Marketing and Retail Analysis

MRA PROJECT - MILESTONE 1

07/12/2021

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

- HARSH ALKESH PANDYA
- PGP DSBA FEB_A 2021

Data Dictionary

Here we observe a detailed description of all cells available in an excel file attached to a problem statement.

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 Quantity and Item Price
MSRP :	Manufacturer's Suggested Retail Price		
PRODUCTCODE :	Code of Product		

Explore the Dataset

	ORDERNUMBER	QUANTITYORDERED	PRICEEACH	ORDERLINENUMBER	SALES	ORDERDATE	DAYS_SINCE_LASTORDER	STATUS	PRODUCTLINE	MSRP	PRODUCTCODE	CUSTOMERNAME	PHONE	ADDRESSLINE1	CITY	POSTALCODE	COUNTRY	CONTACTLASTNAME
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
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
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
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
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

- Here is the list available in the data set, here I extract here top 5 rows that are available in a Data set. To read the file in python I used the “Read” command this function is available in the “pandas” library.

Exploratory Data Analysis And Inferences

Describe Data

	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

- The dataset is measured using a central measure for all the columns with integer values.
- It tells how the data is been distributed, deviated, or centrally aligned

Data information and Data type

- Here we observe that there one data type is based on date time which is showing us a date and time of order.
- Here we can observe that most of the columns are object type, and the rest are the int and float type
- The datatype of ORDERDATE is datetime64[ns] format.

```
<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
```

```
ORDERNUMBER                      int64
QUANTITYORDERED                  int64
PRICEEACH                      float64
ORDERLINENUMBER                  int64
SALES                          float64
ORDERDATE               datetime64[ns]
DAYS_SINCE_LASTORDER             int64
STATUS                          object
PRODUCTLINE                     object
MSRP                             int64
PRODUCTCODE                     object
CUSTOMERNAME                    object
PHONE                           object
ADDRESSLINE1                    object
CITY                            object
POSTALCODE                      object
COUNTRY                         object
CONTACTLASTNAME                 object
CONTACTFIRSTNAME                object
DEALSIZE                        object
dtype: object
```

Size of Data frame

```
(2747, 20)
```

- Df.shape gives us the output in the total number of rows and columns are available in the data frame.
- In the above image you can observe that the total numbers of rows and columns are available in the data frame
- In the original data set have 2747 rows and 20 columns.

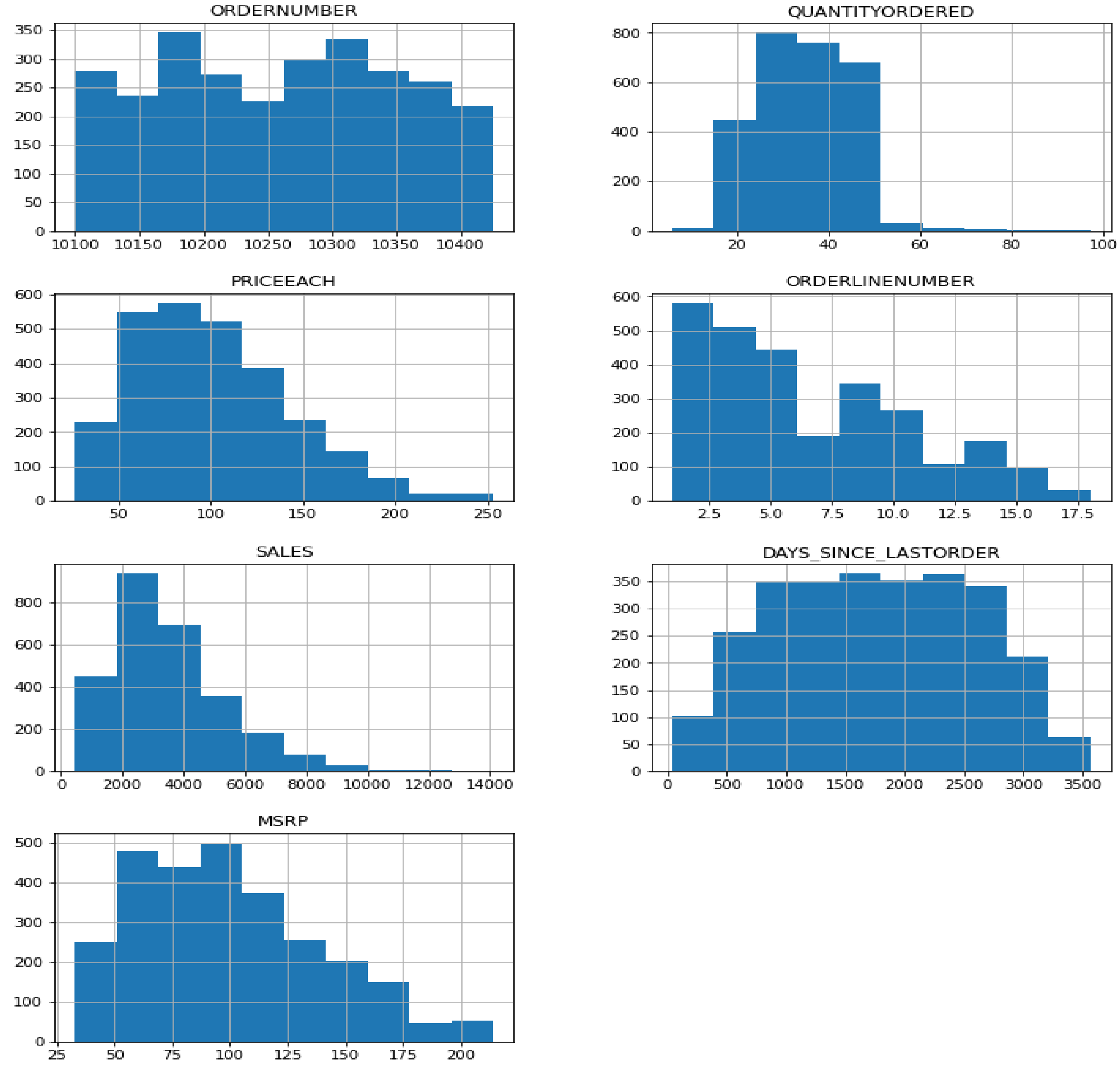
Check Missing values

- *Here we see that the dataset doesn't have any null values.*
- *If we can observe any presence of missing values or duplicate values, we would have treated it before performing any calculations.*

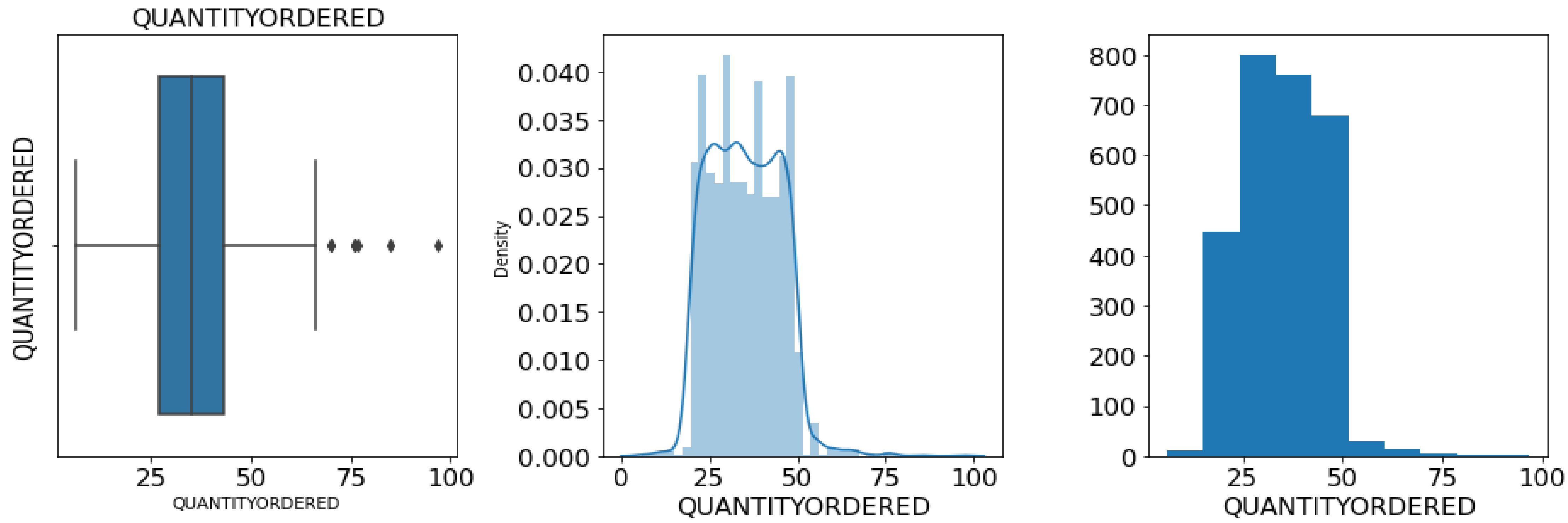
```
ORDERNUMBER              0
QUANTITYORDERED          0
PRICEEACH                0
ORDERLINENUMBER          0
SALES                    0
ORDERDATE                0
DAYS_SINCE_LASTORDER     0
STATUS                   0
PRODUCTLINE              0
MSRP                     0
PRODUCTCODE              0
CUSTOMERNAME             0
PHONE                    0
ADDRESSLINE1             0
CITY                     0
POSTALCODE               0
COUNTRY                  0
CONTACTLASTNAME          0
CONTACTFIRSTNAME         0
DEALSIZE                 0
dtype: int64
```

Univariate Analysis

- By using this analysis we can observe the outliers and skewness of the data.
- Here we have three different kind of graphs who gives us a different kind of information's.
- By using box plot we can observe that there is so many outliers are there.
- And by using histogram plot we can observe that this plot is normally distributed.

