MRA PROJECT ML 1

USER NAME : DINYA ANTONY

BATCH : AUG21_A1

EMAIL : VOLLAFREEBIRD@GMAIL.COM

TABLE OF CONTENTS

❖ ***Problem statement.***

❖ ***Analyzing the data.***

  ➢ **Info, Shape, Describe, datatype, null values**

❖ ***EDA***

  ➢ **Univariate, Bivariate, and multivariate analysis.**

❖ ***Sales Trends***

  ➢ **Weekly, Monthly, Quarterly, Yearly**

❖ ***Sales Across different Categories***

v ***RFM***

  Ø ***Meaning, parameters used, output head***

❖ ***Inferences from RFM Analysis and identified segments***

  ➢ ***Best 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.

ANALYSING THE DATA

- ***Data information:***

```
Range Index: 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   int64
9      MSRP                    2747 non-null   int64
```

ANALYSING THE DATA

❖ Data information:

```
#          Column                    Non-Null Count    Dtype
---         ------                     --------------      -----
10         PRODUCTCODE               2747 non-null   float64
11         CUSTOMERNAME              2747 non-null   int64
12          PHONE                    2747 non-null   float64
13          ADDRESSLINE1             2747 non-null   datetime64[ns]
14          CITY                     2747 non-null   int64
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
```

ANALYSING THE DATA

- **Data head:**

ORDERNUMBER	QTYORDERED	PRICEEACH	ORDERLINENUMBER	SALES	ORDERDATE	DAYS_SINCE_LASTORDER	STATUS	PRODUCTLINE	MSRP	PRODUCTCODE	CUSTOMERNAME	PHONE	ADDRLINE1	CITY	POSTALCODE	COUNTRY	CONTACTLASTNAME	CONTACTFIRSTNAME	DEALSIZE
10107	30	95.7	2	2871	2018-02-24	828	Shipped	Motorcycles	95	S10_1678	Land of Toys Inc.	2125557818	897 Long Airport Avenue	NYC	10022	USA	Yu	Kwai	Small
10121	34	81.35	5	2765.9	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
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
10145	45	83.26	6	3746.7	2018-08-25	649	Shipped	Motorcycles	95	S10_1678	Toys4GrownUps.com	6265557265	78934 Hillside Dr.	Pasadena	90003	USA	Young	Julie	Medium
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

ANALYSING THE DATA

- *Data shape: (2747, 20)*
- **Describe the data: Numeric data:**

	Count	*Mean*	*STD*	*MIN*	*25.00%*	*50.00%*	*75.00%*	*MAX*
ORDERNUMBER	*2747*	*10259.761558*	*91.877521*	*10100*	*10181*	*10264*	*10334.5*	*10425*
QUANTITYORDERED	*2747*	*35.103021*	*9.762135*	*6*	*27*	*35*	*43*	*97*
PRICEEACH	*2747*	*101.098951*	*42.042548*	*26.88*	*68.745*	*95.55*	*127.1*	*252.87*
ORDERLINENUMBER	*2747*	*6.491081*	*4.230544*	*1*	*3*	*6*	*9*	*18*
SALES	*2747*	*3553.047583*	*1838.953901*	*482.13*	*2204.35*	*3184.8*	*4503.095*	*14082.8*
DAYS_SINCE_LASTORDER	*2747*	*1757.085912*	*819.280576*	*42*	*1077*	*1761*	*2436.5*	*3562*
MSRP	*2747*	*100.691664*	*40.114802*	*33*	*68*	*99*	*124*	*214*

ANALYSING THE DATA

❖ Describe the data: Categorical data:

	Count	Unique	Top	Freq
STATUS	*2747*	*6*	*Shipped*	*2541*
PRODUCTLINE	*2747*	*7*	*Classic Cars*	*949*
PRODUCTCODE	*2747*	*109*	*S18_3232*	*51*
CUSTOMERNAME	*2747*	*89*	*Euro Shopping Channel*	*259*
PHONE	*2747*	*88*	*(91) 555 94 44*	*259*
ADDRESSLINE1	*2747*	*89*	*C/ Moralzarzal, 86*	*259*
CITY	*2747*	*71*	*Madrid*	*304*
POSTALCODE	*2747*	*73*	*28034*	*259*
COUNTRY	*2747*	*19*	*USA*	*928*
CONTACTLASTNAME	*2747*	*76*	*Freyre*	*259*
CONTACTFIRSTNAME	*2747*	*72*	*Diego*	*259*
DEALSIZE	*2747*	*3*	*Medium*	*1349*

ANALYSING THE DATA

❖ Interpretation:

- ***The data has 2747 rows and 20 columns with int, Float and object as the data type.***
- ***We have no non-null data with 20 variables. Numeric 7 variables, 1 date-time and 12 object types.***
- ***The summary stats: average item price is approximate 101, varies from 26-253.***
- ***The orders that are line, its average is around 6.***
- ***The sales average is 3553.***
- ***The a***utomobile parts manufacturing company has customer re-order interval from 42 days to 3562 days.
- ***The MSRP average is in close range to the item price average 100.***
- ***It shows that the manufacturing company sell the items within a small range difference from making cost.***
- ***7-category we have in product line and deal size is small, medium or large.***
- ***There are 6 different status, stage of the order.***
- ***Also our data set features 19 countries data of manufacturing company.***
    - ***With 71 different cities.***
- ***Order size, base of quantity varies from 6 to 97, gives us a sense that may sell B2B and B2C.***

EXPLORATORY DATA ANALYSIS

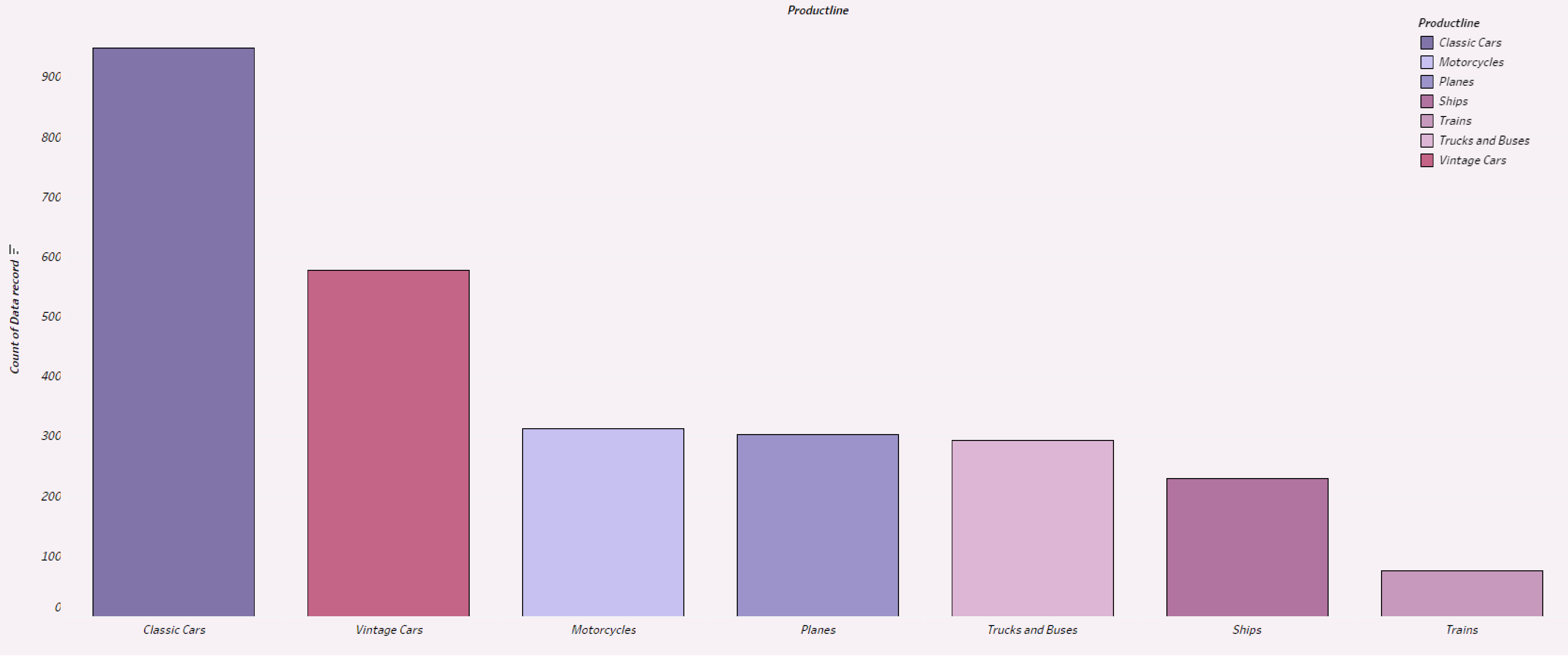
❖ Pre_check_eda Table:

	Null values	Data types
ORDERNUMBER	0	int64
QUANTITYORDERED	0	int64
PRICEEACH	0	float64
ORDERLINENUMBER	0	int64
SALES	0	float64
ORDERDATE	0	datetime64[ns]
DAYS_SINCE_LASTORDER	0	int64
STATUS	0	object
PRODUCTLINE	0	object
MSRP	0	int64
PRODUCTCODE	0	object
CUSTOMERNAME	0	object
PHONE	0	object
ADDRESSLINE1	0	object
CITY	0	object
POSTALCODE	0	object
COUNTRY	0	object
CONTACTLASTNAME	0	object
CONTACTFIRSTNAME	0	object
DEALSIZE	0	object

