MRA Project – Milestone 1

- 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

Fact check about given data

- Shape of dataset – Number of rows – 2747; Number of columns – 20
- Null Values – All the 20 columns doesn't have any null values in it.

Dataset statistics

Number of variables	19
Number of observations	2747
Missing cells	0
Missing cells (%)	0.0%
Duplicate rows	0
Duplicate rows (%)	0.0%
Total size in memory	407.9 KiB
Average record size in memory	152.0 B

Variable types

Numeric	6
DateTime	1
Categorical	12

High Cardinality -High-cardinality refers to columns with values that are very uncommon or unique. High-cardinality column values are typically identification numbers, email addresses, or user names

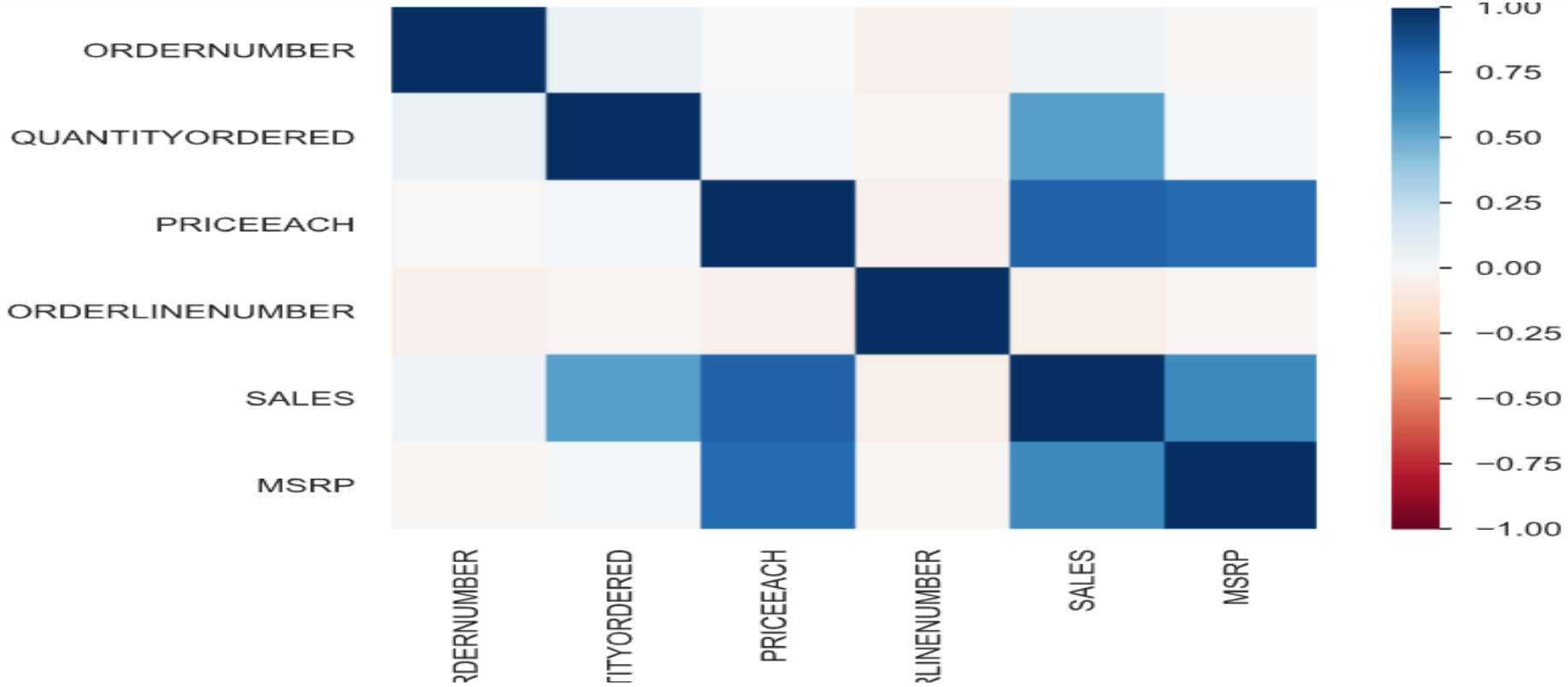
`PRODUCTCODE` has a high cardinality: 109 distinct values	High cardinality
`CUSTOMERNAME` has a high cardinality: 89 distinct values	High cardinality
`PHONE` has a high cardinality: 88 distinct values	High cardinality
`ADDRESSLINE1` has a high cardinality: 89 distinct values	High cardinality
`CITY` has a high cardinality: 71 distinct values	High cardinality
`POSTALCODE` has a high cardinality: 73 distinct values	High cardinality
`CONTACTLASTNAME` has a high cardinality: 76 distinct values	High cardinality
`CONTACTFIRSTNAME` has a high cardinality: 72 distinct values	High cardinality

High Correlation - When two sets of data are strongly linked together we say they have a High Correlation. The following variables are highly correlated

QUANTITYORDERED is highly correlated with SALES	High correlation
PRICEEACH is highly correlated with SALES and 1 other fields	High correlation
SALES is highly correlated with QUANTITYORDERED and 2 other fields	High correlation
MSRP is highly correlated with PRICEEACH and 1 other fields	High correlation
QUANTITYORDERED is highly correlated with SALES	High correlation
PRICEEACH is highly correlated with SALES and 1 other fields	High correlation
SALES is highly correlated with QUANTITYORDERED and 2 other fields	High correlation
MSRP is highly correlated with PRICEEACH and 1 other fields	High correlation
PRICEEACH is highly correlated with SALES and 1 other fields	High correlation
SALES is highly correlated with PRICEEACH	High correlation

MSRP is highly correlated with PRICEEACH	High correlation
ADDRESSLINE1 is highly correlated with CUSTOMERNAME and 9 other fields	High correlation
QUANTITYORDERED is highly correlated with SALES and 1 other fields	High correlation
CUSTOMERNAME is highly correlated with ADDRESSLINE1 and 9 other fields	High correlation
POSTALCODE is highly correlated with ADDRESSLINE1 and 9 other fields	High correlation
MSRP is highly correlated with SALES and 3 other fields	High correlation
CONTACTLASTNAME is highly correlated with ADDRESSLINE1 and 9 other fields	High correlation
CONTACTFIRSTNAME is highly correlated with ADDRESSLINE1 and 9 other fields	High correlation
COUNTRY is highly correlated with ADDRESSLINE1 and 7 other fields	High correlation
STATUS is highly correlated with ADDRESSLINE1 and 7 other fields	High correlation
CITY is highly correlated with ADDRESSLINE1 and 9 other fields	High correlation
ORDERNUMBER is highly correlated with ADDRESSLINE1 and 8 other fields	High correlation
SALES is highly correlated with QUANTITYORDERED and 3 other fields	High correlation
PRODUCTLINE is highly correlated with ADDRESSLINE1 and 7 other fields	High correlation
PRICEEACH is highly correlated with MSRP and 2 other fields	High correlation
PHONE is highly correlated with ADDRESSLINE1 and 9 other fields	High correlation
DEALSIZE is highly correlated with QUANTITYORDERED and 3 other fields	High correlation
ADDRESSLINE1 is highly correlated with CUSTOMERNAME and 6 other fields	High correlation
CUSTOMERNAME is highly correlated with ADDRESSLINE1 and 6 other fields	High correlation

POSTALCODE is highly correlated with ADDRESSLINE1 and 6 other fields	High correlation
CONTACTLASTNAME is highly correlated with ADDRESSLINE1 and 6 other fields	High correlation
CONTACTFIRSTNAME is highly correlated with ADDRESSLINE1 and 6 other fields	High correlation
COUNTRY is highly correlated with ADDRESSLINE1 and 6 other fields	High correlation
CITY is highly correlated with ADDRESSLINE1 and 6 other fields	High correlation
PHONE is highly correlated with ADDRESSLINE1 and 6 other fields	High correlation
PRODUCTCODE is uniformly distributed	Uniform



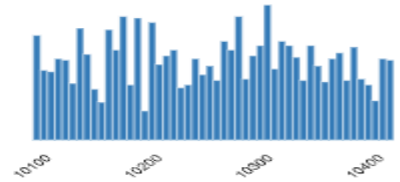
ORDERNUM...