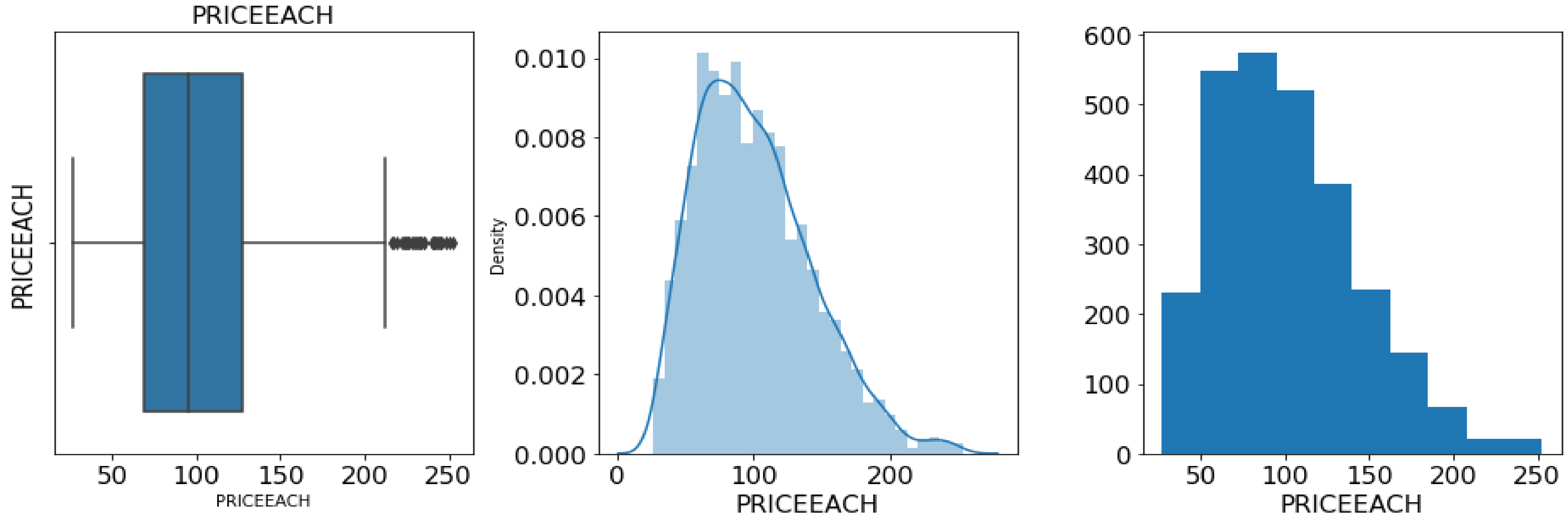


Order Quantity



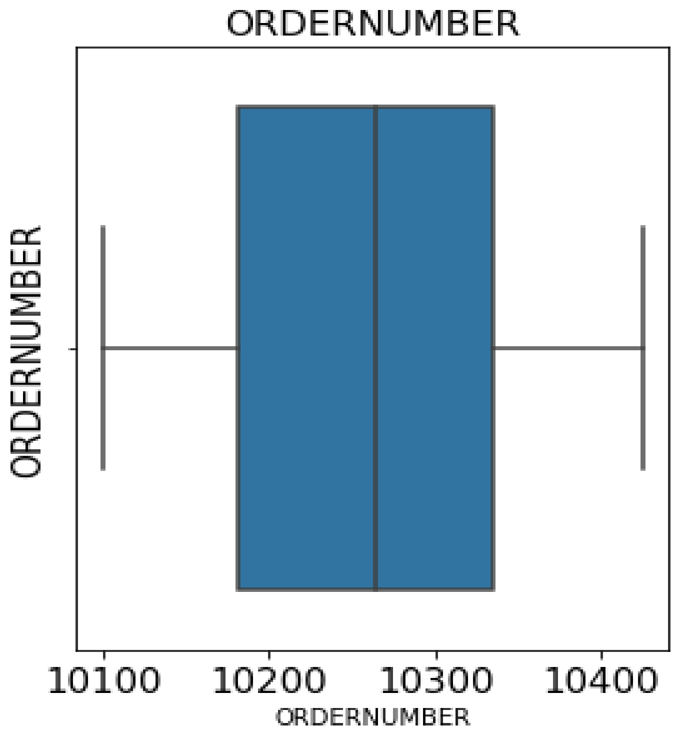
- *The box plot of the min order quantity amount variable shows few outliers.*
- *Min each price amount is normal skewed - 0.04025*
- *The dist. plot shows the distribution of data from 10 to 100.*

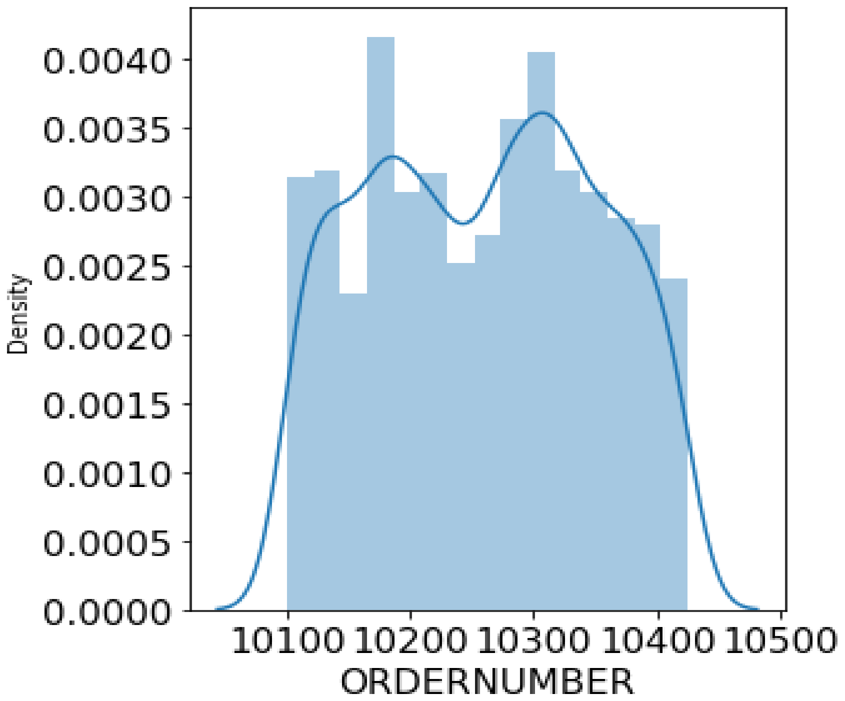
Each item price

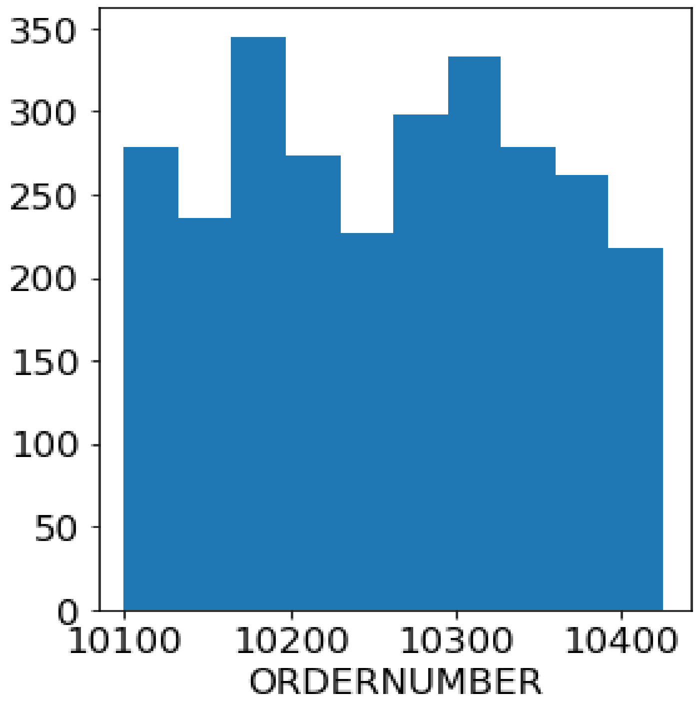


- *The box plot of the min Each product price amount variable shows few outliers.*
- *Min each price amount is negatively skewed - 0.01016*
- *The dist. plot shows the distribution of data from 45 to 250*

