

10/16/2022

MRA Project

Milestone 1

Shailesh Pande
Pune

Table of Contents

• Problem Statement.....	3
• Executive Summary of Data	3
• Statistical Summary of Numeric Data:	4
• Summary of Important Nominal Features:.....	4
• Exploratory Analysis.....	5
• Outlier Analysis	7
• Sales Trends	8
• Product Line and Deal Size Sales Analysis.....	9
• Country Wise Sales Analysis.....	10
• Country Wise – Product Line Wise Sales Analysis.....	11
• Country Wise – Concern Areas	12
• Status Type Analysis.....	13
• Customer Segregation into 4 Segments using RFM analysis	14
• What is RFM?	14
• Parameters used, and assumptions made for RFM analysis	15
• KNIME Workflow Image of Output table head	16
• Customer Segmentation of 89 Customers.....	17
• INFERENCES FROM RFM ANALYSIS AND NAMED SEGMENTS	18
• BEST CUSTOMERS LIST.....	18
• CUSTOMERS ON THE VERGE OF CHURNING	19
• LIST OF LIKELY LOST CUSTOMERS.....	19
• LIST OF LOYAL CUSTOMERS	20

Figures and Tables index:

Figure 1 : Count Plots of Important Features	5
Figure 2 : Order Quantity Analysis	6
Figure 3 : Order Date Analysis.....	6
Figure 4: Outlier Visualization	7
Figure 5 : Sales Trends - Annual/Quarterly/Mthly/Weekly	8
Figure 6 : Product Line and Deal Size Sales Analysis.....	9
Figure 7 : Country wise Sales	10
Figure 8 : Country-Product line Sales.....	11
Figure 9 : Annual Country Sales	12
Figure 10 : Deal Status Type Analysis.....	13
Figure 11 : RFM Explanation	14
Figure 12 : KNIME WORK FLOW OUTPUT	16

Table 1 : Data Summary	3
Table 2 : Feature Details with data type	3
Table 3 : Statistical Summary of Numeric Features	4
Table 4 : Value Counts and Description of Categorical Features.....	4
Table 5 : Product Line and Deal Size Sales Crosstab	9
Table 6 : Segment Name with RFM Score Criteria	16
Table 7 : Segmentation of all 89 customers.....	17
Table 8 : Best Customers List	18
Table 9 : Customers Likely to Churn	19
Table 10 : List of Lost Customers	19
Table 11 : List of Loyal Customers.....	20

- 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.

- Executive Summary of Data

TABLE 1 : DATA SUMMARY

Dataset statistics		Variable types	
Number of variables	20	Numeric	7
Number of observations	2747	Date Time	1
Missing cells	0	Categorical	12
Missing cells (%)	0.0%		
Duplicate rows	0		
Duplicate rows (%)	0.0%		
Total size in memory	429.3 KiB		
Average record size in memory	160.0 B		

TABLE 2 : FEATURE DETAILS WITH DATA TYPE

No	Feature Name	Description	Data Type	No	Feature Name	Description	Data Type
1	ORDERNUMBER :	Order Number	Number (integer)	11	PRODUCTCODE :	Code of Product	String
2	QUANTITYORDERED :	Quantity ordered	Number (integer)	12	CUSTOMERNAME :	customer	String
3	PRICEEACH :	Price of Each item	Number (double)	13	PHONE :	Phone of the customer	String
4	ORDERLINENUMBER :	order line	Number (integer)	14	ADDRESSLINE1 :	Address of customer	String
5	SALES :	Sales amount	Number (double)	15	CITY :	City of customer	String
6	ORDERDATE :	Order Date	Local Date	16	POSTALCODE :	Postal Code of customer	String
7	DAYS_SINCE_LASTORDER :	Days Sinc Lastorder	Number (integer)	17	COUNTRY :	Country customer	String
8	STATUS :	Status of order like Shipped or not	String	18	CONTACTLASTNAME :	Contact person customer	String
9	PRODUCTLINE :	Product line – CATEGORY	String	19	CONTACTFIRSTNAME :	Contact person customer	String
10	MSRP :	Manufacturer's Suggested Retail Price	Number (integer)	20	DEALSIZE :	Size of the deal based on Quantity and Item Price	String

- Statistical Summary of Numeric Data:

TABLE 3 : STATISTICAL SUMMARY OF NUMERIC FEATURES

	count	mean	std	min	25%	50%	75%	max
ORDERNUMBER	2747.0	10259.76	91.88	10100.00	10181.00	10264.00	10334.50	10425.00
QUANTITYORDERED	2747.0	35.10	9.76	6.00	27.00	35.00	43.00	97.00
PRICEEACH	2747.0	101.10	42.04	26.88	68.74	95.55	127.10	252.87
ORDERLINENUMBER	2747.0	6.49	4.23	1.00	3.00	6.00	9.00	18.00
SALES	2747.0	3553.05	1838.95	482.13	2204.35	3184.80	4503.09	14082.80
DAYS_SINCE_LASTORDER	2747.0	1757.09	819.28	42.00	1077.00	1761.00	2436.50	3562.00
MSRP	2747.0	100.69	40.11	33.00	68.00	99.00	124.00	214.00

Inferences on Numeric Data

1. Sales Value for a product code in an order varies from \$ 482 to \$ 14,083
2. Days since last order varies from 42 to 3562 days
3. Quantity ordered for a product code varies from 6 – 97
4. Data is from 6th Jan 2018 to 5th May 2020

- Summary of Important Nominal Features:

TABLE 4 : VALUE COUNTS AND DESCRIPTION OF CATEGORICAL FEATURES

Column	Unique Values	All nominal values
STATUS	6	Shipped, Cancelled, Resolved, On Hold, In Process, Disputed
PRODUCTLINE	7	Classic Cars, Vintage Cars, Motorcycles, Planes, Trucks and Buses, Ships, Trains
PRODUCTCODE	109	
CUSTOMERNAME	89	
CITY	71	
COUNTRY	19	
DEALSIZE	3	Medium, Small, Large

Inferences on important Nominal Features

1. There are a total of eighty-nine customers, who are spread in nineteen countries over seventy-one cities.
2. There are 7 Product lines with a total of 109 products.
3. There are three categories of Deal Size – Large / Medium / Small
4. Status type of a deal are six as listed above.

- Exploratory Analysis

- Count Plots of notable features showing their distribution.