EXPLORATORY DATA ANALYSIS

EDA - Univariate Analysis: Categorical variables: Product Categories

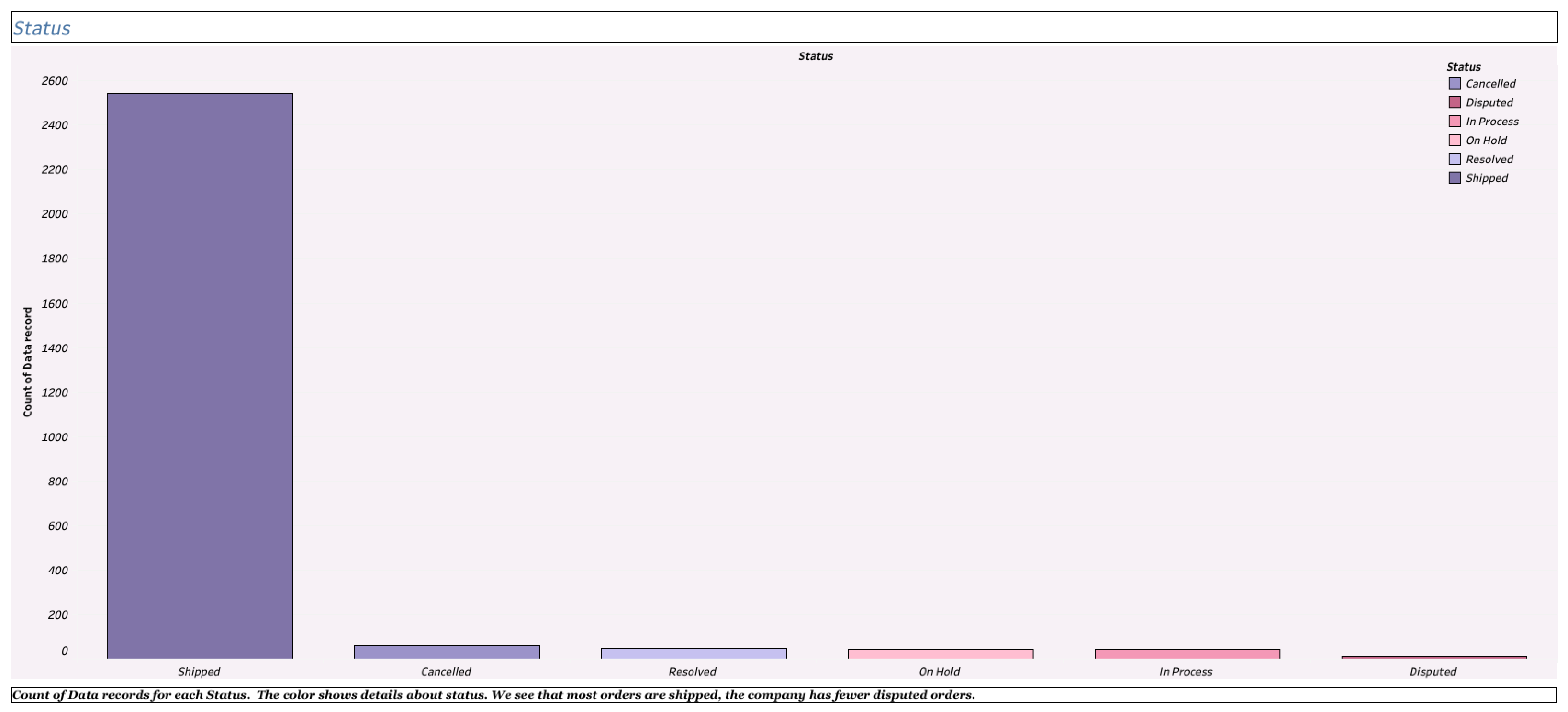
Product Category



Count of Data records for each Productline. The color shows details about Productline. We see that classic Cars have the most orders and Trains have the least.

EXPLORATORY DATA ANALYSIS

Univariate Analysis: Categorical variables: Status

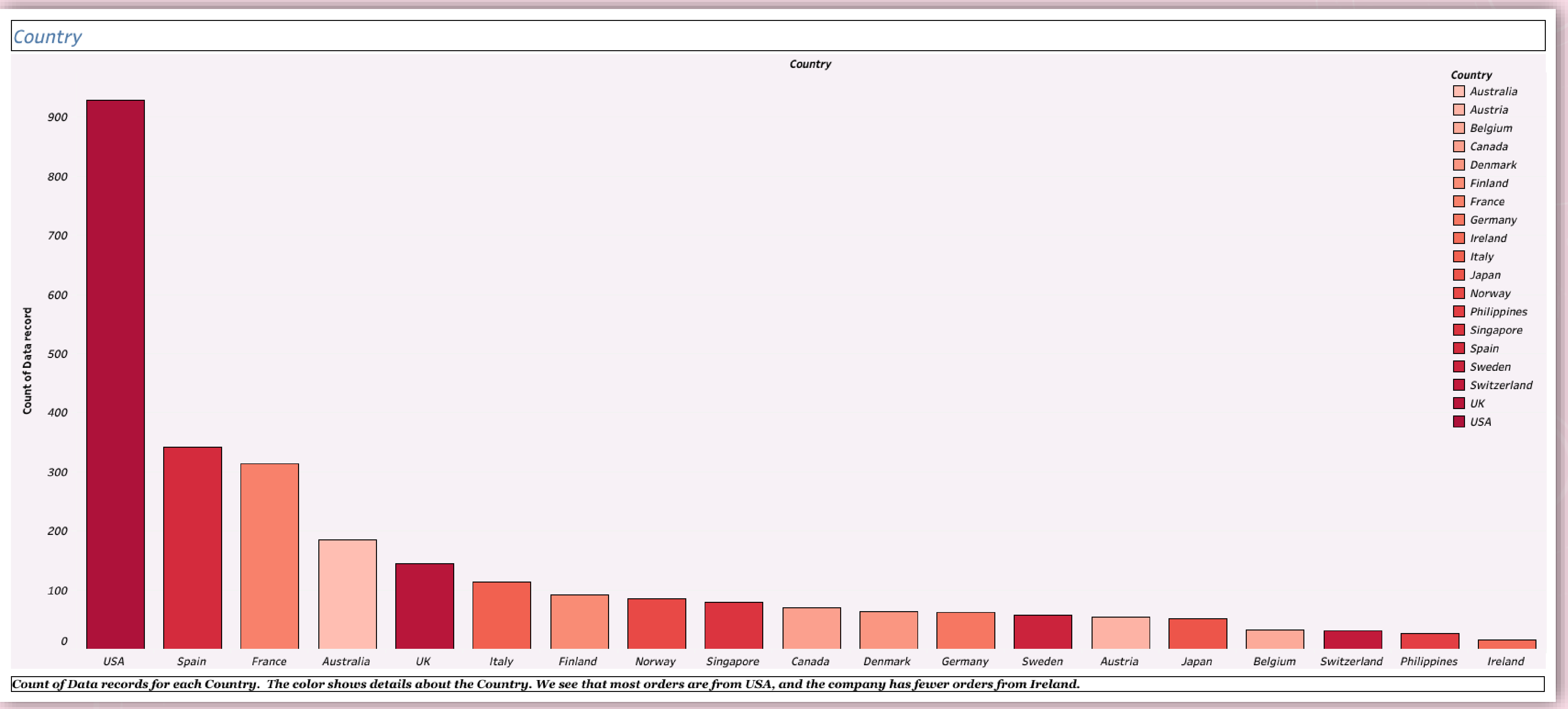


Status

Count of Data records for each Status. The color shows details about status. We see that most orders are shipped, the company has fewer disputed orders.

EXPLORATORY DATA ANALYSIS

Univariate Analysis: Categorical variables: Country



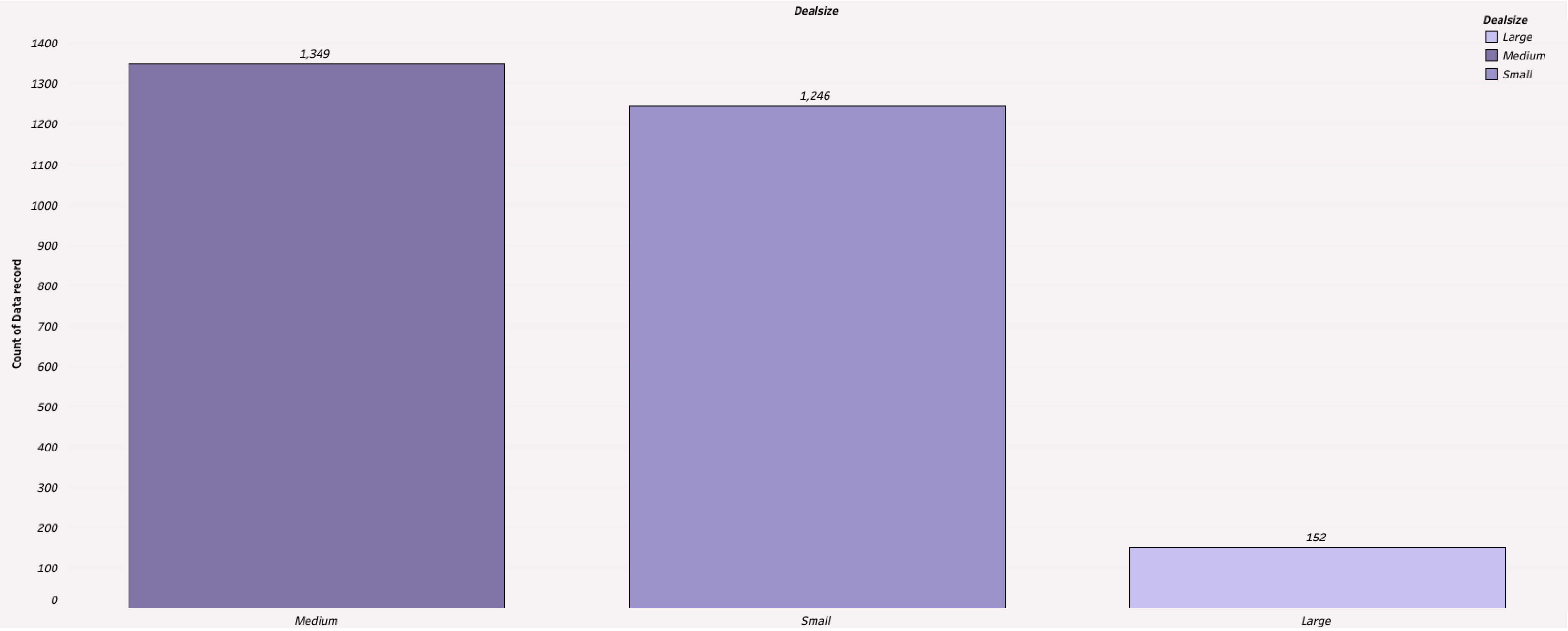
Country

Count of Data records for each Country. The color shows details about the Country. We see that most orders are from USA, and the company has fewer orders from Ireland.

