Marketing and Retail Analytics Project – Part 1

CHETAN DUDHANE

PGP DSBA GRP 2 JULY B

30-MAY-2021

Problem Statement

RFM Segmentation

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.

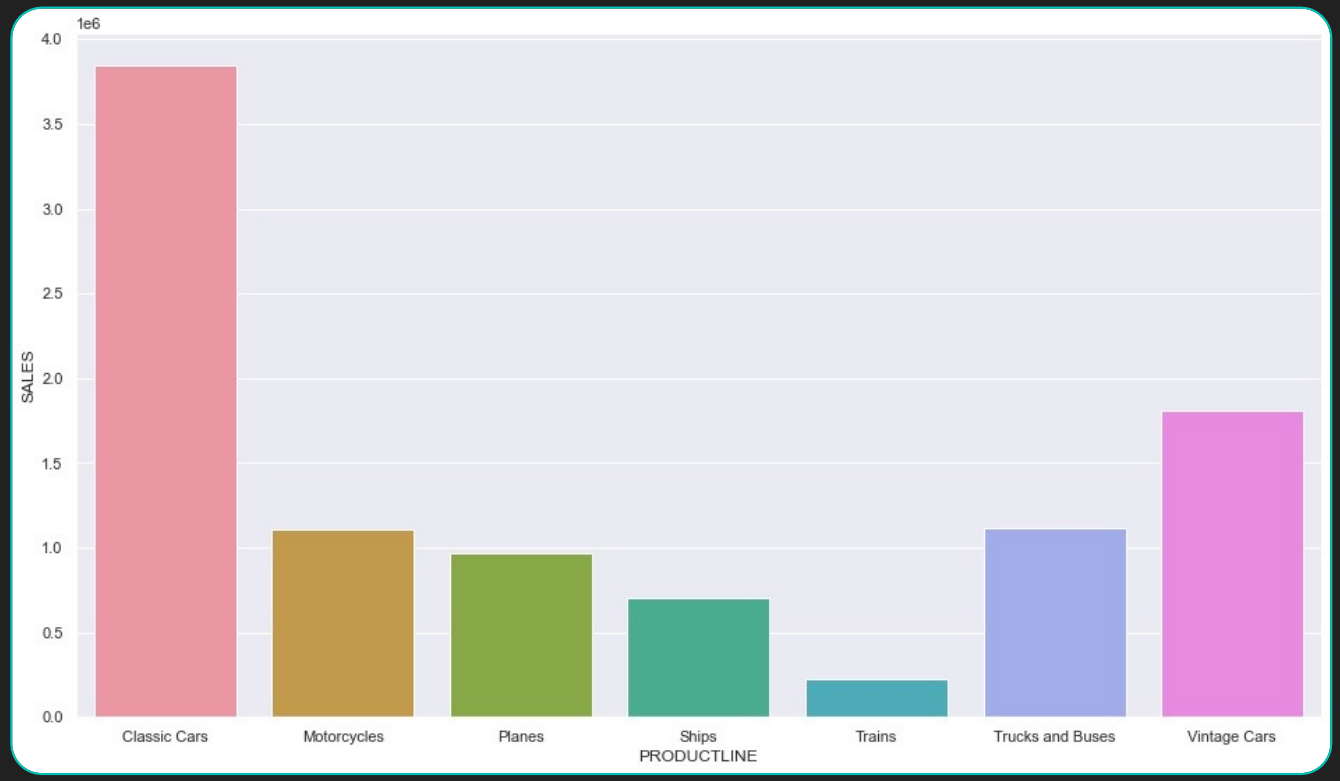
Your job is to use your magical data science skills to provide them with suitable insights about their data and their customers.

Synopsis

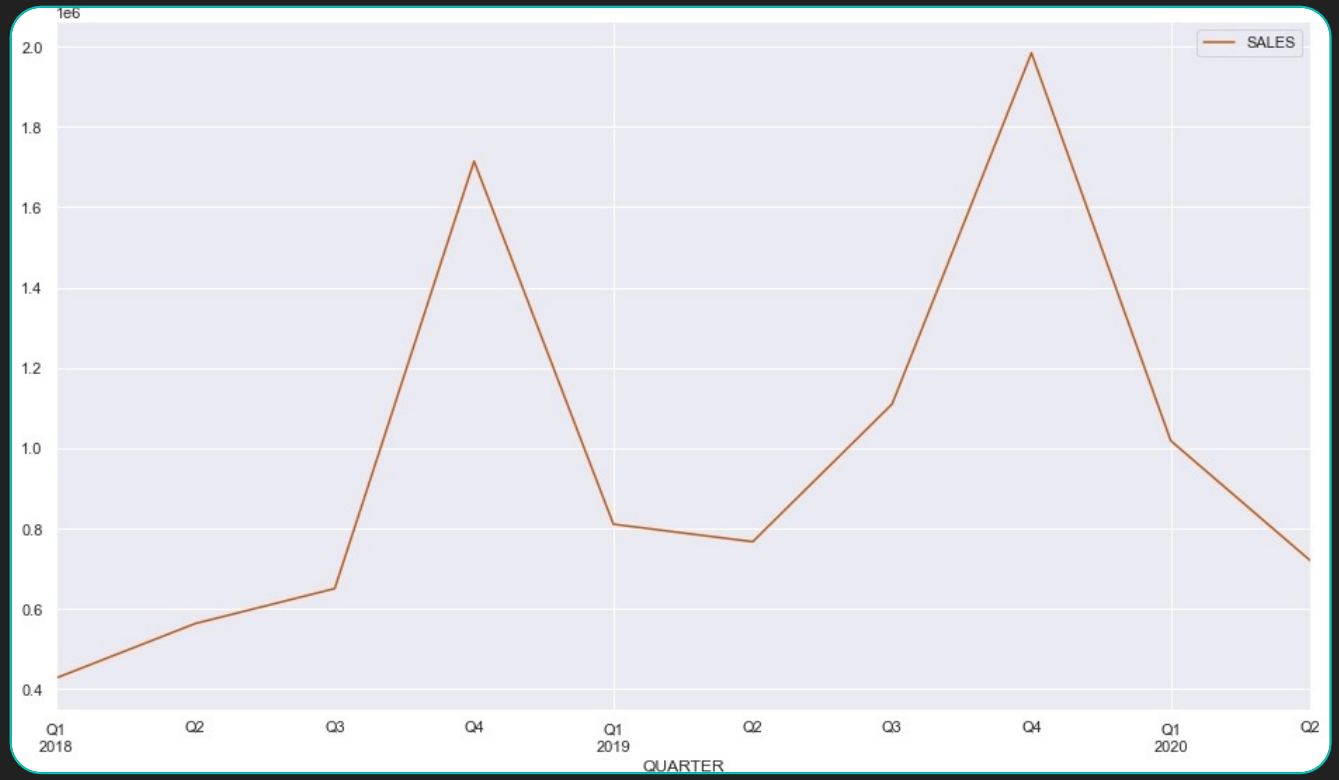
- Total No. of Sales Entries = 2747
- Total No. of Variables = 20
- No. of Missing Entries = 0
- No. of Duplicate Entries = 0
- Sales Data in the dataset is -
  - From 6th Jan 2018 to 31st May 2020
- Range of Products = 109
- Number of Customers = 89
- Number of Ordering Countries = 19
- Sales entries of spare parts of following Product Line categories (7) -
  - Classic cars
  - Motorcycles
  - Planes
  - Ships
  - Trains
  - Trucks and Buses
  - Vintage Cars

Synopsis contd..

- Top 3 selling Categories are –
  - Classic cars
  - Vintage cars
  - Trucks and Buses
- Least selling category is
  - Trains



Synopsis contd..



- Total sales start to increase from Q2
- Spike in sales is seen in Q4 of every year
- Sharp drop in sales in Q1

Synopsis contd..

- Every parameter of RFM is binned into 3 groups by percentiles –
  - Bin 1 – 0 to 0.25 percentile
  - Bin 2 – 0.25 to 0.75 percentile
  - Bin 3 – 0.75 to 1 percentile

- Final 4 segments are created –
  - Champions – Star Performers
  - Silver – At Risk Customers
  - Gold – Level 2 Customers
  - Bronze – Lost Customers

Synopsis contd..