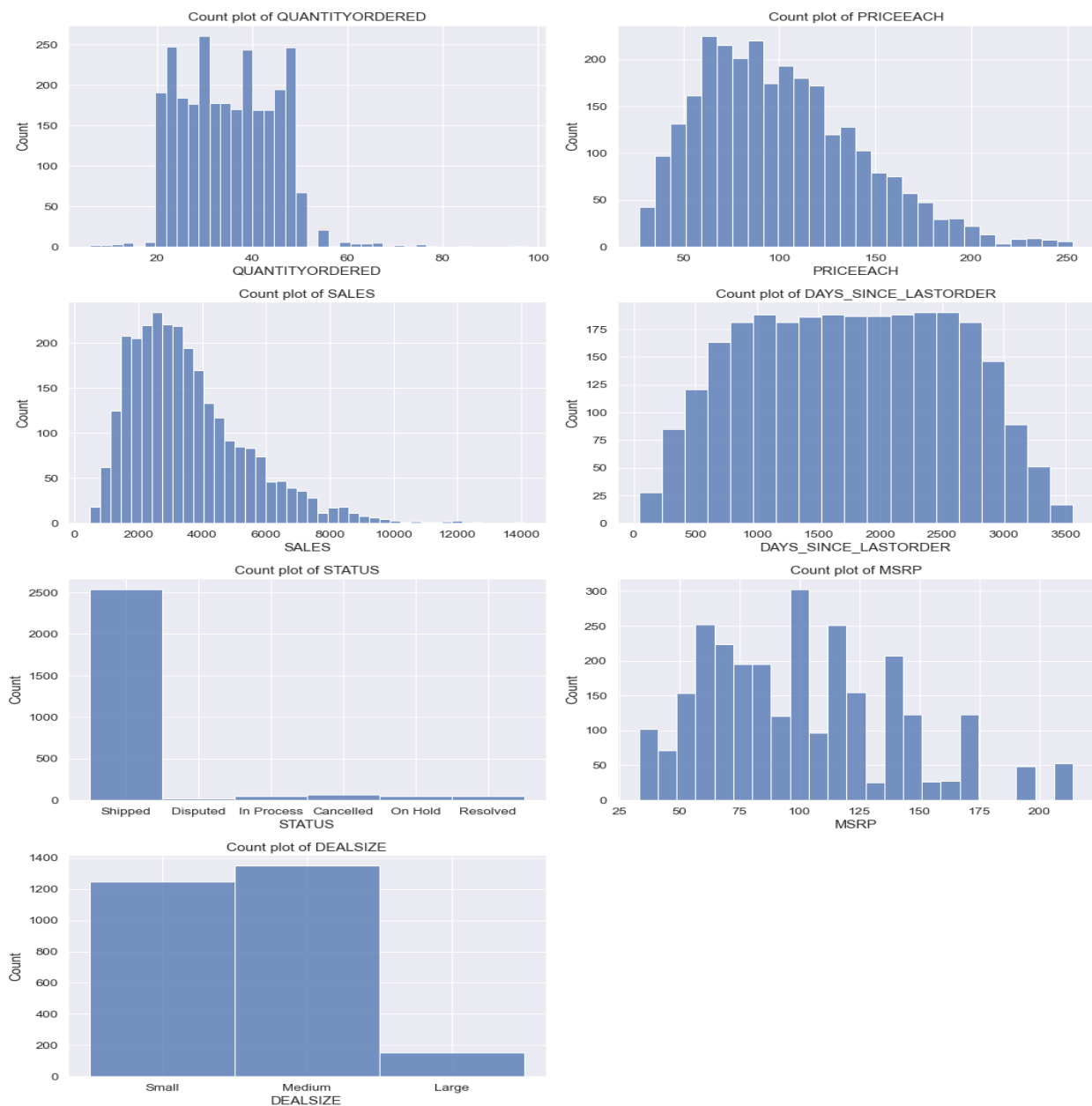


FIGURE 1 : COUNT PLOTS OF IMPORTANT FEATURES

- Inferences:

- Sales Data is right skewed
- Large Deals are very few
- Shipped is the most popular status

- Order Quantity Analysis

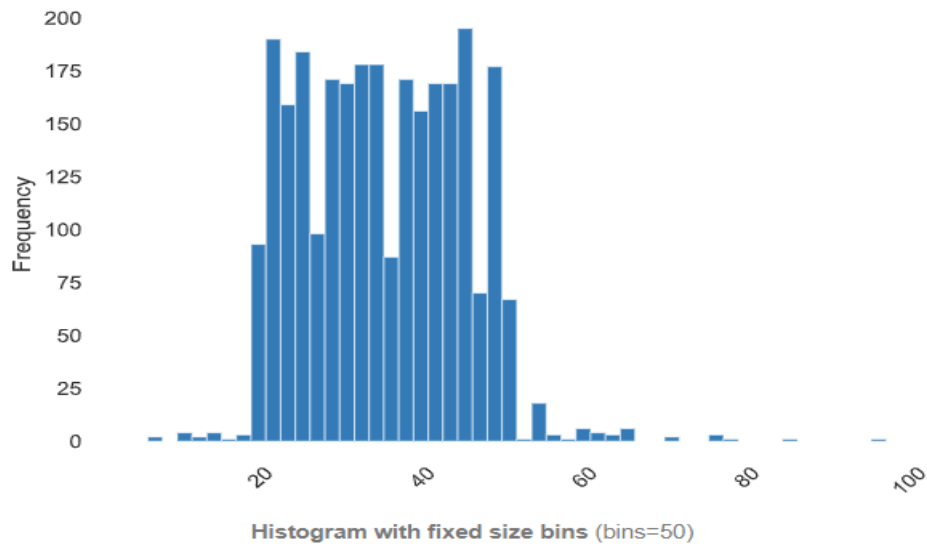


FIGURE 2 : ORDER QUANTITY ANALYSIS

- Quantity ordered for a part number in an Order ranges from 6 to 97, maximum order quantity range being from 20 to 50

