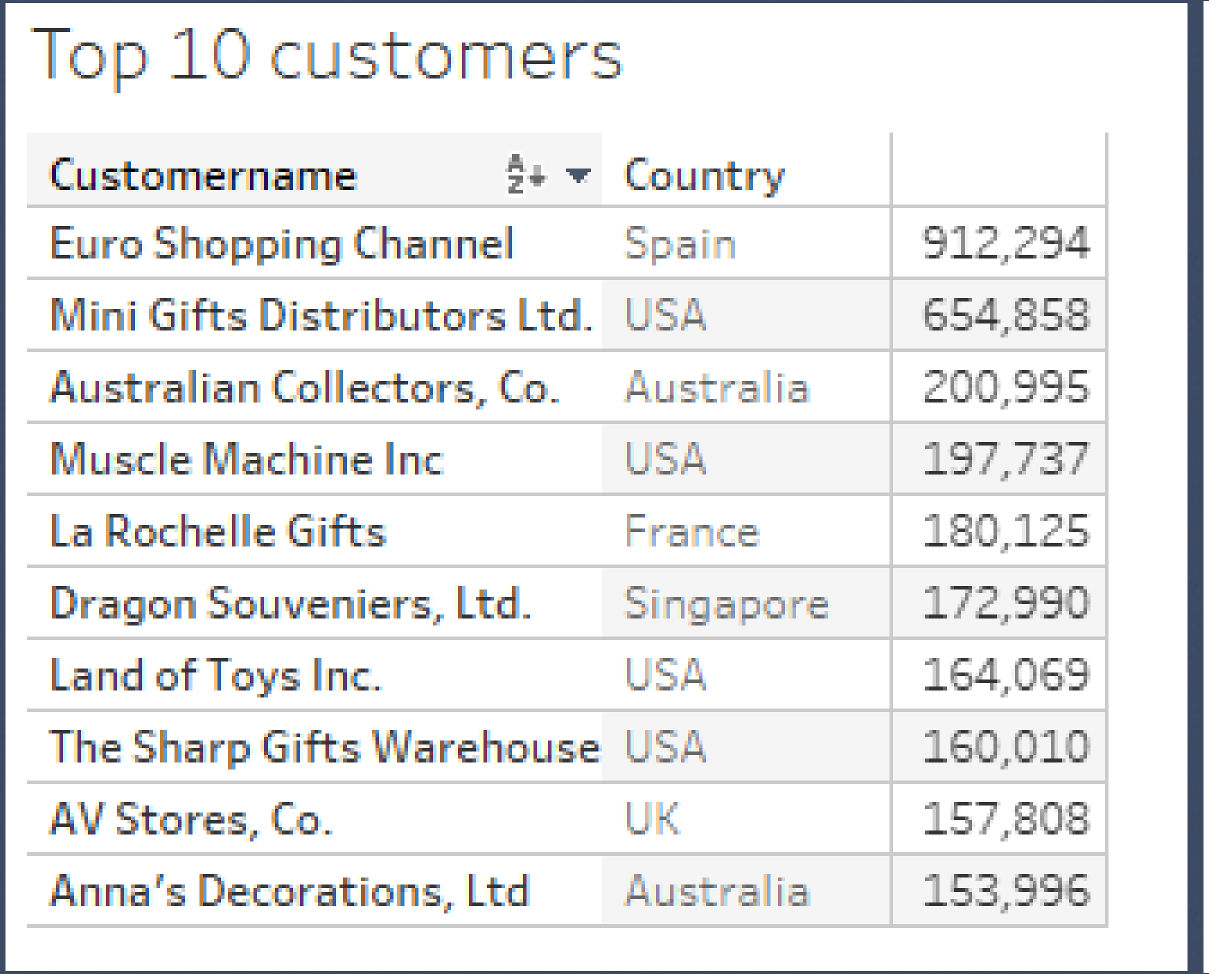
Count of orders

	Orderdate		
Count..	2018	2019	2020
USA	320	458	150
Spain	116	137	89
France	87	163	64
Australia	77	65	43
UK	52	83	9
Italy	44	55	14
Finland	32	22	38
Norway	53	32	
Singapore	41	35	3
Canada	15	46	9
Denmark	27	33	3
Germany	21	41	
Sweden	18	31	8
Austria	26	12	17
Japan		42	10
Belgium	2	23	8
Switzerland		31	
Philippines	22	4	
Ireland		16	
Grand Total	953	1,329	465

