MARKETING & RETAIL ANALYTICS

Project Business Report – DSBA

Submitted By : Maheep Singh

Batch : PGP-DSBA (PGPDSBA.O.AUG24.A)

CONTENT

1. PART A – Automobile Sales Problem

- Business Context & Problem Statement
- Data Summary
- Exploratory Data Analysis
- Customer Segmentation using RFM analysis
- Inferences & Business Recommendations

2. PART B – Grocery Retail Problem

- Business Context & Problem Statement
- Data Summary
- Exploratory Data Analysis
- Market Basket Analysis
- Inferences & Business Recommendations

PART A – Automobile Sales Problem

BUSINESS CONTEXT & PROBLEM STATEMENT

Business Context

- Automobile Parts Manufacturing Company – Sells products to a diverse range of customers for the past three years
- Wish to leverage Analytics & uncover hidden patterns and trends in their customer transactions
- Aims to better understand customer behavior, improve customer segmentation, and implement targeted marketing strategies
- Carry out Customer Segmentation using RFM analysis
- Ultimate goal is to enhance customer satisfaction, while, driving revenue growth by offering more personalized and efficient services.

Problem Statement / Objective

- Identify underlying patterns in customer purchasing behavior
- Segment customers based on their transactional data
- Provide actionable insights to optimize the company's marketing efforts
- Recommend personalized marketing strategies for each customer segment to maximize sales and customer retention

DATA SUMMARY (1/2)

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 Information

```
 #   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
```

```
Missing Values:-
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
```

```
Duplicated Values:-
0
```

DATA SUMMARY (2/2)

Data Description

	count	unique	top	freq	mean	min	25%	50%	75%	max	std
QUANTITYORDERED	2747.0	NaN	NaN	NaN	35.103021	6.0	27.0	35.0	43.0	97.0	9.762135
PRICEEACH	2747.0	NaN	NaN	NaN	101.098951	26.88	68.745	95.55	127.1	252.87	42.042548
ORDERLINENUMBER	2747.0	NaN	NaN	NaN	6.491081	1.0	3.0	6.0	9.0	18.0	4.230544
SALES	2747.0	NaN	NaN	NaN	3553.047583	482.13	2204.35	3184.8	4503.095	14082.8	1838.953901
ORDERDATE	2747	NaN	NaN	NaN	2019-05-13 21:56:17.211503360	2018-01-06 00:00:00	2018-11-08 00:00:00	2019-06-24 00:00:00	2019-11-17 00:00:00	2020-05-31 00:00:00	NaN
DAYS_SINCE_LASTORDER	2747.0	NaN	NaN	NaN	383.085912	0.0	196.0	342.0	570.0	876.0	230.231295
STATUS	2747	6	Shipped	2541	NaN	NaN	NaN	NaN	NaN	NaN	NaN
PRODUCTLINE	2747	7	Classic Cars	949	NaN	NaN	NaN	NaN	NaN	NaN	NaN
MSRP	2747.0	NaN	NaN	NaN	100.691664	33.0	68.0	99.0	124.0	214.0	40.114802
PRODUCTCODE	2747	109	S18_3232	51	NaN	NaN	NaN	NaN	NaN	NaN	NaN
CUSTOMERNAME	2747	89	Euro Shopping Channel	259	NaN	NaN	NaN	NaN	NaN	NaN	NaN
PHONE	2747	88	(91) 555 94 44	259	NaN	NaN	NaN	NaN	NaN	NaN	NaN
ADDRESSLINE1	2747	89	C/ Moralzarzal, 86	259	NaN	NaN	NaN	NaN	NaN	NaN	NaN
CITY	2747	71	Madrid	304	NaN	NaN	NaN	NaN	NaN	NaN	NaN
POSTALCODE	2747	73	28034	259	NaN	NaN	NaN	NaN	NaN	NaN	NaN
COUNTRY	2747	19	USA	928	NaN	NaN	NaN	NaN	NaN	NaN	NaN
CONTACTLASTNAME	2747	76	Freyre	259	NaN	NaN	NaN	NaN	NaN	NaN	NaN
CONTACTFIRSTNAME	2747	72	Diego	259	NaN	NaN	NaN	NaN	NaN	NaN	NaN
DEALSIZE	2747	3	Medium	1349	NaN	NaN	NaN	NaN	NaN	NaN	NaN

Data Inference

- **Data**: Past 3 years.
- **Dataset**: 20 columns and 2747 rows
- **Missing values and Duplicate values**: None
- **No abnormalities** detected in the data that required pre-processing

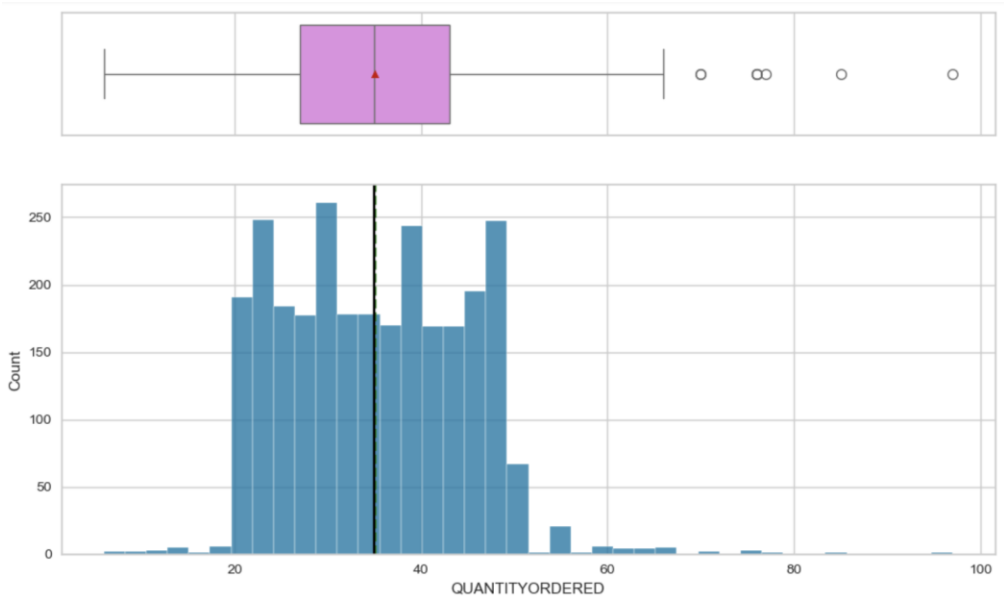
- **Mean** of No. of items ordered = 35 | **Standard Deviation** = 9.76 | **Range covering 68%** values = 25-45
- **Mean** of Price of each item = 101.9 | **Standard Deviation** = 42.04 | **Range covering 68%** values = 60-144
- **Mean** of Sales amount per order = 3553.05 |**Standard Deviation** = 1838.95 | **Range covering 68%** values = 1714-5392
- **Mean** of Days since the last order = 383.09 |**Standard Deviation** = 230.23 | **Range covering 68%** values = 153-613
- **Mean** of Manufacturer's Suggested Retail Price per order = 100.69 |**Standard Deviation** = 40.11 | **Range covering 68%** values = 61-141

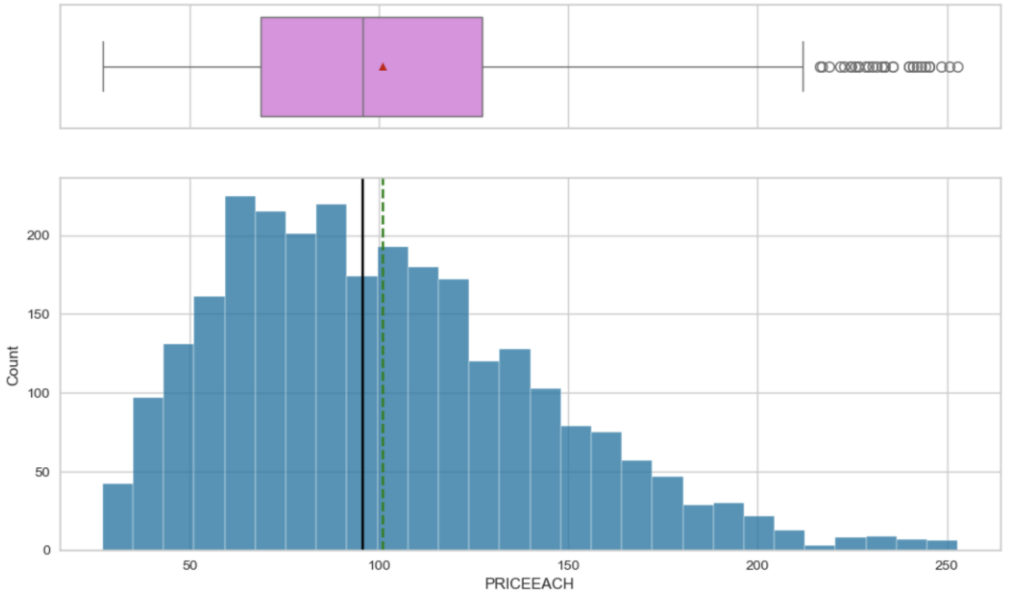
Assumptions: -

- Each row in the data represents a unique transaction made by a customer
- 'DAYS_SINCE_LAST_ORDER' is ignored and recomputed as '[Max(ORDERDATE)- ORDERDATE]'
- The recommendations provided are based on the insights gained from the analysis of the transaction data.

EXPLORATORY DATA ANALYSIS (1/13)

Univariate Analysis





Quantity Ordered:-

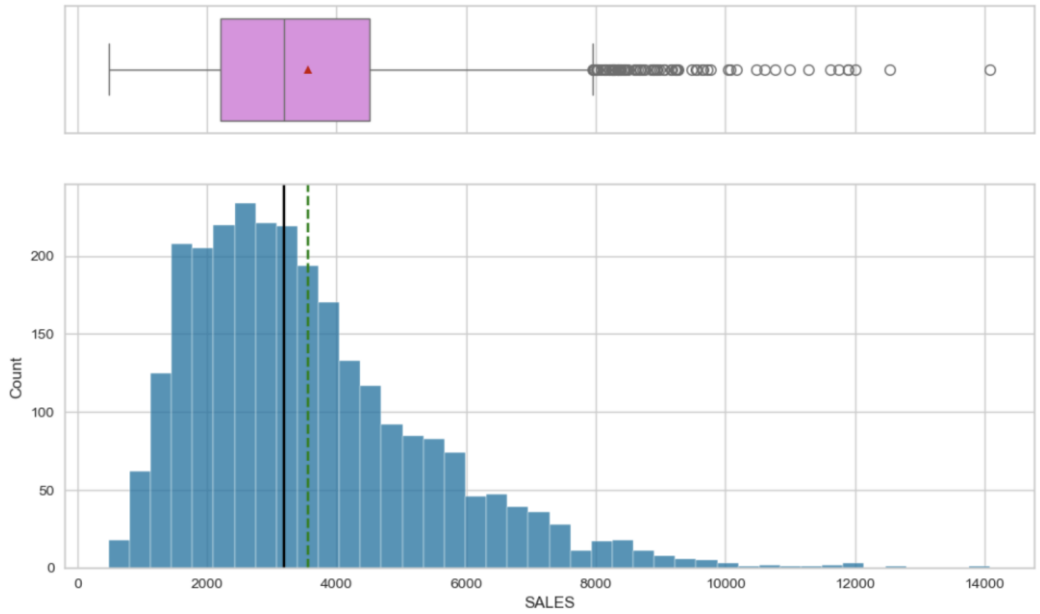
- There is a large variation
- Distribution seems to be symmetric with similar mean & median
- Seems to be multi-modal distribution having multiple peaks.
- Few outliers observed, but we would keep the data as-is to avoid any loss of information.

Price of Each Item:-

- There is a large variation
- Distribution seems to be slightly right-skewed
- Seems to be multi-modal distribution having multiple peaks
- Lot of outliers observed, but we would keep the data as-is to avoid any loss of information

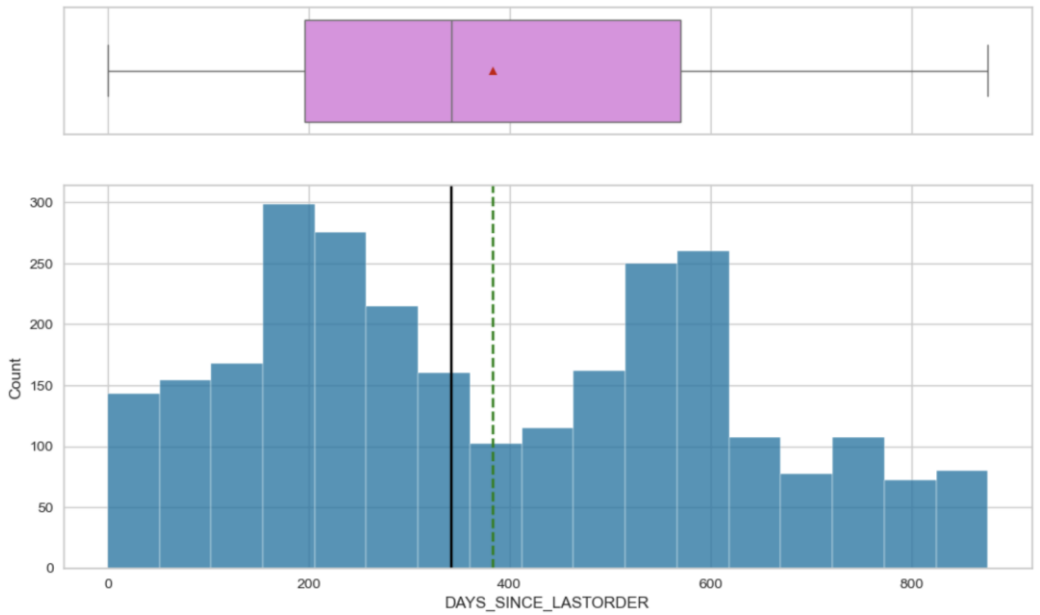
EXPLORATORY DATA ANALYSIS (2/13)

Univariate Analysis



Sales:-

- There is a large variation
- Distribution seems to be right-skewed
- Seems to be unimodal distribution
- Lot of outliers observed, but we would keep the data as-is to avoid any loss of information.

