greatlearning
Learning for Life

Business report

Sales dataset

DATA SCIENCE AND BUSINESS ANALYTICS

Content:

1. Exploratory Data Analysis

1.1. Problem definition:

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.

Objectives:

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

1. Identify underlying patterns in customer purchasing behaviour.
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.

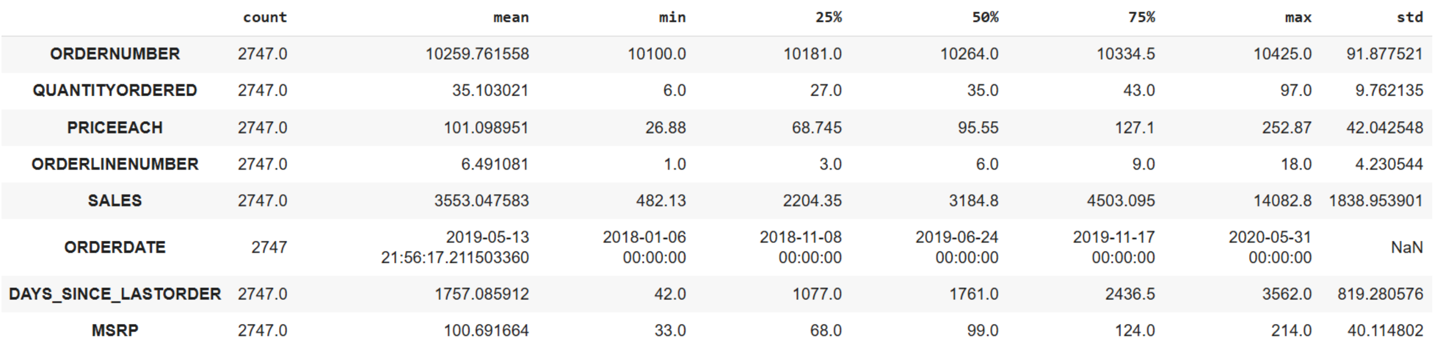
Data Description:

The dataset provided contains three years of transactional data from the company, with each row representing a unique order.

Below is an explanation of the key attributes:

- **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).

1.2. Statistical Summary:

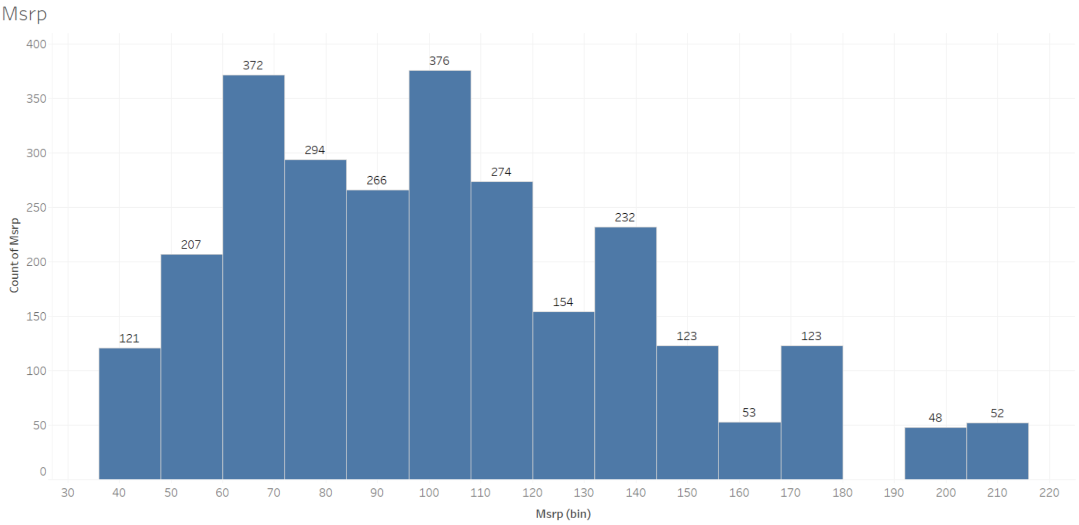


	count	mean	min	25%	50%	75%	max	std
ORDERNUMBER	2747.0	10259.761558	10100.0	10181.0	10264.0	10334.5	10425.0	91.877521
QUANTITYORDERED	2747.0	35.103021	6.0	27.0	35.0	43.0	97.0	9.762135
PRICEEACH	2747.0	101.098951	26.88	68.745	95.55	127.1	252.87	42.042548
ORDERLINENUMBER	2747.0	6.491081	1.0	3.0	6.0	9.0	18.0	4.230544
SALES	2747.0	3553.047583	482.13	2204.35	3184.8	4503.095	14082.8	1838.953901
ORDERDATE	2747	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	1757.085912	42.0	1077.0	1761.0	2436.5	3562.0	819.280576
MSRP	2747.0	100.691664	33.0	68.0	99.0	124.0	214.0	40.114802

- **Order number:** This is the order number when customers which starts with around 10100 and ends with around 10425.
- **Quantity ordered:** The number of quantity ordered by the customers which might start from 6 and ends with 97 with average quantity of 35.
- **Price each:** This column shows the price per product which might starts from 26 to the end of 253 with average price of 101.
- **Order line number:** It is the sequence number of product with min of, max of 18 and average of 6
- **Sales:** Data has min sales of 482 and max of 14083 with average of 3553.
- **Order date:** min order date is 2018-01-06 and max is 2020-05-31 with average of the 2019-05-13
- **Days since last order:** number of days since the customer's previous order, min days are 42 and the max are 3562 with average of 1757.
- **MSRP:** This column shows manufacturer’s suggested retail price for the product with min of 100, max of 214 and average of 100.

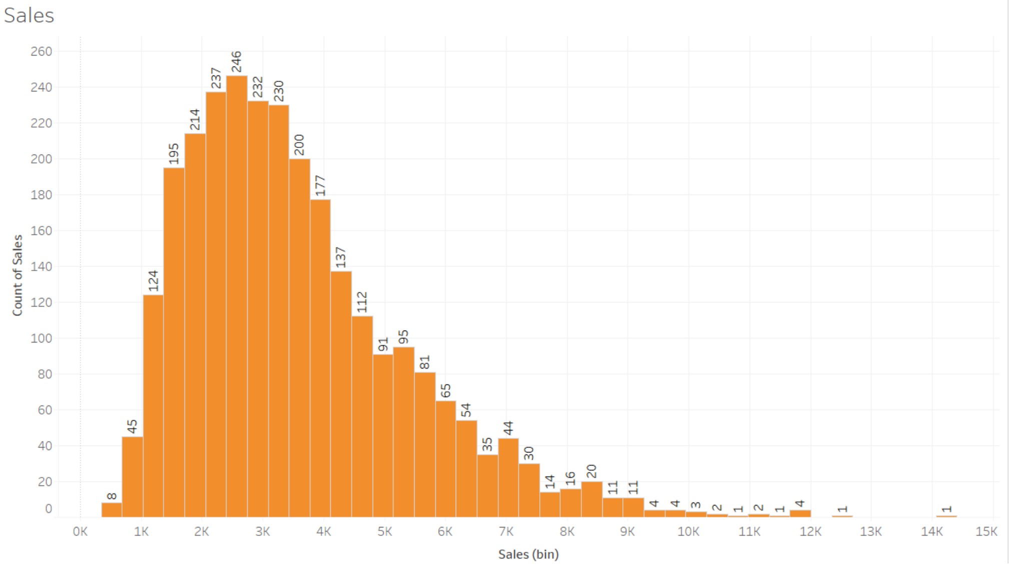
1.3. Univariate analysis.

▪ Msrp:



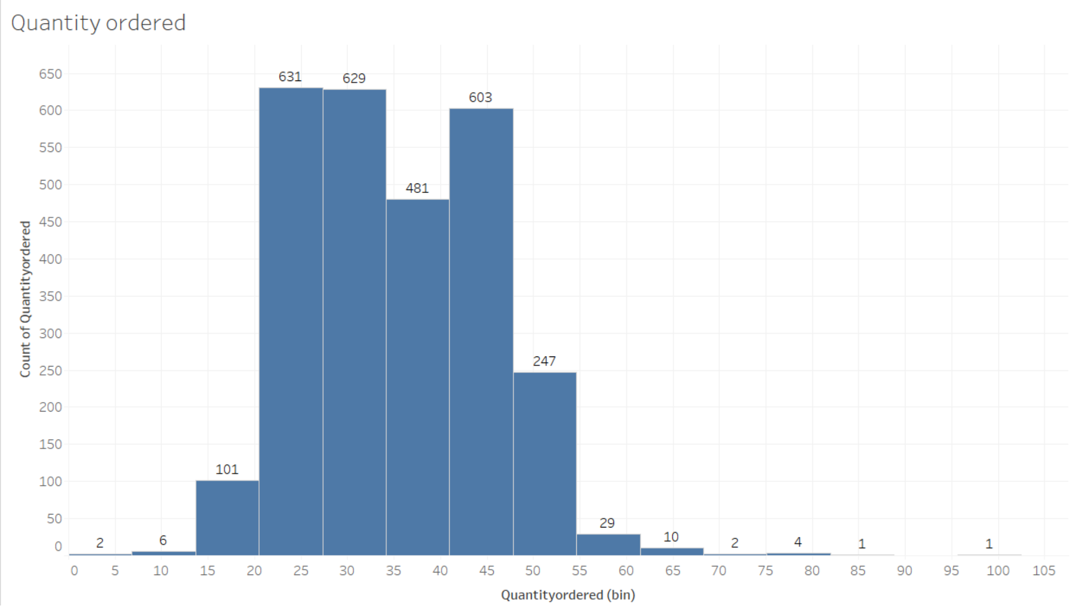
- The 60–70 and 100–110 bins have the highest frequency (~370 items), showing these price points are very popular.
- Very few products are priced above 180, indicating that extremely high MSRP products are rare.