EXPLORATORY DATA ANALYSIS

Univariate Analysis: Categorical variables: Deal Size

DEALSIZE

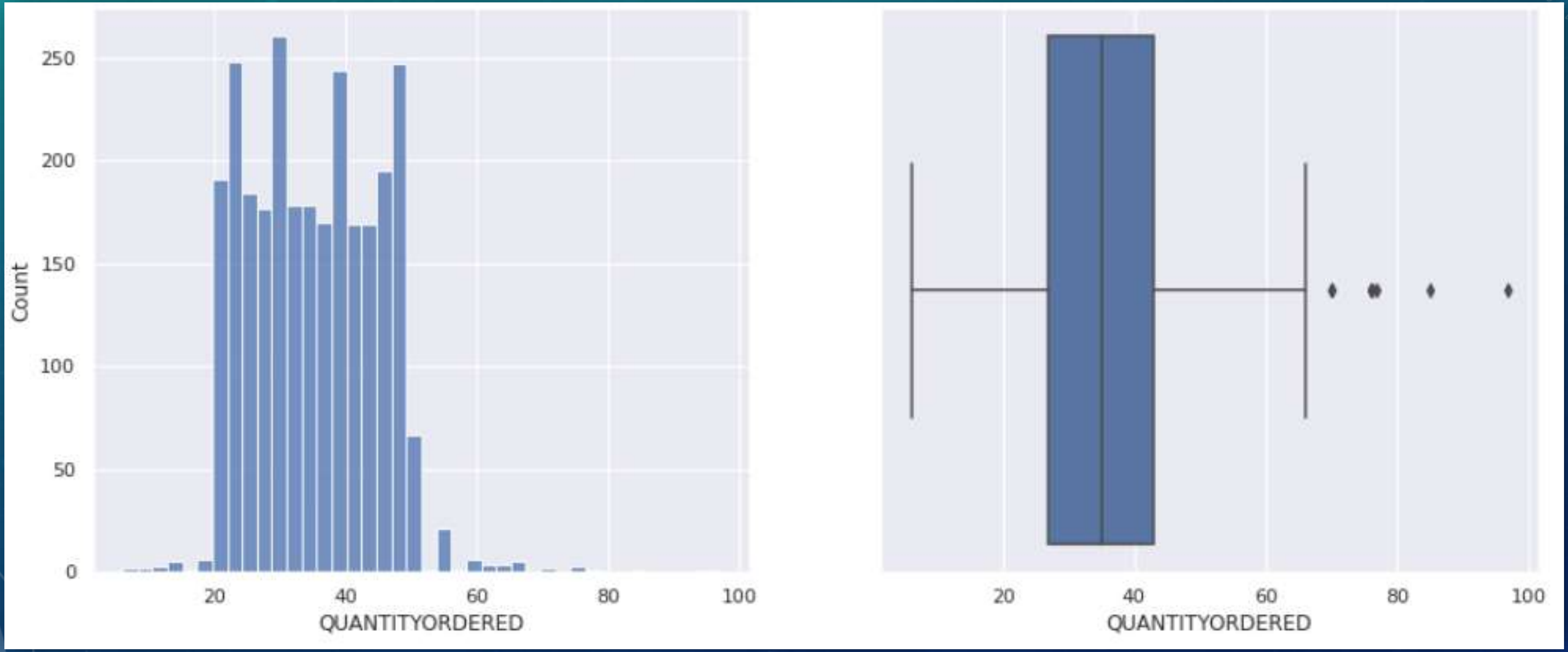


Count of Data records for each Dealsize. The color shows details about Dealsize. Company gets fewer orders that are large in size.

EXPLORATORY DATA ANALYSIS

Univariate Analysis: Numerical variables: Quantity ordered

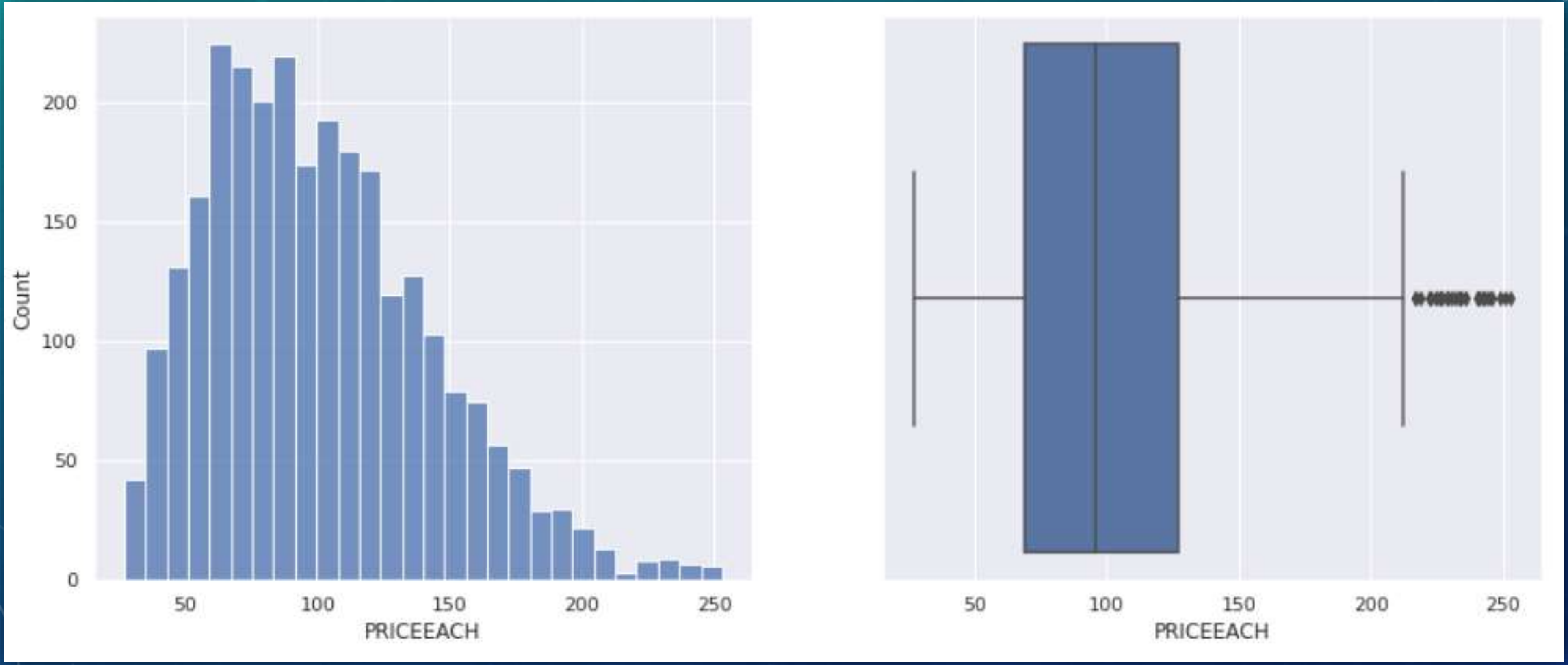
- ***The Boxplot tells us there are few outliers Quantity ordered distribution.***
- ***The distplot distribution can be said to be a mostly normal distribution. The distribution ranges mostly between 20 to 50, with few outlier, below and above the range.***



EXPLORATORY DATA ANALYSIS

Univariate Analysis: Numerical variables: PRICEEACH

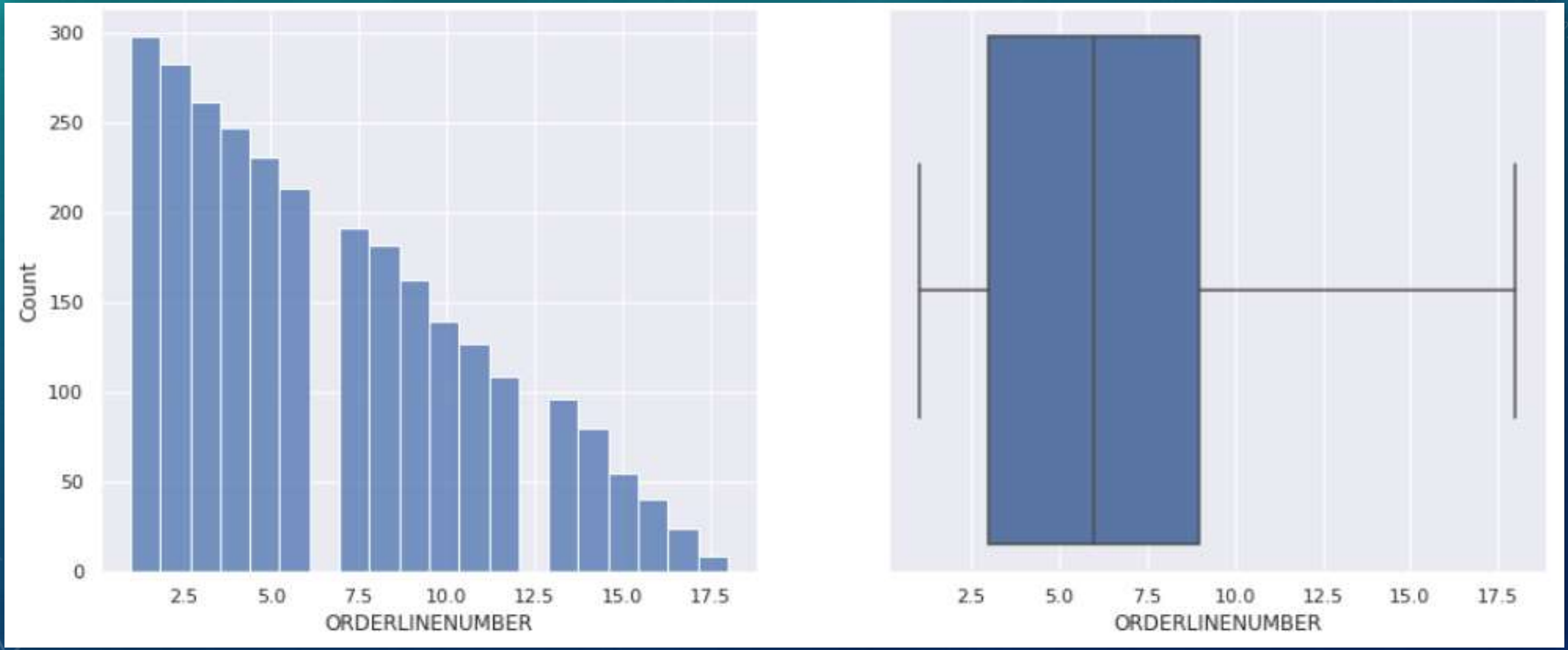
- ***The Boxplot tells us there are few outliers Price of each item distribution.***
- ***The distribution can be said to be left-skewed. The distribution ranges between 26 to 252.***



