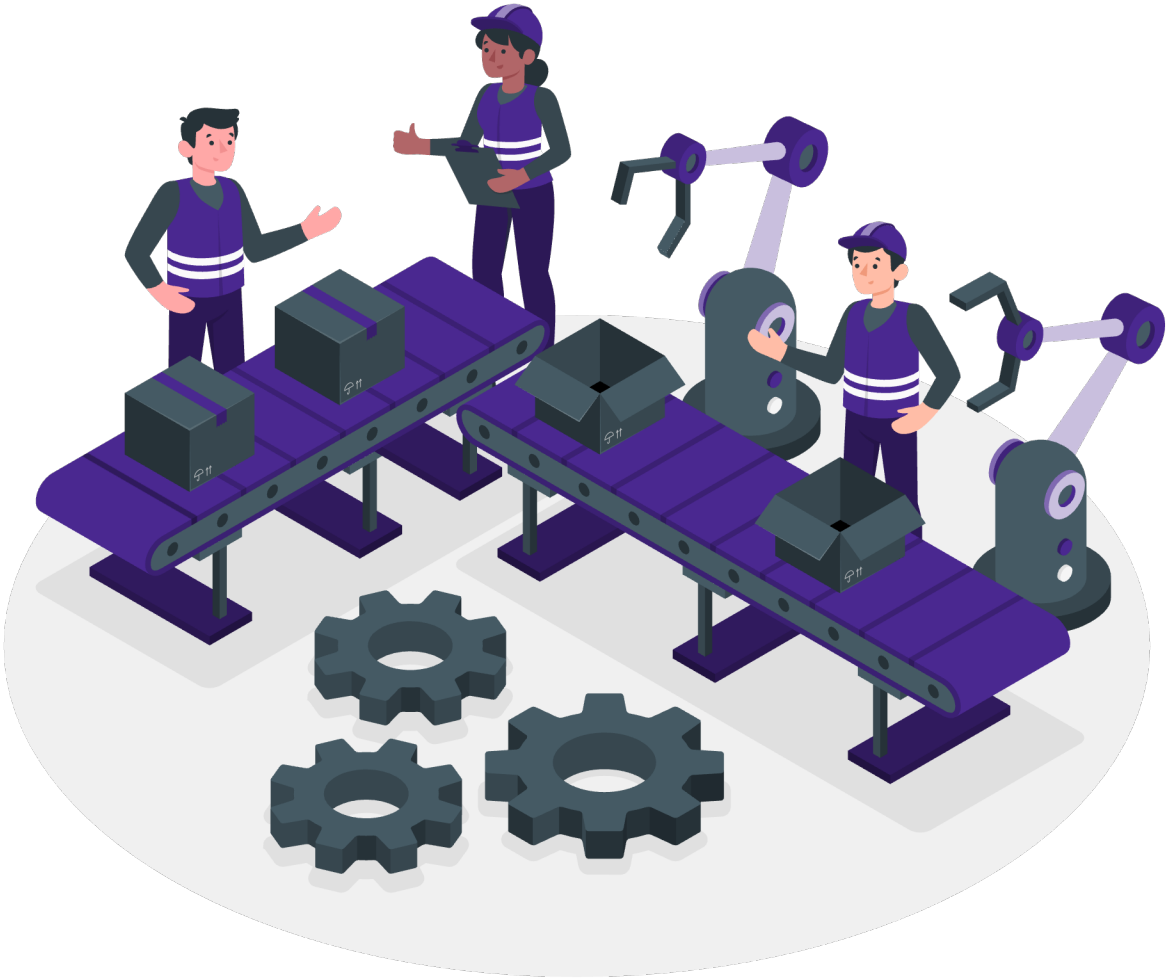
MRA Project - Milestone 1



Ankita Mishra

PGP - DSBA

Agenda

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 us as their consultant. Our job is to use the magical data science skills to provide them with suitable insights about their data and their customers.

Data Dictionary

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		

Data Summary

Data Info

```
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
```

Shape Info – (2747, 20)

Data Description

	ORDERNUMBER	QUANTITYORDERED	PRICEEACH	ORDERLINENUMBER	SALES	DAYS_SINCE_LASTORDER	MSRP
count	2747.000000	2747.000000	2747.000000	2747.000000	2747.000000	2747.000000	2747.000000
mean	10259.761558	35.103021	101.098951	6.491081	3553.047583	1757.085912	100.691664
std	91.877521	9.762135	42.042548	4.230544	1838.953901	819.280576	40.114802
min	10100.000000	6.000000	26.880000	1.000000	482.130000	42.000000	33.000000
25%	10181.000000	27.000000	68.745000	3.000000	2204.350000	1077.000000	68.000000
50%	10264.000000	35.000000	95.550000	6.000000	3184.800000	1761.000000	99.000000
75%	10334.500000	43.000000	127.100000	9.000000	4503.095000	2436.500000	124.000000
max	10425.000000	97.000000	252.870000	18.000000	14082.800000	3562.000000	214.000000

	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

Skewness

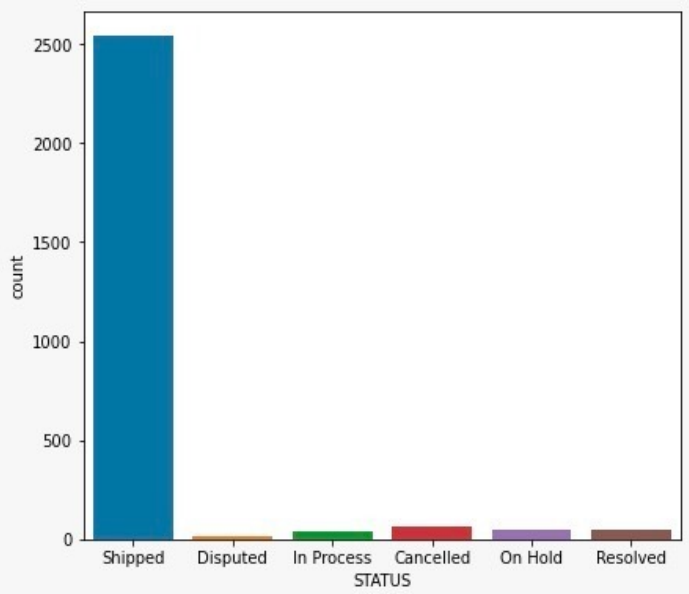
```
ORDERNUMBER          -0.006995
DAYS_SINCE_LASTORDER -0.002983
QUANTITYORDERED       0.369286
ORDERLINENUMBER       0.575327
MSRP                  0.575646
PRICEEACH             0.697222
SALES                 1.155940
dtype: float64
```

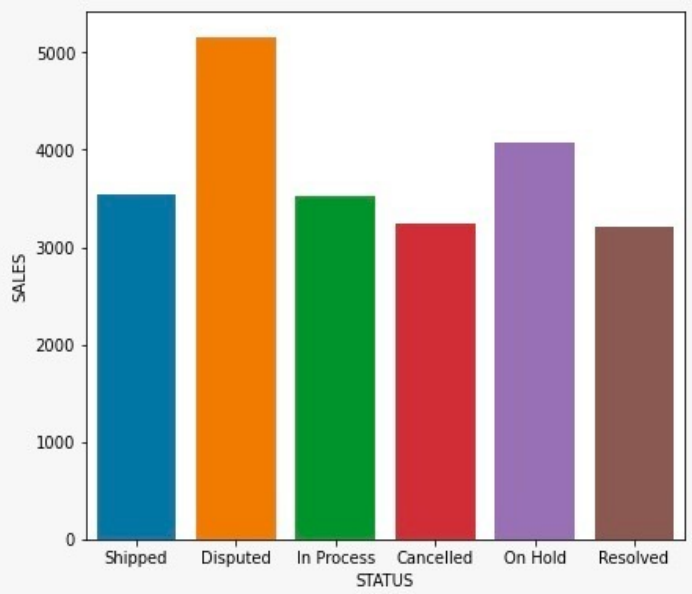
Data Summary

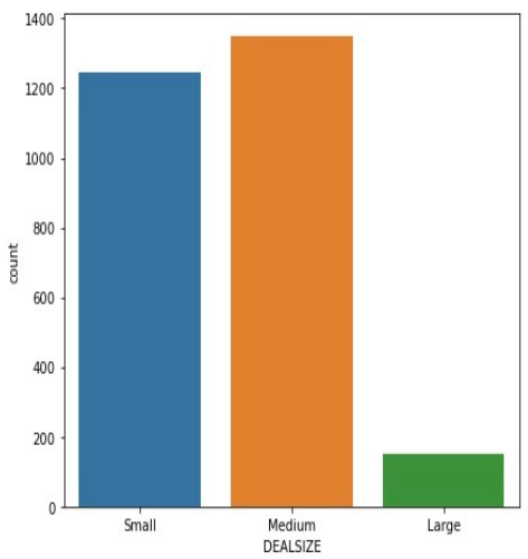
Data Inference

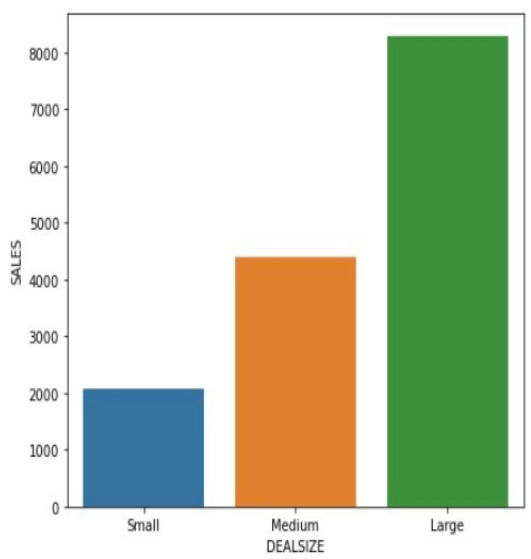
- The datatypes of the variables/columns along with the respective count are: datetime(1), float(2), int(5), object(12)
- There are no null and missing values in the dataset.
- The dataset does not have any duplicate values as well
- ORDERNUMBER, DAYS_SINCE_LASTORDER are negatively skewed (left-skewed)
- While ORDERLINENUMBER, MSRP, PRICEEACH are slightly right skewed. SALES is highly right skewed
- And finally, if the value is between -0.5 and 0.5, we consider the distribution to be approximately symmetric which is the case for QUANTITYORDERED