Order numbers

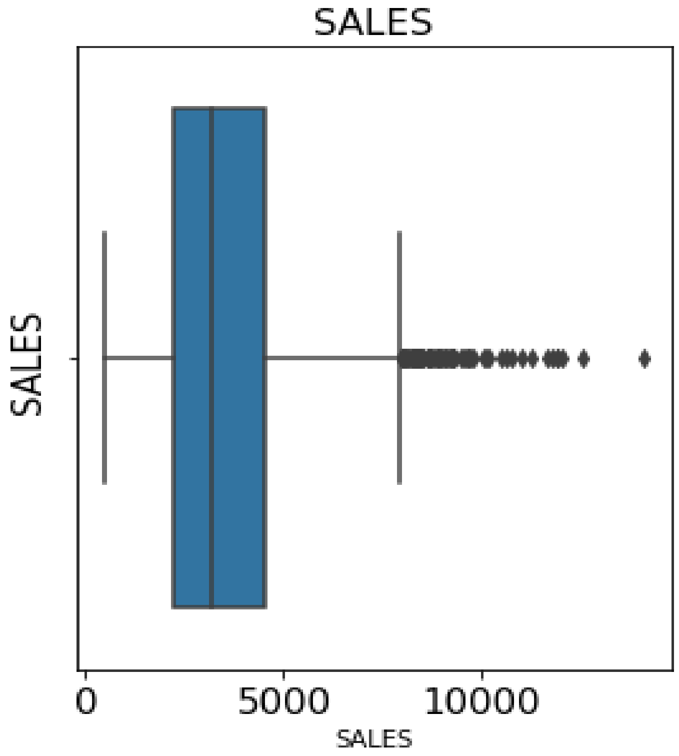


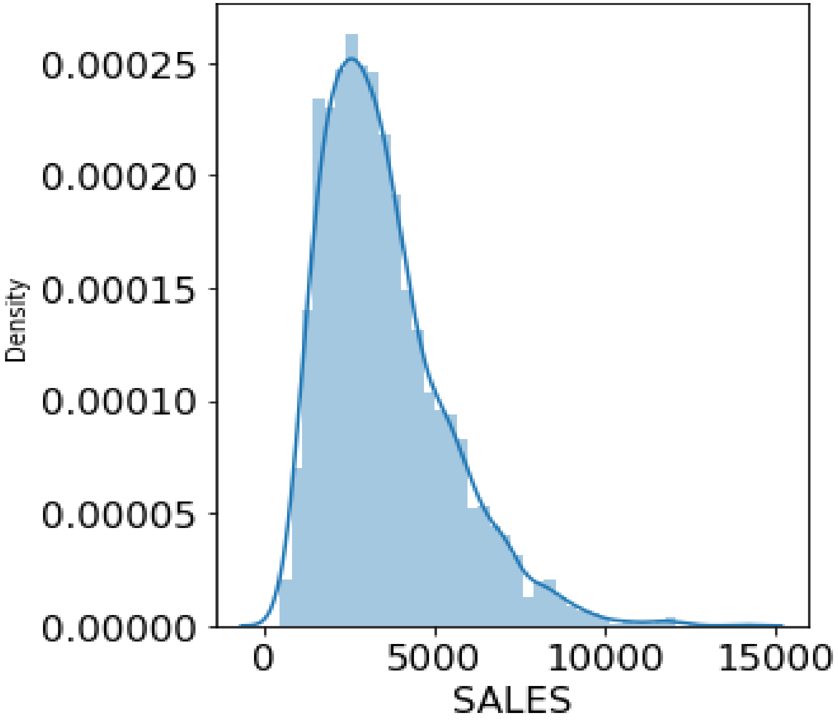


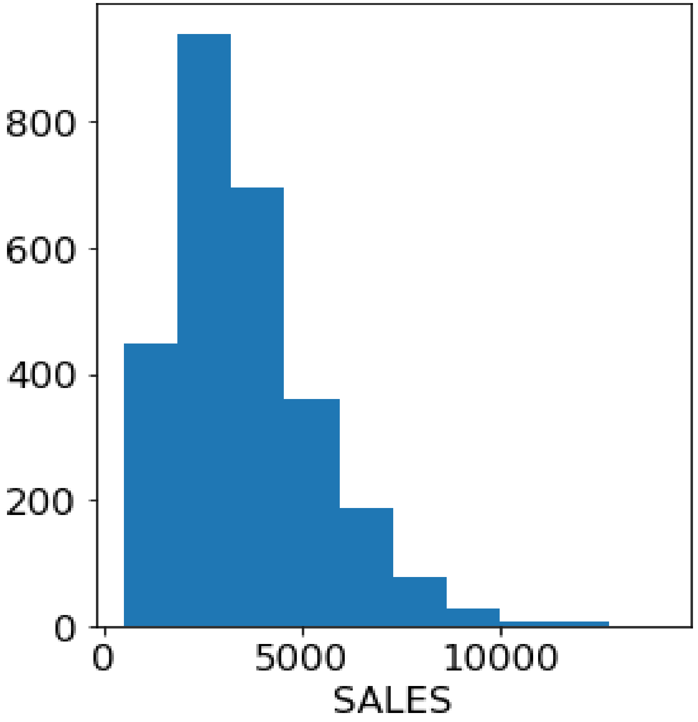


- *The box plot of the min order quantity amount variable shows few outliers.*
- *Min each price amount is normal skewed - 0.0040*
- *The dist. plot shows the distribution of data from 10110 to 10400.*

Sales

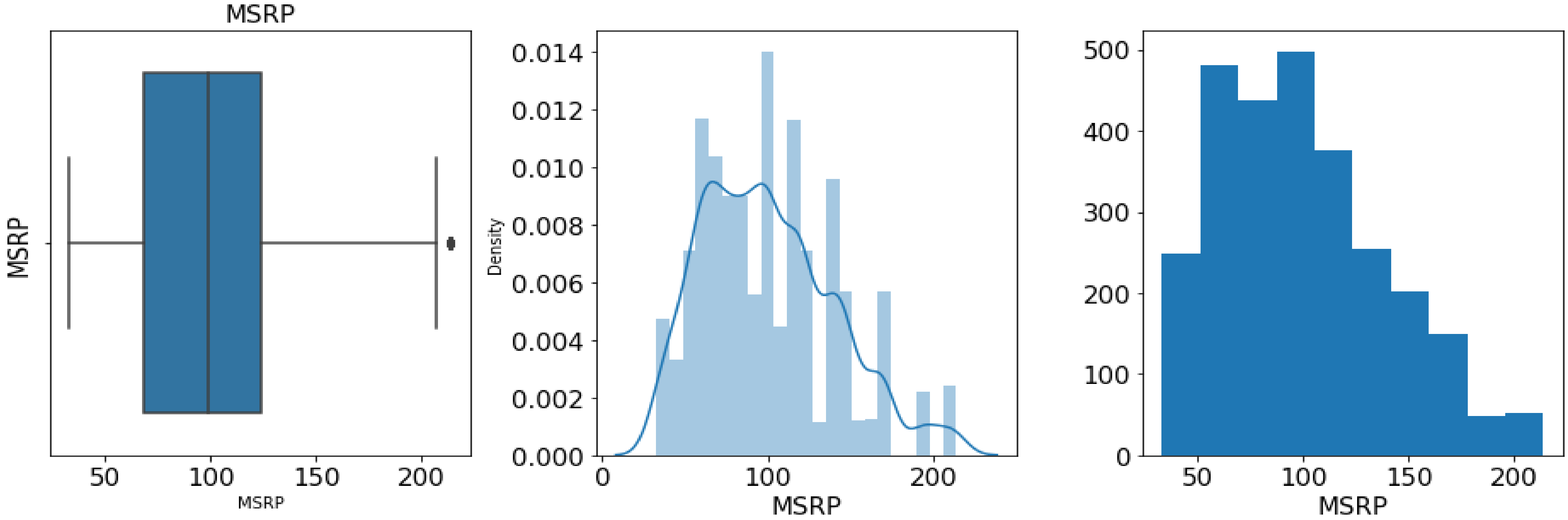






- *The box plot of the min Sales amount variable shows few outliers.*
- *Min each price amount is negatively skewed - 0.000254*
- *The dist. plot shows the distribution of data from 5 to 10564.*

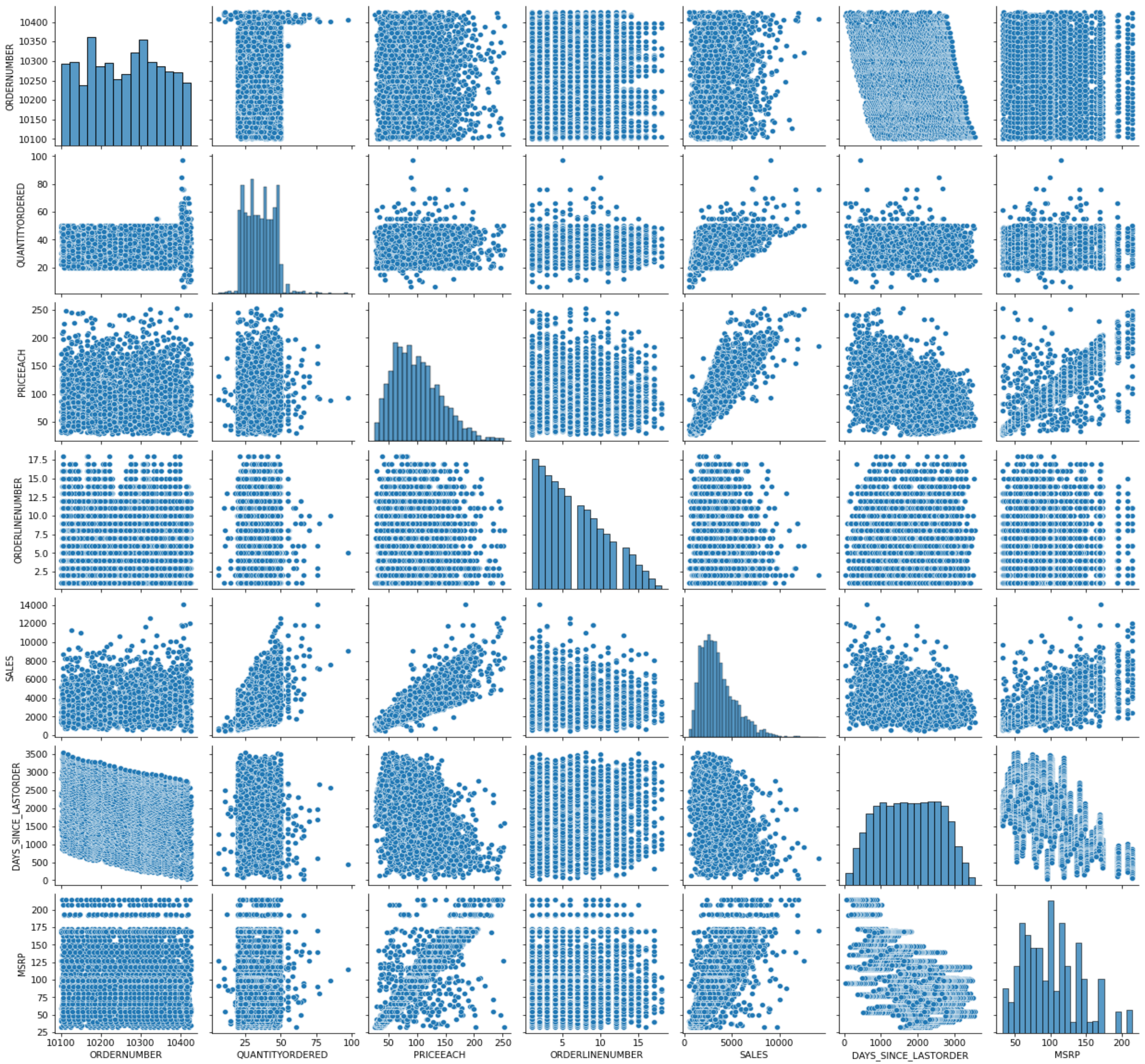
MSRP



- *The box plot of the min MSRP variable shows one outliers.*
- *Min each price amount is normal skewed - 0.01402*
- *The dist. plot shows the distribution of data from 24 to 250.*

