MRA PROJECT ML 1

Boby Sinha

PGP DSBA JAN_A 22 BATCH

TABLE OF CONTENT

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.

ORDERNUMBER :	Order Number	CUSTOMER NAME:	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

DATA SUMMARY

- An automobile parts manufacturing company has collected data of transactions for 3 years.
- The data has 2747 entries (0 To 2746) of rows and 20 columns. The data has 1 datetime64 , 2 float64, 5 int64, and 12 Object data types.
- The dataset contains customer geographical information and transaction history.

```
The Sales_Data has 2747 rows and 20 columns
```

```
The size of Sales_Data 54940
```

```
<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
```

DESCRIPTION OF THE DATA

- 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.

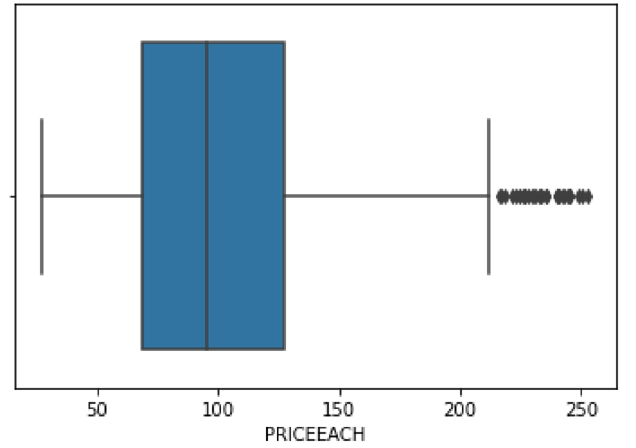
- Manufacturer's Suggested Retail Price(MSRP) for each product code is decided but we found that this is not matching with Price of Each item & is inconsistent with MSRP.

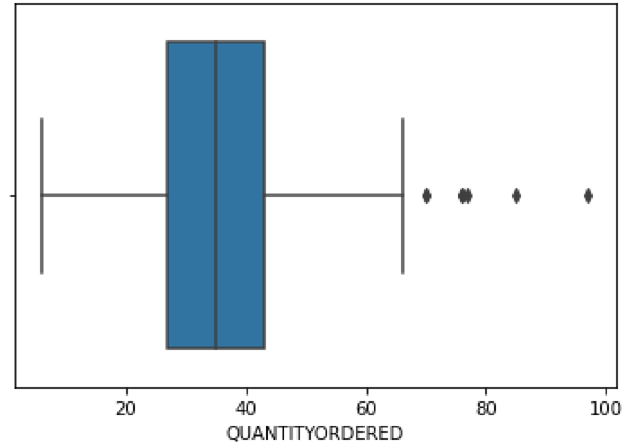
	count	unique	top	freq	first	last	mean	std	min	25%	50%	75%	max
ORDERNUMBER	2747	NaN	NaN	NaN	NaT	NaT	10259.8	91.8775	10100	10181	10264	10334.5	10425