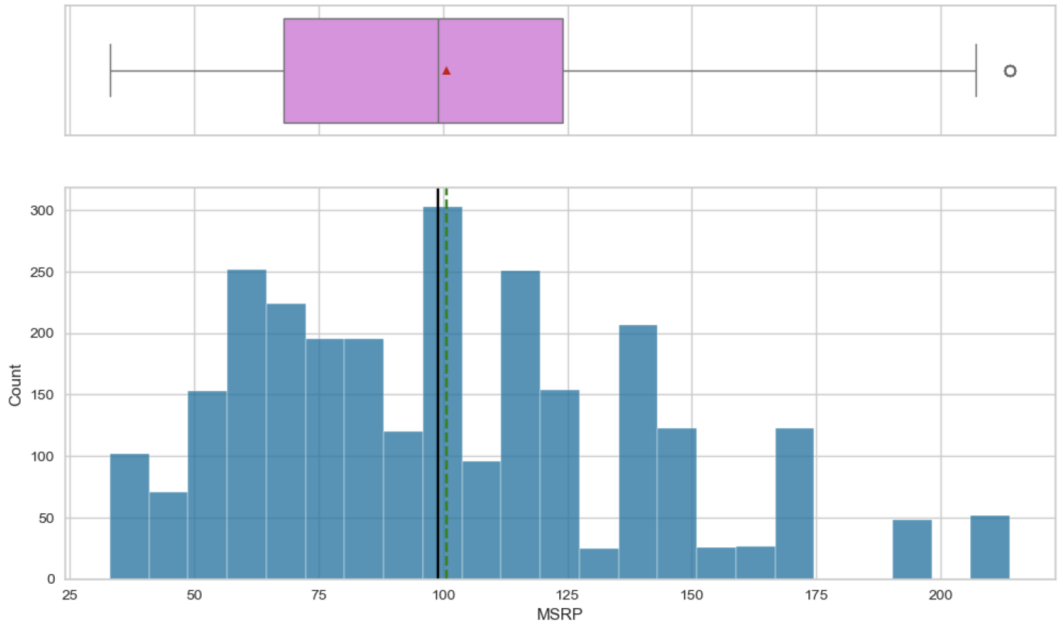


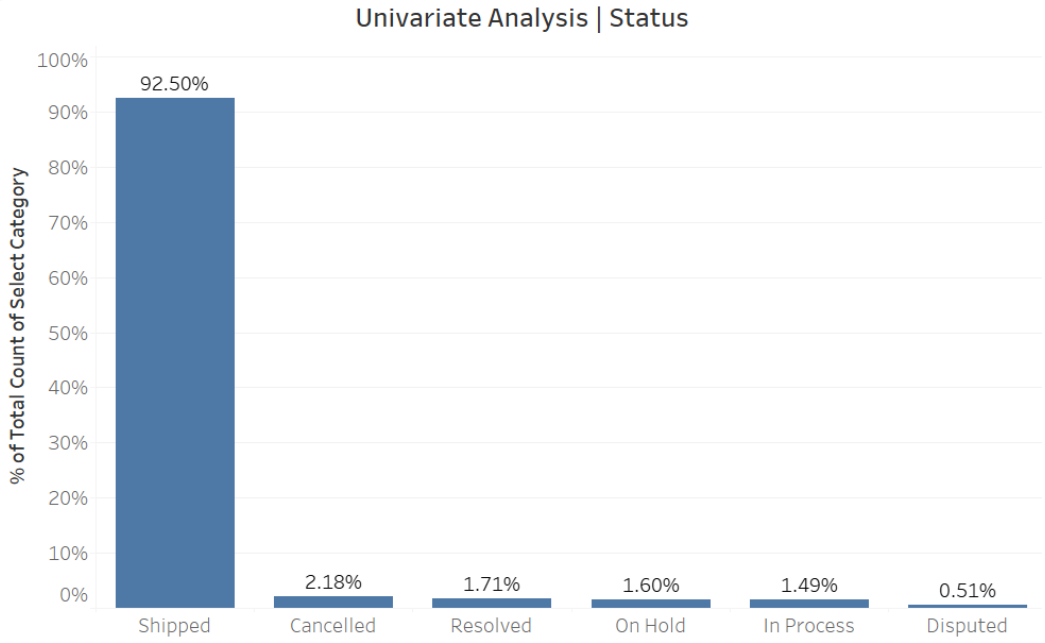
Days Since Last Order:-

- There is a large variation
- Distribution seems to be slightly right-skewed
- Seems to be Bi-modal distribution having 2 peaks
- No outliers observed

EXPLORATORY DATA ANALYSIS (3/13)

Univariate Analysis





Manufacturer's Suggested Retail Price:-

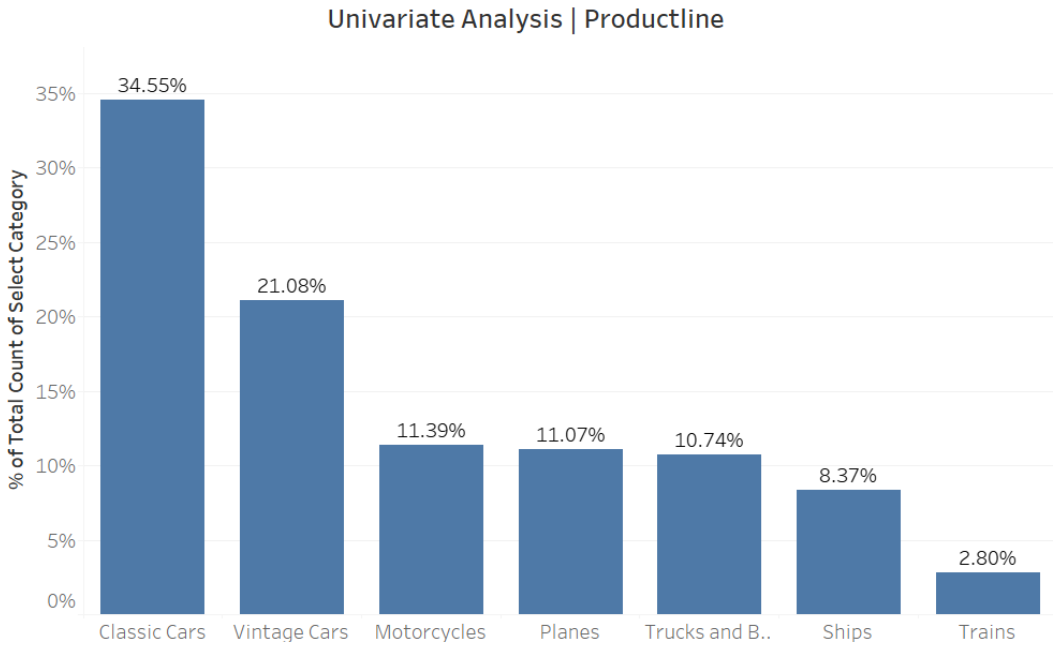
- There is a large variation
- Distribution seems to be almost symmetrical
- Seems to be unimodal distribution
- No outliers observed

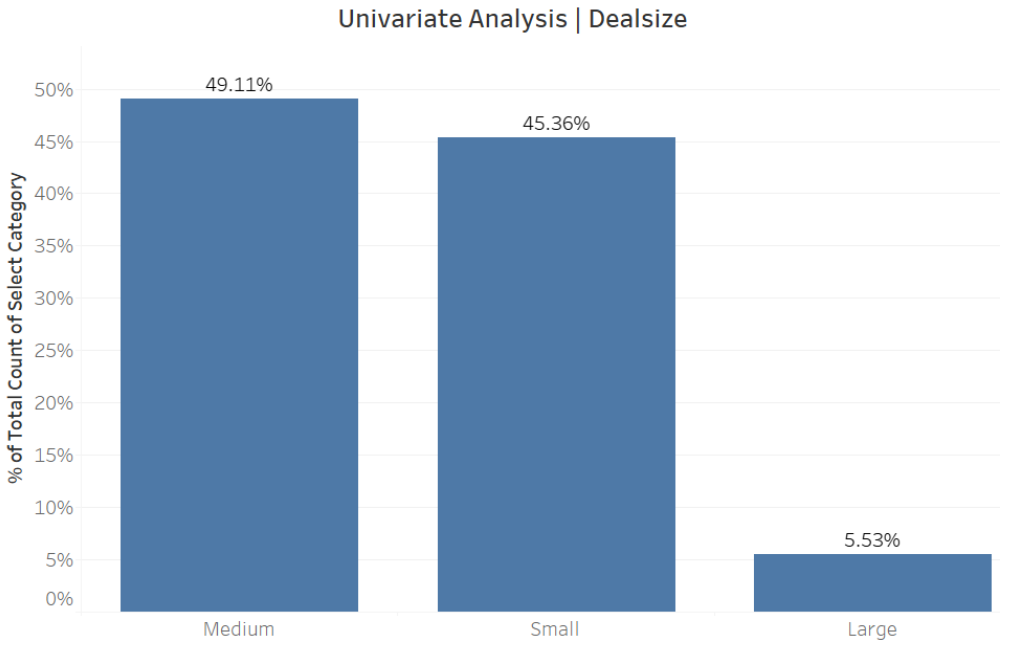
Status of the Order:-

- Majority of the orders are in the Shipped state (~93%)
- 1-2% orders each are in Cancelled, Disputed In-progress, On-hold or Resolves State

EXPLORATORY DATA ANALYSIS (4/13)

Univariate Analysis





Product Line (Category of Product):-

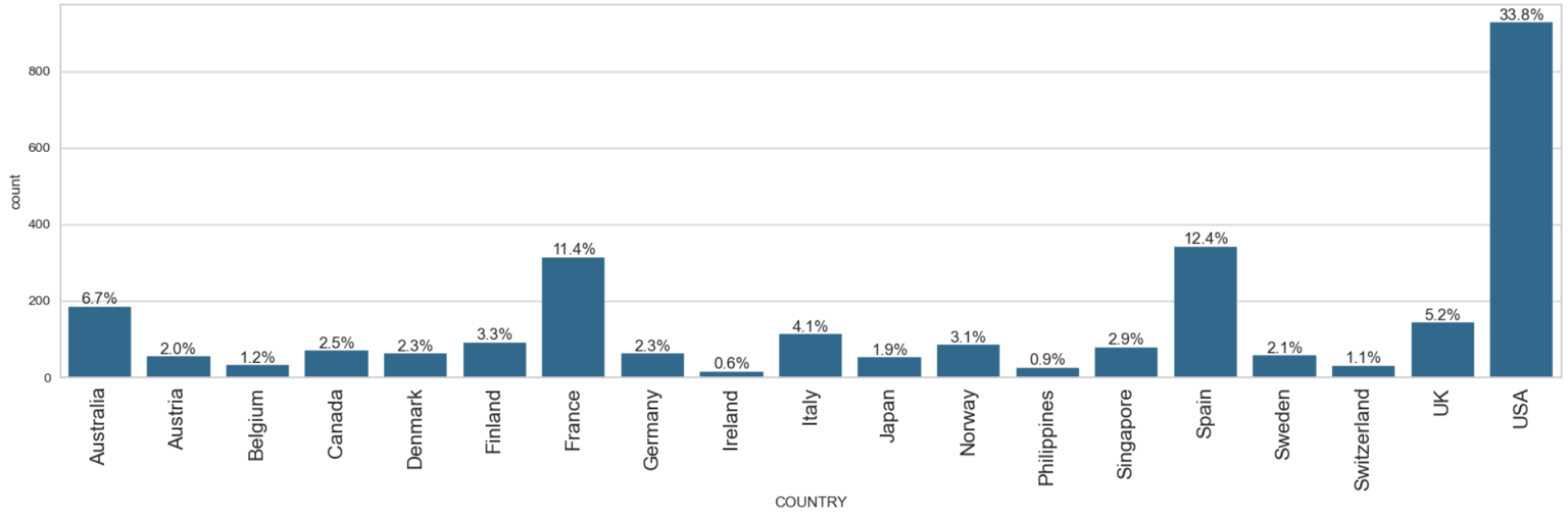
- Majority of the products are Classic Car Parts (35%), followed by Vintage Car Parts (21%)
- Other categories constitute 8-12% each.

Deal Size:-

- Majority of the deals are either Medium or Small (45-49%)
- Only 5% constitute Large sized-deals

EXPLORATORY DATA ANALYSIS (5/13)

Univariate Analysis

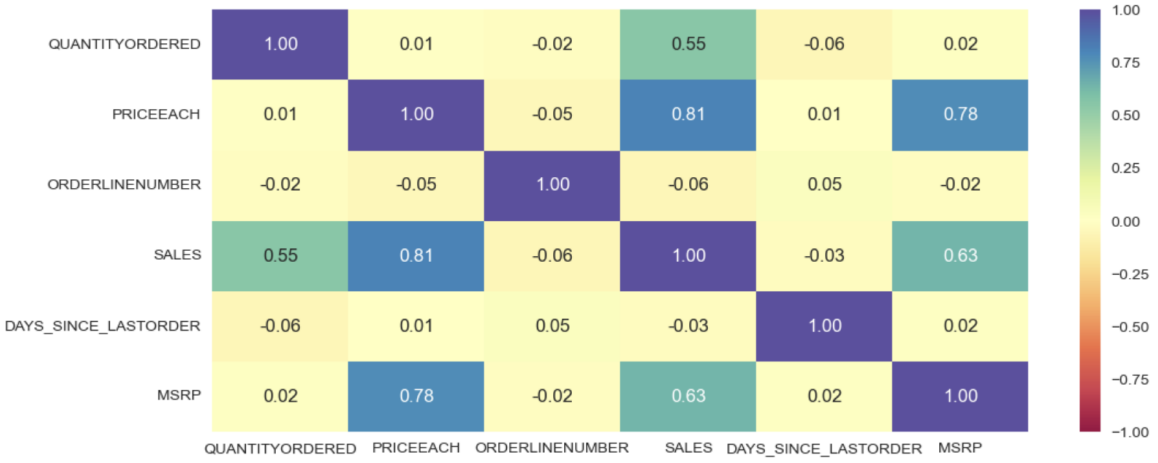


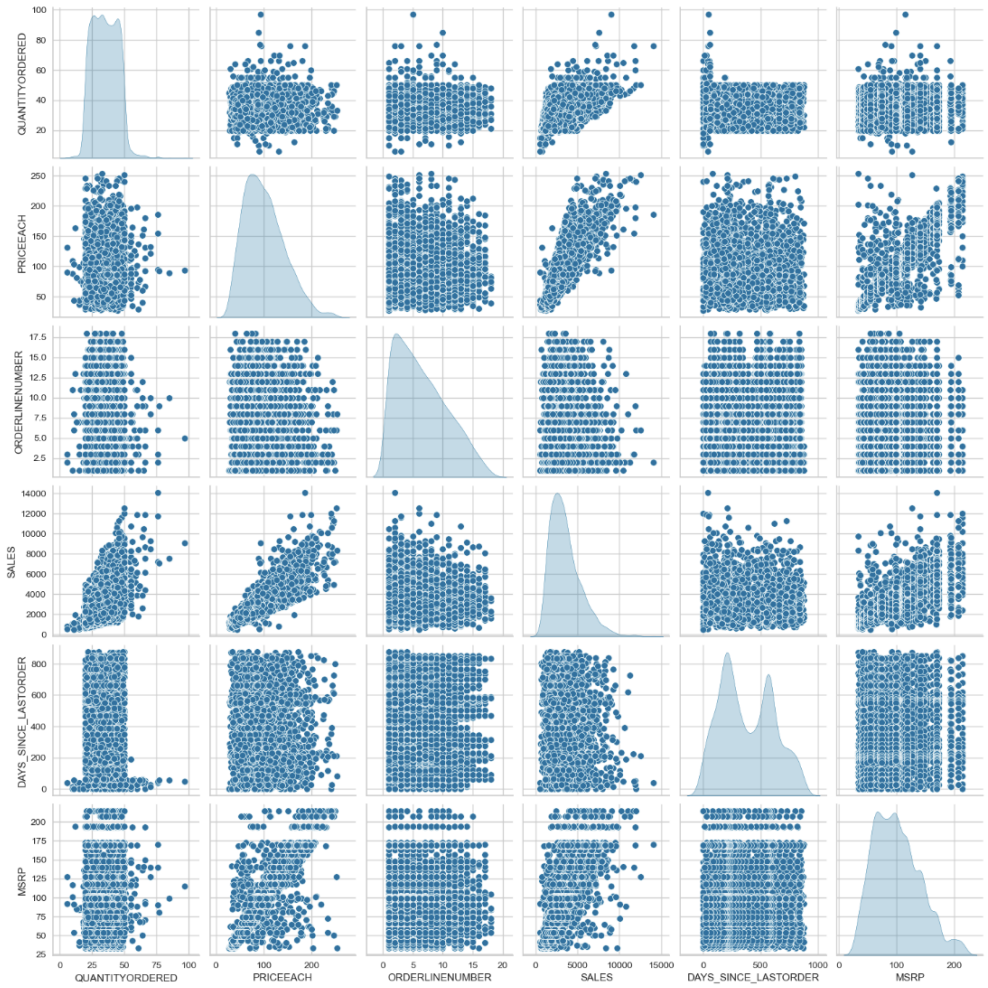
Country of Customer:-

- Majority of the customers belong to US (34%), followed by Spain (12%) & France (11%)
- Other countries constitute 1-7% each.