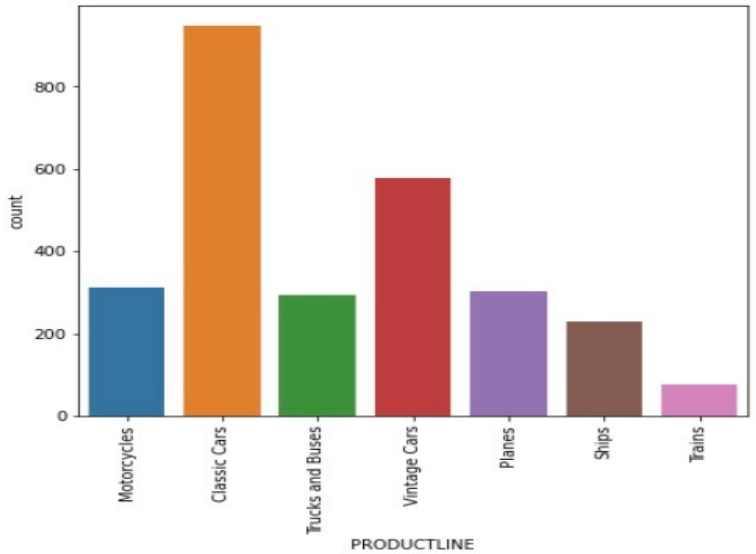
Exploratory Data Analysis - Univariate

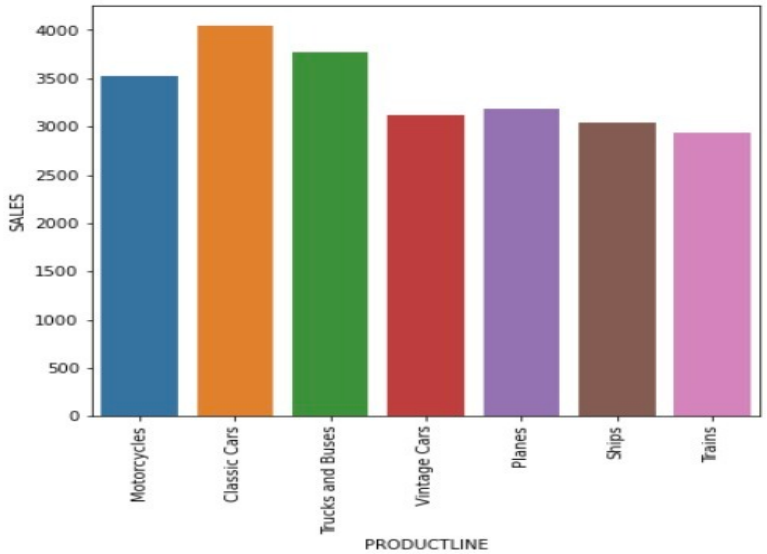






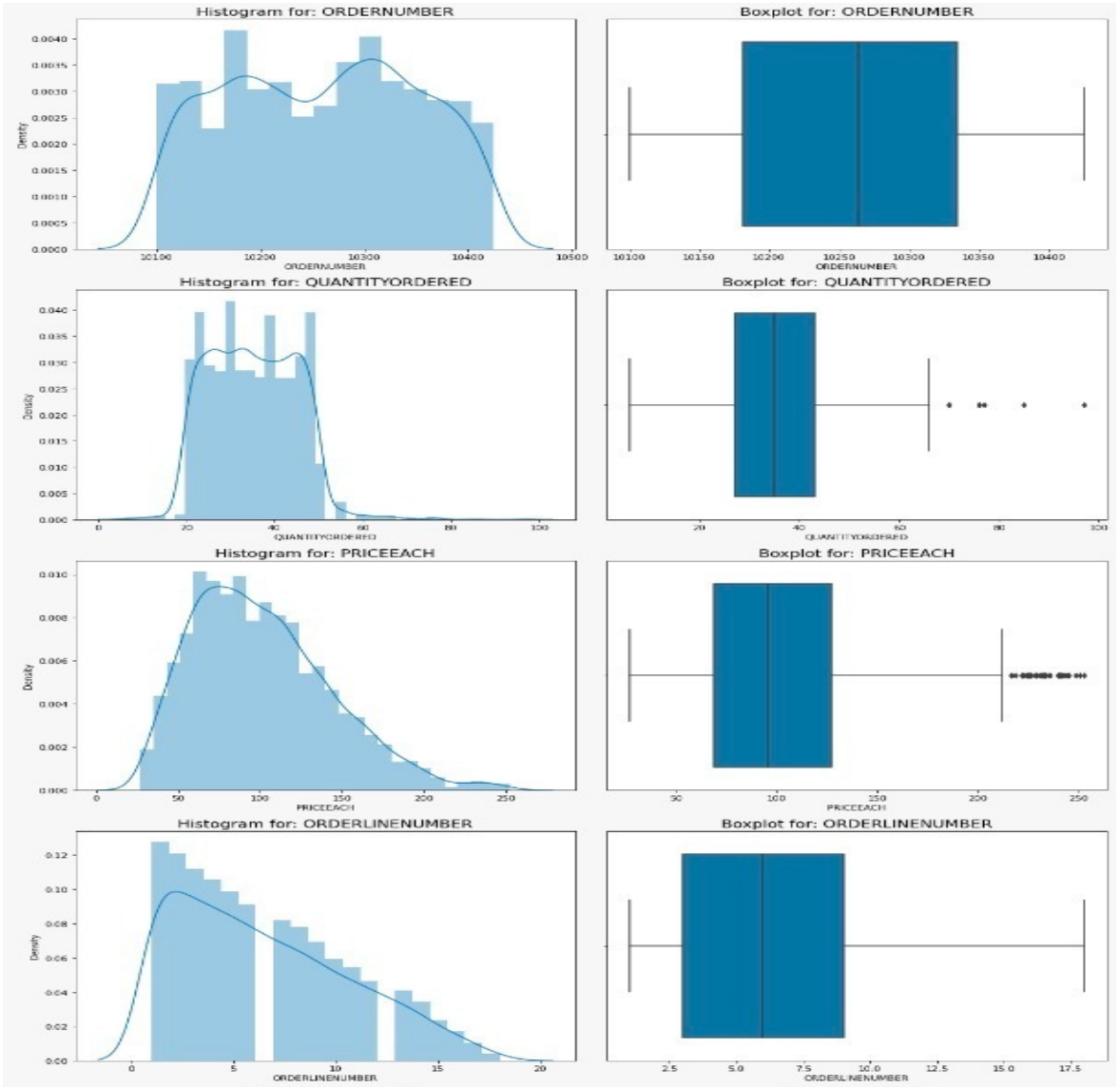


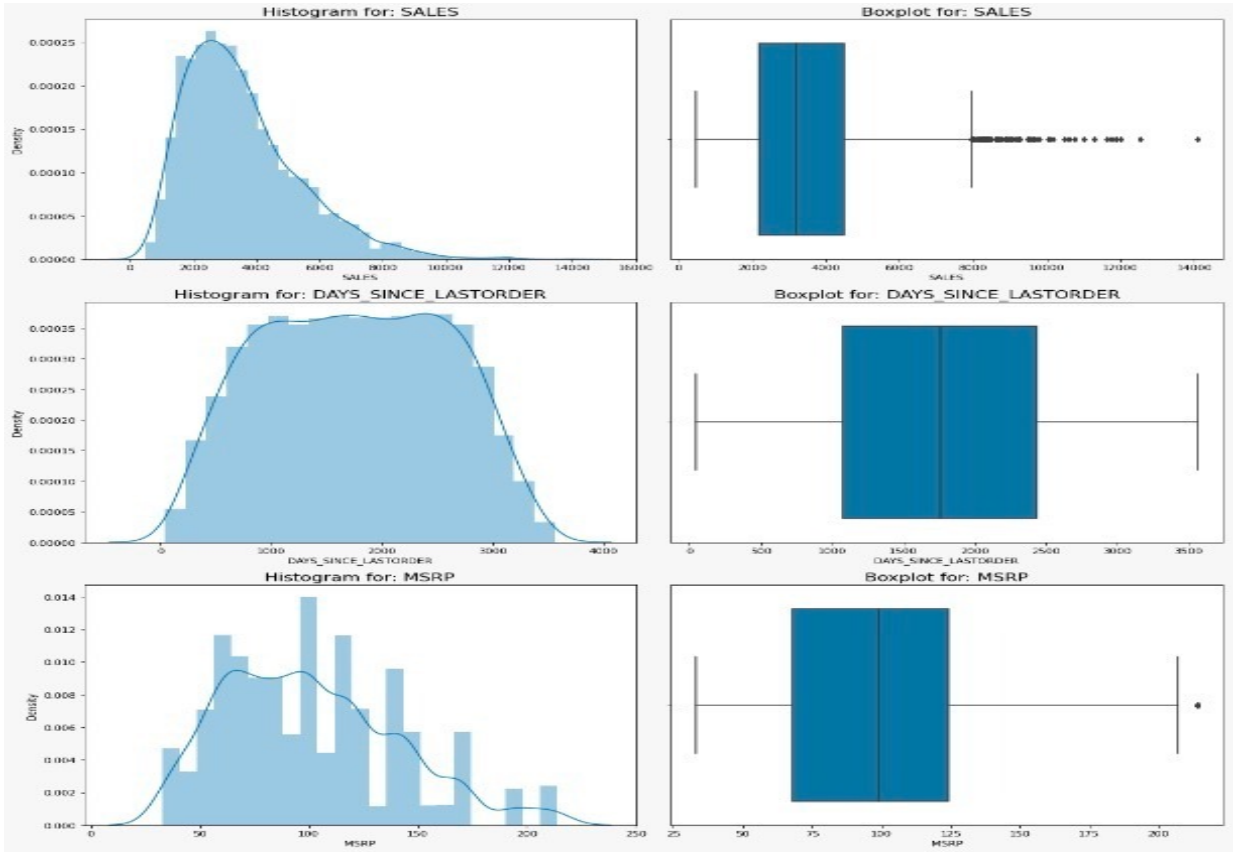




From the above univariate analysis against Sales it can be inferred that the number of products Shipped are more but the amount of Sales is highest for the Order Status of Disputed products. The Sales for Classic Cars is the highest with the most number of Sale while the Trucks and Buses have lower number of Sales but the amount is second highest in terms