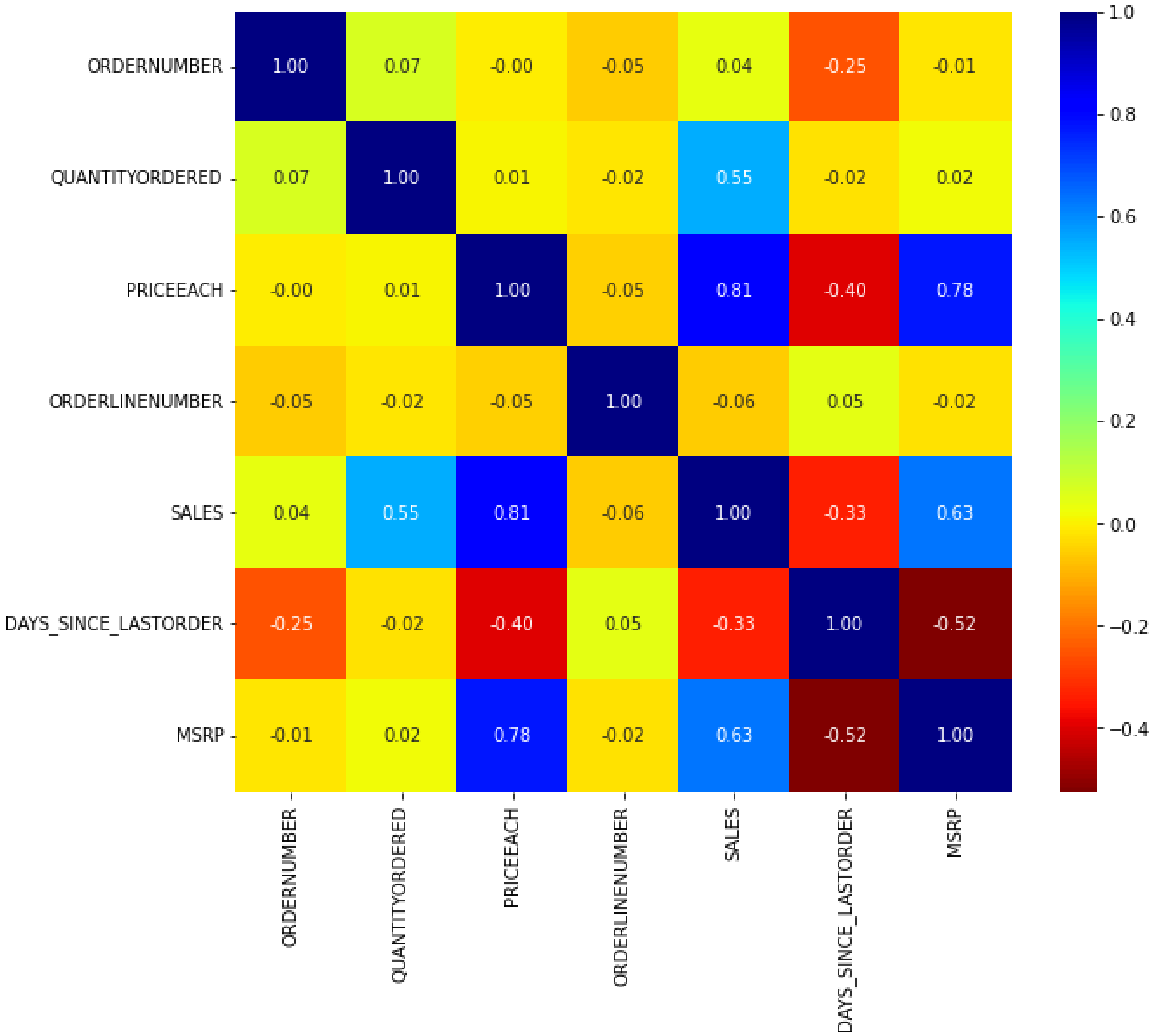
Bivariate Analysis

- *In histogram we can observe there are some relation between each other but we need a clear picture of this so in next stap we can observe the correlation table and heat map.*



Multi-variate Analysis

- The correlation across the variables can be found using corr() function in a matrix form.
- To indicate and visualize the clusters within the data using the heat map (from sns package).
- There is good correlation (0.78) between MSRP and PRICEEACH. The customers paid advance amount nearly equal to their spending amount.
- Next good correlation (0.63) is between SALES and MSRP. If the customers had no spending for that month, the advance amount is stored as current balance in his account.
- The next good correlation (0.55) is between QUANTITYORDER and SALES. The customer plans for the spending based on the current balance amount available.



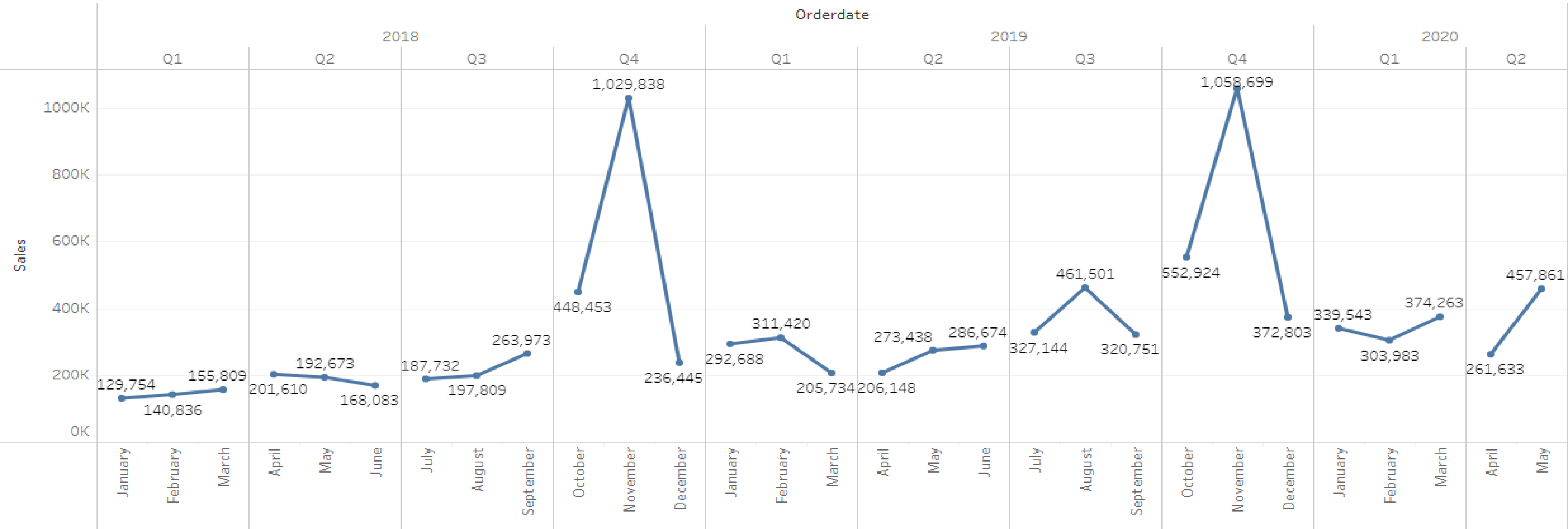
Perfect Correlation Table

- *By using this table we got a clear picture of what is the maximum correlation with witch segment.*
- *As we observe that priceeach and sales have highest correlation (0.8082)*

		correlation
PRICEEACH	SALES	0.808287
MSRP	PRICEEACH	0.778393
	SALES	0.634849
QUANTITYORDERED	SALES	0.553359
MSRP	DAYS_SINCE_LASTORDER	0.524285
DAYS_SINCE_LASTORDER	PRICEEACH	0.397092
SALES	DAYS_SINCE_LASTORDER	0.334274

Data Visualizations

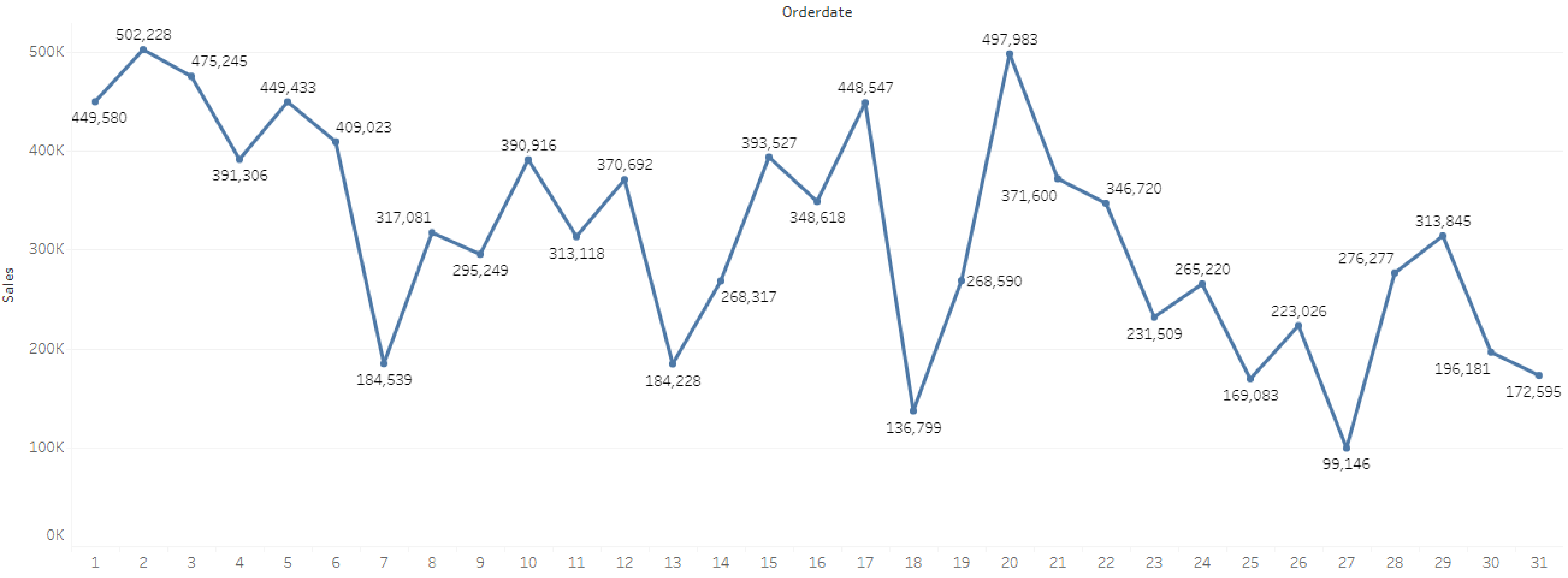
Trends in Sales



- *Sales is highest in the 4th Quarter for the year 2018 and for the year 2019 when compared to all other quarters.*

- *For the year 2020 sales is highest in the 2nd Quarter.*

- *For the rest of the quarters, sales is low and is on an increasing and decreasing trend.*