QUANTITYORDERED	2747	NaN	NaN	NaN	NaT	NaT	35.103	9.76214	6	27	35	43	97
PRICEEACH	2747	NaN	NaN	NaN	NaT	NaT	101.099	42.0425	26.88	68.745	95.55	127.1	252.87
ORDERLINENUMBER	2747	NaN	NaN	NaN	NaT	NaT	6.49108	4.23054	1	3	6	9	18
SALES	2747	NaN	NaN	NaN	NaT	NaT	3553.05	1838.95	482.13	2204.35	3184.8	4503.09	14082.8
ORDERDATE	2747	246	2018-11-14 00:00:00	38	2018-01-06	2020-05-31	NaN	NaN	NaN	NaN	NaN	NaN	NaN
DAYS_SINCE_LASTORDER	2747	NaN	NaN	NaN	NaT	NaT	1757.09	819.281	42	1077	1761	2436.5	3562
STATUS	2747	6	Shipped	2541	NaT	NaT	NaN	NaN	NaN	NaN	NaN	NaN	NaN
PRODUCTLINE	2747	7	Classic Cars	949	NaT	NaT	NaN	NaN	NaN	NaN	NaN	NaN	NaN
MSRP	2747	NaN	NaN	NaN	NaT	NaT	100.692	40.1148	33	68	99	124	214
PRODUCTCODE	2747	109	S18_3232	51	NaT	NaT	NaN	NaN	NaN	NaN	NaN	NaN	NaN
CUSTOMERNAME	2747	89	Euro Shopping Channel	259	NaT	NaT	NaN	NaN	NaN	NaN	NaN	NaN	NaN
PHONE	2747	88	(91) 555 94 44	259	NaT	NaT	NaN	NaN	NaN	NaN	NaN	NaN	NaN
ADDRESSLINE1	2747	89	C/ Moralzarzal, 86	259	NaT	NaT	NaN	NaN	NaN	NaN	NaN	NaN	NaN
CITY	2747	71	Madrid	304	NaT	NaT	NaN	NaN	NaN	NaN	NaN	NaN	NaN
POSTALCODE	2747	73	28034	259	NaT	NaT	NaN	NaN	NaN	NaN	NaN	NaN	NaN
COUNTRY	2747	19	USA	928	NaT	NaT	NaN	NaN	NaN	NaN	NaN	NaN	NaN