Order status

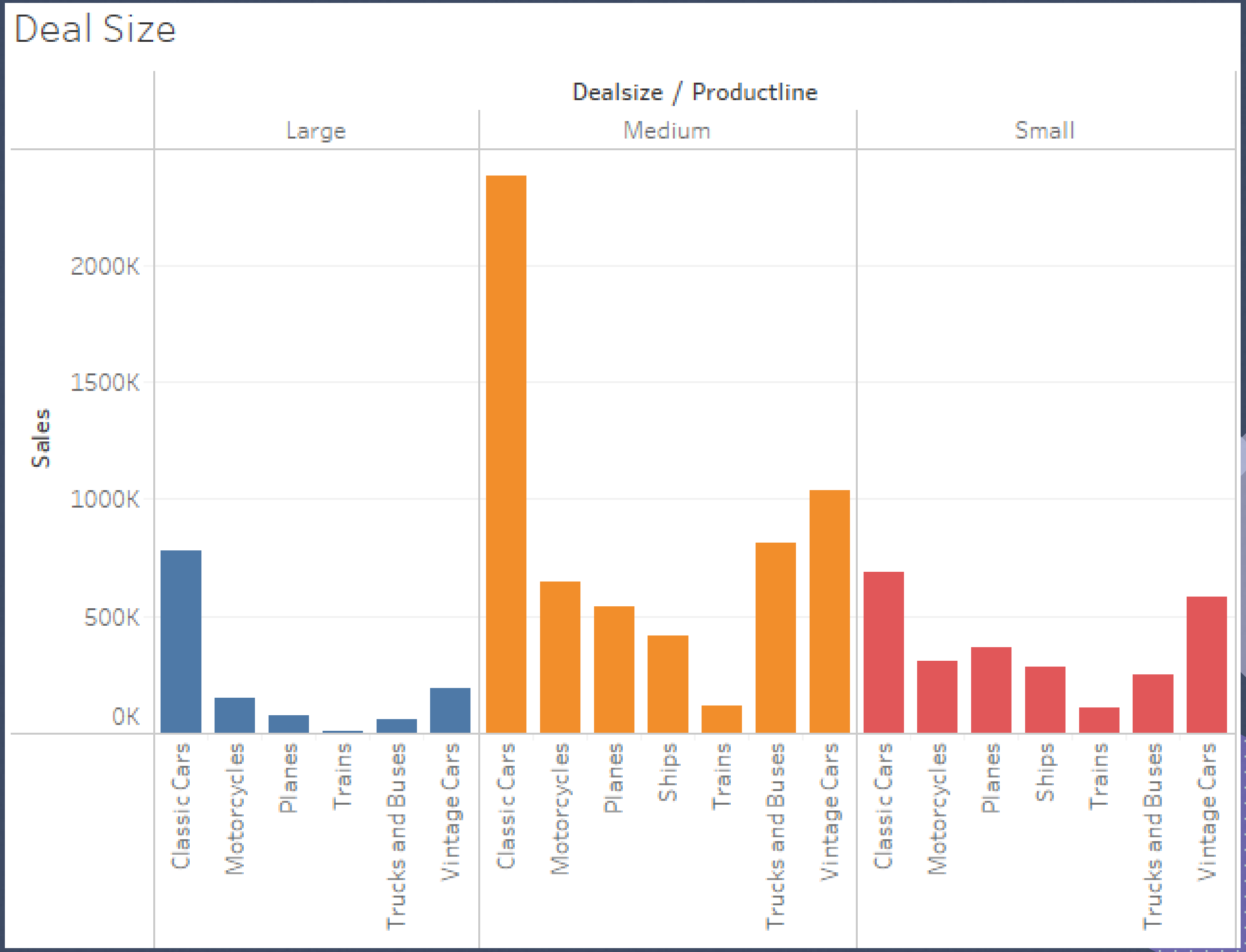
Order status by Product line

EDA



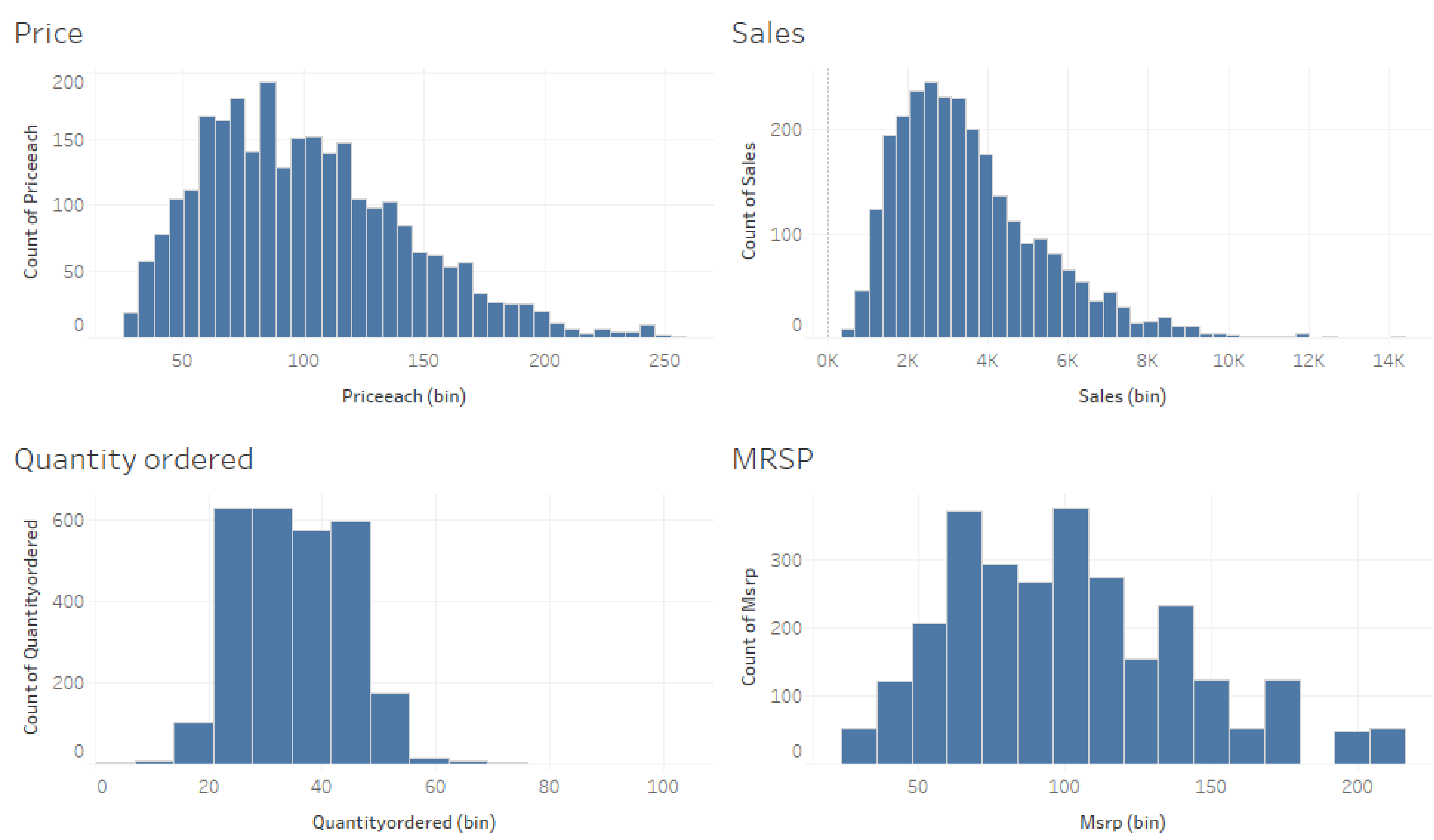
Top 10 customers

Customername	Country	
Euro Shopping Channel	Spain	912,294
Mini Gifts Distributors Ltd.	USA	654,858
Australian Collectors, Co.	Australia	200,995
Muscle Machine Inc	USA	197,737
La Rochelle Gifts	France	180,125
Dragon Souveniers, Ltd.	Singapore	172,990
Land of Toys Inc.	USA	164,069
The Sharp Gifts Warehouse	USA	160,010
AV Stores, Co.	UK	157,808
Anna's Decorations, Ltd	Australia	153,996



1. Euro shopping channel and Mini Gifts ltd are the top customers.
2. Most of the sales are happening through medium size deals.

