MRA - Main Project

Problem Statement - MRA Main Project

Part - A

Business Context:

An automobile parts manufacturing company has been actively selling products to a diverse range of customers for the past three years. Despite its growth, the company lacks the in-house expertise to derive actionable insights from its transaction data. As a result, they wish to uncover hidden patterns and trends in their customer transactions. By analyzing this data, the company aims to better understand customer behavior, improve customer segmentation, and implement targeted marketing strategies. These insights will help the company not only enhance customer satisfaction but also drive revenue growth by offering more personalized and efficient services.

Objective:

The primary objective of this analysis is to leverage data science techniques to:

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

Data Description:

ORDERNUMBER: Unique identifier for each order.

QUANTITYORDERED: Number of items ordered in a specific transaction.

PRICEEACH: Price per unit of the product in the order.

ORDERLINENUMBER: Sequence number of the product in the order.

SALES: Total sales value for the order.

ORDERDATE: Date when the order was placed.

DAYS_SINCE_LASTORDER: Number of days since the customer's previous order.

STATUS: Current status of the order (e.g., Shipped, Disputed).

PRODUCTLINE: Product category to which the item belongs (e.g., Motorcycles, Classic Cars).

MSRP: Manufacturer's Suggested Retail Price for the product.

PRODUCTCODE: Unique identifier for the product.

CUSTOMERNAME: Name of the customer placing the order.

PHONE: Customer's contact phone number.

ADDRESSLINE1: Customer's primary address.

CITY: City of the customer's address.

POSTALCODE: Postal code of the customer's address.

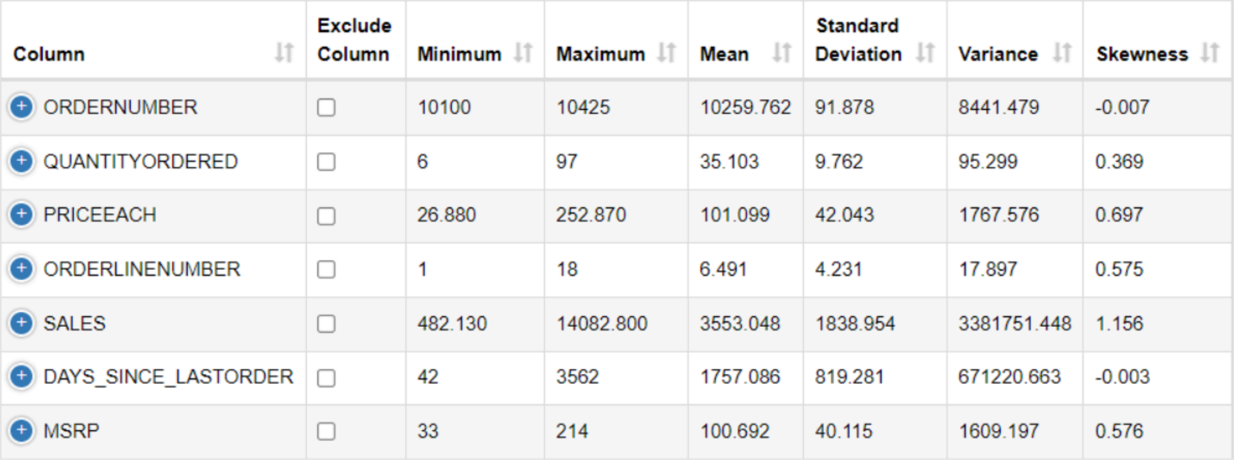
COUNTRY: Country of the customer's address.

CONTACTLASTNAME: Last name of the customer's contact person.

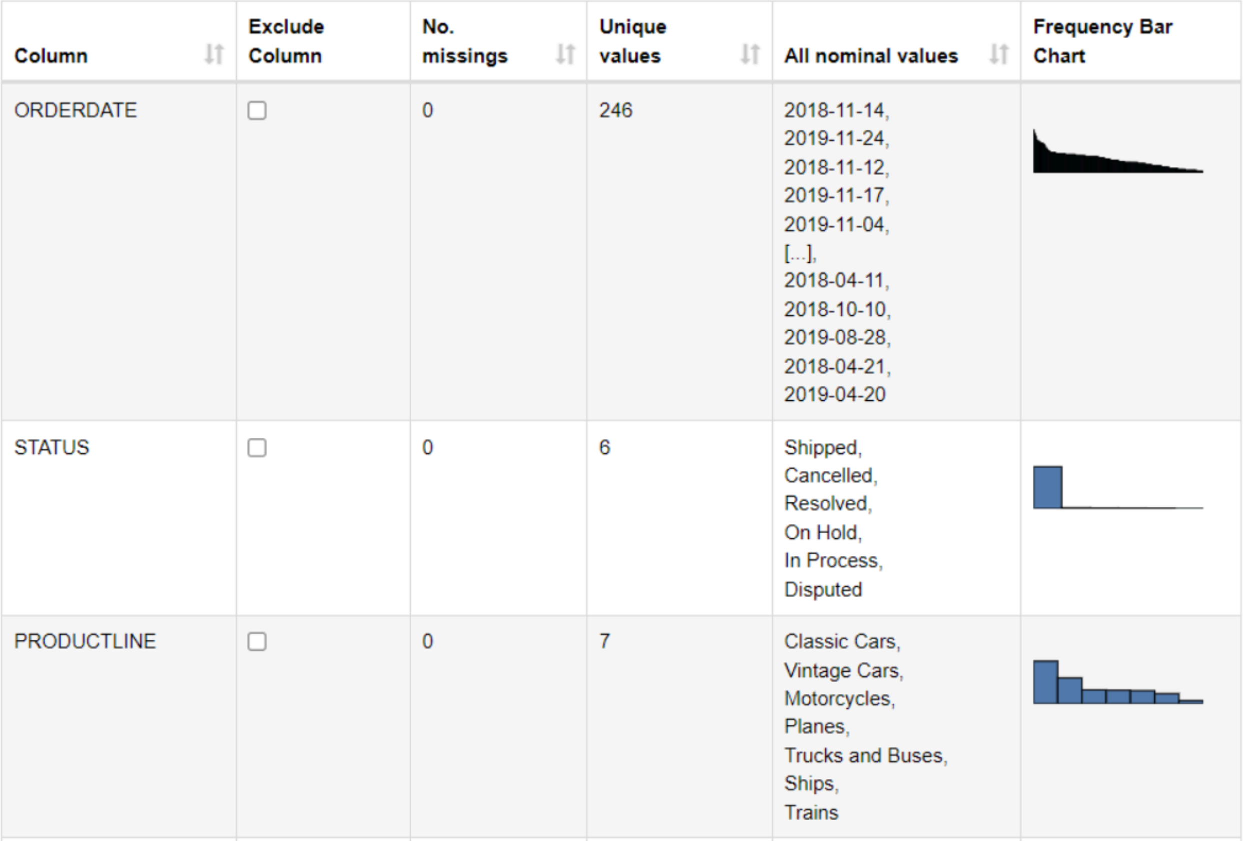
CONTACTFIRSTNAME: First name of the customer's contact person.