- Order Date Data

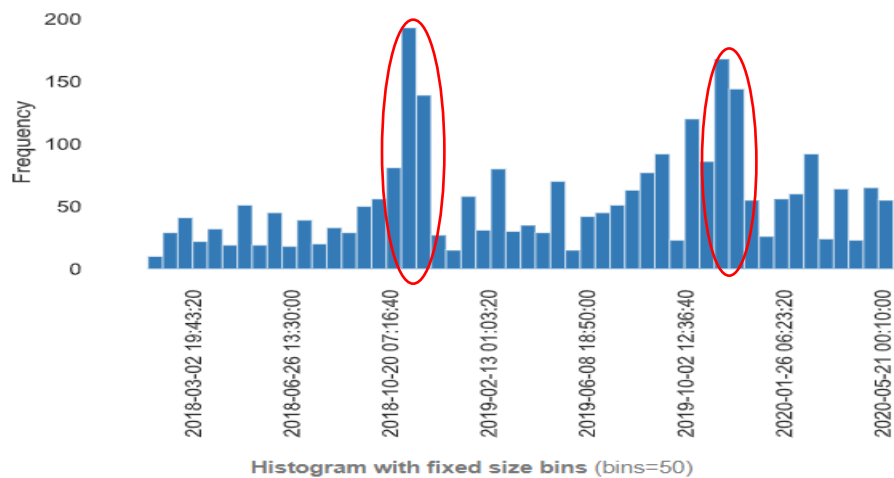


FIGURE 3 : ORDER DATE ANALYSIS

- Nov and Dec months in 2018 and 2019 were peak months

- Outlier Analysis

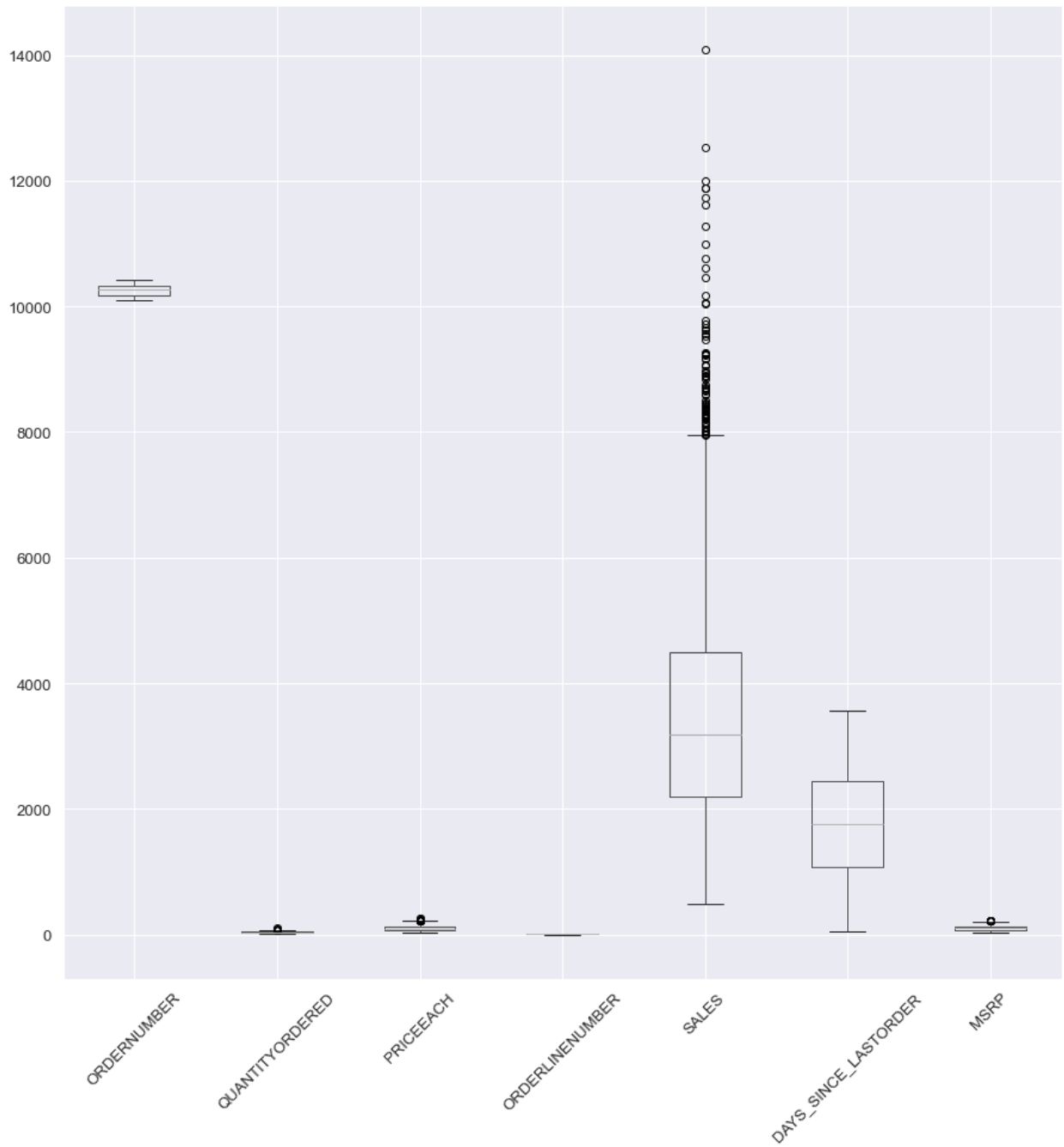


FIGURE 4: OUTLIER VISUALIZATION

- Sales Data has considerable number of outliers

- Sales Trends

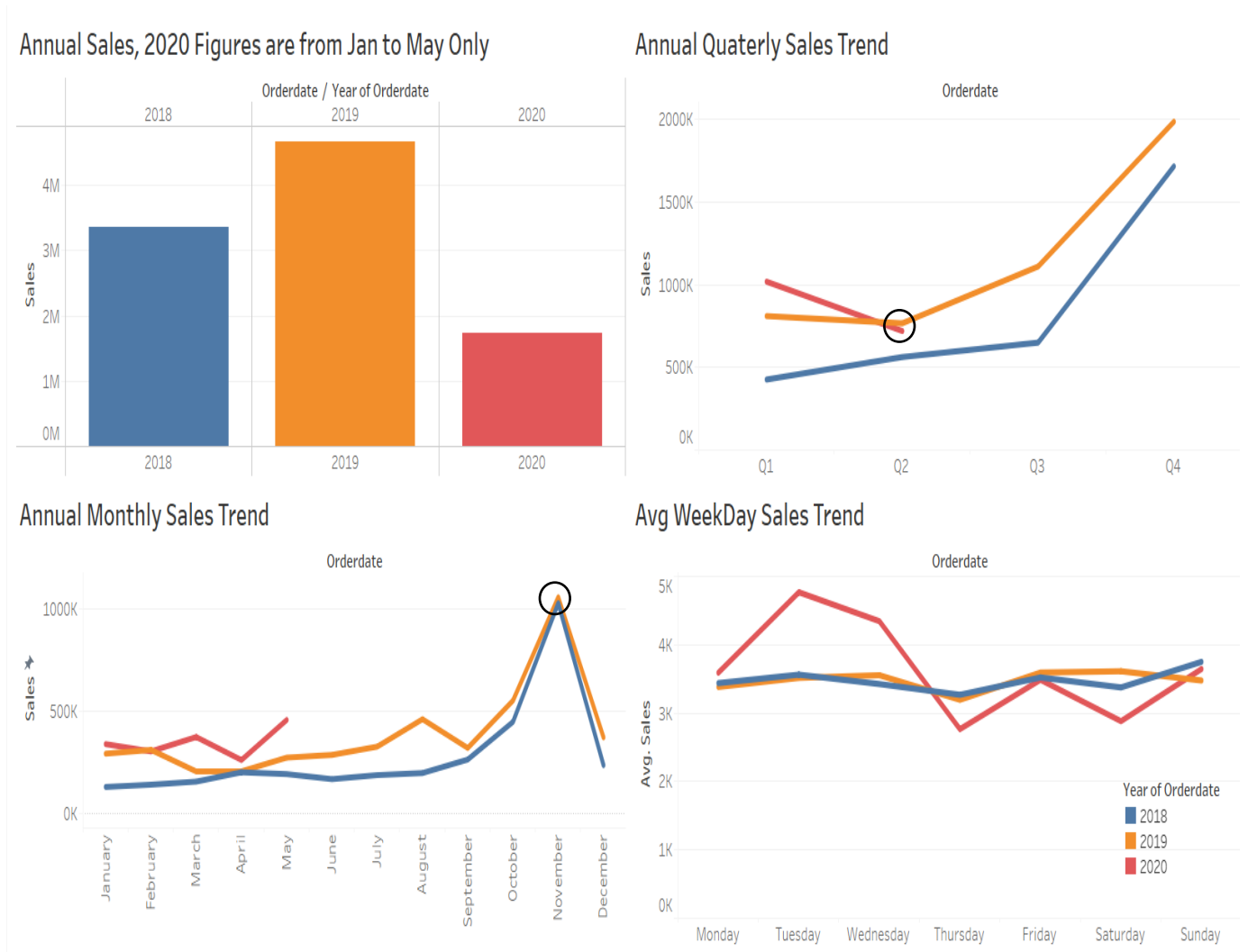


FIGURE 5 : SALES TRENDS - ANNUAL/QUARTERLY/MTHLY/WEEKLY

- 2020 YTD May figures are showing growth over 2019.
- Q2 sales of 2022 are only for April and May and do not include June sales. Hence the Quarterly graph of 2020 is showing a degrowth over Q2, 2019
- November month has been the peak sales both in 2018 and 2019
- Avg Sales on Thursdays are the lowest And highest on Tuesdays