CONTACTLASTNAME	2747	76	Freyre	259	NaT	NaT	NaN	NaN	NaN	NaN	NaN	NaN	NaN
CONTACTFIRSTNAME	2747	72	Diego	259	NaT	NaT	NaN	NaN	NaN	NaN	NaN	NaN	NaN
DEALSIZE	2747	3	Medium	1349	NaT	NaT	NaN	NaN	NaN	NaN	NaN	NaN	NaN

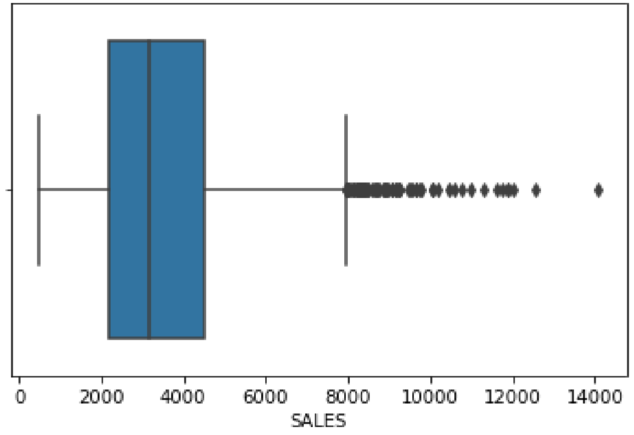
- No missing values in the dataset

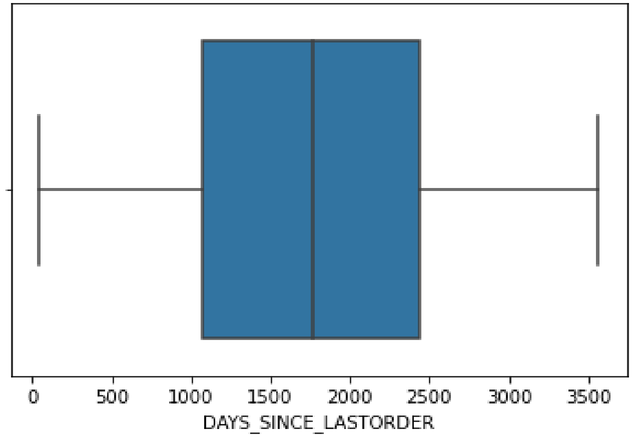
```
ORDERNUMBER             0
QUANTITYORDERED         0
PRICEEACH               0
ORDERLINENUMBER         0
SALES                   0
ORDERDATE               0
DAYS_SINCE_LASTORDER    0
STATUS                  0
PRODUCTLINE             0
MSRP                    0
PRODUCTCODE             0
CUSTOMERNAME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

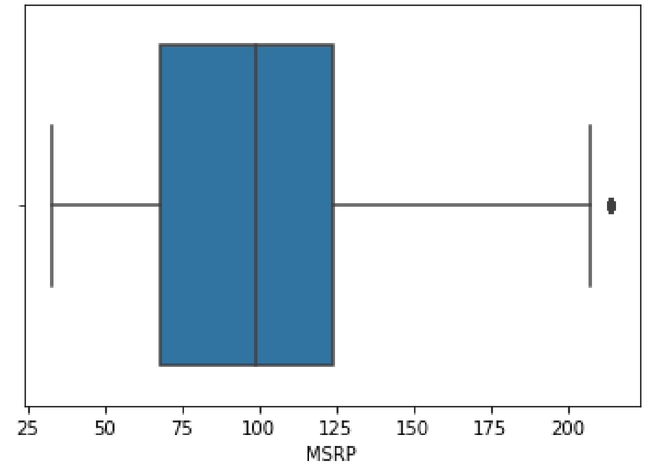
EDA – UNIVARIATE ANALYSIS

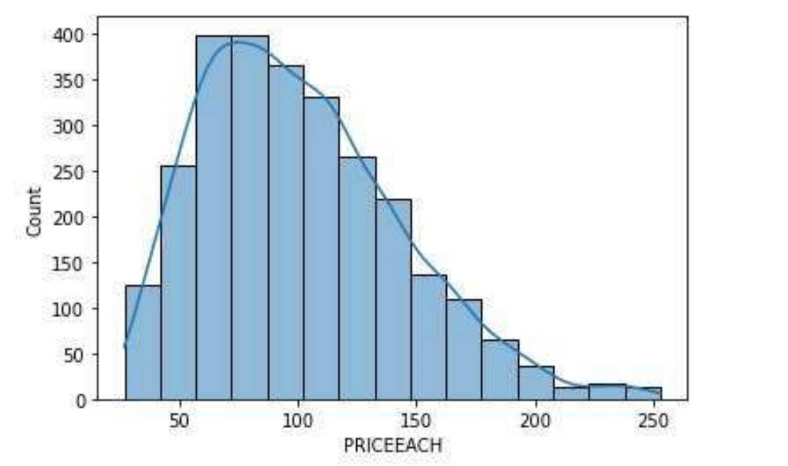


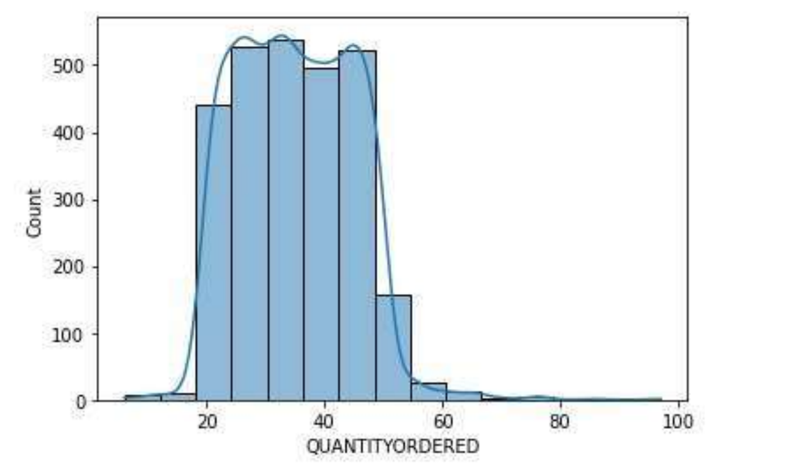


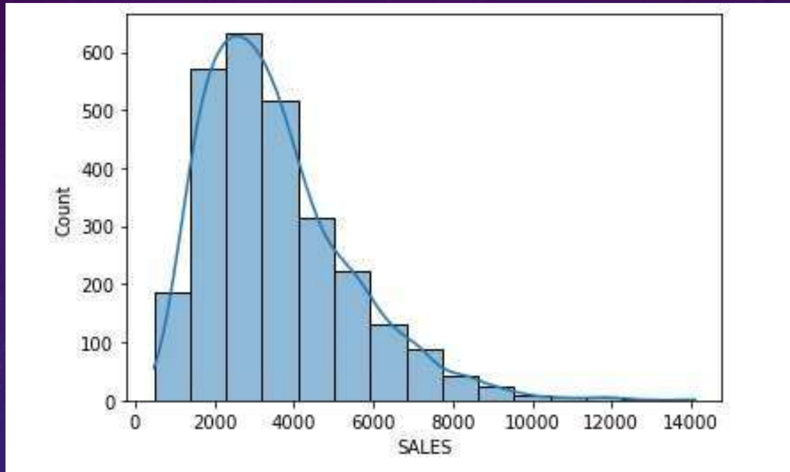


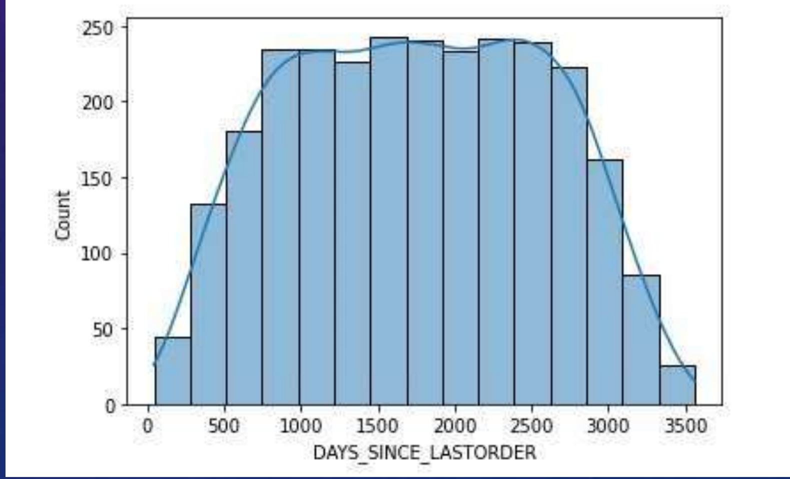


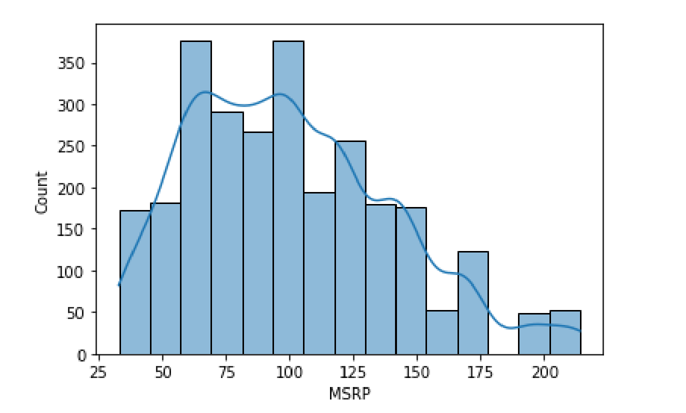


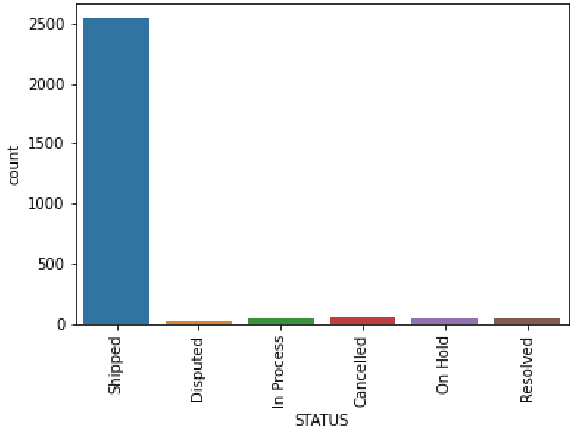

Count
PRICEEACH

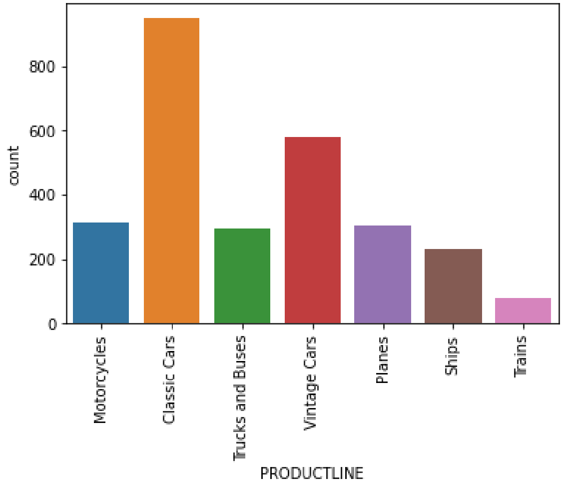

Count
QUANTITYORDERED

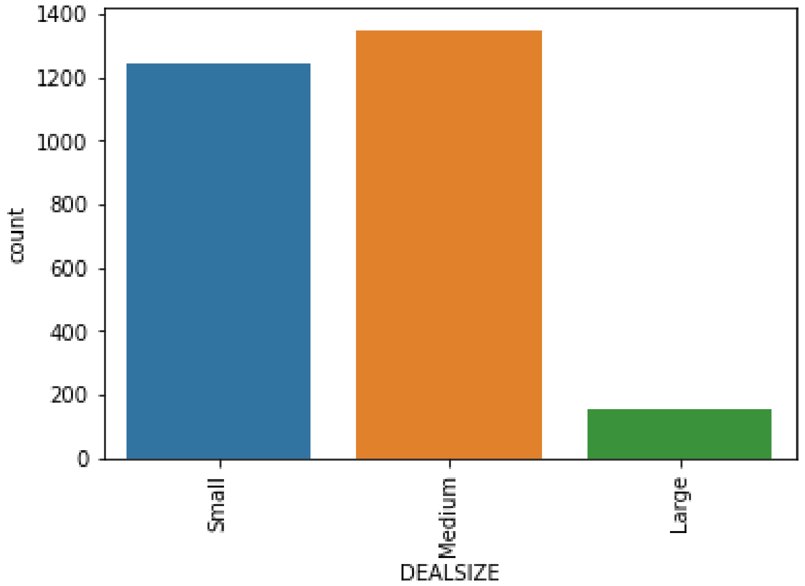

Count
SALES


Count
DAYS_SINCE_LASTORDER

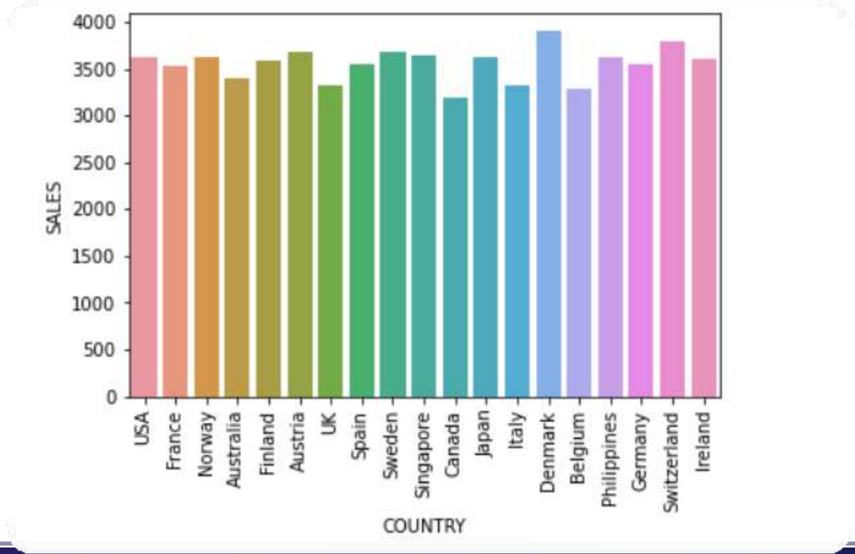

Count