EDA



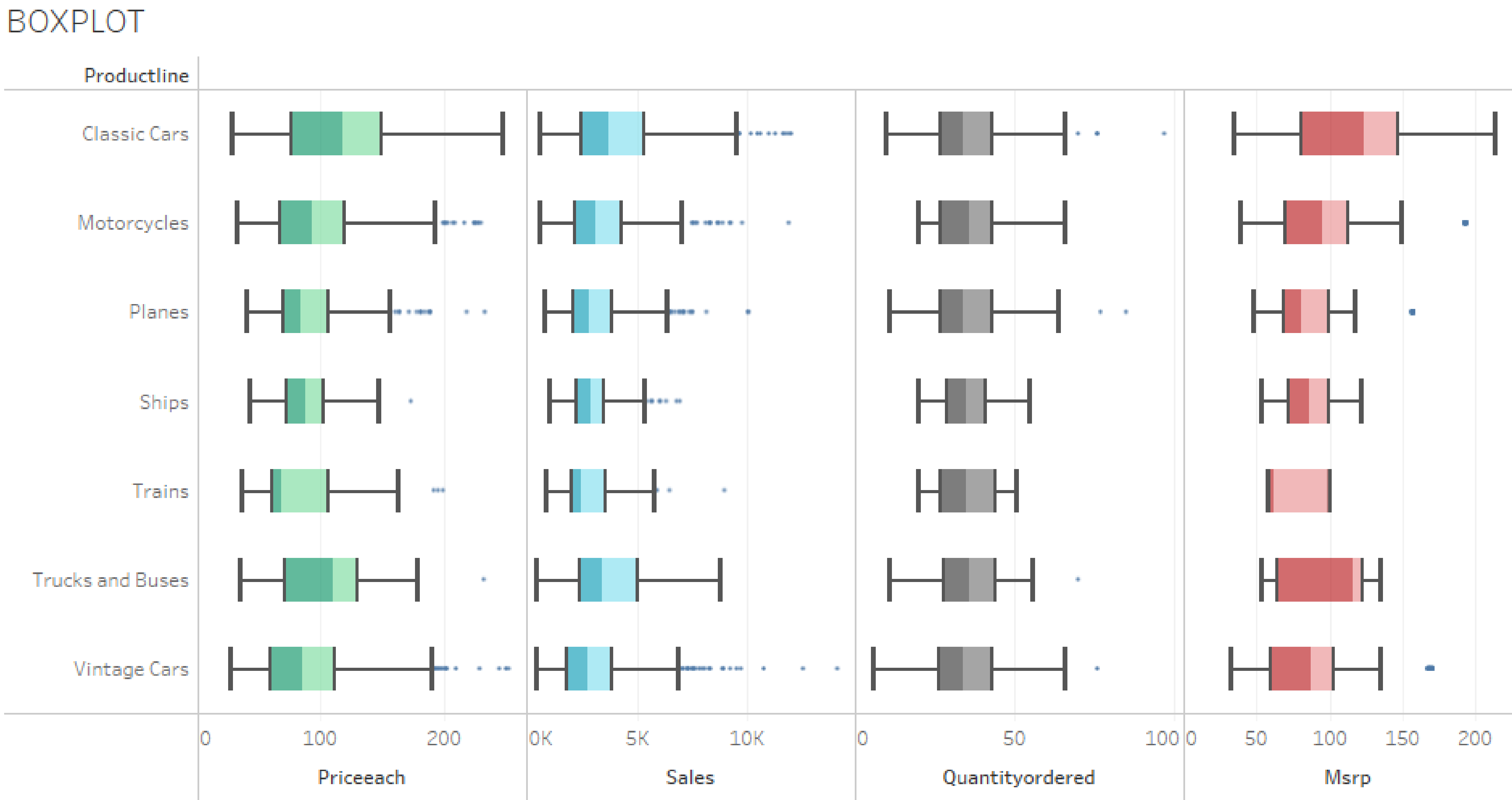
Price

Sales

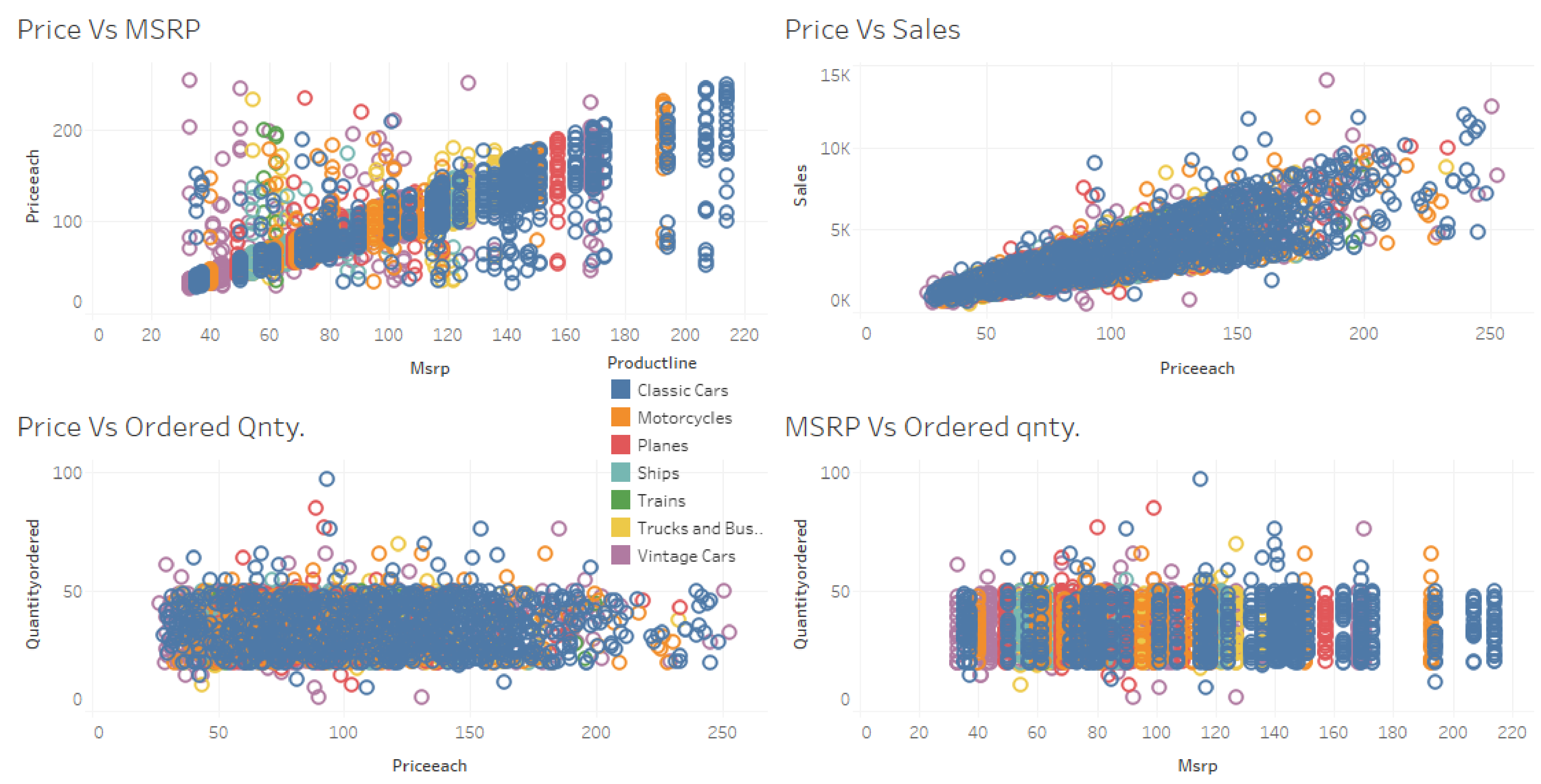
Quantity ordered

MRSP

EDA


BOXPLOT
Productline
Classic Cars
Motorcycles
Planes
Ships
Trains
Trucks and Buses
Vintage Cars
0
100
200
Priceeach
0K
5K
10K
Sales
0
50
100
Quantityordered
0
50
100
150
200
Msrp

EDA



Price Vs MSRP

Price Vs Sales

Price Vs Ordered Qnty.

MSRP Vs Ordered qnty.