- Product Line and Deal Size Sales Analysis

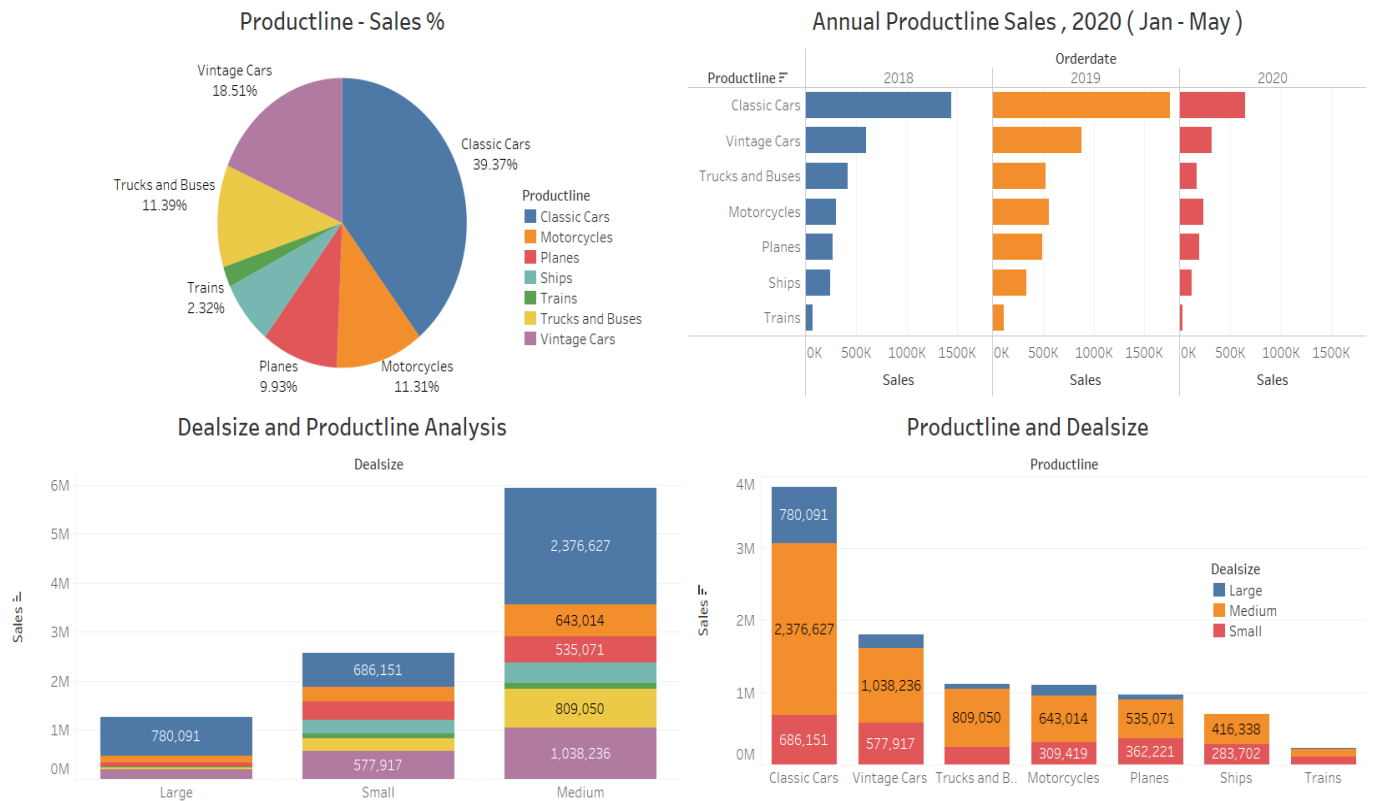


FIGURE 6 : PRODUCT LINE AND DEAL SIZE SALES ANALYSIS

TABLE 5 : PRODUCT LINE AND DEAL SIZE SALES CROSSTAB

Deal Size and Productline Crosstab

Deal Size	Productline							Grand Total
	Classic Cars	Vintage Cars	Trucks and B..	Motorcycles	Planes	Ships	Trains	
Medium	24.35%	10.64%	8.29%	6.59%	5.48%	4.27%	1.16%	60.77%
Small	7.03%	5.92%	2.52%	3.17%	3.71%	2.91%	1.07%	26.33%
Large	7.99%	1.95%	0.58%	1.55%	0.74%	0.09%	0.09%	12.90%
Grand Total	39.37%	18.51%	11.39%	11.31%	9.93%	7.17%	2.32%	100.00%

- Classic Car parts contribute to 39 % Sales which is the most. Train Parts at 2% contribute the least
- 2020 (Jan to May) is showing same trends in Product Line proportion sales.
- Medium Deal Size @ 61 % contribute to maximum Sales. Large Deals contribute only 13 %. There are no large deals in Ships parts business.

Country Wise Sales Analysis

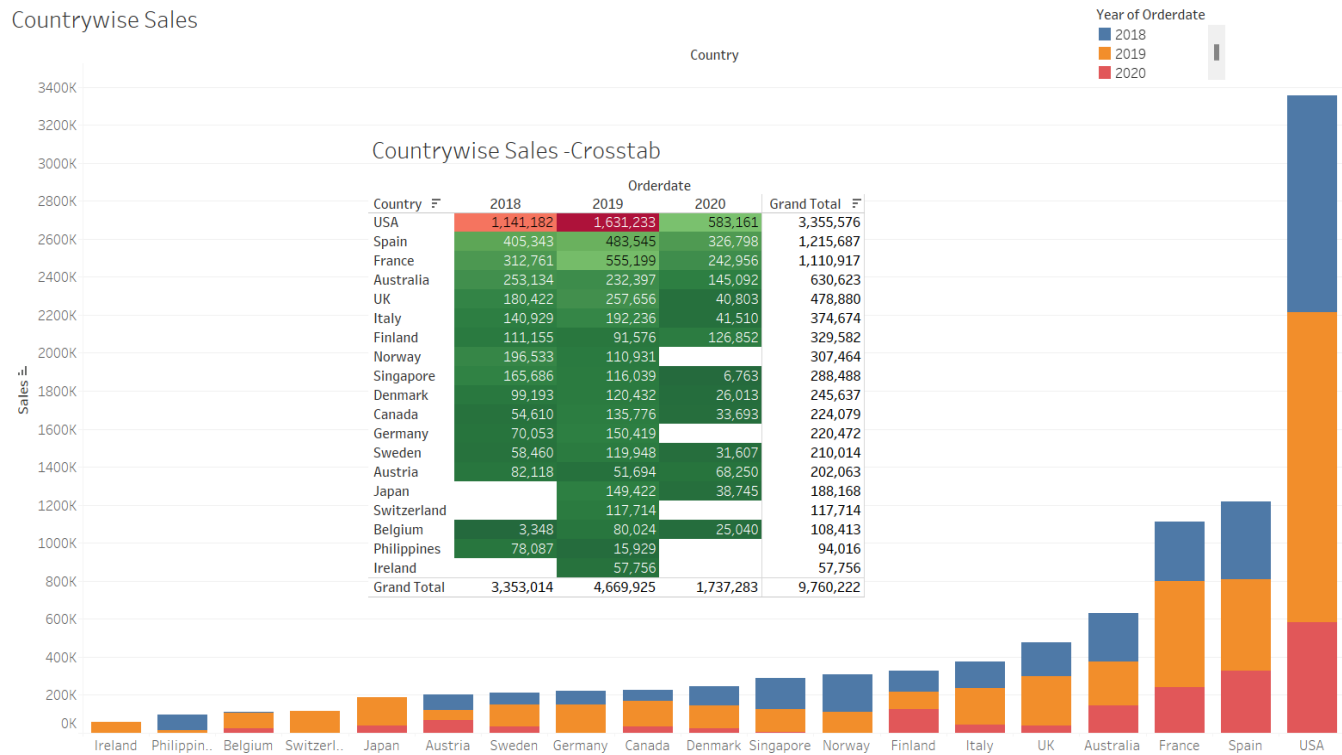


FIGURE 7 : COUNTRY WISE SALES

- USA is the largest market size, followed by Spain and France. In 2019 France had a higher figure than Spain.
- 5 months into 2020, markets of Norway, Germany, Switzerland, Philippines, and Ireland have still not opened their account.
- Japan, Switzerland, and Ireland opened only in 2019. There was no business transaction with these markets in 2018.