- Tools Used in this Project –
  - Python – For basic exploration, EDA and Time Series
  - KNIME – For Segmentation, RFM Analysis
  - Excel – For Summary Pivots and Views

Raw Sales Data

ORDERNUMBER	QUANTITYORDERED	PRICEEACH	ORDERLINENUMBER	SALES	ORDERDATE	DAYS_SINCE_LASTORDER	STATUS	PRODUCTLINE	MSRP	PRODUCTCODE	CUSTOMERNAME	PHONE	ADDRESSLINE1	CITY	POSTALCODE	COUNTRY	CONTACTLASTNAME	CONTACTFIRSTNAME	DEALSIZE
10107	30	95.70	2	2871	24/02/18	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	07/05/18	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	01/07/18	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	25/08/18	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	28/10/18	586	Shipped	Motorcycles	95	S10_1678	Technics Stores Inc.	6505556809	9408 Furth Circle	Burlingame	94217	USA	Hirano	Juri	Medium

Data Description – Numeric Variables

	count	mean	std	min	0.25	0.50	0.75	max
ORDERNUMBER	2747.00	10259.76	91.88	10100.00	10181.00	10264.00	10334.50	10425.00
QUANTITYORDERED	2747.00	35.10	9.76	6.00	27.00	35.00	43.00	97.00
PRICEEACH	2747.00	101.10	42.04	26.88	68.75	95.55	127.10	252.87
ORDERLINENUMBER	2747.00	6.49	4.23	1.00	3.00	6.00	9.00	18.00
SALES	2747.00	3553.05	1838.95	482.13	2204.35	3184.80	4503.10	14082.80
DAYS_SINCE_LASTORDER	2747.00	1757.09	819.28	42.00	1077.00	1761.00	2436.50	3562.00
MSRP	2747.00	100.69	40.11	33.00	68.00	99.00	124.00	214.00

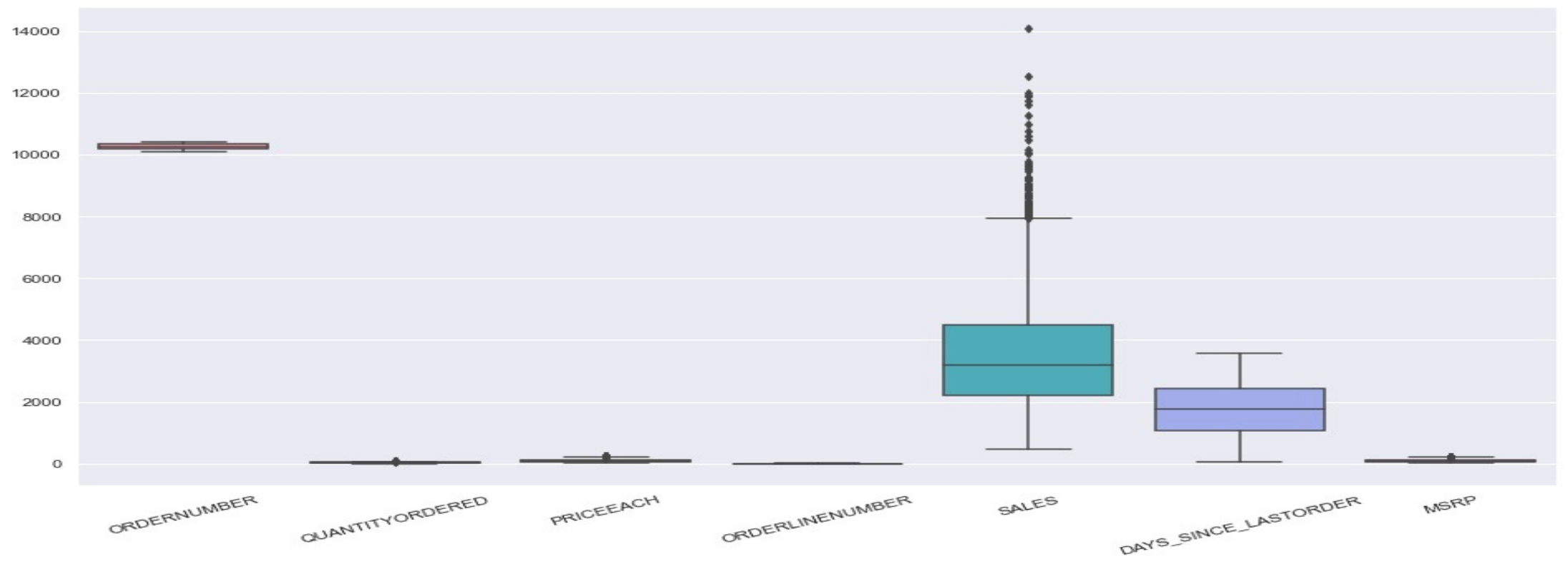
Data Description – Categorical Variables

	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

Summary Info

```
<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)
```

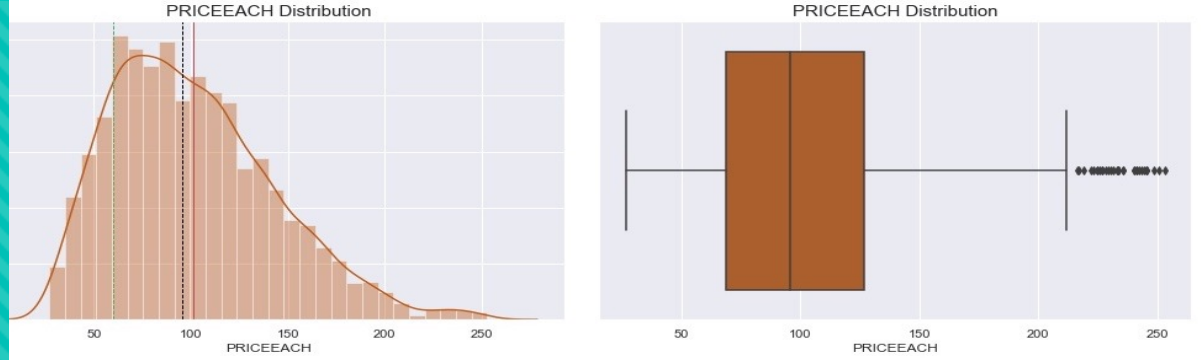
Data Exploration - Outliers

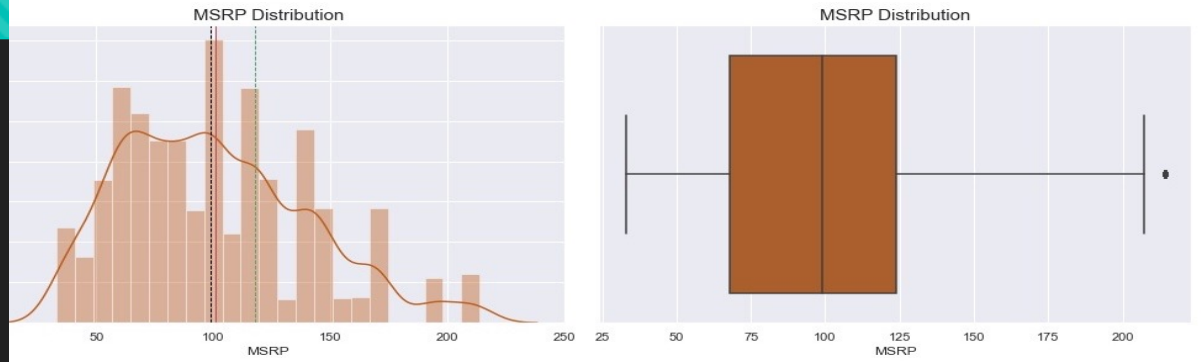


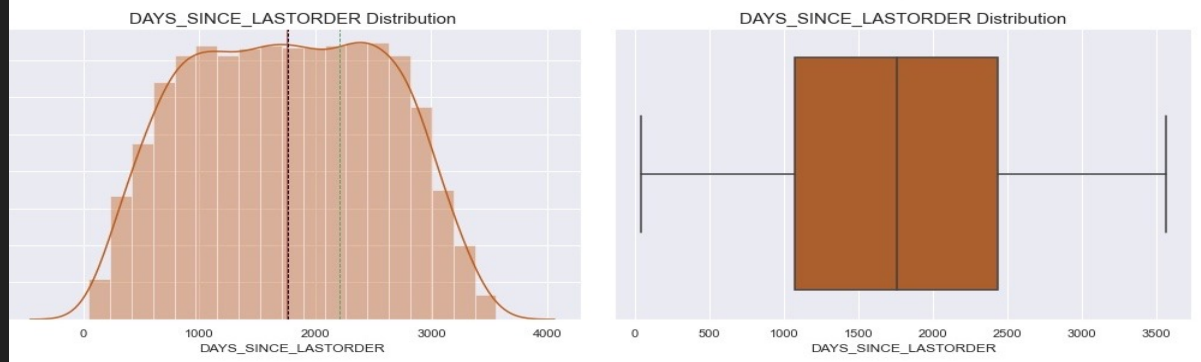
- Variables have Outliers – For this analysis we'll not treat them