MSRP

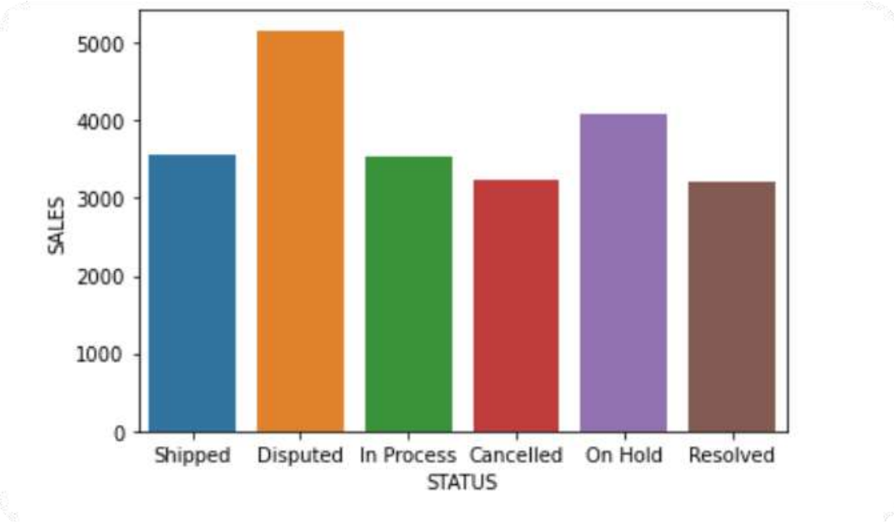

count
2500
2000
1500
1000
500
0
Shipped
Disputed
In Process
Cancelled
On Hold
Resolved
STATUS

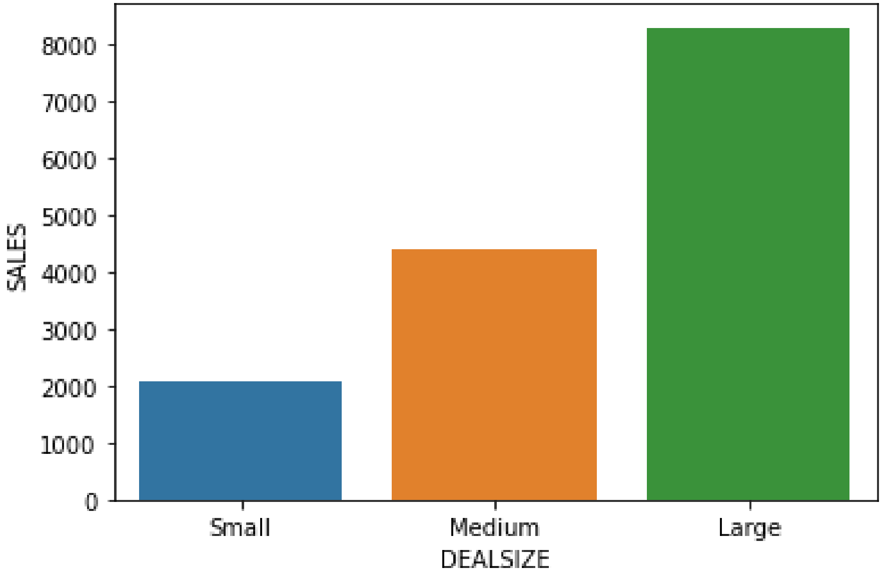

count
800
600
400
200
0
Motorcycles
Classic Cars
Trucks and Buses
Vintage Cars
Planes
Ships
Trains
PRODUCTLINE


count
1400
1200
1000
800
600
400
200
0
Small
Medium
Large
DEALSIZE

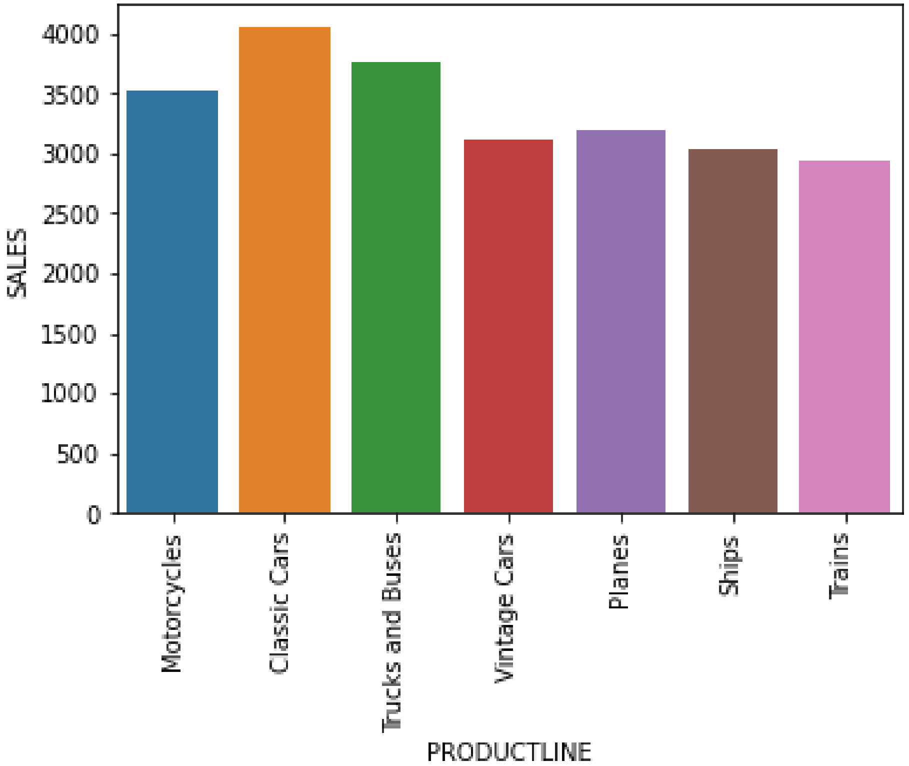
EDA – BIVARIATE ANALYSIS



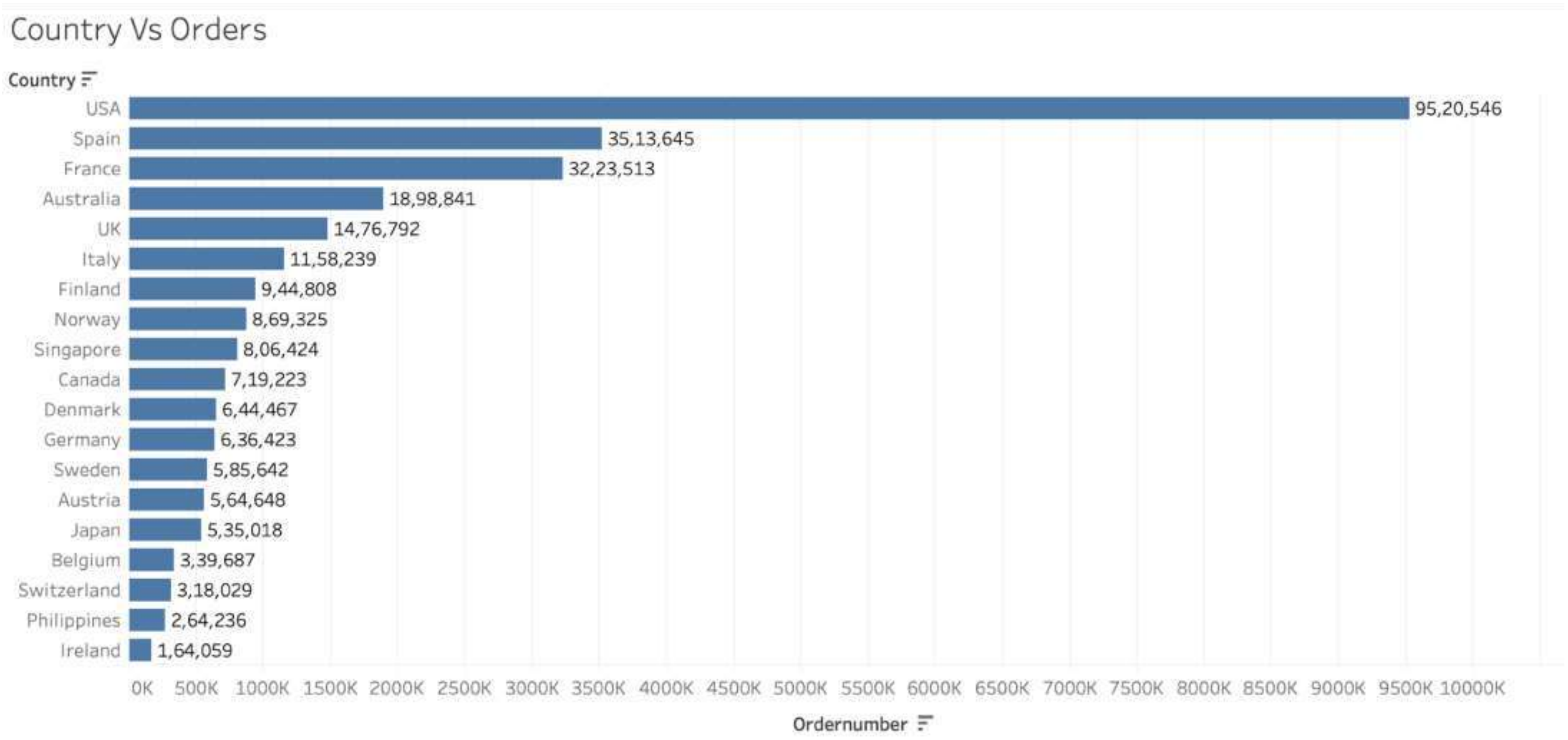




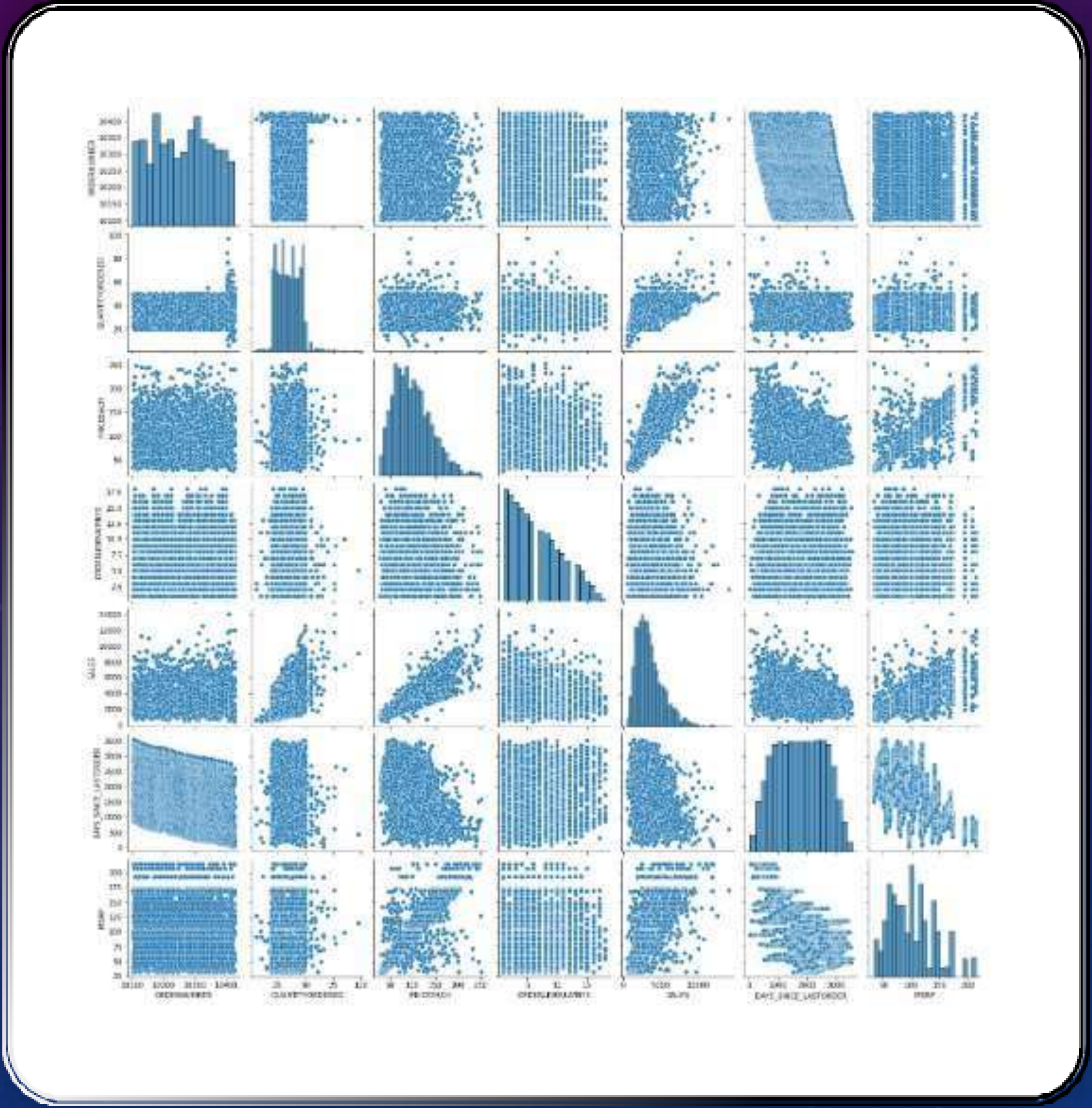
8000
7000
6000
5000
4000
3000
2000
1000
0
SALES
Small
Medium
Large
DEALSIZE

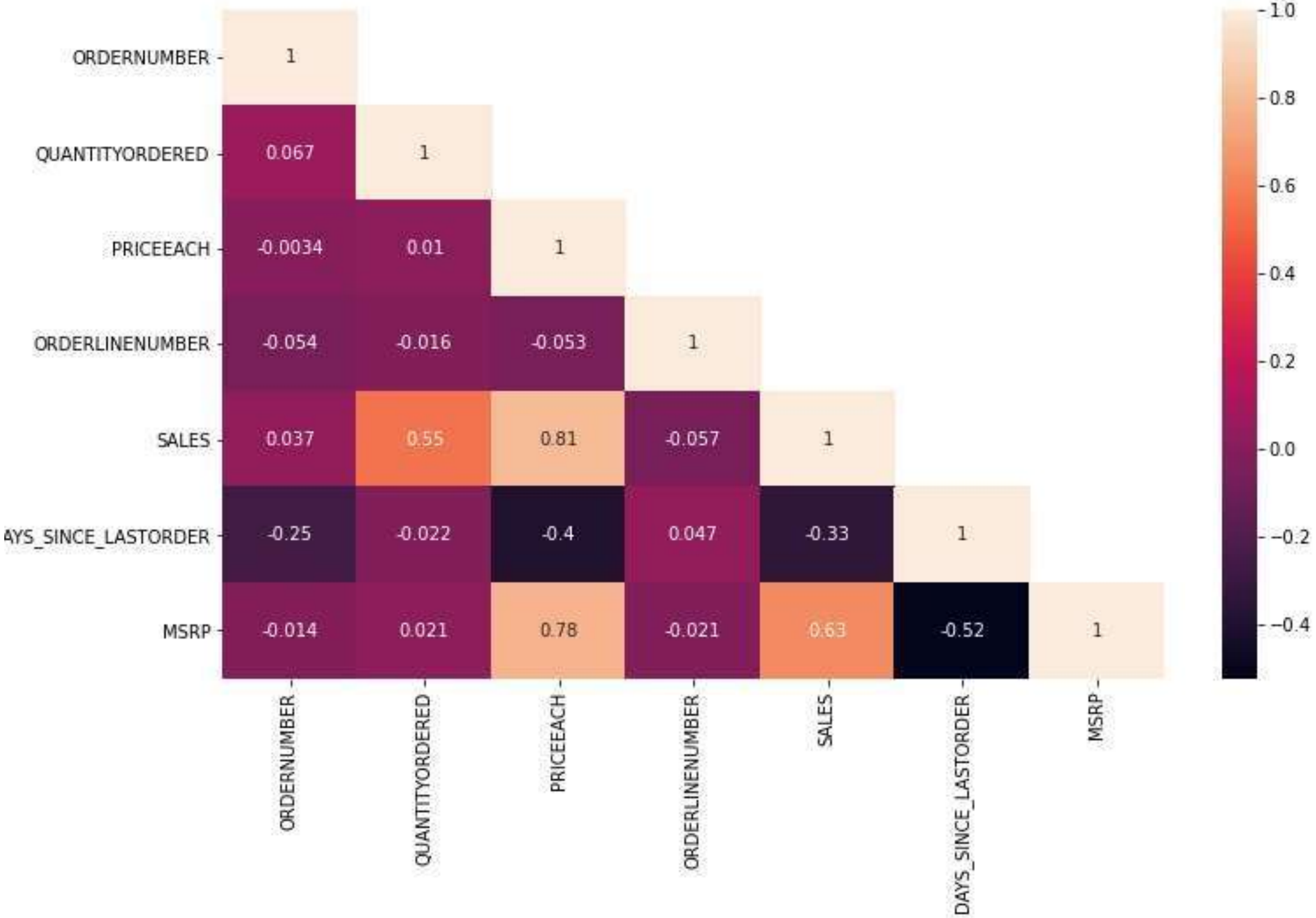


4000
3500
3000
2500
2000
1500
1000
500
0
SALES
Motorcycles
Classic Cars
Trucks and Buses
Vintage Cars
Planes
Ships
Trains
PRODUCTLINE


Country Vs Orders
Country
USA
Spain
France
Australia
UK
Italy
Finland
Norway
Singapore
Canada
Denmark
Germany
Sweden
Austria
Japan
Belgium
Switzerland
Philippines
Ireland
95,20,546
35,13,645
32,23,513
18,98,841
14,76,792
11,58,239
9,44,808
8,69,325
8,06,424
7,19,223
6,44,467
6,36,423
5,85,642
5,64,648
5,35,018
3,39,687
3,18,029
2,64,236
1,64,059
0K 500K 1000K 1500K 2000K 2500K 3000K 3500K 4000K 4500K 5000K 5500K 6000K 6500K 7000K 7500K 8000K 8500K 9000K 9500K 10000K