DEALSIZE: Size category of the transaction (e.g., Small, Medium, Large).

Summary Statistics



Column	Exclude Column	Minimum	Maximum	Mean	Standard Deviation	Variance	Skewness
ORDERNUMBER	□	10100	10425	10259.762	91.878	8441.479	-0.007
QUANTITYORDERED	□	6	97	35.103	9.762	95.299	0.369
PRICEEACH	□	26.880	252.870	101.099	42.043	1767.576	0.697
ORDERLINENUMBER	□	1	18	6.491	4.231	17.897	0.575
SALES	□	482.130	14082.800	3553.048	1838.954	3381751.448	1.156
DAYS_SINCE_LASTORDER	□	42	3562	1757.086	819.281	671220.663	-0.003
MSRP	□	33	214	100.692	40.115	1609.197	0.576



Column	Exclude Column	No. missings	Unique values	All nominal values	Frequency Bar Chart
ORDERDATE	□	0	246	2018-11-14, 2019-11-24, 2018-11-12, 2019-11-17, 2019-11-04, [...], 2018-04-11, 2018-10-10, 2019-08-28, 2018-04-21, 2019-04-20	
STATUS	□	0	6	Shipped, Cancelled, Resolved, On Hold, In Process, Disputed	
PRODUCTLINE	□	0	7	Classic Cars, Vintage Cars, Motorcycles, Planes, Trucks and Buses, Ships, Trains	

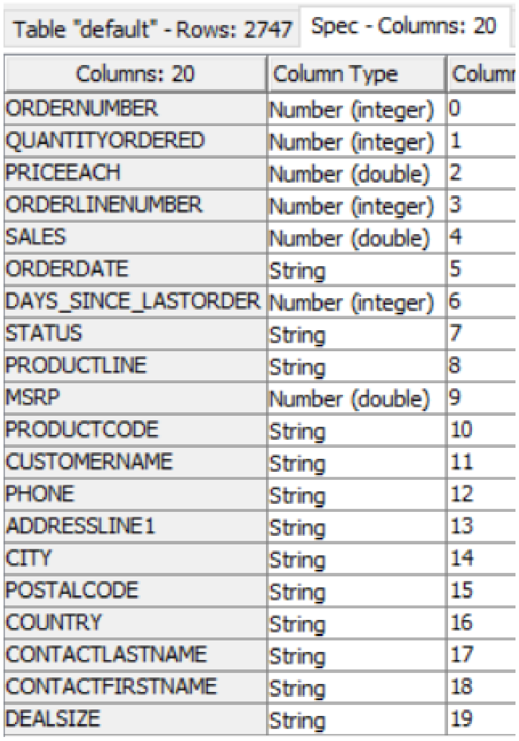


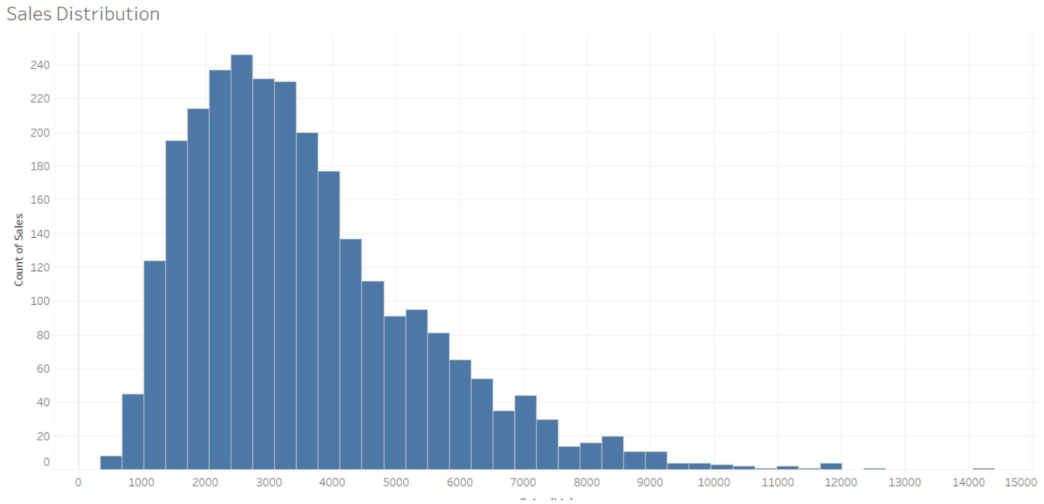
Table "default" - Rows: 2747 | Spec - Columns: 20

Columns: 20	Column Type	Colum
ORDERNUMBER	Number (integer)	0
QUANTITYORDERED	Number (integer)	1
PRICEEACH	Number (double)	2
ORDERLINENUMBER	Number (integer)	3
SALES	Number (double)	4
ORDERDATE	String	5
DAYS_SINCE_LASTORDER	Number (integer)	6
STATUS	String	7
PRODUCTLINE	String	8
MSRP	Number (double)	9
PRODUCTCODE	String	10
CUSTOMERNAME	String	11
PHONE	String	12
ADDRESSLINE1	String	13
CITY	String	14
POSTALCODE	String	15
COUNTRY	String	16
CONTACTLASTNAME	String	17
CONTACTFIRSTNAME	String	18
DEALSIZE	String	19

- There are a total of 2747 rows in the dataset with 20 variables.
- Each variable has its datatype; there are 7 numerical and 13 categorical.
- There are no missing values.