▪ Sales:



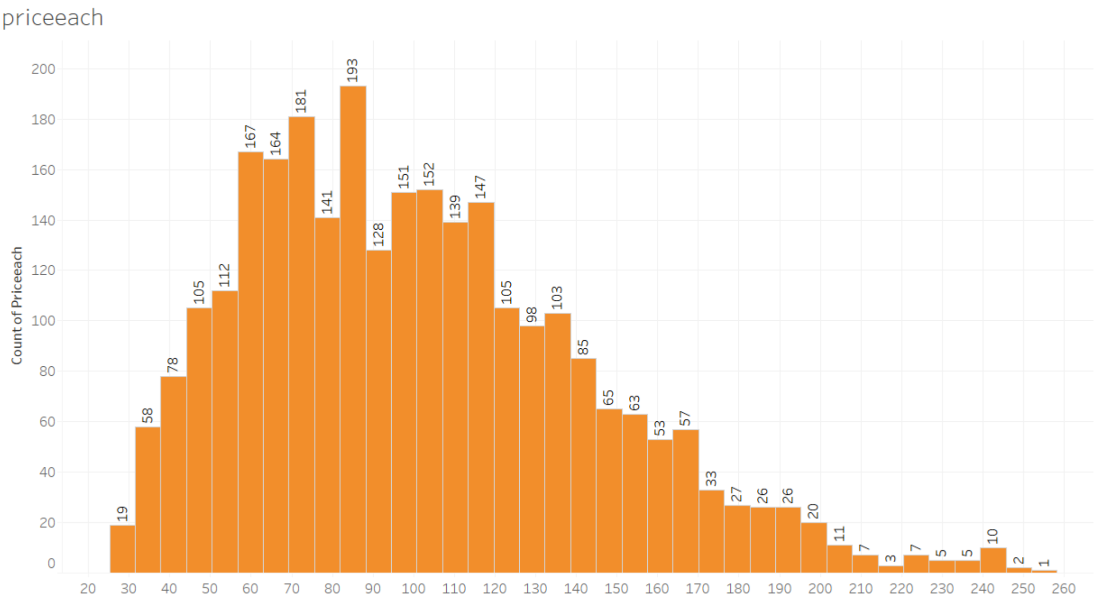
- Highest number of entries in 2500–3000 range.
- Sales start tapering after 4000, and very few go beyond 6500+

▪ Quantity order:



- Distribution Shape is fairly balanced with a dip at the extreme low (15–20) and a minor dip at the high end.
- Popular Quantity Ranges peaks between 22–32 and 42–47 units per order.

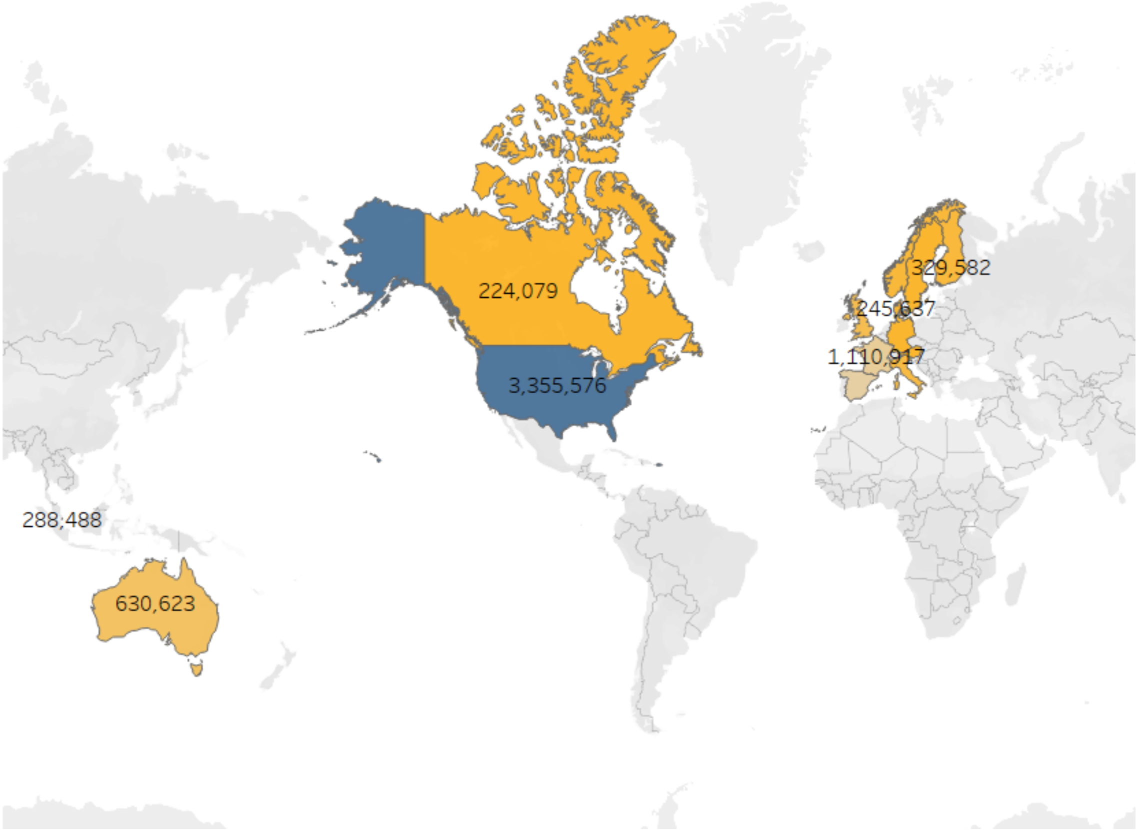
▪ QUANTITY ORDER:



- Distribution Shape is slightly right-skewed.
- Popular Unit Price Ranges between ₹60–₹90 and ₹100–₹120 have consistent counts, peaking around ₹80–₹90.

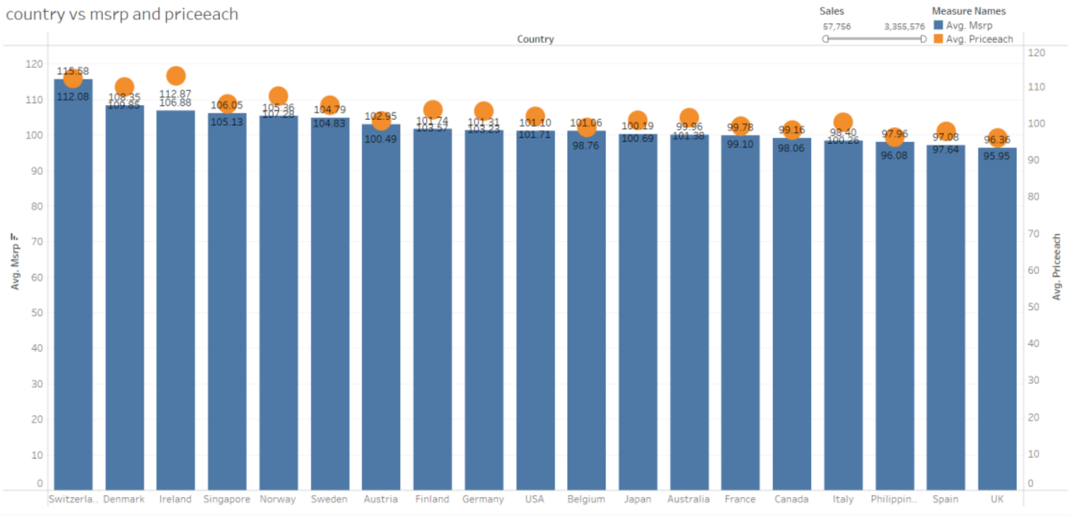
1.4. Bivariate analysis.

- **Country vs sales:**



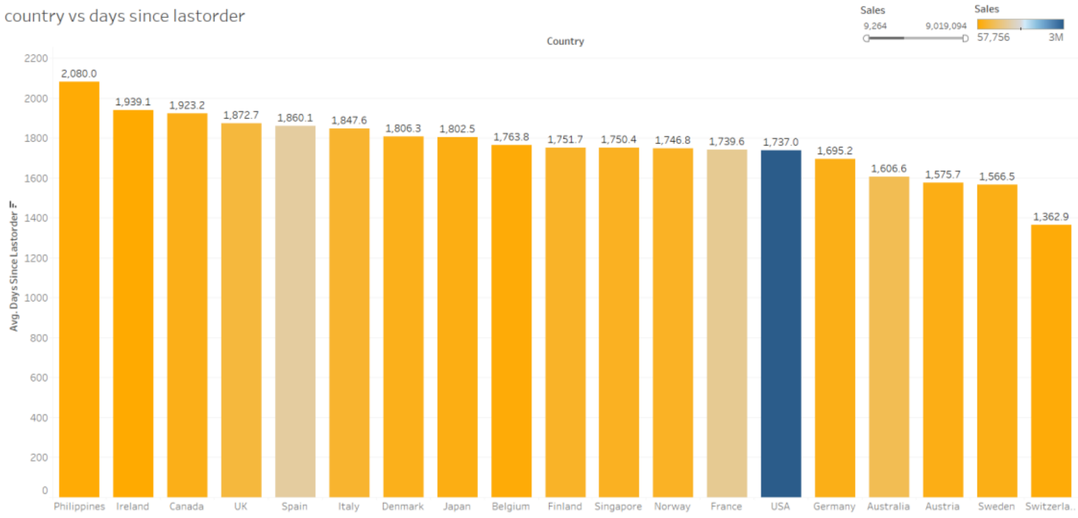
- USA has the highest sales of $3355576.
- Spain has the 2nd highest sales of $1215686.
- It has a sales connection with 19 countries.
- Top 5 countries in sales are USA, Spain, France, Australia and UK.
- Company should expand marketing and partnerships in Canada to mirror US success.

▪ Country vs msrp and priceeach:



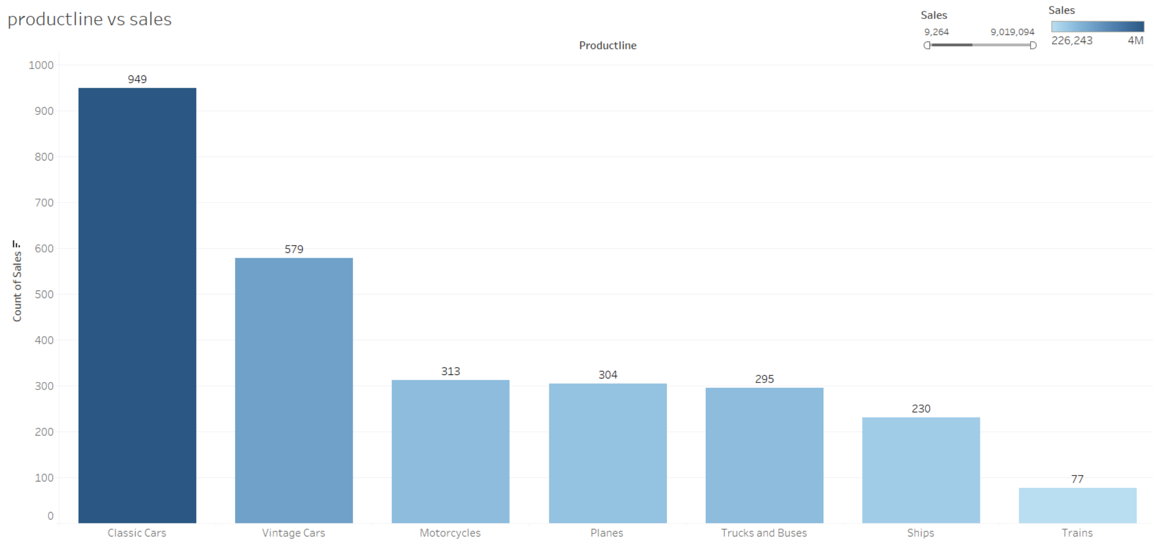
- As we saw that the Switzerland has the highest price per product and highest mrsp across 19 countries.
- At the end UK has the lowest price per product and lowest mrsp .
- Second highest mrsp has in Denmark but second highest price per product has in Ireland.