Ordernumber

EDA –MULTIVARIATE ANALYSIS




ORDERNUMBER
QUANTITYORDERED
PRICEEACH
ORDERLINENUMBER
SALES
AYS_SINCE_LASTORDER
MSRP
1
0.067
1
-0.0034
0.01
1
-0.054
-0.016
-0.053
1
0.037
0.55
0.81
-0.057
1
-0.25
-0.022
-0.4
0.047
-0.33
1
-0.014
0.021
0.78
-0.021
0.63
-0.52
1
ORDERNUMBER
QUANTITYORDERED
PRICEEACH
ORDERLINENUMBER
SALES
DAYS_SINCE_LASTORDER
MSRP
1.0
0.8
0.6
0.4
0.2
0.0
−0.2
−0.4

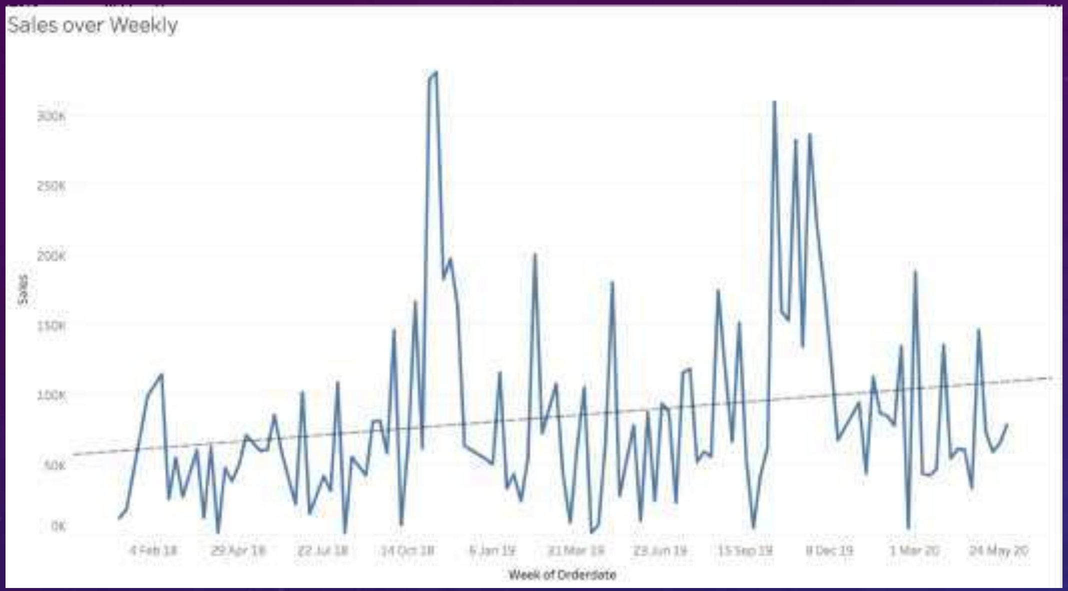
EDA SUMMARY [INFERENCES]

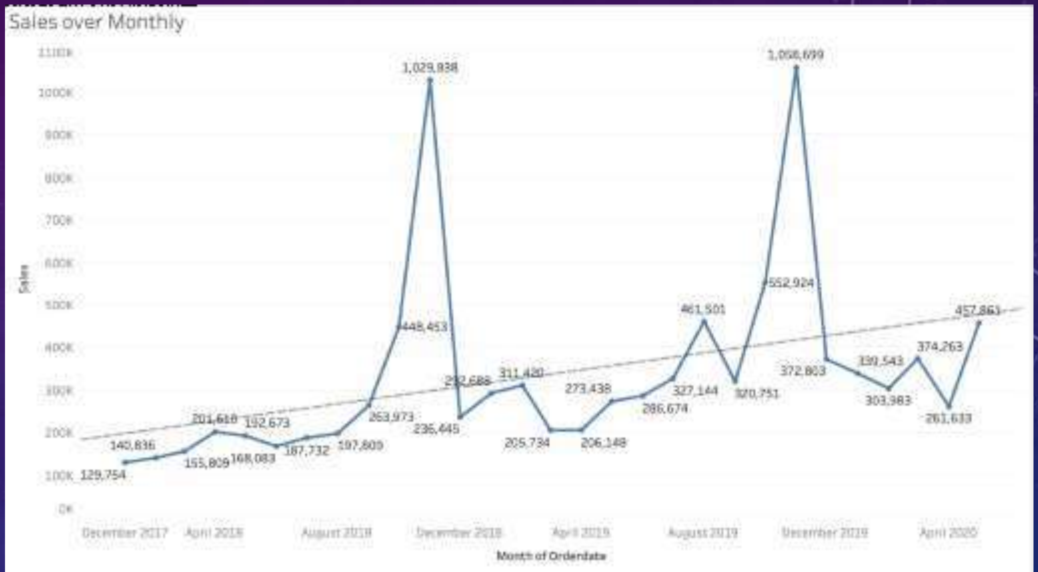
- We can clearly see that outliers are present.
- Most of the data is normally distributed. Histogram of sales is right skewed.
- We have noticed that the sales of classic cars products are high followed by vintage car product sales.
- The number of medium deal size seems to be higher than small and large.
- we can see the larger portion of classic cars followed by vintage cars were as trains has the least demand.
- The sales of the Disputed Status is high.

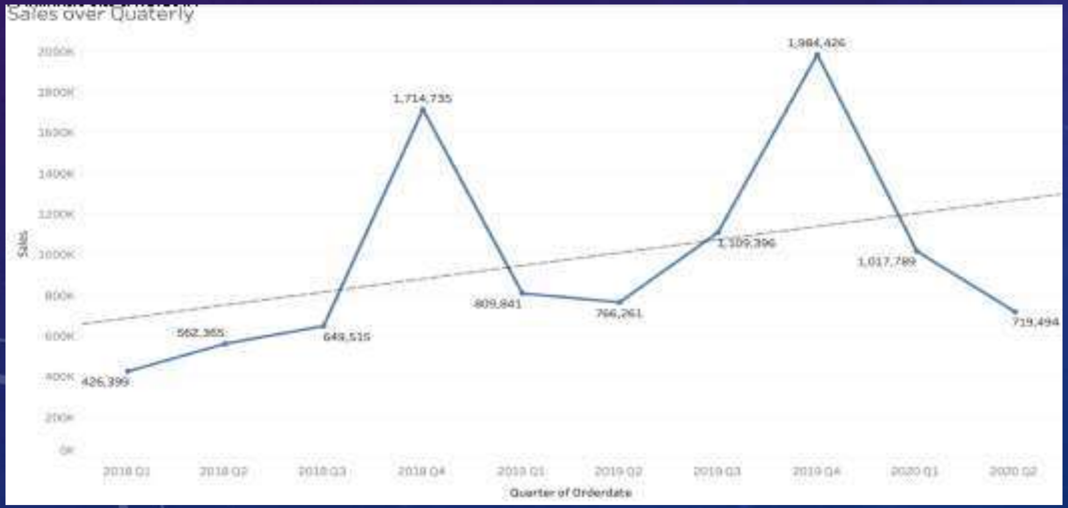
EDA SUMMARY [INFERENCES] – CONTD…

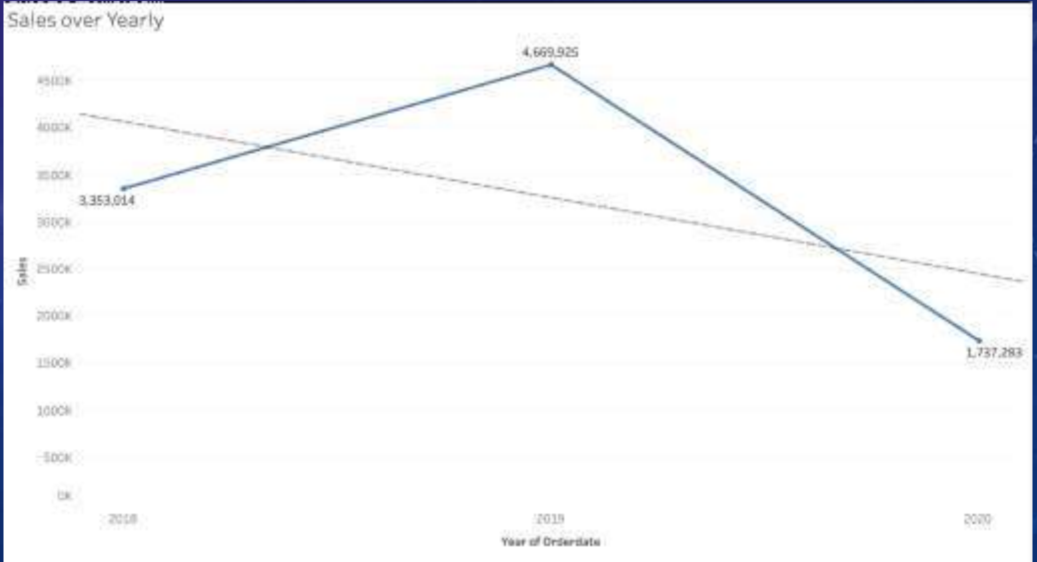
- USA has high number of orders following by Spain and France.