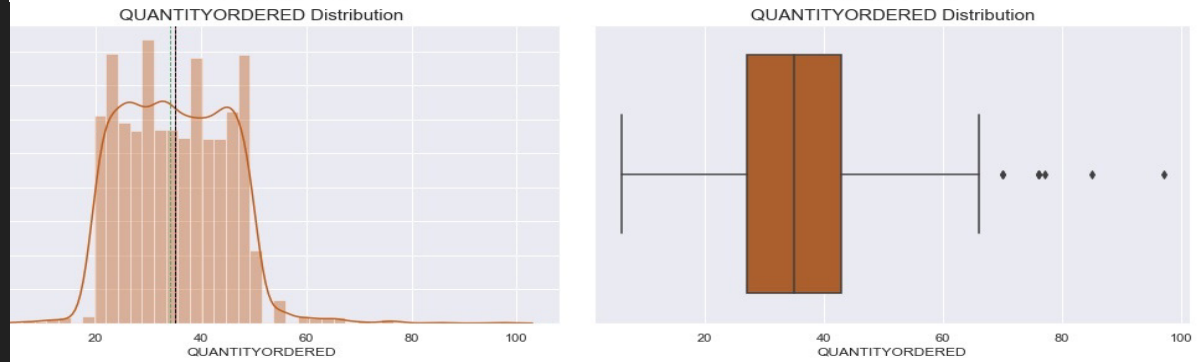
EDA - Univariate

- Histogram with KDE and Boxplots shown for -
  - QUANTITYORDERED
  - PRICEEACH
  - DAYS_SINCE_LAST_ORDER
  - MSRP
- All variables have approx normal distribution – indicating max data evenly distributed about its mean
- QUANTITYORDERED shows many outliers on the higher side



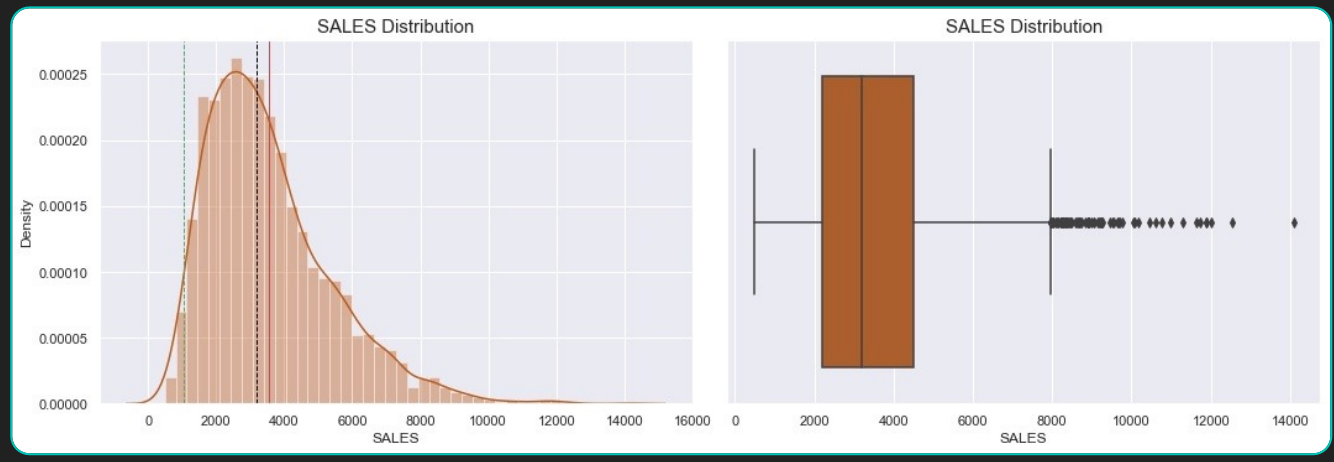






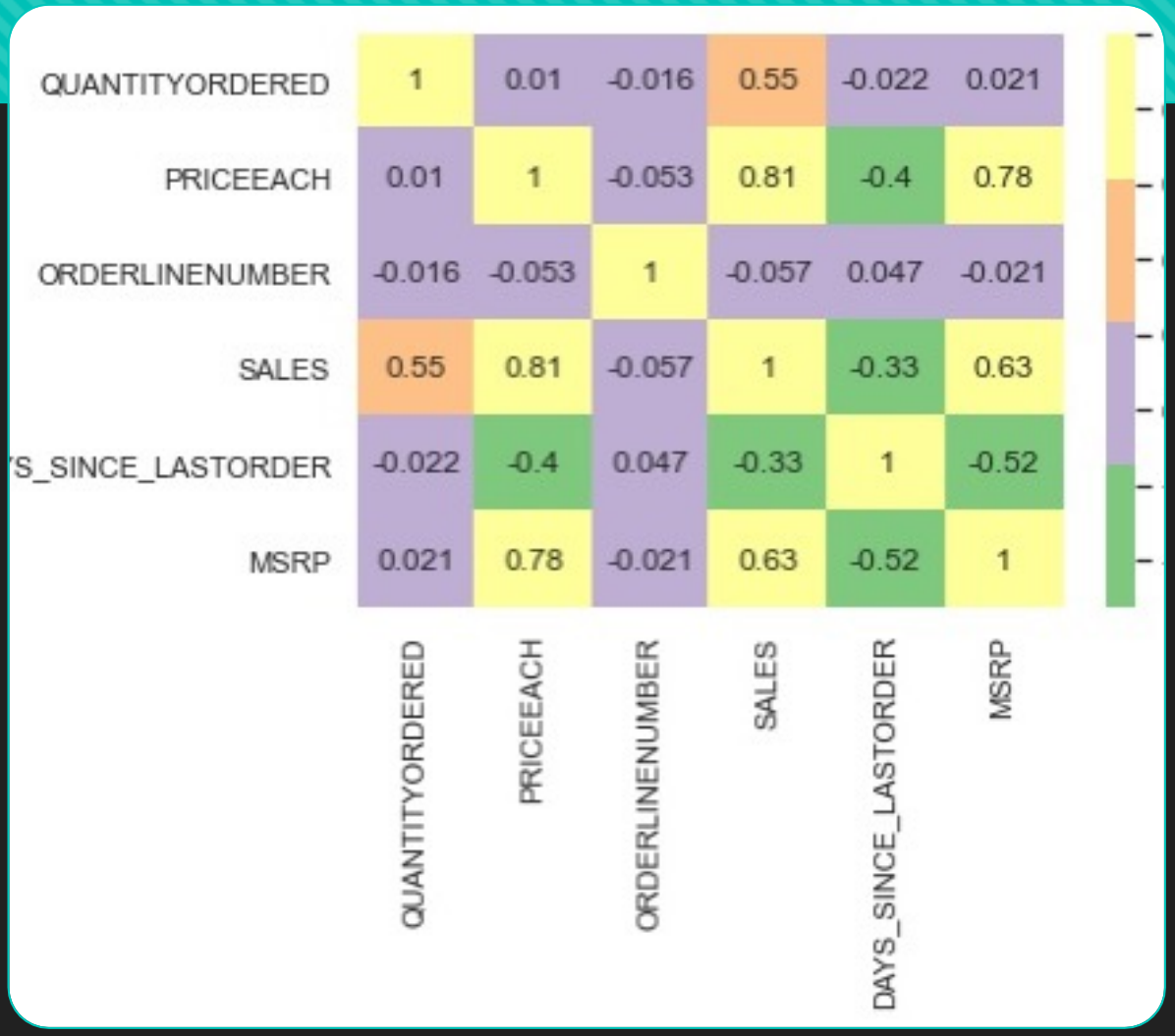
EDA – Univariate (SALES)

- SALES has a slight right skew - showing existence of many large values
- Coefficient of Variation = 51.76
  - This is high compared to others
  - Indicating large variation and inconsistency
- There are many Outliers on the higher side

