

Marketing & Retail Analytics. Milestone 1

THILAK RAJ | Batch: 18 FEB 2022

Content of the Presentation:

- **Problem Statement.**
- **Data Analysis**
 - Data Info, Shape and Summary of the Data
- **Exploratory Data Analysis and Inferences.**
 - Univariate, Bivariate and Multivariate Analysis
- **Customer Segmentation using RFM analysis.**
 - Tool used, Parameter used and Assumptions made
 - Output table head
 - KNIME work flow
- **KNIME Inferences and identified segments.**
 - Best Customers
 - Customers on the verge of churning
 - Lost customers
 - Loyal 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.

Agenda:

Agenda 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 recommend them customized marketing strategies for different segments of customers.

Executive Summary of the data:

We have received the 3 years data of automobile part manufacture . Consisting 2747 entries with 20 variable details regarding the demography of the product and customer information.

Data Analysis:

```
<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
```

Data Info:

- Dataset is having 20 variables.
- 12 are Categorical variables, 7 are numerical variables and one is date and time field variable.
- Found no missing values.
- Found No duplicated data.
- Total number of observations are 2747.

Data Analysis:

Data Summary:

	count	mean	std	min	25%	50%	75%	max
ORDERNUMBER	2747.0	10259.761558	91.877521	10100.00	10181.000	10264.00	10334.500	10425.00
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
ORDERLINENUMBER	2747.0	6.491081	4.230544	1.00	3.000	6.00	9.000	18.00
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

Data Shape:

- The Shape of the data is (2747, 20).

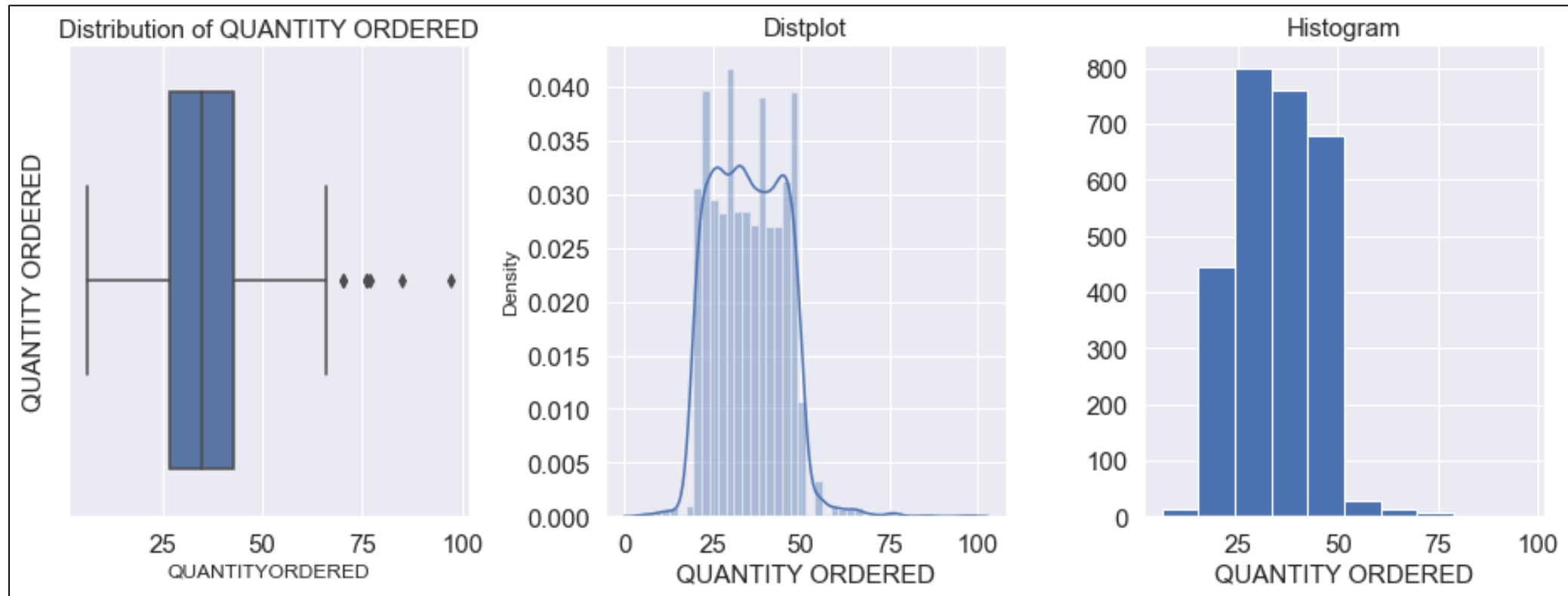
Data Assumption:

- The data is about an automobile parts manufacturing company. They have provided the data collected of transactions for 3 years with 2747 entries
- This data more or less reflects the purchasing behaviour of customers in different categories . 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.
- 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.
- Manufacturer's Suggested Retail Price(MSRP) for each product code is decided but we found that this is not matching with Price of Each item & is inconsistent with MSRP.

Exploratory Data Analysis

Univariate Analysis:

QUANTITY ORDERED:

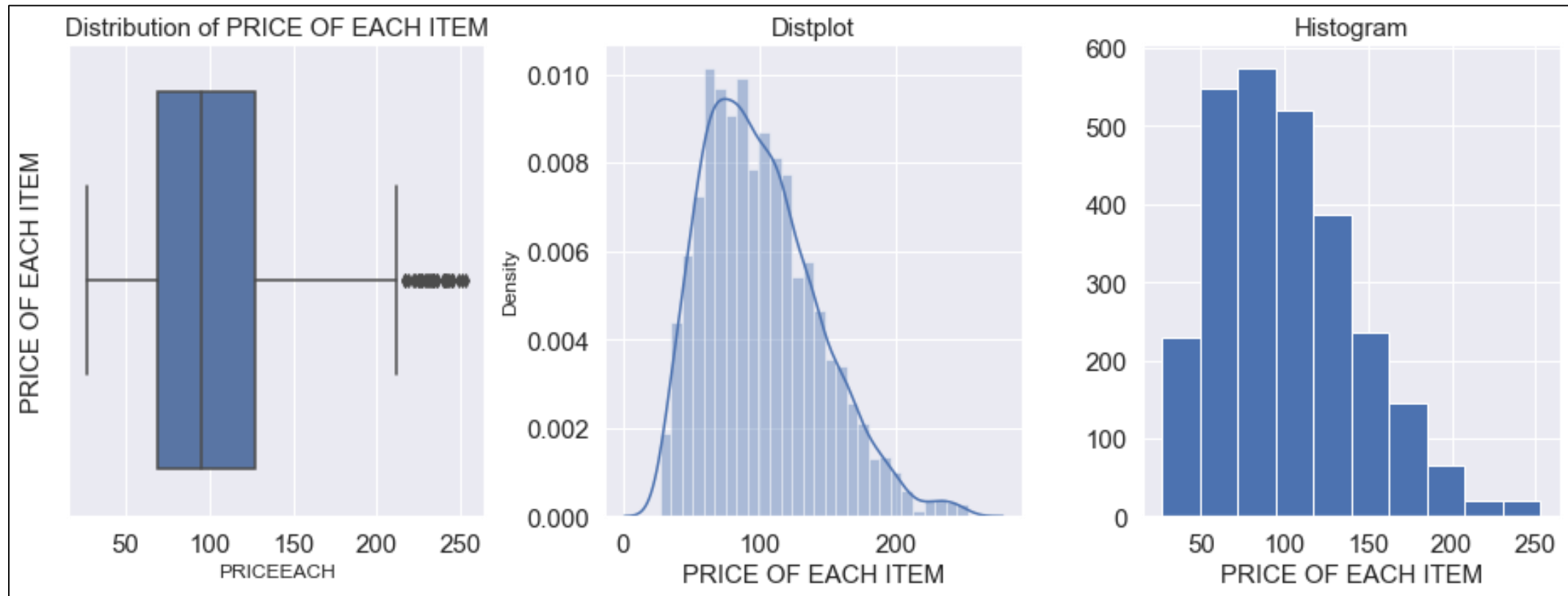


- Outliers are present in this variable and the data is not perfectly Normally distributed

Exploratory Data Analysis

Univariate Analysis:

PRICE OF EACH ITEM:

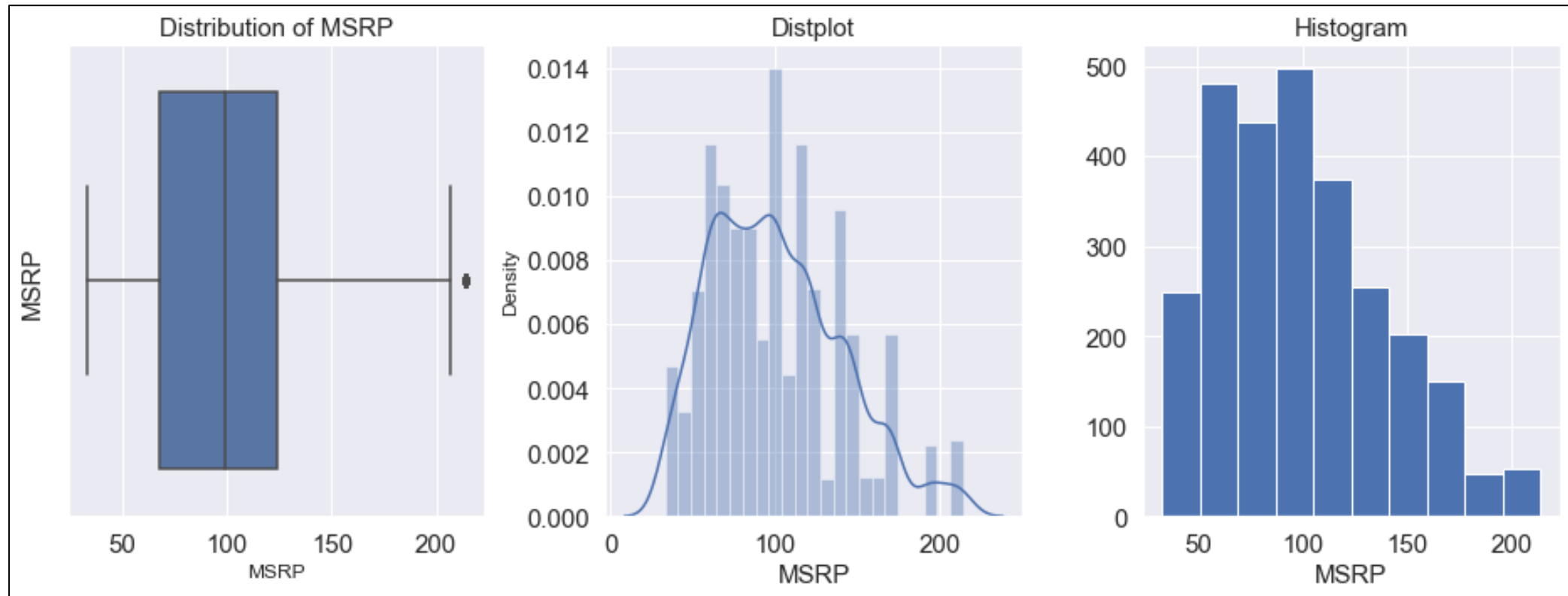


- Outliers are present in this variable and the data has slightly right skewed distribution.