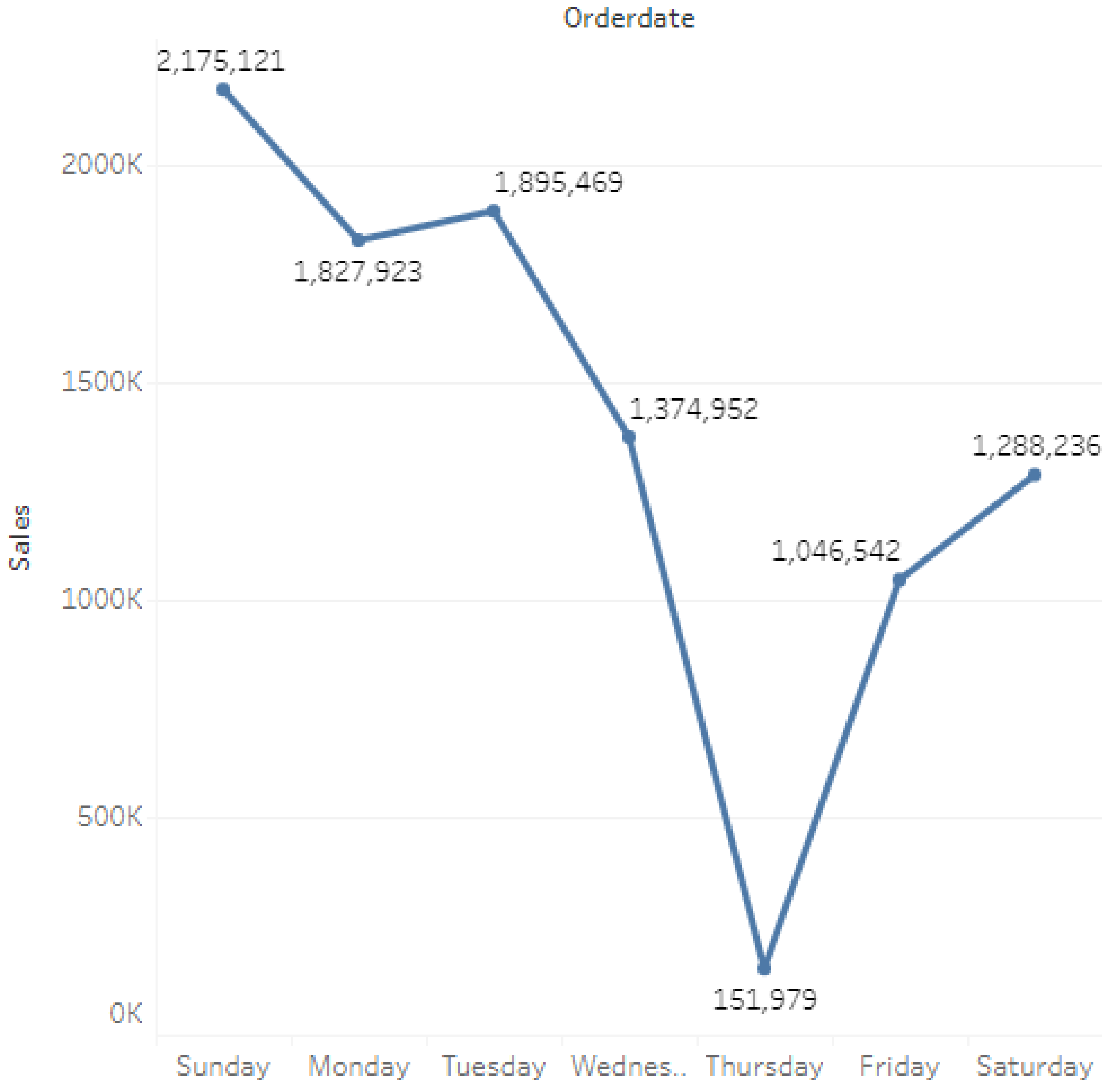


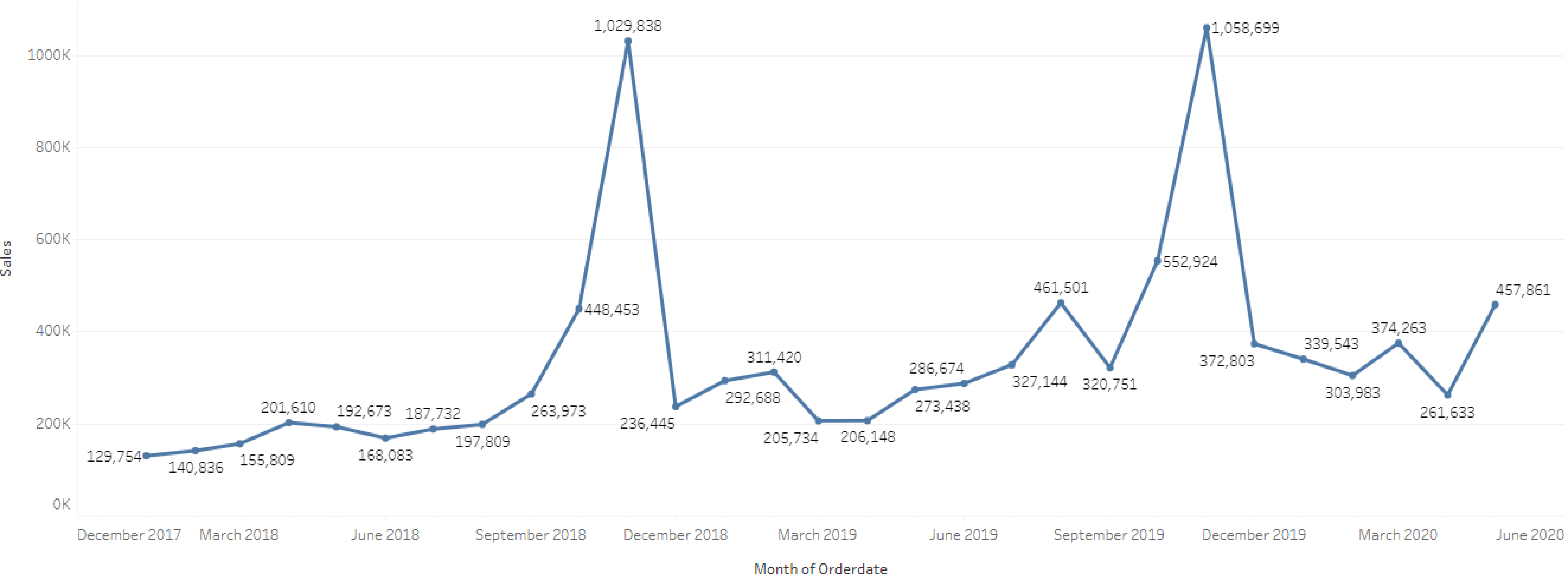
Treads in Sales Daily Basis

- *We see that on day basis, 20th Day in month has the highest sales followed by Tuesday.*
- *But 27th has the least sales compared to other week days.*

Treads in Sales Weekly Basis

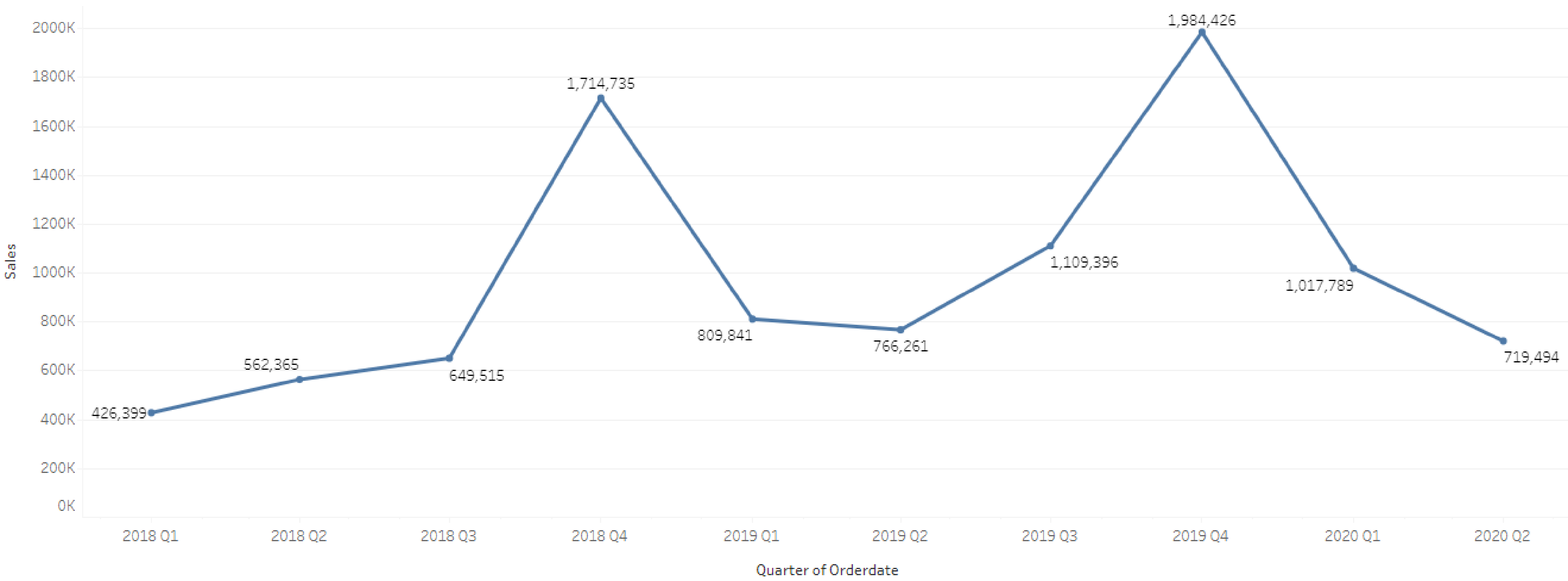
- We see that on Weekly basis, Sunday has the highest sales followed by the week.
- But Thursday has the least sales compared to other week days.





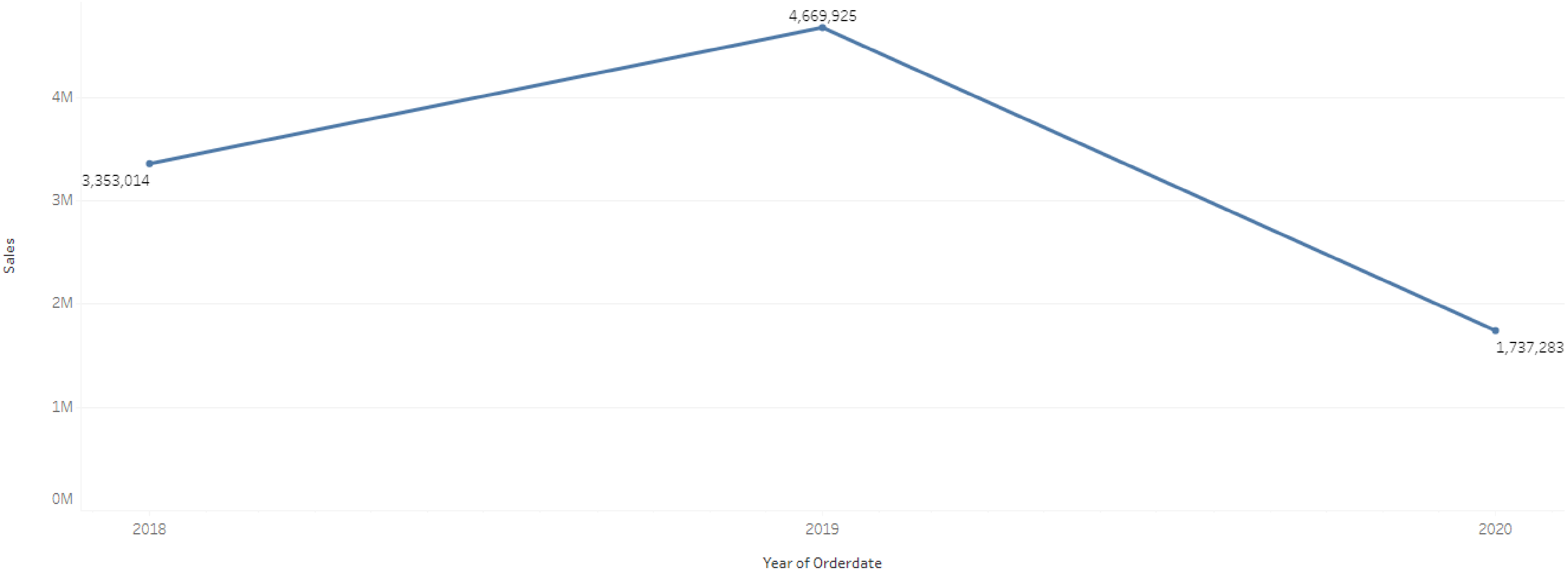
Treads in Sales Monthly Basis

- *We observe that on monthly basis, November 2018 and 2019 month has the highest sales followed by January 2017.*
- *But January 2017-2018 has the least sales compared to other months.*