- PRICEEACH is highly correlated with MSRP.
- SALES is highly correlated to PRICEEACH.
- As sales are high for classic cars the company has even sold below MSRP, there might be a chances that the company has given more discounts to its customers and vice versa for vintage cars were the company has sold above MSRP.
- Ship, vintage car & train are been sold above the MSRP. By looking at the given data almost all the transactions are been shipped.

TIME SERIES FORECASTING





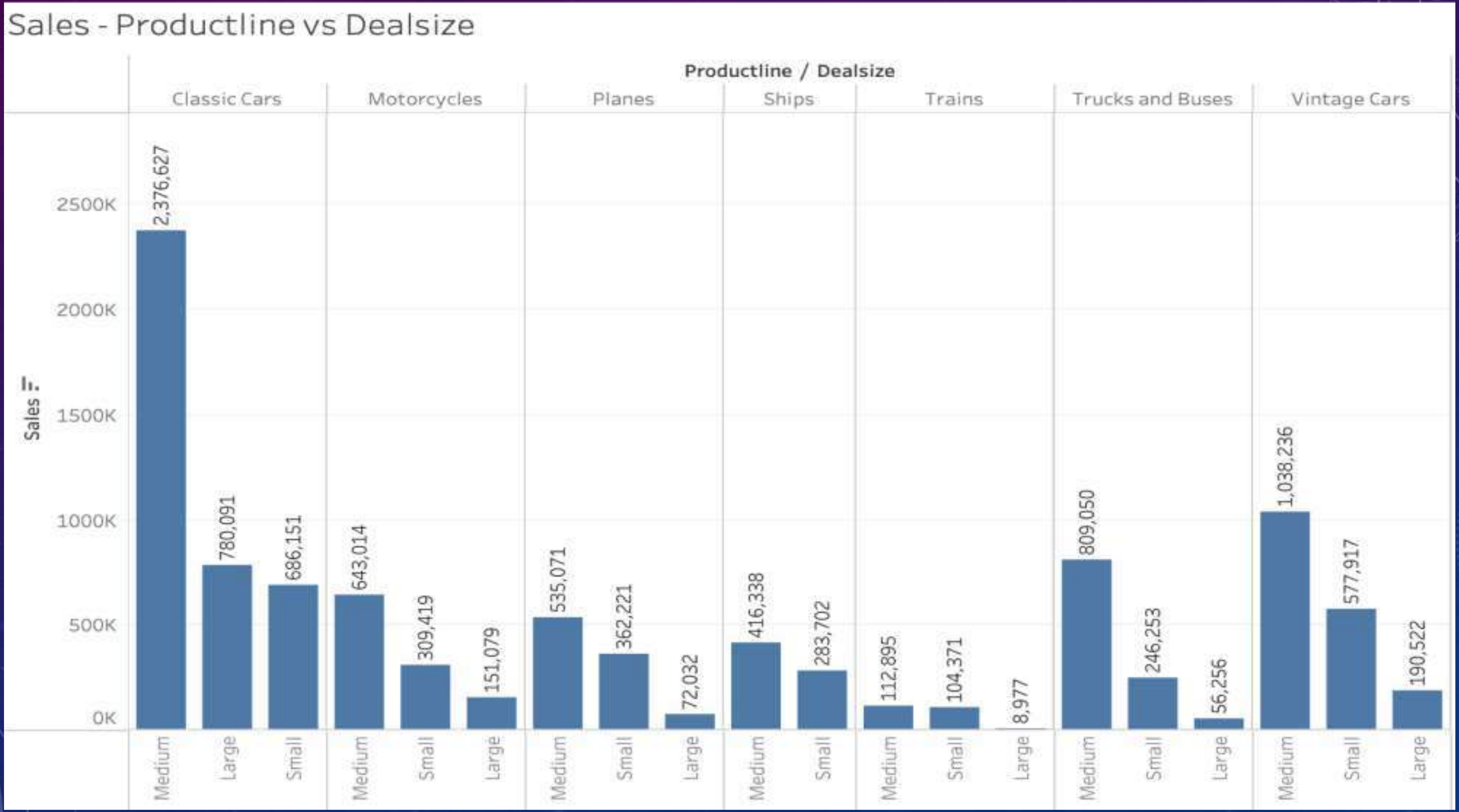




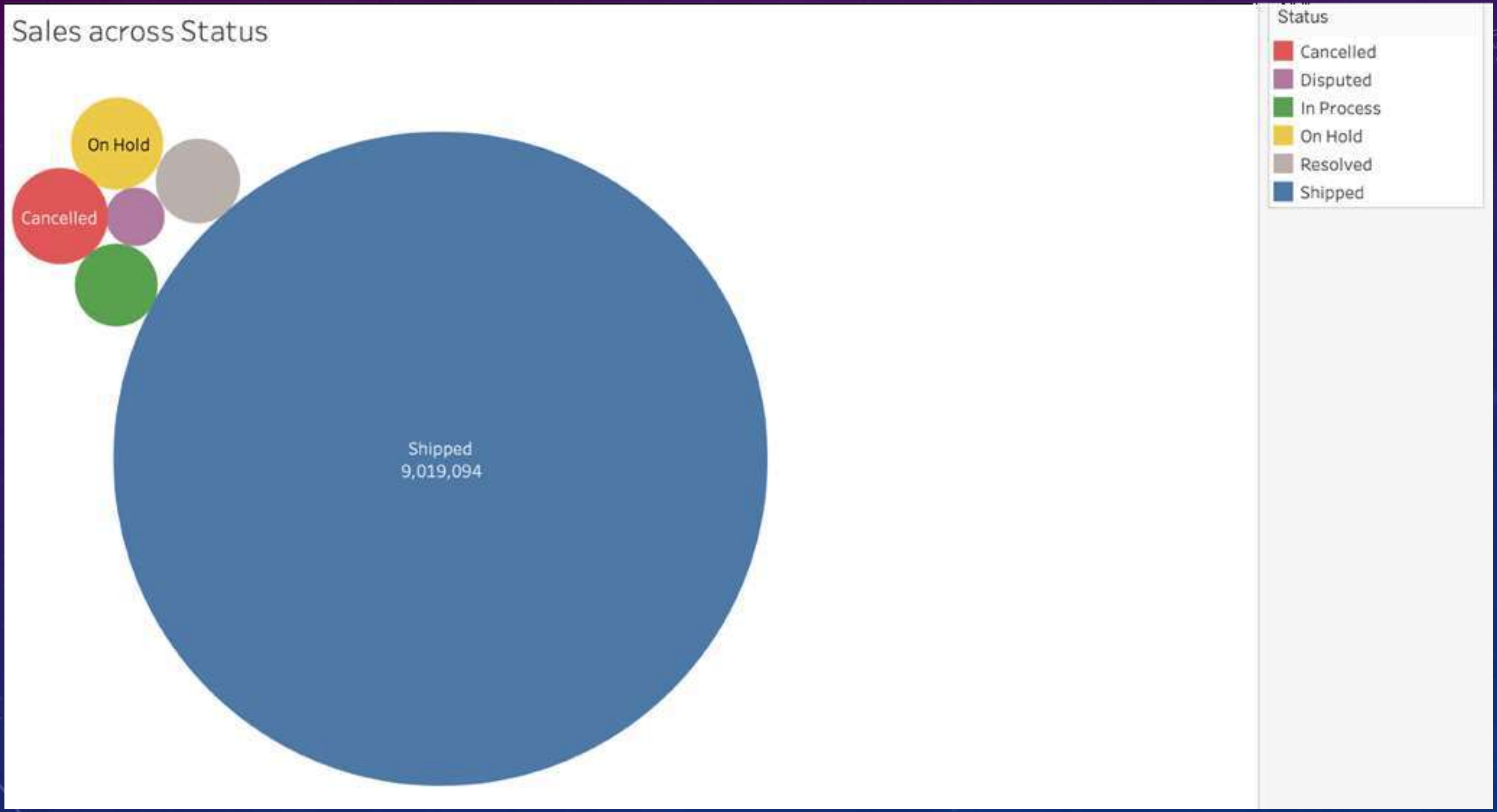
TIME SERIES FORECASTING

- 2020 has highest sales in first quarter.