EXPLORATORY DATA ANALYSIS (6/13)

Bivariate Analysis – Heatmap & Pairplot



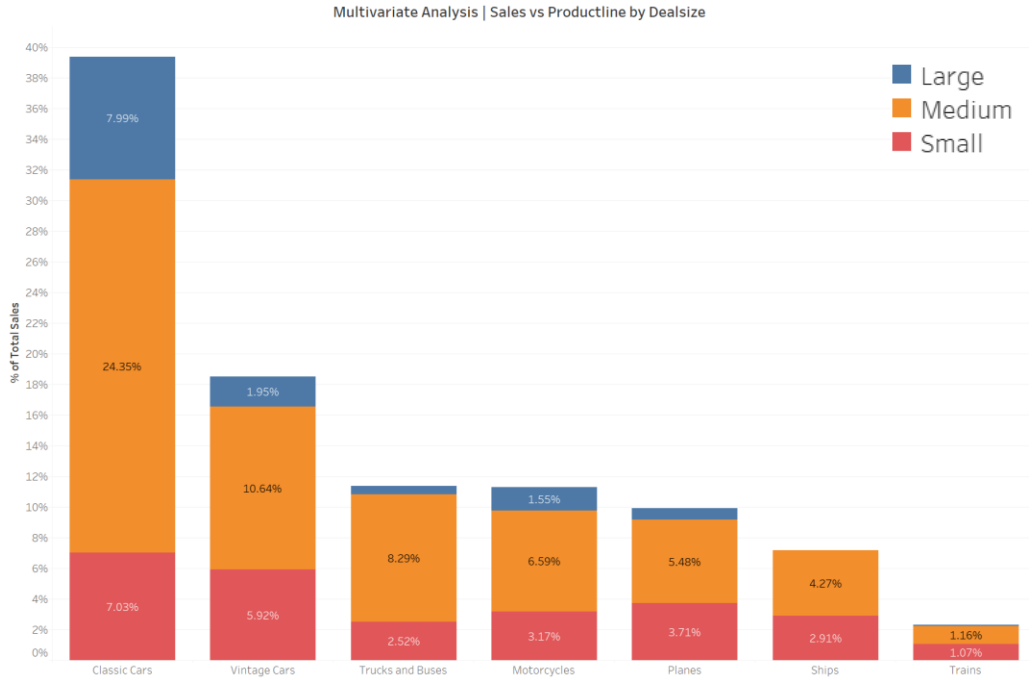


Correlation Inference between Numerical Variables:-

- No significant correlation observed between variables of interest
- Expected correlations observed:-
  - ✓ Between Sales & Price of Each Item – Expected as Price would impact sales directly
  - ✓ Between Price of Each Item & Manufacturer's Suggested Retail Price – Expected correlation
  - ✓ Between Sales & Manufacturer's Suggested Retail Price – Expected as Manufacturer's Retail Price would impact sales directly

EXPLORATORY DATA ANALYSIS (7/13)

Bivariate/Multivariate Analysis



Sales vs Product Line & Deal Size:-

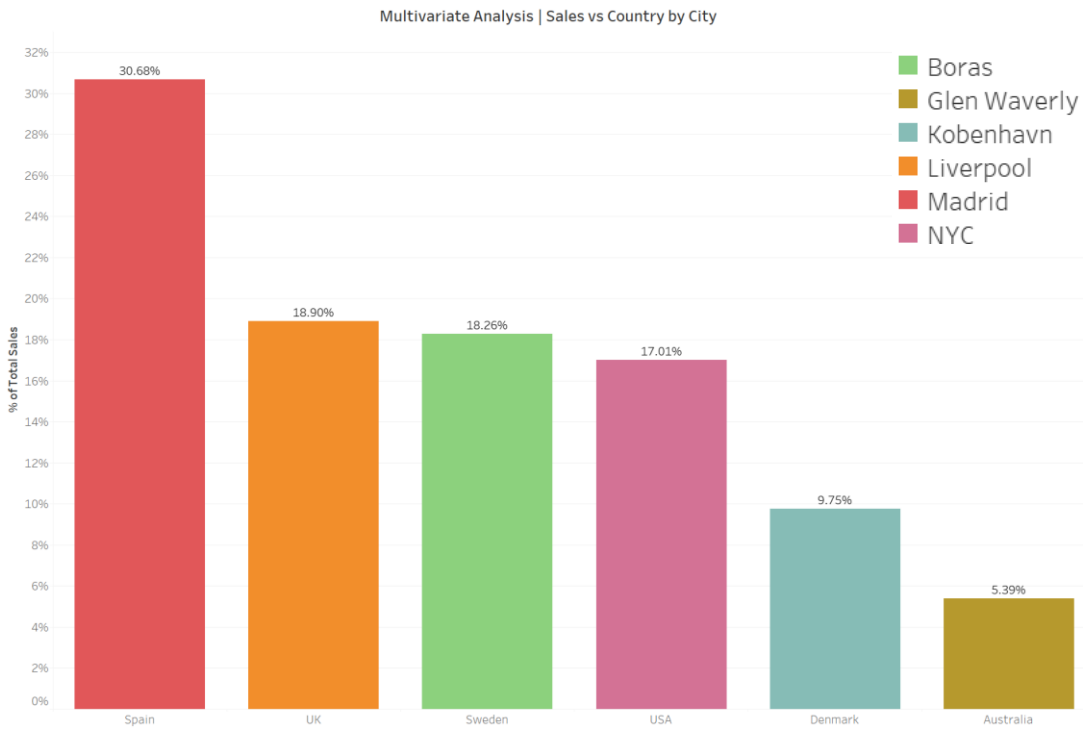
- Clearly, Classic Car Parts contribute the most in terms of revenue (39%), followed by, Vintage Car Parts (19%)
- Large & Medium Deal Size is mostly coming from Classic Cars category whereas Small Deal Size from Classic Car & Vintage Car parts

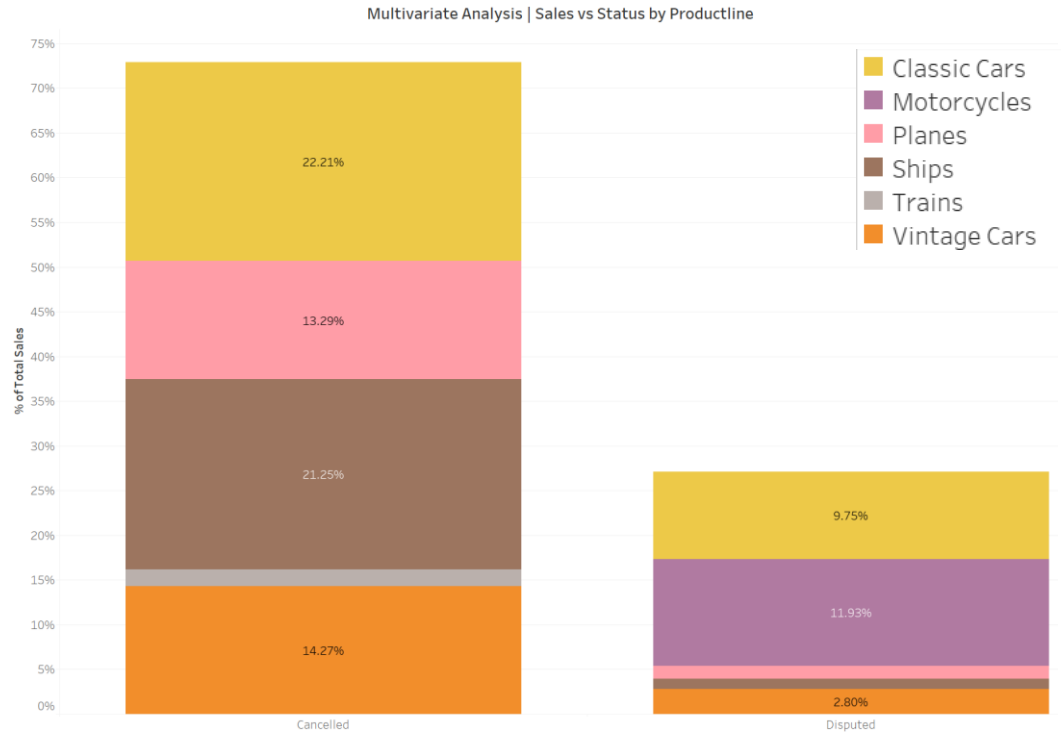
Sales vs Status & Product Line:-

- 92% of revenue is contributed by Shipped orders
- Of the orders shipped, Classic Car & Vintage Car parts contribute the most (37% & 17% respectively)

EXPLORATORY DATA ANALYSIS (8/13)

Bivariate/Multivariate Analysis – Disputed & Cancelled Orders



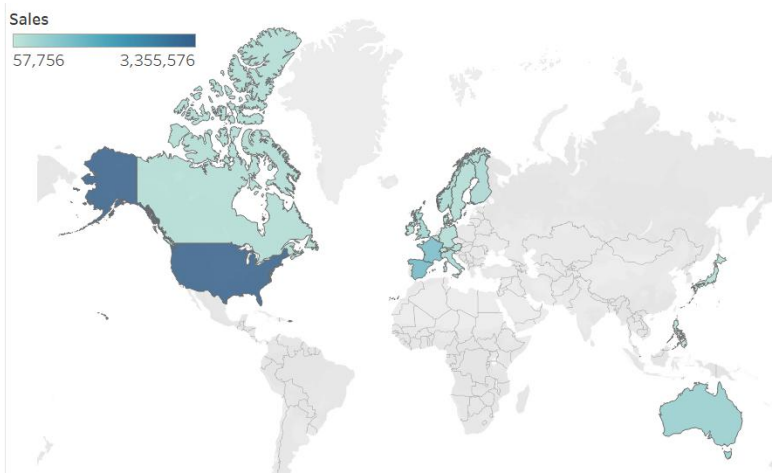


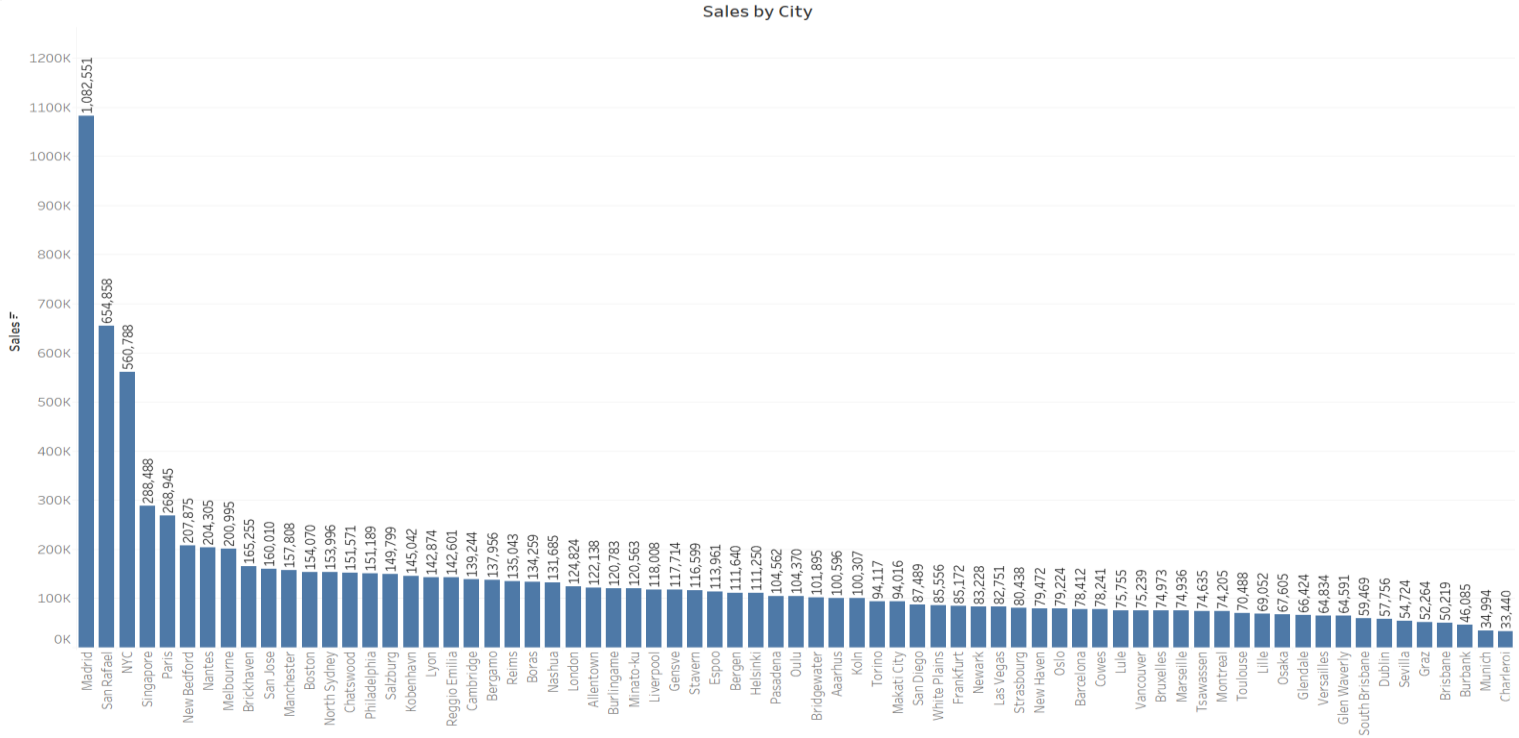
Sales Lost due to Disputed/Cancelled Orders by Country/City & Product Line:-

- Max. Revenue Loss due to Disputed/Cancelled orders contributed by Spain (Madrid), followed by UK (Liverpool), Sweden (Boras), USA (NYC), Denmark (Kobenhavn) & Australia (Glen Waverly)
- Ship parts suffer maximum Cancellations & Motorcycle parts suffer from maximum Disputed Orders

EXPLORATORY DATA ANALYSIS (9/13)

Bivariate/Multivariate Analysis – By Region



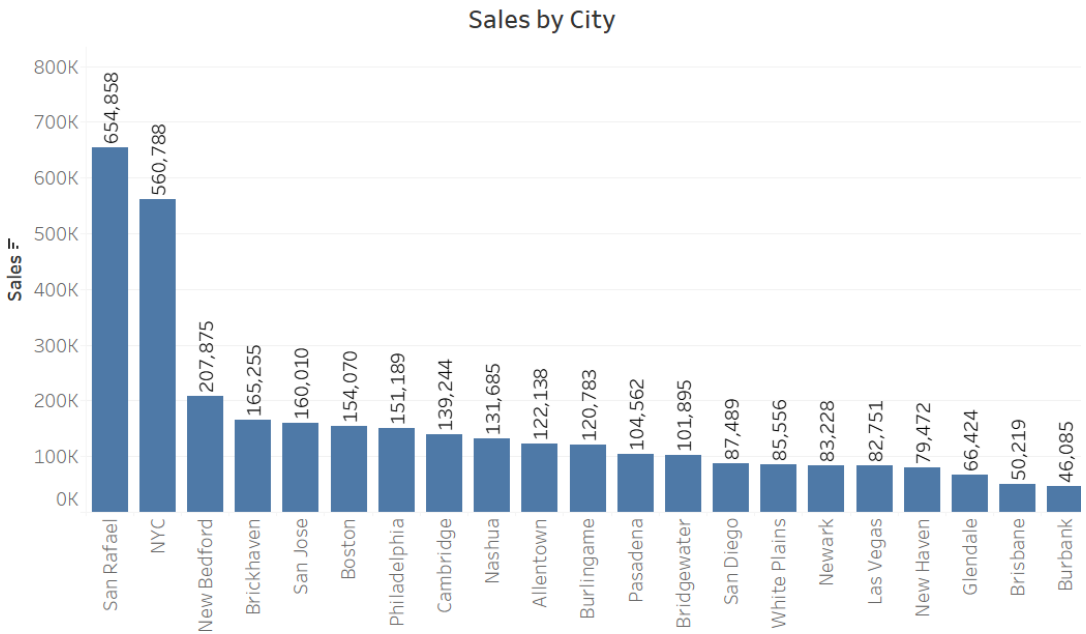


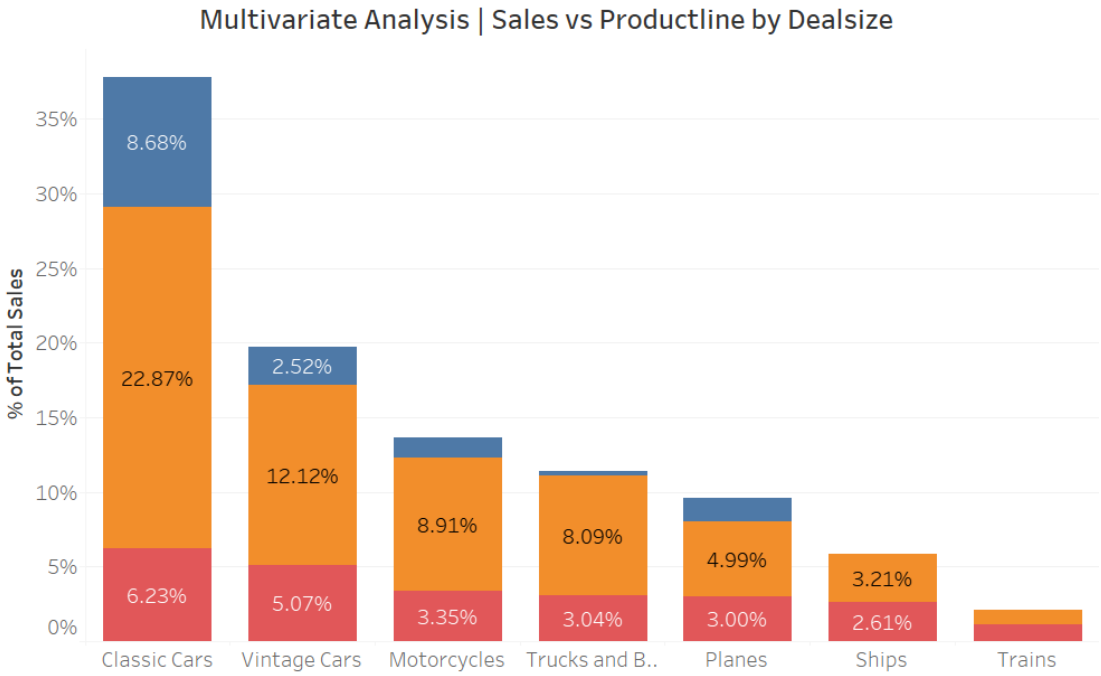
Sales by Country & City:-

- In the previous slides we observed that majority of the orders came from US, Spain & France. However, US & Canada are contributing to majority of Revenue
- US is dominating in terms of no. of orders & Revenue.
- Top 3 Cities contributing to Revenue – Madrid, San Rafael & NYC