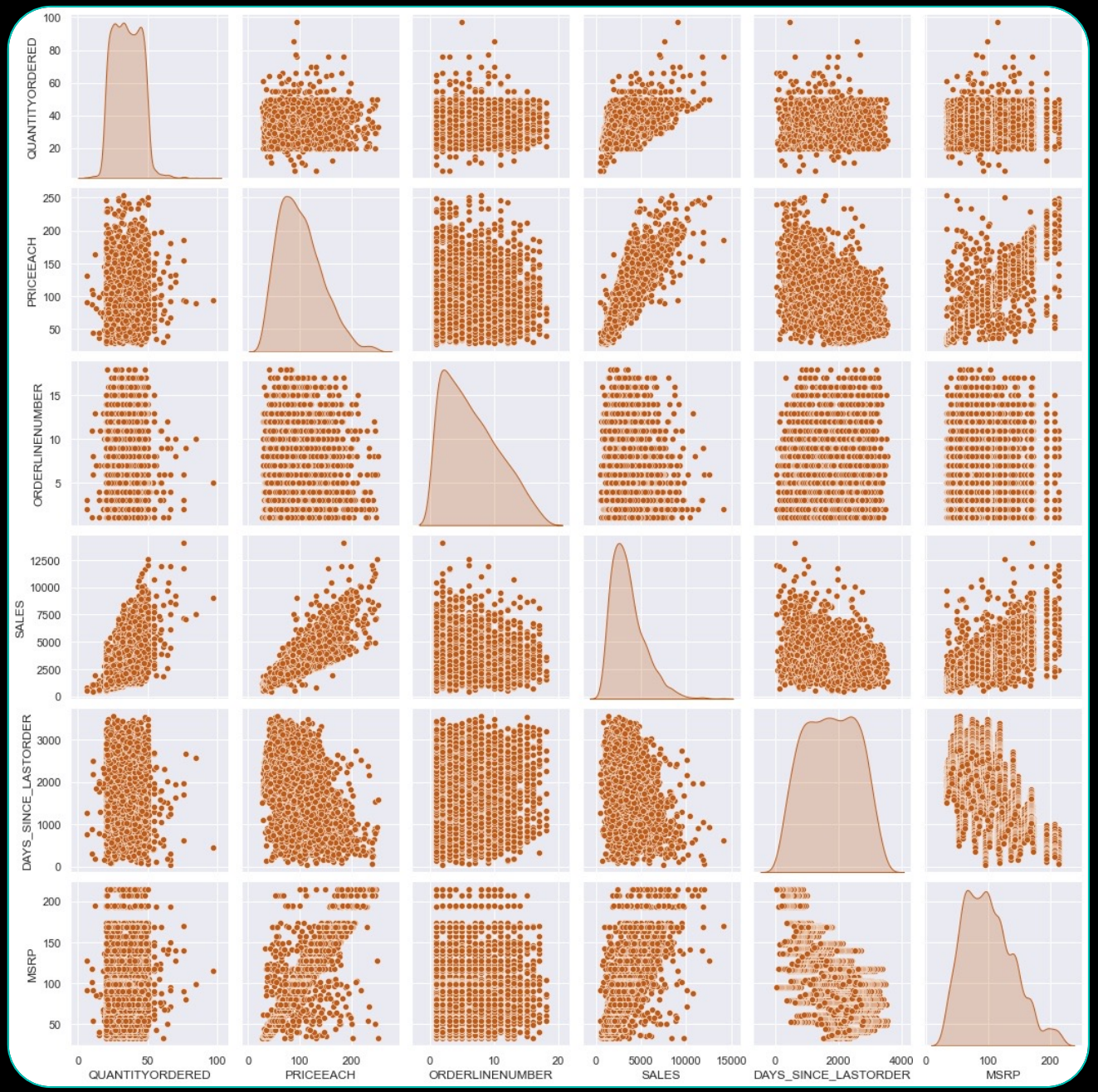


EDA – Correlation Heatmap

- High Correlation (0.81) between – SALES and PRICEEACH
- Also, High Correlation (0.78) between – MSRP and PRICEEACH
- The first one is obvious
- But second one indicates that the Actual price is very close to the suggested price by the manufacturers