- Fourth quarter of 2018 and 2019 has highest sales when compared to other quarters.
- For the rest of the quarters, sales is low and is on an increasing and decreasing trend.
- 2019 has highest sales when compared to 2018 and 2020.
- Weekly and Monthly sales follows seasonality.

SALES ACROSS DIFFERENT CATEGORIES

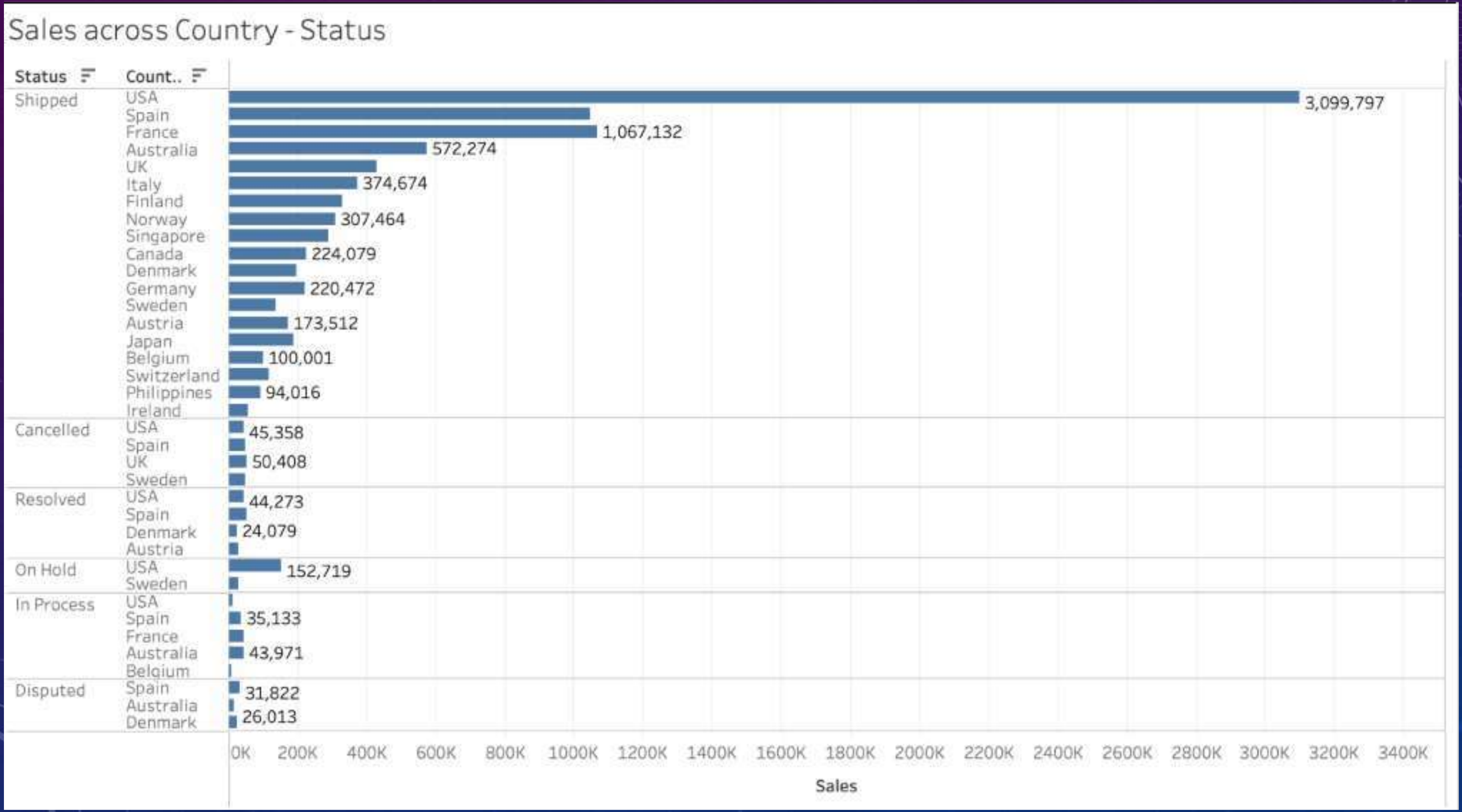


SALES ACROSS DIFFERENT CATEGORIES



SALES ACROSS DIFFERENT CATEGORIES

SALES ACROSS DIFFERENT CATEGORIES



- From the Time series forecast, it is evident 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.

- Countries with least sales, mega offers with low EMI facilities can be incorporated to promote the sales.

- Classic cars have majority sales whereas the sales on trucks and buses can be expanded.

- Cancelled orders must be considered and validate the cancellation reasons thoroughly.

- The large size deals are the lowest and almost stagnant. Steps should be taken to promote and attract the customers to buy more of large size deals.

- Pending disputed orders should be resolved at the earliest so it doesn’t give a negative feel to the customer.

RFM ANALYSIS

KNIME is used here for RFM analysis.

- What all parameters used, and assumptions made?

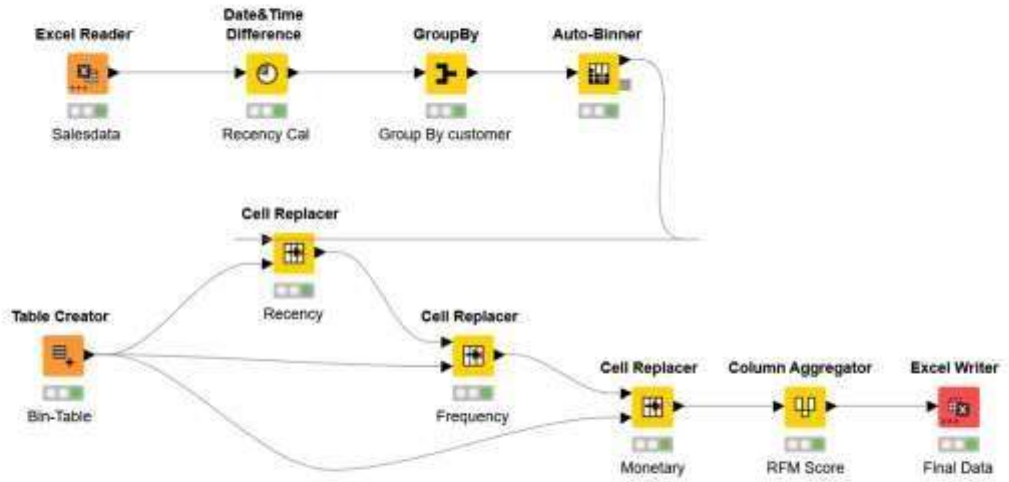
(a) created new column name “**Recency**” as "[Max(order date) - order date)]“ assumed “**01-06-2020**“ as a reference date.

(b) Order number has been repetitive for different products. So, count of each order number has been considered as “**Frequency”** of an order number.

(c) Aggregation : Sum of Sales has been considered as “**Monetary”.**

- Created four bins for Recency, frequency & Monetary using percentile range(0,0.25,0.50,0.75,0.1)
- Based on above 4 bins, 4 segments like High (H) , Medium (M) , Low (L) and Churn (C) are considered.