EXPLORATORY DATA ANALYSIS (10/13)

Bivariate/Multivariate Analysis – US Region



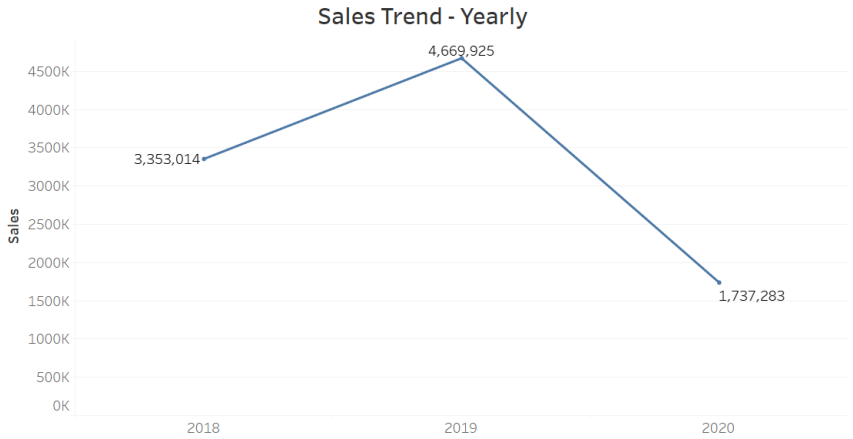


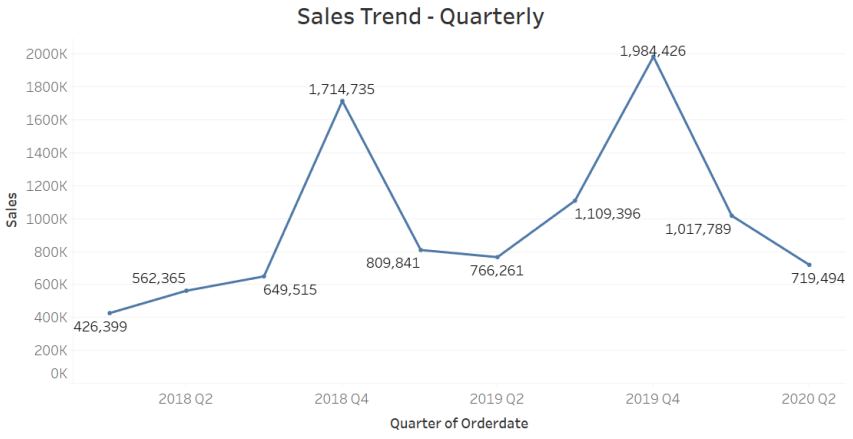
Sales by City & Product line/Deal size – US Region Drilldown

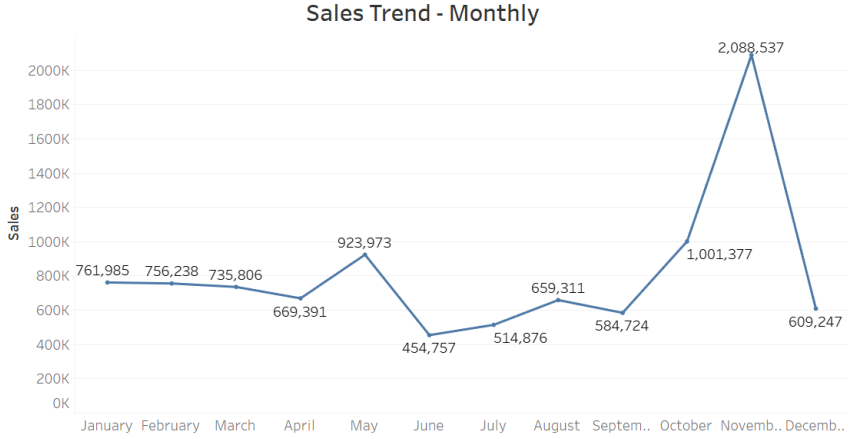
- In the US Region, San Rafael & NYC significantly contribute to the revenue with 5X sales relative to other regions
- Classic Car & Vintage Car partss are the top 2 contributors in terms of both the Product Category & Deal Size

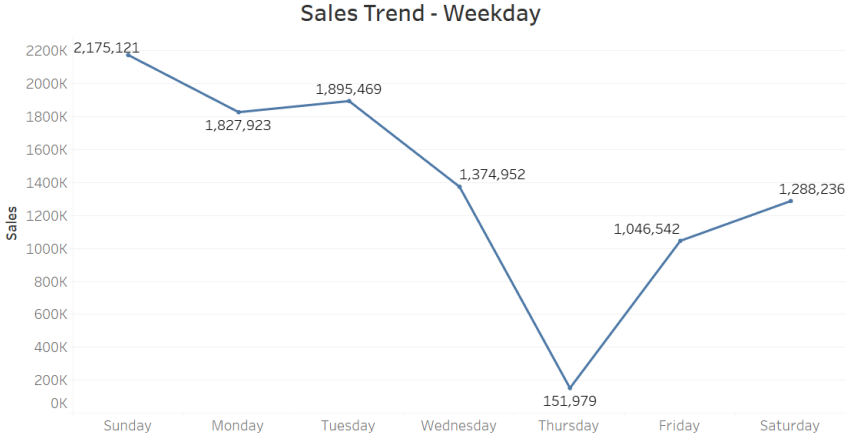
EXPLORATORY DATA ANALYSIS (11/13)

Bivariate/Multivariate Analysis – Sales Trend









Sales Trend –
Yearly, Quarterly, Monthly & Weekly

- 3 years data being analyzed
- Even though the yearly trend suggests a decline, it cannot be ascertained as the data provided for 2020 is not complete (only 2 quarters data provided). If we interpolate the data for remaining quarters, the trend would seem stable (with slight negative bias).
- Clear pattern of uptick in the Sales observed in the month of November
- On weekly basis, sales on Thursdays is the lowest, relative to other weekdays. In fact, sales increase from Friday to Sun & drops from Monday to Thursday.

EXPLORATORY DATA ANALYSIS (12/13)

Key Takeaways

Insights – Seasonality

Stable (slight negative bias) Annual Sales

Q4 (predominantly, Nov.) has relatively higher Sales

Sales increase from Friday to Sunday and dip from Monday to Thursday

Thursday records the lowest Sales out of 7 days

Insights – General

'Classic Car' & 'Vintage Car' Parts record the highest Sales

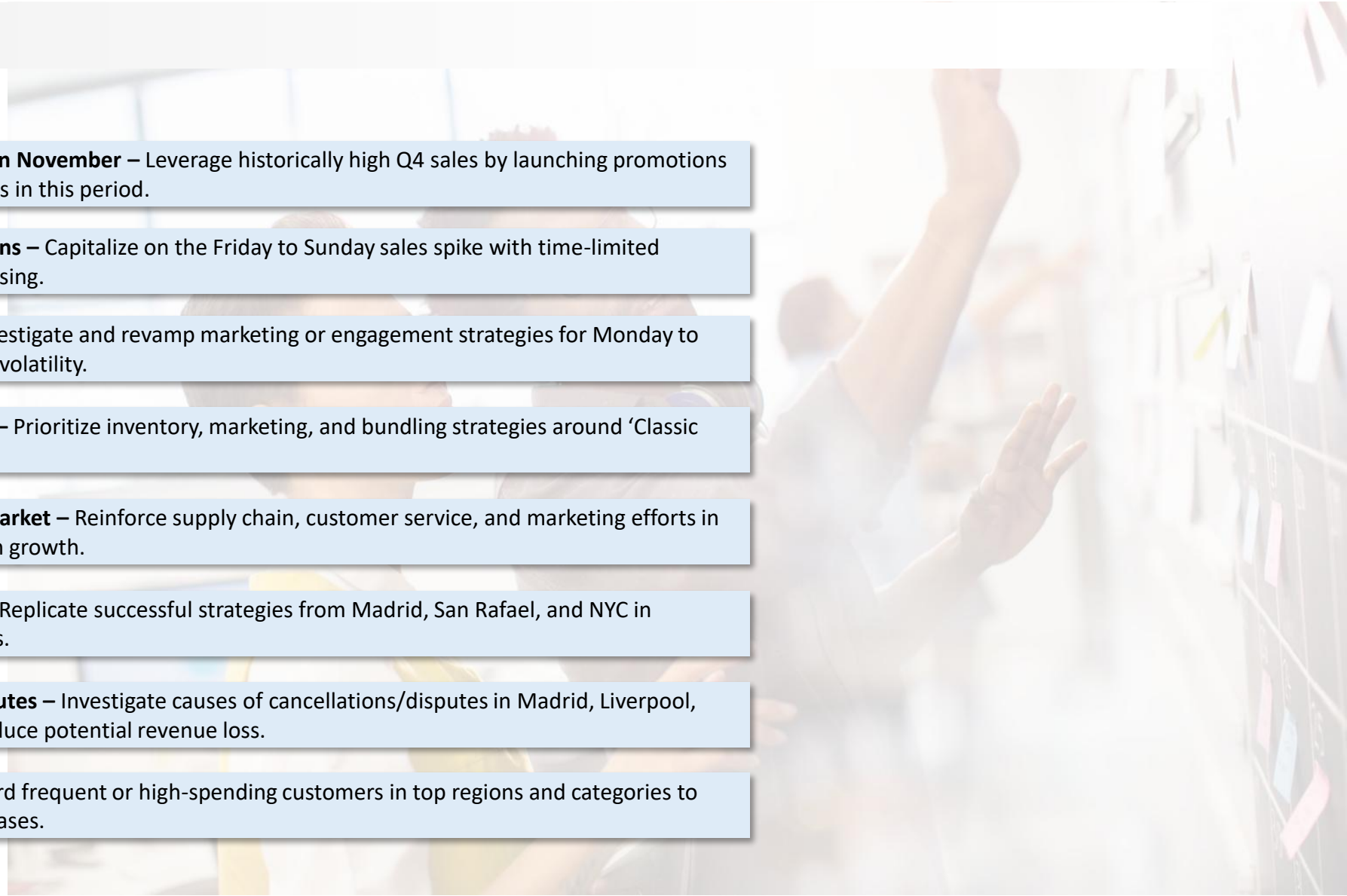
US Region dominates – No. of orders & Revenue

Sales Superstars – Madrid (Spain), San Rafael (US) & NYC (US)

Top 5 Cities with Revenue Loss (Disputed/Cancelled Orders) – Madrid, Liverpool, Boras, NYC & Kobenhavn

EXPLORATORY DATA ANALYSIS (13/13)

Business Recommendations

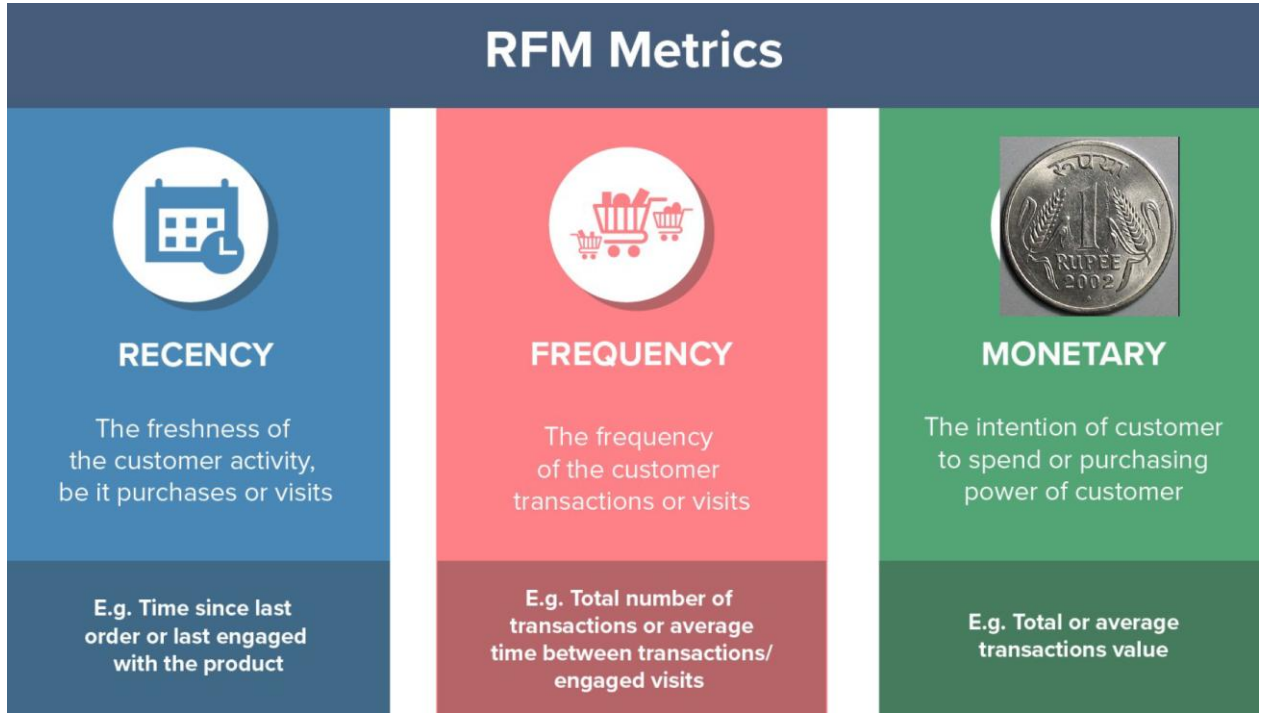


- **Boost Q4 Campaigns, Especially in November –** Leverage historically high Q4 sales by launching promotions and targeted marketing campaigns in this period.
- **Run Weekend-Focused Promotions –** Capitalize on the Friday to Sunday sales spike with time-limited weekend deals and online advertising.
- **Address Weekday Sales Dip –** Investigate and revamp marketing or engagement strategies for Monday to Thursday to smooth weekly sales volatility.
- **Promote Best-Selling Categories –** Prioritize inventory, marketing, and bundling strategies around 'Classic Car' and 'Vintage Car' parts.
- **Strengthen Presence in the US Market –** Reinforce supply chain, customer service, and marketing efforts in the dominant US region to sustain growth.
- **Expand High-Performing Cities –** Replicate successful strategies from Madrid, San Rafael, and NYC in underperforming but similar cities.
- **Analyze and Mitigate Order Disputes –** Investigate causes of cancellations/disputes in Madrid, Liverpool, Boras, NYC, and Kobenhavn to reduce potential revenue loss.
- **Launch Loyalty Programs –** Reward frequent or high-spending customers in top regions and categories to boost retention and repeat purchases.

CUSTOMER SEGMENTATION USING RFM (1/9)

What is RFM?

- Recency, Frequency, Monetary value (RFM) is a marketing analysis tool used to identify a firm's best clients based on the nature of their spending habits.
- An RFM analysis evaluates clients and customers by scoring them in three categories: how recently they've made a purchase, how often they buy, and the size of their purchases.