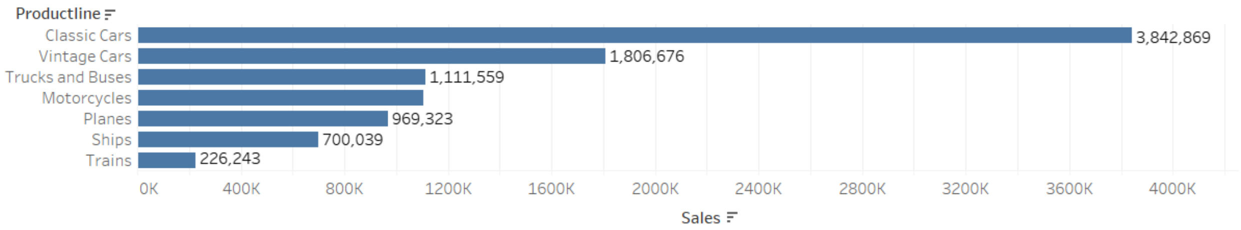
Univariate Analysis

1. Sales distribution



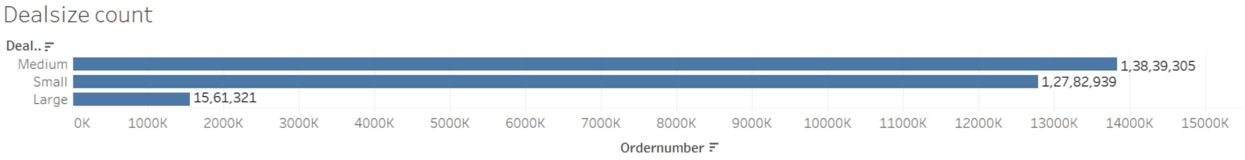
- It is right-skwed so there are many large purchases.

2. ProductLine Popularity



- Classic cars are sold higher. The sales of classic cars are much higher than the others.

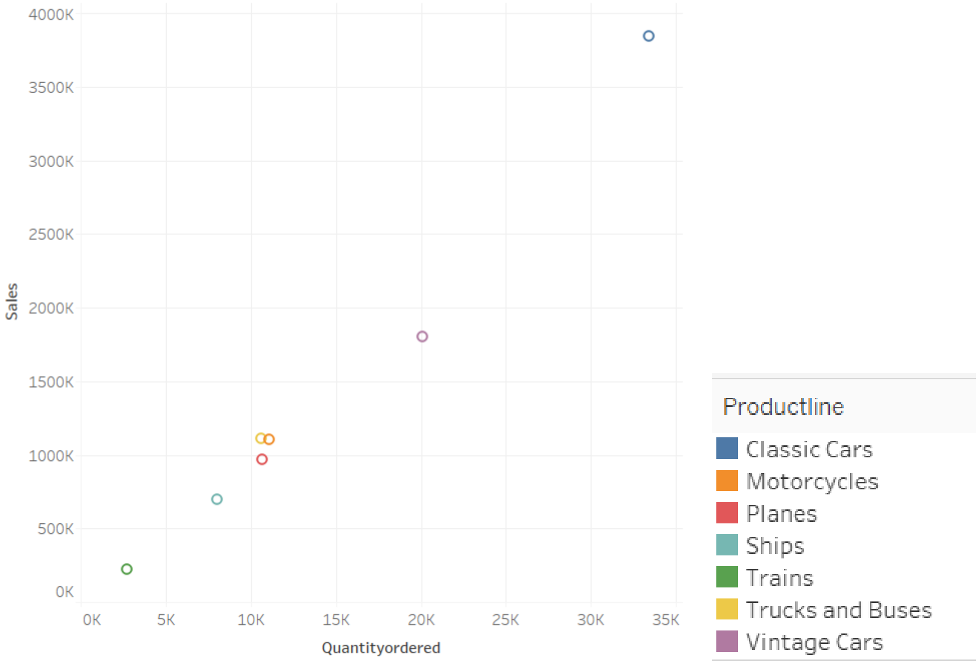
3. Deal Size Count



- Medium sized deals are most common.

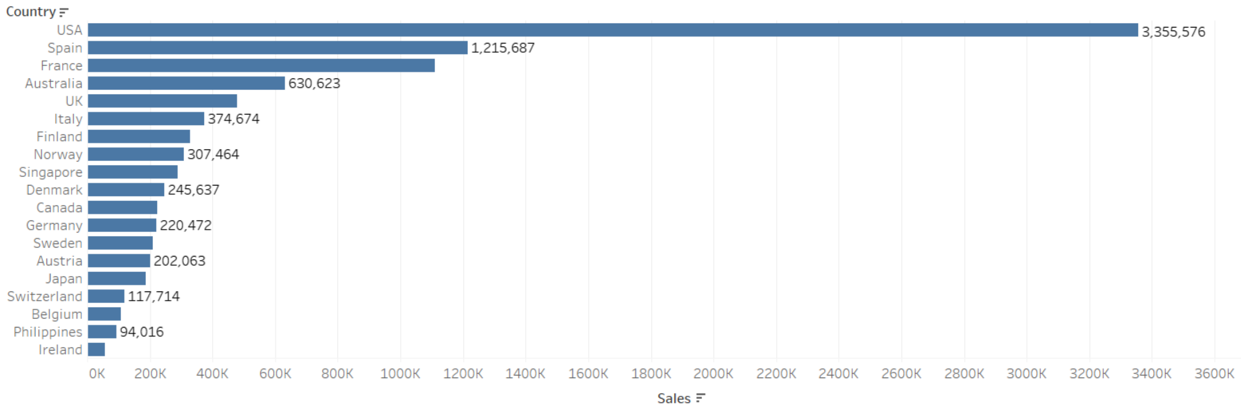
Bivariate Analysis (Two-Variable Relationships)