EXPLORATORY DATA ANALYSIS

Univariate Analysis: Numerical variables: ORDERLINENUMBER

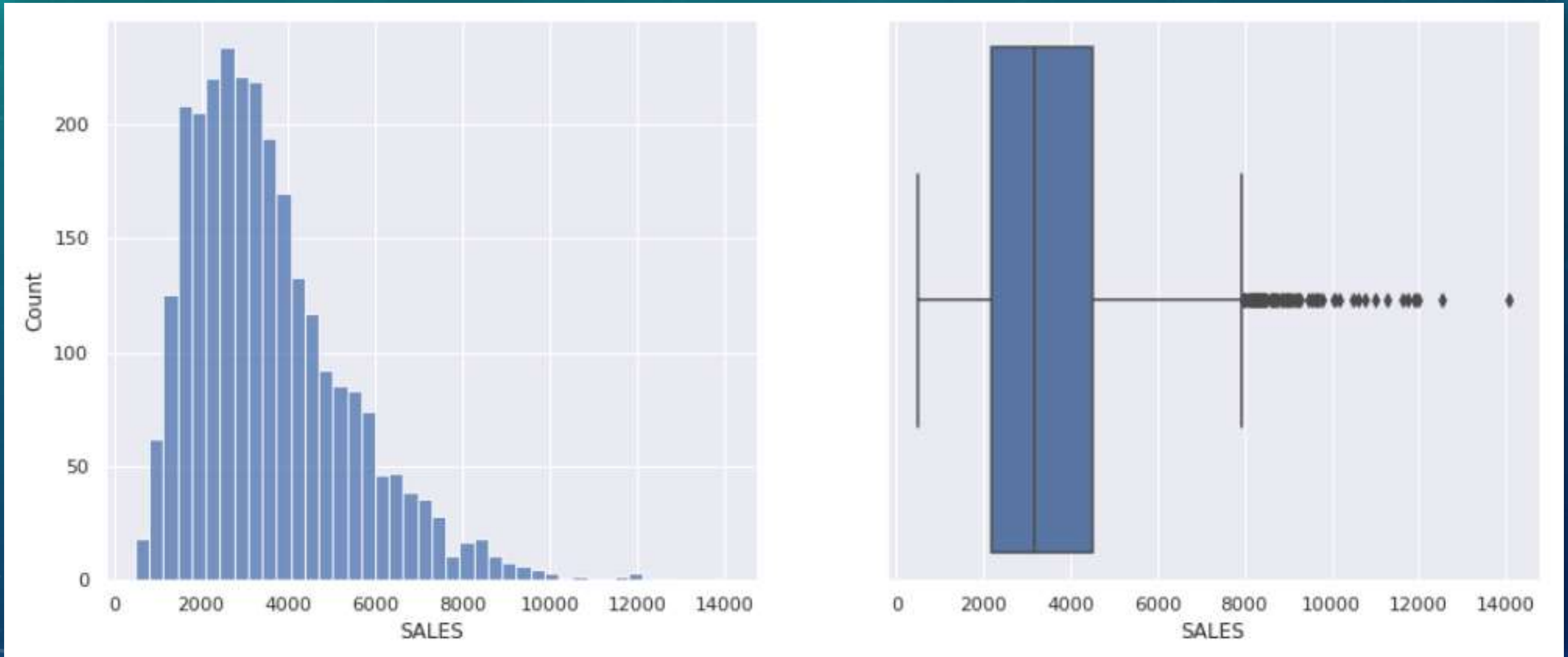
- ***The Boxplot tells us there are no outliers in line order number distribution.***
- ***The distribution can be said to be highly left-skewed. The distribution ranges between 1 to 18.***



EXPLORATORY DATA ANALYSIS

Univariate Analysis: Numerical variables: SALES

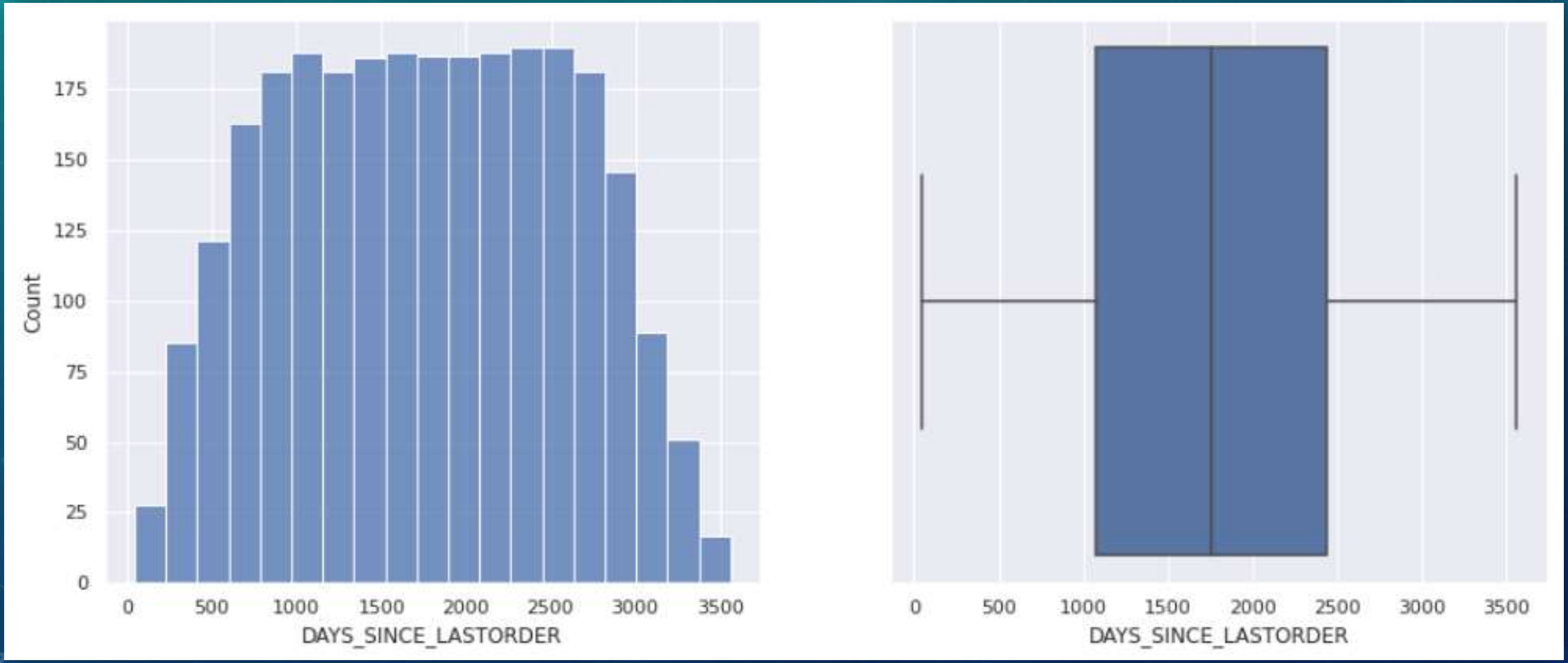
- *The Boxplot tells us there are quick a few outliers sales distribution.*
- *The distribution can be said to be highly left-skewed. The distribution ranges between 482 to 14082.*



EXPLORATORY DATA ANALYSIS

Univariate Analysis: Numerical variables: DAYS_SINCE_LASTORDER

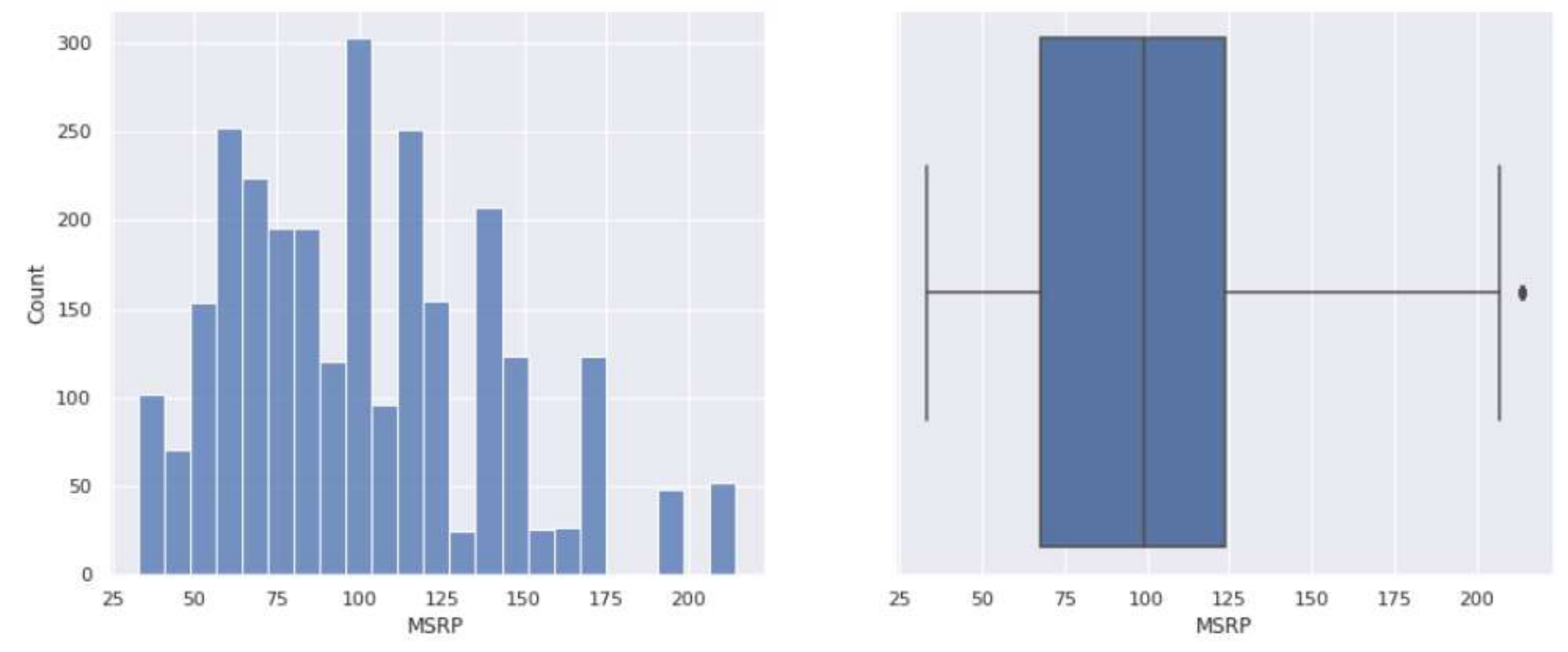
- *The Boxplot tells us there are no outliers* DAYS_SINCE_LAST_ORDER *distribution.*
- *The distplot distribution can be said to be a mostly normal distribution. The distribution ranges between 42 to 3562.*