- Country Wise – Product Line Wise Sales Analysis

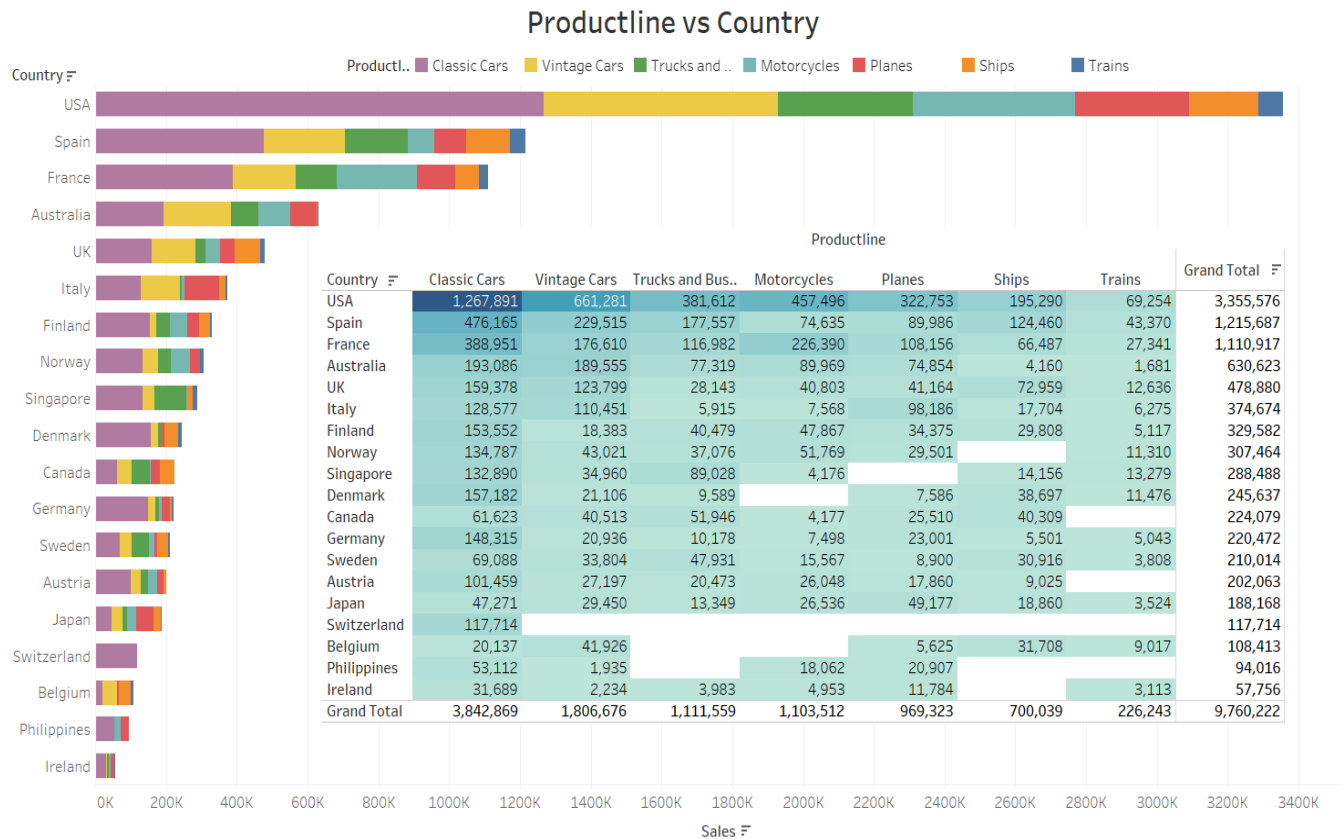


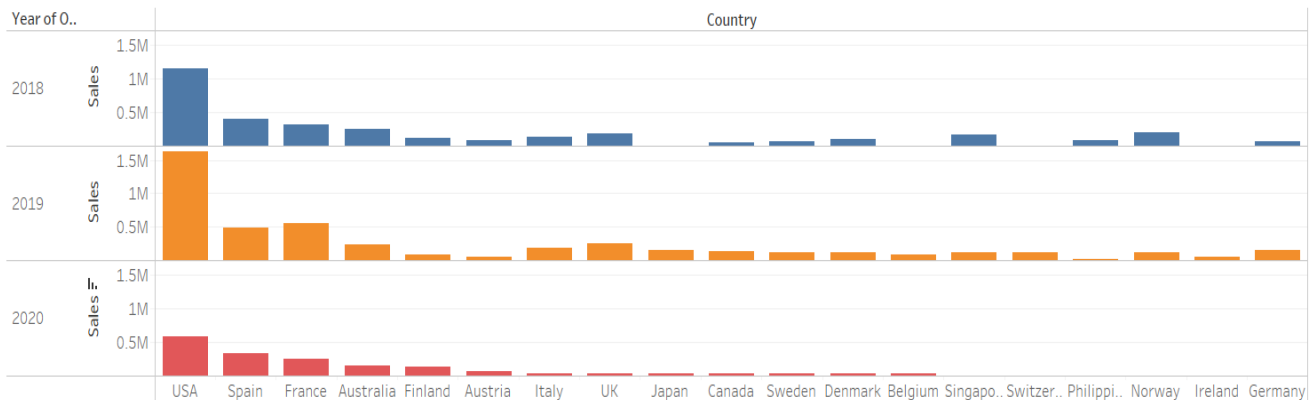
FIGURE 8 : COUNTRY-PRODUCT LINE SALES

- Classic Cars – USA @ \$1.3M is the biggest market
- White Blank space above show zero presence in those respective markets. E.g., Switzerland has a market only for Classic Car Parts.
- Classic Cars @ \$ 3.8 M constitutes 39% of the total sales volume from Jan 2018 to May 2020 of \$ 9.8 M

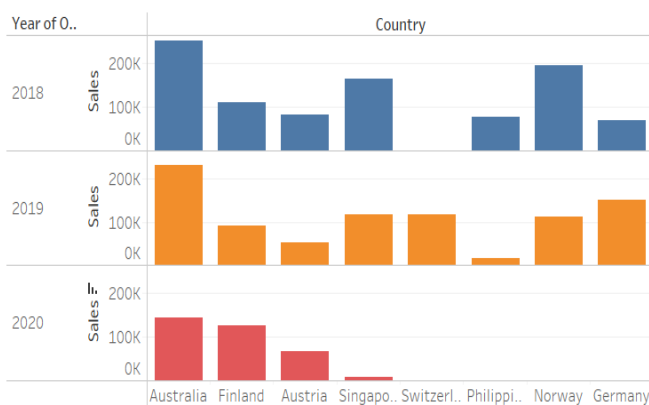
Train Parts business @ \$ 0.2 M constitutes 2% of the total and is the lowest