Exploratory Data Analysis

Univariate Analysis:

Manufacturer's Suggested Retail Price(MSRP):

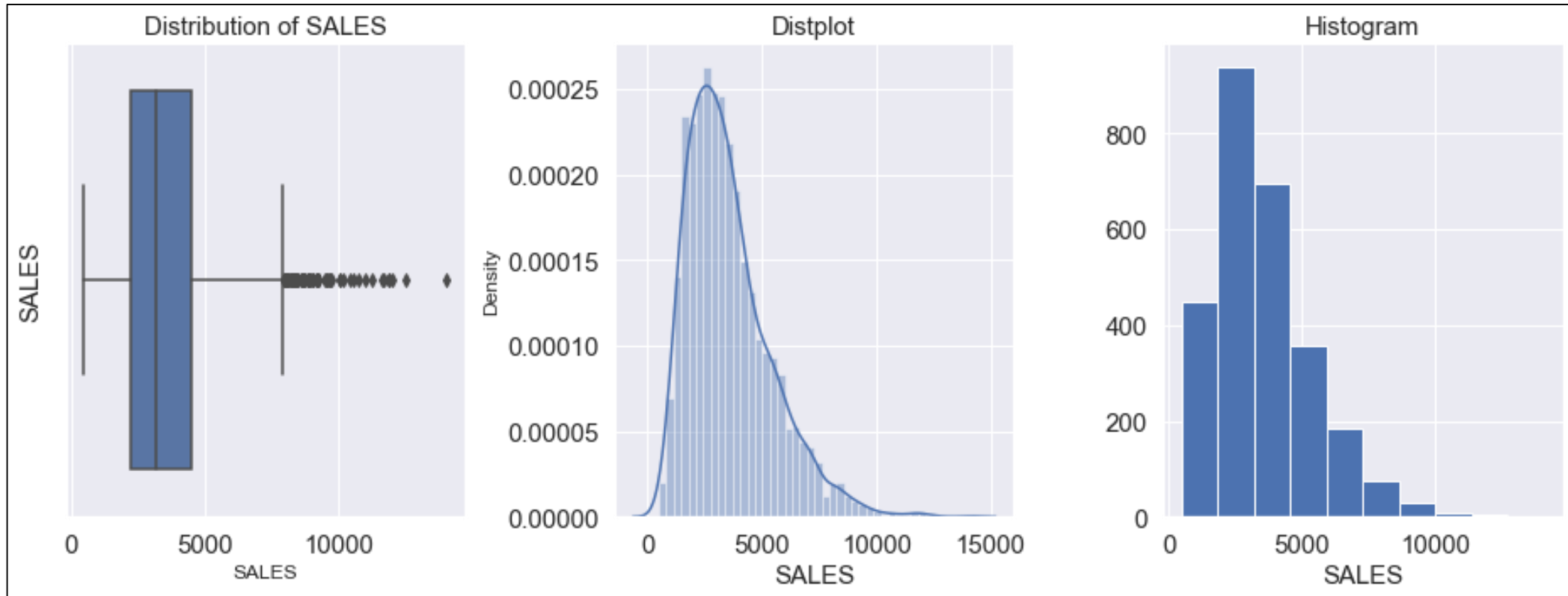


- Very less number of outliers are present in this variable and the data is right skewed.

Exploratory Data Analysis

Univariate Analysis:

SALES:



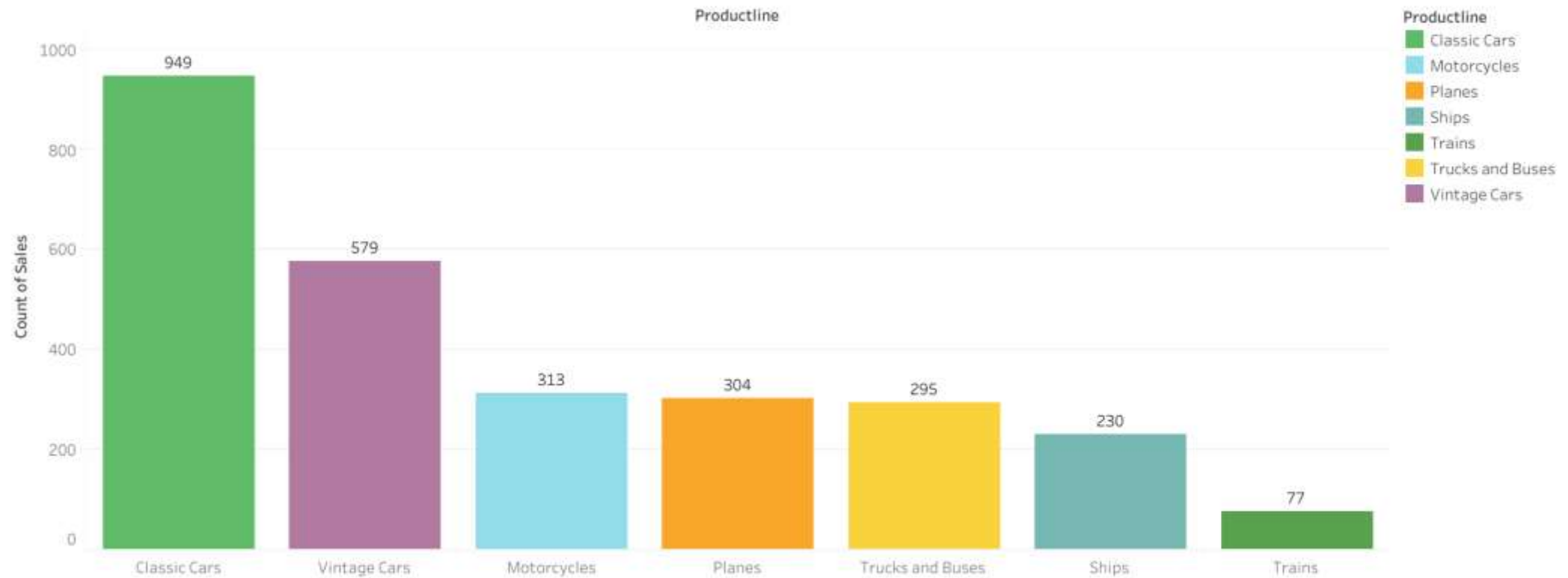
- Outliers are present in this variable and the data has slightly right skewed distribution.

Exploratory Data Analysis

Univariate Analysis:

PRODUCT LINE:

Sales vs Productline



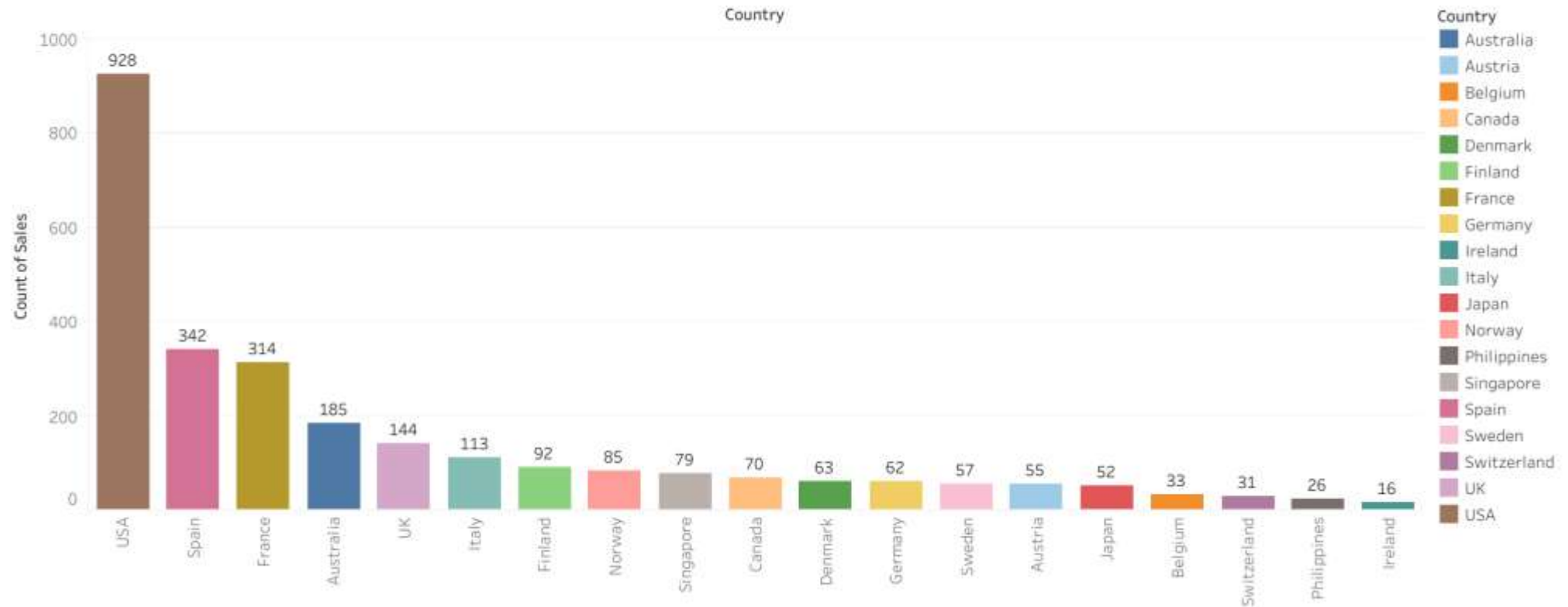
- Classic car parts are the most frequently ordered product lines and the trains parts are less compared to others (Excluding the ordered qty).

Exploratory Data Analysis

Univariate Analysis:

COUNTRY (Order Lines) :