▪ Country vs days since last order:



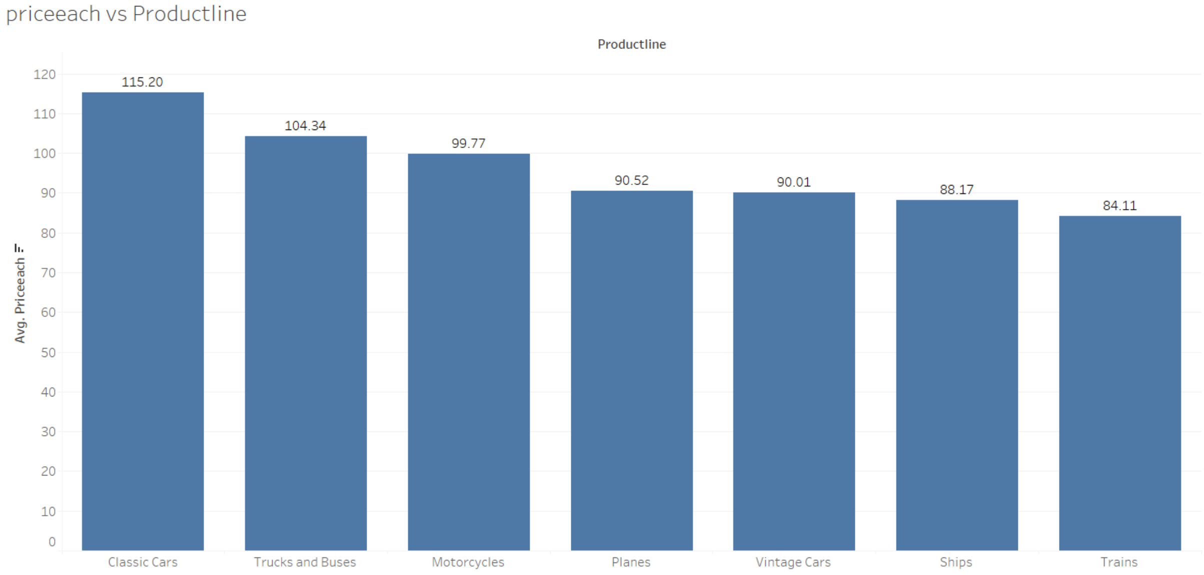
- As per the above plot the Philippines has the highest less occurring sales which has third lowest price per each product.
- Whereas the people in Switzerland purchase often as compared to other countries even though it has the highest price per each product which means company can sell high price products in Switzerland compared to Philippines.

▪ Product line VS Sales:



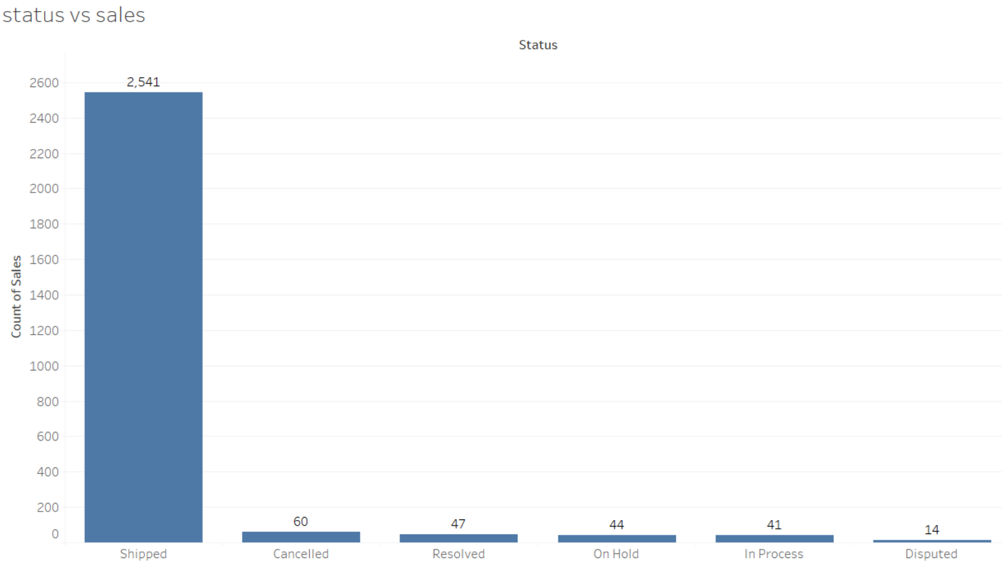
- Purchasing Classic cars much more as compared to other product line,949 times purchases as per the above plot. And the second highest purchasing car is vintage cars, which has frequency of 579 times.
- Trains has less frequency purchases which make sense as well with the rate of 77 times.
- Company should focus on Classic cars and try to improve sales in Vintage car.

▪ Price each VS Product line:



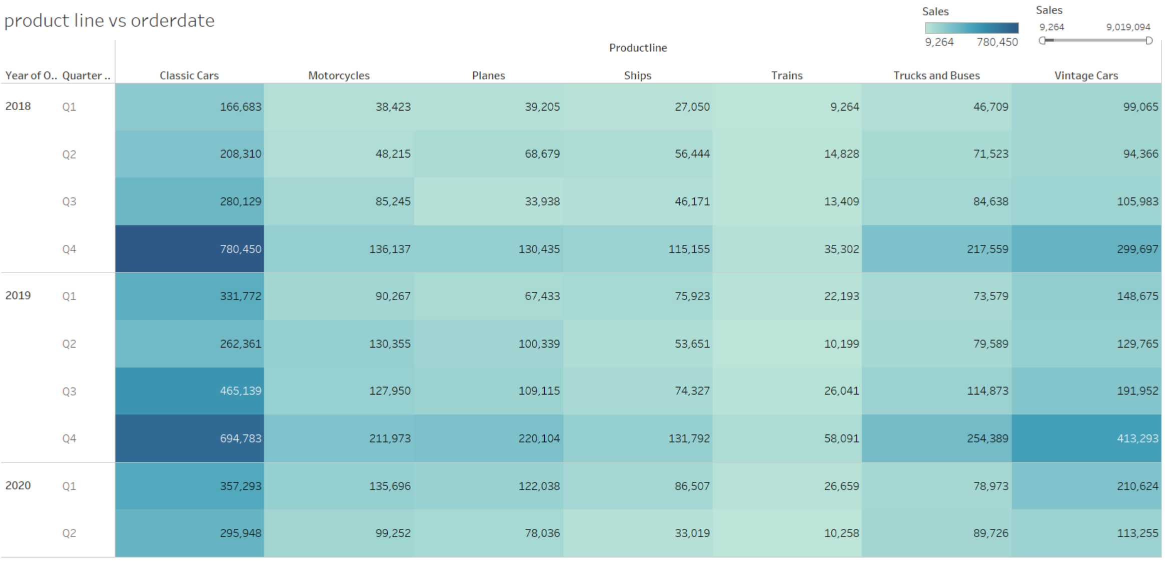
- As said above the classic cars have the highest avg price which is 115 per product. And the trains have lowest price per product.

▪ Status VS Sales:



- Those products which are shipped are more, sales count of 2541.
- Cancellation are 2nd highest so we need to check why the cancellation are higher than resolved and must focus on it to increase sales.

▪ Product line vs order date:

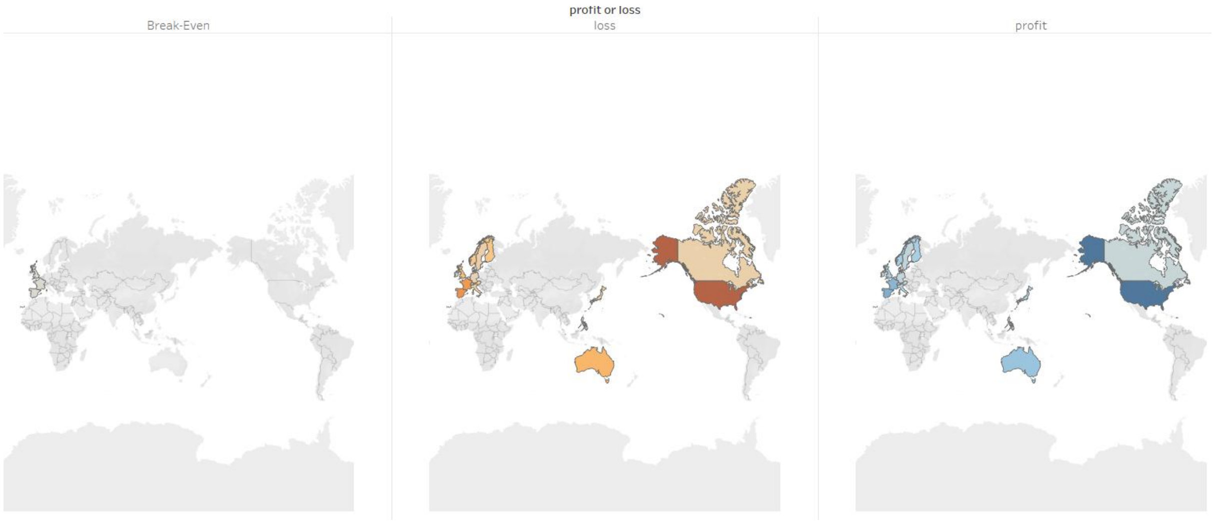

product line vs orderdate

Sales
9,264
780,450
Sales
9,264
9,019,094

		Productline						
Year of O..	Quarter ..	Classic Cars	Motorcycles	Planes	Ships	Trains	Trucks and Buses	Vintage Cars
2018	Q1	166,683	38,423	39,205	27,050	9,264	46,709	99,065
	Q2	208,310	48,215	68,679	56,444	14,828	71,523	94,366
	Q3	280,129	85,245	33,938	46,171	13,409	84,638	105,983
	Q4	780,450	136,137	130,435	115,155	35,302	217,559	299,697
2019	Q1	331,772	90,267	67,433	75,923	22,193	73,579	148,675
	Q2	262,361	130,355	100,339	53,651	10,199	79,589	129,765
	Q3	465,139	127,950	109,115	74,327	26,041	114,873	191,952
	Q4	694,783	211,973	220,104	131,792	58,091	254,389	413,293
2020	Q1	357,293	135,696	122,038	86,507	26,659	78,973	210,624
	Q2	295,948	99,252	78,036	33,019	10,258	89,726	113,255