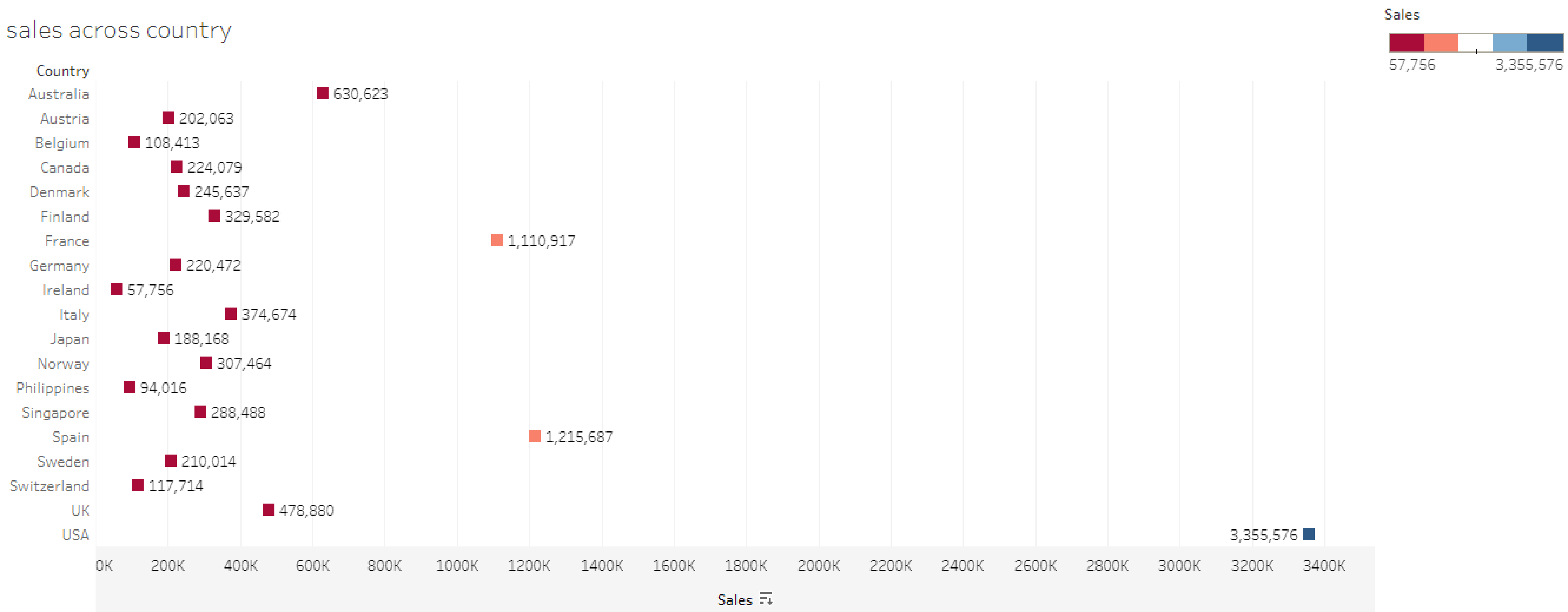
Treads in Sales Quarterly Basis

- *We observe that on Quarterly basis, 2019 Q4 has the highest sales followed by 2018 Q4.*
- *But 2018 Q1 has the least sales compared to other months.*



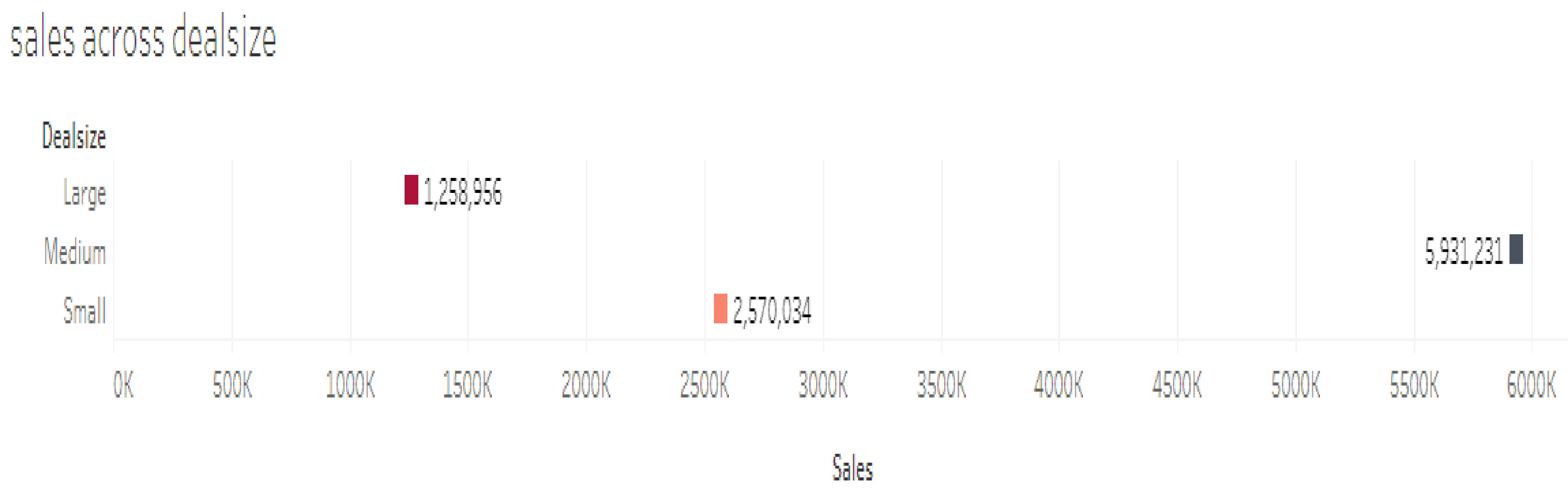
Treads in Sales Yearly Basis

- *We observe that on Yearly basis, 2019 has the highest sales followed by 2018.*
- *But Year 2020 has the least sales compared to other months.*



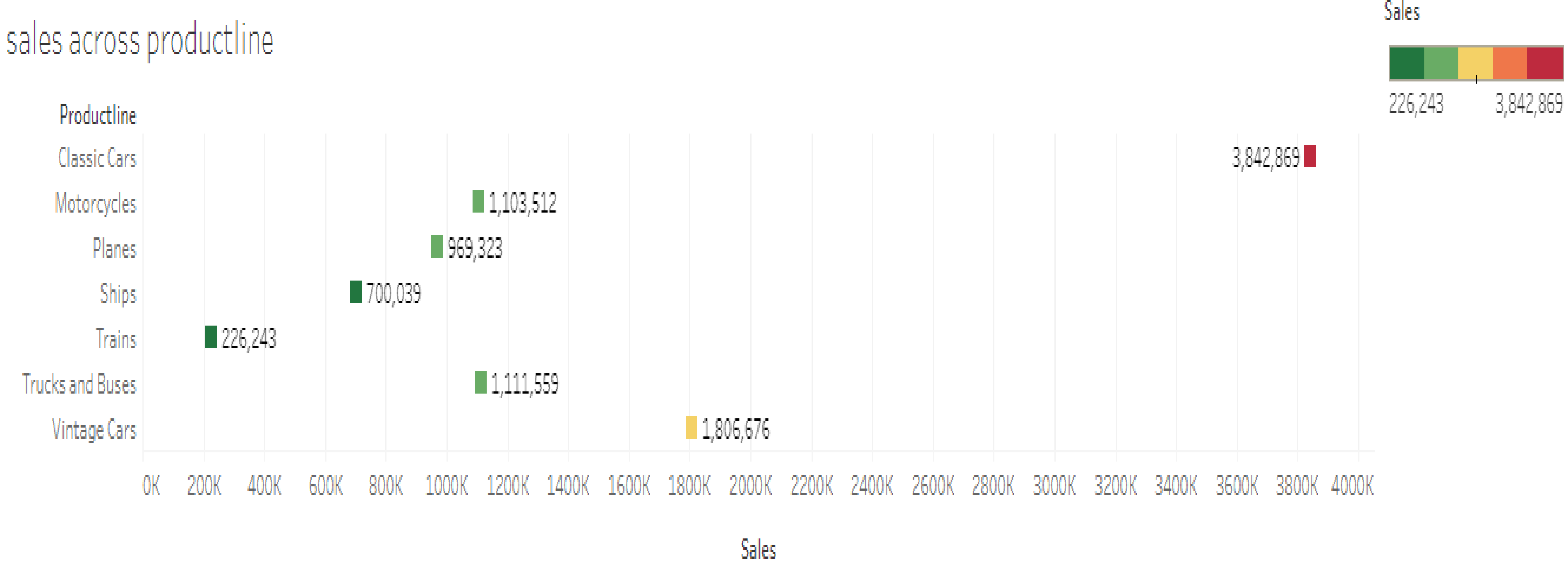
Sales Across Different Countries

- *We see that in sales across the country, USA has the maximum sales followed by Spain.*
- *The county with least or lowest sales is Ireland.*