of Sale. Similarly, the amount for Large Dealsize is highest even the the count is the lowest among others.

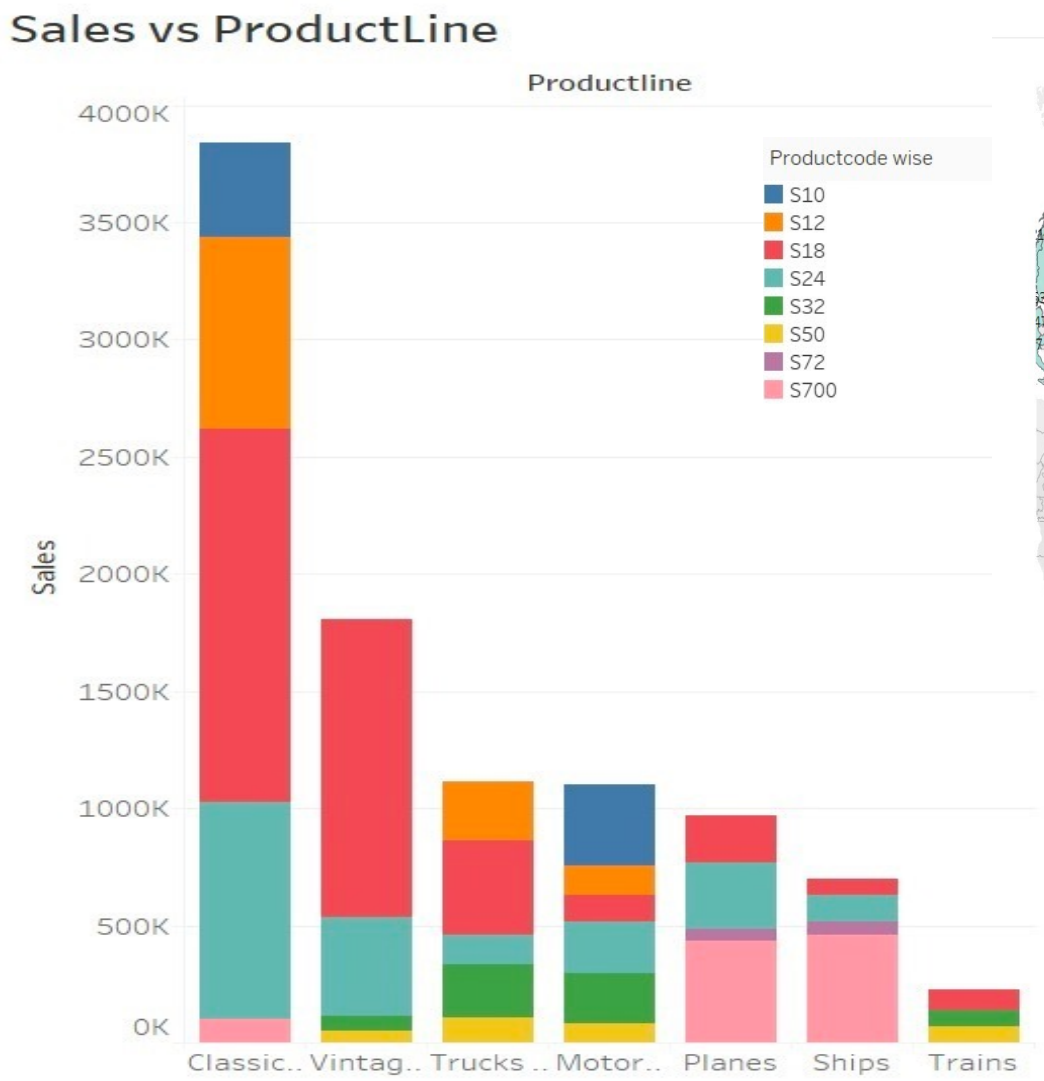
Exploratory Data Analysis - Univariate

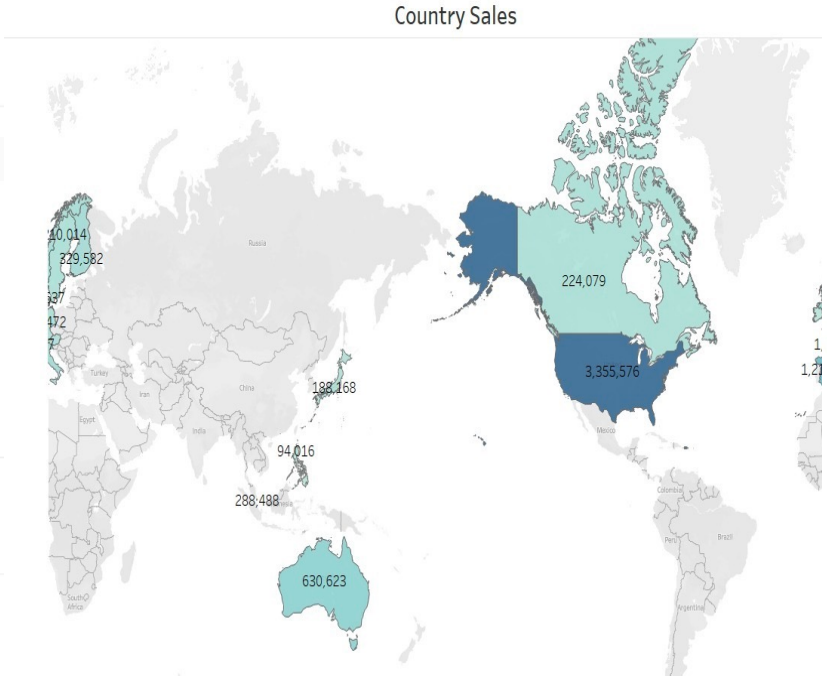


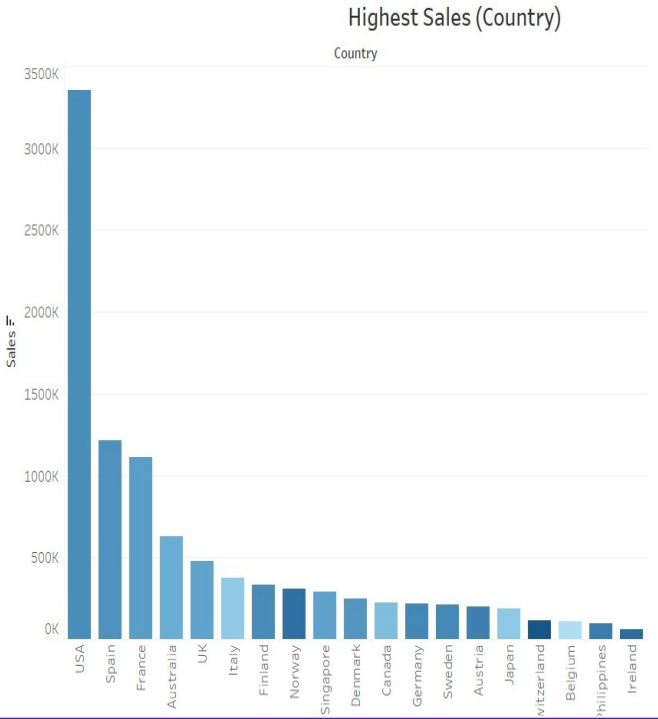


A type of distribution in which most values are clustered around the left tail of the distribution while the right tail of the distribution is longer and vice-versa can be inferred from Histogram and in Boxplot if external points are present outside these 1.5*IQR values are flagged as outliers.