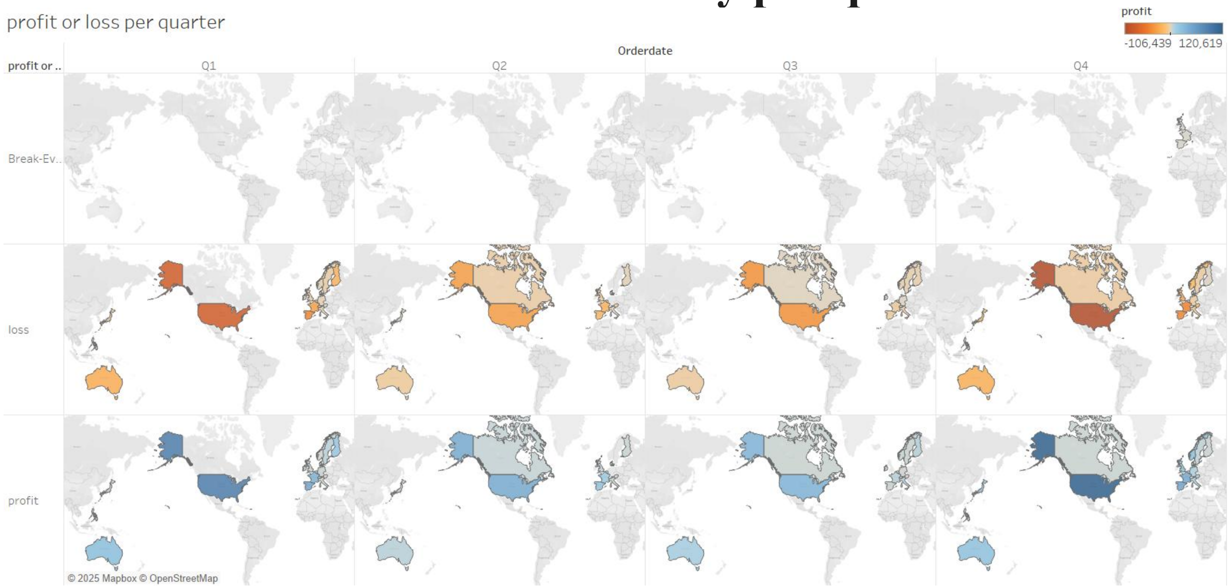
- On 4th quarter has the highest sales. As said above Classic cars are the highest sale, vintage cars are 2nd highest.
- Sales are increasing year by year. But in 2019 quarter 4 has reduced from 780450 to 694783.

Profit or loss in each countries:



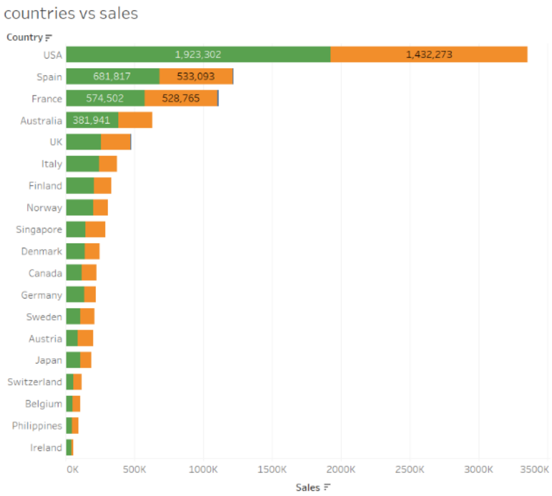
- Many of the countries have both losses and the profits. And some of the countries such as UK, France and Spain has break even sales, no loss no profit.
- USA itself has highest loss and profit as compared to other countries but the profits are more than losses.
- Canada has more loss compared to profits which leads to loss to the company.

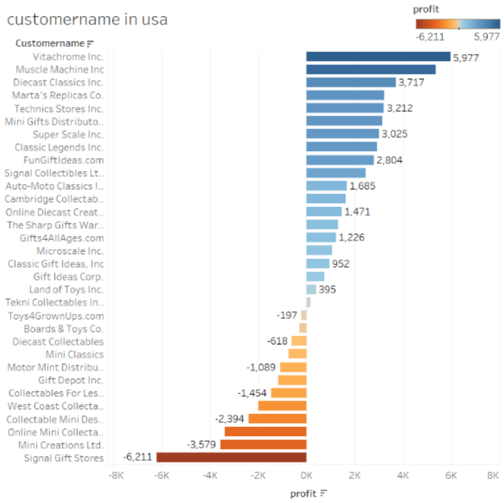
Profit or loss in each country per quarter



- For countries like USA the is huge loss at 1st quarter and decreased in 2nd quarter and then has continuous increase in 3rd and 4th quarter. Same for profits.
- For countries like Canada there is huge loss at 2nd quarter which decreased in 3rd and again increased in 4th quarter and the profits are decreasing quarter by quarter.
- Japan has both profit and loss in 1st, 2nd but no sales in 3rd and has profit and loss in 4th.

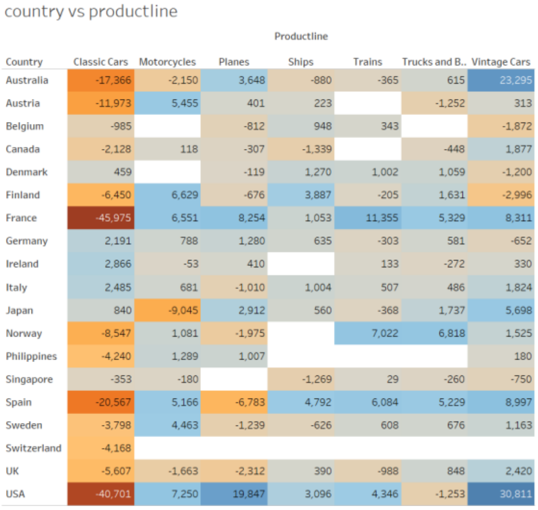
▪ Profits and losses



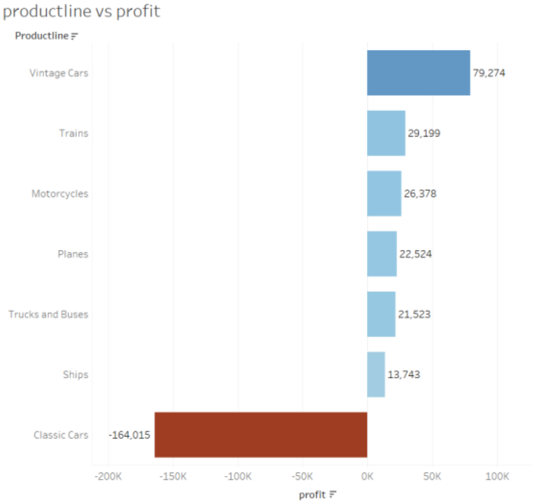


- USA has the highest sales across all whereas told earlier there is both profits and losses.
- Vitachrome inc is giving the highest profits whereas Signal gift stores is giving the highest losses.
- Muscle machine inc is giving the second highest profits to the company.

▪ Productline affecting profit and loss.


country vs productline

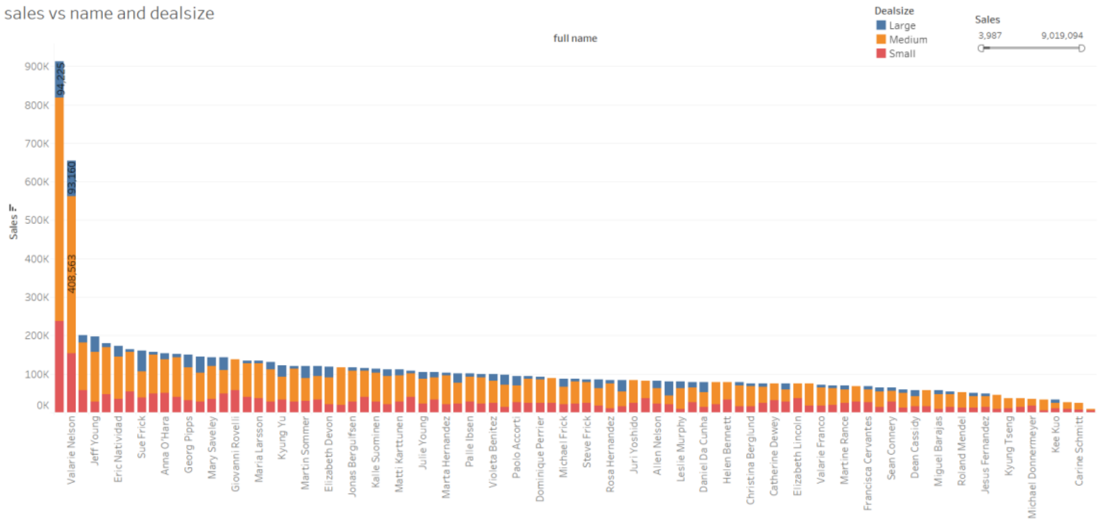
	Productline						
Country	Classic Cars	Motorcycles	Planes	Ships	Trains	Trucks and B..	Vintage Cars
Australia	-17,366	-2,150	3,648	-880	-365	615	23,295
Austria	-11,973	5,455	401	223		-1,252	313
Belgium	-985		-812	948	343		-1,872
Canada	-2,128	118	-307	-1,339		-448	1,877
Denmark	459		-119	1,270	1,002	1,059	-1,200
Finland	-6,450	6,629	-676	3,887	-205	1,631	-2,996
France	-45,975	6,551	8,254	1,053	11,355	5,329	8,311
Germany	2,191	788	1,280	635	-303	581	-652
Ireland	2,866	-53	410		133	-272	330
Italy	2,485	681	-1,010	1,004	507	486	1,824
Japan	840	-9,045	2,912	560	-368	1,737	5,698
Norway	-8,547	1,081	-1,975		7,022	6,818	1,525
Philippines	-4,240	1,289	1,007				180
Singapore	-353	-180		-1,269	29	-260	-750
Spain	-20,567	5,166	-6,783	4,792	6,084	5,229	8,997
Sweden	-3,798	4,463	-1,239	-626	608	676	1,163
Switzerland	-4,168						
UK	-5,607	-1,663	-2,312	390	-988	848	2,420
USA	-40,701	7,250	19,847	3,096	4,346	-1,253	30,811



- Most of the losses are due to classic cars in most of the countries France and USA has the highest loss. Company should stop selling classic cars in such companies to avoid huge losses.
- As said above USA has the highest losses as compared to other countries where classic cars has the most losses and 2nd loss is in trucks and buses.
- Vintage cars has the highest profit which should try to sell the most and should avoid classic cars.