EXPLORATORY DATA ANALYSIS

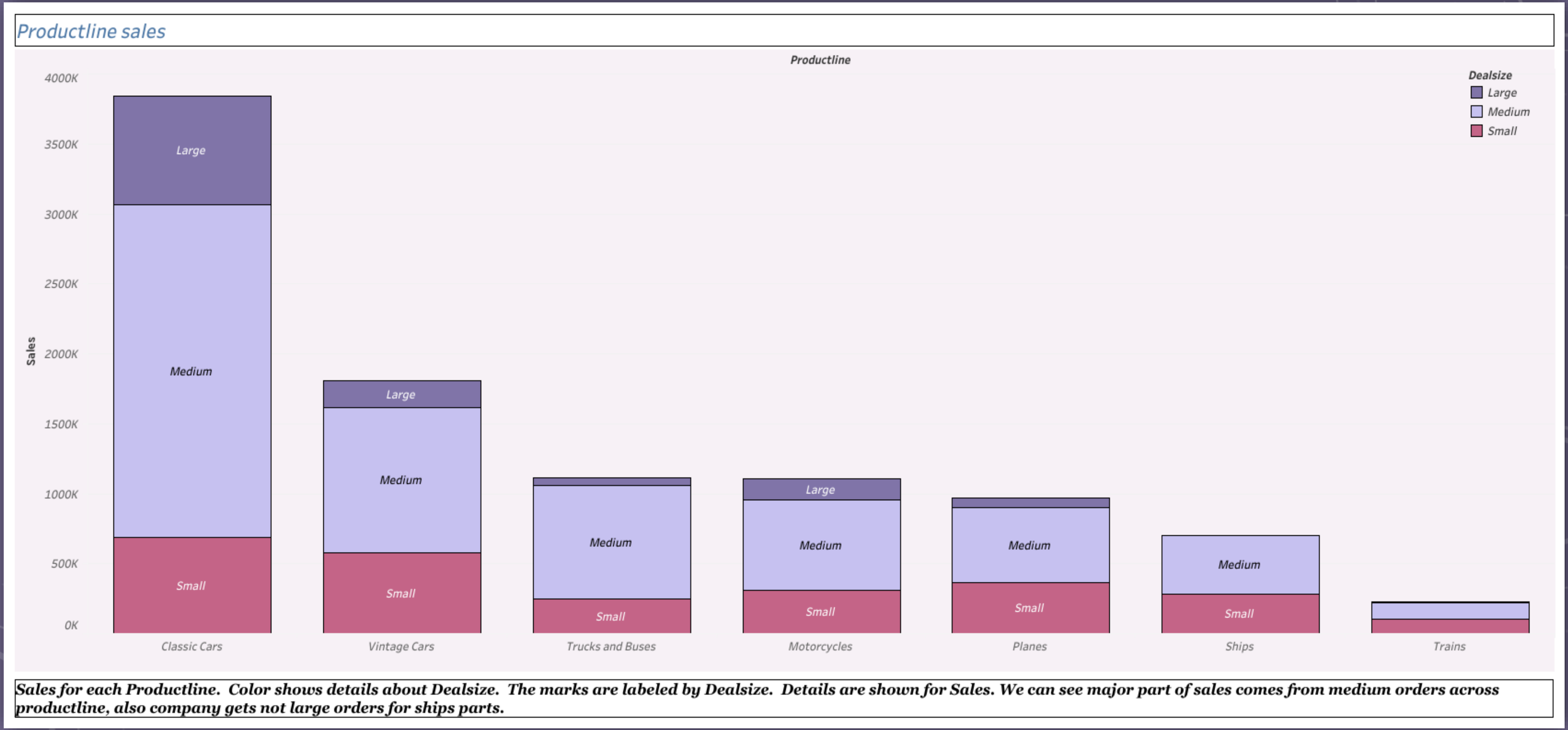
Univariate Analysis: Numerical variables: MSRP

- ***The Boxplot tells us there are no outliers Manufacturer's Suggested Retail Price distribution.***
- ***The distplot distribution can be said to be a left-skewed distribution. The distribution ranges between 33 to 214.***



EXPLORATORY DATA ANALYSIS

Bivariate Analysis: Product line sales.

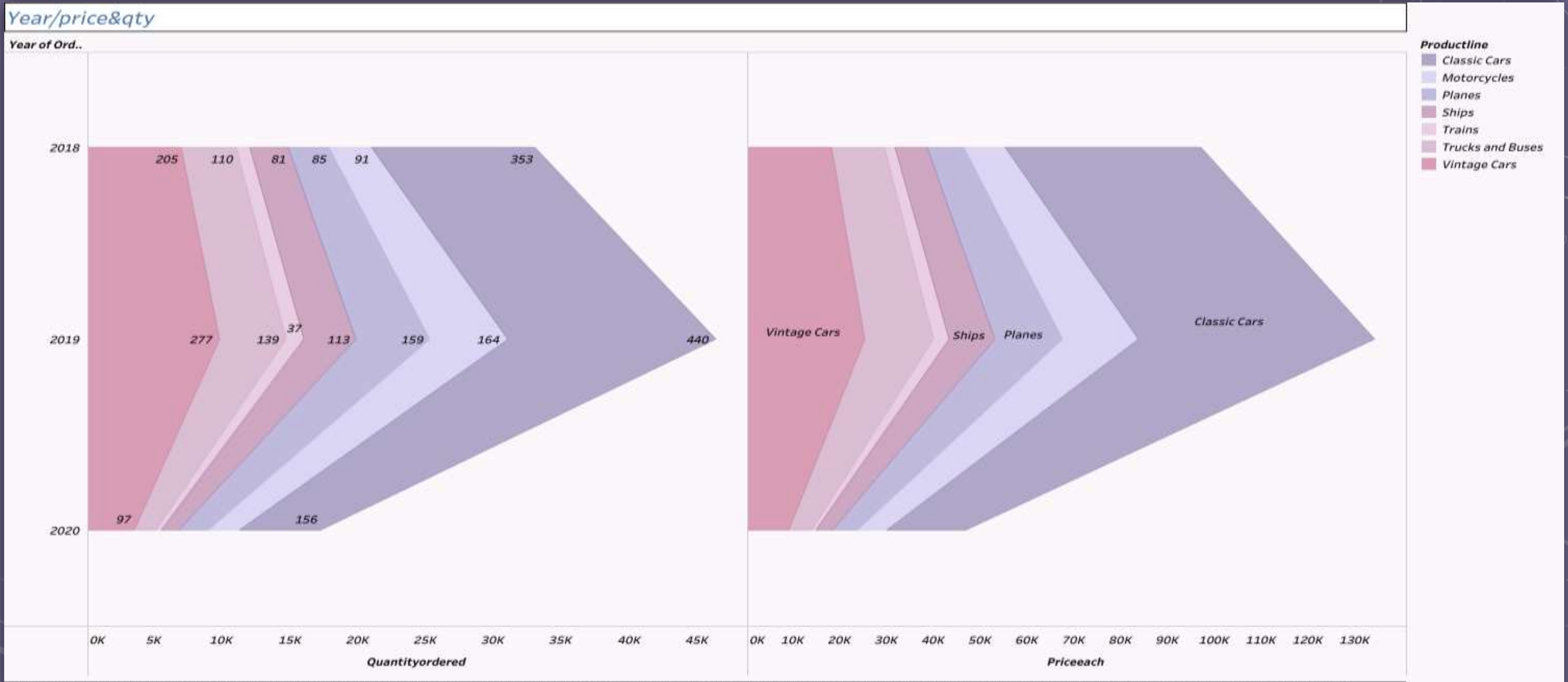


Productline sales

Sales for each Productline. Color shows details about Dealsize. The marks are labeled by Dealsize. Details are shown for Sales. We can see major part of sales comes from medium orders across productline, also company gets not large orders for ships parts.

EXPLORATORY DATA ANALYSIS

Bivariate Analysis: Order Qty and unit price over 3 years.



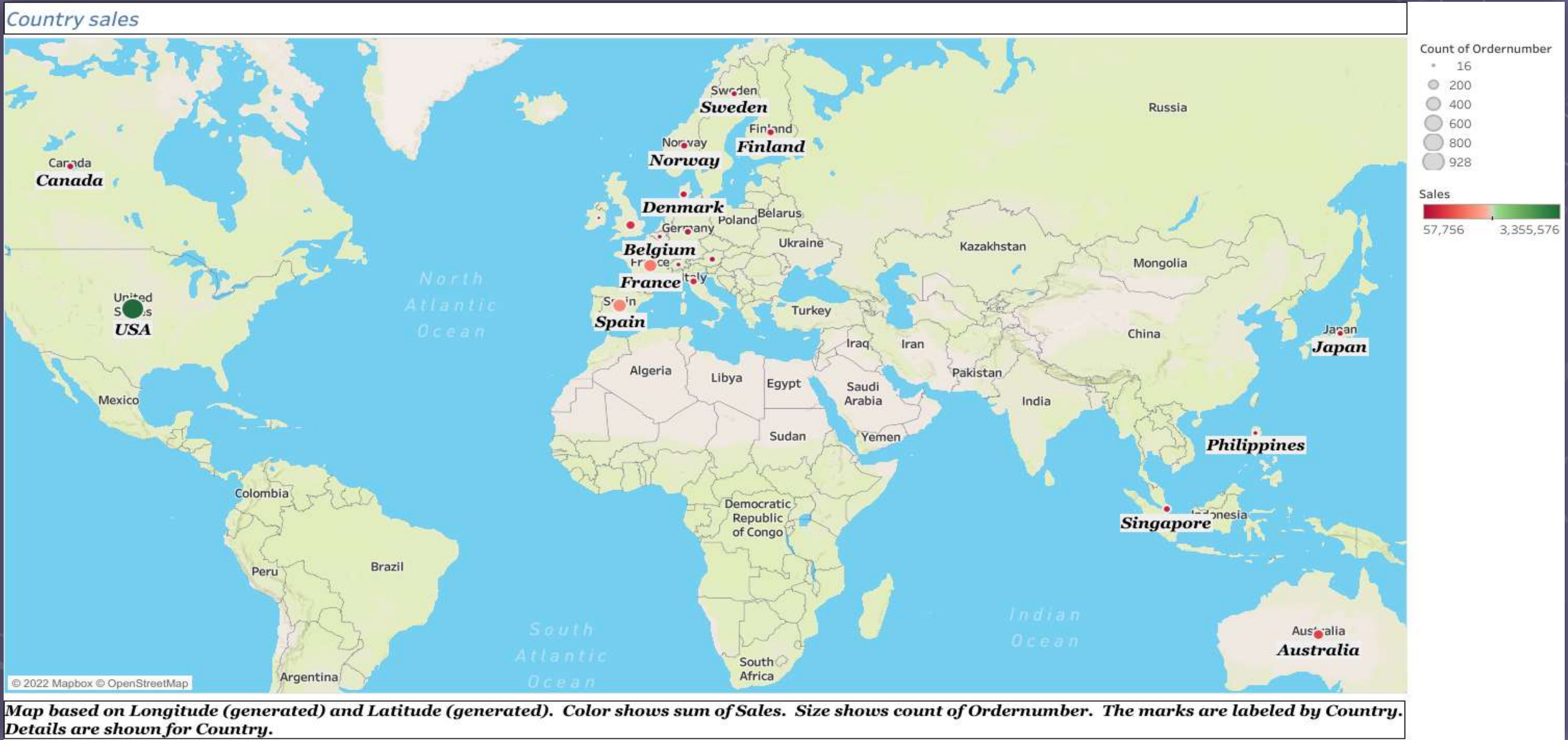
Year/price&qty

Sum of Quantityordered and sum of Priceeach for each Orderdate Year. Color shows details about Productline. For pane Sum of Priceeach: The marks are labeled by Productline. For pane Sum of Quantityordered: The marks are labeled by count of Ordernumber. Both price unit and order quantity has rise during 2019.