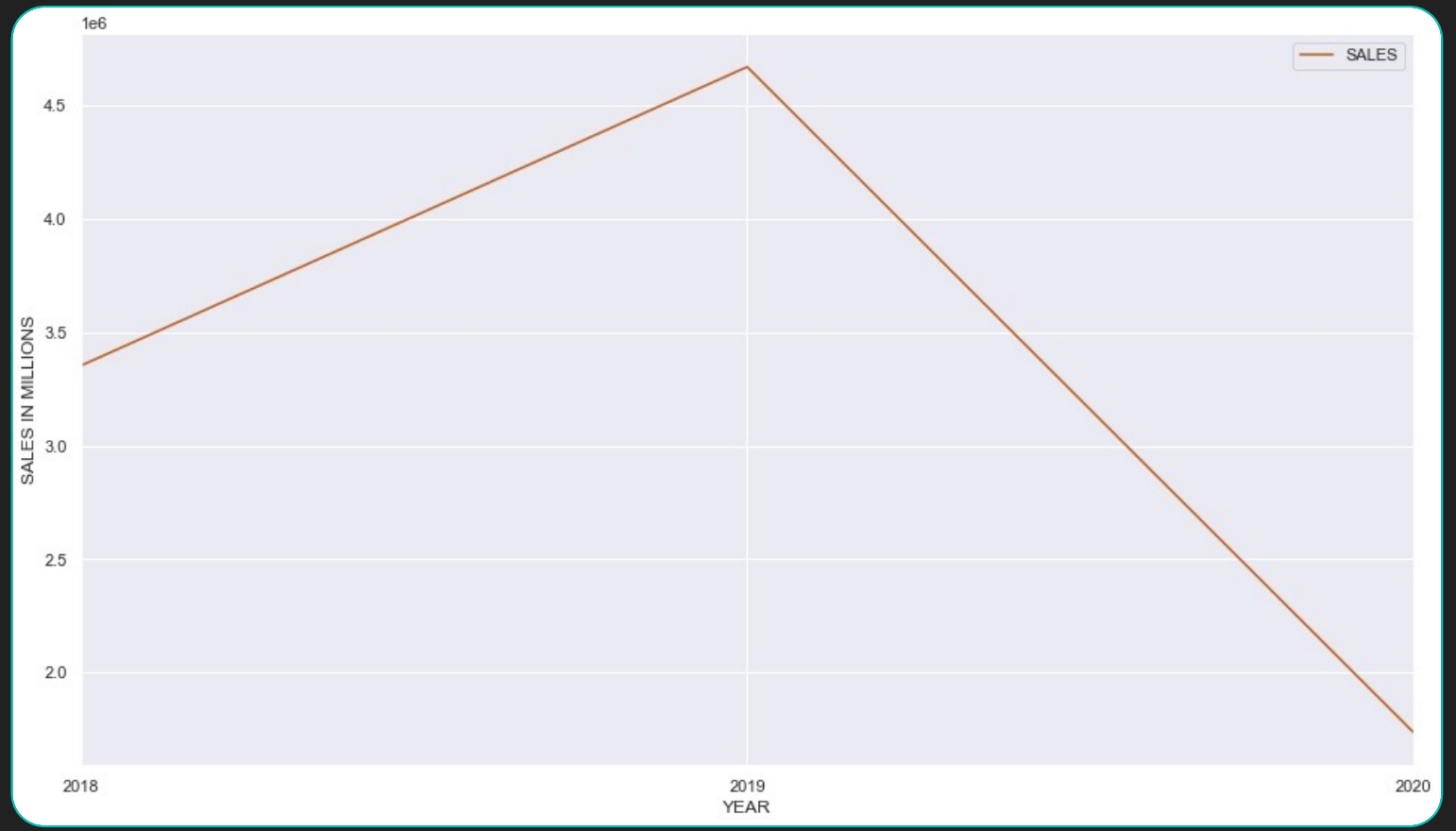
EDA - Pairplot



Trend in SALES - YEARLY

- Sales rise from 2018 to 2019

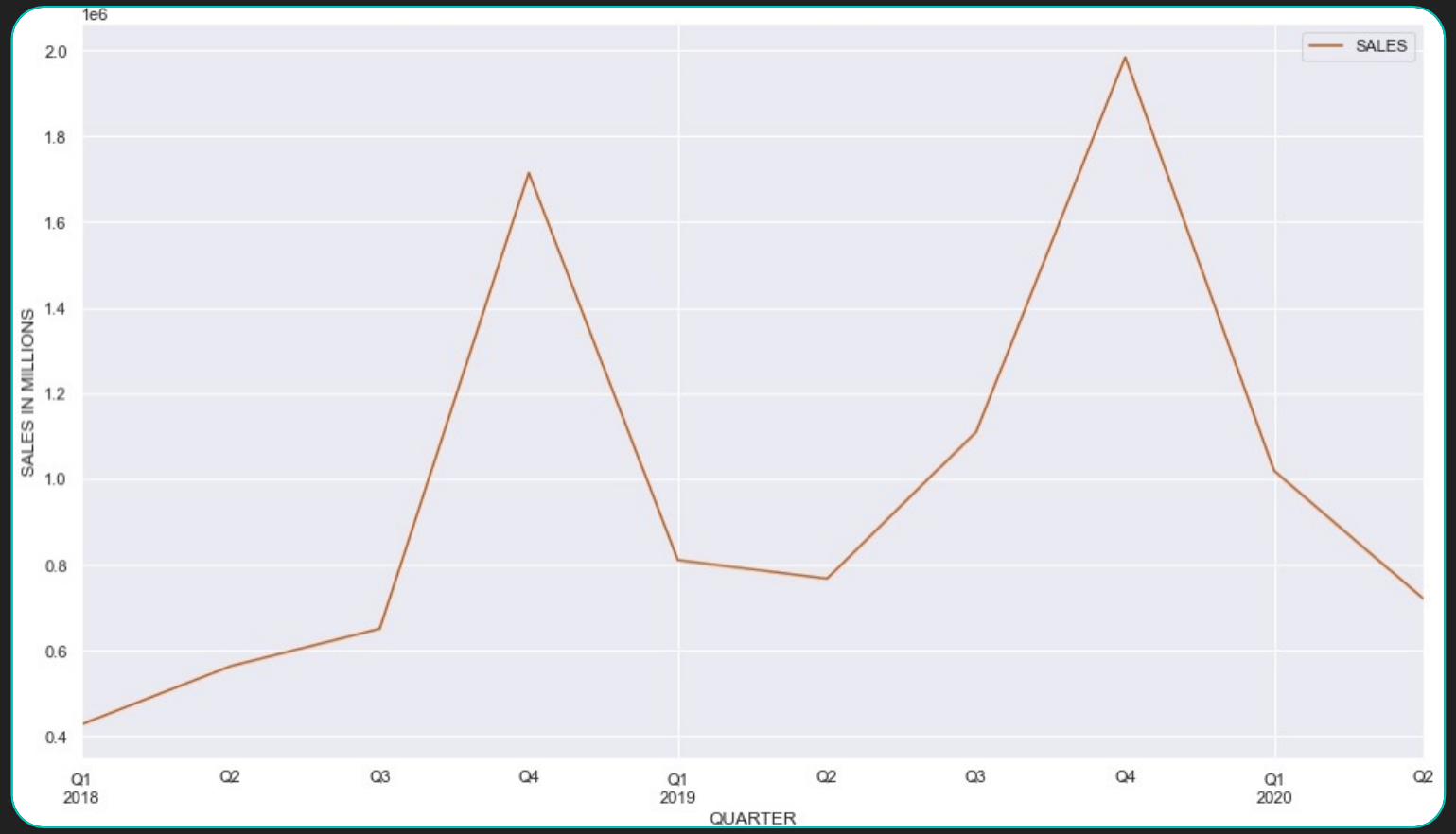
- Please Note - Data consists of full years of 2018, 2019 and only first 5 months of 2020



Trend in SALES - QUARTERLY

- Increasing trend year to year
- Seasonal spikes in Q4 of every year
- Seasonality seems Multiplicative

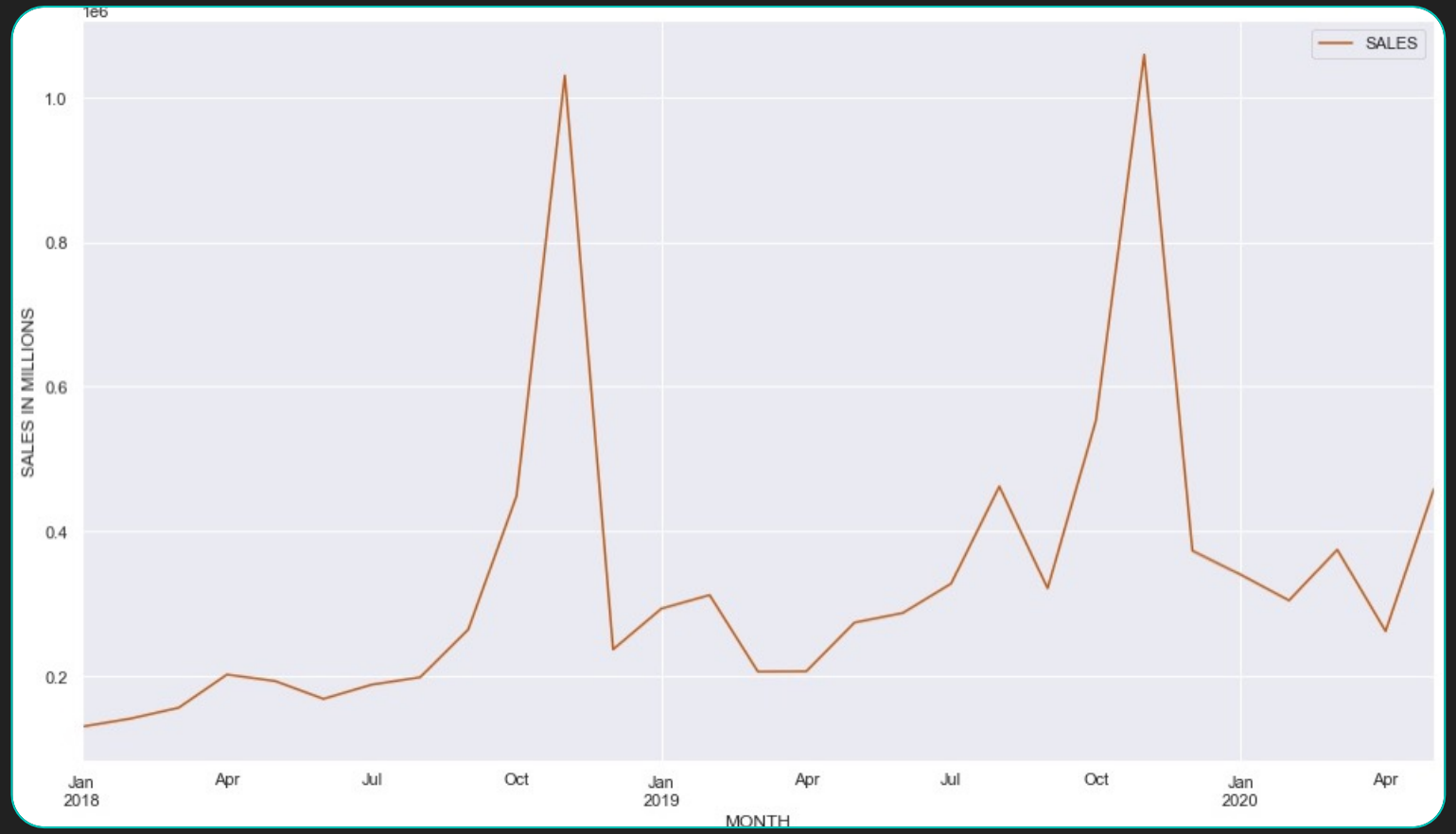
- Please Note - Data consists of full years of 2018, 2019 and only first 5 months of 2020



Trend in SALES - MONTHLY

- Every year, increasing trend from Jan with a spike in Oct-Nov
- Spike crashes around Dec end