1. Sales vs Quantity Ordered



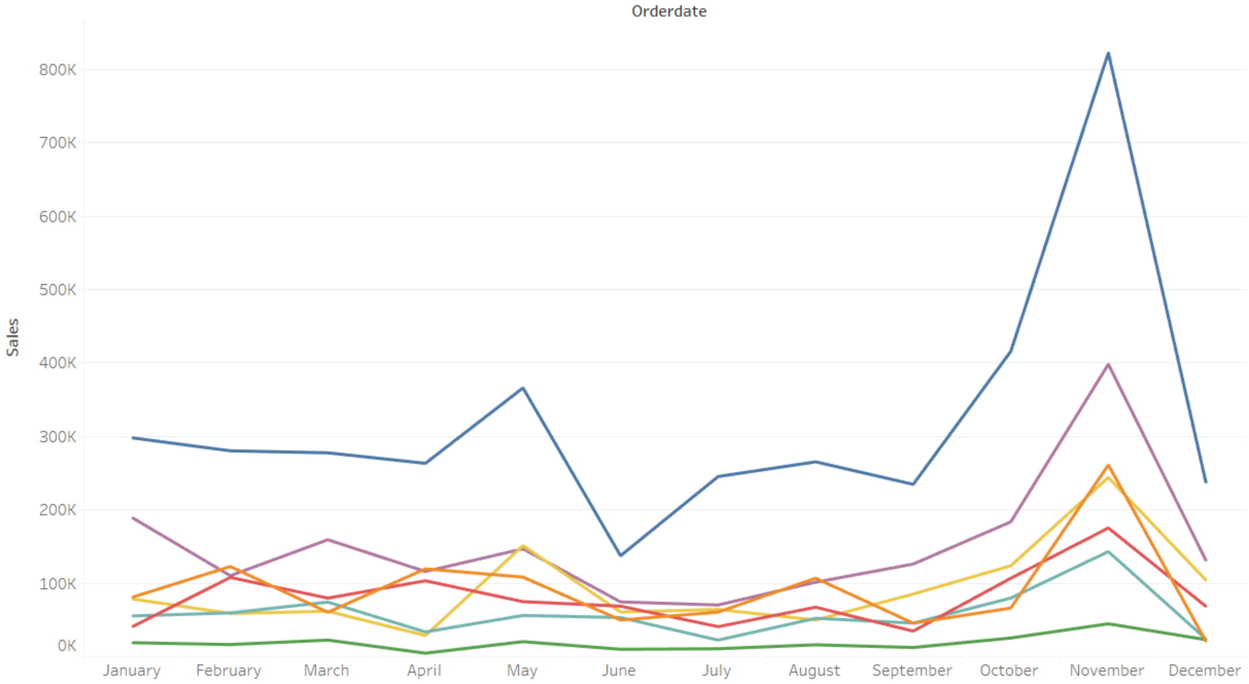
- More sales and quantities are ordered for Classic cars in the automobile sector.

2. Sales by Country



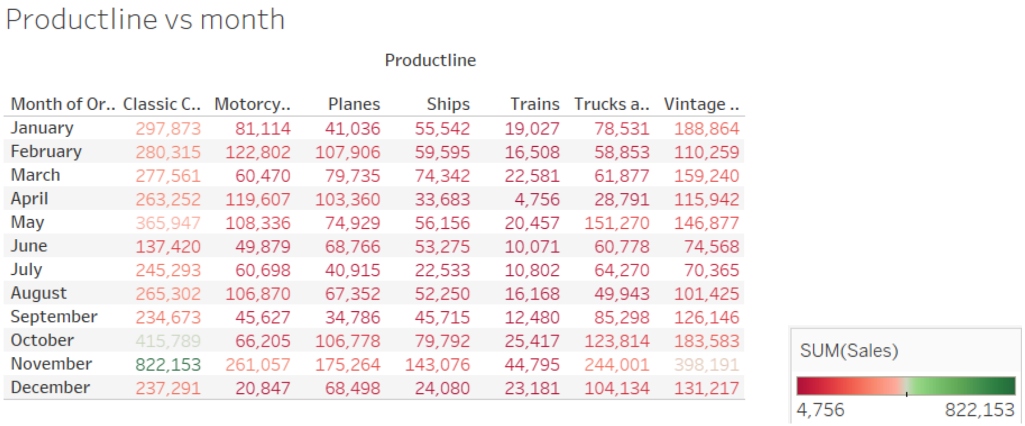
- USA generates the highest revenue among all.

4. Monthly Sales Trend



- The seasonal spikes are mostly at the end of the year. The last 4 months have the highest sales for all the automobiles.

5. Productline vs Month : Heatmap



Productline vs month

Productline

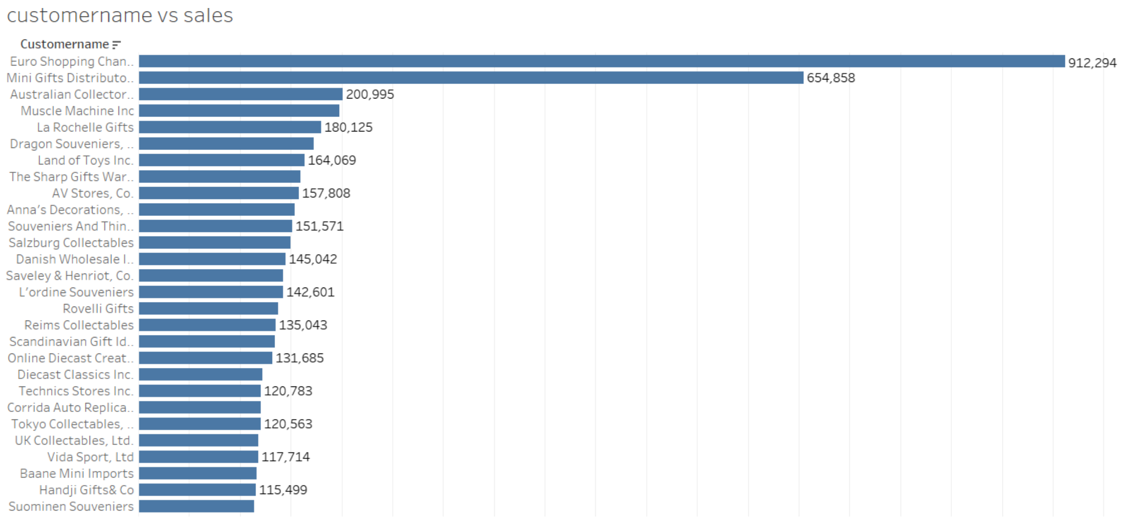
Month of Or..	Classic C..	Motorcy..	Planes	Ships	Trains	Trucks a..	Vintage ..
January	297,873	81,114	41,036	55,542	19,027	78,531	188,864
February	280,315	122,802	107,906	59,595	16,508	58,853	110,259
March	277,561	60,470	79,735	74,342	22,581	61,877	159,240
April	263,252	119,607	103,360	33,683	4,756	28,791	115,942
May	365,947	108,336	74,929	56,156	20,457	151,270	146,877
June	137,420	49,879	68,766	53,275	10,071	60,778	74,568
July	245,293	60,698	40,915	22,533	10,802	64,270	70,365
August	265,302	106,870	67,352	52,250	16,168	49,943	101,425
September	234,673	45,627	34,786	45,715	12,480	85,298	126,146
October	415,789	66,205	106,778	79,792	25,417	123,814	183,583
November	822,153	261,057	175,264	143,076	44,795	244,001	398,191
December	237,291	20,847	68,498	24,080	23,181	104,134	131,217

SUM(Sales)

4,756 822,153

- The Classic cars are the most sold in the month of november .