Real number ($\mathbb{R}_{\geq 0}$)

HIGH CORRELATION

Distinct	298
Distinct (%)	10.8%
Missing	0
Missing (%)	0.0%
Infinite	0
Infinite (%)	0.0%
Mean	10259.76156

Minimum	10100
Maximum	10425
Zeros	0
Zeros (%)	0.0%
Negative	0
Negative (%)	0.0%
Memory size	21.6 KiB



QUANTITYO...

Real number ($\mathbb{R}_{\geq 0}$)

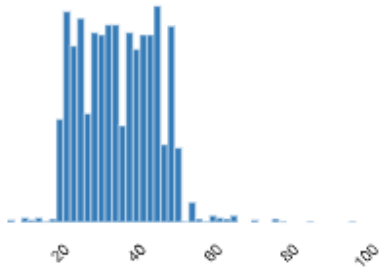
HIGH CORRELATION

HIGH CORRELATION

HIGH CORRELATION

Distinct	58
Distinct (%)	2.1%
Missing	0
Missing (%)	0.0%
Infinite	0
Infinite (%)	0.0%
Mean	35.10302148

Minimum	6
Maximum	97
Zeros	0
Zeros (%)	0.0%
Negative	0
Negative (%)	0.0%
Memory size	21.6 KiB



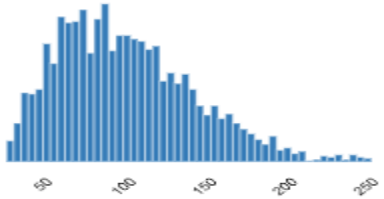
PRICEEACH

Real number ($\mathbb{R}_{\geq 0}$)

HIGH CORRELATION
HIGH CORRELATION
HIGH CORRELATION
HIGH CORRELATION

Distinct	1984
Distinct (%)	72.2%
Missing	0
Missing (%)	0.0%
Infinite	0
Infinite (%)	0.0%
Mean	101.0989511

Minimum	26.88
Maximum	252.87
Zeros	0
Zeros (%)	0.0%
Negative	0
Negative (%)	0.0%
Memory size	21.6 KiB

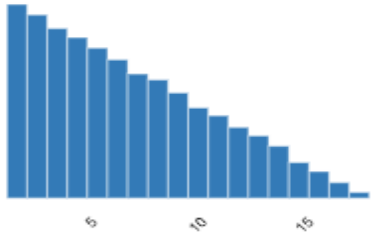


ORDERLINE...

Real number ($\mathbb{R}_{\geq 0}$)

Distinct	18
Distinct (%)	0.7%
Missing	0
Missing (%)	0.0%
Infinite	0
Infinite (%)	0.0%
Mean	6.491081179

Minimum	1
Maximum	18
Zeros	0
Zeros (%)	0.0%
Negative	0
Negative (%)	0.0%
Memory size	21.6 KiB



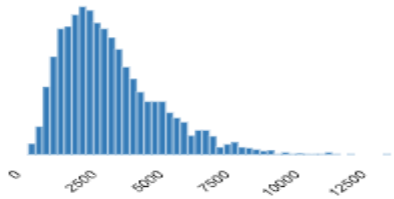
SALES

Real number (ℝ≥0)

HIGH_CORRELATION
HIGH_CORRELATION
HIGH_CORRELATION
HIGH_CORRELATION

Distinct	2690
Distinct (%)	97.9%
Missing	0
Missing (%)	0.0%
Infinite	0
Infinite (%)	0.0%
Mean	3553.047583

Minimum	482.13
Maximum	14082.8
Zeros	0
Zeros (%)	0.0%
Negative	0
Negative (%)	0.0%
Memory size	21.6 KiB

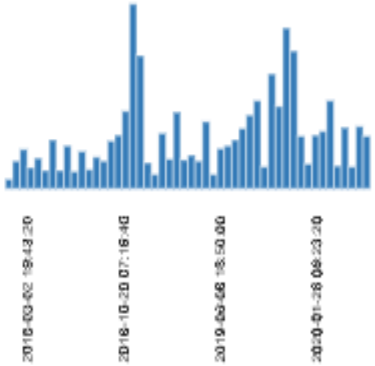


ORDERDATE

Date

Distinct	246
Distinct (%)	9.0%
Missing	0
Missing (%)	0.0%
Memory size	21.6 KiB

Minimum	2018-01-06 00:00:00
Maximum	2020-05-31 00:00:00

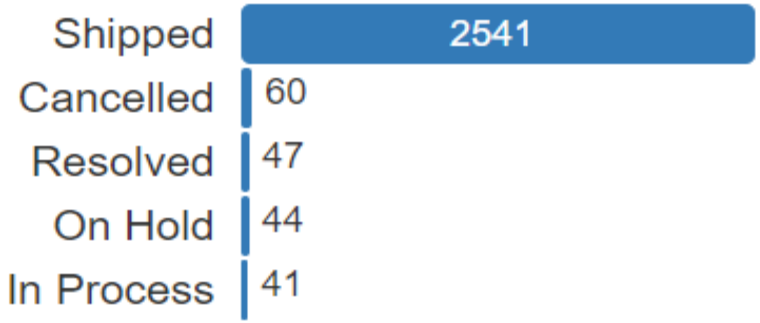


STATUS

Categorical

HIGH CORRELATION

Distinct	6
Distinct (%)	0.2%
Missing	0
Missing (%)	0.0%
Memory size	21.6 KiB

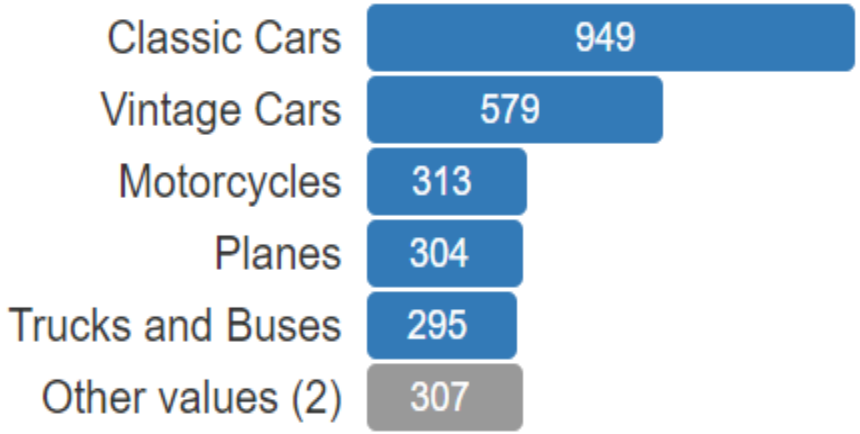


PRODUCTLINE

Categorical

HIGH CORRELATION

Distinct	7
Distinct (%)	0.3%
Missing	0
Missing (%)	0.0%
Memory size	21.6 KiB



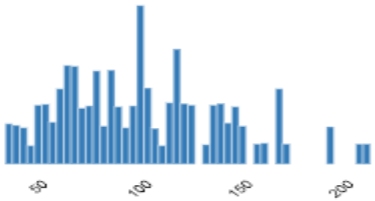
MSRP

Real number (ℝ≥0)

HIGH CORRELATION
HIGH CORRELATION
HIGH CORRELATION
HIGH CORRELATION

Distinct	80
Distinct (%)	2.9%
Missing	0
Missing (%)	0.0%
Infinite	0
Infinite (%)	0.0%
Mean	100.6916636

Minimum	33
Maximum	214
Zeros	0
Zeros (%)	0.0%
Negative	0
Negative (%)	0.0%
Memory size	21.6 KiB