Exploratory Data Analysis - Bivariate

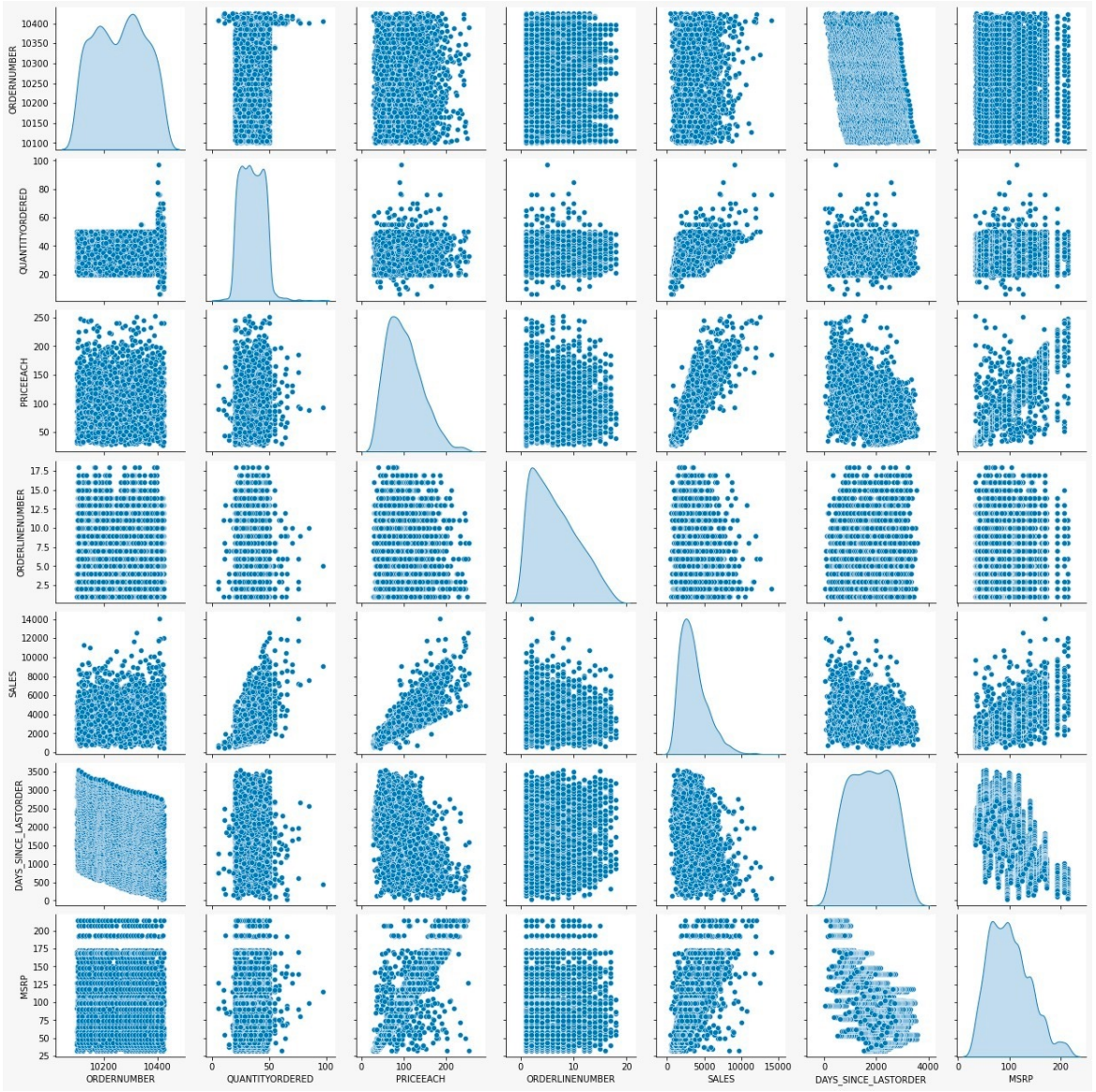


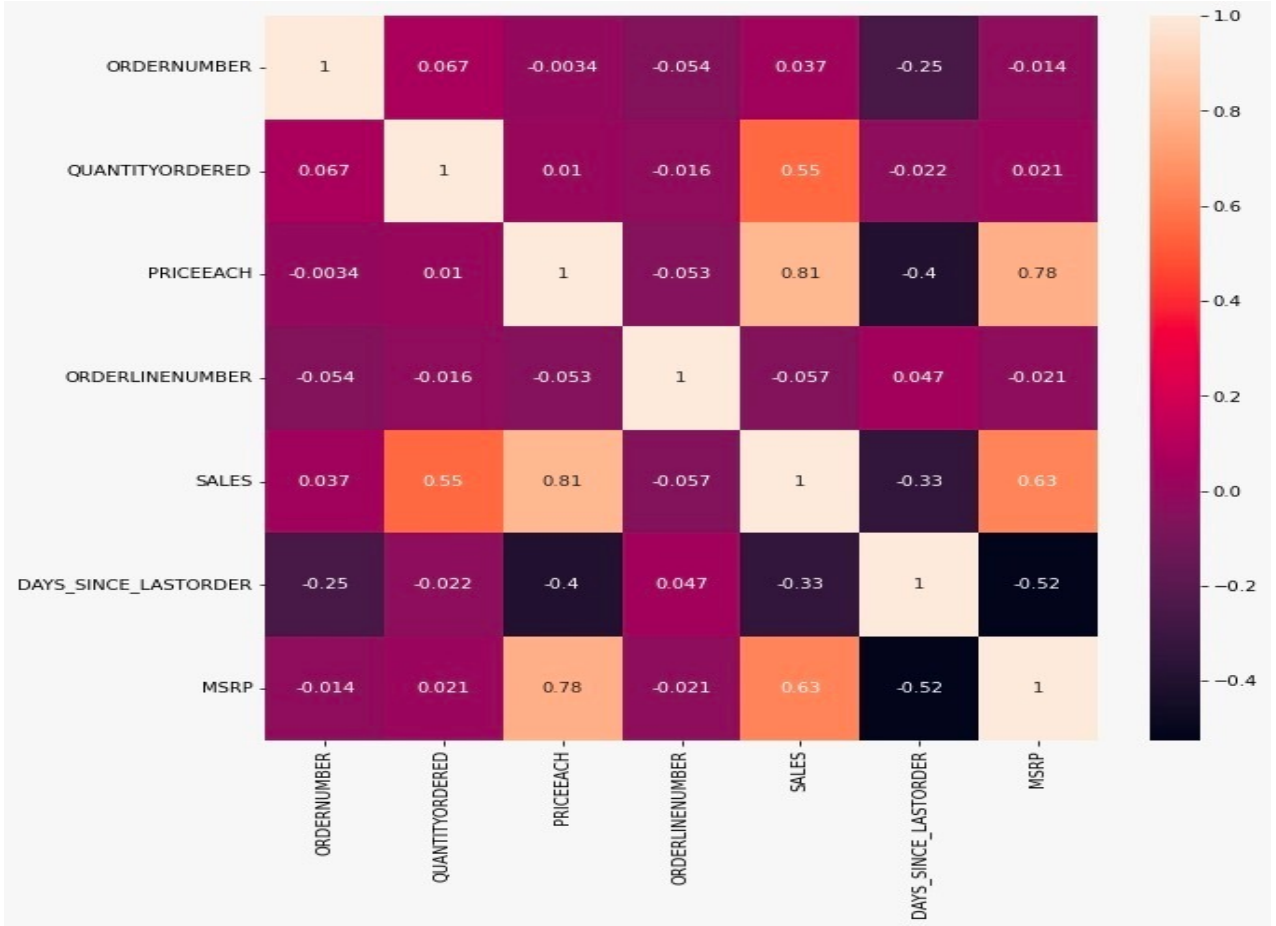




The highest SALES for the PRODUCTLINE is for Classic Cars. The Sales vs ProductLine has been further categorized using PRODUCTCODE. The USA has the highest SALES followed by Spain and France. Color coding has been done on the basis of median SALES which is the highest for Switzerland. Using average SALES instead of median shows different figures which indicates the existence of outliers.