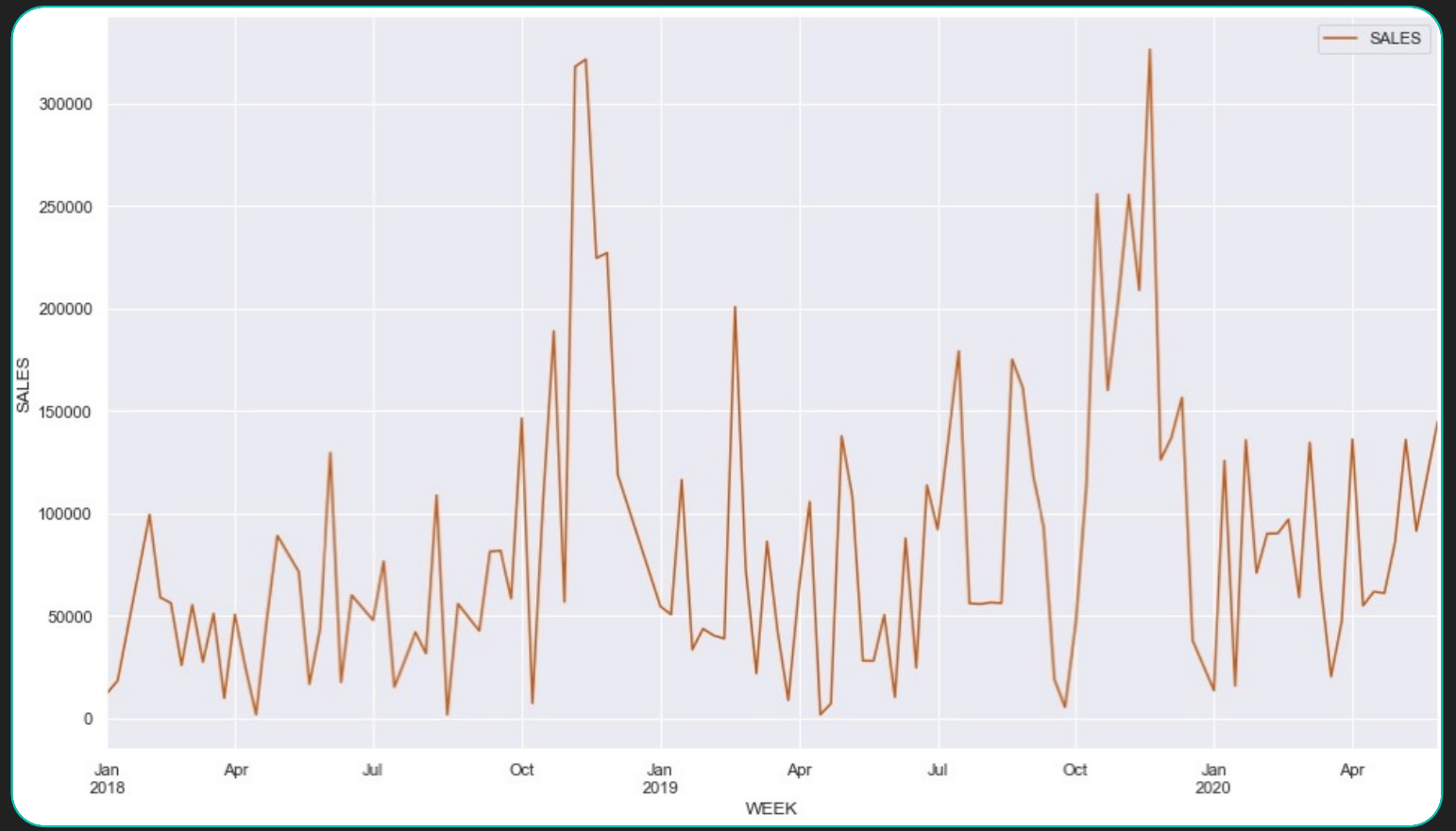
- Please Note - Data consists of full years of 2018, 2019 and only first 5 months of 2020



Trend in SALES - WEEKLY

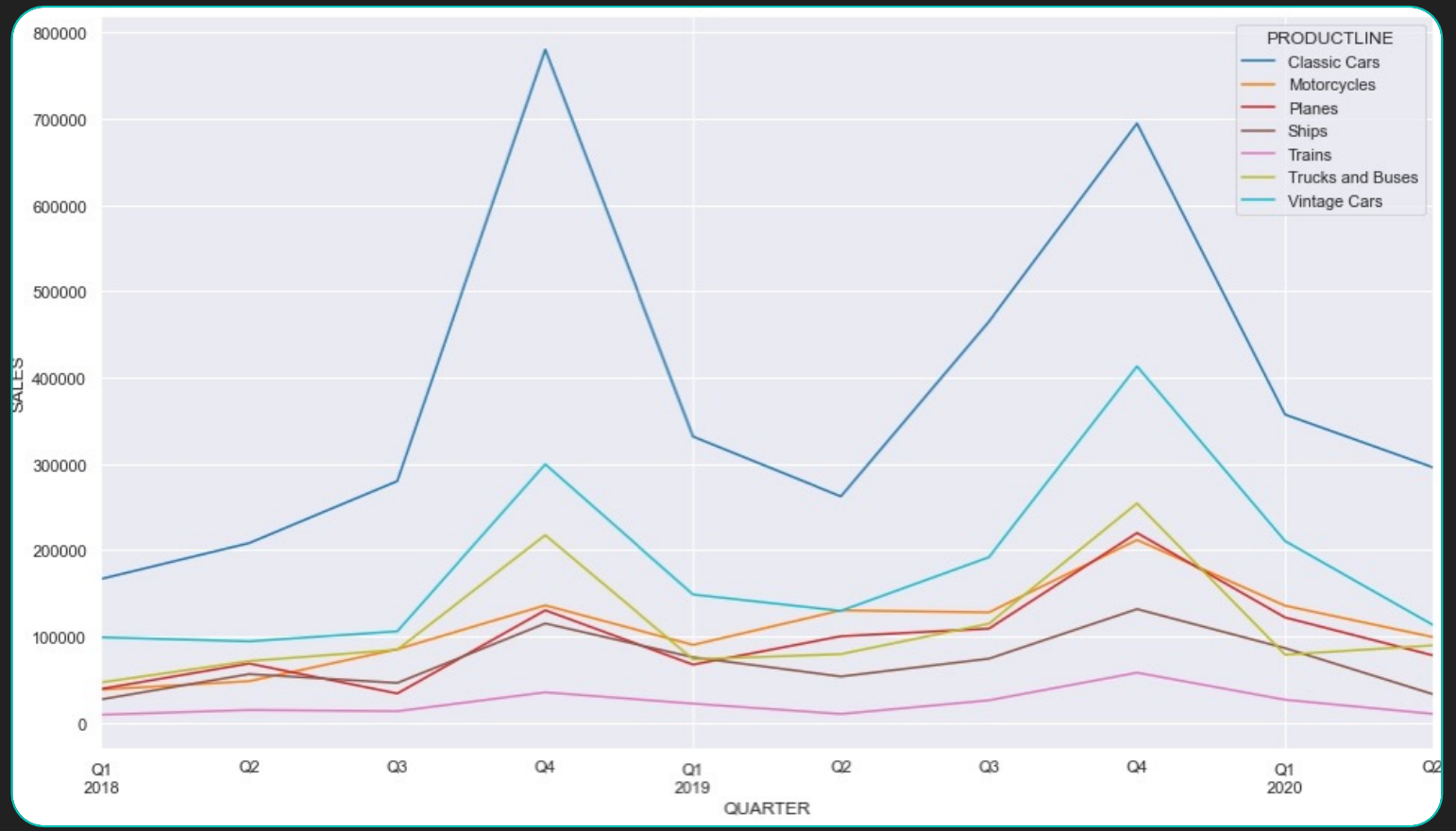
- Spike in Q4
- Graph consistent with earlier findings

- Please Note - Data consists of full years of 2018, 2019 and only first 5 months of 2020