EDA Summary

- The dataset has 7 numerical, 1 datetime and 12 object variables
- There is no null value in the dataset
- There is no duplicates of any of the entries
- There are outliers in most of the numerical variables
- Most of the countries have less sales in terms of sum totals. US has highest sales followed by Spain and France
- Classics cars are most saleable cars among countries with highest sales such as USA, Spain, France. Vintage car has significant sales in USA
- Madrid has the highest sales city wise in last 3 years. Classics cars followed by Vintage car have highest sales in city such as Madrid, San Rafael, NYC etc.
- Overall the sales has been in a uptrend on monthly, quarterly and weekly data.
- Yearly sales was highest in year 2019, has declined in 2020. In monthly as well as quarterly data, seasonality is observed in month of September to November.
- Classic cars and Vintage cars are leaders in sales as well as quantity ordered. Planes has significant sales , an interesting fact.

EDA Summary

- Euro shopping channel and Mini Gifts ltd are the top customers based on sales. Most of the sales are happening through medium size deals.
- The numerical variables are having skewed distribution
- There are corelations between Price Vs Sales, Price Vs MRSP etc..

RFM Analysis

RFM Analysis Procedure

For RFM Analysis, the basic variables/parameters to consider are Quantity ordered, Sales, Order date, Customer name from the provided dataset.

The Recency refers to the number of days to recent purchase which is derived by subtracting Order date from Max order date (1st June 2020).

The Frequency is assumed to be the total number of quantity ordered till reference date which is 1st June 2020.

The Monetary is the total sales amount till reference date.

There is one important parameter which is "Status". Different labels of "Status" affects the overall RFM analysis, e.g. the status "Cancelled" order should not be considered while calculating the RFM scores.

RFM Analysis

Few **Assumptions**

1. The Recency is derived by subtracting Order date from Max order date (1st June 2020). Recent date could have been considered, but it will give the recency numbers very high, which may cause the dataset looks unrealistic.

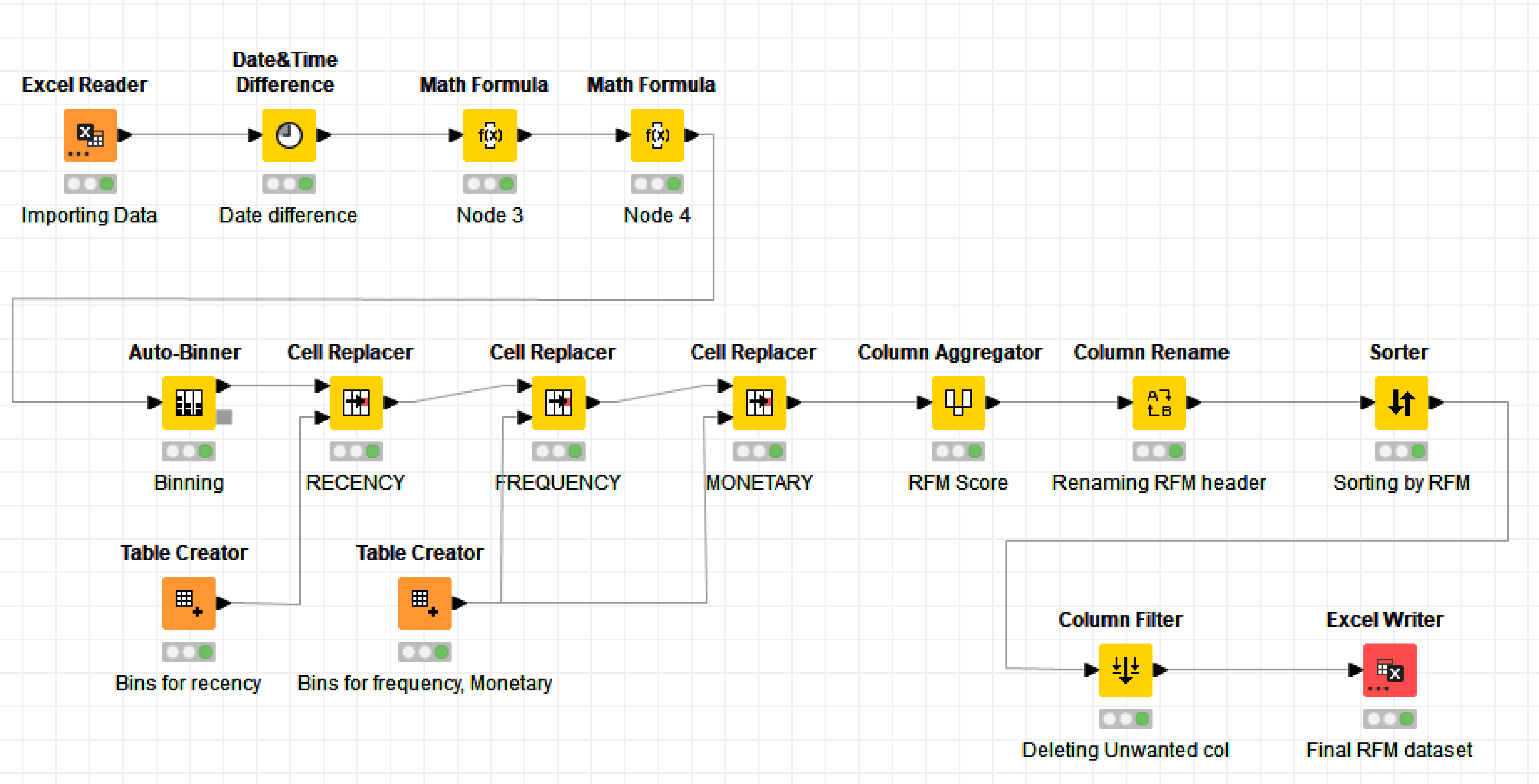
2. Frequency is taken as the total quantity ordered till 1st June 2020.