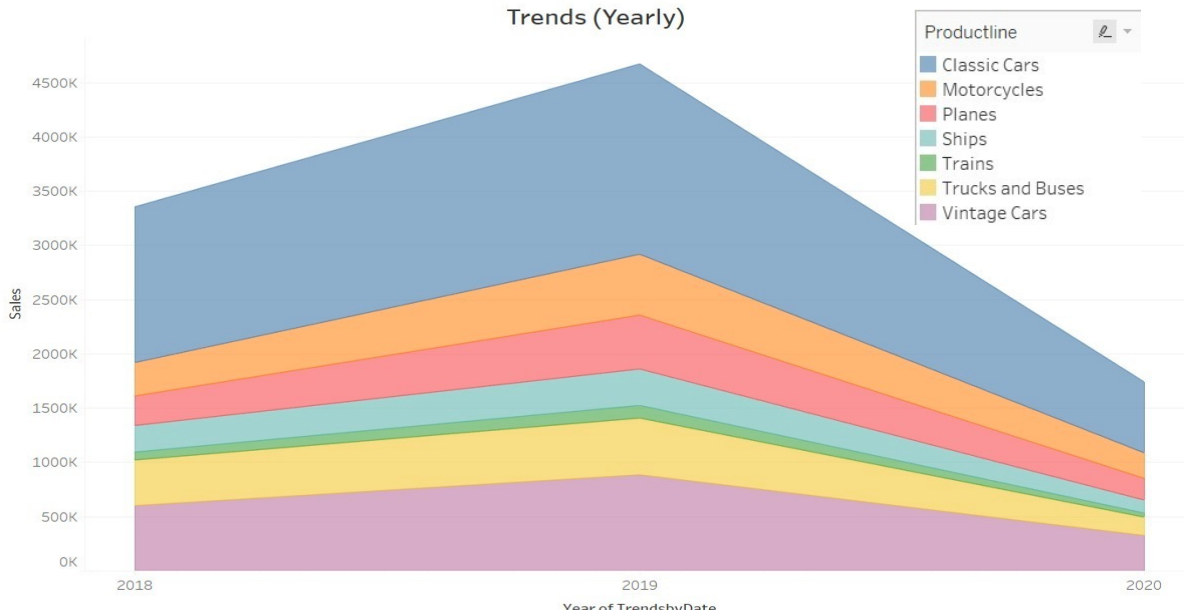
Exploratory Data Analysis - Multivariate

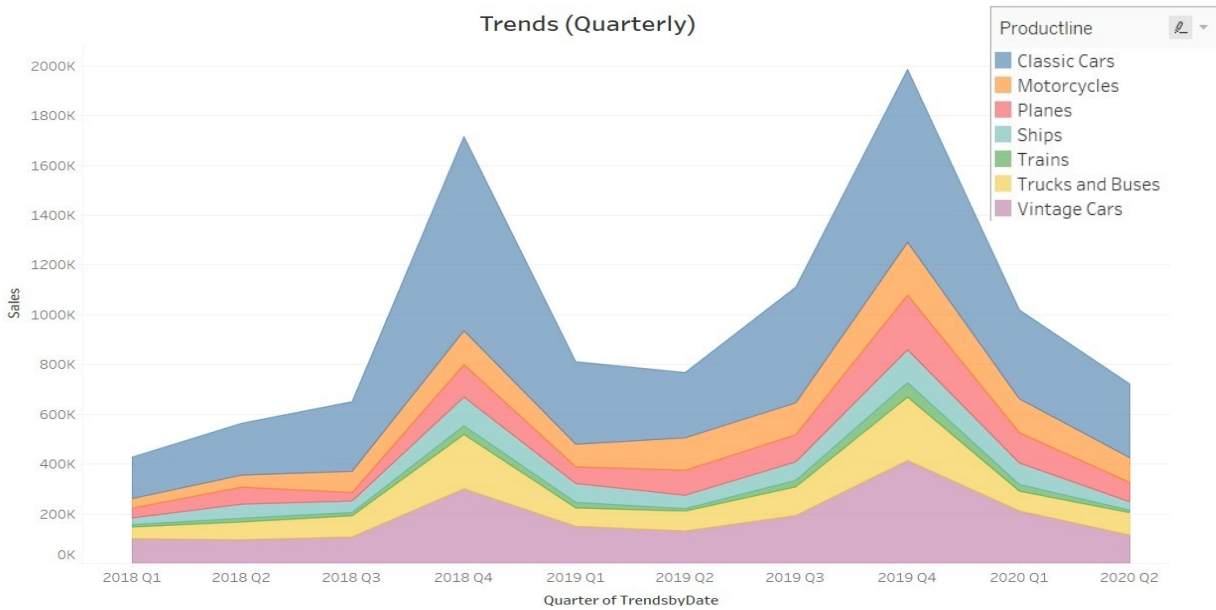


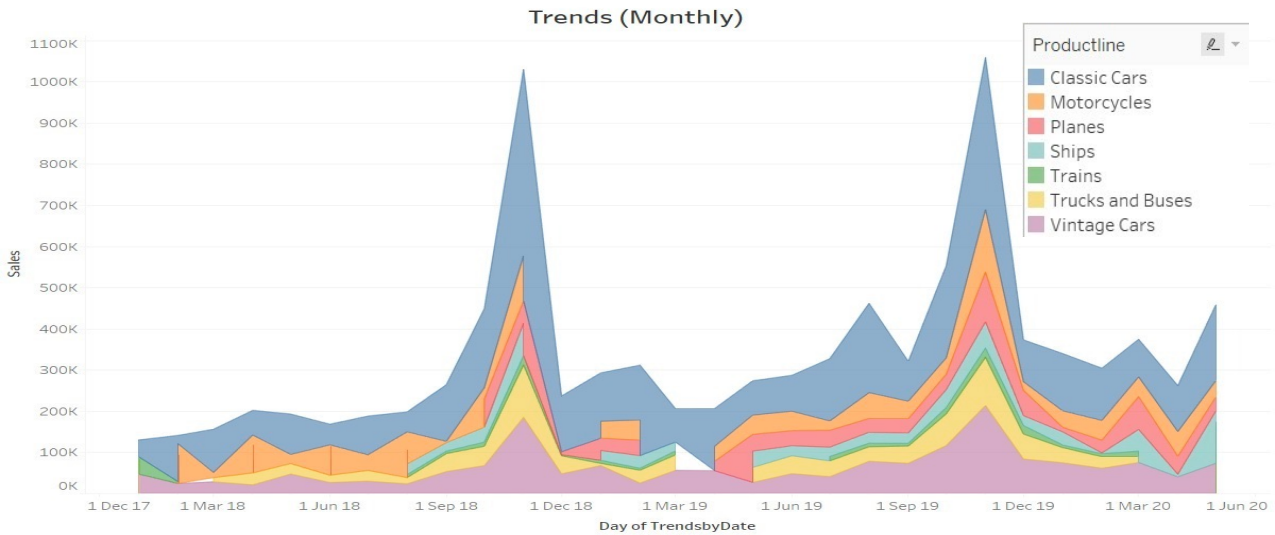


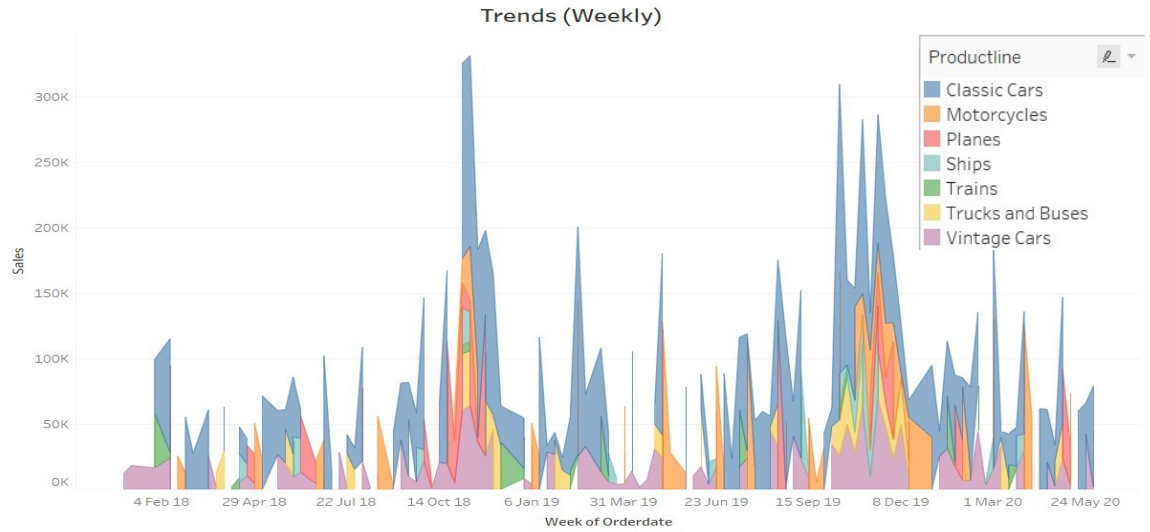
There is hardly any correlation between the any two variables from the pair plot or the heatmap. The only correlation can be seen is between SALES vs QUANTITYORDERED and SALES vs MSRP.

Exploratory Data Analysis - Trends









Recency Frequency Monetary Analysis

What is RFM

RFM analysis is a marketing technique used to quantitatively rank and group customers based on the recency, frequency and monetary total of their recent transactions to identify the best customers and perform targeted marketing campaigns. The system assigns each customer numerical scores based on these factors to provide an objective analysis. RFM analysis is based on the marketing adage that "80% of your business comes from 20% of your customers.

RFM stands for Recency, Frequency, and Monetary value, each corresponding to some key customer trait. These RFM metrics are important indicators of a customer's behavior because frequency and monetary value affects a customer's lifetime value, and recency affects retention, a measure of engagement.