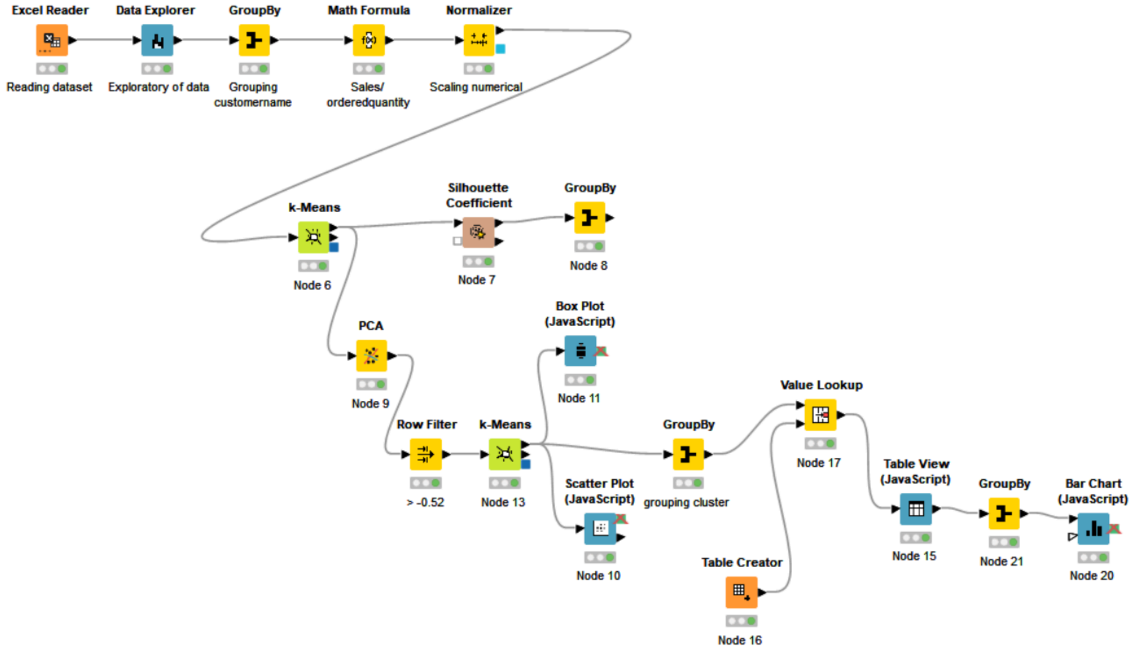
6. Customername vs Sales



- The most valuable customers are Euro Shopping Channel with 912,294.

https://public.tableau.com/app/profile/akshay.sreekumar8075/viz/MRA_main-Analysis/Dashboard1

Customer Segmentation using RFM analysis



After performing RFM-based clustering on the transaction data, **4 customer segments** were identified using the K-Means clustering algorithm in KNIME.

The number of clusters was determined using a combination of:

- **Elbow Method**: To identify the point of diminishing returns in the within-cluster sum of squares.

- **Business Interpretability**: Ensuring each cluster represented a distinct, actionable customer group.

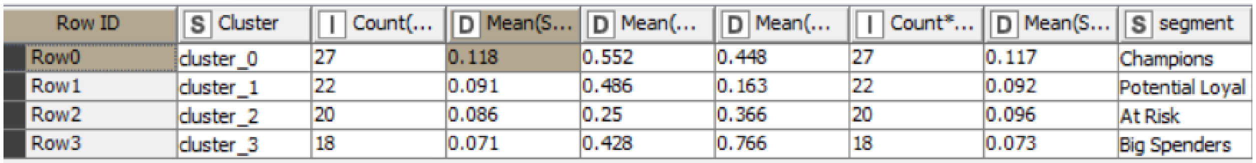
The final 4 segments are:

1. **Champions** – Highly active and valuable customers

2. **Potential Loyal** – Fairly active but with lower monetary value

3. **At Risk** – Declining frequency, need re-engagement

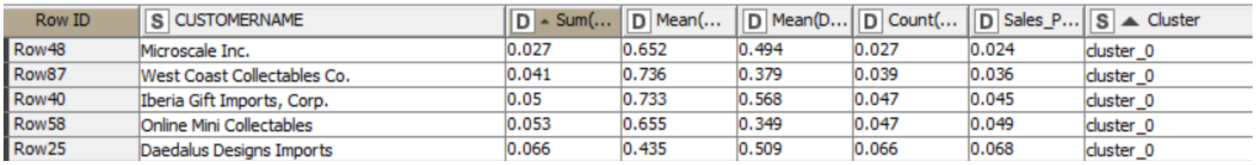
4. **Big Spenders** – High monetary value but recent inactivity



Row ID	S Cluster	I Count(...	D Mean(S...	D Mean(...	D Mean(...	I Count*...	D Mean(S...	S segment
Row0	cluster_0	27	0.118	0.552	0.448	27	0.117	Champions
Row1	cluster_1	22	0.091	0.486	0.163	22	0.092	Potential Loyal
Row2	cluster_2	20	0.086	0.25	0.366	20	0.096	At Risk
Row3	cluster_3	18	0.071	0.428	0.766	18	0.073	Big Spenders

Q) Mention the top five customers

=>



Row ID	S CUSTOMERNAME	D ▲ Sum(...	D Mean(...	D Mean(D...	D Count(...	D Sales_P...	S ▲ Cluster
Row48	Microscale Inc.	0.027	0.652	0.494	0.027	0.024	cluster_0
Row87	West Coast Collectables Co.	0.041	0.736	0.379	0.039	0.036	cluster_0
Row40	Iberia Gift Imports, Corp.	0.05	0.733	0.568	0.047	0.045	cluster_0
Row58	Online Mini Collectables	0.053	0.655	0.349	0.047	0.049	cluster_0
Row25	Daedalus Designs Imports	0.066	0.435	0.509	0.066	0.068	cluster_0

Cluster 0 – Champions

- **Recency:** 0.1179 → Very recent buyers (good).
- **Frequency:** 0.5524 → Highest frequency.
- **Monetary:** 0.4477 → High spenders.
- **Sales per item:** 0.1168 → Consistently buying at a good value.

  Best Overall Customers – Active, frequent, and high-value.

Cluster 1 – Potential Loyal

- **Recency:** 0.0914 → Recent buyers.
- **Frequency:** 0.4863 → Moderate to high frequency.
- **Monetary:** 0.1626 → Low spenders.

Engaged but Low-Value – May be regular small buyers. These can be up-sold or nurtured into Champions through bundled discounts or targeted campaigns.