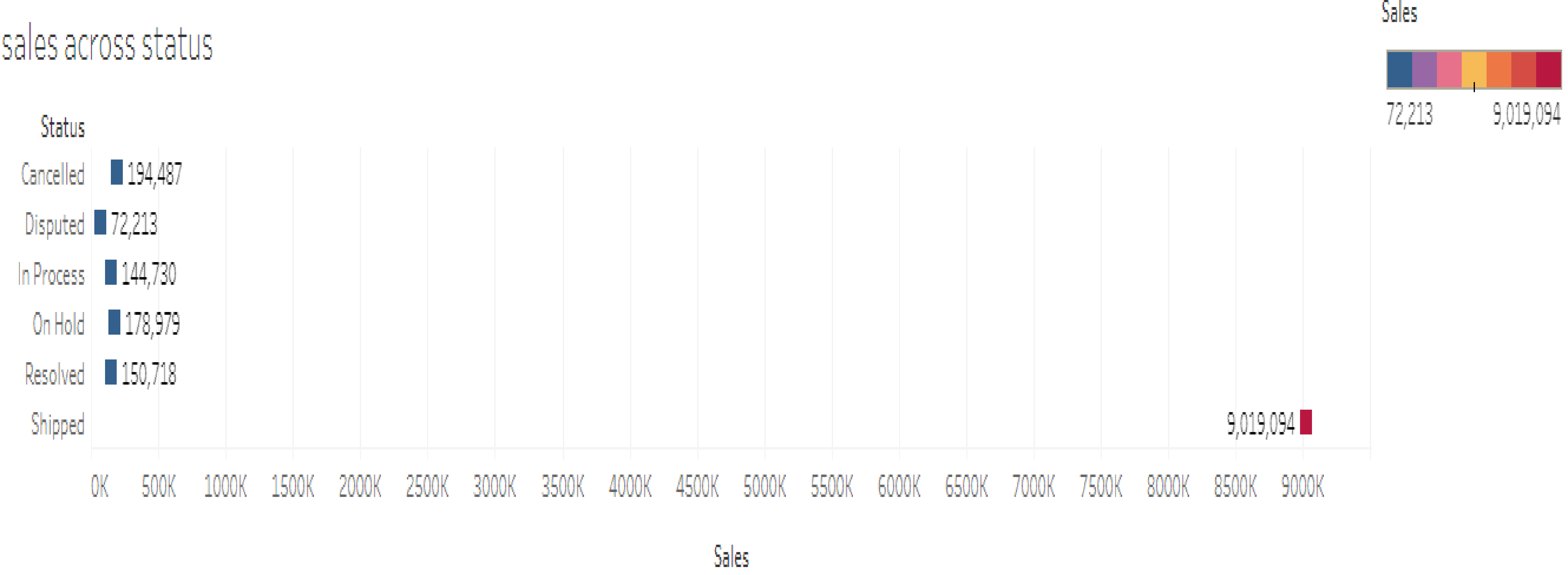
Sales Across Different Deal size

- *Most of the sales is happening in medium deal size. This means that most of the orders are of medium size and not too large or small.*
- *The flow of sales as per deal size is Medium > Small > Large.*
- *This means that least orders are of large deal size.*



Sales Across Different Product line

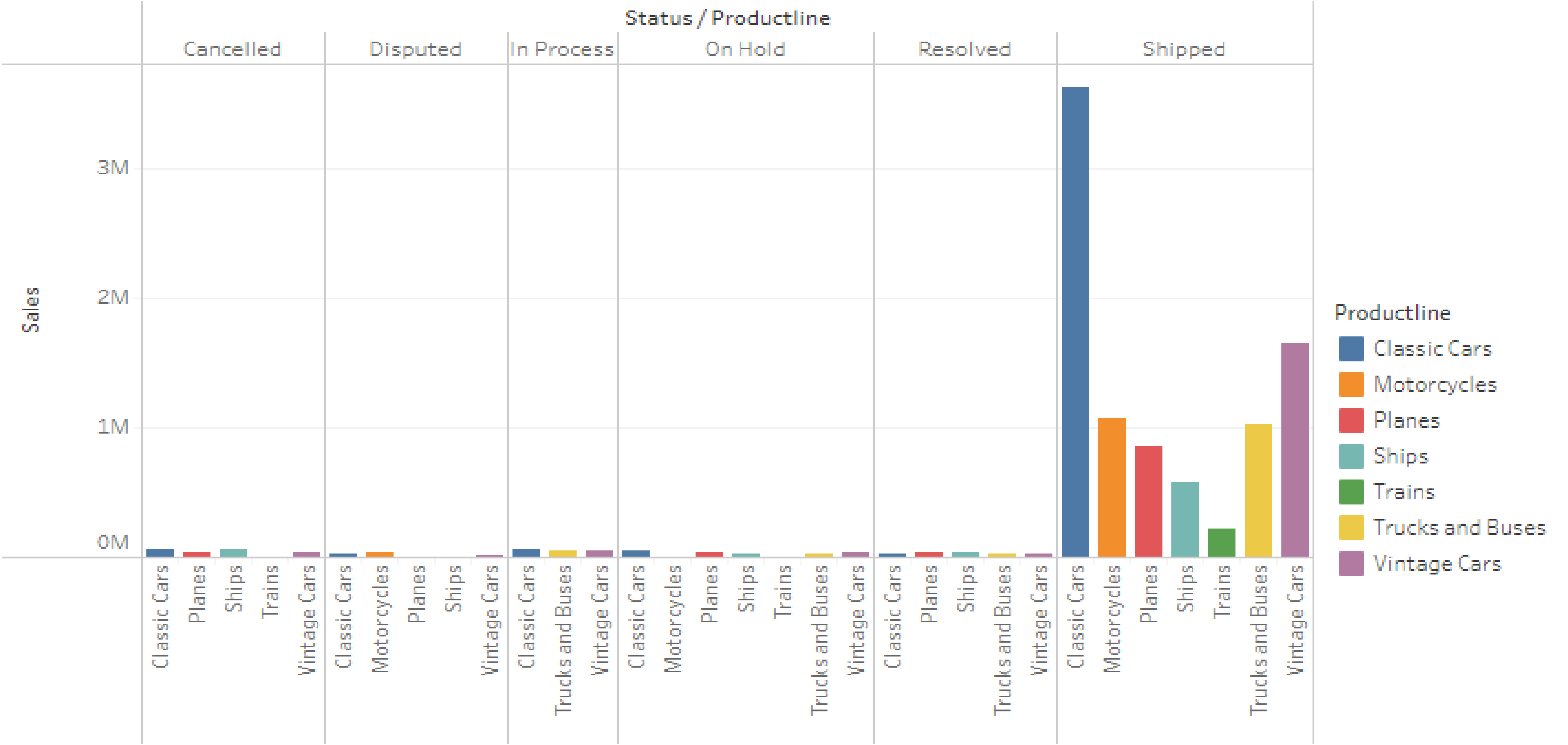
- *From the chart below we see that most of the sales is happening in Classic Cars.*
- *The flow of sales as per product line is Classic Cars > Vintage cars> Trucks and buses> motorcycles> planes> ships> trains.*
- *This means that least orders are of trains and highest of classic cars.*



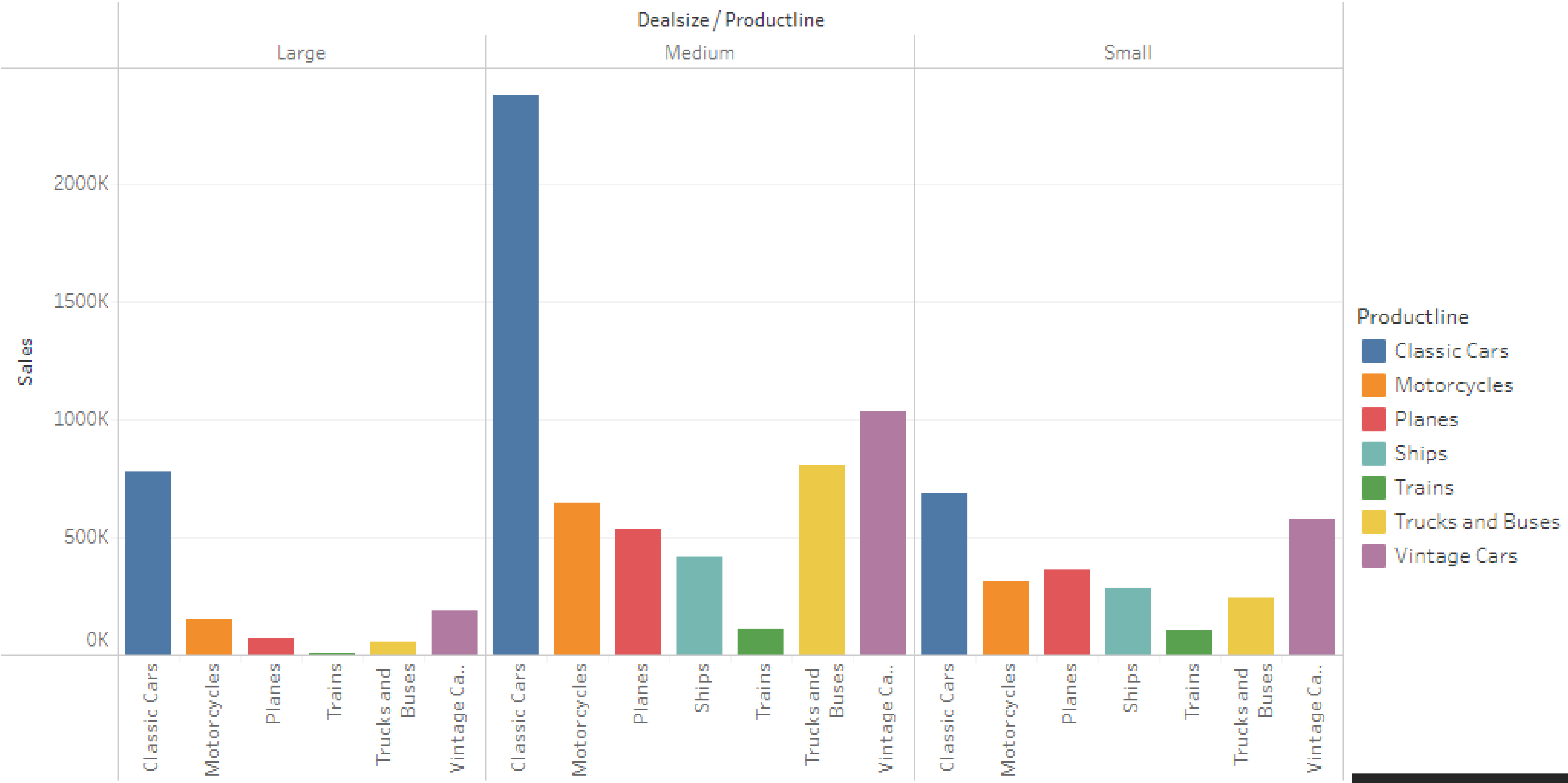
Sales Across Different Status

- *From the below chart we see that most of the orders are shipped.*
- *Few orders have been cancelled where as few are in process. .*
- *But we also see that orders which are under dispute are been resolved*
- *but still some are yet to be resolved.*