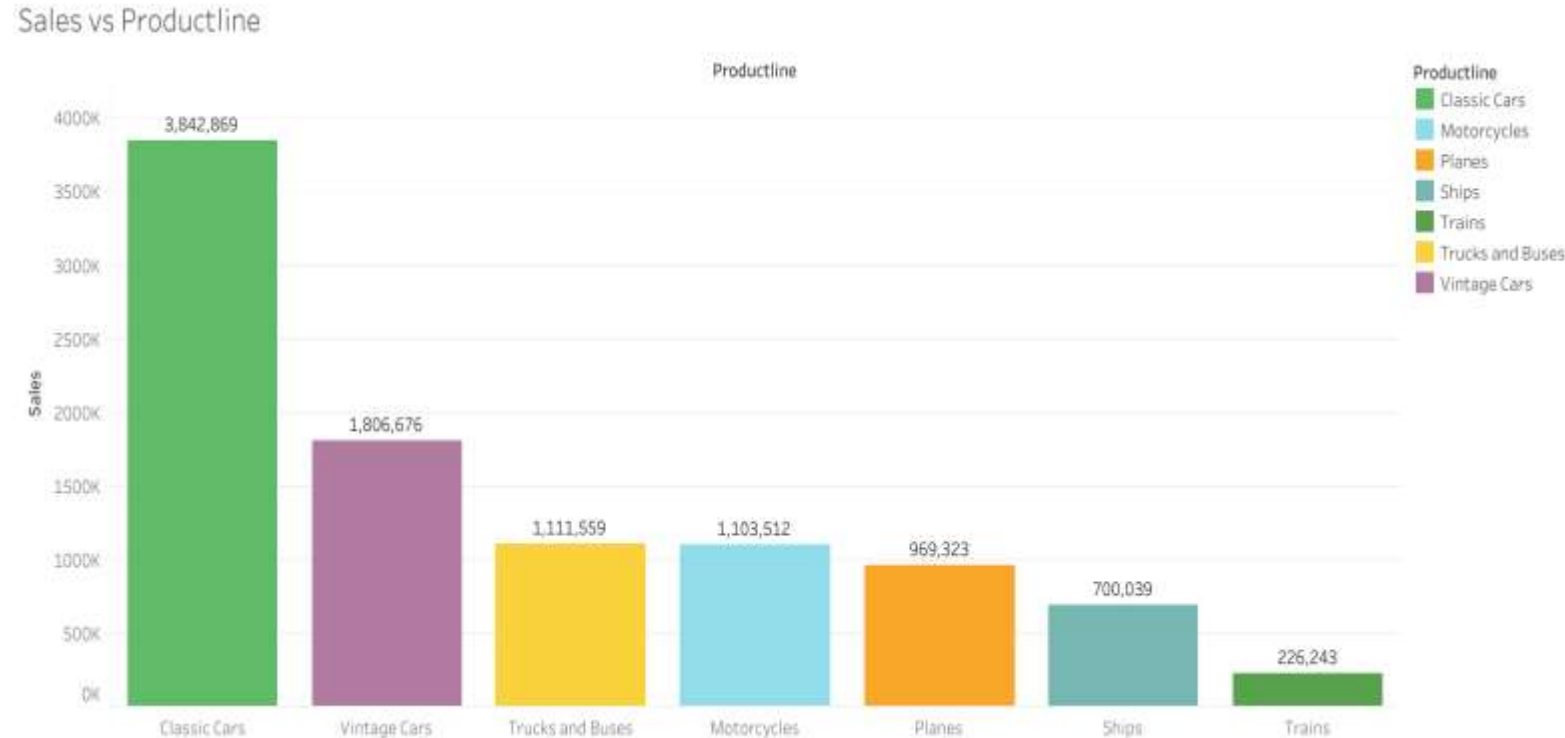
Sales vs Country



- USA has most number of order lines and Ireland has the less number of order lines (Excluding the qty and)

Exploratory Data Analysis

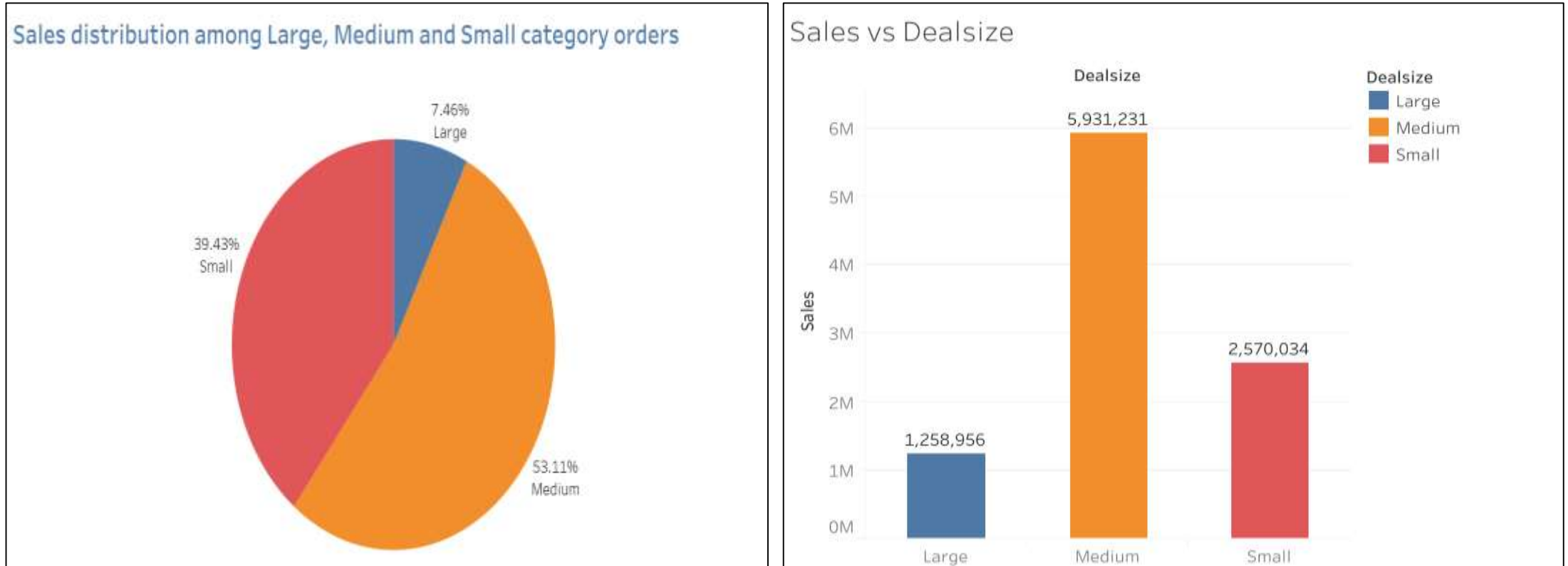
Bivariate Analysis:



- The sale 'Classic Cars' is highest among all other product line categories with the sale of almost 4Millions followed by 'Vintage Cars' which is less than 50% of the 'Classic cars'.
- The 'Trains' product line type got the least number of sales in 3 years (approx. 0.2Million).

Exploratory Data Analysis

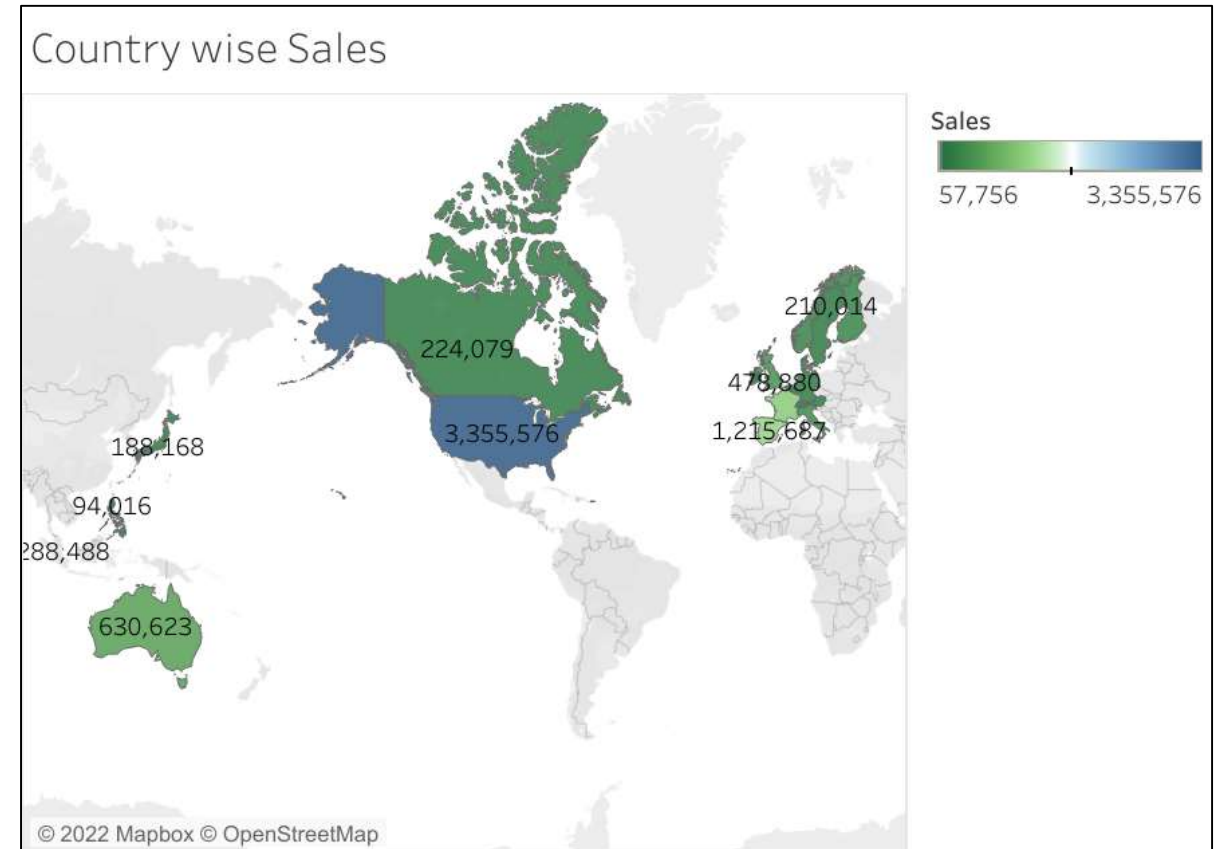
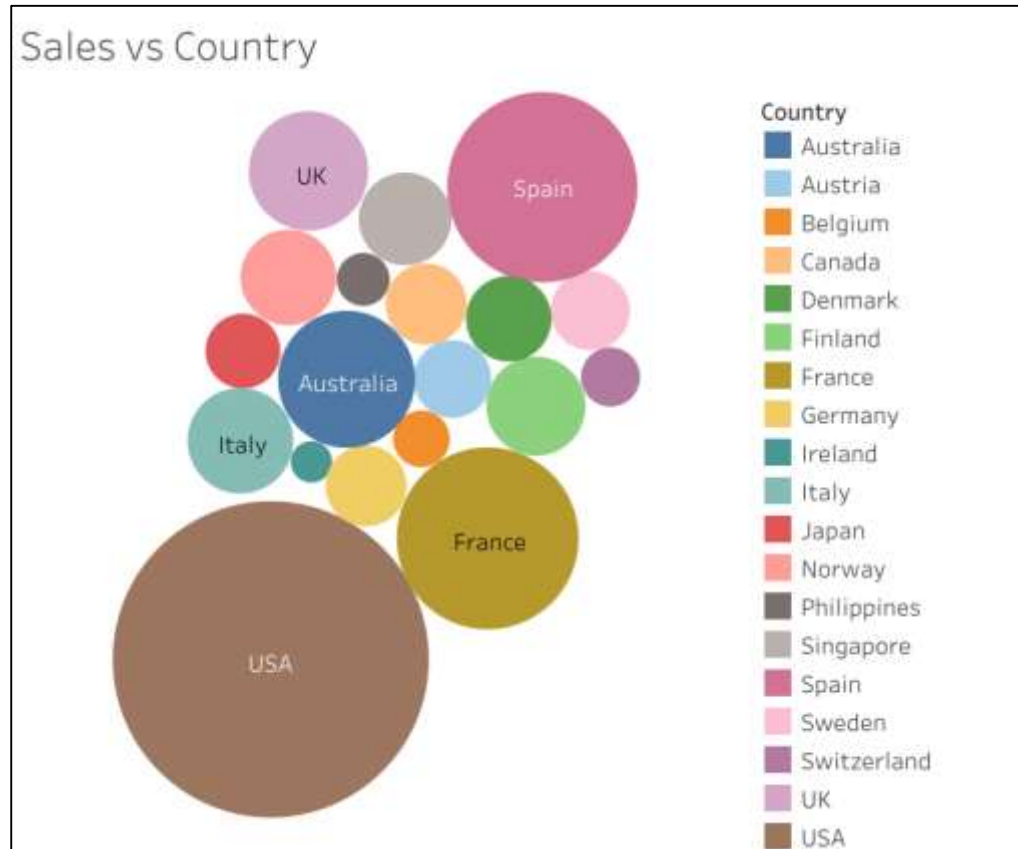
Bivariate Analysis:



- Sales Distribution is skewed. The medium Deal size has more share with 53.11% followed by small deals with 39.43% of the total sale.