Cluster 2 – At Risk

- **Recency:** 0.0863 → Recent enough (surprising for At Risk).
- **Frequency:** 0.2499 → Low frequency.
- **Monetary:** 0.3664 → Medium spenders.

Infrequent Buyers – Spend a decent amount but rarely buy. A drop in frequency could lead to churn. They are ideal for win-back offers or limited-time promotions.

Cluster 3 – Big Spenders

- **Recency:** 0.0706 → Least recent (danger zone).
- **Frequency:** 0.4284 → Decent frequency.
- **Monetary:** 0.7662 → Highest spending cluster.

High Value, Low Engagement – These customers are very profitable but not recent. Immediate follow-up needed. They should be targeted with personal outreach, VIP programs, or renewal deals.

Problem Statement - MRA Main Project

Part - B

Business Context:

In the highly competitive grocery retail industry, understanding customer buying patterns is crucial for enhancing sales, increasing customer satisfaction, and improving profitability. By identifying frequently purchased item combinations, grocery stores can craft effective marketing strategies, optimize inventory management, and tailor promotions to meet customer needs. Leveraging Point of Sale (POS) data can unlock valuable insights that drive customer-centric offerings, such as combo packs, discounts, and targeted promotions, which can increase basket size and improve customer retention. This analysis aligns with business goals by maximizing revenue, reducing operational costs, and boosting customer loyalty.

Objective:

The goal is to analyze the POS transactional data to identify frequently purchased item combinations. Using association rule mining or similar techniques, the aim is to uncover patterns that will help the store create targeted combo offers and discounts, ultimately driving revenue growth by increasing customer purchases and average basket size.

Data Description:

Date: The date when the transaction took place.

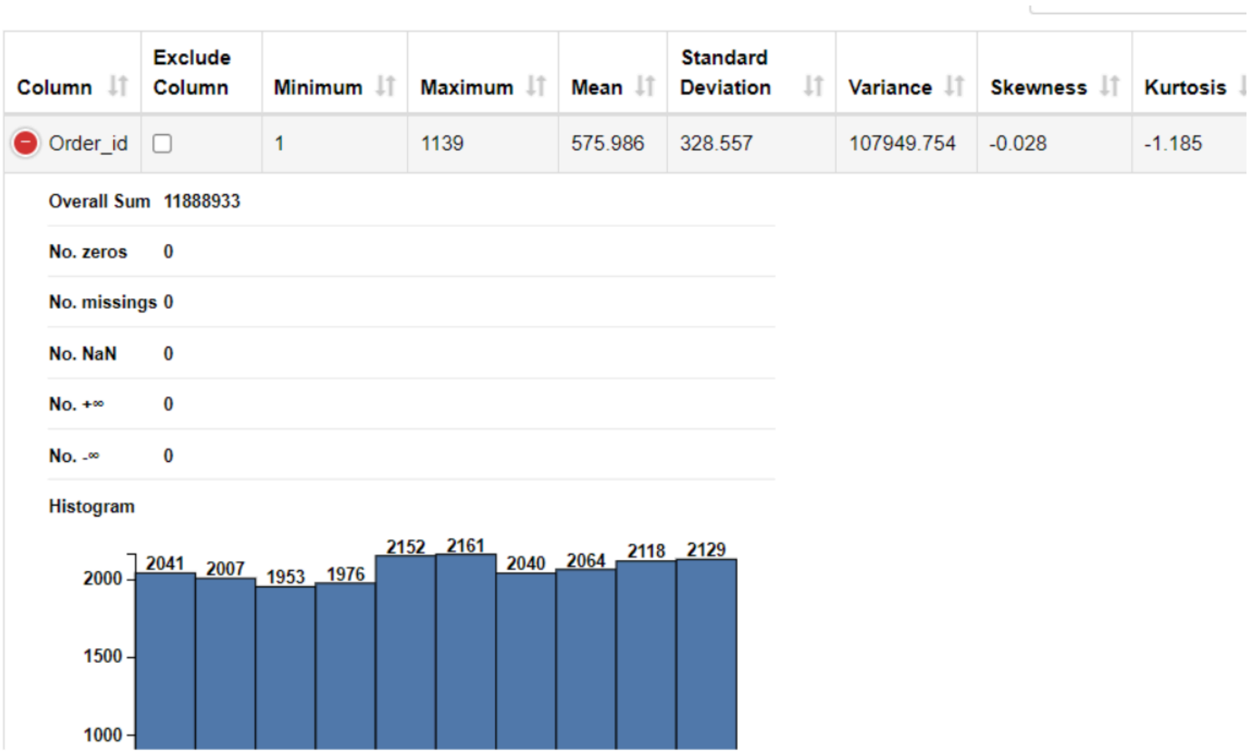
Order_id: A unique identifier for each customer order.

Product: The individual item purchased in the transaction.

Exploratory Data Analysis (EDA)

Summary Statistics

- The dataset includes the following columns Date, Order_ID,Product.
- It contains 1 numerical and 2 nominal values.
- There are a total 20641 rows.
- Dates ranging from 2018 - 2020
- There are 640 unique values.
- There are no missing values

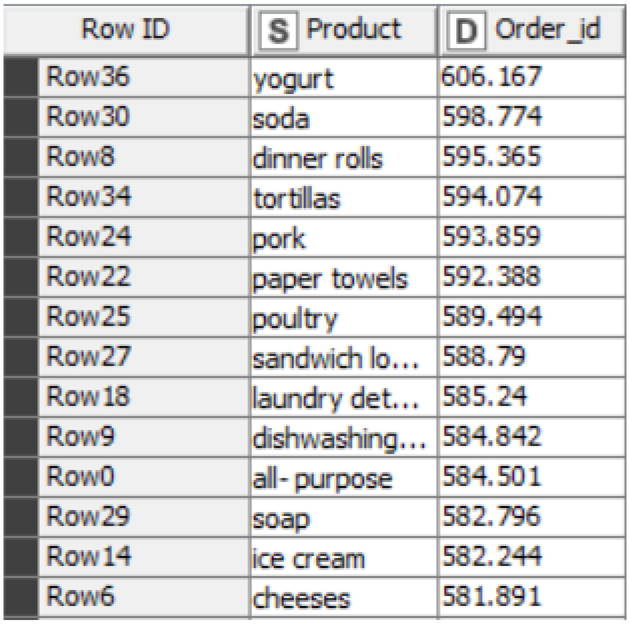


Column	Exclude Column	Minimum	Maximum	Mean	Standard Deviation	Variance	Skewness	Kurtosis
Order_id	□	1	1139	575.986	328.557	107949.754	-0.028	-1.185