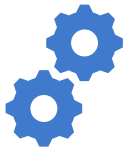
Tool Used for Analysis:-

- KNIME, the Konstanz Information Miner, is a free and open-source data analytics, reporting and integration platform**.**
- KNIME shall be used to carry our customer segmentation in subsequent slides

CUSTOMER SEGMENTATION USING RFM (2/9)

Parameters Used

- **Recency:-**
  - ✓ As per instructions the column 'Days since last order' has been ignored
  - ✓ New column Recency has been created as '[Max(order date)-order date]'
  - ✓ As per above formula, we have used '01-06-2020' as a reference date and created Recency column.
- **Frequency:-**
  - ✓ Frequency value is calculated by counting all the Orders made by a Customer, i.e. Aggregate Count(ORDERNUMBER) Group By (CUSTOMERNAME)
- **Monetary:-**
  - ✓ Monetary value is the same as SALES (i.e. Quantity X Price). Since this field is already present, need not compute it again.
- **Clustering:-**
  - ✓ K-Means Clustering carried out to create additional clusters based on Recency, Frequency & Monetary Values computed above, which can be correlated later for final segmentation.

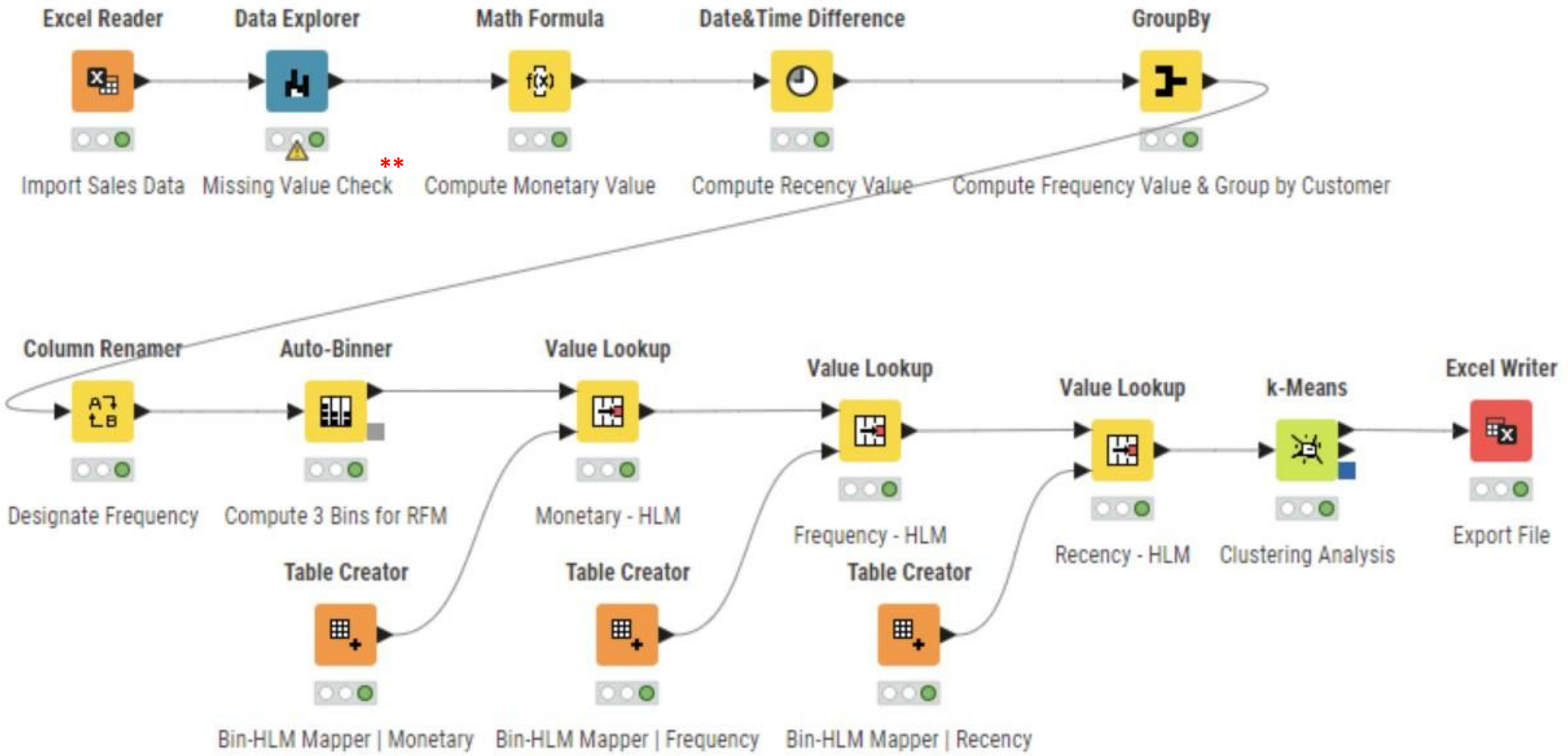


Assumptions Made

- All the orders having status apart form 'SHIPPED' were also considered as a valid business transactions.
- For calculating recency, we have taken the last 'ORDERDATE' as reference.
- 3 Bins (Quantiles-logic) has been created for Recency, Frequency & Monetary values.
- 3 Bins have been labeled as High (H), Medium (M), and Low (L).
- For Recency, Bin 1 has been considered highest as lower value is better.
- For Frequency & Monetary, Bin 3 has been considered highest as higher is better.

CUSTOMER SEGMENTATION USING RFM (3/9)

KNIME Workflow for RFM Analysis



** ORDERDATE is excluded from Data Explorer Node calculation

CUSTOMER SEGMENTATION USING RFM (4/9)

RFM Output Table from KNIME Workflow (Top 30 Rows)

Rows: 89 | Columns: 11

#	RowID	CUSTOMERNA... String	Frequency Number (integer)	Monetary Number (double)	Recency Number (long)	Frequency [Binn... String	Monetary [Binn... String	Recency [Binned] String	Monetary [HLM] String	Frequency [HLM] String	Recency [HLM] String	Cluster String
1	Row0	AV Stores, Co.	51	157,807.81	197	Bin 3	Bin 3	Bin 2	H	H	M	cluster_1
2	Row1	Alpha Cognac	20	70,488.44	65	Bin 1	Bin 1	Bin 1	L	L	H	cluster_3
3	Row2	Amica Models & Co.	26	94,117.26	266	Bin 2	Bin 2	Bin 3	M	M	L	cluster_2
4	Row3	Anna's Decorations, Lt	46	153,996.13	84	Bin 3	Bin 3	Bin 2	H	H	M	cluster_1
5	Row4	Atelier graphique	7	24,179.96	189	Bin 1	Bin 1	Bin 2	L	L	M	cluster_3
6	Row5	Australian Collectable	23	64,591.46	23	Bin 2	Bin 1	Bin 1	L	M	H	cluster_3
7	Row6	Australian Collectors,	55	200,995.41	185	Bin 3	Bin 3	Bin 2	H	H	M	cluster_1
8	Row7	Australian Gift Networ	15	59,469.12	120	Bin 1	Bin 1	Bin 2	L	L	M	cluster_3
9	Row8	Auto Assoc. & Cie.	18	64,834.32	234	Bin 1	Bin 1	Bin 3	L	L	L	cluster_3
10	Row9	Auto Canal Petit	27	93,170.66	55	Bin 2	Bin 2	Bin 1	M	M	H	cluster_2
11	Row10	Auto-Moto Classics In	8	26,479.26	181	Bin 1	Bin 1	Bin 2	L	L	M	cluster_3
12	Row11	Baane Mini Imports	32	116,599.19	209	Bin 2	Bin 2	Bin 2	M	M	M	cluster_2
13	Row12	Bavarian Collectables	14	34,993.92	260	Bin 1	Bin 1	Bin 3	L	L	L	cluster_3
14	Row13	Blauer See Auto, Co.	22	85,171.59	209	Bin 2	Bin 2	Bin 2	M	M	M	cluster_2
15	Row14	Boards & Toys Co.	3	9,129.35	114	Bin 1	Bin 1	Bin 2	L	L	M	cluster_3
16	Row15	CAF Imports	13	49,642.05	440	Bin 1	Bin 1	Bin 3	L	L	L	cluster_3
17	Row16	Cambridge Collectabl	11	36,163.62	390	Bin 1	Bin 1	Bin 3	L	L	L	cluster_3
18	Row17	Canadian Gift Exchan	22	75,238.92	223	Bin 2	Bin 2	Bin 2	M	M	M	cluster_2
19	Row18	Classic Gift Ideas, Inc	21	67,506.97	231	Bin 2	Bin 1	Bin 2	L	M	M	cluster_3
20	Row19	Classic Legends Inc.	20	77,795.2	193	Bin 1	Bin 2	Bin 2	M	L	M	cluster_2
21	Row20	Clover Collections, Co	16	57,756.43	259	Bin 1	Bin 1	Bin 3	L	L	L	cluster_3
22	Row21	Collectable Mini Desig	25	87,489.23	461	Bin 2	Bin 2	Bin 3	M	M	L	cluster_2
23	Row22	Collectables For Less	24	81,577.98	133	Bin 2	Bin 2	Bin 2	M	M	M	cluster_2
24	Row23	Corrida Auto Replicas	32	120,615.28	213	Bin 2	Bin 3	Bin 2	H	M	M	cluster_2
25	Row24	Cruz & Sons Co.	26	94,015.73	198	Bin 2	Bin 2	Bin 2	M	M	M	cluster_2
26	Row25	Daedalus Designs Imp	20	69,052.41	466	Bin 1	Bin 1	Bin 3	L	L	L	cluster_3
27	Row26	Danish Wholesale Imp	36	145,041.6	47	Bin 3	Bin 3	Bin 1	H	H	H	cluster_1
28	Row27	Diecast Classics Inc.	31	122,138.14	2	Bin 2	Bin 3	Bin 1	H	M	H	cluster_2
29	Row28	Diecast Collectables	18	70,859.78	402	Bin 1	Bin 2	Bin 3	M	L	L	cluster_3
30	Row29	Double Decker Gift St	12	36,019.04	496	Bin 1	Bin 1	Bin 3	L	L	L	cluster_3

CUSTOMER SEGMENTATION USING RFM (5/9)

Customer Segmentation Rationale

- RFM Analysis has been primarily used to distinguish between Best, Verge of Churning, Lost & Loyal Customers.
- K-Means Clustering was only used as a secondary recommendation model to validate segmentation.

- Customer Segmentation rationale has been summarized in the table below:-

Category	Recency	Frequency	Monetary	Notes
Best Customers	H	H	H	Most valuable and recently active customers
Customers on the Verge of Churning	L	H / M	H / M	Previously strong customers now inactive
Lost Customers	L	L	L	Inactive, low frequency and value
Loyal Customers	M	H	H	Consistently valuable, not most recent

CUSTOMER SEGMENTATION USING RFM (6/9)

Best Customers

- Citeria – Recency = H | Frequency = H | Monetary = H
- These customers **Buy frequently**, **Spend a lot** and **Purchased recently**
- They are engaged, valuable, and loyal right now — **the core of the business**.

- Below are the top 5 Best Customers, sorted by Recency value (ascending order)

CUSTOMERNAME	Frequency	Monetary	Recency	Monetary [HLM]	Frequency [HLM]	Recency [HLM]
Euro Shopping Channel	259	912294.11	1	H	H	H
La Rochelle Gifts	53	180124.90	1	H	H	H
Mini Gifts Distributors Ltd.	180	654858.06	3	H	H	H
Souveniers And Things Co.	46	151570.98	3	H	H	H
Salzburg Collectables	40	149798.63	15	H	H	H

CUSTOMER SEGMENTATION USING RFM (7/9)

Customers on the Verge of Churning

- Citeria – Recency = L | Frequency = H or M | Monetary = H or M
- These customers **Used to be active and valuable**, **Haven't purchased in a long time** and there's a **Risk they may stop buying permanently**
- They once offered value to the business, but now are at **the risk of getting churned**

- Below are the top 5 Customers on the Verge of Churning, sorted by Monetary value (descending order)

CUSTOMERNAME	Frequency	Monetary	Recency	Monetary [HLM]	Frequency [HLM]	Recency [HLM]
Saveley & Henriot, Co.	41	142874.25	457	H	H	L
Vida Sport, Ltd	31	117713.56	276	M	M	L
Herkku Gifts	29	111640.28	272	M	M	L
Marta's Replicas Co.	27	103080.38	232	M	M	L
Amica Models & Co.	26	94117.26	266	M	M	L

CUSTOMER SEGMENTATION USING RFM (8/9)

Lost Customers

- Citeria – Recency = L | Frequency = L | Monetary = L
- These customers **Rarely bought**, **Didn't spend much** and **Haven't returned in a long time**
- They have **low value across the board** and likely churned completely

- Below are the top 5 Lost Customers, sorted by Recency value (descending order)

CUSTOMERNAME	Frequency	Monetary	Recency	Monetary [HLM]	Frequency [HLM]	Recency [HLM]
Double Decker Gift Stores, Ltd	12	36019.04	496	L	L	L
West Coast Collectables Co.	13	46084.64	489	L	L	L
Signal Collectibles Ltd.	15	50218.51	477	L	L	L
Daedalus Designs Imports	20	69052.41	466	L	L	L
CAF Imports	13	49642.05	440	L	L	L

CUSTOMER SEGMENTATION USING RFM (9/9)

Loyal Customers

- Citeria – Recency = M | Frequency = H | Monetary = H
- These customers **Buy often**, **Spend a lot** and **May not always be very recent but are consistently valuable**
- They form the **backbone of the revenue stream**

- Below are the top 5 Loyal Customers, sorted by Monetary value (descending order)

CUSTOMERNAME	Frequency	Monetary	Recency	Monetary [HLM]	Frequency [HLM]	Recency [HLM]
Australian Collectors, Co.	55	200995.41	185	H	H	M
Muscle Machine Inc	48	197736.94	183	H	H	M
Dragon Souveniers, Ltd.	43	172989.68	91	H	H	M
Land of Toys Inc.	49	164069.44	199	H	H	M
AV Stores, Co.	51	157807.81	197	H	H	M

INFERENCES & BUSINESS RECOMMENDATIONS

Business Recommendations based on Customer Segmentation

BEST CUSTOMERS

GOAL

Maximize their **lifetime value**, reduce the chance of defection, and turn them into **brand advocates**.

STRATEGY

- **Retention-focused:** Keep them loyal and delighted
- **VIP Programs:** Offer early access, exclusive products, or events
- **Surprise & Delight:** Thank-you gifts, handwritten notes, birthday offers.
- **Referral Incentives:** Encourage them to bring in others with similar traits.
- **Feedback Loops:** Use their input to improve products/services.

LOYAL CUSTOMERS

GOAL

Deepen the relationship, encourage higher spending and longer retention.