EXPLORATORY DATA ANALYSIS

Bivariate Analysis: Country Sales

- ***We can see most number of orders are from European part of the globe, the highest sales is given by USA.***



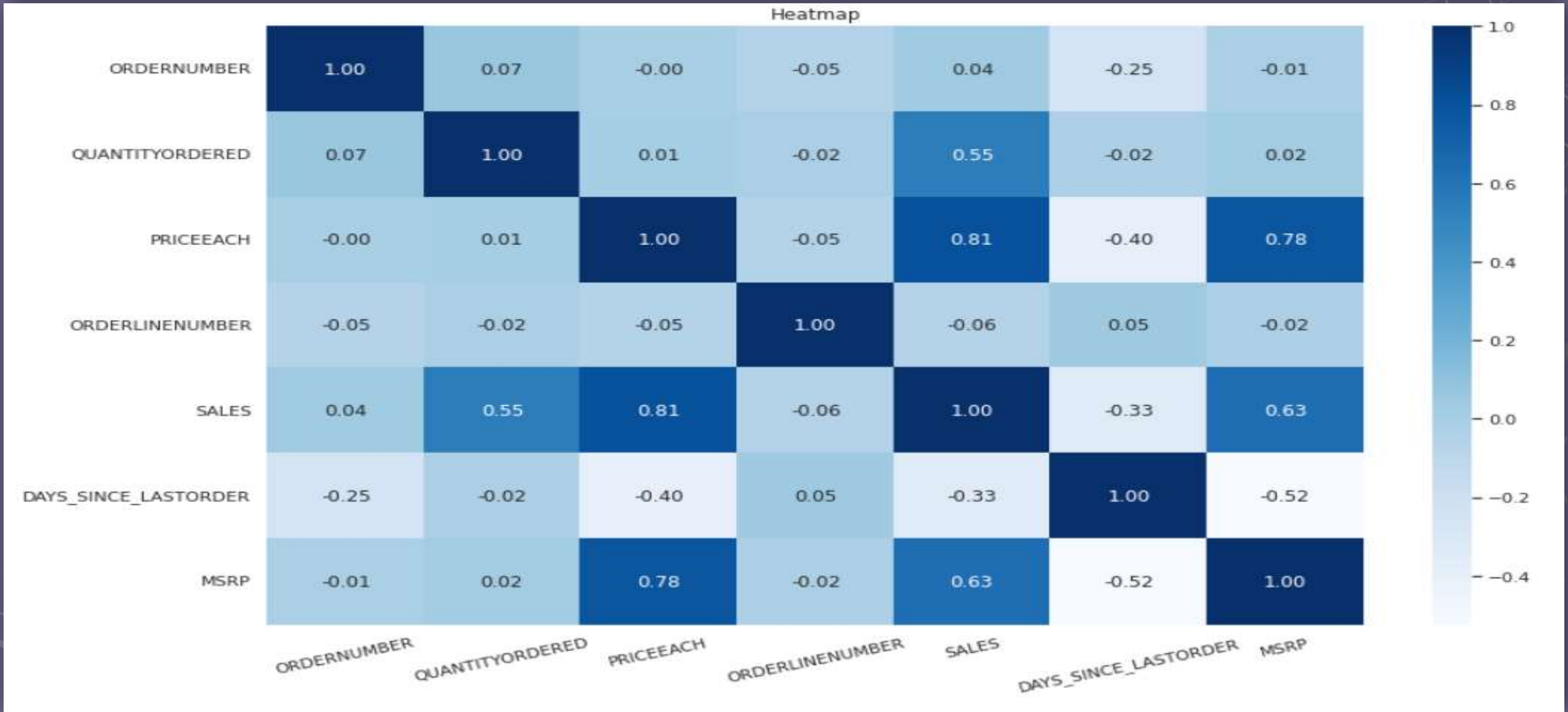
Country sales

Map based on Longitude (generated) and Latitude (generated). Color shows sum of Sales. Size shows count of Ordernumber. The marks are labeled by Country. Details are shown for Country.

EXPLORATORY DATA ANALYSIS

Multivariate Analysis: Heatmap

- ***We can see that PriceEach(unit) is highly corelated to Sales and MSRP.***



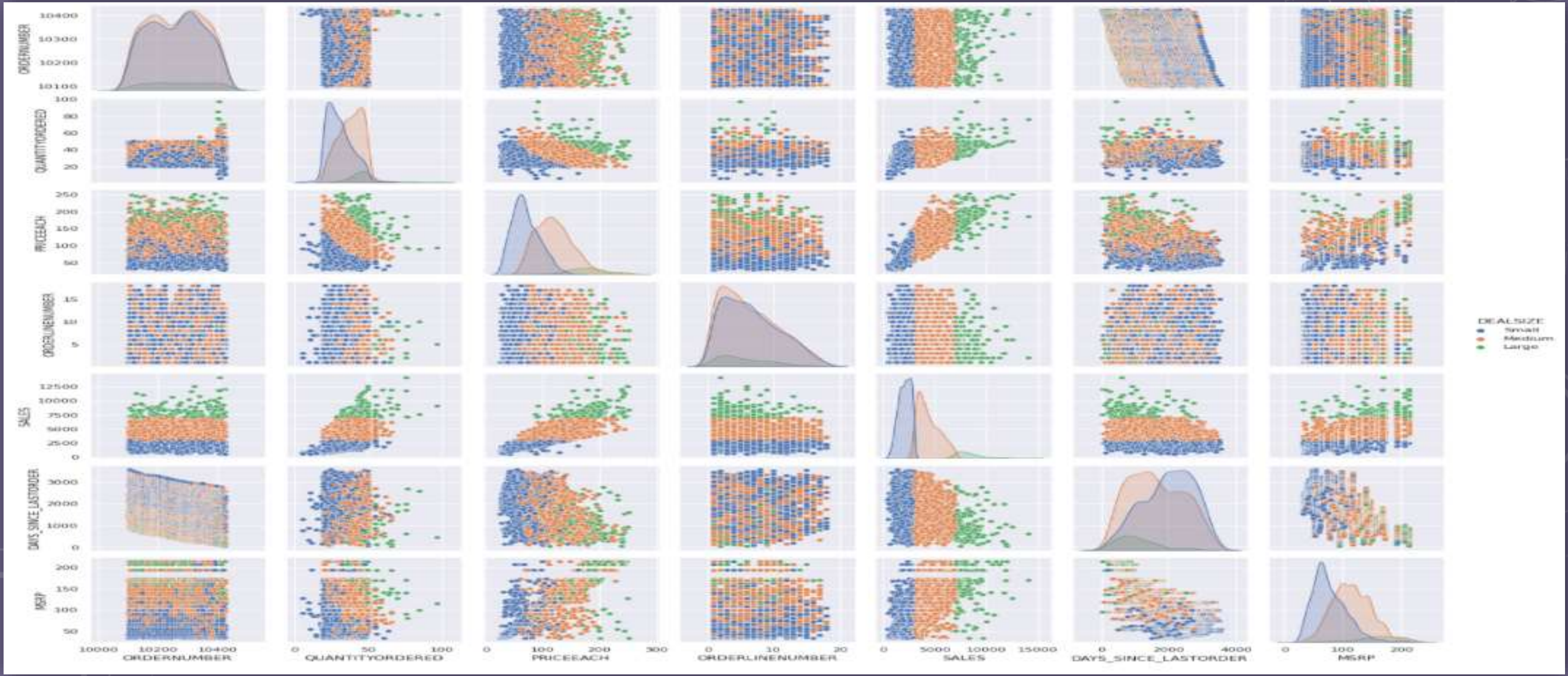
Heatmap

	ORDERNUMBER	QUANTITYORDERED	PRICEEACH	ORDERLINENUMBER	SALES	DAYS_SINCE_LASTORDER	MSRP
ORDERNUMBER	1.00	0.07	-0.00	-0.05	0.04	-0.25	-0.01
QUANTITYORDERED	0.07	1.00	0.01	-0.02	0.55	-0.02	0.02
PRICEEACH	-0.00	0.01	1.00	-0.05	0.81	-0.40	0.78
ORDERLINENUMBER	-0.05	-0.02	-0.05	1.00	-0.06	0.05	-0.02
SALES	0.04	0.55	0.81	-0.06	1.00	-0.33	0.63
DAYS_SINCE_LASTORDER	-0.25	-0.02	-0.40	0.05	-0.33	1.00	-0.52
MSRP	-0.01	0.02	0.78	-0.02	0.63	-0.52	1.00

EXPLORATORY DATA ANALYSIS

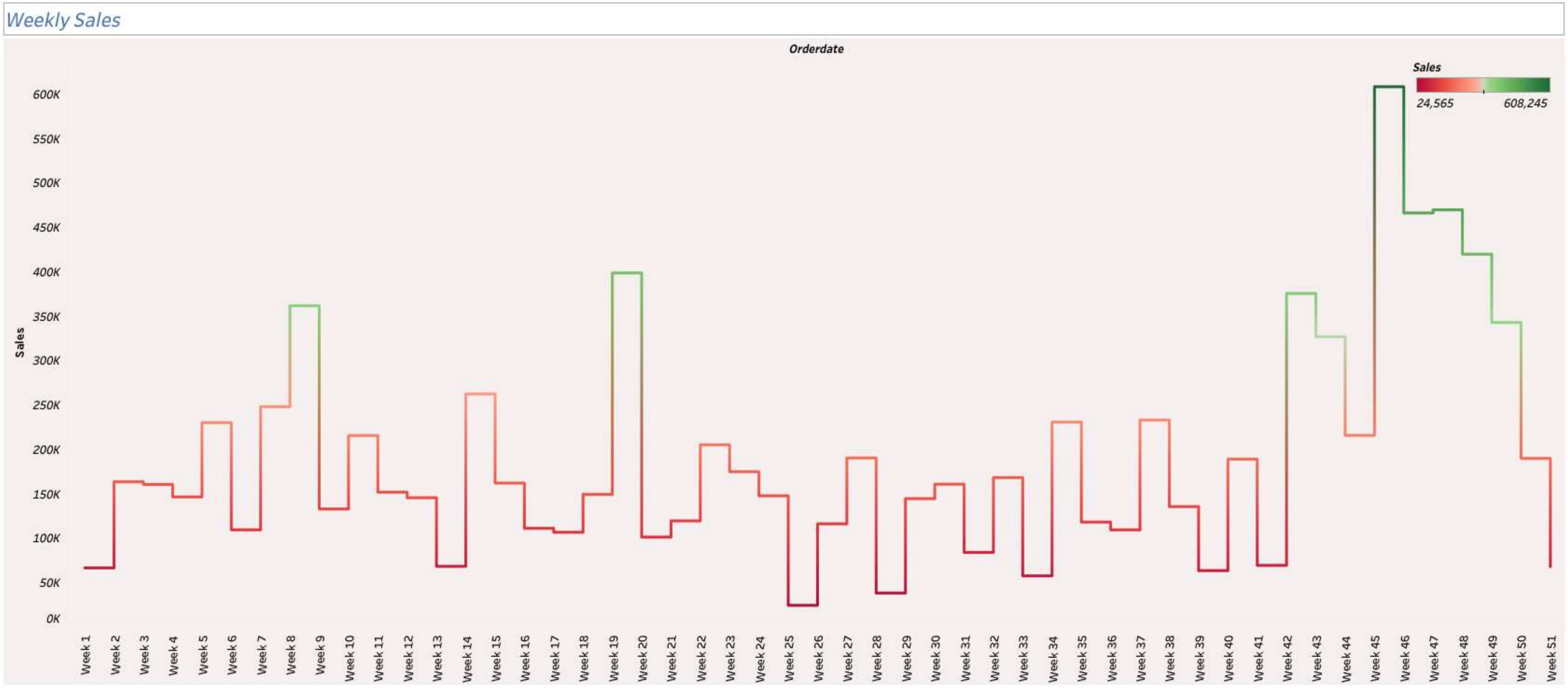
Multivariate Analysis: Pair Plot

- ***We see that medium order size have a lead on number of orders and in line orders, whereas, all other parameters Small size order have the lean.***