3. In parameter "Status", the data with labels "Shipped" is used for RFM analysis. Rest of the data is deleted from the provided data set. The other labels such as "Resolved", "Disputed" etc. do not have exact definition. Moreover, inclusion of labels such as "cancelled", "In process" will bring in unnecessary variability to the final RFM score. Moreover the deleted data is around 8% of the total entries, so no significant impact overall.

RFM Analysis

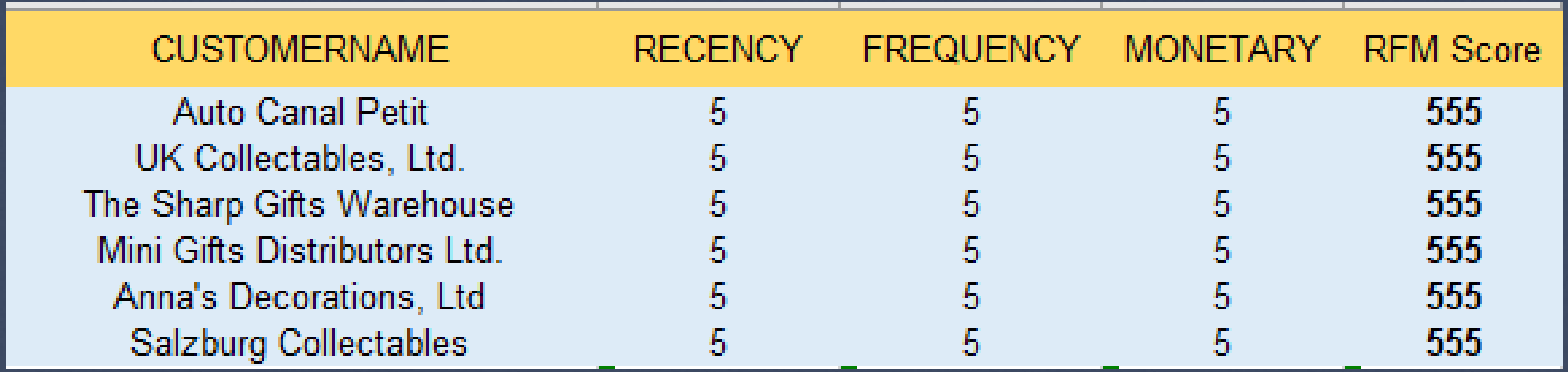
- KNIME and Excel have been used for RFM analysis. KNIME output has been analysed in Excel for better visualisation and conclusion.
- Recency, frequency and monetary values have been binned into 5 categories for Scoring purpose.

KNIME Workflow



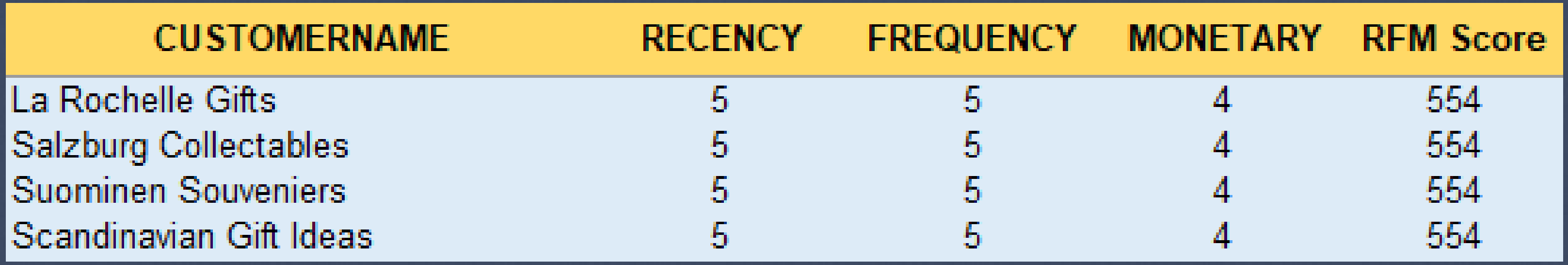
RFM Analysis

- Once Recency, frequency and monetary scores are calculated, these 3 numbers are concatenated into a single number. For example, if Recency, frequency and monetary values are 5,4,5 resp., then after concatenation the RFM score is 545.
- Below is the head of the final RFM scores along with customer name (after Sorting by descending order).



CUSTOMERNAME	RECENCY	FREQUENCY	MONETARY	RFM Score
Auto Canal Petit	5	5	5	555
UK Collectables, Ltd.	5	5	5	555
The Sharp Gifts Warehouse	5	5	5	555
Mini Gifts Distributors Ltd.	5	5	5	555
Anna's Decorations, Ltd	5	5	5	555
Salzburg Collectables	5	5	5	555

Some more RFM score



CUSTOMERNAME	RECENCY	FREQUENCY	MONETARY	RFM Score
La Rochelle Gifts	5	5	4	554
Salzburg Collectables	5	5	4	554
Suominen Souveniers	5	5	4	554
Scandinavian Gift Ideas	5	5	4	554

RFM Analysis

- With R, F and M values ranging from 1 to 5, there will be around 125 possible combination of RFM score, e.g. 555, 454, 234, 121, 111 etc..