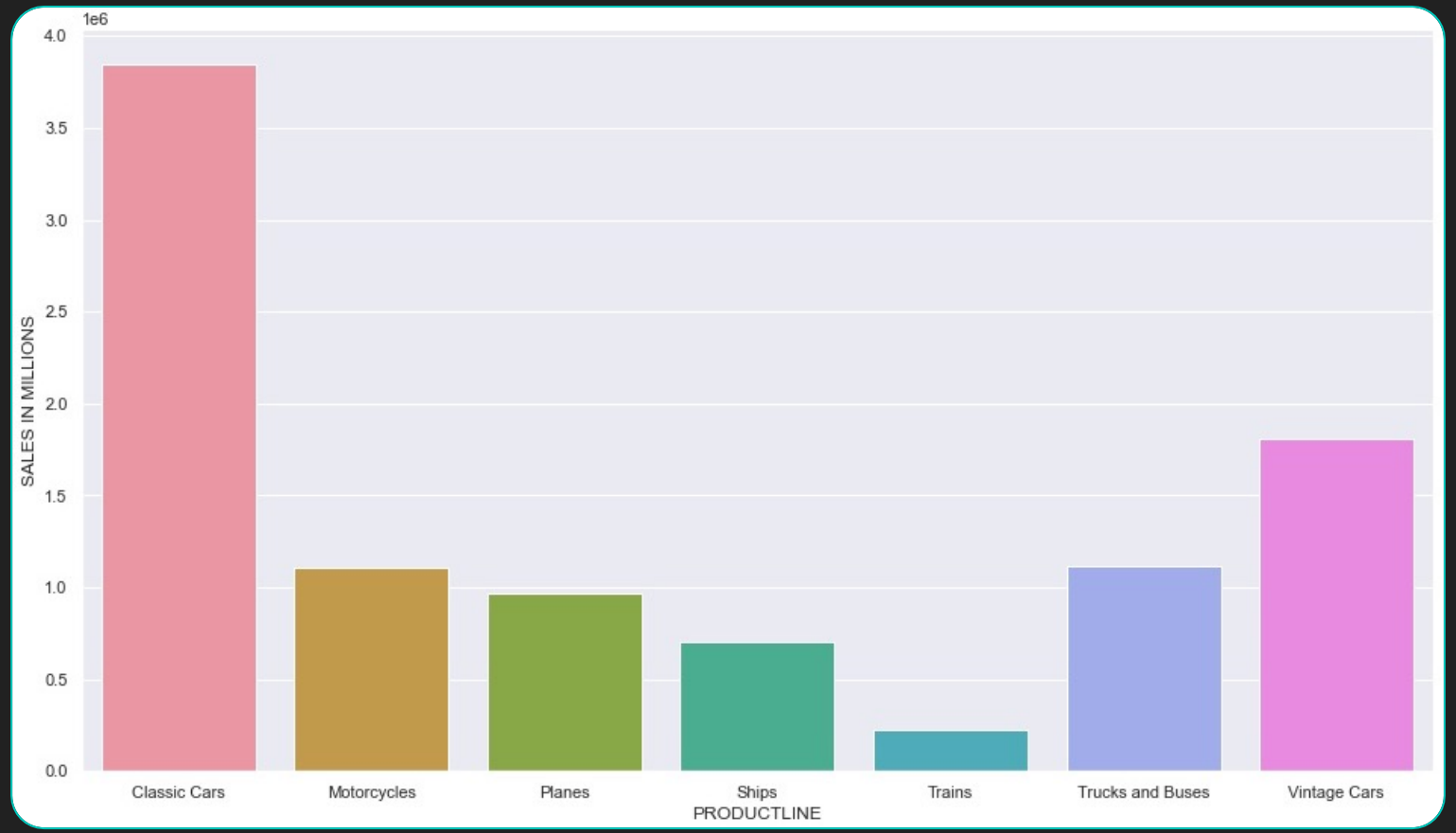


SALES – Product Category

- Total Sales are mainly driven by Classic Cars
- Q4 spikes is seen in all categories
- Pronounced Spike in Classic Cars compared to Other categories



SALES – Product Category contd..



- Only Classic Cars account for 40% of Total Sales
- Along with Vintage Cars – Top 2 contribute about 60 % to Total Sales

EDA - Summary

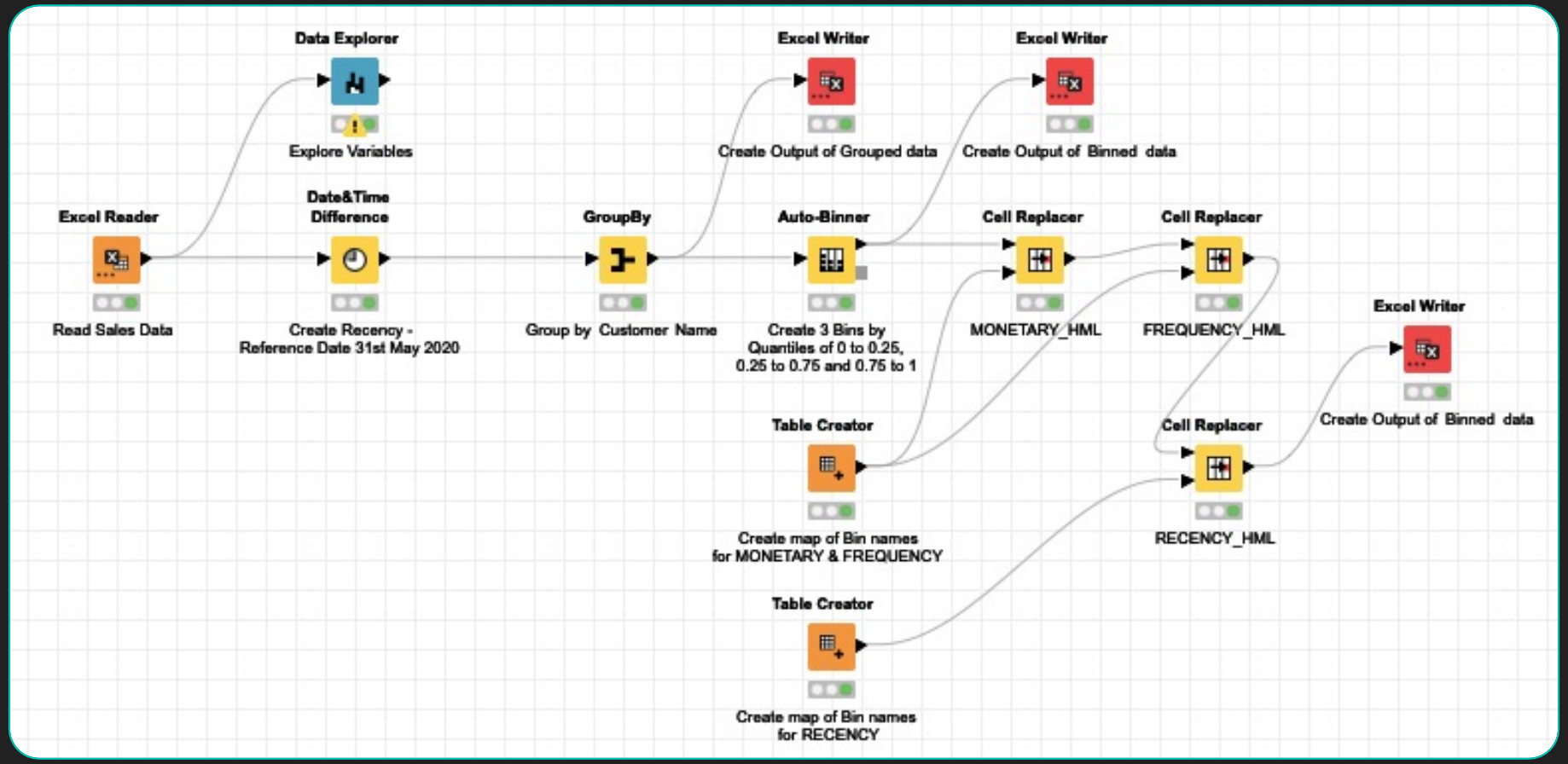
- Sales are Seasonal with spikes in Q4 every year
- Top 3 Categories by Sales are –
  - Classic cars
  - Vintage cars
  - Trucks and Buses
- Least Selling category – Trains
- Classic Cars account for 40% and top 3 together make 70% of Total Sales
- One-third of Total orders are from USA, with Spain and France following behind
- Also, One-third of Total Sales comes from USA
- 95% of Sales are of Medium and Small Deal-size
  - Shows a pattern of multiple retail orders
  - No Wholesale Bulk orders

Segmentation - RFM

- We create 4 segments using Recency – Frequency - Monetary (RFM) analysis
- We do segmentation on Customer Names
- Scores are binned by percentiles –
  - Bin 1 – 0 to 0.25
  - Bin 2 – 0.25 to 0.75
  - Bin 3 – 0.75 to 1
- Tool used – KNIME

- Recency –
  - Reference Date is last date in data – 31-May-20
- Frequency –
  - This is Total No. of Products (Items) per customer
- Monetary –
  - This is Total Sales by each Customer