SALES TRENDS

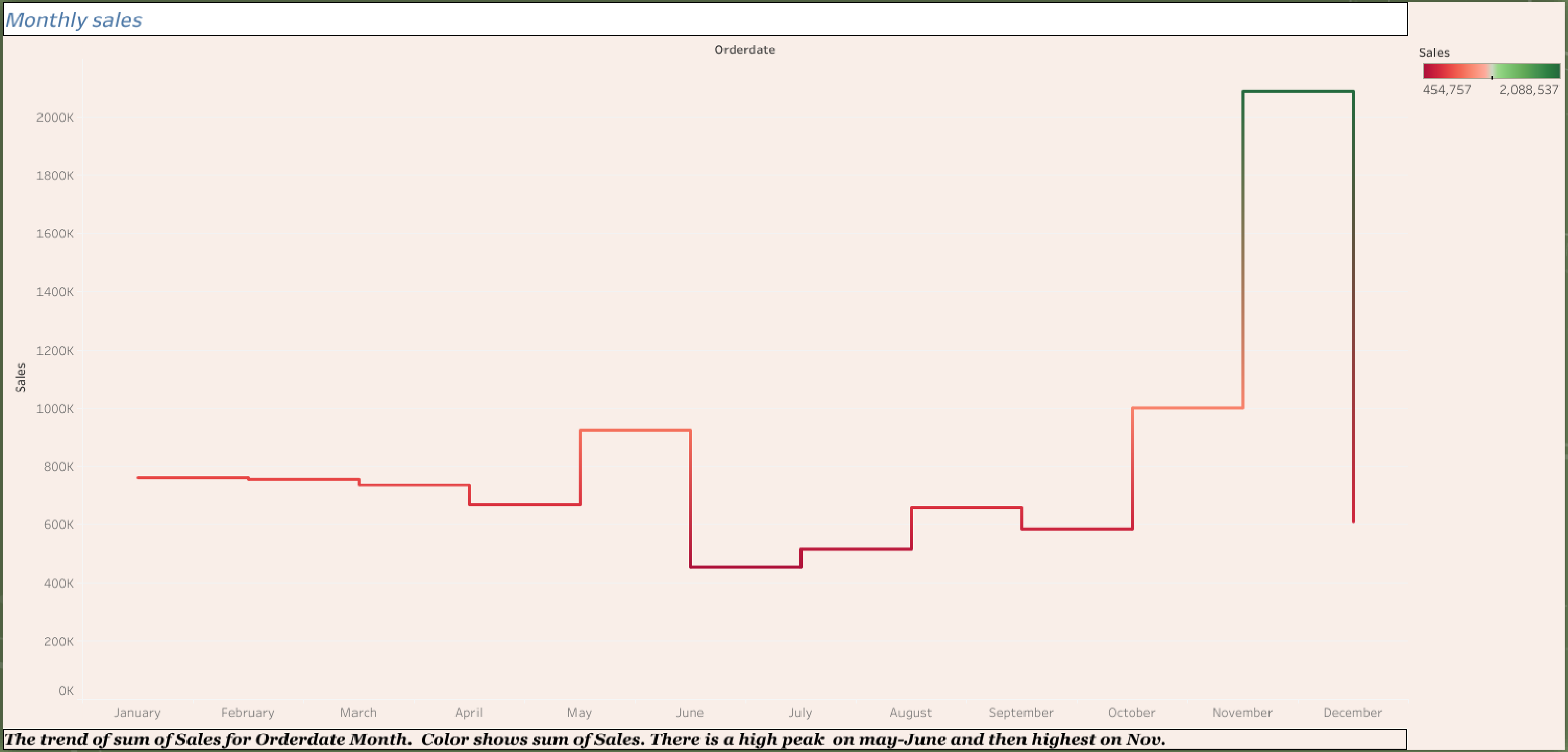
- **Weekly:**



The trend of sum of Sales for Orderdate Week. Color shows sum of Sales. There is a low-rise trend weekly, also the highest sale are in Week 46.

SALES TRENDS

- **Monthly:**

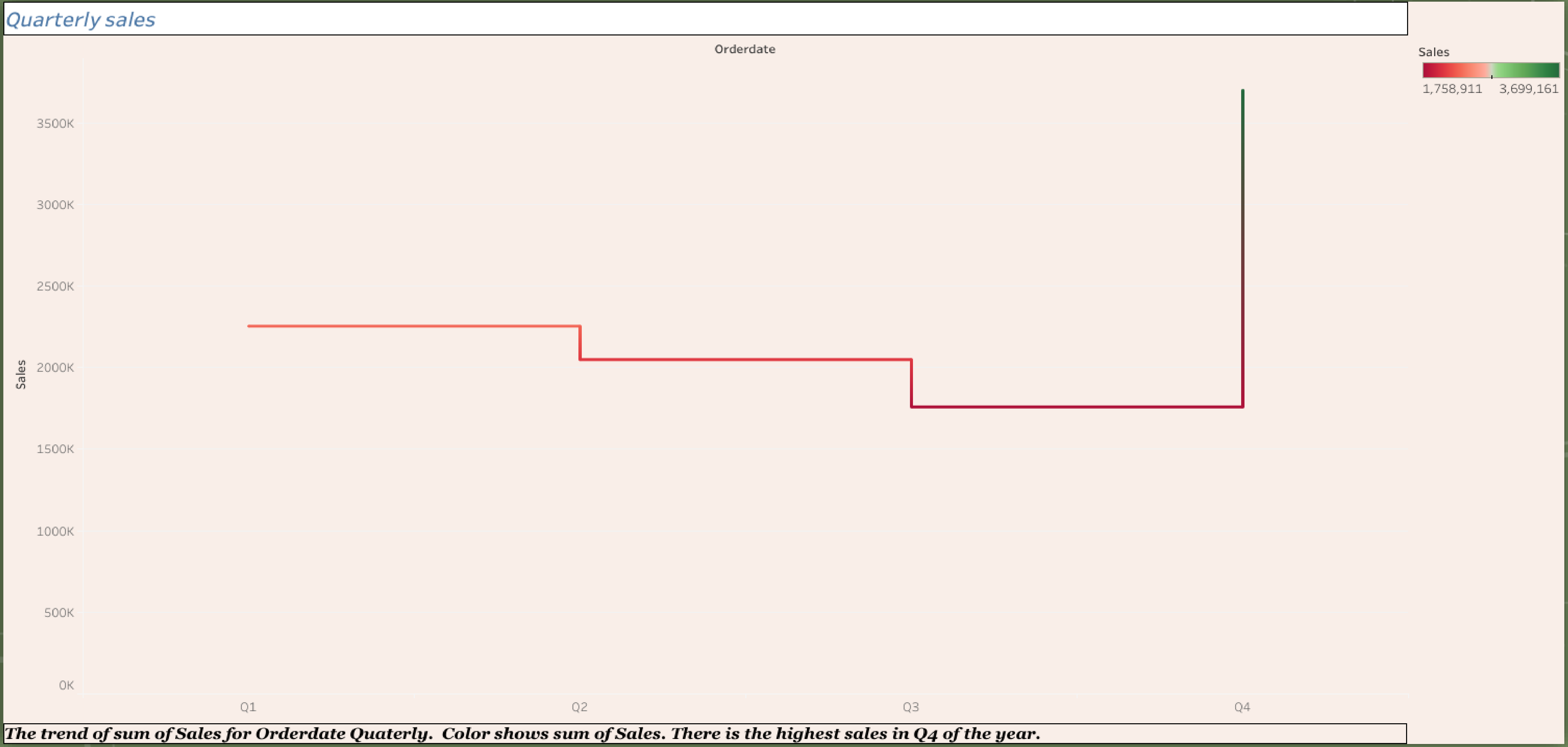


Monthly sales

The trend of sum of Sales for Orderdate Month. Color shows sum of Sales. There is a high peak on may-June and then highest on Nov.

SALES TRENDS

- ***Quarterly*:**



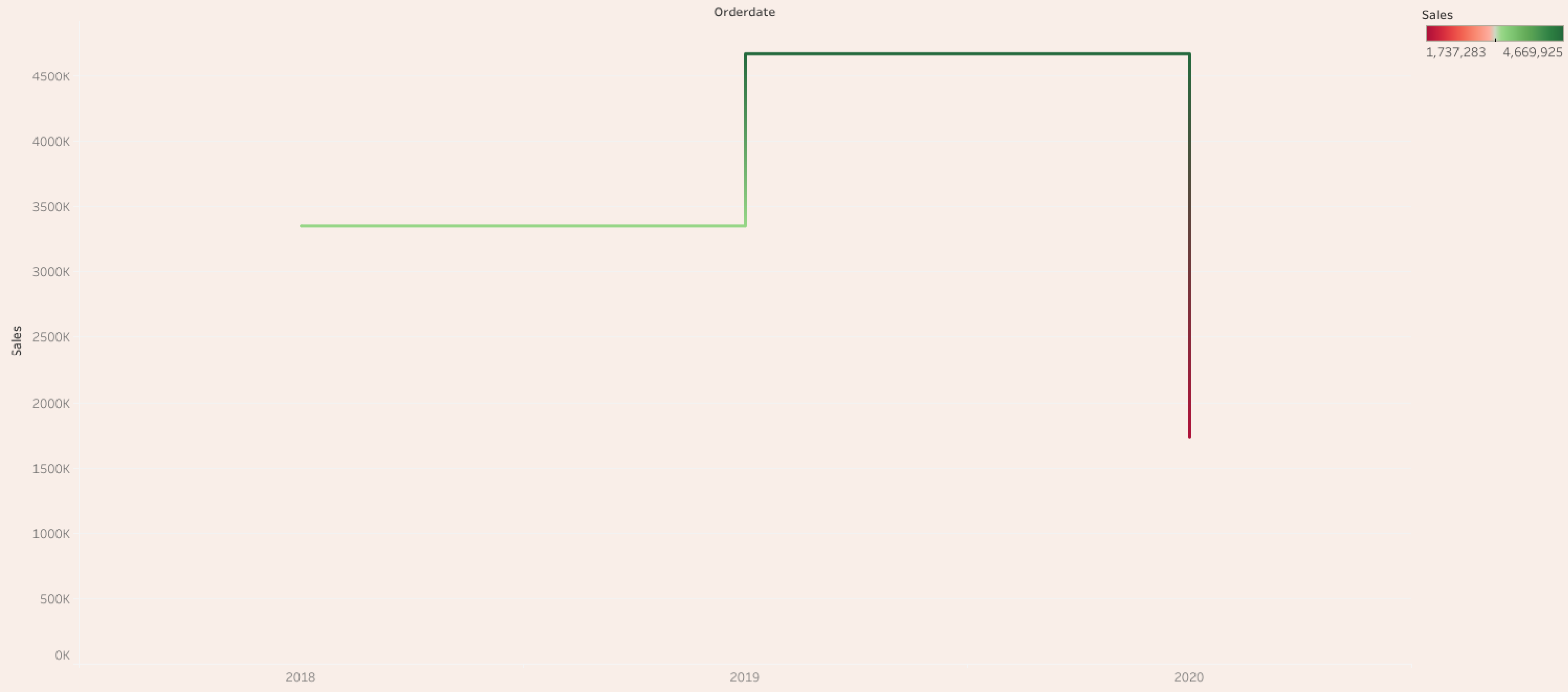
Quarterly sales

The trend of sum of Sales for Orderdate Quaterly. Color shows sum of Sales. There is the highest sales in Q4 of the year.