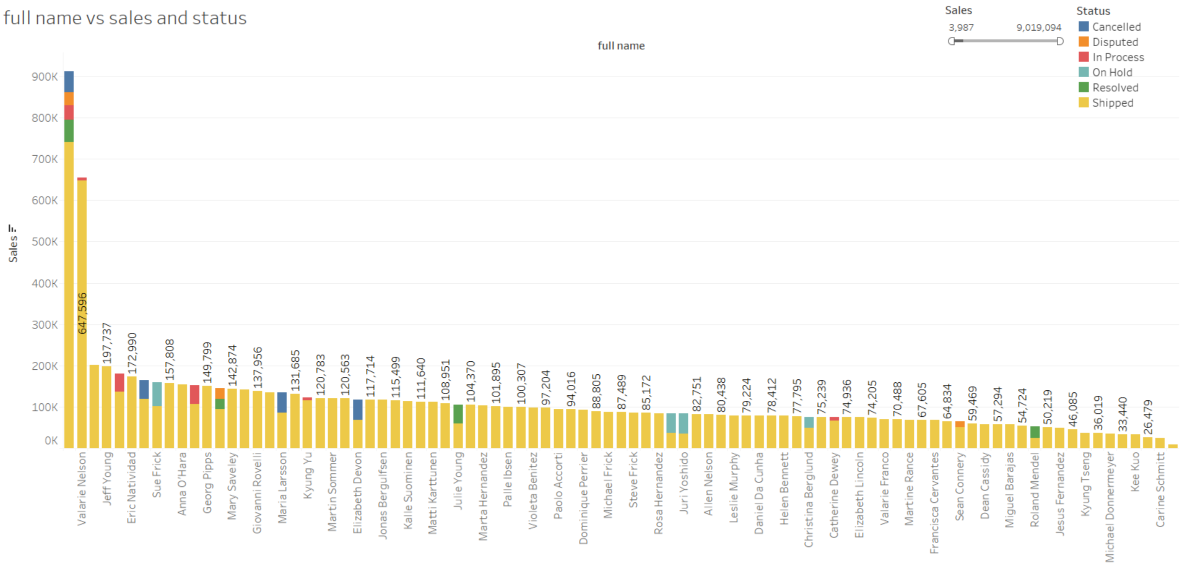
1.5. Multivariate analysis.

▪ Sales VS Name and Deal size:



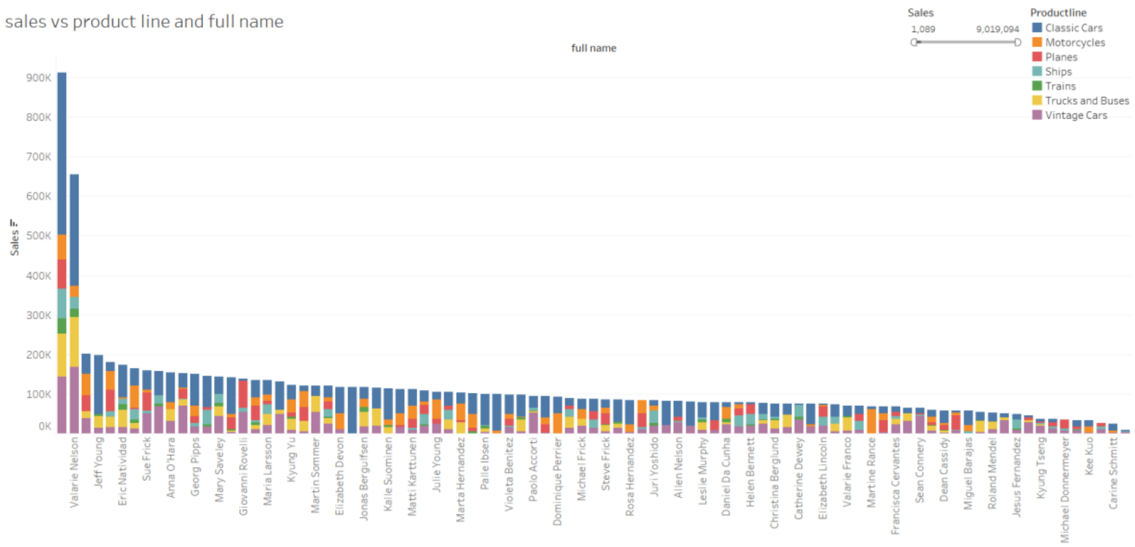
- Deigo Freyre is the top customer who purchase much more as compared to others.
- Valarie Nelson is the 2nd top customer and very less sale customer is Leslie Young.
- Medium deal size is the highest than other deal size among all the customers.

▪ Full name VS Sales and Status:



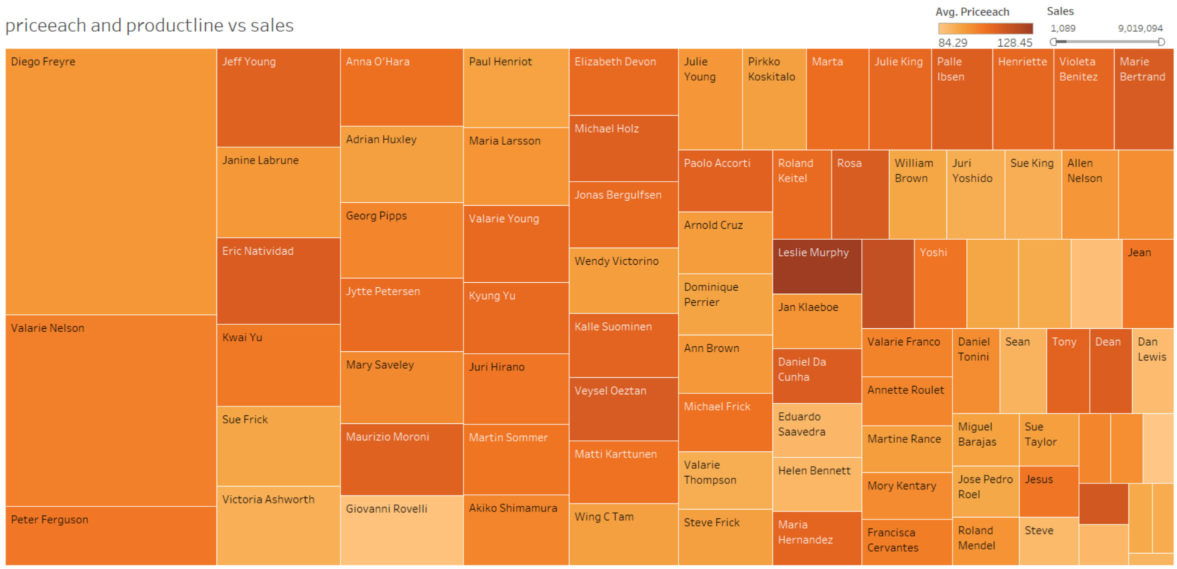
- As said above those are customers with highest and the lowest sales.
- Shipped are more and many of the cancellation are also there should try to reduce the cancellation rate which can automatically increases the sales.

▪ Full name VS Product line and Sales:



- Deigo Freyre is investing on all types of product line sounds like a rich who is interested in buying such things.
- 4 customers are investing in trains if a company wants to increase trains selling should focus on these 4 Deigo Freyre, Valarie Nelson, Eric Natividad and Jonas Bergulfsen.
- Classic cars are highly sold product line.

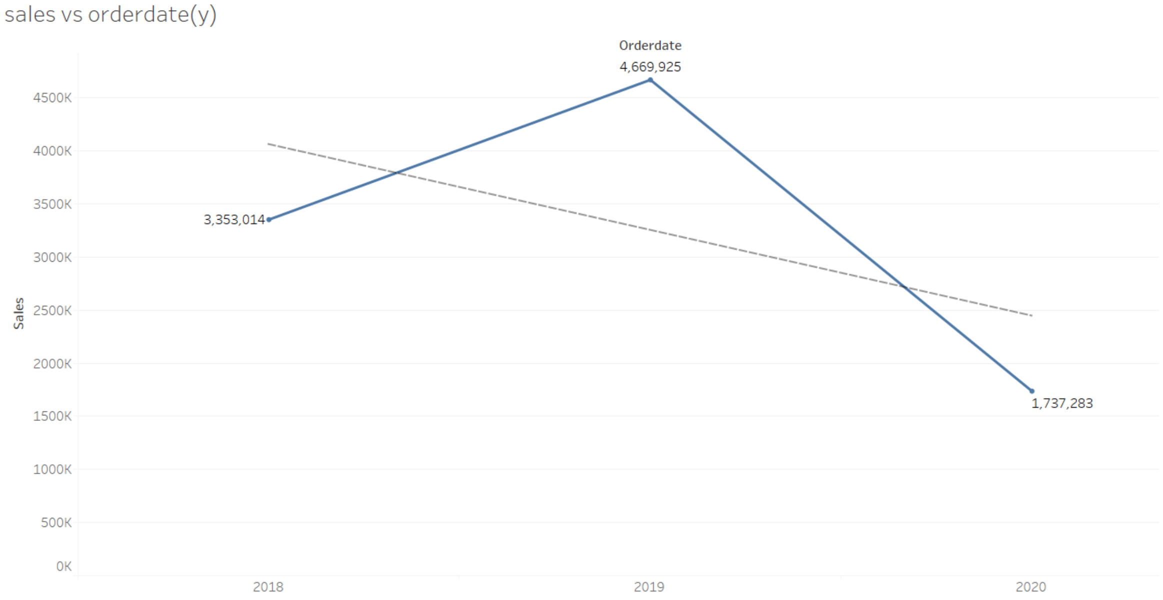
▪ GENRE VS VIEWS TRAILER:



- Leslie Murphy has the highest avg price per each product who has not much sales.
- Which shows that customer purchases high priced products as well as medium priced products, that customer has less purchases of low priced products.