PRODUCTCO…

Categorical

HIGH CARDINALITY
UNIFORM

Distinct	109
Distinct (%)	4.0%
Missing	0
Missing (%)	0.0%
Memory size	21.6 KiB

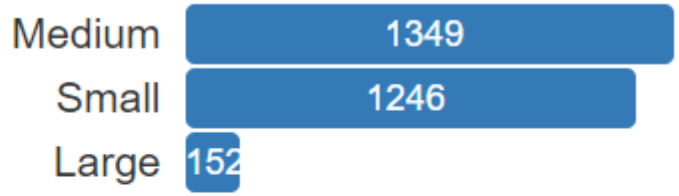
S18_3232	51
S32_2509	28
S24_1444	28
S50_1392	28
S24_2840	28
Other values (104)	2584

DEALSIZE

Categorical

HIGH CORRELATION

Distinct	3
Distinct (%)	0.1%
Missing	0
Missing (%)	0.0%
Memory size	21.6 KiB

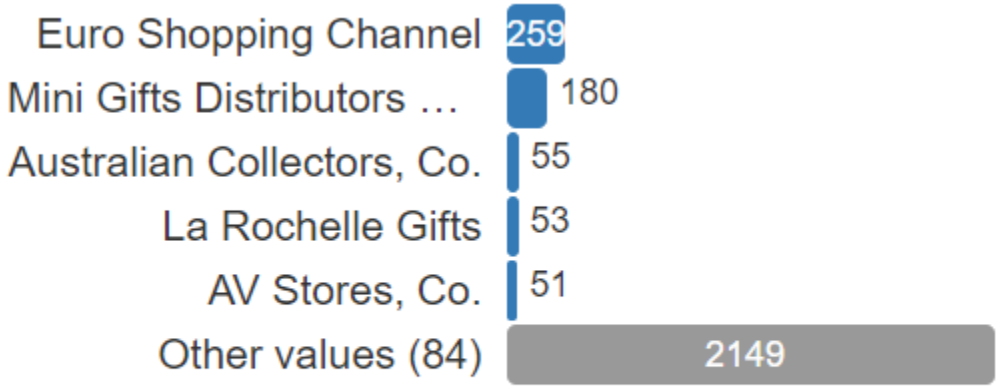


CUSTOMERN…

Categorical

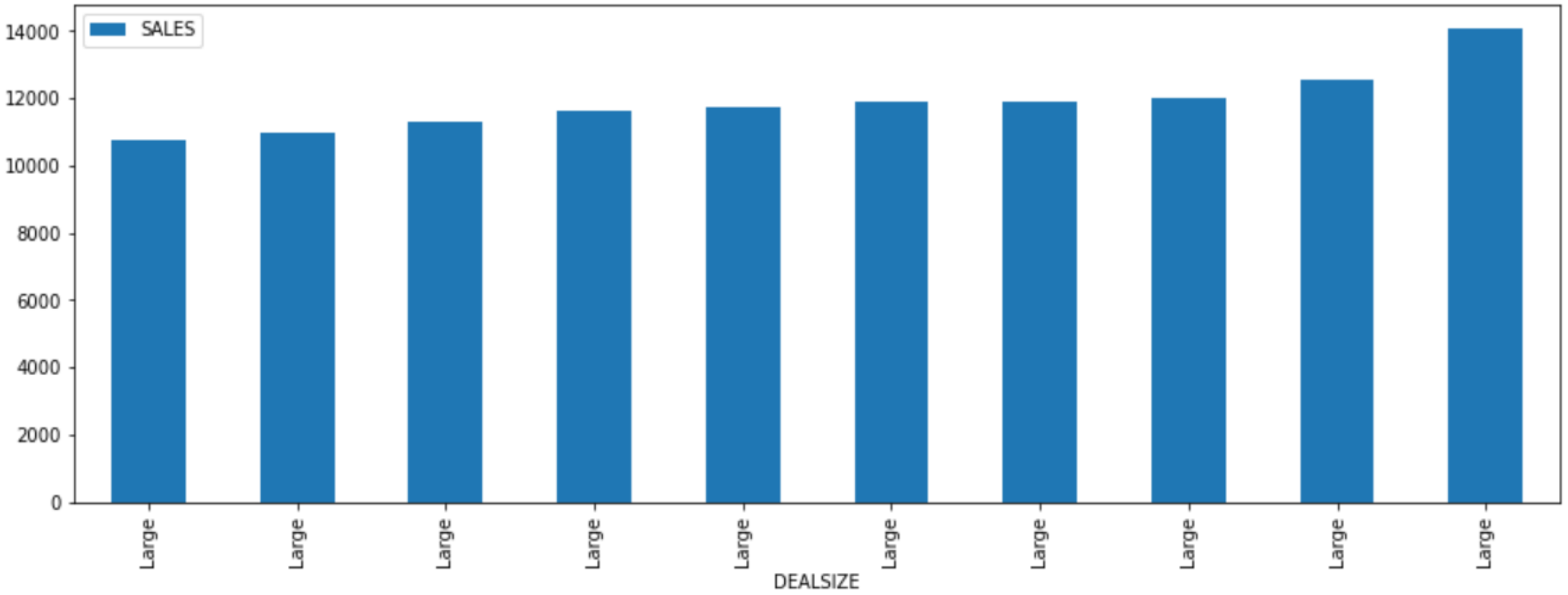
HIGH CARDINALITY
HIGH CORRELATION
HIGH CORRELATION

Distinct	89
Distinct (%)	3.2%
Missing	0
Missing (%)	0.0%
Memory size	21.6 KiB

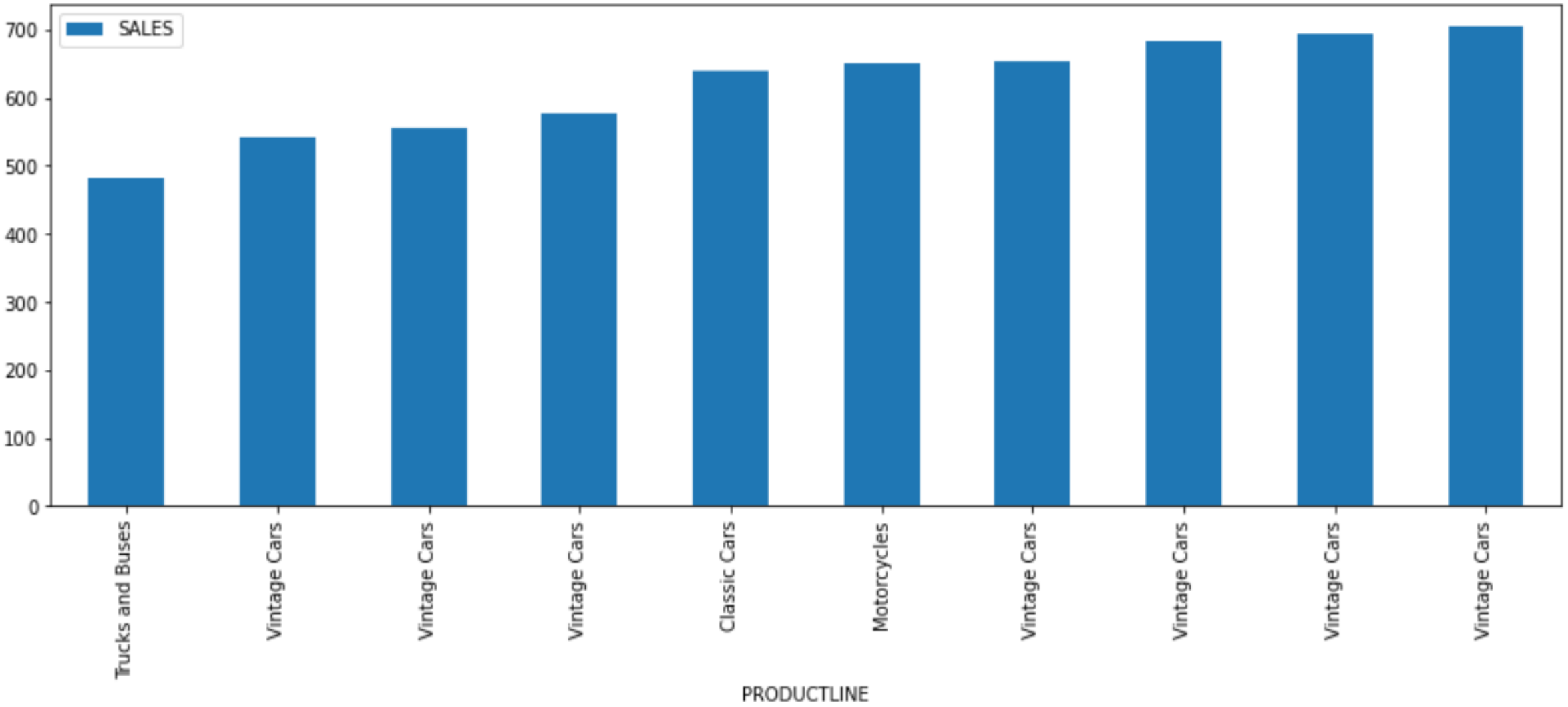


Bi-variate Analysis

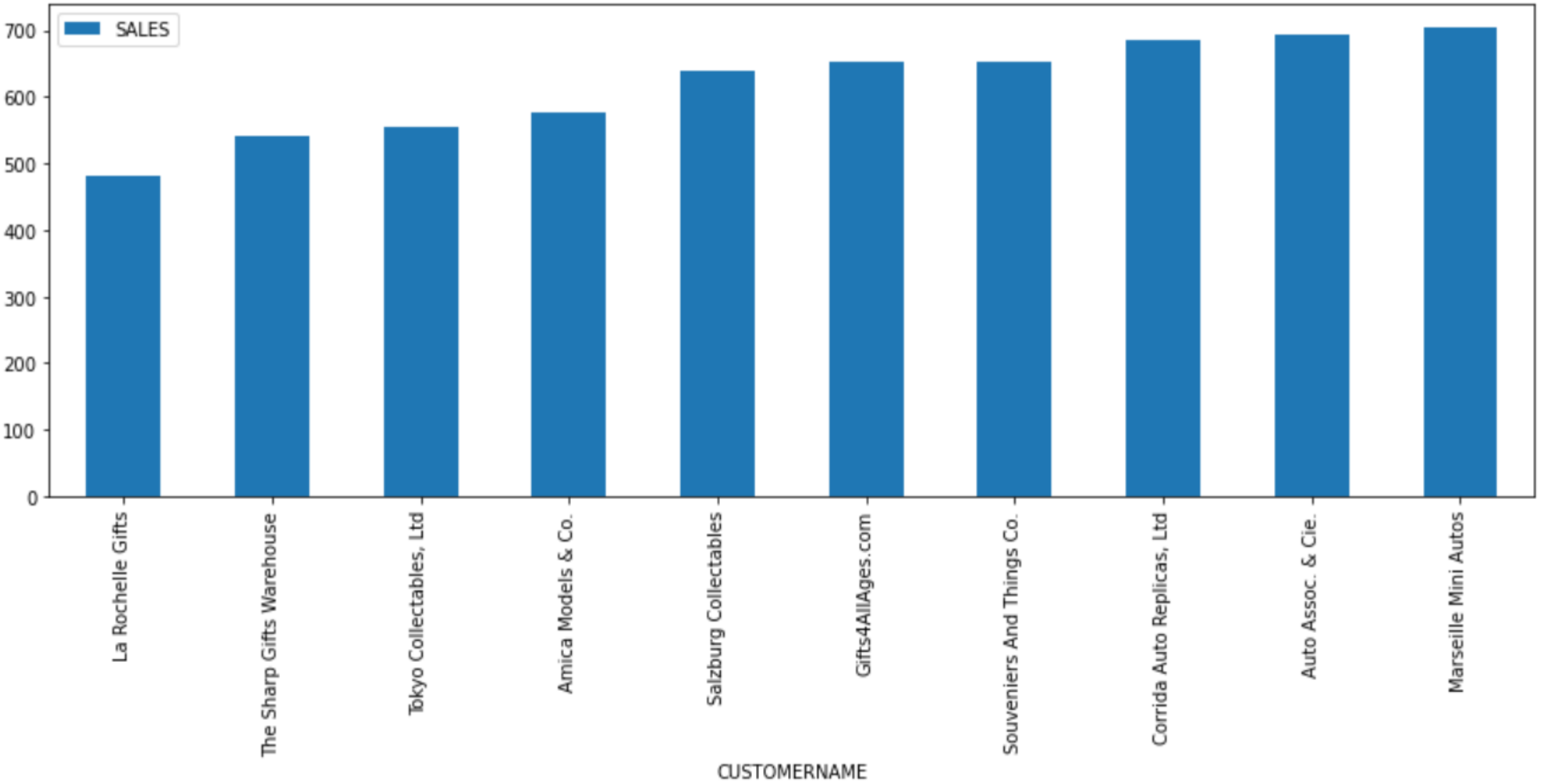
Numerical Vs Categorical