- Country Wise – Concern Areas

Yearwise Country Sales



Concern Market Areas - Degrowth in 2019 or 2020



Concern Market Areas Jan to May business record

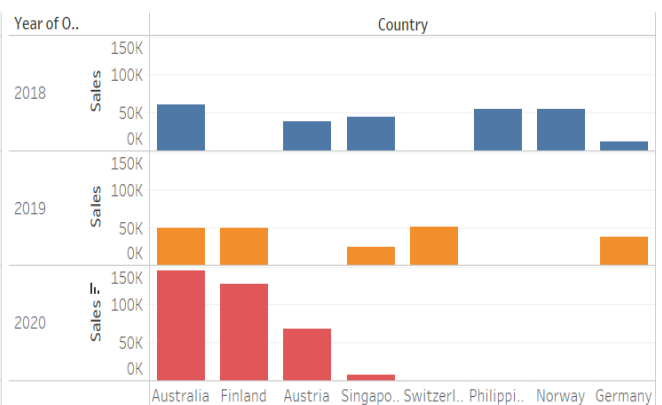


FIGURE 9 : ANNUAL COUNTRY SALES

- Australia, Finland, Austria, Singapore, Philippines, Norway showed a degrowth in 2019 over 2018.
- Up until May 2022, countries like Singapore, Switzerland, Philippines, Norway, Ireland, and Germany have recorded zero sales in 2020.
- Decline of Philippines market is alarming since 2018.
 2018 - \$ 78 k
 2019 - \$ 16 k
 2020 - Nil sales until May 2020
- German market fall is also note worthy
 2018 - \$ 70 k
 2019 - \$ 150 K
 2020 - Nil sales until May 2020

- Status Type Analysis

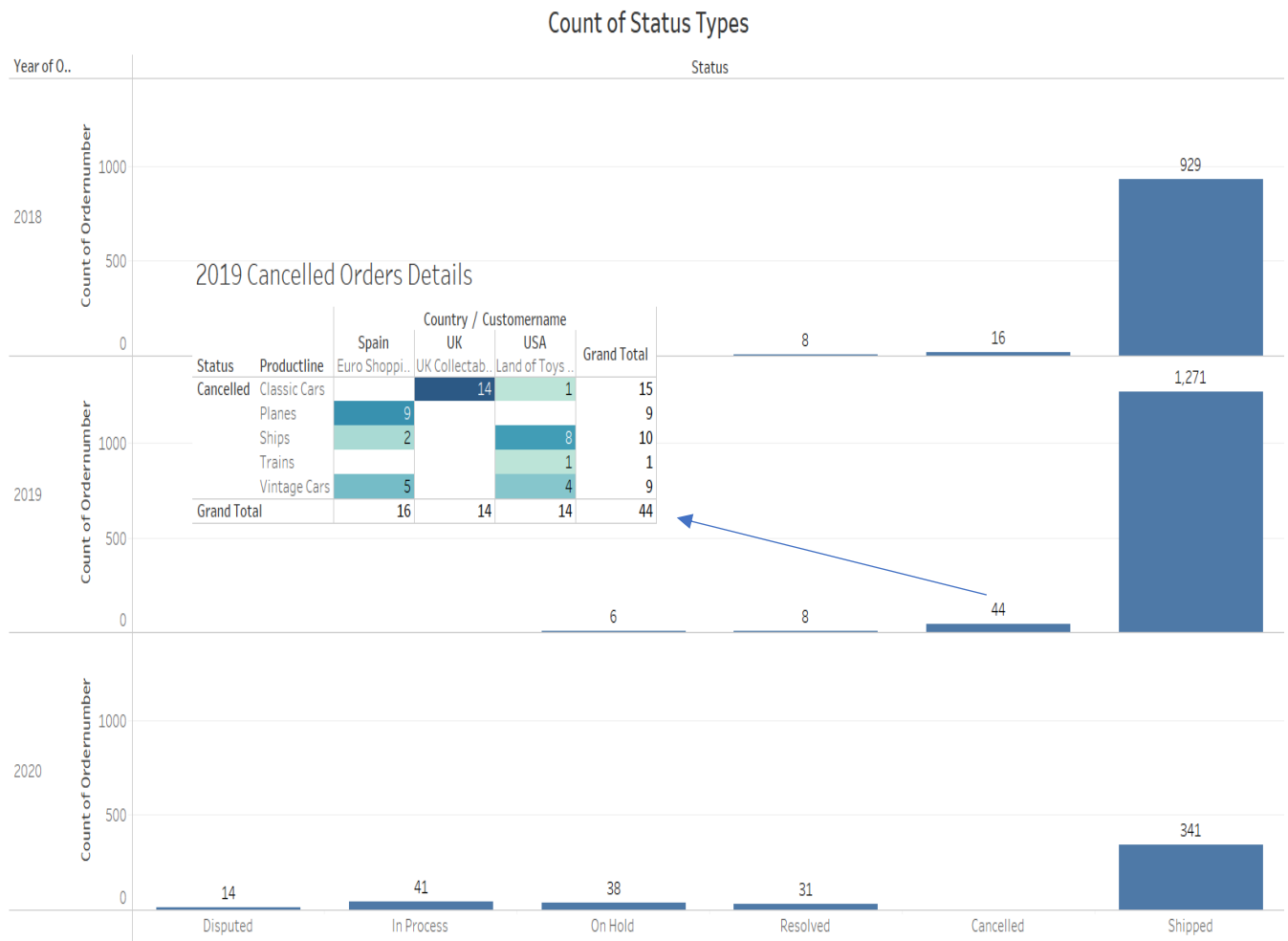


FIGURE 10 : DEAL STATUS TYPE ANALYSIS

- 2019 saw the highest number of cancelled orders
- Cancelled orders were from 3 customers only
 - Euro Shopping Channel- Spain
 - UK Collectables – UK
 - Land of Toys Inc

- Customer Segregation into 4 Segments using RFM analysis
 - What is RFM?