Exploratory Data Analysis

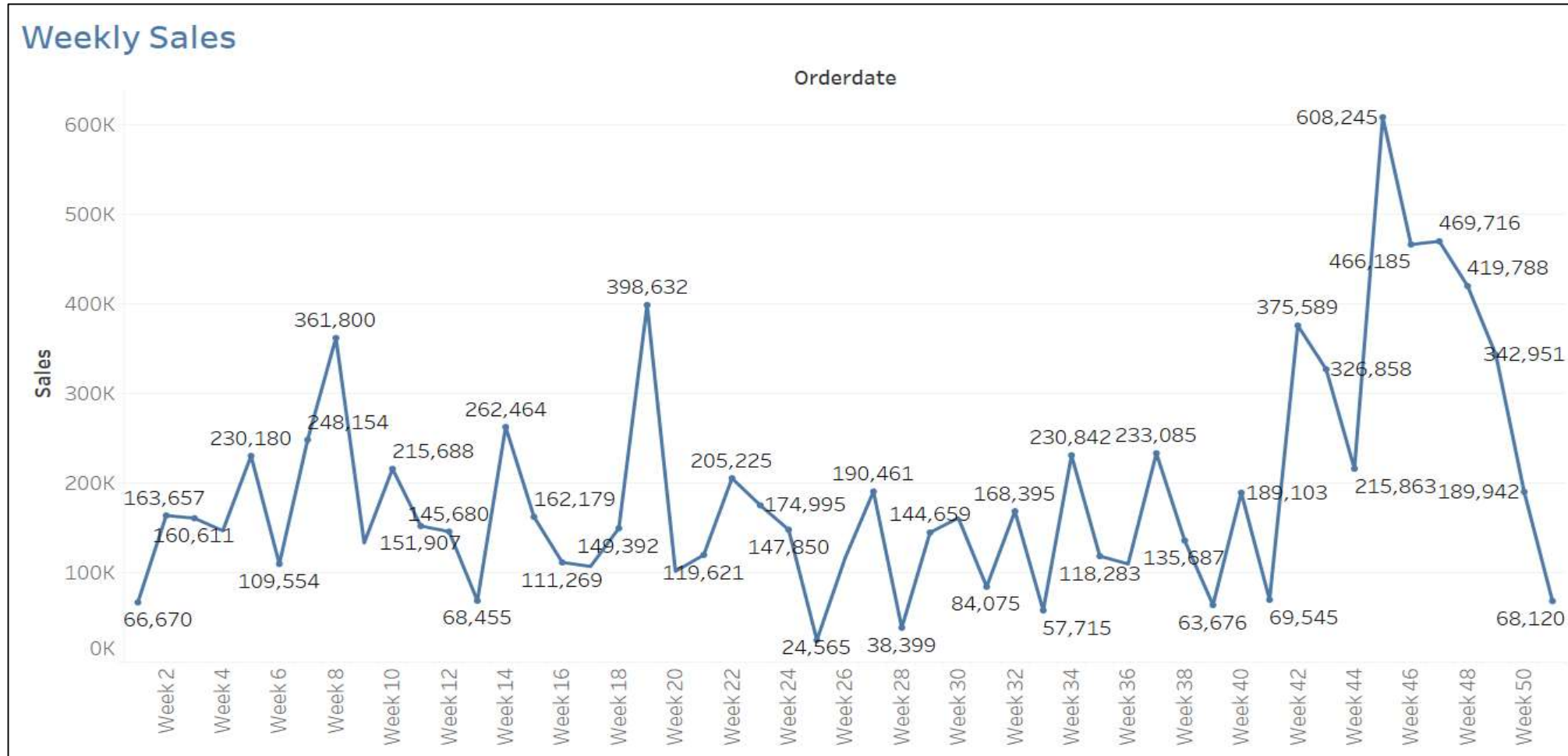
Bivariate Analysis:



- USA is the large market for parts compare to other countries with the sale of 3.3M followed by Spain (1.2M) and France (1.1M). The least sale is registered by Ireland with approximately 0.06M sales.

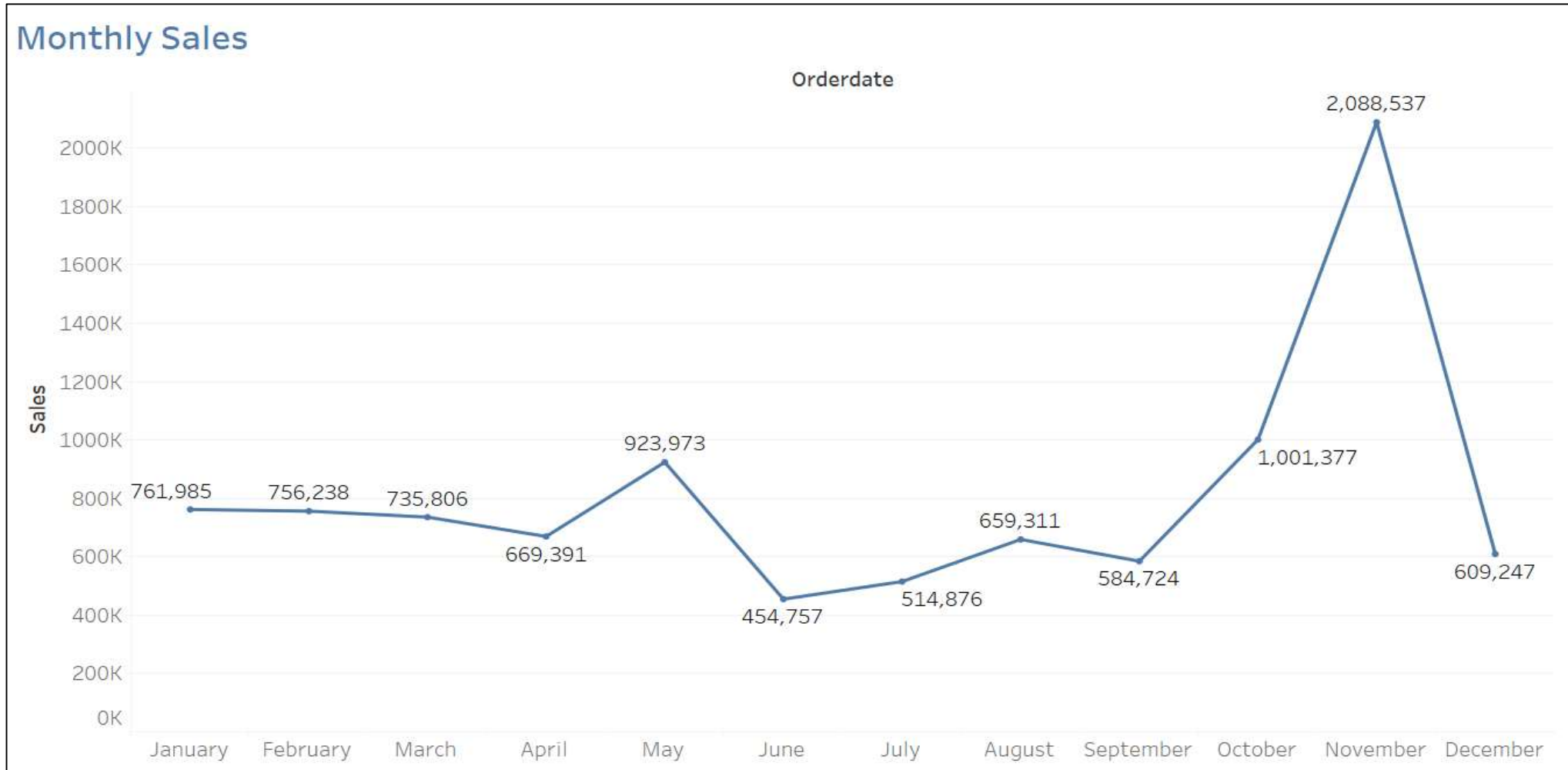
Exploratory Data Analysis

Bivariate Analysis:



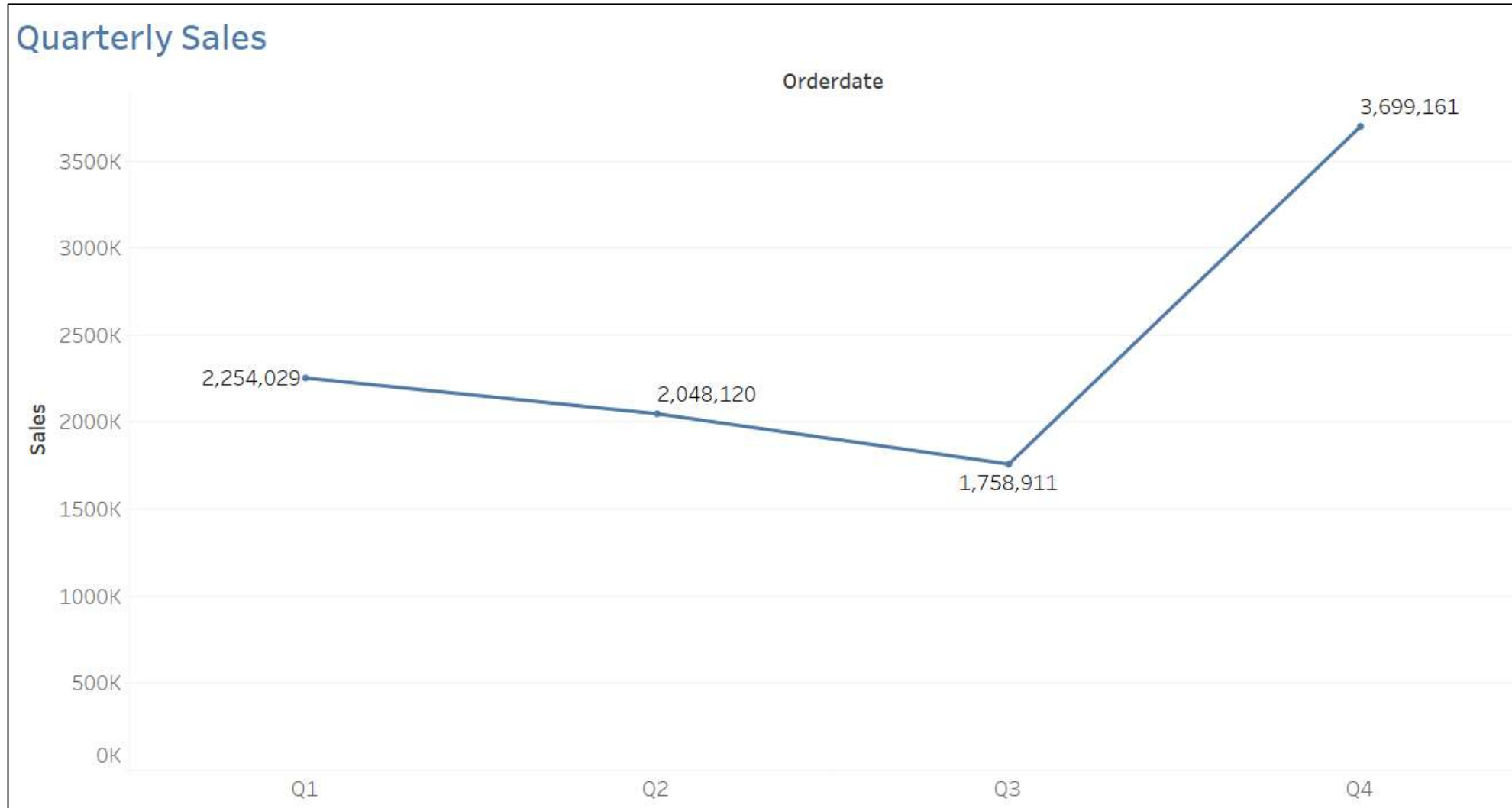
Exploratory Data Analysis

Bivariate Analysis:



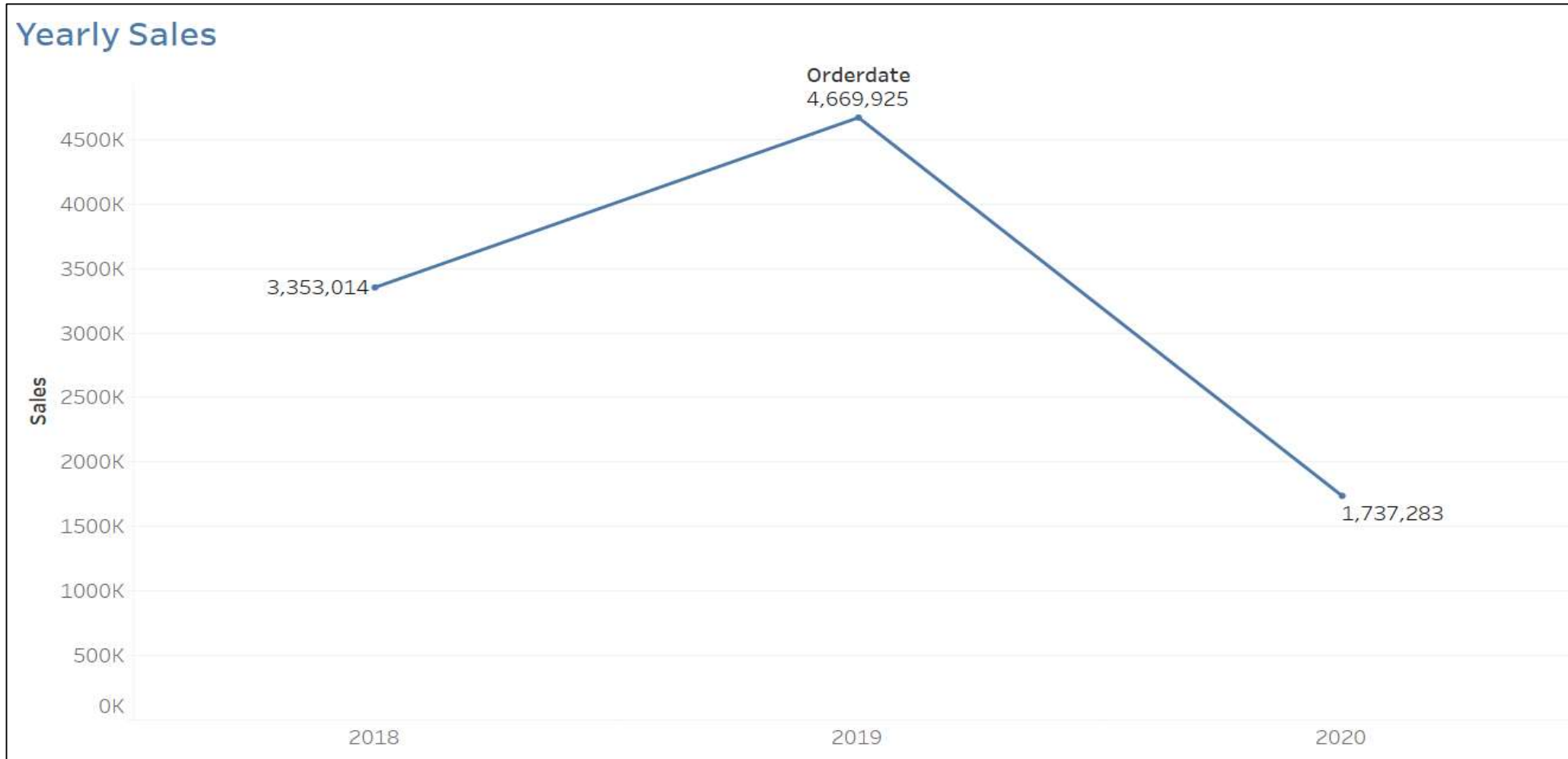
Exploratory Data Analysis

Bivariate Analysis:



Exploratory Data Analysis

Bivariate Analysis:



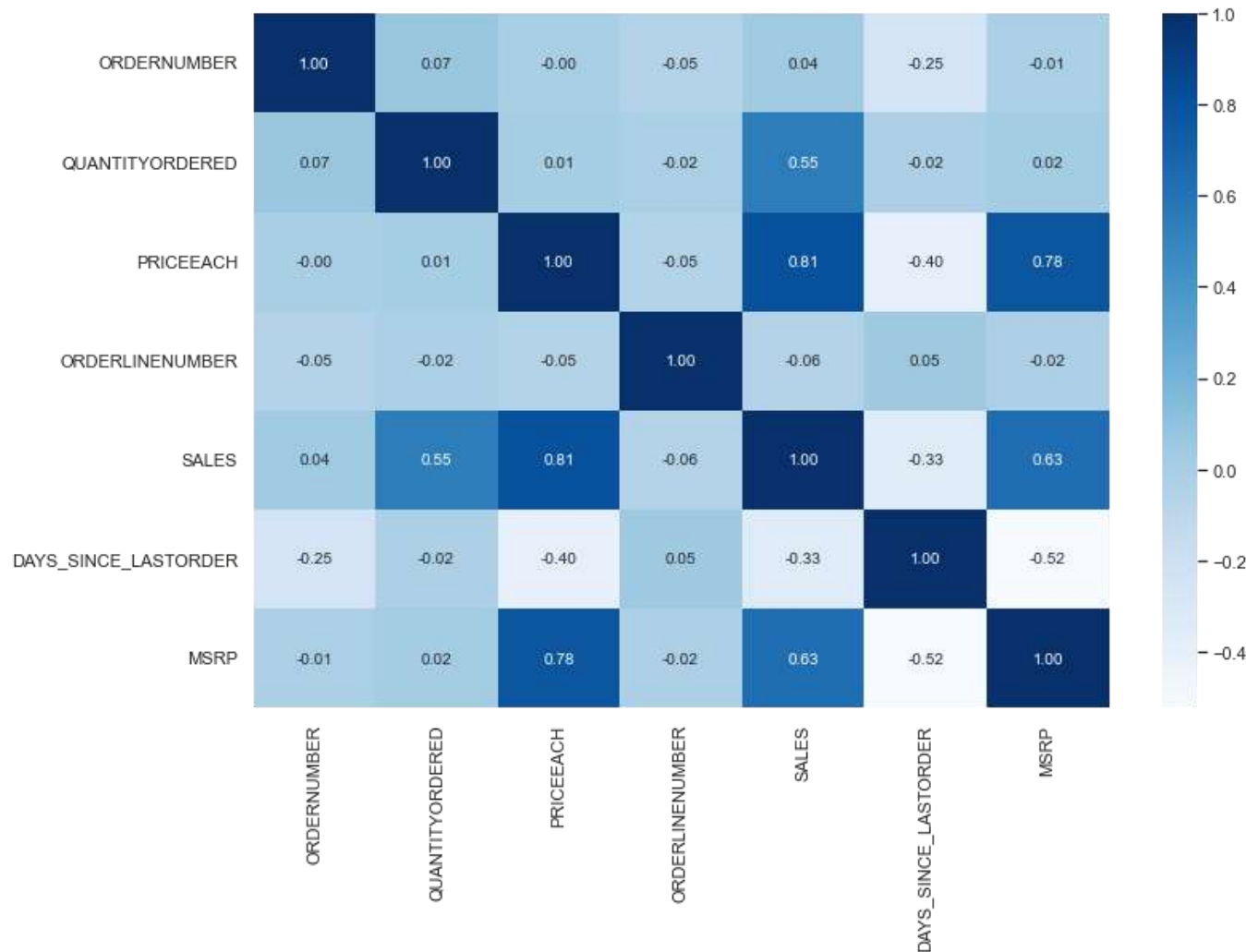
Trends in the Sales Summary:

- Week 45 seems to have the highest sales as compared to the other weeks. The total sale is 0.6M in this week.
- The Sale is increasing after September and falling after December.
- November month gives out the highest sales and the also found that the week 45 falls in November hence, it is also validate the first point(2.1M).
- Sale is decreasing from Q1 to Q3 and gained the strength in Q4.
- 2019 seems to deliver the highest sale(4.67M) and seen a huge drop in sale in 2020(1.7M)

Exploratory Data Analysis

Multivariate Analysis:

Corelation Heatmap

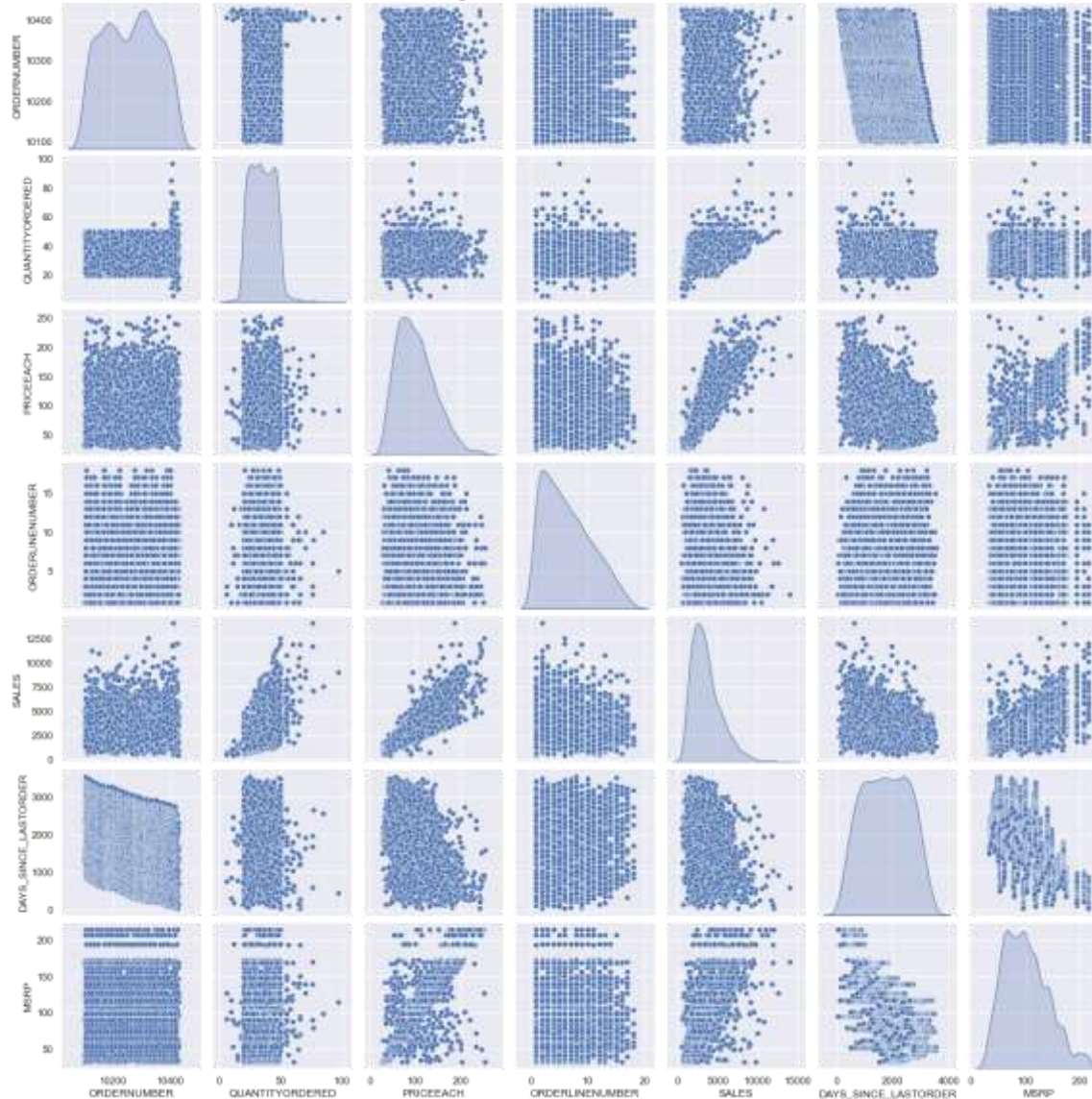


Below are the highly correlated 5 value combinations.