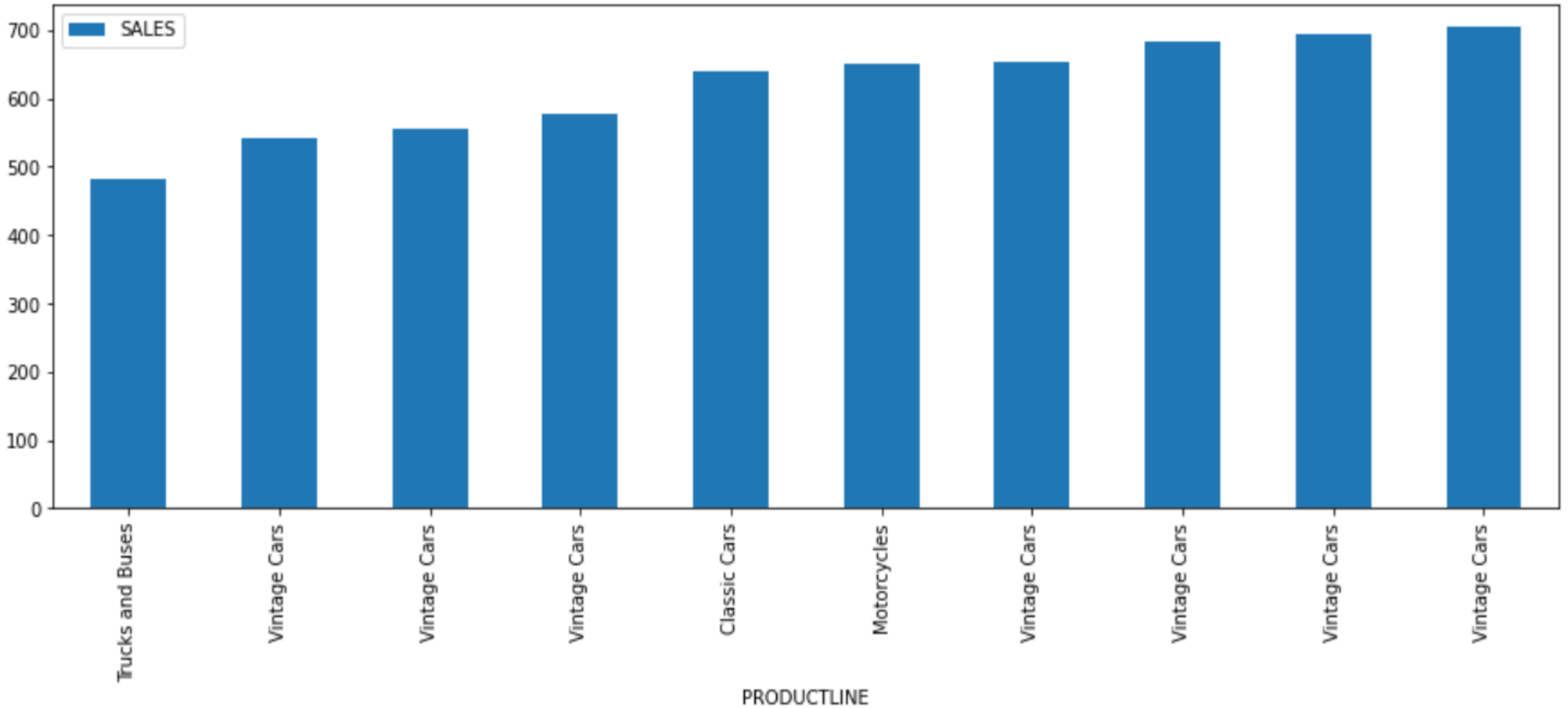


Bi variate analysis... contd



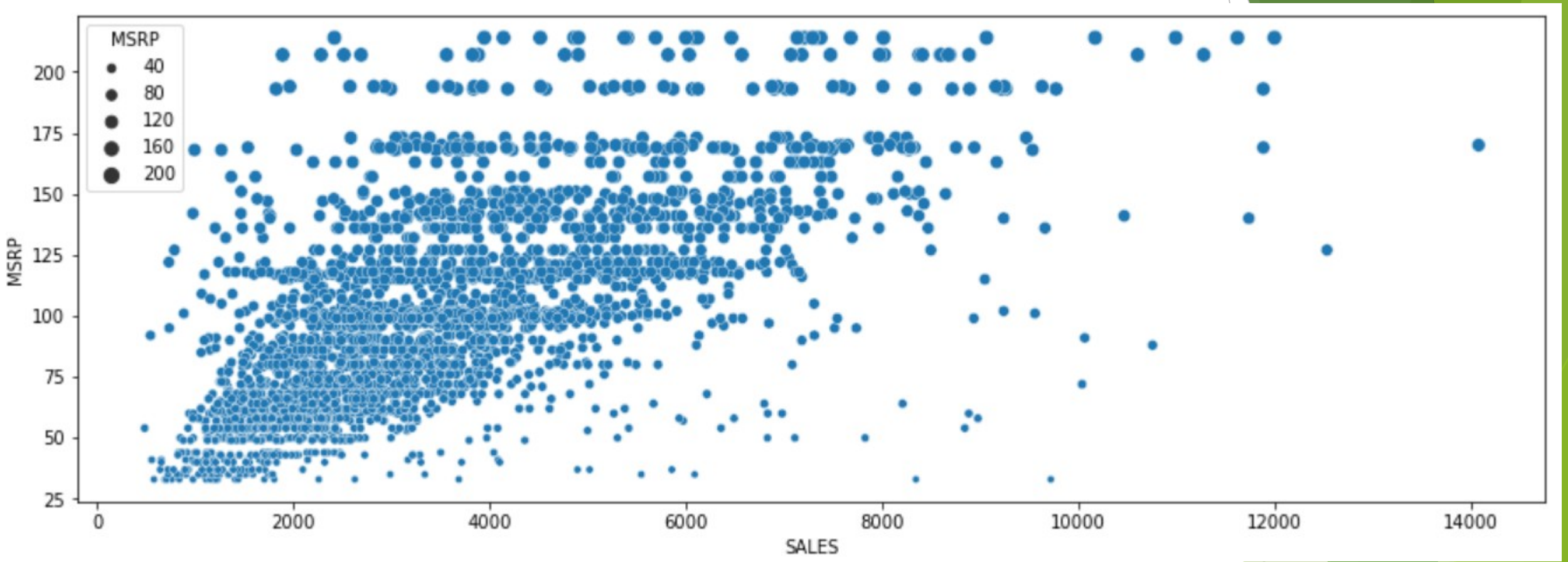
Bi variate analysis... contd



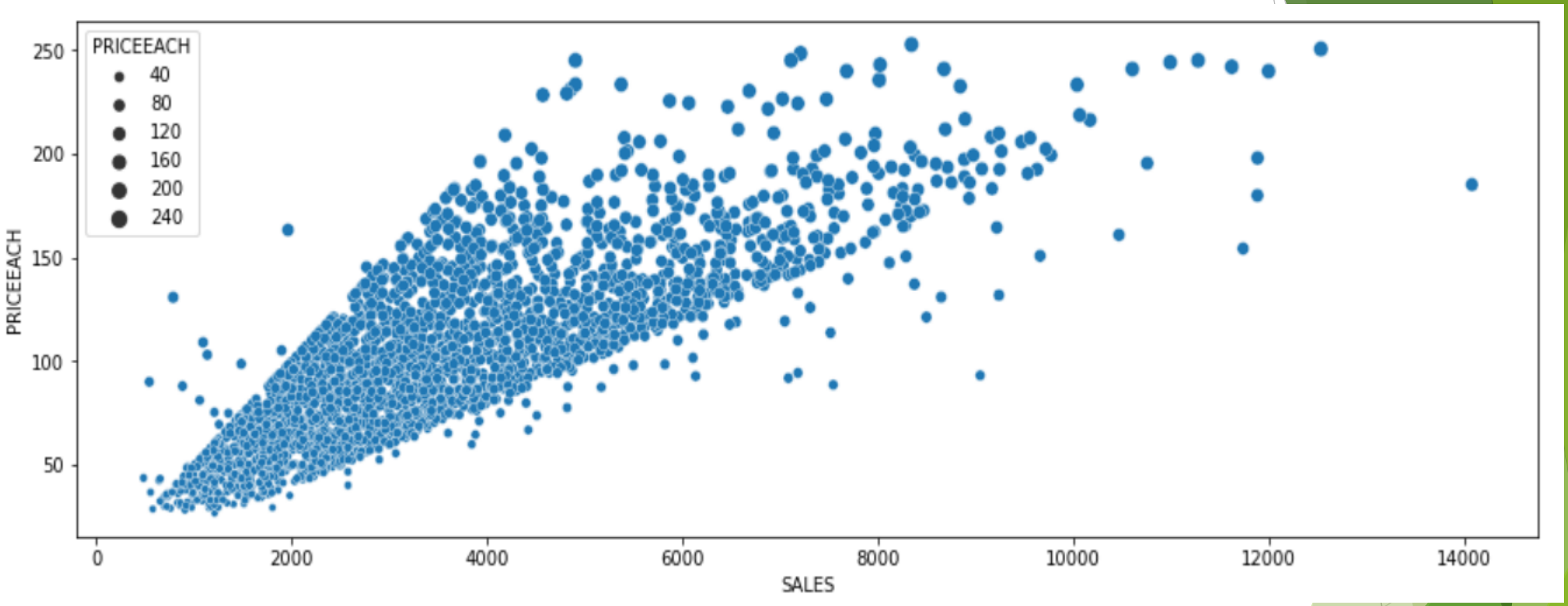

SALES
700
600
500
400
300
200
100
0
Trucks and Buses
Vintage Cars
Vintage Cars
Vintage Cars
Classic Cars
Motorcycles
Vintage Cars
Vintage Cars
Vintage Cars
Vintage Cars
PRODUCTLINE