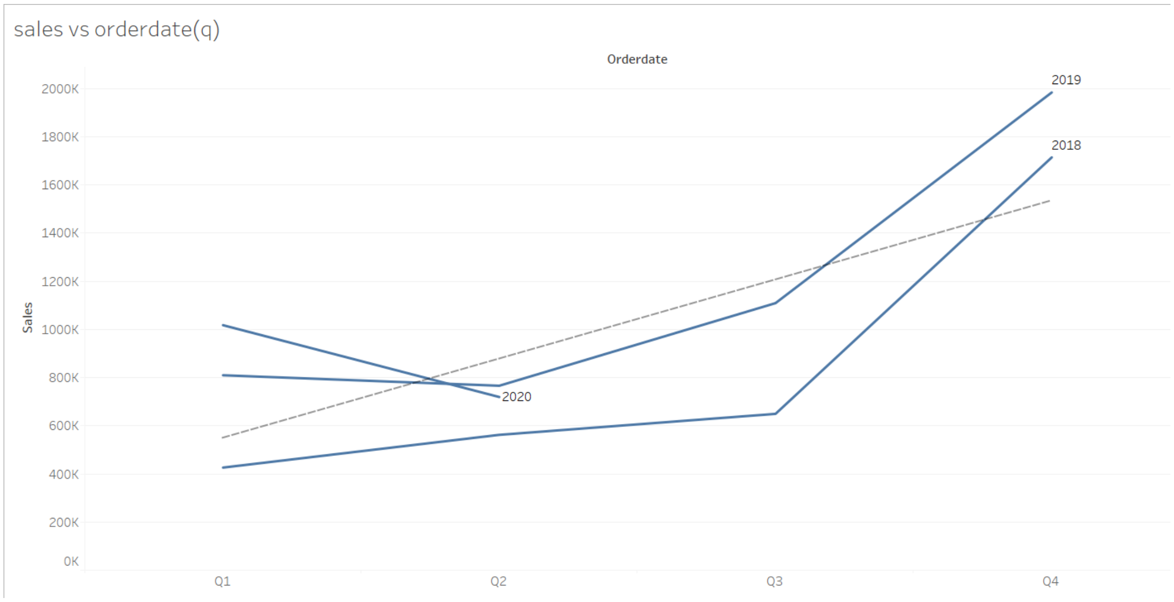
1.6. Time series

▪ Yearly time series



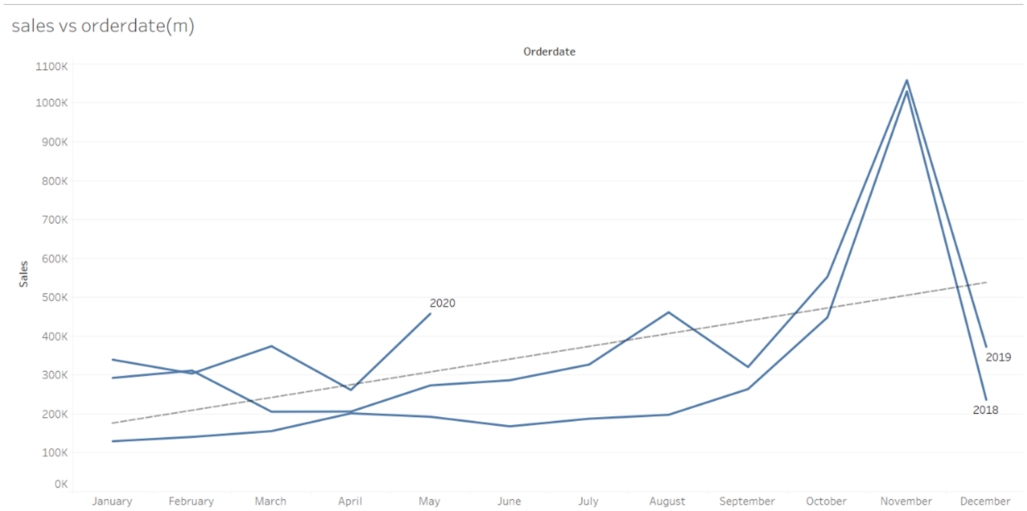
- Sales is increasing year by year. In dataset 2020 has only 2 quarters, that's why its showing less sales but the sales had increased in 2020 as well as increased in 2019.

▪ Quarterly time series



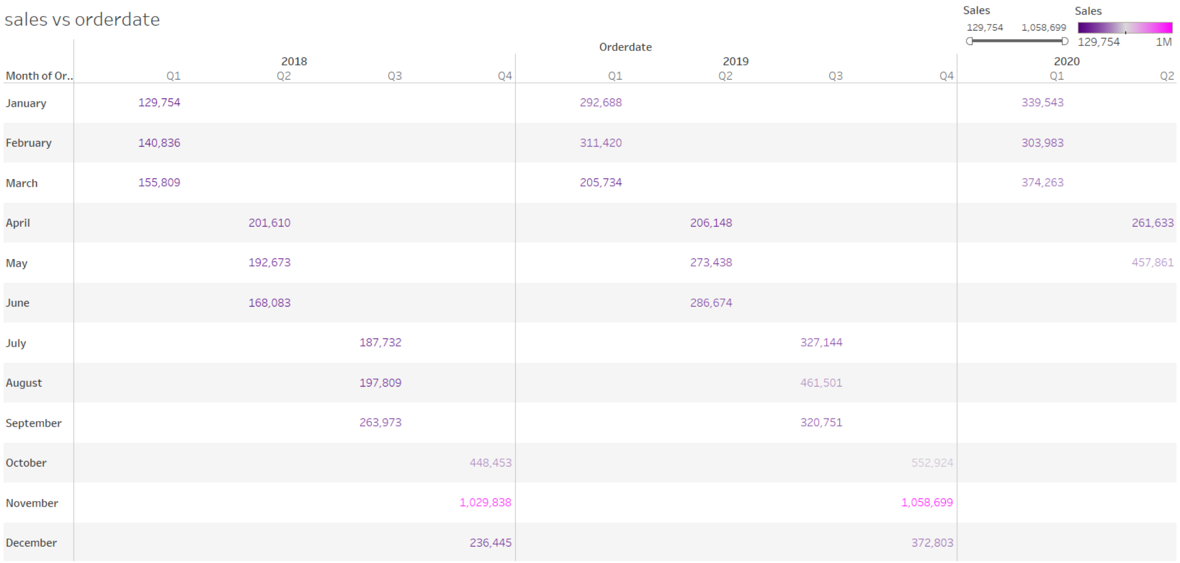
- As year-by-year increase there is also quarter by quarter increase.
- There is an increasing trend from quarter 1 to quarter 4.

▪ Monthly time series



- There is a continuous increase in months, increasing trend.
- There is a steady increase in sales from the month of August to Nov but dropped down in December. Should conduct promotional activities in December to get high sales in that month also.
- Highest sales month is Nov, where should stop marketing and focus on selling and stock

▪ Detailed time series:


sales vs orderdate

Sales: 129,754 – 1,058,699
Sales: 129,754 – 1M

	Orderdate									
	2018				2019				2020	
Month of Or..	Q1	Q2	Q3	Q4	Q1	Q2	Q3	Q4	Q1	Q2
January	129,754				292,688				339,543	
February	140,836				311,420				303,983	
March	155,809				205,734				374,263	
April		201,610				206,148				261,633
May		192,673				273,438				457,861
June		168,083				286,674				
July			187,732				327,144			
August			197,809				461,501			
September			263,973				320,751			
October				448,453				552,924		
November				1,029,838				1,058,699		
December				236,445				372,803		

- As said above there is an increasing trend from year to year, quarter to quarter.
- Highest sales in the month of Nov and lowest in the month of January which much increased from 2018 to 2020. The sales techniques which now using is effective should continue that.

2. Customer Segmentation using RMF analysis

2.1. Proper justification for the number of customer segments

The customers were segmented into four distinct groups based on their RFM (Recency, Frequency, and Monetary) scores. The segmentation logic was developed to classify customers meaningfully by purchasing behaviour and engagement patterns.

The **four segments** identified are:

- **Best Customers** – Recent purchasers who buy frequently and spend the most.
- **Loyal Customers** – Customers with consistent frequency and monetary value, even if not extremely recent.
- **Customers on the Verge of Churning** – Customers who used to buy more often but haven't done so recently.
- **Lost Customers** – Low in all three metrics, indicating minimal engagement and likelihood of churn.

This segmentation allows targeted marketing strategies for retention, reactivation, or loyalty rewards.

2.2. Mention the parameters used and the assumptions made

Parameters used are:

- **Recency**: Days since the last purchase. Lower values are better.
- **Frequency:** Number of transactions made. Higher values are better.
- **Monetary**: Total purchase value. Higher values are better.

Assumptions made:

Each RFM component was categorized into three levels:

- **L (Low)** – Bottom third of customers based on value
- **M (Medium)** – Middle third
- **H (High)** – Top third

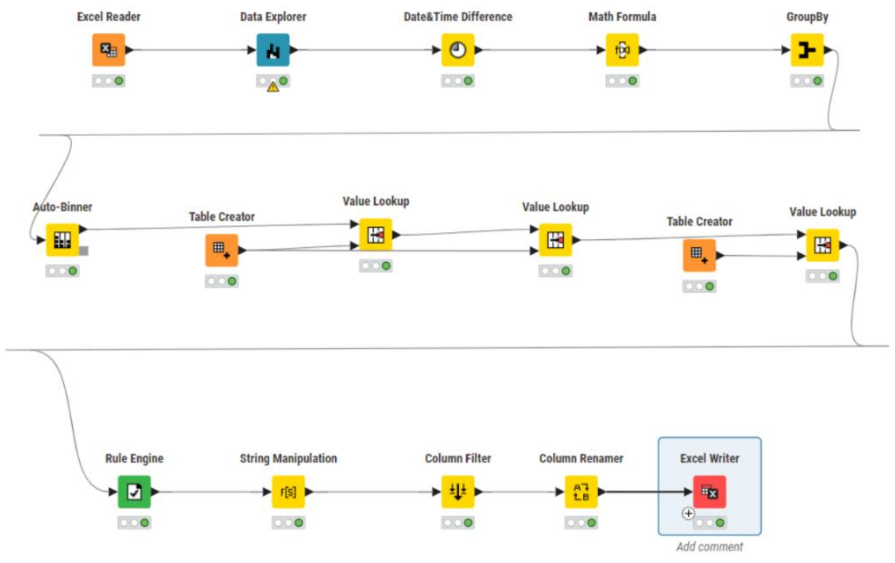
Assumption: The RFM scores were equally weighted in grouping customers, and each combination (e.g., HHH, LMM) holds interpretative value when mapped to behavioural segments.