STRATEGY

- **Loyalty programs:** Points-based rewards or tiered benefits
- **Subscription or bundling options:** Make it easier to buy regularly
- **Exclusive experiences:** Beta testing, private previews, “insider” content
- **Recognition:** Feature them in customer stories or shout-outs
- **Cross-selling/Upselling:** Suggest complementary high-margin products

CUSTOMERS ON THE VERGE OF CHURNING

GOAL

Re-engage them before they slip into the "lost" category

STRATEGY

- **Reactivation campaigns:** Send personalized "We miss you" messages
- **Time-limited offers:** Discounts or incentives with urgency
- **Product reminders:** Recommend items based on their past purchases
- **Multi-channel outreach:** Use SMS, email, social, or retargeted ads
- **Ask why:** Quick surveys to understand their disengagement

LOST CUSTOMERS

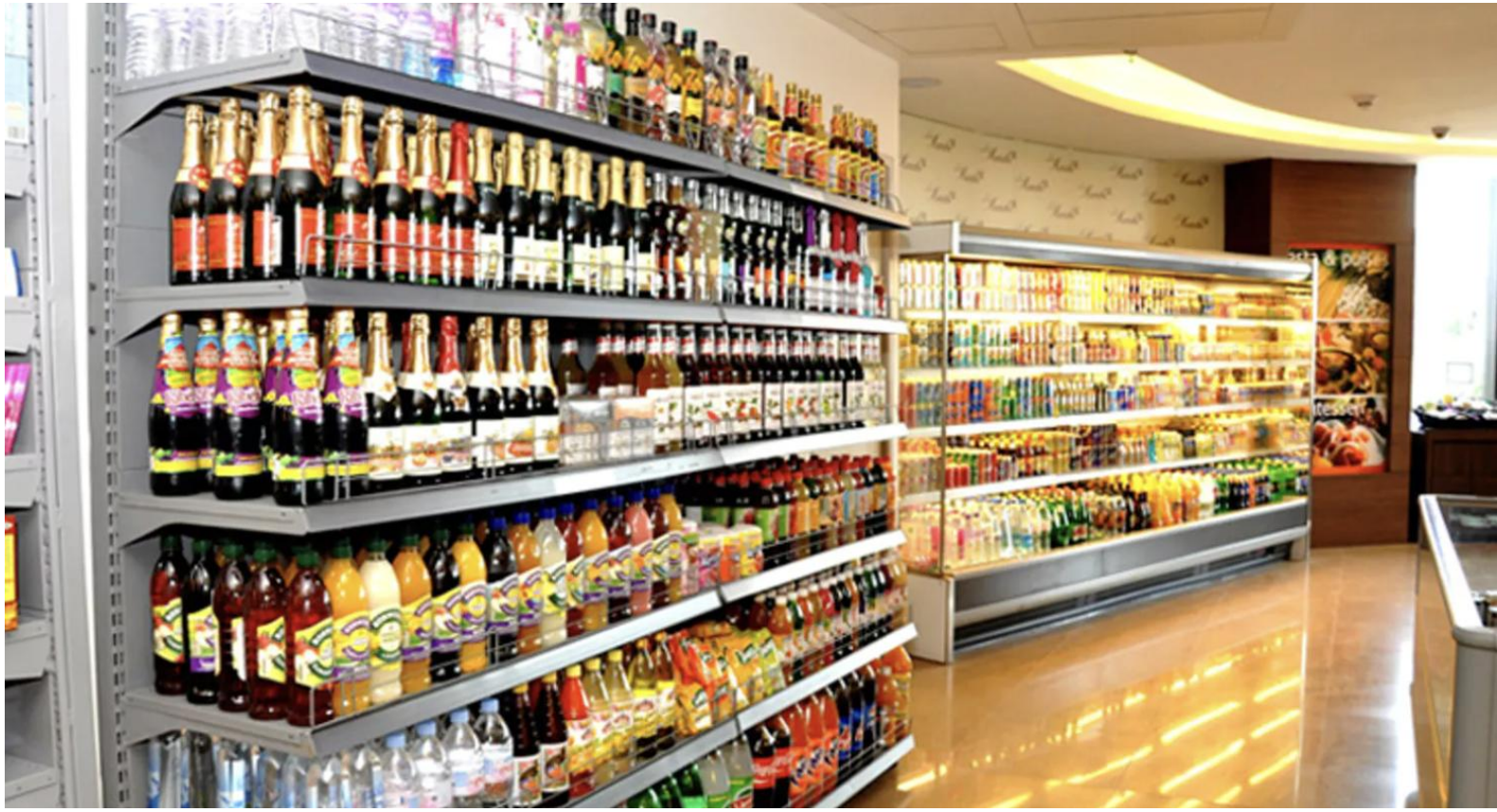
GOAL

Limit acquisition/retention costs; only re-engage if ROI is favorable.

STRATEGY

- **Low-cost reactivation:** Batch-target them with seasonal or clearance promotions.
- **Re-acquisition campaigns:** Treat like new leads with onboarding-style emails
- **Exit feedback:** Ask why they left (if willing)
- **Segment wisely:** Don’t overspend trying to win them back

PART B – Grocery Retail Problem



BUSINESS CONTEXT & PROBLEM STATEMENT

Business Context

- A Grocery Store – shared transactional data
- Wish to leverage Analytics to understand customer buying patterns for enhancing sales, increasing customer satisfaction and improving profitability
- Aims to identifying frequently purchased item combinations to craft effective marketing strategies, optimize inventory management and tailor promotions to meet customer needs
- Drive customer-centric offerings which can increase basket size and improve customer retention.

Problem Statement / Objective

- Analyze the POS transactional data to identify frequently purchased item combinations
- Using association rule mining or similar techniques to uncover patterns that will help the store create targeted combo offers and discounts
- Ultimate goal is to drive revenue growth by increasing customer purchases and average basket size

DATA SUMMARY (1/2)

Data Dictionary & Data Information

Data Dictionary:-

- Date: The date when the transaction took place
- Order_id: A unique identifier for each customer order
- Product: The individual item purchased in the transaction

```
Data columns (total 3 columns):
 #   Column    Non-Null Count  Dtype
---  ------    --------------  -----
 0   Date      20641 non-null  object
 1   Order_id  20641 non-null  int64
 2   Product   20641 non-null  object
dtypes: int64(1), object(2)
```

```
Missing Values:-
Date        0
Order_id    0
Product     0
```

Duplicate Values:-

- Dropping duplicate rows may not be appropriate as there is no unique identifier for each row.
- Each row consists of a date, a customer ID, and a product purchased, but the **same product can be purchased by multiple customers on the same date**.
- If we drop duplicate rows; it may inadvertently remove valid information from the dataset.
- Hence, **duplicate values have not been removed** from the dataset

```
Duplicated Values:-
4730
```

DATA SUMMARY (2/2)

Data Description

	count	unique	top	freq	mean	std	min	25%	50%	75%	max
Date	20641	603	08-02-2019	183	NaN	NaN	NaN	NaN	NaN	NaN	NaN
Order_id	20641.0	NaN	NaN	NaN	575.986289	328.557078	1.0	292.0	581.0	862.0	1139.0
Product	20641	37	poultry	640	NaN	NaN	NaN	NaN	NaN	NaN	NaN

Data Inference

- **Data**: Past 3 years.
- **Dataset**: 3 columns and 20641 rows
- **Missing values**: None
- **Duplicate values:** 4730 | More than one quantity of a product bought in a single order. Hence, not removed (please refer previous slide)

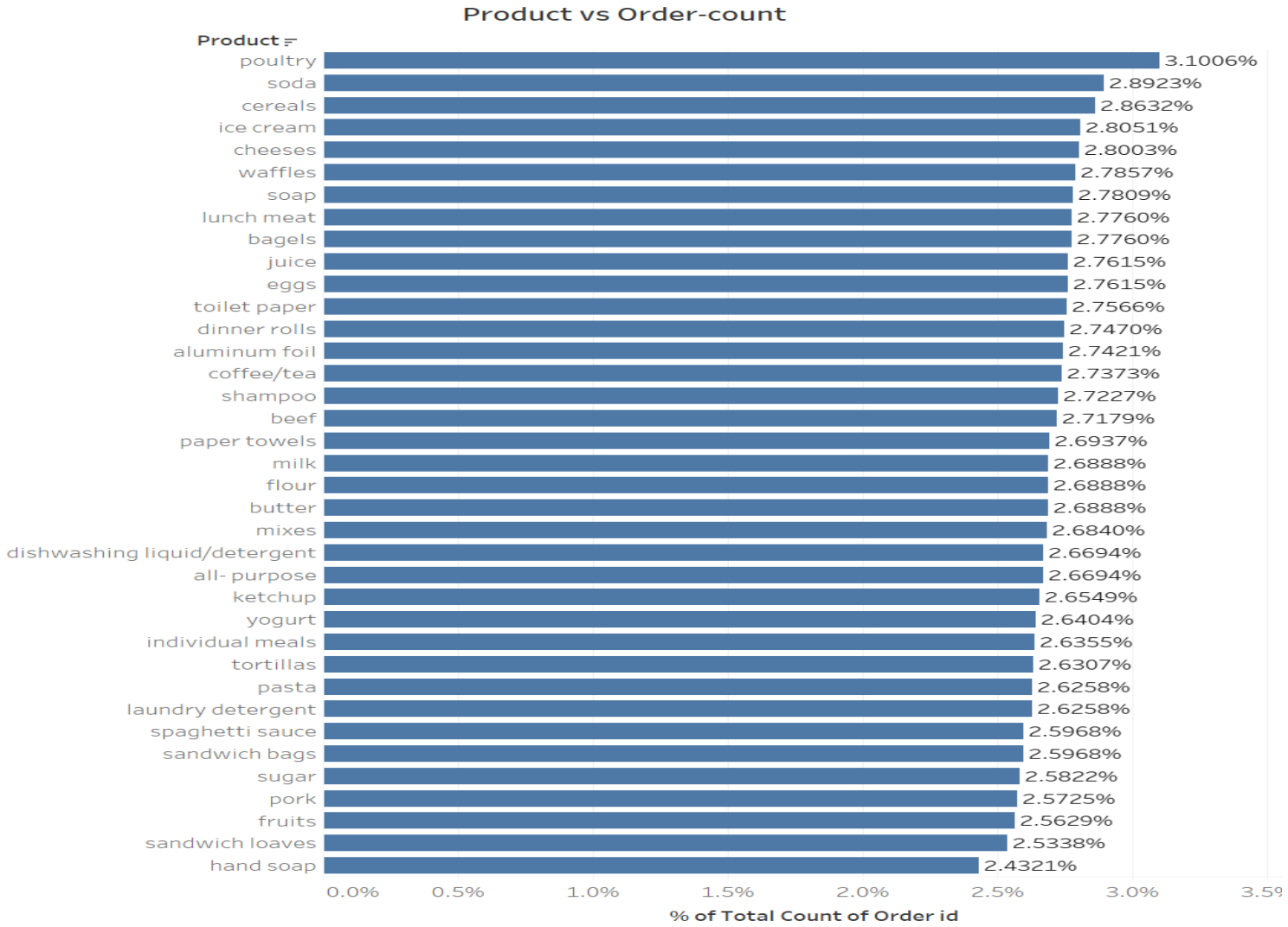
- **Order_id is an identifier** and finding outliers is meaningless
- There are a total of **37 unique products** in the Grocery Store, among which **poultry** is the highest sold product.
- Maximum no. of products were bought on **08/02/2019** (183 items)

Assumptions: -

- Data represents a list of items purchased at a grocery store on various dates.
- Each entry represents a single item purchased.
- The same item can be purchased by multiple customers on different dates.
- Duplicated values were not dropped to avoid any loss of information

EXPLORATORY DATA ANALYSIS (1/8)

Univariate Analysis



Product:-

- Poultry dominates the item list purchased, followed by Soda & Cereals
- Top 3 products – poultry, soda, cereals

Order ID:-

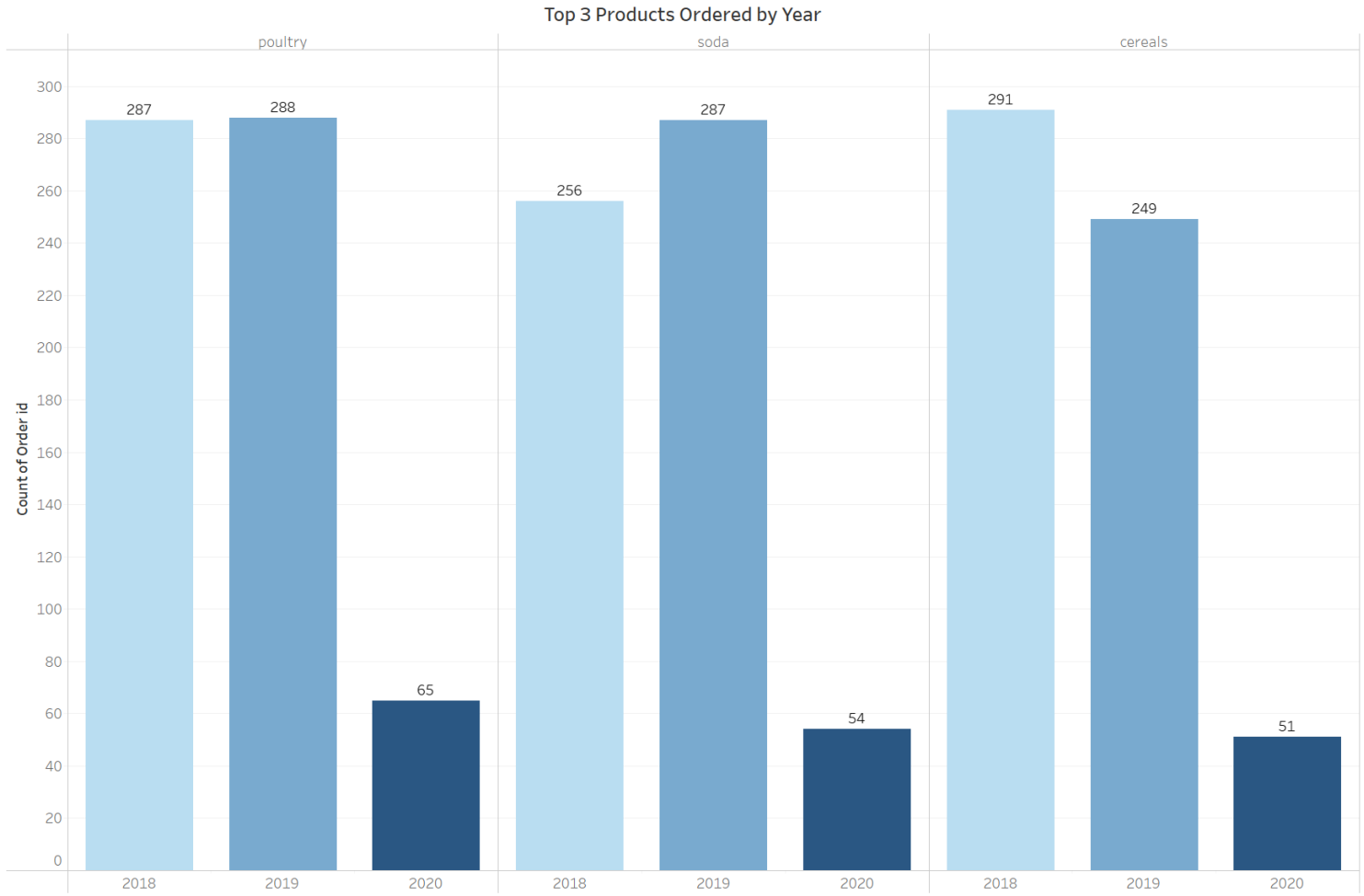
- Since, it is an identifier, its distribution is meaningless

Date:-

- Distribution of date is meaningless.
- Bivariate analysis with Date would be covered under Trends section

EXPLORATORY DATA ANALYSIS (2/8)