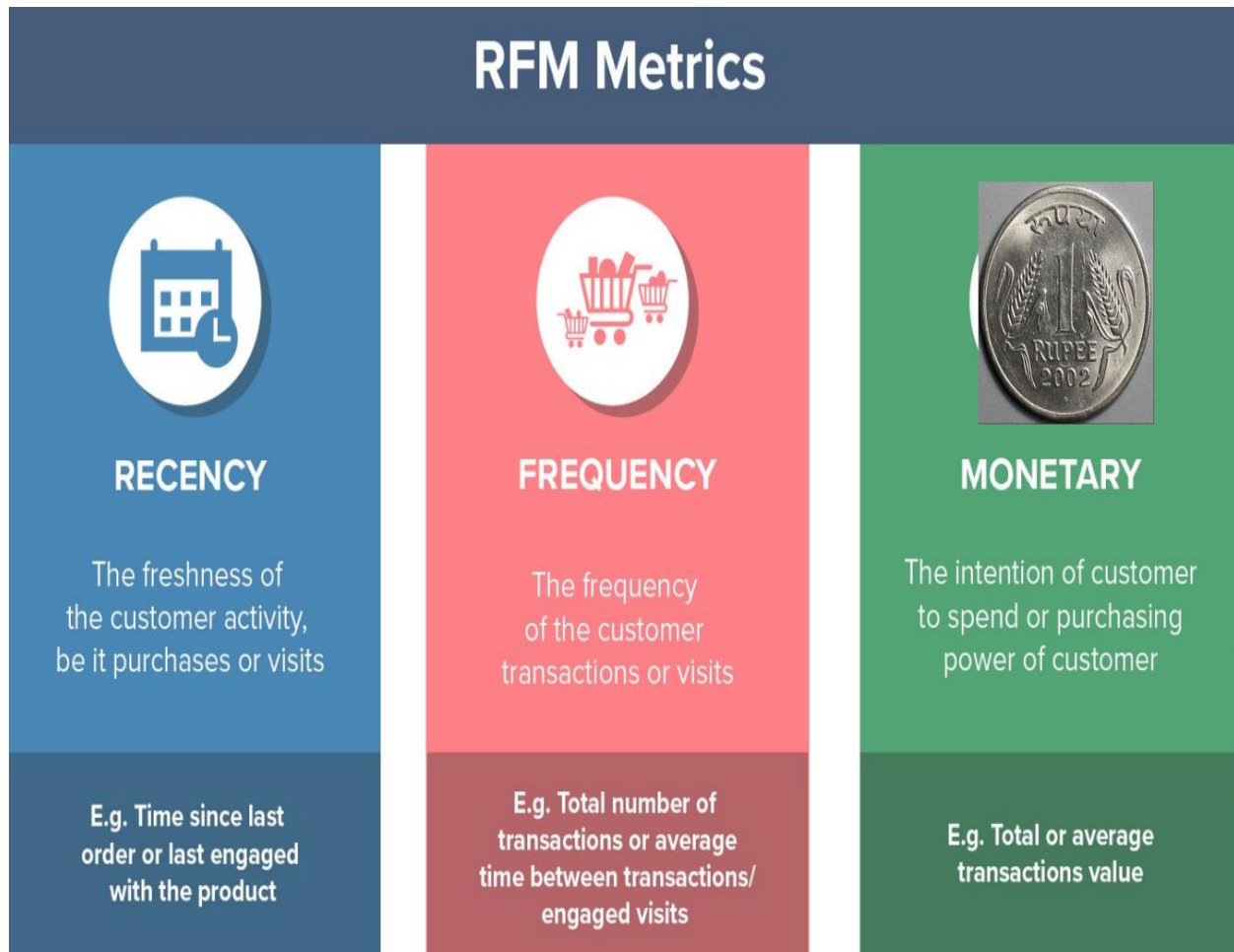


FIGURE 11 : RFM EXPLANATION

RFM segmentation is a great technique to find sets of customers for distinctive treatment. RFM segmentation allows marketers to target specific groups of customers with promotions that are more significant for their actual behavior – and hence supplies much higher proportions of response and increased loyalty.

- Parameters used, and assumptions made for RFM analysis
 - Recency – Minimum of 'Days Since Last Order,' after grouping the data by 'Customer Name.' This would give us the number of days from his last purchase.
 - Frequency – Count of 'Order Numbers' after Grouping the data by 'Customer Name '.
 - Monetary – Sum of 'Sales' after grouping the data by Customer Name
 - For the above features ,4 Bins were created as per the following quantiles
 - Bin 1: 0 – 0.25
 - Bin 2: 0.25 – 0.5
 - Bin 3: 0.5 – 0.75
 - Bin 4: 0.75 -1.0
 - The bins were labeled 0 to 3 for Count of Order Quantity (Frequency) and Sum of Sales (Monetary) from Bin 1 to Bin 4. Any value in the top 25 Percentile (Bin4) would get the highest score of 3 and Bin 1 would get the lowest score of 0.
 - For Recency metric, Bin 1 (lowest number of days since last order) would get the highest score of 3 and Bin 4 would get a score of 0.
 - After obtaining the scores for the three metrics – Recency, Frequency and Monetary, the same are added to obtain the RFM score.
 - Customers are put into 4 Segments as per their RFM scores as under

TABLE 6 : SEGMENT NAME WITH RFM SCORE CRITERIA

Segment Name	RFM Score Criteria	Segment Description	Sample RFM Labels
Platinum	8 to 9	All the 3 metrics of RFM are High	333 , 332 , 233 , 323
Gold	6 to 7	RFM scores are a combination of High and Medium	133 , 223 , 322 , 033 , 132 , 222 , 312
Silver	4 to 5	RFM scores are a combination of Low and Medium	122 , 212 , 311 , 022 , 112 , 211 , 301 , 012 , 021 , 111 , 201 , 210 , 300 ,
Bronze	1 to 3	All the 3 metrics of RFM are low	011 , 101 , 200 , 001 , 010 , 100 , 000