Overall Sum	11888933
No. zeros	0
No. missings	0
No. NaN	0
No. +∞	0
No. -∞	0

Histogram

Q) Top 10 Frequently Purchased Products

=> 

Row ID	S Product	D Order_id
Row36	yogurt	606.167
Row30	soda	598.774
Row8	dinner rolls	595.365
Row34	tortillas	594.074
Row24	pork	593.859
Row22	paper towels	592.388
Row25	poultry	589.494
Row27	sandwich lo...	588.79
Row18	laundry det...	585.24
Row9	dishwashing...	584.842
Row0	all- purpose	584.501
Row29	soap	582.796
Row14	ice cream	582.244
Row6	cheeses	581.891

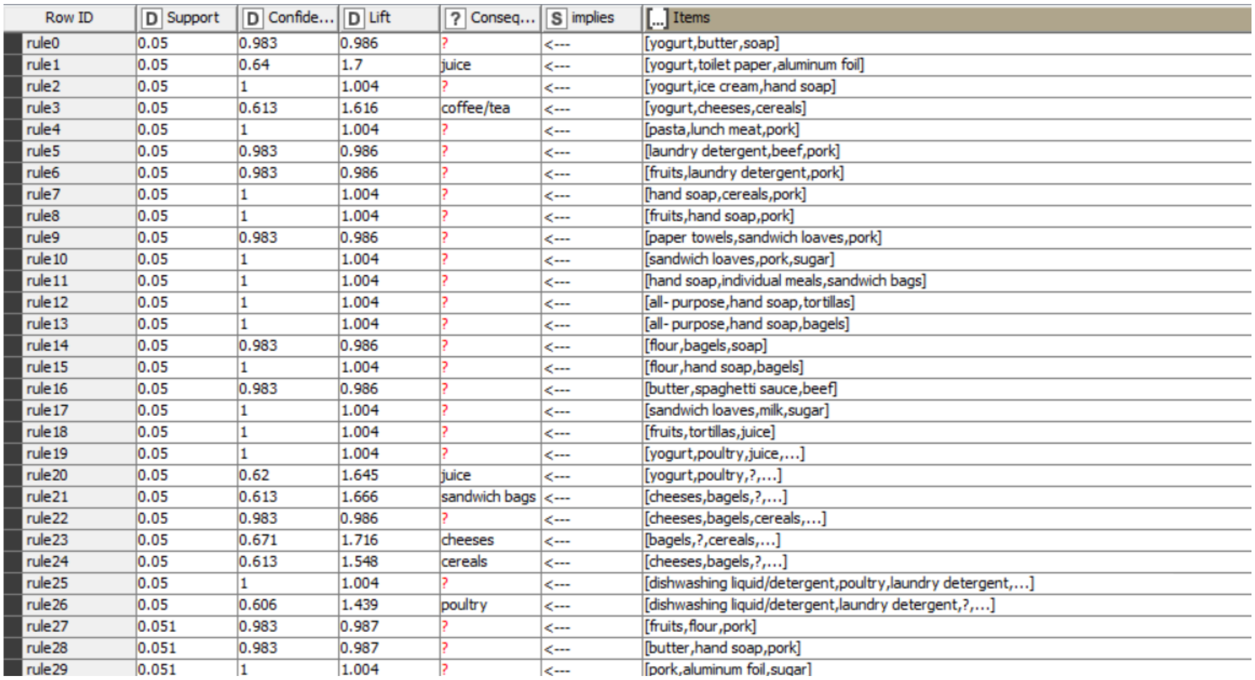
- These are the top 10 items purchased.

Thresholds Used

- Minimum Support: `0.05` (5% of transactions)
- Minimum Confidence: `0.60` (60% confidence required)

Explanation of Metrics

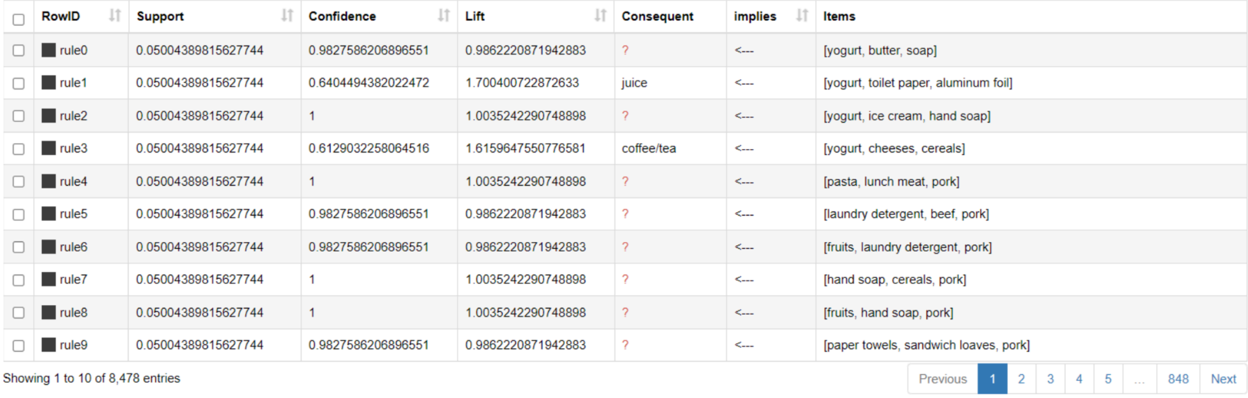
- Support: Fraction of transactions containing both Items and consequent.
- Confidence: How often the consequent appears when the Items are present.
- Lift: How much more likely the consequent is with the items compared to random chance.