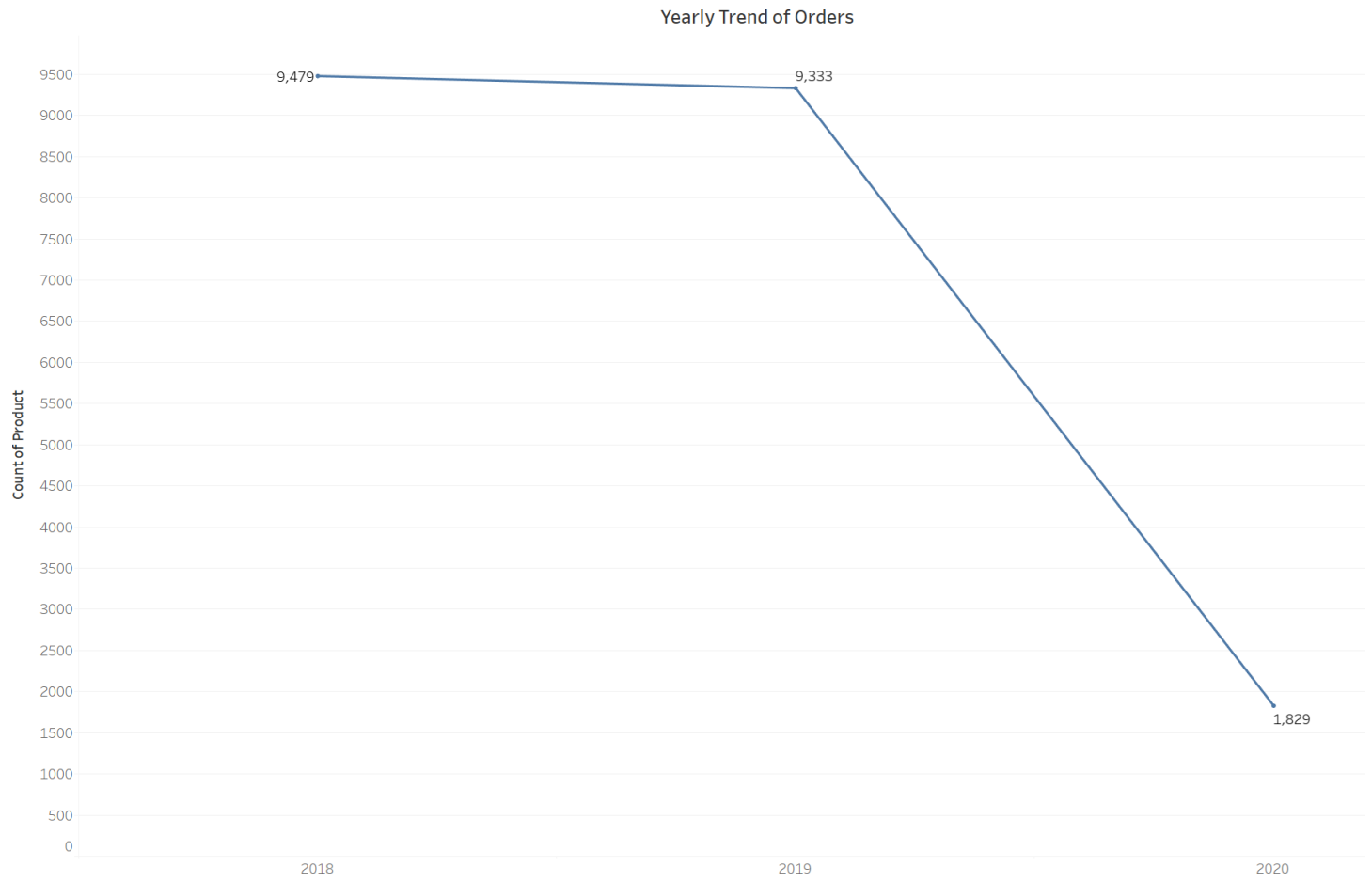
Bivariate/Multivariate Analysis



- Even though the data suggests a sharp fall in each of the product orders (poultry, soda, cereals), it cannot be ascertained as the data for 2020 is not compete (only till Q1). **Interpolating** the data may suggest a slight **decline on the overall trend**.
- **Poultry** – stable orders in 2018 & 2019
- **Soda** – slight increase in orders from 2018 to 2019
- **Cereals** – Sharp decline seen in orders from 2018 to 2019

EXPLORATORY DATA ANALYSIS (3/8)

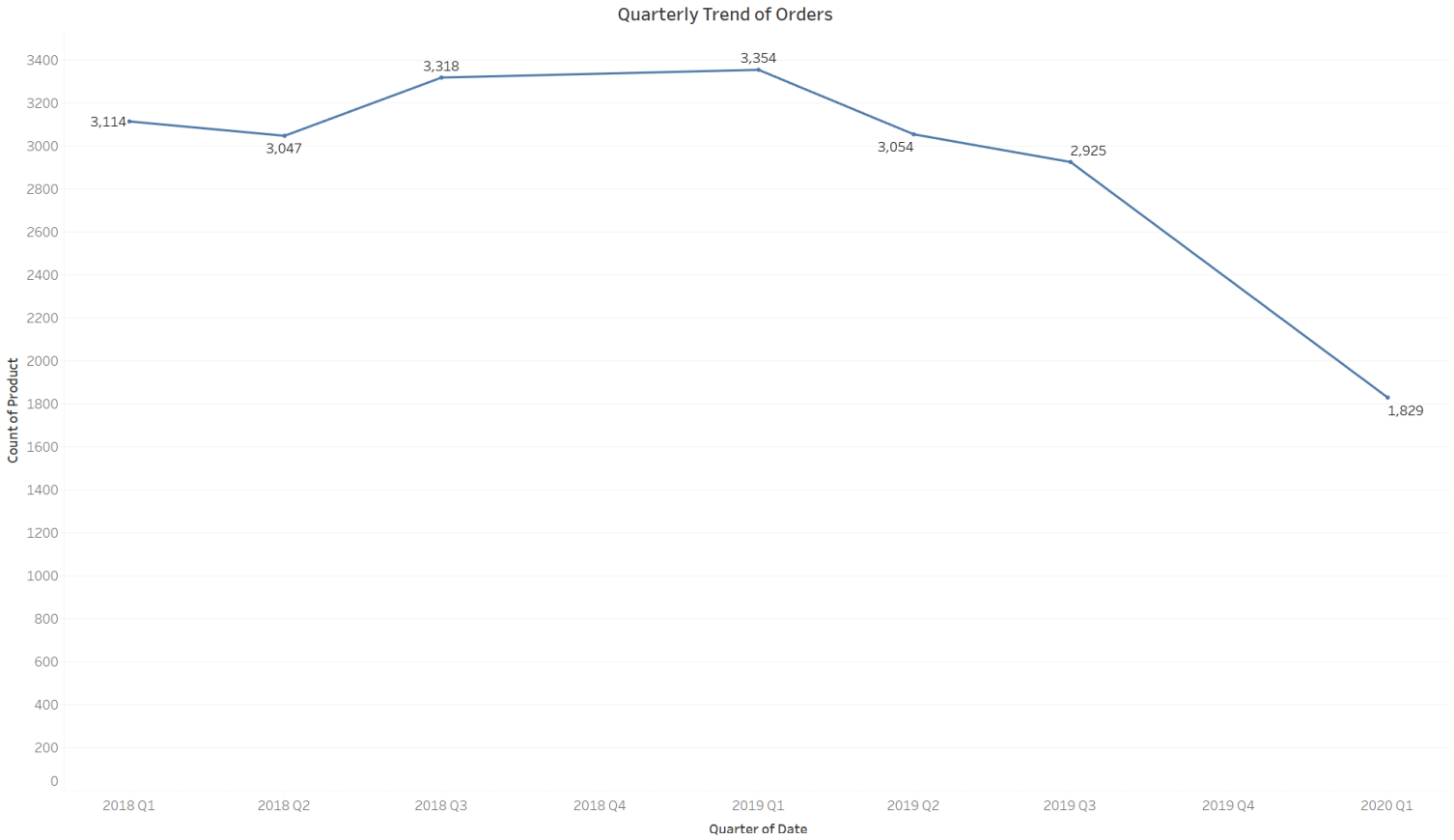
Bivariate/Multivariate Analysis – Yearly Trends



- Products sold in 2018 & 2019 remained fairly stable at around 9300-9500 orders
- The chart shows a decline, but cannot be ascertained as data is not complete for 2020 (data till Q1 only)
- However, if we interpolate, it may imply 2020 observing decline in the order trend. However, this is just a conjecture & cannot be ascertained at this moment. Time series forecasting techniques can be deployed to establish a level of confidence in this conjecture.

EXPLORATORY DATA ANALYSIS (4/8)

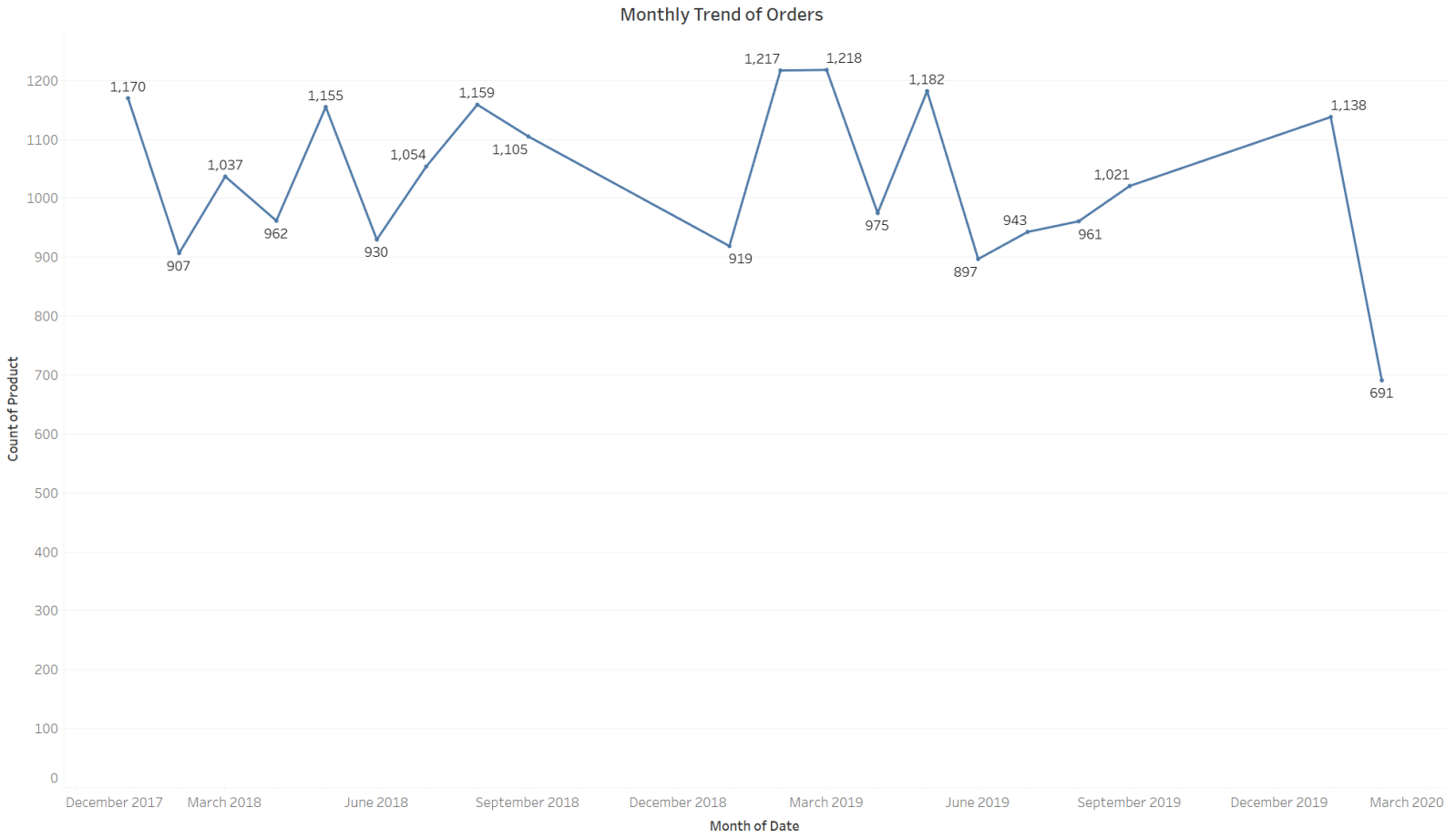
Bivariate/Multivariate Analysis – Quarterly Trends



- Products sold in 2018 over Q1 through Q4 was fairly stable
- Products sold in 2019 Q1 & Q2 were similar to 2018, however, a sharp fall observed in 2019 Q3 & Q4.
- Poor performance & further decline extended till 2020 Q1.
- Data clearly suggest a fall in performance, requiring immediate attention.

EXPLORATORY DATA ANALYSIS (5/8)

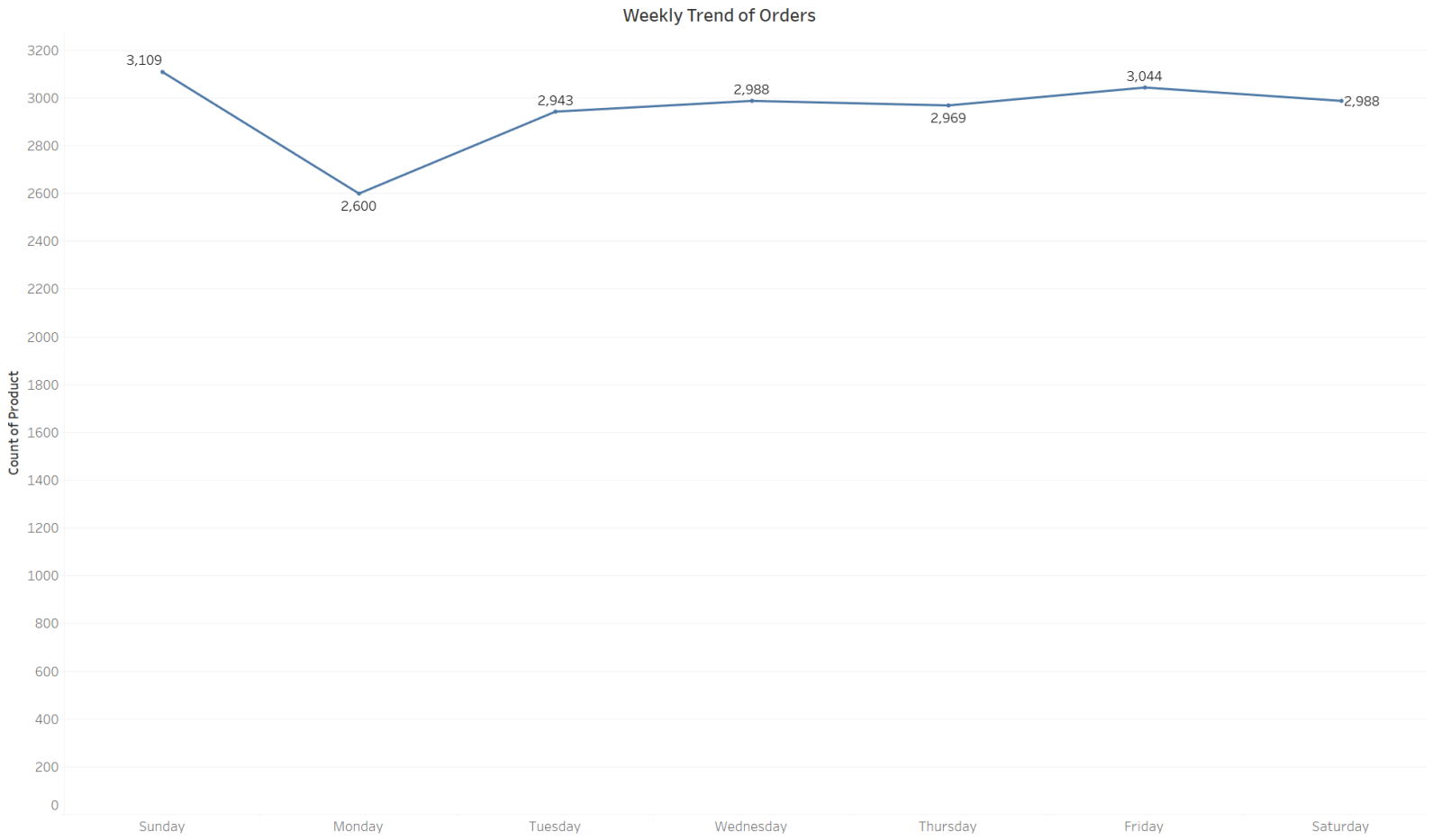
Bivariate/Multivariate Analysis – Monthly Trends



- Monthly trend is erratic & suggests no pattern
- No seasonality observed across the year
- This makes sense also since most of the products sold are FMCG, that have pretty much the same demand all round the year.

EXPLORATORY DATA ANALYSIS (6/8)

Bivariate/Multivariate Analysis – Weekday Trends



- Products sold on Sunday is the highest, which makes sense as it is an off day for most of the working people
- Friday is the second highest after Sunday, suggesting people buying Friday evening as they gear up for the weekend.
- Lowest no. of products are sold on Monday.