- KNIME Workflow Image of Output table head

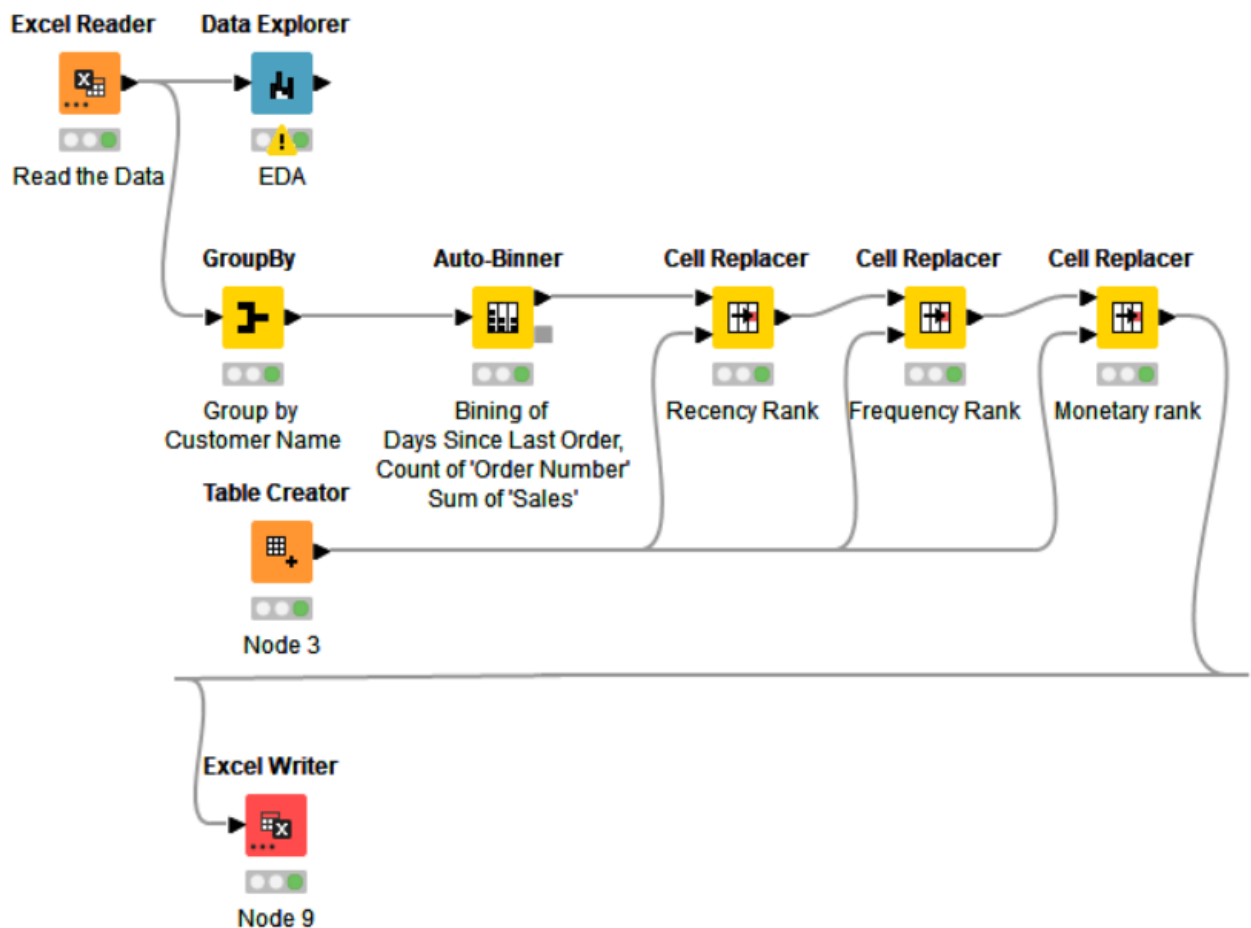


FIGURE 12 : KNIME WORK FLOW OUTPUT

- Customer Segmentation of 89 Customers

TABLE 7 : SEGMENTATION OF ALL 89 CUSTOMERS

Customer Segmentation	RFM Sco	RFM_Lab	Customer Count
Platinum	9	333	11
	9 Total		11
	8	233	2
		323	1
	8 Total		3
Platinum Total			14
Gold	7	133	5
		223	1
		322	4
	7 Total		10
	6	033	2
		132	2
		222	5
		312	2
	6 Total		11
Gold Total			21
Silver	5	122	2
		212	3
		311	3
	5 Total		8
	4	022	1
		112	1
		211	5
		301	1
	4 Total		8
	3	012	2
		021	1
		111	3
		201	2
		210	1
		300	1
	3 Total		10
Silver Total			26
Bronze	2	011	4
		101	2
		200	3
	2 Total		9
	1	001	1
		010	1
		100	7
	1 Total		9
	0	000	10
	0 Total		10
Bronze Total			28
Grand Total			89

- Out of 14 Platinum customers eleven have the highest RFM rating of 333
- Out of 28 Bronze customers ten have lowest RFM rating of 000

- INFERENCES FROM RFM ANALYSIS AND NAMED SEGMENTS
- BEST CUSTOMERS LIST
 - Customers with high Ratings in Recency, Frequency and Monetary metrics, with an RFM score of 8 or 9

TABLE 8 : BEST CUSTOMERS LIST

CUSTOMERNAME	CITY	COUNTRY	RFM Label	Customer Segmentation	RFM Score	DAYS SINCE LAST ORDER	Sum of Sales
Euro Shopping Channel	Madrid	Spain	333	Platinum	9	42	\$912,294
Mini Gifts Distributors Ltd.	San Rafael	USA	333	Platinum	9	219	\$654,858
Australian Collectors, Co.	Melbourne	Australia	333	Platinum	9	229	\$200,995
La Rochelle Gifts	Nantes	France	333	Platinum	9	139	\$180,125
Land of Toys Inc.	NYC	USA	333	Platinum	9	216	\$164,069
The Sharp Gifts Warehouse	San Jose	USA	333	Platinum	9	182	\$160,010
Anna's Decorations, Ltd	North Sydney	Australia	333	Platinum	9	131	\$153,996
Souvenirs And Things Co.	Chatswood	Australia	333	Platinum	9	186	\$151,571
Salzburg Collectables	Salzburg	Austria	333	Platinum	9	188	\$149,799
Reims Collectables	Reims	France	233	Platinum	8	287	\$135,043
Scandinavian Gift Ideas	Boras	Sweden	233	Platinum	8	262	\$134,259
Online Diecast Creations Co.	Nashua	USA	333	Platinum	9	253	\$131,685
Diecast Classics Inc.	Allentown	USA	323	Platinum	8	228	\$122,138
Technics Stores Inc.	Burlingame	USA	333	Platinum	9	241	\$120,783

- CUSTOMERS ON THE VERGE OF CHURNING

- Customers who with high frequency (Frequency Score =3) and high monetary value (Monetary Score = 3) but have not transacted in recent past (Recency Score 0 or 1)

TABLE 9 : CUSTOMERS LIKELY TO CHURN

CUSTOMERNAME	CITY	COUNTRY	RFM Label	Customer Segmentation	DAYS SINCE LAST ORDER	Sum of Sales
Dragon Souveniers, Ltd.	Singapore	Singapore	033	Gold	649	\$172,990
Rovelli Gifts	Bergamo	Italy	033	Gold	1032	\$137,956
Muscle Machine Inc	NYC	USA	133	Gold	502	\$197,737
AV Stores, Co.	Manchest	UK	133	Gold	421	\$157,808
Danish Wholesale Import	Kobenhavn	Denmark	133	Gold	499	\$145,042
Saveley & Henriot, Co.	Lyon	France	133	Gold	586	\$142,874
L'ordine Souveniers	Reggio Em	Italy	133	Gold	493	\$142,601

- LIST OF LIKELY LOST CUSTOMERS

- Customers whose Recency Metric Score is zero and Frequency and Monetary scores are also low

TABLE 10 : LIST OF LOST CUSTOMERS

CUSTOMERNAME	CITY	COUNTRY	RFM Label	Customer Segmentation	RFM Score	DAYS SINCE LAST ORDER	Sum of Sales
Diecast Collectables	Boston	USA	001	Bronze	1	672	\$70,860
Alpha Cognac	Toulouse	France	000	Bronze	0	675	\$70,488
Australian Collectables, Ltd	Glen Waverly	Australia	010	Bronze	1	1018	\$64,591
Clover Collections, Co.	Dublin	Ireland	000	Bronze	0	659	\$57,756
Gift Ideas Corp.	Glendale	USA	000	Bronze	0	947	\$57,294

- LIST OF LOYAL CUSTOMERS

- These are customers whose frequency is highest but recency and monetary metrics may not be the highest

TABLE 11 : LIST OF LOYAL CUSTOMERS

CUSTOMERNAME	CITY	COUNTRY	RFM Label	Customer Segmentati on	RFM Score	DAYS SINCE LAST ORDER	Sum of Sales
Euro Shopping Channel	Madrid	Spain	333	Platinum	9	42	\$912,294
Mini Gifts Distributors Ltd.	San Rafael	USA	333	Platinum	9	219	\$654,858
Australian Collectors, Co.	Melbourne	Australia	333	Platinum	9	229	\$200,995
La Rochelle Gifts	Nantes	France	333	Platinum	9	139	\$180,125
Land of Toys Inc.	NYC	USA	333	Platinum	9	216	\$164,069
The Sharp Gifts Warehouse	San Jose	USA	333	Platinum	9	182	\$160,010
Anna's Decorations, Ltd	North Sydney	Australia	333	Platinum	9	131	\$153,996
Souvenirs And Things Co.	Chatswood	Australia	333	Platinum	9	186	\$151,571
Salzburg Collectables	Salzburg	Austria	333	Platinum	9	188	\$149,799
Online Diecast Creations Co.	Nashua	USA	333	Platinum	9	253	\$131,685
Technics Stores Inc.	Burlingame	USA	333	Platinum	9	241	\$120,783
Reims Collectables	Reims	France	233	Platinum	8	287	\$135,043
Scandinavian Gift Ideas	Boras	Sweden	233	Platinum	8	262	\$134,259
Muscle Machine Inc	NYC	USA	133	Gold	7	502	\$197,737
AV Stores, Co.	Manchester	UK	133	Gold	7	421	\$157,808
Danish Wholesale Imports	Kobenhavn	Denmark	133	Gold	7	499	\$145,042
Saveley & Henriot, Co.	Lyon	France	133	Gold	7	586	\$142,874
L'ordine Souvenirs	Reggio Emilia	Italy	133	Gold	7	493	\$142,601
Dragon Souvenirs, Ltd.	Singapore	Singapore	033	Gold	6	649	\$172,990
Rovelli Gifts	Bergamo	Italy	033	Gold	6	1032	\$137,956
Handji Gifts& Co	Singapore	Singapore	132	Gold	6	488	\$115,499
Mini Creations Ltd.	New Bedford	USA	132	Gold	6	595	\$108,951