Multi variate Analysis

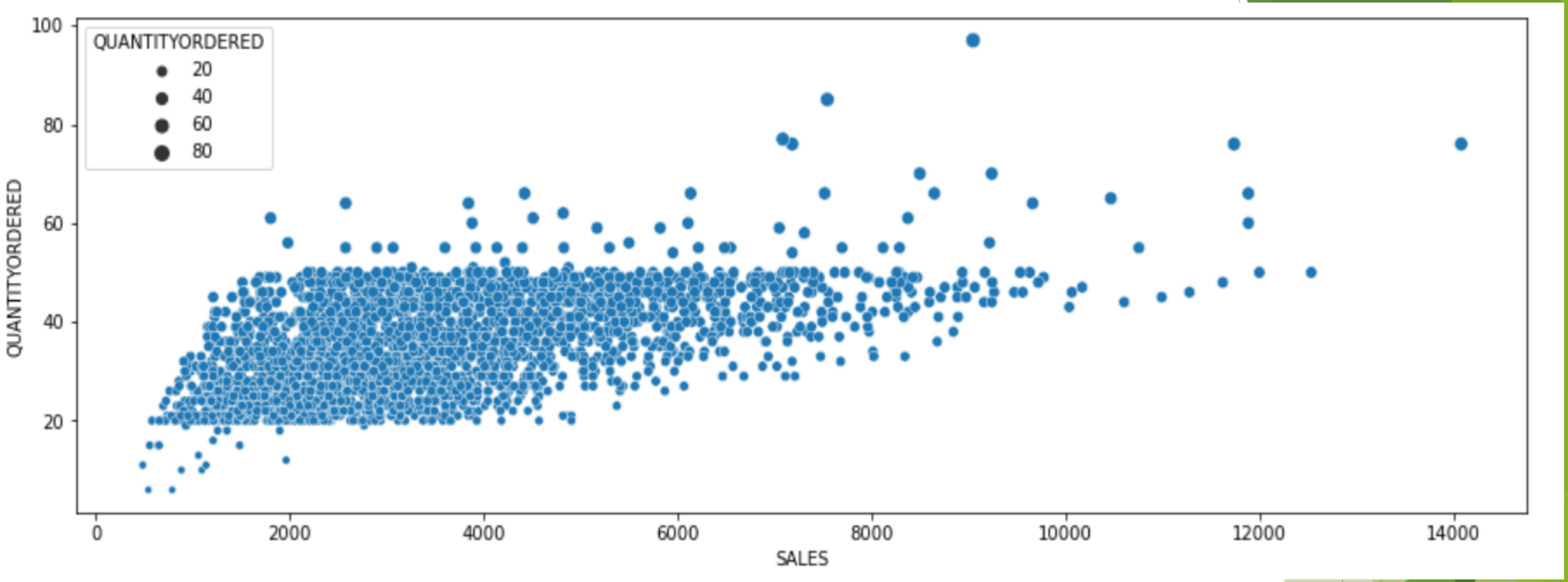
Numerical Vs Numerical



Multi variate analysis… contd

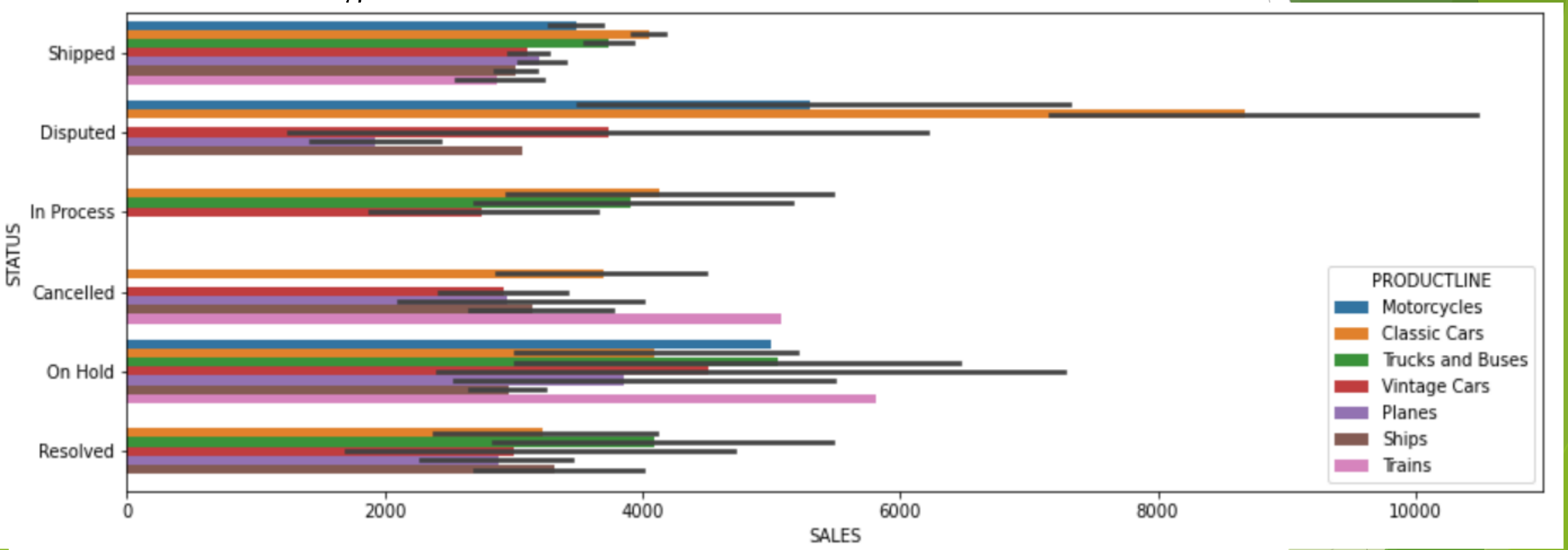


Multi variate analysis… contd

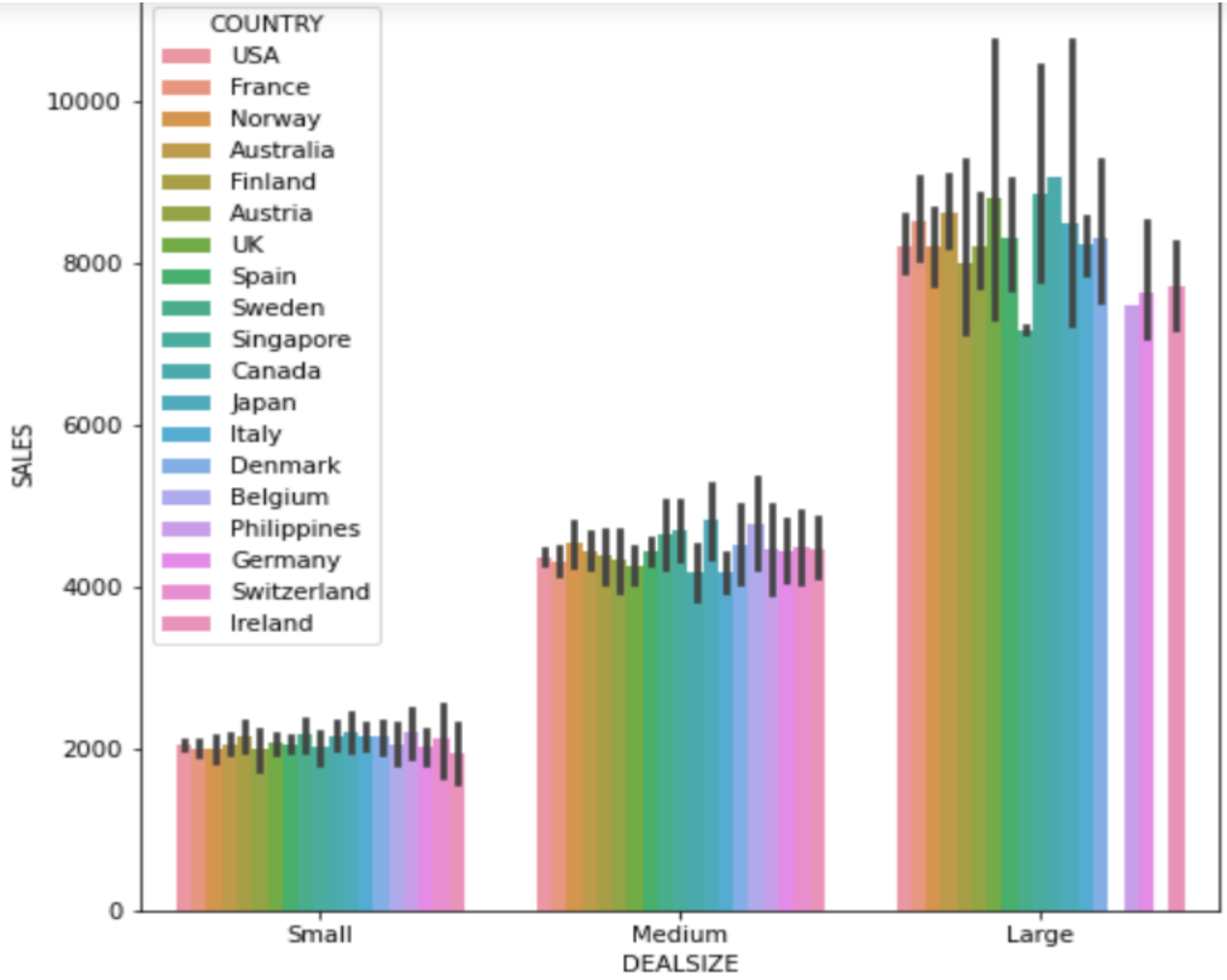


Bi variate analysis… contd

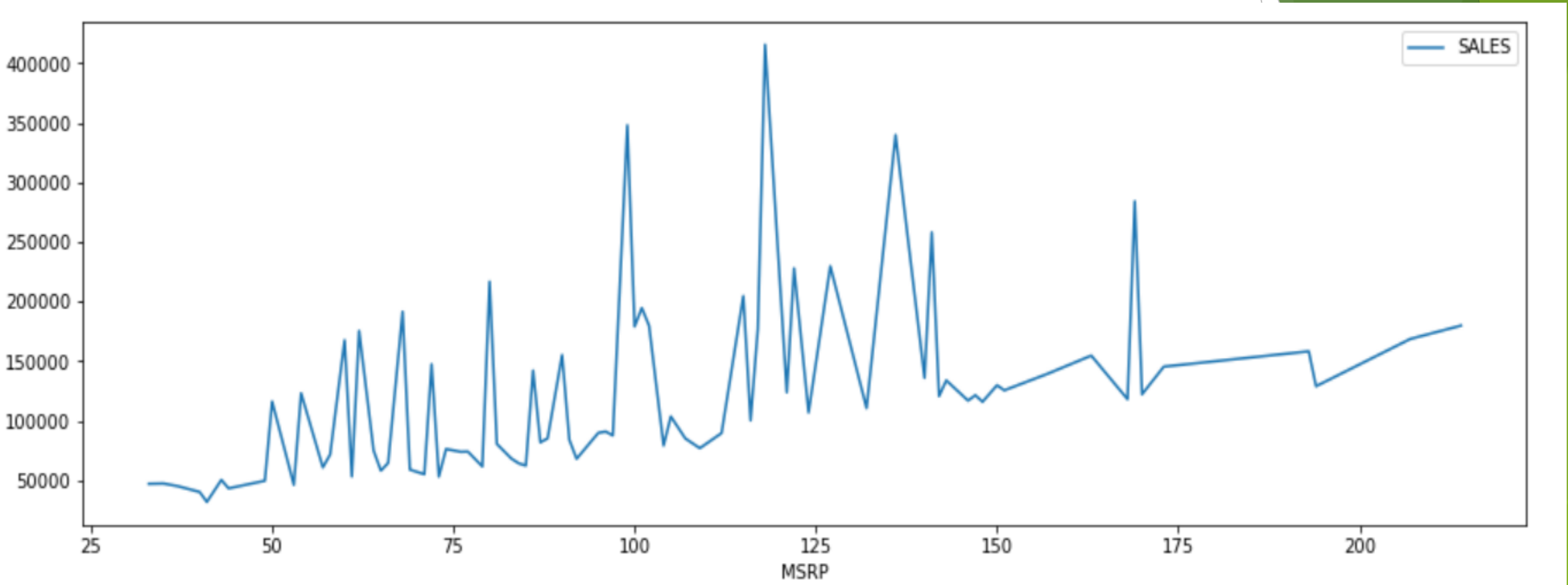
Numerical Vs 2 Categorical



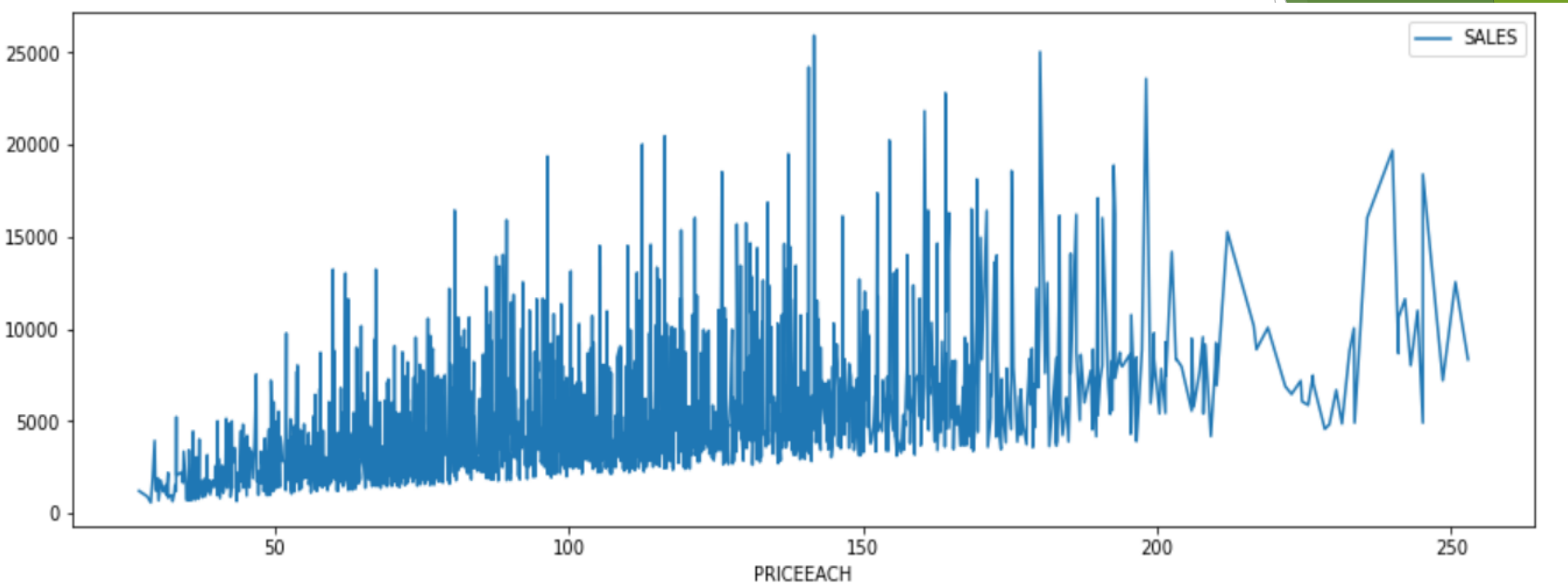
Bi variate analysis… contd



Trend Analysis



Trend analysis



Inferences

- 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 sales of product line like vintage cars and motorcycles is in saturation stage and there is no scope in them, however trucks and buses product line is still have scoop for sales.
- 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 rationale way because in case there is client churn's it will impact the sales of the company grossly.
- Company has huge pile of orders which are disputed in classic cars category followed by motorcycles category, so therefore management should try to resolve these disputes so that sales could not be impacted.
- Company drives export revenue mainly from large size deals followed by medium size deals and small size deals therefore it can be inferenced that large size deals is mainly from foreign customers and company should focus on domestic sales as well.