EXPLORATORY DATA ANALYSIS (7/8)

Key Takeaways

Insights – General

Product Sales declining trend from 2019 Q3 onwards

Top 3 Products Sold – Poultry, Soda, Cereals

Poultry, Soda – stable/increasing trend in sales; Cereals – declining trend

No monthly seasonal trend

Sunday & Friday record maximum sales relative to other days

Monday records lowest sales relative to other days

EXPLORATORY DATA ANALYSIS (8/8)

Business Recommendations

- **Investigate Causes of Overall Sales Decline** since Q3 2019 and implement corrective strategies such as refreshed product offerings or targeted promotions
- **Leverage High-Performing Products like Poultry and Soda** through prominent shelf placement and bundled offers to boost basket size
- **Reevaluate Cereals Category** by analyzing customer preferences, pricing, and competitor offerings to revive sales
- **Avoid Overreliance on Seasonal Campaigns**, as no monthly seasonality exists; instead, maintain consistent promotions year-round
- **Focus Promotions on High-Traffic Days (Sunday & Friday)** to capitalize on peak sales periods and increase average spend
- **Introduce Monday-Specific Deals or Loyalty Rewards** to uplift sales on the lowest-performing day
- **Enhance Inventory Planning for Poultry and Soda**, ensuring availability during high-demand days to prevent stockouts
- **Use POS Data for Personalized Offers**, targeting customers based on purchase history to increase frequency and retention

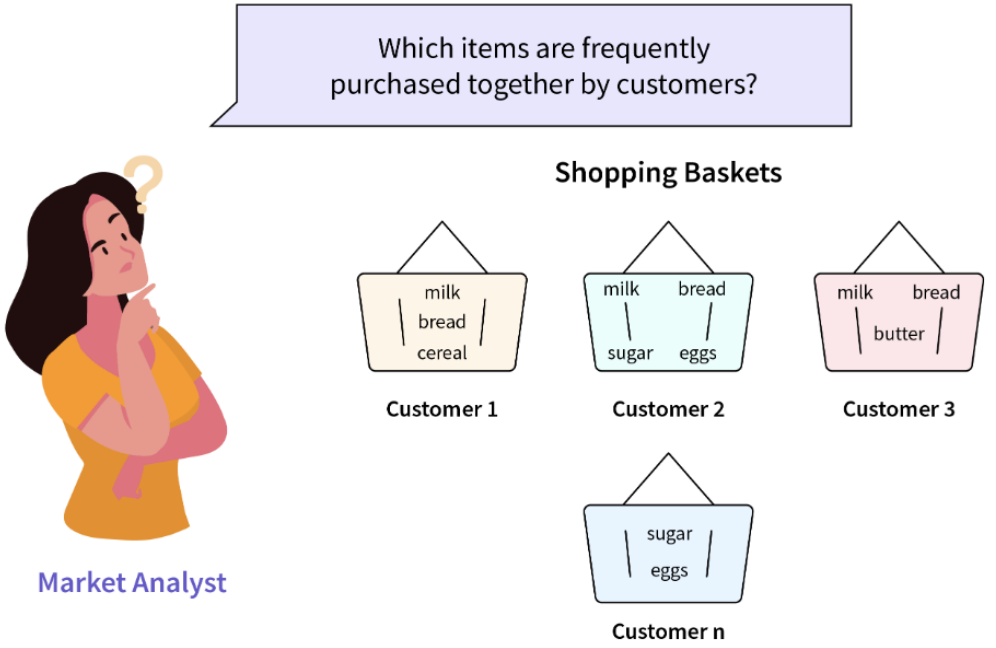
MARKET BASKET ANALYSIS (1/5)

What is Market Basket Analysis?

- Market Basket Analysis is a techniques used by large retailers to uncover the association between items or to **identify the relationship between items** which are bought together more frequently.
- **Support** – Support measures the frequency of occurrence of an item or itemset within a dataset of transactions. It's the proportion of transactions that contain the specific item or itemset.
- **Confidence** – Confidence quantifies the probability that, given the purchase of an antecedent item, the consequent item is also purchased in the same transaction. It's the ratio of transactions containing both items (antecedent and consequent) to the transactions containing only the antecedent.
- **Lift** – Lift measures the strength of the association between two items by comparing the observed probability of their co-occurrence to the expected probability if they were independent. It's the ratio of the observed support of the association to the product of the individual supports of the antecedent and consequent.

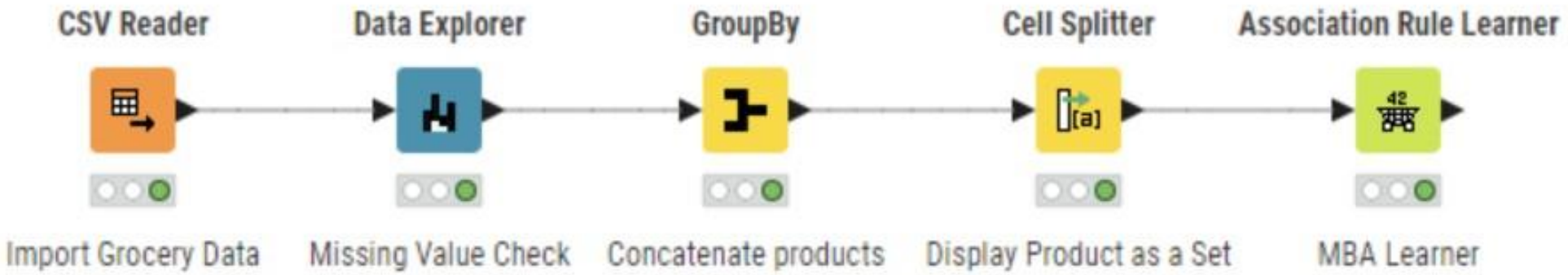
Tool Used for Analysis:-

- KNIME, the Konstanz Information Miner, is a free and open-source data analytics, reporting and integration platform**.**
- KNIME shall be used to carry our Market Basket Analysis in subsequent slides



MARKET BASKET ANALYSIS (2/5)

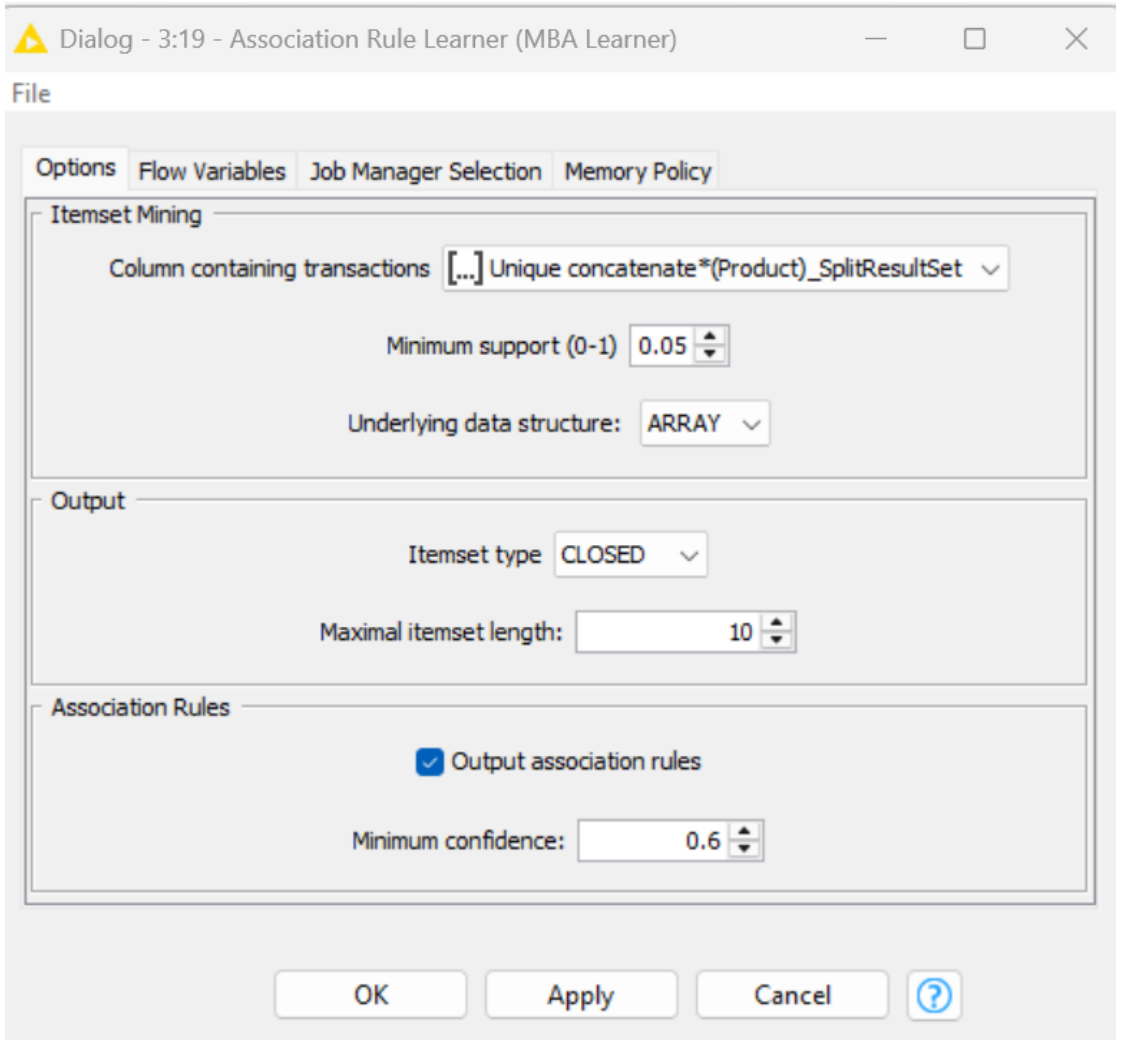
KNIME Workflow for Market Basket Analysis



MARKET BASKET ANALYSIS (3/5)

Threshold Values Used

- The above KNIME Workflow has been used to determine the rule using **Association Learner Rule**.
- Following Thresholds have been used: -
  - ✓ **Minimum Support** = 0.05
    - Def: The minimum percentage of transactions in which an itemset must appear to be considered frequent.
    - Value 0.05 means: The itemset must appear in at least 5% of all transactions.
    - Implication: Filters out very rare itemsets; Helps focus only on patterns that are relatively common; A lower value gives more rules (including rare ones), while, a higher value gives fewer, more significant rules.
  - ✓ **Minimum Confidence** = 0.6
    - Def: The minimum probability that the consequent occurs given the antecedent.
    - Value 0.6 means: At least 60% of the time, when the antecedent items are bought, the consequent item is also bought.
    - Implication: Filters out weak or uncertain rules; Ensures the rules one gets are reliable for making recommendations; A higher confidence means stronger and more actionable rules.
  - ✓ **Maximum itemset length** = 10
    - Def: The maximum number of items in any frequent itemset or rule (antecedent + consequent).
    - Value 10 means: The rule can include up to 10 items.
    - Implication: Allows for more complex and broader rules, capturing long purchase patterns



MARKET BASKET ANALYSIS (4/5)

Market Basket Analysis Output Table from KNIME Workflow

#	RowID	Support *Number (double)*	Confidence *Number (double)*	Lift *Number (double)*	Consequent *String*	implies *String*	Items *Set*
1	rule0	0.05	0.64	1.7	juice	<---	[yogurt,toilet paper,aluminum foil]
2	rule1	0.05	0.62	1.645	juice	<---	[yogurt,poultry,aluminum foil]
3	rule2	0.05	0.613	1.616	coffee/tea	<---	[yogurt,cheeses,cereals]
4	rule3	0.05	0.6	1.424	poultry	<---	[dishwashing liquid/detergent,laundry dete
5	rule4	0.051	0.63	1.678	mixes	<---	[yogurt,poultry,aluminum foil]
6	rule5	0.051	0.611	1.66	sandwich bags	<---	[cheeses,bagels,cereals]
7	rule6	0.051	0.674	1.726	cheeses	<---	[bagels,cereals,sandwich bags]
8	rule7	0.051	0.617	1.558	cereals	<---	[cheeses,bagels,sandwich bags]
9	rule8	0.051	0.63	1.621	dinner rolls	<---	[spaghetti sauce,poultry,cereals]
10	rule9	0.051	0.637	1.512	poultry	<---	[dinner rolls,spaghetti sauce,cereals]
11	rule10	0.051	0.604	1.589	milk	<---	[poultry,laundry detergent,cereals]
12	rule11	0.052	0.628	1.61	eggs	<---	[dinner rolls,poultry,soda]
13	rule12	0.052	0.641	1.649	dinner rolls	<---	[spaghetti sauce,poultry,ice cream]
14	rule13	0.052	0.686	1.628	poultry	<---	[dinner rolls,spaghetti sauce,ice cream]
15	rule14	0.052	0.628	1.614	dinner rolls	<---	[spaghetti sauce,poultry,juice]
16	rule15	0.052	0.602	1.429	poultry	<---	[dinner rolls,spaghetti sauce,juice]
17	rule16	0.052	0.634	1.627	eggs	<---	[paper towels,dinner rolls,pasta]
18	rule17	0.052	0.602	1.621	pasta	<---	[paper towels,eggs,dinner rolls]
19	rule18	0.054	0.642	1.651	dinner rolls	<---	[spaghetti sauce,poultry,laundry detergent
20	rule19	0.054	0.656	1.556	poultry	<---	[dinner rolls,spaghetti sauce,laundry deter
21	rule20	0.055	0.624	1.565	ice cream	<---	[paper towels,eggs,pasta]
22	rule21	0.055	0.63	1.616	eggs	<---	[paper towels,ice cream,pasta]
23	rule22	0.055	0.643	1.731	pasta	<---	[paper towels,eggs,ice cream]
24	rule23	0.055	0.649	1.791	paper towels	<---	[eggs,ice cream,pasta]

How to interpret/read an association? (For instance, lets interpret Rule 0 with Consequent juice, and, Antecedents yogurt, toilet paper, aluminium foil):-

- **Support of 0.05** signifies the itemset (yogurt, toilet paper, aluminium foil) has appeared in 5% of all transactions.
- **Confidence of 0.64** signifies that 64% of the time, when the antecedent items (yogurt, toilet paper, aluminum foil) are bought, the consequent item (juice) is also bought.
- **Lift of 1.7** means that items are bought together [yogurt, toilet paper, aluminum foil & juice] 1.7 times more frequently than would be expected if the item was purchased independently [juice]

MARKET BASKET ANALYSIS (5/5)

Key Insights from the Analysis from Market Basket Analysis Output Table

1. **Popular Core Items:**
   - Poultry, Dinner Rolls, Eggs, Juice, Ice Cream, and Pasta appear frequently in both antecedents and consequents, indicating they are core products in shopping baskets.

2. **Strong Product Associations:**
   - Spaghetti Sauce + Poultry → Juice (high confidence and lift)
   - Paper Towels + Eggs → Ice Cream, also appears repeatedly.
   - These indicate reliable product pairings in consumer behavior.

3. **Repeated Antecedent Patterns:**
   - Itemsets like Yogurt, Poultry, Aluminum Foil occur multiple times with different consequents – suggesting strong bundle potential.

4. **Lift Values > 1.6 in Many Rules:**
   - Lift values above 1 suggest that the presence of the antecedent strongly increases the chance of the consequent being bought—ideal for cross-selling.

INFERENCES & BUSINESS RECOMMENDATIONS (1/3)

Business Recommendations

1

Combo Bundles with Discounts

❑ Create curated combo packs based on highly associated items.

Combo Name	Items	Offer Type
Sunday Roast Pack	Poultry + Dinner Rolls + Juice	Flat 10% off
Breakfast Basics	Eggs + Paper Towels + Pasta	Buy 2 get 1 free
Family Dinner Delight	Spaghetti Sauce + Poultry + Ice Cream	Save ₹30 combo
Weekly Essentials	Yogurt + Toilet Paper + Aluminum Foil	Bundle offer at ₹199

INFERENCES & BUSINESS RECOMMENDATIONS (2/3)

Business Recommendations

2

Cross-Sell Strategies

❑ In-Store / App Recommendations:
- ✓ If a customer adds Poultry → suggest Juice or Dinner Rolls.
- ✓ If cart has Eggs + Pasta → promote Ice Cream as a sweet ending.

❑ Email / SMS Offers:
- ✓ "Buying Pasta? Get 20% off on Eggs when you add both!"

3

Store Layout Optimization

❑ Place frequently co-purchased items in close proximity:
- ✓ Pasta → Eggs → Paper Towels
- ✓ Poultry → Dinner Rolls → Juice

❑ For online apps: Show "Frequently Bought Together" banners using these combinations

INFERENCES & BUSINESS RECOMMENDATIONS (3/3)

Business Recommendations

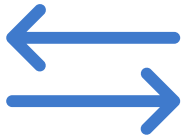
4

Loyalty and Seasonal Promotions

❑ Festival / Holiday Kits:
- ✓ Use high-confidence associations for curated holiday kits.

❑ Loyalty Points:
- ✓ Extra points on combos that include 2 or more associated items.

5



Stocking and Inventory Planning

❑ Use rule strength (support, confidence) to stock Poultry, Eggs, Juice, etc., in proportion to associated items.

❑ Ensure associated items are not out-of-stock together to prevent lost sales opportunities.


Thank you!