Parameters Used

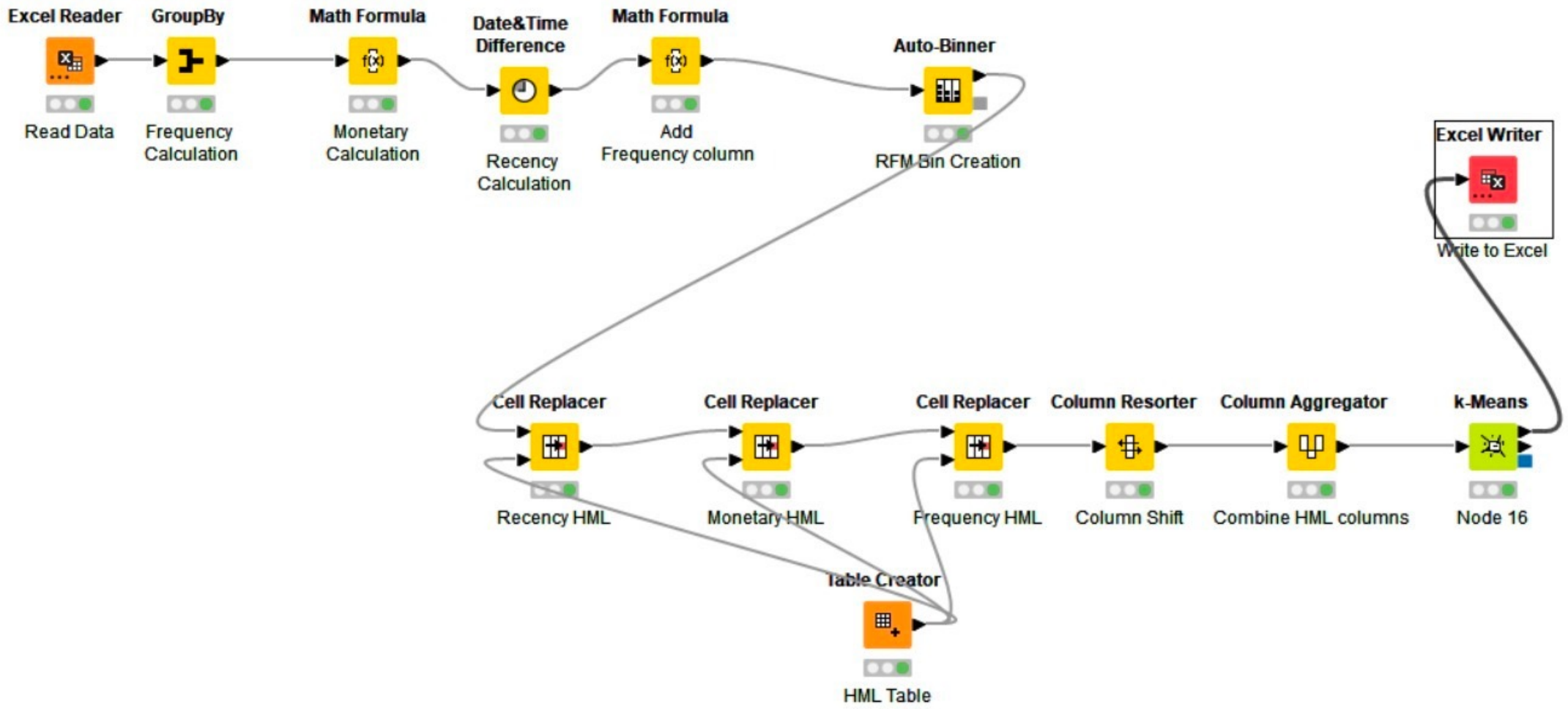
- Recency has been calculated by by first calculating the latest ORDERDATE in all the ORDERNUMBER. This was used as the reference date to calculate the recency of the customer by subtracting his last purchase date from this date. The reference date is 31/05/2020
- Frequency is calculated by counting all the orders made by an induvial customer
- Monetary is calculated by grouping order by customer name and then summing the multiplication of all his QUNATITYORDDERED and PRICEEACH

Assumptions

- All the orders having status apart form SHIPPED were also considered as a valid business transaction as customer has raised the request
- For calculating recency, we have taken the last ORDERDATE as the reference
- 3 Bins of equal frequency has been created for recency, monetary, and frequency. They have been labeled High (H), Medium (M), and Low (L)
- For recency Bin1 has been considered highest as lower value is better
- For monetary and frequency Bin 3 is considered highest as higher is better