KNIME WORKFLOW



OUTPUT TABLE HEAD

Table "default" - Rows: 89 Spec - Columns: 17 Properties Flow Variables

Row ID	S CUSTO...	S Recency	S Freque...	S Monetary	S RFM Sc...	I ORDER...	I QUANT...	D PRICEE...	D SALES	I STATUS	I PRODU...	I PRODU...	L RECENCY	I DEALSIZE	S ORDER...	S S
Row0	AV Stores, Co.	L	H	H	LHH	51	1778	91.085	157,807.81	51	51	51	197	51	Bin 4	Bin 4
Row1	Alpha Cognac	H	C	C	HCC	20	687	101.16	70,488.44	20	20	20	65	20	Bin 1	Bin 1
Row2	Amica Model...	C	L	M	CLM	26	843	110.853	94,117.26	26	26	26	266	26	Bin 2	Bin 3
Row3	Anna's Decor...	M	H	H	MHH	46	1469	106.424	153,996.13	46	46	46	84	46	Bin 4	Bin 4
Row4	Atelier graph...	L	C	C	LCC	7	270	92.239	24,179.96	7	7	7	189	7	Bin 1	Bin 1
Row5	Australian C...	H	L	C	HLC	23	705	90.042	64,591.46	23	23	23	23	23	Bin 2	Bin 1
Row6	Australian C...	M	H	H	MHH	55	1926	104.59	200,995.41	55	55	55	185	55	Bin 4	Bin 4
Row7	Australian Gi...	M	C	C	MCC	15	545	110.554	59,469.12	15	15	15	120	15	Bin 1	Bin 1
Row8	Auto Assoc. ...	C	C	C	CCC	18	637	99.488	64,834.32	18	18	18	234	18	Bin 1	Bin 1
Row9	Auto Canal P...	H	M	M	HMM	27	1001	94.255	93,170.66	27	27	27	55	27	Bin 3	Bin 3
Row10	Auto-Moto Cl...	M	C	C	MCC	8	287	92.8	26,479.26	8	8	8	181	8	Bin 1	Bin 1
Row11	Baane Mini I...	L	M	M	LMM	32	1082	108.574	116,599.19	32	32	32	209	32	Bin 3	Bin 3
Row12	Bavarian Coll...	C	C	C	CCC	14	401	84.289	34,993.92	14	14	14	260	14	Bin 1	Bin 1
Row13	Blauer See A...	L	L	L	LLL	22	811	108.031	85,171.59	22	22	22	209	22	Bin 2	Bin 2
Row14	Boards & To...	M	C	C	MCC	3	102	89.807	9,129.35	3	3	3	114	3	Bin 1	Bin 1
Row15	CAF Imports	C	C	C	CCC	13	468	104.963	49,642.05	13	13	13	440	13	Bin 1	Bin 1
Row16	Cambridge C...	C	C	C	CCC	11	357	101.329	36,163.62	11	11	11	390	11	Bin 1	Bin 1
Row17	Canadian Gif...	L	L	L	LLL	22	703	105.341	75,238.92	22	22	22	223	22	Bin 2	Bin 2
Row18	Classic Gift I...	L	L	C	LLC	21	668	103.32	67,506.97	21	21	21	231	21	Bin 2	Bin 1
Row19	Classic Lege...	L	C	L	LCL	20	720	109.803	77,795.2	20	20	20	193	20	Bin 1	Bin 2
Row20	Clover Collec...	C	C	C	CCC	16	490	112.87	57,756.43	16	16	16	259	16	Bin 1	Bin 1
Row21	Collectable M...	C	L	L	CLL	25	954	91.535	87,489.23	25	25	25	461	25	Bin 2	Bin 2
Row22	Collectables ...	M	L	L	MLL	24	795	97.237	81,577.98	24	24	24	133	24	Bin 2	Bin 2
Row23	Corrida Auto...	L	M	H	LMH	32	1163	105.175	120,615.28	32	32	32	213	32	Bin 3	Bin 4
Row24	Cruz & Sons ...	L	L	M	LLM	26	961	96.08	94,015.73	26	26	26	198	26	Bin 2	Bin 3
Row25	Daedalus De...	C	C	C	CCC	20	699	95.474	69,052.41	20	20	20	466	20	Bin 1	Bin 1
Row26	Danish Whol...	H	H	H	HHH	36	1315	108.038	145,041.6	36	36	36	47	36	Bin 4	Bin 4
Row27	Diecast Class...	H	M	H	HMH	31	1111	108.566	122,138.14	31	31	31	2	31	Bin 3	Bin 4
Row28	Diecast Colle...	C	C	L	CCL	18	695	101.783	70,859.78	18	18	18	402	18	Bin 1	Bin 2
Row29	Double Deck...	C	C	C	CCC	12	357	99.108	36,019.04	12	12	12	496	12	Bin 1	Bin 1
Row30	Dragon Souv...	M	H	H	MHH	43	1524	113.106	172,989.68	43	43	43	91	43	Bin 4	Bin 4
Row31	Enaco Distrib...	L	L	L	LLL	23	882	88.783	78,411.86	23	23	23	190	23	Bin 2	Bin 2
Row32	Euro Shoppin...	H	H	H	HHH	259	9327	97.383	912,294.11	259	259	259	1	259	Bin 4	Bin 4
Row33	FunGiftIdeas...	M	L	M	MLM	26	903	109.587	98,923.73	26	26	26	90	26	Bin 2	Bin 3
Row34	Gift Depot Inc.	H	L	M	HLM	25	903	108.932	101,894.79	25	25	25	27	25	Bin 2	Bin 3

INFERENCES FROM RFM ANALYSIS AND IDENTIFIED SEGMENTS

Best Customers

CUSTOMERNAME	ORDERNUMBE	SALES	RECENC	ORDERNUMBER	SALES	RECENCY	Recency	Freque	Moneta	RFM Score
Euro Shopping Channel	259	912294.11	1	Bin 4	Bin 4	Bin 1	H	H	H	HHH
La Rochelle Gifts	53	180124.9	1	Bin 4	Bin 4	Bin 1	H	H	H	HHH
Mini Gifts Distributors Ltd.	180	654858.06	3	Bin 4	Bin 4	Bin 1	H	H	H	HHH
Souveniers And Things Co.	46	151570.98	3	Bin 4	Bin 4	Bin 1	H	H	H	HHH
Salzburg Collectables	40	149798.63	15	Bin 4	Bin 4	Bin 1	H	H	H	HHH

On basis on Recency, frequency & monetary top customers are grouped, given the most significance to recency parameter as these customers has recently purchased our products. According to RFM model the most importance metric is recency. Hence, considered it for selecting the top customers.

Example : Customer name -Euro Shopping Channel, they have recently made a purchase, also has high frequency with a high monetary transaction.

INFERENCES FROM RFM ANALYSIS AND IDENTIFIED SEGMENTS

Loyal Customers

CUSTOMERNAME	ORDERNUMBER	SALES	RECENC	ORDERNUMBER	SALES	RECENCY	Recency	Freque	Moneta	RFM Score
Euro Shopping Channel	259	912294.11	1	Bin 4	Bin 4	Bin 1	H	H	H	HHH
Mini Gifts Distributors Ltd.	180	654858.06	3	Bin 4	Bin 4	Bin 1	H	H	H	HHH
Australian Collectors, Co.	55	200995.41	185	Bin 4	Bin 4	Bin 2	M	H	H	MHH
Muscle Machine Inc	48	197736.94	183	Bin 4	Bin 4	Bin 2	M	H	H	MHH
La Rochelle Gifts	53	180124.9	1	Bin 4	Bin 4	Bin 1	H	H	H	HHH

On basis on Recency, frequency & monetary loyal customers are grouped. These customers have purchased multiple times with good monetary value.

To be focused this segment, so that the Loyal Customers turn to be the Best Customers.

INFERENCES FROM RFM ANALYSIS AND IDENTIFIED SEGMENTS

Verge of Churning Customers

CUSTOMERNAME	ORDERNUMBER	SALES	RECENC	ORDERNUMBER	SALES	RECENCY	Recency	Freque	Moneta	RFM Score
AV Stores, Co.	51	157807.81	197	Bin 4	Bin 4	Bin 3	L	H	H	LHH
Land of Toys Inc.	49	164069.44	199	Bin 4	Bin 4	Bin 3	L	H	H	LHH
Rovelli Gifts	48	137955.72	202	Bin 4	Bin 4	Bin 3	L	H	H	LHH
Saveley & Henriot, Co.	41	142874.25	457	Bin 4	Bin 4	Bin 4	C	H	H	CHH
Online Diecast Creations Co.	34	131685.3	210	Bin 4	Bin 4	Bin 3	L	H	H	LHH

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

Example : Customer Name : AV Stores, Co. – The frequency is good with good monetary value, but low recency made them stand in this group. If the company pays more attention and fulfil their requirement, then we can easily turn them into our regular customer and we can save them from churning out.

INFERENCES FROM RFM ANALYSIS AND IDENTIFIED SEGMENTS

Lost Customers

CUSTOMERNAME	ORDERNUMBER	SALES	RECENC	ORDERNUMBER	SALES	RECENCY	Recency	Frequei	Moneta	RFM Score
Double Decker Gift Stores, Ltd	12	36019.04	496	Bin 1	Bin 1	Bin 4	C	C	C	CCC
West Coast Collectables Co.	13	46084.64	489	Bin 1	Bin 1	Bin 4	C	C	C	CCC
Signal Collectibles Ltd.	15	50218.51	477	Bin 1	Bin 1	Bin 4	C	C	C	CCC
Daedalus Designs Imports	20	69052.41	466	Bin 1	Bin 1	Bin 4	C	C	C	CCC
CAF Imports	13	49642.05	440	Bin 1	Bin 1	Bin 4	C	C	C	CCC

On basis on Recency, frequency & monetary parameters the customers who we lost are grouped. Their recency is very low and hasn’t made any purchase since long. So, we can say these are our lost customers.

Regular feedback and understanding their requirements in detail, we might bring them back to been a good customer.

RECOMMENDATIONS

- Recency, frequency & monetary parameters are used to group top , loyal, on the verge of churning and lost customers.
- Customers with good recency has been our top customers.
- Customer with low recency and low monetary were lost customer.
- Customers on verge of churning can be saved and can be converted to either top or loyal customers.
- The RFM model helps the companies to understand the sales based on top, loyal and verge of churning and lost customers and they can act upon it well in advance and produce the strategies to convert the churning customer as regular customers, Lost Customers to Good Customer and Loyal Customer to Best Customer.