KNIME Workflow - RFM



Final Output Table Head

CUSTOMERNAME	Count*(ORDERNUMBER)	Min*(RECENCY)	Last*(ORDERDATE)	Mode*(STATUS)	Mode*(PRODUCTLINE)	First*(PHONE)	First*(ADDRESSLINE1)	First*(CITY)	First*(POSTALCODE)	First*(COUNTRY)	First*(CONTACTLASTNAME)	First*(CONTACTFIRSTNAME)	Mode*(DEALSIZE)	Sum(SALES)	Count*(ORDERNUMBER) [Binned]	Min*(RECENCY) [Binned]	Sum(SALES) [Binned]	MONETARY_HML	FREQUENCY_HML	RECENCY_HML
AV Stores, Co.	51	196	2019-10-14	Shipped	Vintage Cars	(171) 555-1555	Fauntleroy Circus	Manchester	EC2 5NT	UK	Ashworth	Victoria	Medium	157807.81	Bin 3	Bin 2	Bin 3	HIGH	HIGH	MEDIUM
Alpha Cognac	20	64	2020-03-28	Shipped	Ships	61.77.6555	1 rue Alsace-Lorraine	Toulouse	31000	France	Roulet	Annette	Medium	70488.44	Bin 1	Bin 1	Bin 1	LOW	LOW	HIGH
Amica Models & Co.	26	265	2019-09-09	Shipped	Vintage Cars	011-4988555	Via Monte Bianco 34	Torino	10100	Italy	Accorti	Paolo	Small	94117.26	Bin 2	Bin 3	Bin 2	MEDIUM	MEDIUM	LOW
Anna's Decorations, Ltd	46	83	2018-11-04	Shipped	Classic Cars	02 9936 8555	201 Miller Street	North Sydney	2060	Australia	O'Hara	Anna	Small	153996.13	Bin 3	Bin 2	Bin 3	HIGH	HIGH	MEDIUM
Atelier graphique	7	188	2019-11-25	Shipped	Classic Cars	40.32.2555	54, rue Royale	Nantes	44000	France	Schmitt	Carine	Medium	24179.96	Bin 1	Bin 2	Bin 1	LOW	LOW	MEDIUM

RFM – Output Matrix

- RFM Segmentation is done on all 89 Customers
- Matrix shown is RFM Segregation of Customers as per –
  - High – Medium - Low

RECENCY	FREQUENCY	MONETARY		
		H	M	L
H	H	9	1	0
	M	1	8	1
	L	0	2	1
M	H	10	1	0
	M	1	21	1
	L	0	2	8
L	H	1	0	0
	M	0	7	0
	L	0	2	12

RFM Inference - Best Customers

- Best Customers are the ones with the highest score in each segment
- So, Best Customers have RFM scores –
  - High_High_High
- There are 9 Customers in this segment.
- Top 5 by Least Recency are given

CUSTOMER NAME	RECENCY	FREQUENCY	MONETARY
Euro Shopping Channel	0	259	912294.11
La Rochelle Gifts	0	53	180124.9
Mini Gifts Distributors Ltd.	2	180	654858.06
Souveniers And Things Co.	2	46	151570.98
Salzburg Collectables	14	40	149798.63

RFM Inference – At Risk of Churn

- Customers in the middle of the matrix, seem to have reduced our share of their wallet
- These Customers are at risk of churn
- We consider, At Risk Customers to have RFM scores –
  - Medium_Medium_Medium
- There are 21 customers in this segment
- Top 5 by Highest Recency are given

CUSTOMER NAME	RECENCY	FREQUENCY	MONETARY
Mini Classics	229	26	85555.99
Toms Spezialitten, Ltd	228	26	100306.58
Heintze Collectables	222	27	100595.55
Canadian Gift Exchange Network	222	22	75238.92
giftsbymail.co.uk	212	26	78240.84

RFM Inference – Lost Customers

- Customers in the bottom right corner have the least scores on all 3
- These Customers are Lost Customers
- Lost Customers have RFM scores –
  - Low_Low_Low
- There are 12 customers in this segment
- Top 5 by Highest Recency are given

CUSTOMER NAME	RECENCY	FREQUENCY	MONETARY
Double Decker Gift Stores, Ltd	495	12	36019.04
West Coast Collectables Co.	488	13	46084.64
Signal Collectibles Ltd.	476	15	50218.51
Daedalus Designs Imports	465	20	69052.41
CAF Imports	439	13	49642.05

RFM Inference – Loyal Customers

- Customers who frequent regularly with good Recency can be called as Loyal Customers
- So, Loyal Customers have –
  - High Frequency
  - High Recency
  - Any Monetary
- There are 10 customers in this segment
- Top 5 by Highest Frequency are given

CUSTOMER NAME	RECENCY	FREQUENCY	MONETARY
Euro Shopping Channel	0	259	912294.11
Mini Gifts Distributors Ltd.	2	180	654858.06
La Rochelle Gifts	0	53	180124.9
Souveniers And Things Co.	2	46	151570.98
Reims Collectables	62	41	135042.94

RFM Segmentation

We create 4 Segments based on RFM Scores

- Champions – Star Performers
- Silver – At Risk Customers
- Gold – Level 2 Customers
- Bronze – Lost Customers

SEGMENT 1 - CHAMPIONS

	COUNT%	SALES %
HIGH_HIGH_HIGH	10.11	26.96
HIGH_HIGH_MEDIUM	1.12	1.18
HIGH_MEDIUM_HIGH	1.12	1.25
MEDIUM_HIGH_HIGH	11.24	16.11
	23.60	**45.50**

SEGMENT 2 - GOLD

	COUNT%	SALES %
HIGH_MEDIUM_MEDIUM	8.99	7.68
MEDIUM_MEDIUM_HIGH	1.12	1.24
MEDIUM_HIGH_MEDIUM	1.12	1.12
MEDIUM_MEDIUM_MEDIUM	23.60	19.92
	34.83	**29.95**

RFM Segmentation contd…

SEGMENT 3 – SILVER – AT RISK

	COUNT%	SALES %
HIGH_LOW_HIGH	0.00	0.00
HIGH_LOW_MEDIUM	2.25	1.63
HIGH_LOW_LOW	1.12	0.72
HIGH_HIGH_LOW	0.00	0.00
HIGH_MEDIUM_LOW	1.12	0.66
MEDIUM_LOW_HIGH	0.00	0.00
MEDIUM_HIGH_LOW	0.00	0.00
LOW_HIGH_HIGH	1.12	1.46
LOW_HIGH_MEDIUM	0.00	0.00
LOW_MEDIUM_HIGH	0.00	0.00
LOW_MEDIUM_MEDIUM	7.87	6.84
MEDIUM_MEDIUM_LOW	1.12	0.69
	14.61	**12.01**

SEGMENT 4– BRONZE - LOST

	COUNT%	SALES %
MEDIUM_LOW_MEDIUM	2.25	1.57
MEDIUM_LOW_LOW	8.99	3.03
LOW_LOW_HIGH	0.00	0.00
LOW_LOW_MEDIUM	2.25	1.54
LOW_HIGH_LOW	0.00	0.00
LOW_MEDIUM_LOW	0.00	0.00
LOW_LOW_LOW	13.48	6.40
	26.97	**12.54**

Inference – Insights - Recommendations

- Segment 1 – CHAMPIONS
  - These are the Champion Star performers
  - This segment has High on at least 2 parameters
  - There are 21 customers in this segment
  - These customers account for 45% of Total Sales
  - They need to be nurtured and preserved
  - They should be given first and exclusive access to new offers and new products
  - Dedicated Relationship Managers should be assigned to them
  - These RMs should periodically visit or call to check on customer needs and take feedback
- Segment 2 - GOLD
  - These are Level 2 grade, yet important customers
  - This segment has Medium on atleast 2 parameters
  - There are 31 customers in this segment
  - These customers account for 30% of Total Sales
  - They need to be incentivised to spend more and more frequently
  - Specially customised offers in their product category would help
  - RMs should call or visit them once to understand any issues faced
  - Increasing product range to suit their needs, should be considered

Inference – Insights - Recommendations

- Segment 3 – SILVER – AT RISK
  - These are At-Risk customers
  - This segment has at-least 1 Low
  - There are 13 customers in this segment
  - They account for 12% of Total Sales
  - They need attention
  - Loss Leader offers in 1 principal category might help
  - Mass mailers with – New Offers, New Arrivals, your company achievements – to refresh your company in their memory – should be considered
- Segment 2 – BRONZE - LOST
  - These are Lost customers
  - There are 24 customers in this segment
  - They account for 12% of Total Sales
  - These are relatively small numbers in terms of Sales
  - These customers should be used to identify mistakes and learn from them
  - A detailed Feedback form should be filled from them about their whole experience
  - One time Cash Offer or Free shipping can be offered to incentivise them to give feedback

Thank You

Marketing and Retail Analytics

Milestone 1 Project

By

CHETAN DUDHANE

PGP DSBA JULY B GRP 2