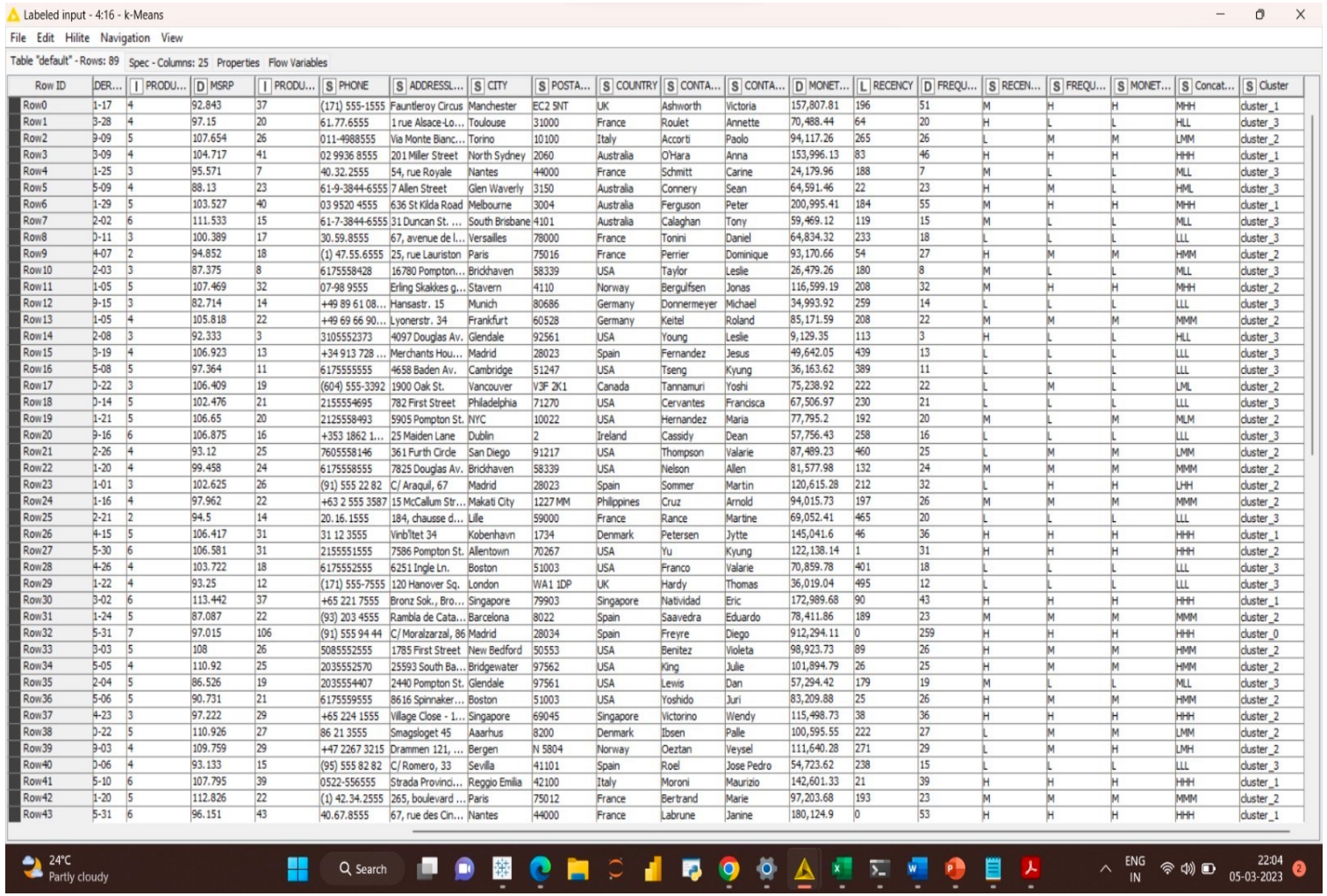
Recency Frequency Monetary Analysis

KNIME Model Workflow



Recency Frequency Monetary Analysis

Output Table Head


Labeled input - 4:16 - k-Means

File Edit Hilite Navigation View

Table "default" - Rows: 89 Spec - Columns: 25 Properties Flow Variables

Row ID	DER...	PRODU...	MSRP	PRODU...	PHONE	ADDRESSL...	CITY	POSTA...	COUNTRY	CONTA...	CONTA...	MONET...	RECENCY	FREQU...	RECEN...	FREQU...	MONET...	Concat...	Cluster
Row0	1-17	4	92.843	37	(171) 555-1555	Fauntleroy Circus	Manchester	EC2 5NT	UK	Ashworth	Victoria	157,807.81	196	51	M	H	H	MHH	cluster_1
Row1	3-28	4	97.15	20	61.77.6555	1 rue Alsace-Lo...	Toulouse	31000	France	Roulet	Annette	70,488.44	64	20	H	L	L	HLL	cluster_3
Row2	9-09	5	107.654	26	011-4988555	Via Monte Bianc...	Torino	10100	Italy	Accorti	Paolo	94,117.26	265	26	L	M	M	LMM	cluster_2
Row3	3-09	4	104.717	41	02 9936 8555	201 Miller Street	North Sydney	2060	Australia	O'Hara	Anna	153,996.13	83	46	H	H	H	HHH	cluster_1
Row4	1-25	3	95.571	7	40.32.2555	54, rue Royale	Nantes	44000	France	Schmitt	Carine	24,179.96	188	7	M	L	L	MLL	cluster_3
Row5	5-09	4	88.13	23	61-9-3844-6555	7 Allen Street	Glen Waverly	3150	Australia	Connery	Sean	64,591.46	22	23	H	M	L	HML	cluster_3
Row6	1-29	5	103.527	40	03 9520 4555	636 St Kilda Road	Melbourne	3004	Australia	Ferguson	Peter	200,995.41	184	55	M	H	H	MHH	cluster_1
Row7	2-02	6	111.533	15	61-7-3844-6555	31 Duncan St. ...	South Brisbane	4101	Australia	Calaghan	Tony	59,469.12	119	15	M	L	L	MLL	cluster_3
Row8	0-11	3	100.389	17	30.59.8555	67, avenue de l...	Versailles	78000	France	Tonini	Daniel	64,834.32	233	18	L	L	L	LLL	cluster_3
Row9	4-07	2	94.852	18	(1) 47.55.6555	25, rue Lauriston	Paris	75016	France	Perrier	Dominique	93,170.66	54	27	H	M	M	HMM	cluster_2
Row10	2-03	3	87.375	8	6175558428	16780 Pompton...	Brickhaven	58339	USA	Taylor	Leslie	26,479.26	180	8	M	L	L	MLL	cluster_3
Row11	1-05	5	107.469	32	07-98 9555	Erling Skakkes g...	Stavern	4110	Norway	Bergulfsen	Jonas	116,599.19	208	32	M	H	H	MHH	cluster_2
Row12	9-15	3	82.714	14	+49 89 61 08...	Hansastr. 15	Munich	80686	Germany	Donnermeyer	Michael	34,993.92	259	14	L	L	L	LLL	cluster_3
Row13	1-05	4	105.818	22	+49 69 66 90...	Lyonerstr. 34	Frankfurt	60528	Germany	Keitel	Roland	85,171.59	208	22	M	M	M	MMM	cluster_2
Row14	2-08	3	92.333	3	3105552373	4097 Douglas Av.	Glendale	92561	USA	Young	Leslie	9,129.35	113	3	H	L	L	HLL	cluster_3
Row15	3-19	4	106.923	13	+34 913 728 ...	Merchants Hou...	Madrid	28023	Spain	Fernandez	Jesus	49,642.05	439	13	L	L	L	LLL	cluster_3
Row16	5-08	5	97.364	11	6175555555	4658 Baden Av.	Cambridge	51247	USA	Tseng	Kyung	36,163.62	389	11	L	L	L	LLL	cluster_3
Row17	0-22	3	106.409	19	(604) 555-3392	1900 Oak St.	Vancouver	V3F 2K1	Canada	Tannamuri	Yoshi	75,238.92	222	22	L	M	L	LML	cluster_2
Row18	0-14	5	102.476	21	2155554695	782 First Street	Philadelphia	71270	USA	Cervantes	Francisca	67,506.97	230	21	L	L	L	LLL	cluster_3
Row19	1-21	5	106.65	20	2125558493	5905 Pompton St.	NYC	10022	USA	Hernandez	Maria	77,795.2	192	20	M	L	M	MLM	cluster_2
Row20	9-16	6	106.875	16	+353 1862 1...	25 Maiden Lane	Dublin	2	Ireland	Cassidy	Dean	57,756.43	258	16	L	L	L	LLL	cluster_3
Row21	2-26	4	93.12	25	7605558146	361 Furth Circle	San Diego	91217	USA	Thompson	Valarie	87,489.23	460	25	L	M	M	LMM	cluster_2
Row22	1-20	4	99.458	24	6175558555	7825 Douglas Av.	Brickhaven	58339	USA	Nelson	Allen	81,577.98	132	24	M	M	M	MMM	cluster_2
Row23	1-01	3	102.625	26	(91) 555 22 82	C/ Araquil, 67	Madrid	28023	Spain	Sommer	Martin	120,615.28	212	32	L	H	H	LHH	cluster_2
Row24	1-16	4	97.962	22	+63 2 555 3587	15 McCallum Str...	Makati City	1227 MM	Philippines	Cruz	Arnold	94,015.73	197	26	M	M	M	MMM	cluster_2
Row25	2-21	2	94.5	14	20.16.1555	184, chausse d...	Lille	59000	France	Rance	Martine	69,052.41	465	20	L	L	L	LLL	cluster_3
Row26	4-15	5	106.417	31	31 12 3555	Vinb'ltet 34	Kobenhavn	1734	Denmark	Petersen	Jytte	145,041.6	46	36	H	H	H	HHH	cluster_1
Row27	5-30	6	106.581	31	2155551555	7586 Pompton St.	Allentown	70267	USA	Yu	Kyung	122,138.14	1	31	H	H	H	HHH	cluster_2
Row28	4-26	4	103.722	18	6175552555	6251 Ingle Ln.	Boston	51003	USA	Franco	Valarie	70,859.78	401	18	L	L	L	LLL	cluster_3
Row29	1-22	4	93.25	12	(171) 555-7555	120 Hanover Sq.	London	WA1 1DP	UK	Hardy	Thomas	36,019.04	495	12	L	L	L	LLL	cluster_3
Row30	3-02	6	113.442	37	+65 221 7555	Bronz Sok., Bro...	Singapore	79903	Singapore	Natividad	Eric	172,989.68	90	43	H	H	H	HHH	cluster_1
Row31	1-24	5	87.087	22	(93) 203 4555	Rambla de Cata...	Barcelona	8022	Spain	Saavedra	Eduardo	78,411.86	189	23	M	M	M	MMM	cluster_2
Row32	5-31	7	97.015	106	(91) 555 94 44	C/ Moralzarzal, 86	Madrid	28034	Spain	Freyre	Diego	912,294.11	0	259	H	H	H	HHH	cluster_0
Row33	3-03	5	108	26	5085552555	1785 First Street	New Bedford	50553	USA	Benitez	Violeta	98,923.73	89	26	H	M	M	HMM	cluster_2
Row34	5-05	4	110.92	25	2035552570	25593 South Ba...	Bridgewater	97562	USA	King	Julie	101,894.79	26	25	H	M	M	HMM	cluster_2
Row35	2-04	5	86.526	19	2035554407	2440 Pompton St.	Glendale	97561	USA	Lewis	Dan	57,294.42	179	19	M	L	L	MLL	cluster_3
Row36	5-06	5	90.731	21	6175559555	8616 Spinnaker...	Boston	51003	USA	Yoshido	Juri	83,209.88	25	26	H	M	M	HMM	cluster_2
Row37	4-23	3	97.222	29	+65 224 1555	Village Close - 1...	Singapore	69045	Singapore	Victorino	Wendy	115,498.73	38	36	H	H	H	HHH	cluster_2
Row38	0-22	5	110.926	27	86 21 3555	Smagsloget 45	Aaarhus	8200	Denmark	Ibsen	Palle	100,595.55	222	27	L	M	M	LMM	cluster_2
Row39	9-03	4	109.759	29	+47 2267 3215	Drammen 121, ...	Bergen	N 5804	Norway	Oeztan	Veysel	111,640.28	271	29	L	M	H	LMH	cluster_2
Row40	0-06	4	93.133	15	(95) 555 82 82	C/ Romero, 33	Sevilla	41101	Spain	Roel	Jose Pedro	54,723.62	238	15	L	L	L	LLL	cluster_3
Row41	5-10	6	107.795	39	0522-556555	Strada Provinci...	Reggio Emilia	42100	Italy	Moroni	Maurizio	142,601.33	21	39	H	H	H	HHH	cluster_1
Row42	1-20	5	112.826	22	(1) 42.34.2555	265, boulevard ...	Paris	75012	France	Bertrand	Marie	97,203.68	193	23	M	M	M	MMM	cluster_2
Row43	5-31	6	96.151	43	40.67.8555	67, rue des Cin...	Nantes	44000	France	Labrune	Janine	180,124.9	0	53	H	H	H	HHH	cluster_1

24°C Partly cloudy Search ENG IN 22:04 05-03-2023

Cluster Sales Analysis

Row Labels	Total SALES
cluster_0	1567152.17
cluster_1	2658560.66
cluster_2	4341369.6
cluster_3	1193139.28

Cluster RFM Analysis

	Recency	Frequency	Monetary
cluster_0	H	H	H
cluster_1	M	H	H
cluster_2	M	M	M
cluster_3	L/M	L	L