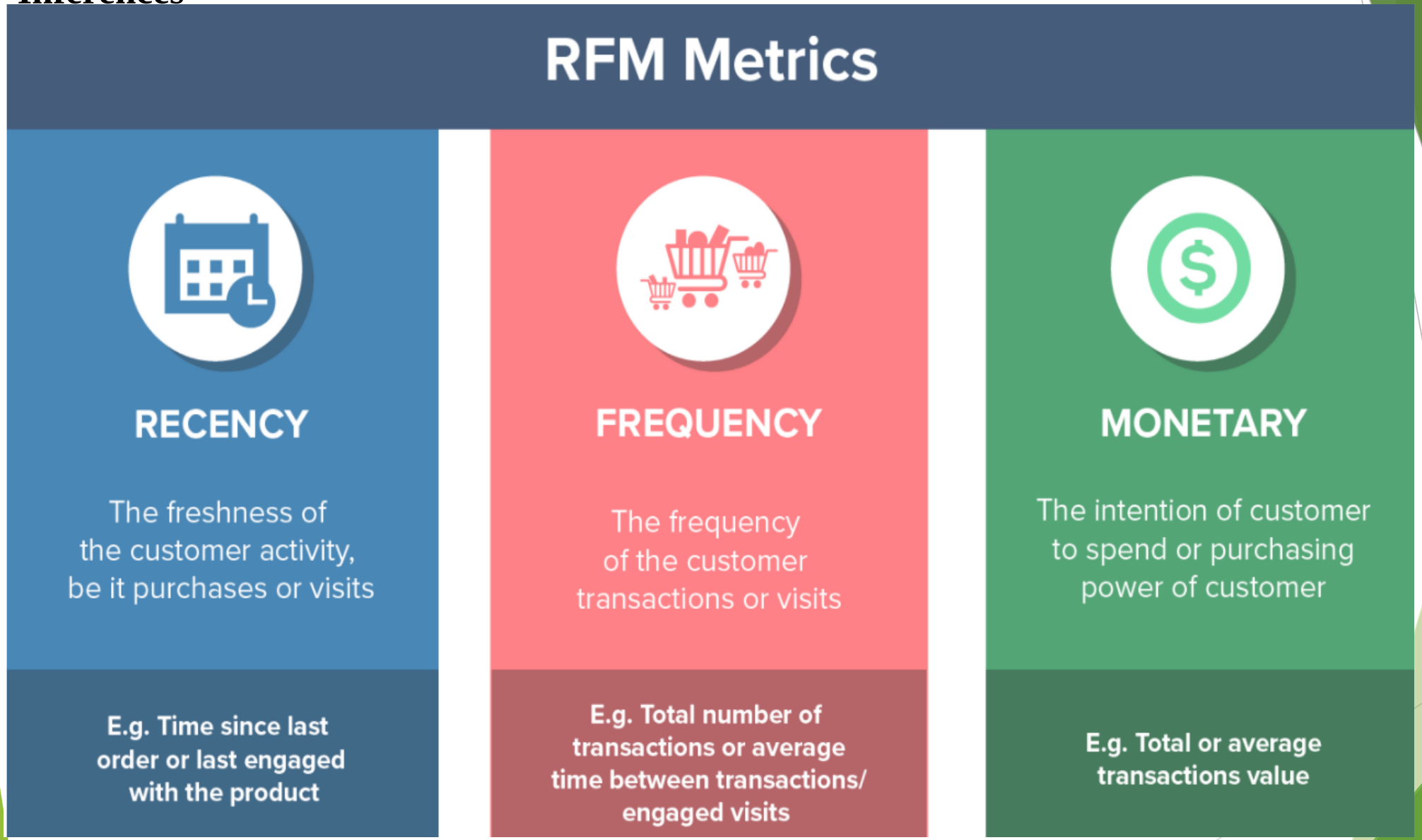
RFM ANALYSIS

Row ID	I QUANT...	I ORDER...	D Monetary	L Recency	S COUNTRY	D PRICEE...	I ORDER...	D SALES	S QUANT...	S Moneta...	S Recenc...	S MONET...	S Freque...	S RECEN...
Row0	6	1	541.14	395	USA	90.19	1	541.14	Bin 1	Bin 1	Bin 2	L	L	M
Row1	11	1	1,135.31	414	USA	103.21	1	1,135.31	Bin 1	Bin 1	Bin 2	L	L	M
Row2	13	1	1,057.29	395	USA	81.33	1	1,057.29	Bin 1	Bin 1	Bin 2	L	L	M
Row3	16	1	1,207.68	381	USA	75.48	1	1,207.68	Bin 1	Bin 1	Bin 1	L	L	H
Row4	19	1	2,764.88	381	USA	145.52	1	2,764.88	Bin 1	Bin 1	Bin 1	L	L	H
Row5	20	31	65,787.2	416	USA	106.108	31	2,122.168	Bin 1	Bin 2	Bin 3	M	L	L
Row6	21	30	66,598.35	414	USA	105.712	30	2,219.945	Bin 1	Bin 2	Bin 2	M	L	M
Row7	22	27	68,924.68	382	USA	116.035	27	2,552.766	Bin 1	Bin 2	Bin 2	M	L	M
Row8	23	24	52,619.86	381	USA	95.326	24	2,192.494	Bin 1	Bin 2	Bin 1	M	L	H
Row9	24	24	61,524.48	382	USA	106.813	24	2,563.52	Bin 1	Bin 2	Bin 2	M	L	M
Row10	25	30	77,388	357	USA	103.184	30	2,579.6	Bin 1	Bin 2	Bin 1	M	L	H
Row11	26	35	95,025.32	395	USA	104.423	35	2,715.009	Bin 1	Bin 2	Bin 2	M	L	M
Row12	27	35	110,643.84	381	USA	117.083	35	3,161.253	Bin 2	Bin 2	Bin 1	M	M	H
Row13	28	24	66,149.44	381	USA	98.437	24	2,756.227	Bin 2	Bin 2	Bin 1	M	M	H
Row14	29	26	66,686.95	548	USA	88.444	26	2,564.883	Bin 2	Bin 2	Bin 3	M	M	L
Row15	30	26	68,910.6	416	USA	88.347	26	2,650.408	Bin 2	Bin 2	Bin 3	M	M	L
Row16	31	38	108,315.86	381	USA	91.949	38	2,850.417	Bin 2	Bin 2	Bin 1	M	M	H
Row17	32	23	70,235.2	460	USA	95.428	23	3,053.704	Bin 2	Bin 2	Bin 3	M	M	L
Row18	33	34	117,086.97	425	USA	104.356	34	3,443.734	Bin 2	Bin 3	Bin 3	H	M	L
Row19	34	28	96,132.62	381	USA	100.98	28	3,433.308	Bin 2	Bin 2	Bin 1	M	M	H
Row20	35	21	70,145.95	358	USA	95.437	21	3,340.283	Bin 2	Bin 2	Bin 1	M	M	H
Row21	36	34	119,235.6	382	USA	97.415	34	3,506.929	Bin 2	Bin 3	Bin 2	H	M	M
Row22	37	29	106,142.64	381	USA	98.921	29	3,660.091	Bin 2	Bin 2	Bin 1	M	M	H
Row23	38	32	130,587.76	414	USA	107.391	32	4,080.867	Bin 2	Bin 3	Bin 2	H	M	M
Row24	39	31	113,528.22	425	USA	93.903	31	3,662.201	Bin 2	Bin 3	Bin 3	H	M	L
Row25	40	26	94,548.4	358	USA	90.912	26	3,636.477	Bin 2	Bin 2	Bin 1	M	M	H

Customer Segmentation using RFM analysis

- For customer segmentation using RFM analysis we have used KNIME tool.
- We have used country column for recency analysis and chosen country as USA.
- RFM Analysis
- We have divided the customers into three bins i.e. Bin-1, Bin-2, Bin-3 and mapped them to recency, frequency and monetary columns.
- After performing RFM analysis, it was found that Recency consists of top 25% customer which can be termed as active customers.
- Next Bin consist of customers which are at risk and comes in Bin-2.
- After that the bottom consists of inactive customers which are in Bin-3 and in lowest quartile.

Inferences


RFM Metrics
RECENCY
The freshness of the customer activity, be it purchases or visits
E.g. Time since last order or last engaged with the product
FREQUENCY
The frequency of the customer transactions or visits
E.g. Total number of transactions or average time between transactions/ engaged visits
MONETARY
The intention of customer to spend or purchasing power of customer
E.g. Total or average transactions value

Inferences cont.…

Champions are those best customers, who bought most recently, most often, and are heavy spenders. They can become early adopters for new products and will help promote the brand.

Potential Loyalists are most recent customers with average frequency and who spent a good amount.

New Customers are those customers who have a high overall RFM score but are not frequent shoppers. Start building relationships with these customers by providing onboarding support and special offers to increase their visits.

At Risk Customers are those customers who purchased often and spent big amounts, but haven't purchased recently.

Can't Lose Them are customers who used to visit and purchase quite often, but haven't been visiting recently.

Conclusions - We found that most of the customers belong to Bin-2 which is categorized as medium meaning that customers in this category are at risk and they might switch to other supplier as well in future if there will be price issue.