2.3. Showcase the KNIME workflow

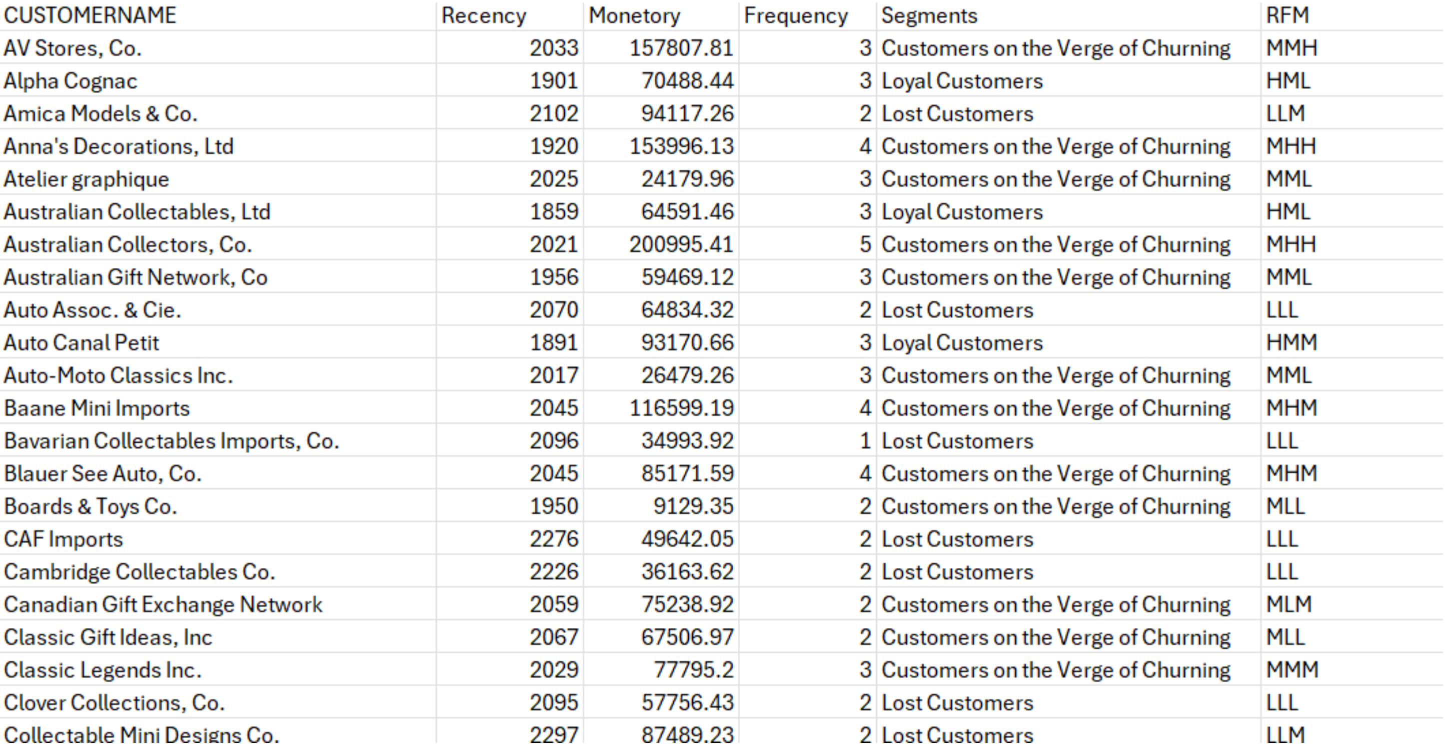
The KNIME workflow used includes the following steps:

1. **Data Preprocessing**: date formatting, and filtering necessary fields.
2. **RFM Calculation**:
   - **Recency** = Difference between the latest date and each customer's most recent purchase.
   - **Frequency** = Count of purchases per customer.
   - **Monetary** = Sum of total purchases per customer.
3. **RFM Binning**: Recency, Frequency, and Monetary were binned into L, M, H using quantile-based classification.
4. **RFM Group Creation**: Combined R, F, and M scores to form strings like "HHH", "LMM".
5. **Segmentation**: Rule Engine node used to assign segments based on RFM groups.
6. **Final Output**: A cleaned table with Customer ID, RFM group, and assigned segment.

Knime workflow:



2.4. Displaying the final output table



CUSTOMERNAME	Recency	Monetory	Frequency	Segments	RFM
AV Stores, Co.	2033	157807.81	3	Customers on the Verge of Churning	MMH
Alpha Cognac	1901	70488.44	3	Loyal Customers	HML
Amica Models & Co.	2102	94117.26	2	Lost Customers	LLM
Anna's Decorations, Ltd	1920	153996.13	4	Customers on the Verge of Churning	MHH
Atelier graphique	2025	24179.96	3	Customers on the Verge of Churning	MML
Australian Collectables, Ltd	1859	64591.46	3	Loyal Customers	HML
Australian Collectors, Co.	2021	200995.41	5	Customers on the Verge of Churning	MHH
Australian Gift Network, Co	1956	59469.12	3	Customers on the Verge of Churning	MML
Auto Assoc. & Cie.	2070	64834.32	2	Lost Customers	LLL
Auto Canal Petit	1891	93170.66	3	Loyal Customers	HMM
Auto-Moto Classics Inc.	2017	26479.26	3	Customers on the Verge of Churning	MML
Baane Mini Imports	2045	116599.19	4	Customers on the Verge of Churning	MHM
Bavarian Collectables Imports, Co.	2096	34993.92	1	Lost Customers	LLL
Blauer See Auto, Co.	2045	85171.59	4	Customers on the Verge of Churning	MHM
Boards & Toys Co.	1950	9129.35	2	Customers on the Verge of Churning	MLL
CAF Imports	2276	49642.05	2	Lost Customers	LLL
Cambridge Collectables Co.	2226	36163.62	2	Lost Customers	LLL
Canadian Gift Exchange Network	2059	75238.92	2	Customers on the Verge of Churning	MLM
Classic Gift Ideas, Inc	2067	67506.97	2	Customers on the Verge of Churning	MLL
Classic Legends Inc.	2029	77795.2	3	Customers on the Verge of Churning	MMM
Clover Collections, Co.	2095	57756.43	2	Lost Customers	LLL
Collectable Mini Designs Co.	2297	87489.23	2	Lost Customers	LLM

Final output has a customer name, recency, monetary, frequency, segments and RFM columns which are necessary to give insights and make decisions for future activities.

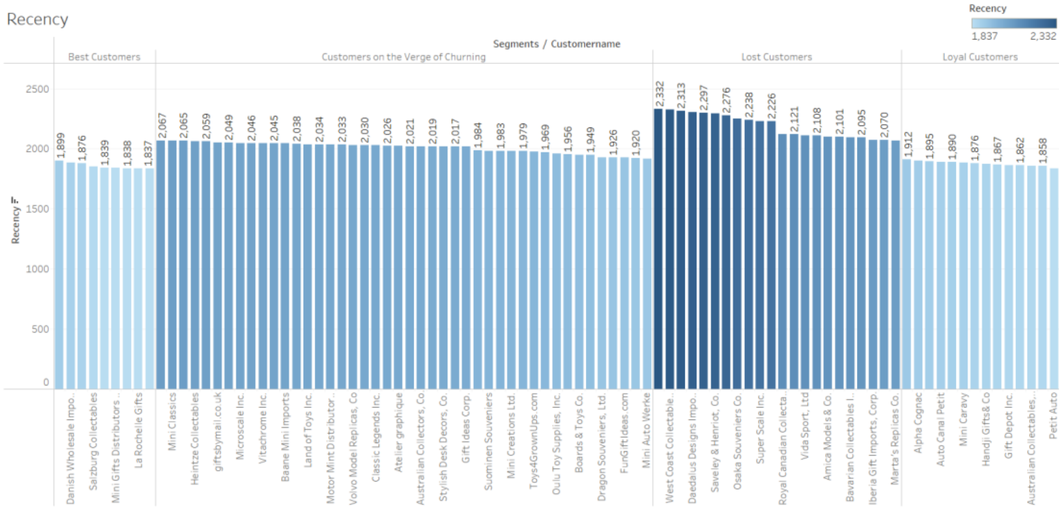
These columns will help to decide which customers to focus on the most and which are not to focus.

Segments column divides customers into segments as Best customers, loyal customers, customers on the verge of churning, and lost customers.

Best customers are those who have High recency, frequency and monetary too. Who are most important to the business who deserves rewards.

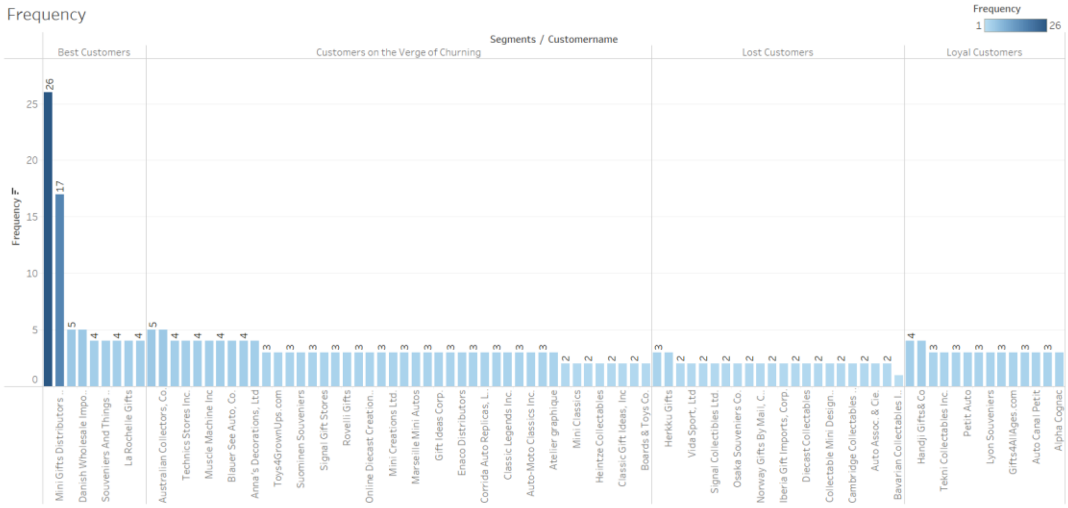
Analysis:

▪ Recency



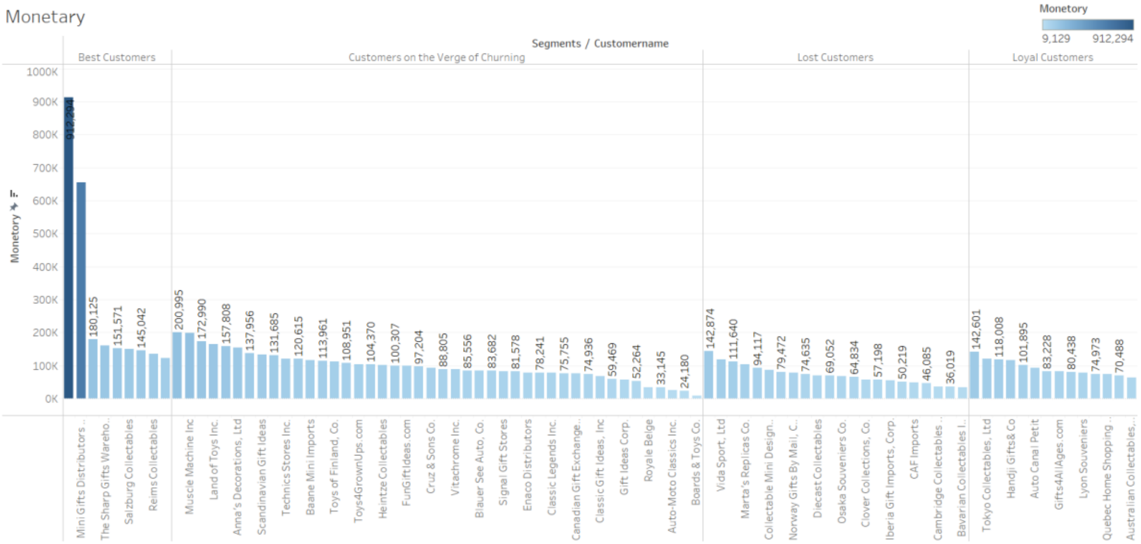
- Best customer has the lowest recency which means they have recently purchased.
- Where as lost customers has the highest recency, they visited so far.
- Loyal customers also have low recency.