Row ID	D Support	D Confide...	D Lift	? Conseq...	S implies	[...] Items
rule0	0.05	0.983	0.986	?	<---	[yogurt,butter,soap]
rule1	0.05	0.64	1.7	juice	<---	[yogurt,toilet paper,aluminum foil]
rule2	0.05	1	1.004	?	<---	[yogurt,ice cream,hand soap]
rule3	0.05	0.613	1.616	coffee/tea	<---	[yogurt,cheeses,cereals]
rule4	0.05	1	1.004	?	<---	[pasta,lunch meat,pork]
rule5	0.05	0.983	0.986	?	<---	[laundry detergent,beef,pork]
rule6	0.05	0.983	0.986	?	<---	[fruits,laundry detergent,pork]
rule7	0.05	1	1.004	?	<---	[hand soap,cereals,pork]
rule8	0.05	1	1.004	?	<---	[fruits,hand soap,pork]
rule9	0.05	0.983	0.986	?	<---	[paper towels,sandwich loaves,pork]
rule10	0.05	1	1.004	?	<---	[sandwich loaves,pork,sugar]
rule11	0.05	1	1.004	?	<---	[hand soap,individual meals,sandwich bags]
rule12	0.05	1	1.004	?	<---	[all- purpose,hand soap,tortillas]
rule13	0.05	1	1.004	?	<---	[all- purpose,hand soap,bagels]
rule14	0.05	0.983	0.986	?	<---	[flour,bagels,soap]
rule15	0.05	1	1.004	?	<---	[flour,hand soap,bagels]
rule16	0.05	0.983	0.986	?	<---	[butter,spaghetti sauce,beef]
rule17	0.05	1	1.004	?	<---	[sandwich loaves,milk,sugar]
rule18	0.05	1	1.004	?	<---	[fruits,tortillas,juice]
rule19	0.05	1	1.004	?	<---	[yogurt,poultry,juice,...]
rule20	0.05	0.62	1.645	juice	<---	[yogurt,poultry,?,...]
rule21	0.05	0.613	1.666	sandwich bags	<---	[cheeses,bagels,?,...]
rule22	0.05	0.983	0.986	?	<---	[cheeses,bagels,cereals,...]
rule23	0.05	0.671	1.716	cheeses	<---	[bagels,?,cereals,...]
rule24	0.05	0.613	1.548	cereals	<---	[cheeses,bagels,?,...]
rule25	0.05	1	1.004	?	<---	[dishwashing liquid/detergent,poultry,laundry detergent,...]
rule26	0.05	0.606	1.439	poultry	<---	[dishwashing liquid/detergent,laundry detergent,?,...]
rule27	0.051	0.983	0.987	?	<---	[fruits,flour,pork]
rule28	0.051	0.983	0.987	?	<---	[butter,hand soap,pork]
rule29	0.051	1	1.004	?	<---	[pork,aluminum foil,sugar]

- If items like yogurt,toilet paper, aluminium foil they recommend for juice.
- Also for yogurt, cheese, cereals are taken they suggest coffee/tea.

The final output table



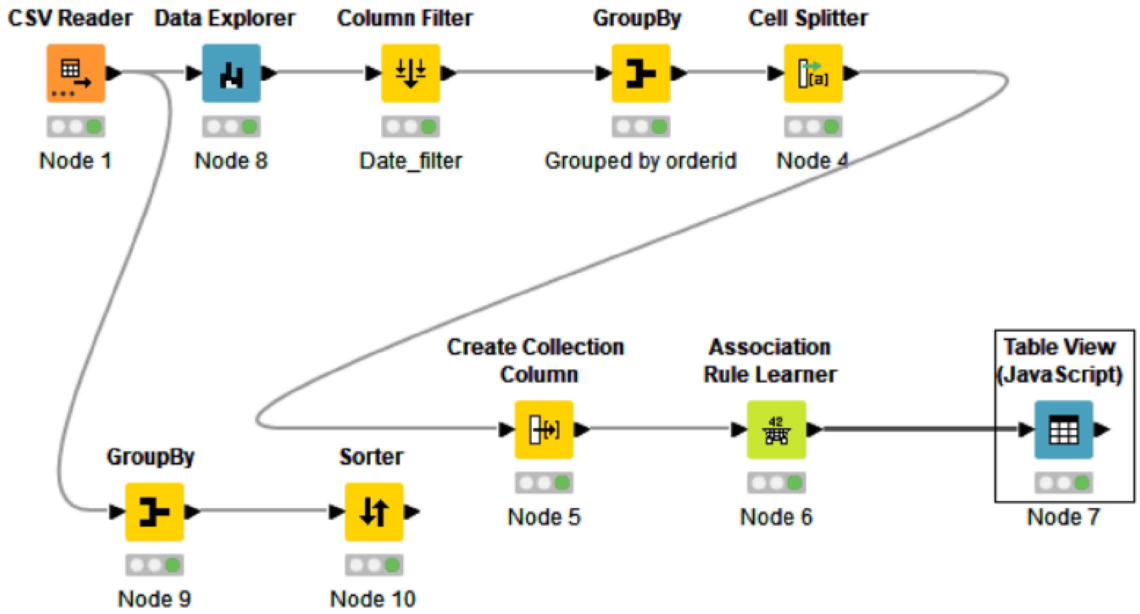
□	RowID	Support	Confidence	Lift	Consequent	implies	Items
□	■ rule0	0.05004389815627744	0.9827586206896551	0.9862220871942883	?	<---	[yogurt, butter, soap]
□	■ rule1	0.05004389815627744	0.6404494382022472	1.700400722872633	juice	<---	[yogurt, toilet paper, aluminum foil]
□	■ rule2	0.05004389815627744	1	1.0035242290748898	?	<---	[yogurt, ice cream, hand soap]
□	■ rule3	0.05004389815627744	0.6129032258064516	1.6159647550776581	coffee/tea	<---	[yogurt, cheeses, cereals]
□	■ rule4	0.05004389815627744	1	1.0035242290748898	?	<---	[pasta, lunch meat, pork]
□	■ rule5	0.05004389815627744	0.9827586206896551	0.9862220871942883	?	<---	[laundry detergent, beef, pork]
□	■ rule6	0.05004389815627744	0.9827586206896551	0.9862220871942883	?	<---	[fruits, laundry detergent, pork]
□	■ rule7	0.05004389815627744	1	1.0035242290748898	?	<---	[hand soap, cereals, pork]
□	■ rule8	0.05004389815627744	1	1.0035242290748898	?	<---	[fruits, hand soap, pork]
□	■ rule9	0.05004389815627744	0.9827586206896551	0.9862220871942883	?	<---	[paper towels, sandwich loaves, pork]

Showing 1 to 10 of 8,478 entries

- Products like pork, yogurt, soap, and cereals appear in multiple strong rules.

- Certain combinations e.g., `[cheeses, cereals, yogurt] => coffee/tea` have high confidence and lift.

- Some rules have perfect confidence (1.00), indicating strong cross-sell potential.


CSV Reader
Data Explorer
Column Filter
GroupBy
Cell Splitter
Node 1
Node 8
Date_filter
Grouped by orderid
Node 4
Create Collection Column
Association Rule Learner
Table View (JavaScript)
GroupBy
Sorter
Node 5
Node 6
Node 7
Node 9
Node 10