- Each customer will have a RFM score depending on their transanction.

- Based on their RFM score, the customers can be segmentated into different groups. Some meaningful names can be assigned to these groups for future reference. These groups can be targeted for future marketing campaigns.

- In our present case study, the customers are segmented into 4 distinct groups.

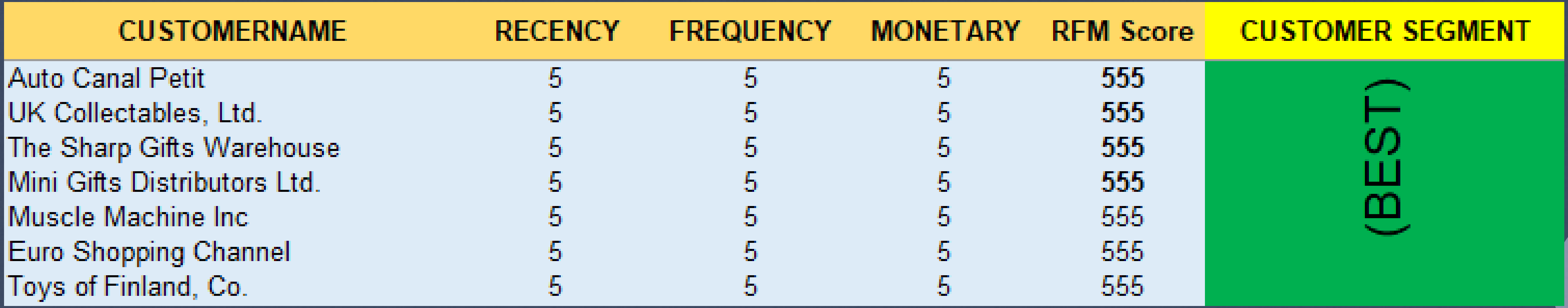
Segmentation based on RFM

Assigning RFM score range for segmentation (4 groups)

Sl. No.	Customer segments	Recency	Frequency	Monetary	Possible combination
1	Champions (best)	4 – 5	4 – 5	4 – 5	555, 545,454,444 etc.
2	Loyal	2 – 4	3 – 4	3 – 4	233, 244, 333, 344 etc.
3	Churning	1 – 2	1 – 3	1 – 2	122, 222, 133, 233 etc.
4	Lost	1	1	1 – 2	111, 112 etc.

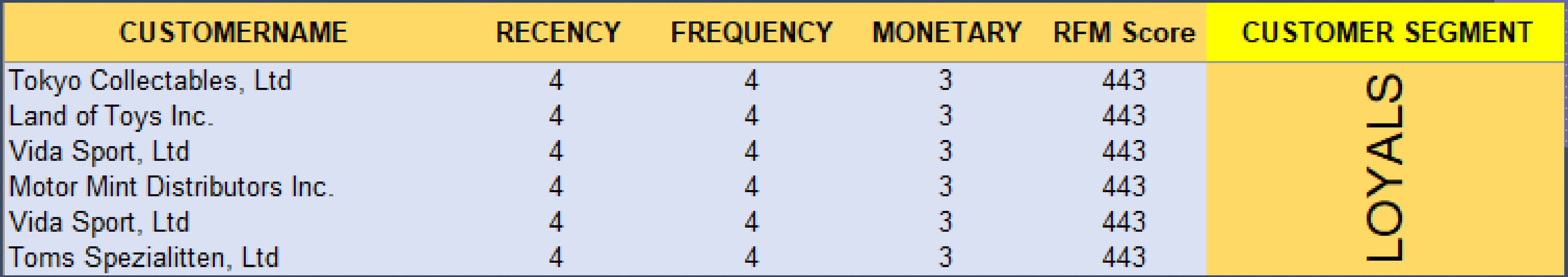
Segmentation based on RFM

❖ BEST 5 CUSTOMERS



CUSTOMERNAME	RECENCY	FREQUENCY	MONETARY	RFM Score	CUSTOMER SEGMENT
Auto Canal Petit	5	5	5	555	(BEST)
UK Collectables, Ltd.	5	5	5	555	
The Sharp Gifts Warehouse	5	5	5	555	
Mini Gifts Distributors Ltd.	5	5	5	555	
Muscle Machine Inc	5	5	5	555	
Euro Shopping Channel	5	5	5	555	
Toys of Finland, Co.	5	5	5	555	