- PRICEEACH and SALES with 0.808
- MSRP and PRICEEACH with 0.778
- MSRP and SALES with 0.635
- QUANTITYORDERED and SALES with 0.553
- MSRP and DAYS_SINCE_LASTORDER with -0.524

Exploratory Data Analysis

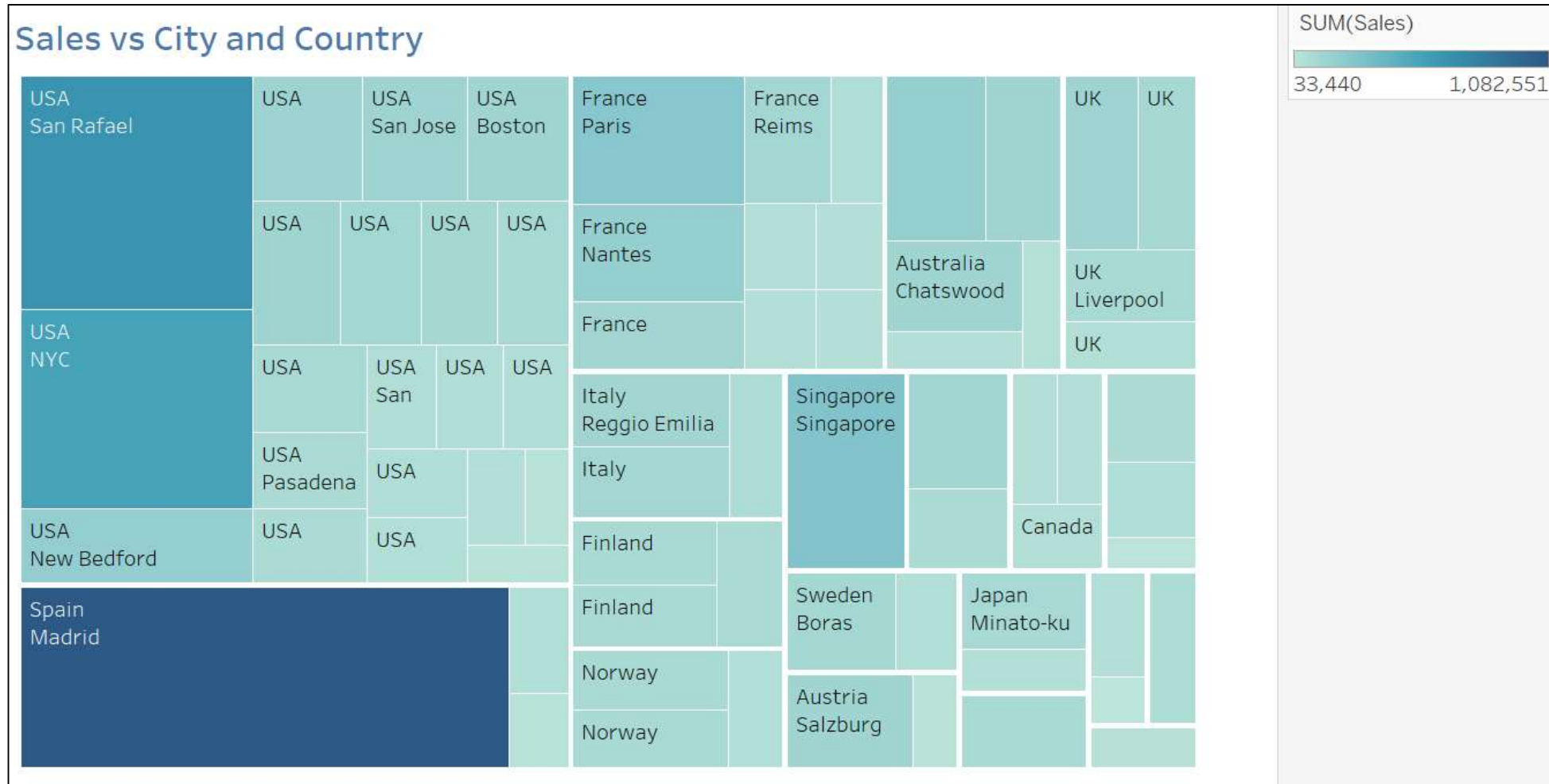
Multivariate Analysis:



- Product Price, Sales and MSRP values are positively high correlated to each other.
- DAYS_SINCE_LAST_ORDER is almost distributed normally. All other data are not normally distributed. Few are right skewed.
- MSRP and DAYS_SINCE_LAST order has negative correlation.

Exploratory Data Analysis

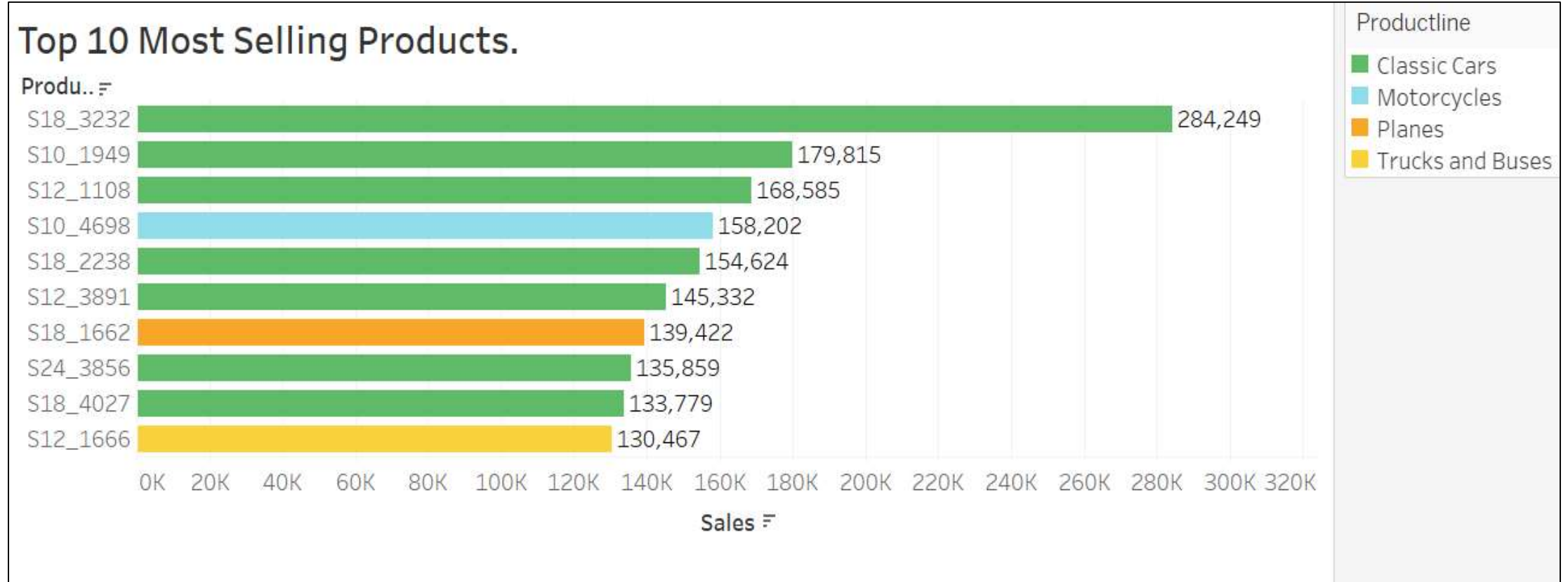
Multivariate Analysis:



- Spain Madrid is the large market for parts compare to other cities with the sale of 1.08M followed by USA San Rafael (0.65M) and USA NYC (0.56M).

Exploratory Data Analysis

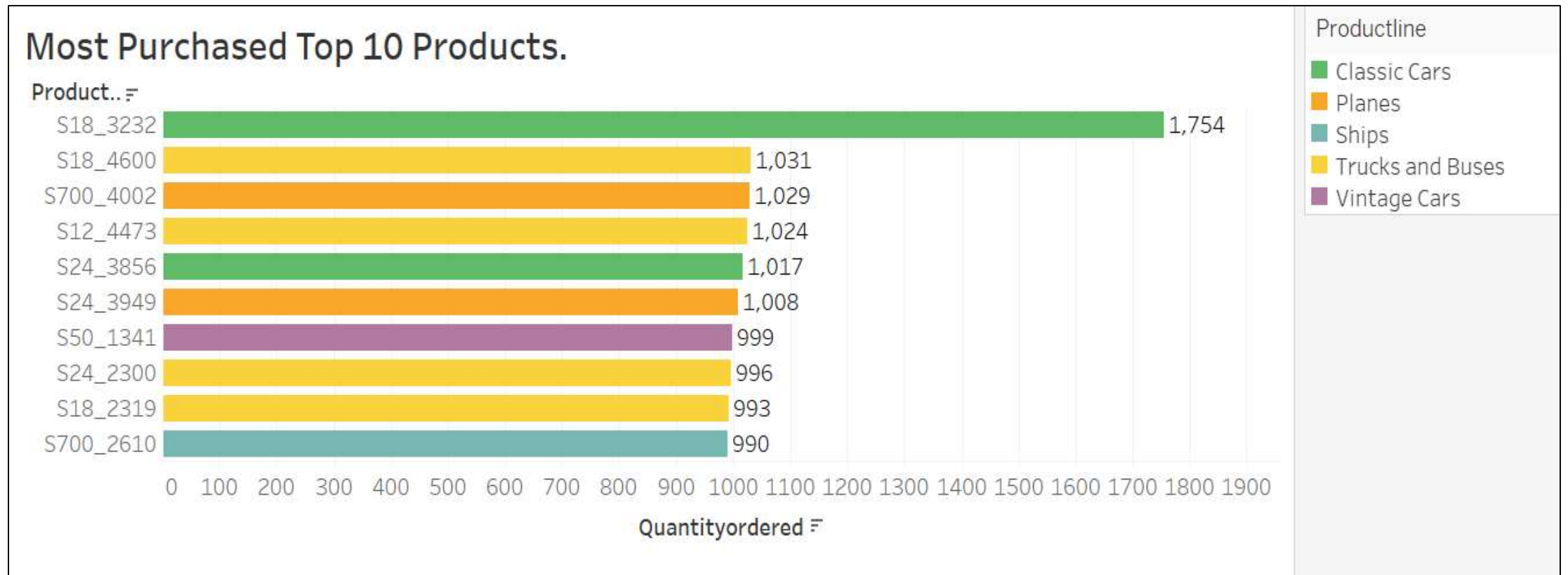
Multivariate Analysis:



- S18_3232 is the most sold item with the Sale of 284,249. followed by S10_1949 and S12_1108.
- We found most of the top selling items are of 'Classic cars' product line categories.