SALES TRENDS

- ***Yearly***:

Yearly sales



The trend of sum of Sales for Orderdate Yearly. Color shows sum of Sales. 2019>2018>2020

EDA AND SALES TRENDS

- **Interpretation:**
  - ***There are received for Classic cars parts, the least from for trains parts.***
  - ***The orders of medium size are received mostly.***
  - ***The company receives orders from different locations of Europe in comparison to other zones.***
  - ***The country that gives most sales to the company is USA.***
  - ***The order quantity and price rise can be seen in 2019, when 3 years are compared.***
  - ***The sales and MSRP is highly corelated to price of each item*.**
  - ***The MSRP average is in close range to the item price average.***
  - ***Currently small size order are leading with highest sales value.***
  - ***Most order are shipped, only few orders get disputed.***
  - ***Rise is sales are seen in year end, some event, discount or feast that would be the reason for the hike.***

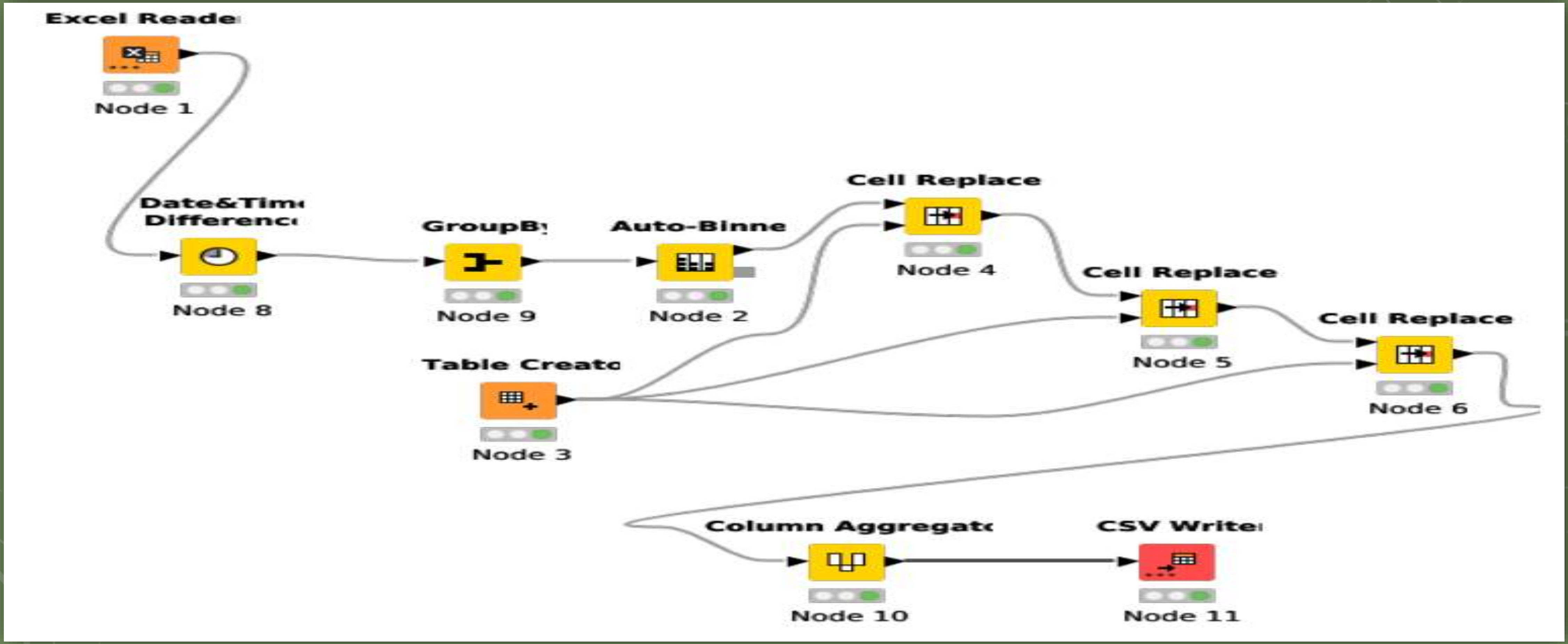
RFM

v *What is RFM?*

- Ø *Recency, frequency, monetary value is a marketing analysis tool used to identify a company's or an organization's best customers by measuring and analysing spending habits.*
- Ø *Recency: How much time has elapsed since a customer’s last activity or transaction with the company.*
- Ø *Frequency: How often has a customer transacted or interacted with the brand during a particular period of time.*
- Ø *Monetary: Also referred to as “monetary value,” this factor reflects how much a customer has spent with the brand during a particular period of time.*
- ➢ *How recently they've made a purchase, how often they buy, and the size of their purchases.*
- ➢ *These are the parameter of customer's behaviour we focus on during RFM Analysis.*

RFM

- *I have used Python and Tableau for data read and EDA*
- *In this project I have used, KNIME is used to perform the RFM Analysis and here is the workflow diagram:*



RFM OUTPUT

- **RFM: output: Head (5 rows × 28 columns)**

CUSTOMERNAME	ORDERNUMBER	QUANTITYORDERED	PRICEEACH	ORDERLINENUMBER	SALES	ORDERDATE	DAYS_SINCE_LASTORDER	STATUS	PRODUCTLINE	MSRP	PRODUCTCODE	PHONE	ADDRESSLINE1	CITY	POSTALCODE	COUNTRY	CONTACTLASTNAME	CONTACTFIRSTNAME	DEALSIZE	Recency	ORDERNUMBER [Binned]	SALES [Binned]	Recency [Binned]	RECENCY	FREQUENCY	MONETARY	Concatenate
AV Stores, Co.	51	34.8627450980392	91.0845098039215	9.0196078431372 5	3094.27078431372	51	1803.80392156863	51	51	92.843137254902	51	51	51	Manchester	51	UK	Ashworth	Victoria	51	197	Bin 4	Bin 1	Bin 3	2	4	1	2, 4, 1
Alpha Cognac	20	34.35	101.16	4.95	3524.422	20	2236.2	20	20	97.15	20	20	20	Toulouse	20	France	Roulet	Annette	20	65	Bin 1	Bin 2	Bin 1	4	1	2	4, 1, 2
Amica Models & Co.	26	32.4230769230769	110.852692307692	7.61538461538462	3619.89461538461	26	1318.6153846153 8	26	26	107.653846153846	26	26	26	Torino	26	Italy	Accorti	Paolo	26	266	Bin 2	Bin 3	Bin 4	1	2	3	1, 2, 3
Anna's Decorations, Ltd	46	31.9347826086956	106.424130434783	6.43478260869565	3347.7419565217 4	46	1463.58695652174	46	46	104.71739130434 8	46	46	46	North Sydney	46	Australia	O'Hara	Anna	46	84	Bin 4	Bin 2	Bin 2	3	4	2	3, 4, 2
Atelier graphique	7	38.57142857142 86	92.23857142857 14	2	3454.28	7	1424.4285714285 7	7	7	95.57142857142 86	7	7	7	Nantes	7	France	Schmitt	Carine	7	189	Bin 1	Bin 2	Bin 3	2	1	2	2, 1, 2

RFM OUTPUT

- *Who are your loyal customers?*
- *Once with high RFM, below 5 are the best customer. High Recency, frequency, monetary*

Concatenate	CUSTOMERNAME
4, 4, 3	L'ordine Souveniers
4, 4, 3	Mini Gifts Distributors Ltd.
4, 4, 3	Salzburg Collectables
4, 4, 4	Danish Wholesale Imports
4, 4, 4	The Sharp Gifts Warehouse

RFM OUTPUT

- *Who are your lost customers?*
- *Least value for Recency, frequency, monetary*

Concatenate	CUSTOMERNAME
1, 1, 1	Double Decker Gift Stores, Ltd
1, 1, 1	Cambridge Collectables Co.
1, 1, 1	Bavarian Collectables Imports, Co.
1, 1, 2	Osaka Souveniers Co.
1, 1, 2	Daedalus Designs Imports

RFM OUTPUT

- ***Who are your best customers?***
- ***These customer have large sale order and have good frequency.***

Concatenate	*CUSTOMERNAME*
3,4,3	***Australian Collectors, Co.***
3,4,4	***Muscle Machine Inc***
3,2,3	***FunGiftIdeas.com***
3,3,3	***Suominen Souveniers***
3,4,4	***Dragon Souveniers, Ltd.***

RFM OUTPUT

- ***Which customers are on the verge of churning?***
- ***These customer have large sale orders in the past, though their current recency is low, some have frequency low too.***

Concatenate	*CUSTOMERNAME*
2,2,4	***Blauer See Auto, Co.***
1,1,4	***CAF Imports***
2,1,4	***Classic Legends Inc.***
1,3,4	***Herkku Gifts***
2,4,4	***Online Diecast Creations Co.***

INTERPRETATION

- ***There are received for Classic cars parts, the least from for trains parts.***
- ***The orders of medium size are received mostly.***
- ***The company receives orders from different locations of Europe in comparison to other zones.***
- ***Sending reminder/long time promotional email for customer with less Recency, will be useful.***
- ***The order quantity and price rise can be seen in 2019, when 3 years are compared.***
- ***The sales and MSRP is highly corelated to price of each item, can push for bundle promotion for large order and loyal customer.***
- ***The MSRP average is in close range to the item price average.***
- ***Rise is sales are seen in year end, some event, discount or feast that would be the reason for the hike.***
- ***Offering sales discount at the start of the year will be helpful.***
- ***Providing more discounts for holiday and large order promotion will benefit the company.***
- ***Focus is needed to push more B2B sales for getting more larger order.***
- ***Company can post reviews from their major sales clients.***