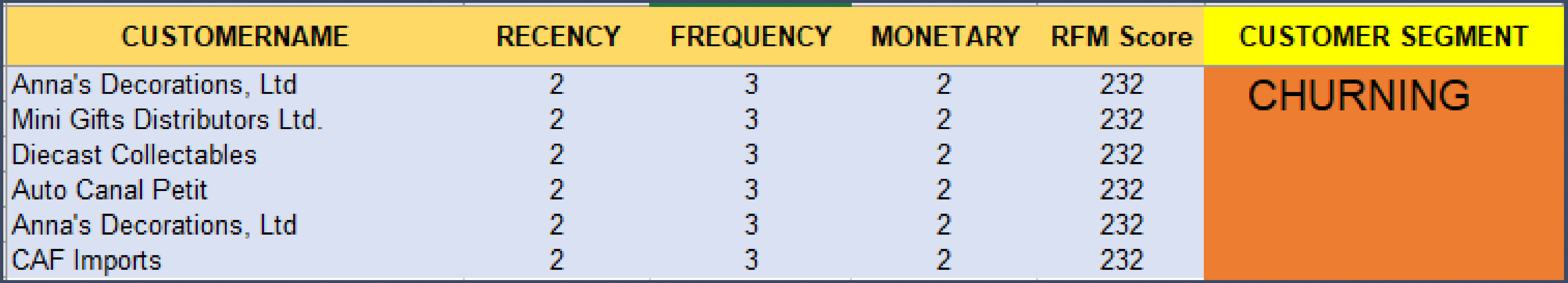
❖ 5 LOYAL CUSTOMERS



CUSTOMERNAME	RECENCY	FREQUENCY	MONETARY	RFM Score	CUSTOMER SEGMENT
Tokyo Collectables, Ltd	4	4	3	443	LOYALS
Land of Toys Inc.	4	4	3	443	
Vida Sport, Ltd	4	4	3	443	
Motor Mint Distributors Inc.	4	4	3	443	
Vida Sport, Ltd	4	4	3	443	
Toms Spezialitten, Ltd	4	4	3	443	

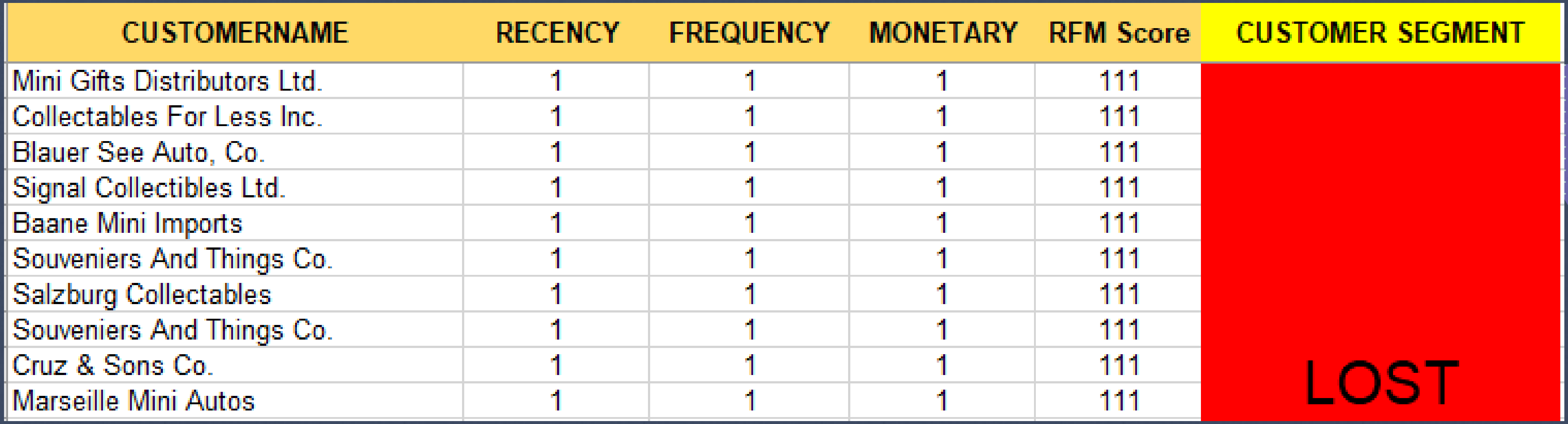
Segmentation based on RFM

❖ TOP 5 CHURNING CUSTOMERS



CUSTOMERNAME	RECENCY	FREQUENCY	MONETARY	RFM Score	CUSTOMER SEGMENT
Anna's Decorations, Ltd	2	3	2	232	CHURNING
Mini Gifts Distributors Ltd.	2	3	2	232	
Diecast Collectables	2	3	2	232	
Auto Canal Petit	2	3	2	232	
Anna's Decorations, Ltd	2	3	2	232	
CAF Imports	2	3	2	232	

❖ TOP 5 LOST CUSTOMERS



CUSTOMERNAME	RECENCY	FREQUENCY	MONETARY	RFM Score	CUSTOMER SEGMENT
Mini Gifts Distributors Ltd.	1	1	1	111	LOST
Collectables For Less Inc.	1	1	1	111	
Blauer See Auto, Co.	1	1	1	111	
Signal Collectibles Ltd.	1	1	1	111	
Baane Mini Imports	1	1	1	111	
Souveniers And Things Co.	1	1	1	111	
Salzburg Collectables	1	1	1	111	
Souveniers And Things Co.	1	1	1	111	
Cruz & Sons Co.	1	1	1	111	
Marseille Mini Autos	1	1	1	111	

Inference/Conclusion

- Customer segmentation can be done into many groups. But in our present case study, we restricted to just 4 core segments.
- Company should start targeted marketing campaigns for these segments.
- For the Best and Loyal customers, face to face interactions, organising customers meet etc. would further strengthen relationships.
- For Churning customers, strategies like cheap deals, max. discounts, enablimg personalized service should help.
- For Lost customers, it is as good as approaching a brand new prospective customer. So, same kind of strategies should work in this scenario.
- Independent clustering techniques like K-means will also help in identifying segments for marketing.