Exploratory Data Analysis

Multivariate Analysis:

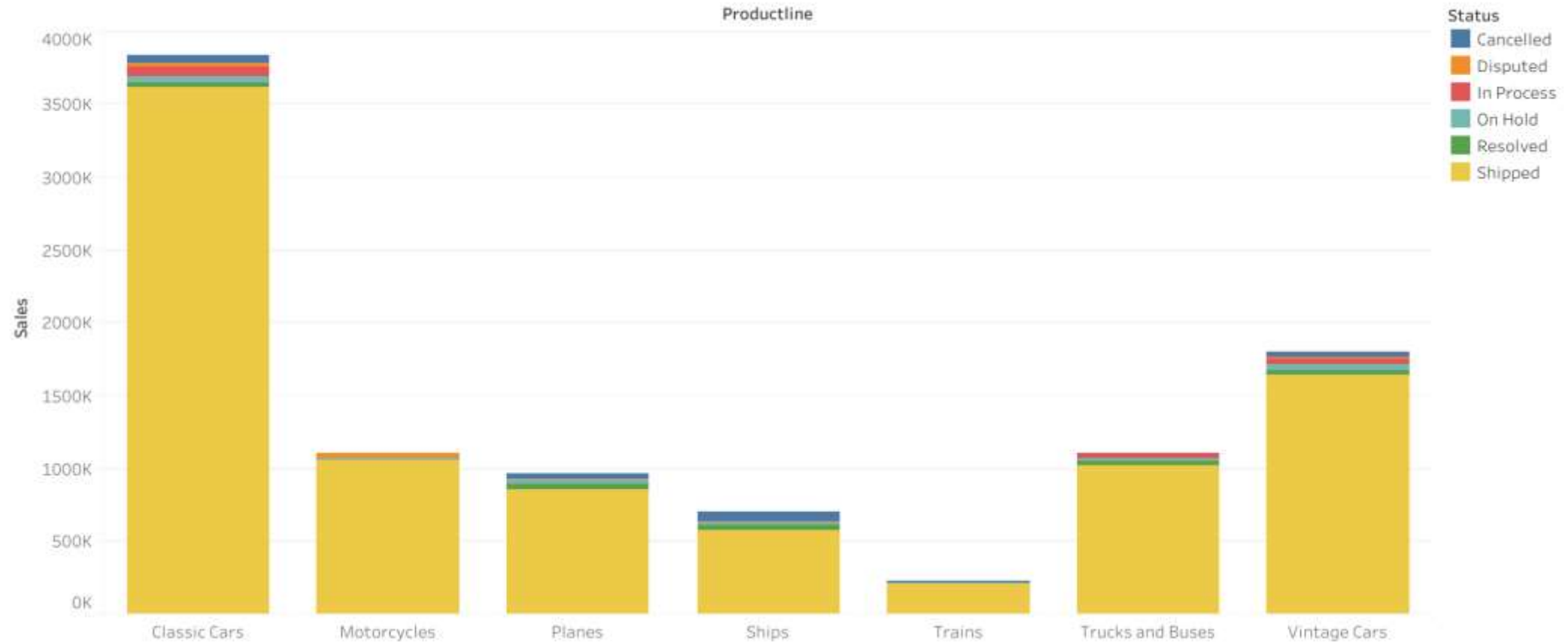


- S18_3232 is the most sold item with the Quantity of 1754. followed by S18_4600 and S700_4002.
- We found no specific product line category is dominating in sales in terms of quantity. Comparatively 'Classic cars', 'Trucks and Buses' and 'Planes' are the categories on demand.

Exploratory Data Analysis

Multivariate Analysis:

Sales, Order status and ProductLine



- In all the product line categories, Almost all the orders are Shipped.
- Also we found the orders which are in 'On progress' or 'On hold' are the recent orders.

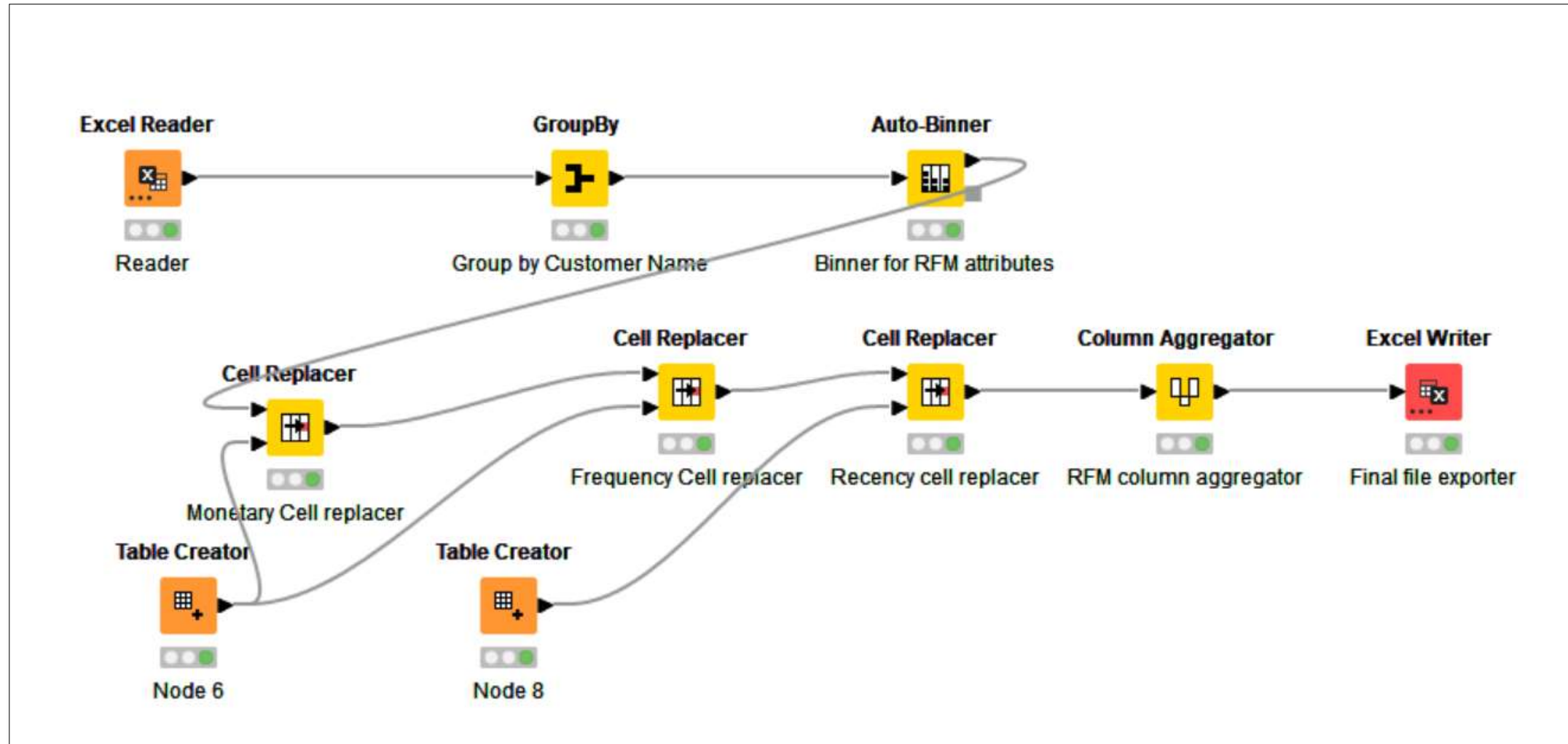
Inferences:

- Classic cars seems to deliver highest sales while Trains have the lowest sales.
- Medium deal size tend to deliver much higher sales as compared to Small & large ones.
- Euro Shopping Channel seems to be the best customer in terms of Sale amount.
- Classic cars are the most purchased and sold products among all.
- USA has the highest sales as compared to other countries.
- City Madrid in country Spain is delivering the highest sales as compared to other cities.
- Product Price, Sales and MSRP values are positively high correlated to each other.
- The sales of large size deal is almost remain stagnant over the years and it can be presumed that company should focus on getting large size chunk projects.
- The company is customer driven because there major chunk of sales comes from 4 - 5 customer. So therefore company should focus more on customer scouting in a rational way because in case there is client churn's it will impact the sales of the company grossly.
- Company is Fulfilling the customers on time. There is old orders which are pending to fulfil.

Customer Segmentation using RFM Analysis.

- KNIME tool is Used for RFM analysis & Customer segmentation.
- Customer name, Quantity ordered, Price each, Order date, order quantity & Sales are the used parameters.
- Predictions:
 - In DAYS SINCE LAST ORDER column we get RECENCY value of the customer. We can use this parameter and using an assumption of minimum of aggregation we created a new column as **Recency** (R).
 - If we can see the data there are same order number repeated for different product Code. So we can assume **count(unique count)** of each order number as **Frequency** of an order number (F).
 - In SALES column we get sales amount for each transaction. We can use SALES parameter and using an assumption of sum of aggregation we created a new column as **Monetary** (M).
 - Then created four different bin for each Recency, Frequency & Monetary using percentile range(0,0.25,0.5,0.75,100).
 - Based on above 4 bin assumption we have considered 4 segments like Low , Medium , High and Excellent.

KNIME Workflow



Output table head For RFM Analysis:

Table "default" - Rows: 89 Spec - Columns: 27 Properties Flow Variables																										
Row ID	S CUSTO...	I ORDER...	I QUANT...	I	I	D SALES	I	I	DAYS_...	I	I	I	I	I	I	I	I	I	S ORDER...	S SALES [...]	S DAYS_...	S SALES_...	S ORDER...	S RECEN...	S Concat...	
Row0	AV Stores, Co.	3	51	51	51	157,807.81	51	421	51	51	...	51	51	51	51	51	51	51	Bin 2	Bin 4	Bin 3	E	M	M	EMM	
Row1	Alpha Cognac	3	20	20	20	70,488.44	20	675	20	20	...	20	20	20	20	20	20	20	Bin 2	Bin 1	Bin 4	L	M	L	LML	
Row2	Amica Model...	2	26	26	26	94,117.26	26	328	26	26	...	26	26	26	26	26	26	26	Bin 1	Bin 3	Bin 2	H	L	H	HLH	
Row3	Anna's Decor...	4	46	46	46	153,996.13	46	131	46	46	...	46	46	46	46	46	46	46	Bin 4	Bin 4	Bin 1	E	E	E	EEE	
Row4	Atelier graph...	3	7	7	7	24,179.96	7	312	7	7	...	7	7	7	7	7	7	7	Bin 2	Bin 1	Bin 2	L	M	H	LMH	
Row5	Australian C...	3	23	23	23	64,591.46	23	1018	23	23	...	23	23	23	23	23	23	23	Bin 2	Bin 1	Bin 4	L	M	L	LML	
Row6	Australian C...	5	55	55	55	200,995.41	55	229	55	55	...	55	55	55	55	55	55	55	Bin 4	Bin 4	Bin 1	E	E	E	EEE	
Row7	Australian Gl...	3	15	15	15	59,469.12	15	190	15	15	...	15	15	15	15	15	15	15	Bin 2	Bin 1	Bin 1	L	M	E	LME	
Row8	Auto Assoc. ...	2	18	18	18	64,834.32	18	275	18	18	...	18	18	18	18	18	18	18	Bin 1	Bin 1	Bin 2	L	L	H	LLH	
Row9	Auto Canal P...	3	27	27	27	93,170.66	27	127	27	27	...	27	27	27	27	27	27	27	Bin 2	Bin 3	Bin 1	H	M	E	HME	
Row10	Auto-Moto Cl...	3	8	8	8	26,479.26	8	1353	8	8	...	8	8	8	8	8	8	8	Bin 2	Bin 1	Bin 4	L	M	L	LML	
Row11	Baane Mini I...	4	32	32	32	116,599.19	32	245	32	32	...	32	32	32	32	32	32	32	Bin 4	Bin 3	Bin 1	H	E	E	HEE	
Row12	Bavarian Coll...	1	14	14	14	34,993.92	14	801	14	14	...	14	14	14	14	14	14	14	Bin 1	Bin 1	Bin 4	L	L	L	LLL	
Row13	Blauer See A...	4	22	22	22	85,171.59	22	705	22	22	...	22	22	22	22	22	22	22	Bin 4	Bin 2	Bin 4	M	E	L	MEL	
Row14	Boards & To...	2	3	3	3	9,129.35	3	410	3	3	...	3	3	3	3	3	3	3	Bin 1	Bin 1	Bin 2	L	L	H	LLH	
Row15	CAF Imports	2	13	13	13	49,642.05	13	625	13	13	...	13	13	13	13	13	13	13	Bin 1	Bin 1	Bin 3	L	L	M	LLM	
Row16	Cambridge C...	2	11	11	11	36,163.62	11	484	11	11	...	11	11	11	11	11	11	11	Bin 1	Bin 1	Bin 3	L	L	M	LLM	
Row17	Canadian Gif...	2	22	22	22	75,238.92	22	364	22	22	...	22	22	22	22	22	22	22	Bin 1	Bin 2	Bin 2	M	L	H	MLH	
Row18	Classic Gift I...	2	21	21	21	67,506.97	21	344	21	21	...	21	21	21	21	21	21	21	Bin 1	Bin 1	Bin 2	L	L	H	LLH	
Row19	Classic Lege...	3	20	20	20	77,795.2	20	309	20	20	...	20	20	20	20	20	20	20	Bin 2	Bin 2	Bin 2	M	M	H	MMH	
Row20	Clover Collec...	2	16	16	16	57,756.43	16	659	16	16	...	16	16	16	16	16	16	16	Bin 1	Bin 1	Bin 4	L	L	L	LLL	
Row21	Collectable M...	2	25	25	25	87,489.23	25	575	25	25	...	25	25	25	25	25	25	25	Bin 1	Bin 2	Bin 3	M	L	M	MLM	
Row22	Collectables ...	3	24	24	24	81,577.98	24	179	24	24	...	24	24	24	24	24	24	24	Bin 2	Bin 2	Bin 1	M	M	E	MME	
Row23	Corrida Auto...	3	32	32	32	120,615.28	32	407	32	32	...	32	32	32	32	32	32	32	Bin 2	Bin 4	Bin 2	E	M	H	EMH	
Row24	Cruz & Sons ...	3	26	26	26	94,015.73	26	971	26	26	...	26	26	26	26	26	26	26	Bin 2	Bin 3	Bin 4	H	M	L	HML	
Row25	Daedalus De...	2	20	20	20	69,052.41	20	573	20	20	...	20	20	20	20	20	20	20	Bin 1	Bin 1	Bin 3	L	L	M	LLM	
Row26	Danish Whol...	5	36	36	36	145,041.6	36	499	36	36	...	36	36	36	36	36	36	36	Bin 4	Bin 4	Bin 3	E	E	M	EEM	
Row27	Diecast Class...	4	31	31	31	122,138.14	31	228	31	31	...	31	31	31	31	31	31	31	Bin 4	Bin 4	Bin 1	E	E	E	EEE	
Row28	Diecast Colle...	2	18	18	18	70,859.78	18	672	18	18	...	18	18	18	18	18	18	18	Bin 1	Bin 2	Bin 4	M	L	L	MLL	
Row29	Double Deck...	2	12	12	12	36,019.04	12	670	12	12	...	12	12	12	12	12	12	12	Bin 1	Bin 1	Bin 4	L	L	L	LLL	
Row30	Dragon Souv...	5	43	43	43	172,989.68	43	649	43	43	...	43	43	43	43	43	43	43	Bin 4	Bin 4	Bin 4	E	E	L	EEL	
Row31	Enaco Distrib...	3	23	23	23	78,411.86	23	659	23	23	...	23	23	23	23	23	23	23	Bin 2	Bin 2	Bin 4	M	M	L	MML	

Inferences from RFM Analysis and identified segments:

1. Best Customers:

On basis on Recency, frequency & monetary we have grouped our top customers. We have given the most significance to recency parameter as these customers has recently purchased our products. Also according to RFM model the most importance is given to recency. Hence we have kept it as our first parameter for selecting top customers.

Below are the sample of Best customers

1	CUSTOMERNAME	Concatenat	ORDERNUMBER	QUANTITYORDERED	PRICEEACH	ORDERLINENUMBER	SALES	ORDERDATE	DAYS_SINCE_LASTORDER	STATUS	PRODUCT
5	Anna's Decorations, Ltd	EEE	4	46	46	46	153996.13	46	131	46	
8	Australian Collectors, Co.	EEE	5	55	55	55	200995.41	55	229	55	
29	Diecast Classics Inc.	EEE	4	31	31	31	122138.14	31	228	31	
34	Euro Shopping Channel	EEE	26	259	259	259	912294.11	259	42	259	
45	La Rochelle Gifts	EEE	4	53	53	53	180124.9	53	139	53	
46	Land of Toys Inc.	EEE	4	49	49	49	164069.44	49	216	49	
55	Mini Gifts Distributors Ltd.	EEE	17	180	180	180	654858.06	180	219	180	
69	Salzburg Collectables	EEE	4	40	40	40	149798.63	40	188	40	
74	Souvenirs And Things Co.	EEE	4	46	46	46	151570.98	46	186	46	
78	Technics Stores Inc.	EEE	4	34	34	34	120783.07	34	241	34	
80	The Sharp Gifts Warehouse	EEE	4	40	40	40	160010.27	40	182	40	

Inferences from RFM Analysis and identified segments:

2: Customers on verge of churning:

On basis on Recency, frequency & monetary we have grouped our Customers who are on verge of churning. We should definitely focus on this group before we lose them and try to convert them into our regular customers.

Below are the sample of such customers:

1	CUSTOMERNAME	Concatenat	ORDERNUMBER	QUANTITYORDERED	PRICEEACH	ORDERLINENUMBER	SALES	ORDERDATE	DAYS_SINCE_LASTORDER	STATUS	PRODI
3	Alpha Cognac	LML	3	20	20	20	70488.44	20	675	20	
6	Atelier graphique	LMH	3	7	7	7	24179.96	7	312	7	
7	Australian Collectables, Ltd	LML	3	23	23	23	64591.46	23	1018	23	
10	Auto Assoc. & Cie.	LLH	2	18	18	18	64834.32	18	275	18	
12	Auto-Moto Classics Inc.	LML	3	8	8	8	26479.26	8	1353	8	
14	Bavarian Collectables Imports, Co.	LLL	1	14	14	14	34993.92	14	801	14	
16	Boards & Toys Co.	LLH	2	3	3	3	9129.35	3	410	3	
17	CAF Imports	LLM	2	13	13	13	49642.05	13	625	13	
18	Cambridge Collectables Co.	LLM	2	11	11	11	36163.62	11	484	11	
19	Canadian Gift Exchange Network	MLH	2	22	22	22	75238.92	22	364	22	
20	Classic Gift Ideas, Inc	LLH	2	21	21	21	67506.97	21	344	21	
21	Classic Legends Inc.	MMH	3	20	20	20	77795.2	20	309	20	
22	Clover Collections, Co.	LLL	2	16	16	16	57756.43	16	659	16	
23	Collectable Mini Designs Co.	MLM	2	25	25	25	87489.23	25	575	25	
24	Collectables For Less Inc.	MME	3	24	24	24	81577.98	24	179	24	
26	Cruz & Sons Co.	HML	3	26	26	26	94015.73	26	971	26	

Inferences from RFM Analysis and identified segments:

3: Lost Customers:

On basis on Recency, frequency & monetary parameters we have grouped our Customers who we'd lost. Their recency is very low and hasn't made any purchase since long. So we can say these are our lost customers. If taken feedback from them and fulfil their demand we might bring them back to been a good customer.

Below are the sample of such customers:

1	CUSTOMERNAME	Concatenat	ORDERNUMBER	QUANTITYORDERED	PRICEEACH	ORDERLINENUMBER	SALES	ORDERDATE	DAYS_SINCE_LASTORDER	STATUS	PRODUCT
14	Bavarian Collectables Imports, Co.	LLL	1	14	14	14	34993.92	14	801	14	
22	Clover Collections, Co.	LLL	2	16	16	16	57756.43	16	659	16	
31	Double Decker Gift Stores, Ltd	LLL	2	12	12	12	36019.04	12	670	12	
42	Iberia Gift Imports, Corp.	LLL	2	15	15	15	54723.62	15	904	15	
72	Signal Collectibles Ltd.	LLL	2	15	15	15	50218.51	15	836	15	

Inferences from RFM Analysis and identified segments:

4: Loyal Customers:

On basis on Recency, frequency & monetary we have grouped our loyal customers. These customers have purchased multiple times with good monetary value. If we focus more on this segment of customers, we can easily turn them into our top best customers too. Also, in this segment we can see the customers for product line - classic cars are many.

Below are the sample of such customers

1	CUSTOMERNAME	Concatenat	ORDERNUMBER	QUANTITYORDERED	PRICEEACH	ORDERLINENUMBER	SALES	ORDERDATE	DAYS_SINCE_LASTORDER	STATUS	PRODL
13	Baane Mini Imports	HEE	4	32	32	32	116599.19	32	245	32	
25	Corrida Auto Replicas, Ltd	EMH	3	32	32	32	120615.28	32	407	32	
28	Danish Wholesale Imports	EEM	5	36	36	36	145041.6	36	499	36	
32	Dragon Souvenirs, Ltd.	EEL	5	43	43	43	172989.68	43	649	43	
57	Muscle Machine Inc	EEM	4	48	48	48	197736.94	48	502	48	
59	Online Diecast Creations Co.	EME	3	34	34	34	131685.3	34	253	34	
65	Reims Collectables	EEH	5	41	41	41	135042.94	41	287	41	
66	Rovelli Gifts	EML	3	48	48	48	137955.72	48	1032	48	
71	Scandinavian Gift Ideas	EMH	3	38	38	38	134259.33	38	262	38	
81	Tokyo Collectables, Ltd	HEH	4	32	32	32	120562.74	32	259	32	
88	Volvo Model Replicas, Co	MEH	4	19	19	19	75754.88	19	338	19	

Recommendations:

- Using Recency, frequency & monetary parameters we have grouped our top , loyal, on the verge of churning and lost customers. Customers with good recency has been our top customers were as we also have lost customer lists.
- Customers on verge of churning can be saved and can be converted into a good buyer.
- RFM model is used for deriving the customers types like Loyal, top or best, on verge of churning & lost customers.
- Recency, frequency & monetary parameters were widely used to bifurcate the types of customers.
- This model can be very helpful to the company to maintain its sales and customers and can focus on how the company has lost the customers & can take various actions to bring back them.
- It is vital for the company to convert the customers who are on verge of churning into a regular customer or at least maintain them.
- And also how to increase the sales ratio can be identified.