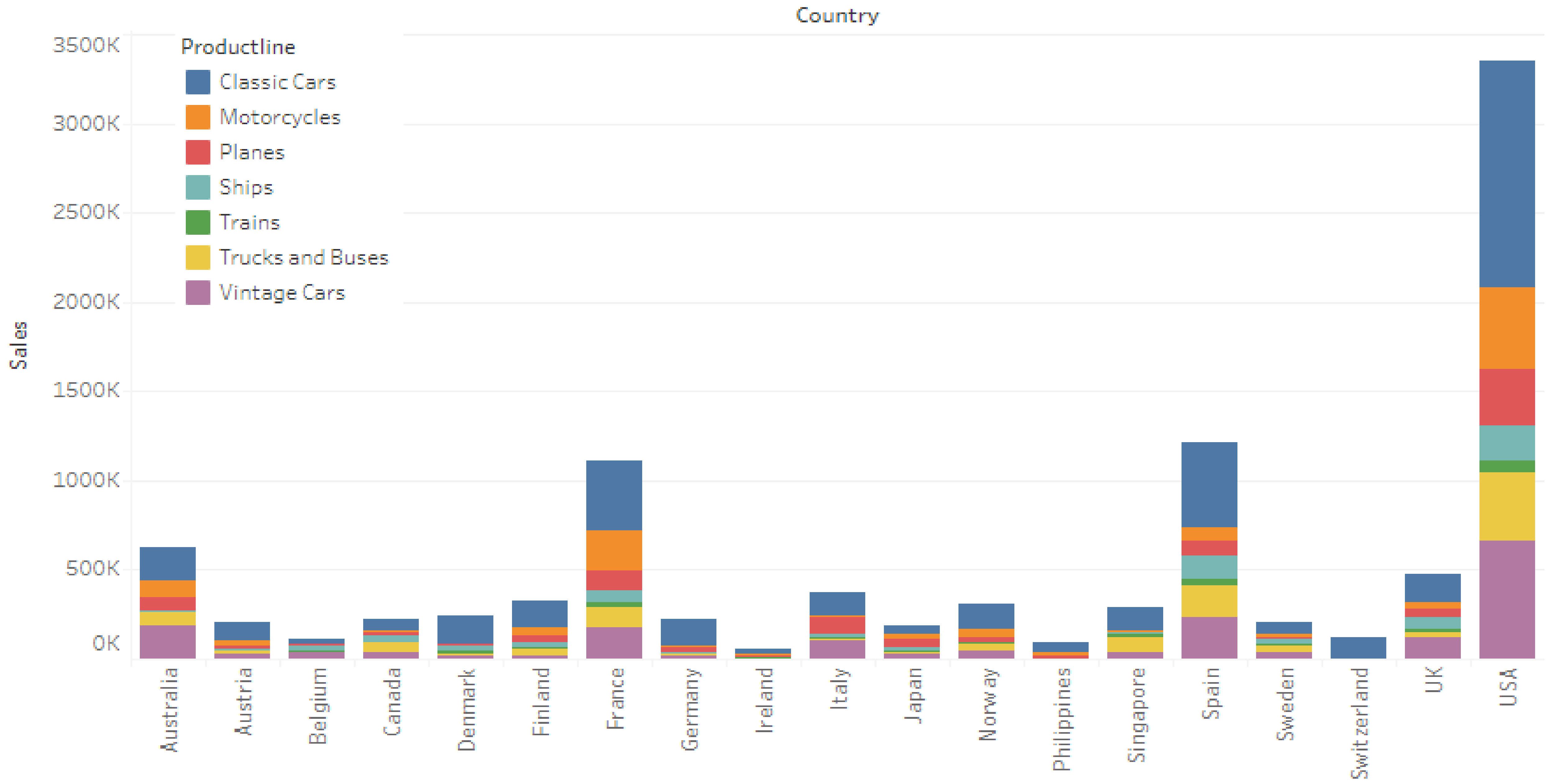
sales across status and productline



sales across dealsize and productline



sales across country and productline



- *From the chart trends across sales we come to know that the sales are maximum during the 4th quarter, to increase over all sales across all the quarters offers or discounts can be given to the customers.*
- *To increase the sales in countries with least sales, mega offers sales with low EMI facilities can be given to promote the sales.*
- *The large size deals are lowest and almost stagnant. Steps should be taken to promote and attract the customers to buy more of large size deals.*
- *Classic cars have majority sales whereas trucks and buses sales can be expanded.*
- *Should focus on cancelled orders to check why it was cancelled and work on it. Pending disputed orders should be resolved at the earliest so it doesn't gives a negative feel to the customer.*
- *Focus should be on expanding sales by taking in more customers by giving more offers like low EMI, festival sale, mega sale, etc.*

INFERENCE

RFM Analysis

- *Customer segmentation using RFM analysis.*
- *Here we are using jupyter notebook with Python codes.*

	ORDERNUMBER	CUSTOMERNAME	DAYS_SINCE_LASTORDER	QUANTITYORDERED	SALES
0	10107	Land of Toys Inc.	828	30	2871.00
1	10121	Reims Collectables	757	34	2765.90
2	10134	Lyon Souveniers	703	41	3884.34
3	10145	Toys4GrownUps.com	649	45	3746.70
4	10168	Technics Stores Inc.	586	36	3479.76

CRM & FRM Score table

- *Using CRM, We assigned weights of 0.33 to each of the factors.*
- *Post this, we fitted the data for the weights and obtained the scores*

	ORDERNUMBER	CUSTOMERNAME	recency	frequency	monetary	r_quartile	f_quartile	m_quartile
0	10107	Land of Toys Inc.	828	30	2871.00	1	3	3
1	10121	Reims Collectables	757	34	2765.90	1	3	3
2	10134	Lyon Souveniers	703	41	3884.34	1	2	2
3	10145	Toys4GrownUps.com	649	45	3746.70	1	1	2
4	10168	Technics Stores Inc.	586	36	3479.76	1	2	2
5	10180	Daedalus Designs Imports	573	29	2497.77	1	3	3
6	10188	Herkku Gifts	567	48	5512.32	1	1	1
7	10211	Auto Canal Petit	510	41	4708.44	1	2	1
8	10223	Australian Collectors, Co.	475	37	3965.66	1	2	2
9	10237	Vitachrome Inc.	432	23	2333.12	1	4	3

	ORDERNUMBER	CUSTOMERNAME	recency	frequency	monetary	r_quartile	f_quartile	m_quartile	RFM_Score
0	10107	Land of Toys Inc.	828	30	2871.00	1	3	3	133
1	10121	Reims Collectables	757	34	2765.90	1	3	3	133
2	10134	Lyon Souveniers	703	41	3884.34	1	2	2	122
3	10145	Toys4GrownUps.com	649	45	3746.70	1	1	2	112
4	10168	Technics Stores Inc.	586	36	3479.76	1	2	2	122
5	10180	Daedalus Designs Imports	573	29	2497.77	1	3	3	133
6	10188	Herkku Gifts	567	48	5512.32	1	1	1	111
7	10211	Auto Canal Petit	510	41	4708.44	1	2	1	121
8	10223	Australian Collectors, Co.	475	37	3965.66	1	2	2	122
9	10237	Vitachrome Inc.	432	23	2333.12	1	4	3	143

Loyal Customers

- *This is the top 10 list of Loyal customers because the RFM Score is as per the requirement.*

	ORDERNUMBER	CUSTOMERNAME	recency	frequency	monetary	r_quartile	f_quartile	m_quartile	RFM_Score
0	10107	Land of Toys Inc.	828	30	2871.00	1	3	3	133
1	10121	Reims Collectables	757	34	2765.90	1	3	3	133
2	10134	Lyon Souveniers	703	41	3884.34	1	2	2	122
3	10145	Toys4GrownUps.com	649	45	3746.70	1	1	2	112
4	10168	Technics Stores Inc.	586	36	3479.76	1	2	2	122
5	10180	Daedalus Designs Imports	573	29	2497.77	1	3	3	133
6	10188	Herkku Gifts	567	48	5512.32	1	1	1	111
7	10211	Auto Canal Petit	510	41	4708.44	1	2	1	121
8	10223	Australian Collectors, Co.	475	37	3965.66	1	2	2	122
9	10237	Vitachrome Inc.	432	23	2333.12	1	4	3	143

Best Customers

- *Here is the list of Top 10 Best Customers because they are frequently buyers and sales are higher then the others*

	ORDERNUMBER	CUSTOMERNAME	recency	frequency	monetary	r_quartile	f_quartile	m_quartile	RFM_Score
571	10407	The Sharp Gifts Warehouse	611	76	14082.8	1	1	1	111
714	10322	Online Diecast Creations Co.	924	50	12536.5	1	1	1	111
49	10424	Euro Shopping Channel	50	50	12001.0	1	1	1	111
1020	10412	Euro Shopping Channel	1049	60	11887.8	1	1	1	111
96	10403	UK Collectables, Ltd.	150	66	11886.6	1	1	1	111
42	10312	Mini Gifts Distributors Ltd.	266	48	11623.7	1	1	1	111
176	10127	Muscle Machine Inc	905	46	11279.2	1	1	1	111
28	10150	Dragon Souveniers, Ltd.	649	45	10993.5	1	1	1	111
186	10247	Suominen Souveniers	579	44	10606.2	1	1	1	111
41	10304	Auto Assoc. & Cie.	275	47	10172.7	1	1	1	111

Lost Customers

- *Here we can observe the Top 10 lost customers because they are not a frequent buyers and they can not order anything for the long time.*

	ORDERNUMBER	CUSTOMERNAME	recency	frequency	monetary	r_quartile	f_quartile	m_quartile	RFM_Score
1797	10288	Handji Gifts& Co	2071	34	2328.66	3	3	3	333
1790	10192	Online Diecast Creations Co.	2349	32	2328.64	3	3	3	333
2197	10314	Heintze Collectables	2420	35	2327.15	3	3	3	333
1636	10124	Signal Gift Stores	2378	32	2326.40	3	3	3	333
1947	10296	Bavarian Collectables Imports, Co.	2207	32	2292.80	3	3	3	333
1786	10148	Anna's Decorations, Ltd	2415	31	2282.22	3	3	3	333
1854	10386	Euro Shopping Channel	1946	35	2231.60	3	3	3	333
1699	10284	Norway Gifts By Mail, Co.	1984	30	2219.70	3	3	3	333
2029	10423	Petit Auto	2031	28	2208.92	3	3	3	333
1814	10203	Euro Shopping Channel	2361	34	2206.60	3	3	3	333

Conclusion

- *In this tutorial, covered a lot of details about Customer Segmentation. I have learned what the customer segmentation is, Need of Customer Segmentation, Types of Segmentation, RFM analysis, Implementation of RFM from scratch in python. Also, covered some basic concepts of pandas such as handling duplicates, groupby, and qcut() for bins based on sample quintiles.*

Tableau Link

https://public.tableau.com/views/MilestoneMRAHARSHPANDYA/trendsacrosssales?:language=en-US&:display_count=n&:origin=viz_share_link