Cluster Customer Segementation

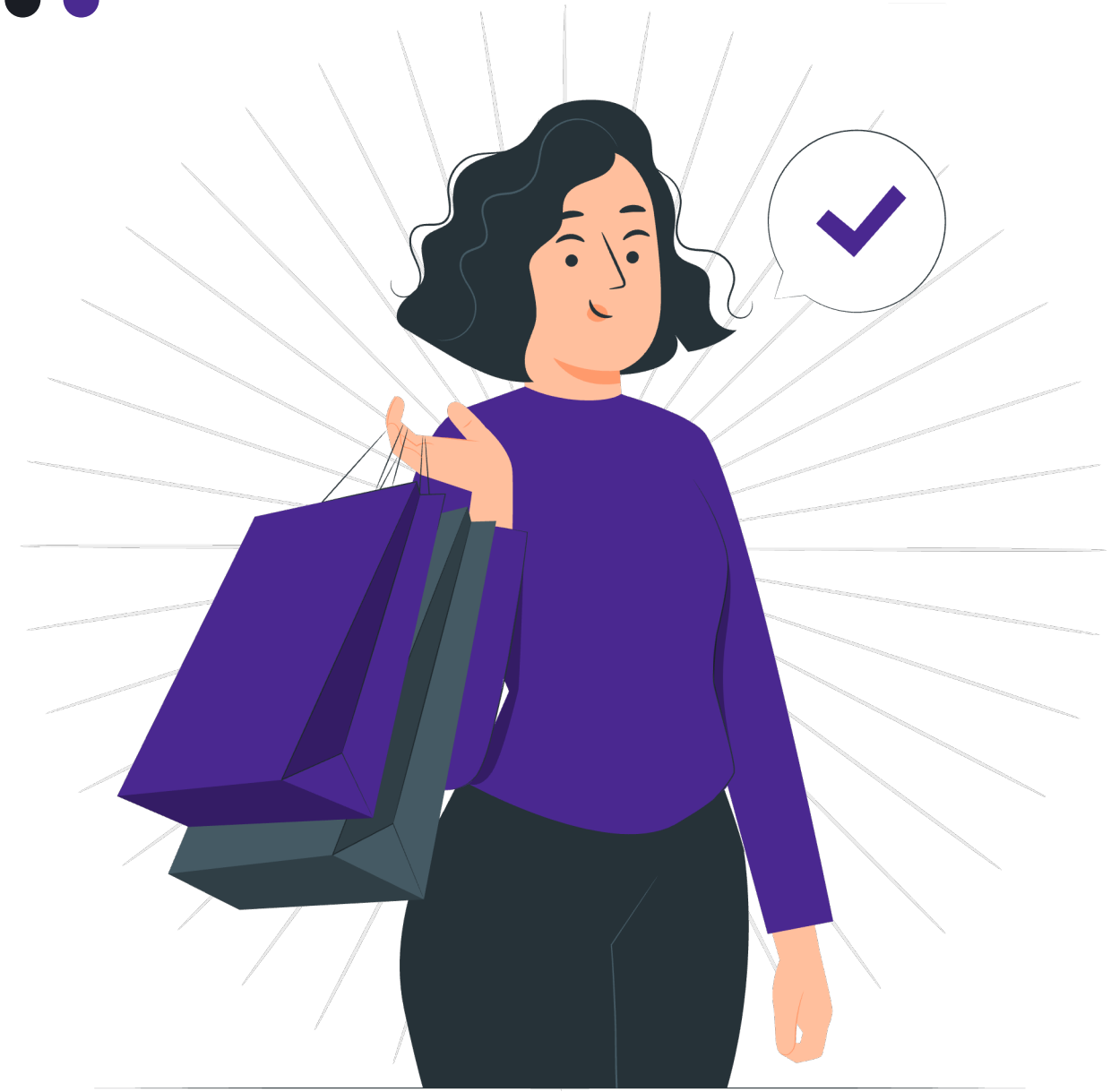
Customer Segment	Clusters
Best Customer	clust_1
Loyal Customer	clust_0
Churn Customer	clust_2
Lost Customer	clust_3

Customer Segmentation - Best Customer

- clust_1 are the best customers as they have the most sum of average sales among the others
- There are a total of 17 customers in this cluster
- Some of them have a have Recency H(10), M(6) and L(1) but the Frequency and Monetary remains high (H) for all
- The top 5 customers in this segment are:
  - Australian Collectors, Co.
  - Muscle Machine Inc
  - La Rochelle Gifts
  - Dragon Souveniers, Ltd.
  - Land of Toys Inc.
- The majority of top customers among this cluster are from Australia and USA

Suggestions

- The organization must focus on providing personalized offering to make this customer buy from them and convert them to loyal customers



Customer Segmentation – Loyal Customer

- clust_0 are the most loyal customers since Recency, Frequency and Monetary are all high for them, concatenated RFM stands at HHH.
- clust_0 has the 2nd best sum of average sales in comparison with the other clusters sum of average sale.
- There are 2 customers in clust_0. The customer names are Euro Shopping Channel and Mini Gifts Distributors Ltd.
- Euro Shopping Channel and Mini Gifts Distributors Ltd. are from Spain and the USA respectively
- Euro Shopping Channel has a larger monetary value than Mini Gifts Distributors Ltd.

Suggestions

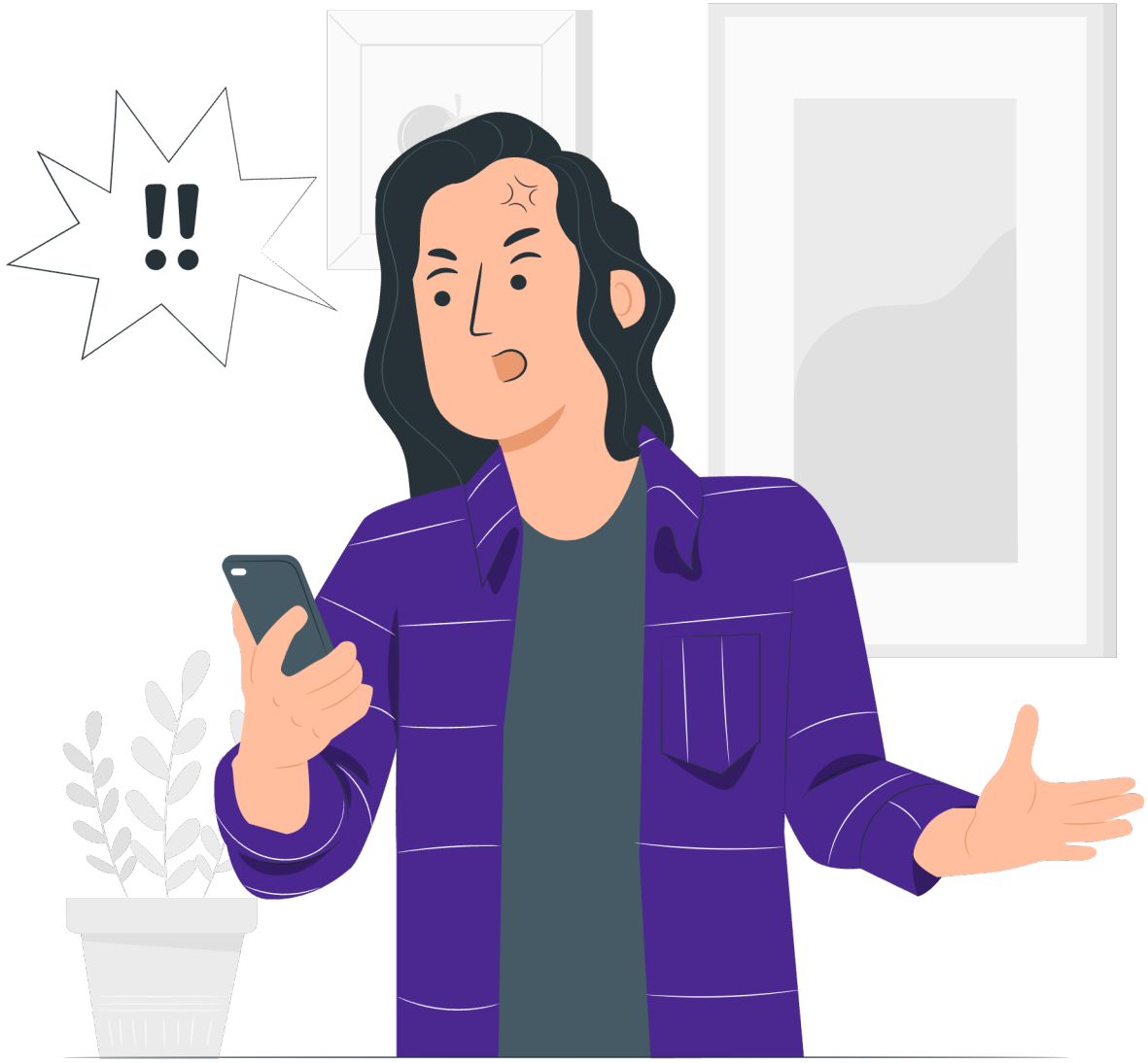
- These are loyal customers and does not require lot to done for selling to them. The organization must try increasing their per order sale quantity more to increase net profits

Customer Segmentation – Churn Customer

- clust_2 are the churn customers
- There are 49 customers in this cluster and probably the highest number of customers in the among any cluster.
- They provide the lowest sum of average sales among the other clusters
- The Recency, Frequency, Monetary are a combined value from H to L. They can be ignored as the difference from the immediate highest sale is too vast.
- The top 3 customers under clust_2 are in order of the descending Monetary value:
  - Amica Models & Co.
  - Auto Canal Petit
  - Baane Mini Imports

Suggestions

- The organization must try to provide these customers personalized offering. The organization must focus on increasing average order quantity of these customer and increase frequency by targeted marketing to convert these customers to Best Customer



Customer Segmentation – Lost Customer

- clust_3 are are the lost customer and surprisingly, they do not have the worst sum of average sales
- There are a total of 24 customers in this cluster
- They have the third highest sum of averge sales, they have provided good business in the past but are lost now
- The most common aggregated RFM is LLL which indicates the average customer are not longer purchasing
- The top 3 customers under clust_3 are in order of the descending Monetary value:
  - Diecast Collectables
  - Alpha Cognac
  - Daedalus Designs Imports

Suggestions

- The customer in this category has low monetary and low frequency so not much needs to be done for customers in this category. However, to expand the business organization might try to provide offering to bring back them as customers only if budget permits.

THANK YOU!