▪ Frequency

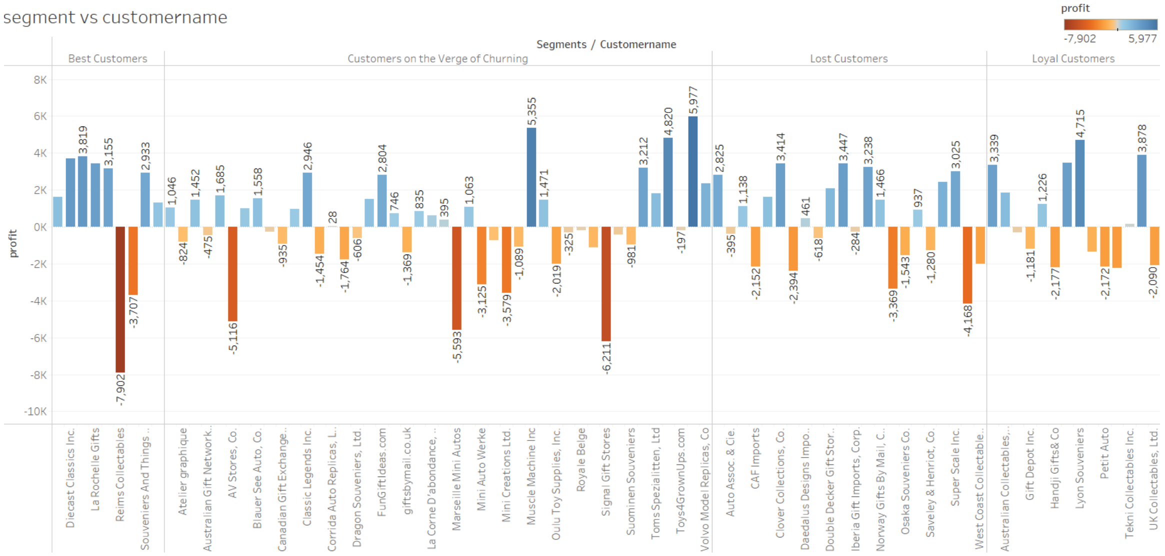


- Two of best customers have the highest frequency which are Euro shopping channel and mini gifts distributors ltd has 26 and 17 times frequency respectively.
- Other customers in best customer segments have 4 and 5 frequency and in loyal customer segments the frequencies are 4 and 3.

- **Monetary**



- Those customers who are said above are those who have the highest monetary
- But some of the customers who have the highest monetary but also belongs to the customers on the verge of churning.



- Reims collectable and Salzburg collectable are the customers who comes under best customers but still are the loss-making customers.
- Some of the customers who are profit-making customers are still belongs to customers on the verge of churning which have to consider and try to increase sales by doing some kind of marketing and discounts.
- Loyal customer segments most of the customers are loss-making ones.
- Even in lost customers there are some customers who are profit-making.

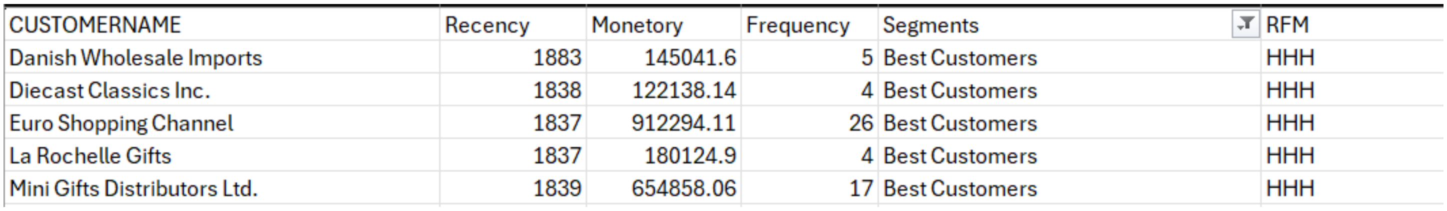
3. Inferences from RFM Analysis:

3.1. Top five customers from each segment.

Based on the RFM segmentation conducted in KNIME, customers have been classified into specific categories that reflect their current engagement with the business. The following are the top five customers from each key category, selected based on a combination of **high or low Recency, Frequency, and Monetary values**, and aligned with behavioral characteristics defined below.

Best customers:

Definition: **High** Recency, **High** Frequency, **High** Monetary
These are the most valuable customers who purchase regularly, recently, and spend significantly.

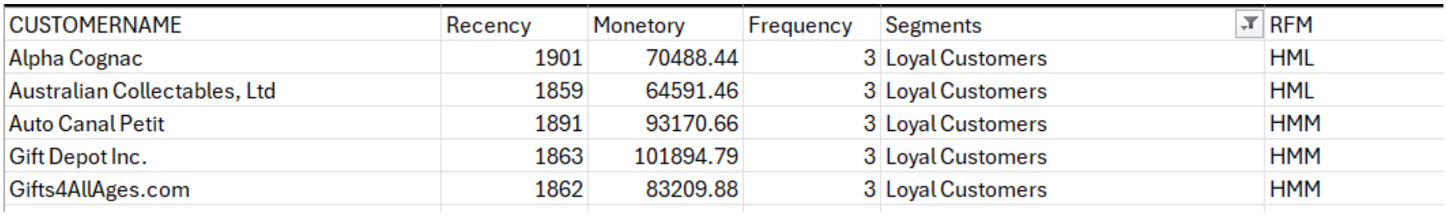


CUSTOMERNAME	Recency	Monetory	Frequency	Segments	RFM
Danish Wholesale Imports	1883	145041.6	5	Best Customers	HHH
Diecast Classics Inc.	1838	122138.14	4	Best Customers	HHH
Euro Shopping Channel	1837	912294.11	26	Best Customers	HHH
La Rochelle Gifts	1837	180124.9	4	Best Customers	HHH
Mini Gifts Distributors Ltd.	1839	654858.06	17	Best Customers	HHH

Insight: These customers should be prioritized for loyalty rewards, premium offers, or early access promotions.

Loyal customers:

Definition: **Medium to High** Frequency and Monetary with **High** Recency. They purchase regularly and bring good value, though recency may vary.

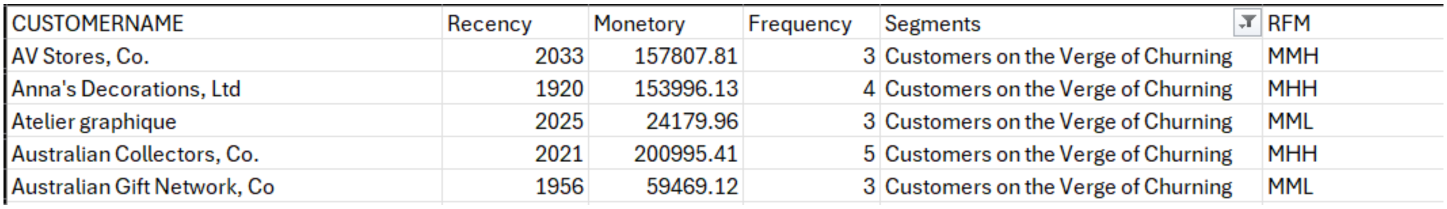


CUSTOMERNAME	Recency	Monetory	Frequency	Segments	RFM
Alpha Cognac	1901	70488.44	3	Loyal Customers	HML
Australian Collectables, Ltd	1859	64591.46	3	Loyal Customers	HML
Auto Canal Petit	1891	93170.66	3	Loyal Customers	HMM
Gift Depot Inc.	1863	101894.79	3	Loyal Customers	HMM
Gifts4AllAges.com	1862	83209.88	3	Loyal Customers	HMM

Insight: Maintain engagement with personalized offers, membership benefits, or referral bonuses.

Customers on verger of churning:

Definition: Medium Recency with Medium/High/ Low Frequency and Monetary in the past. These customers were active but haven't purchased recently, indicating risk of churn.

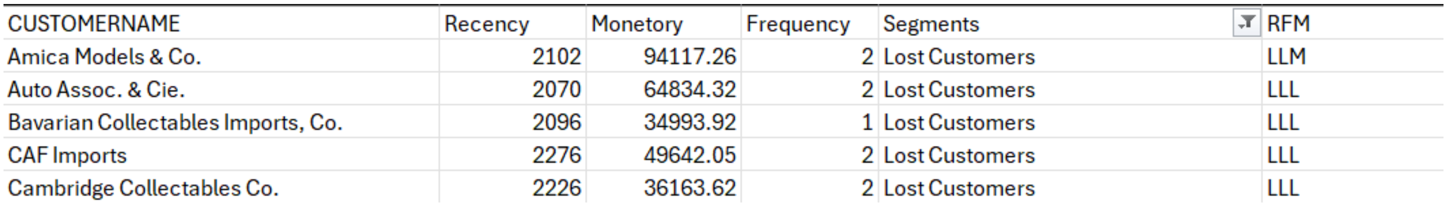


CUSTOMERNAME	Recency	Monetory	Frequency	Segments	RFM
AV Stores, Co.	2033	157807.81	3	Customers on the Verge of Churning	MMH
Anna's Decorations, Ltd	1920	153996.13	4	Customers on the Verge of Churning	MHH
Atelier graphique	2025	24179.96	3	Customers on the Verge of Churning	MML
Australian Collectors, Co.	2021	200995.41	5	Customers on the Verge of Churning	MHH
Australian Gift Network, Co	1956	59469.12	3	Customers on the Verge of Churning	MML

Insight: Consider **win-back campaigns**, such as reactivation discounts, reminders, or surveys to regain their interest.

Lost customers:

Definition: **Low** Recency, **Low** Frequency, **Low** Monetary
Customers who have not engaged for a long time and showed low value even when active.



CUSTOMERNAME	Recency	Monetory	Frequency	Segments	RFM
Amica Models & Co.	2102	94117.26	2	Lost Customers	LLM
Auto Assoc. & Cie.	2070	64834.32	2	Lost Customers	LLL
Bavarian Collectables Imports, Co.	2096	34993.92	1	Lost Customers	LLL
CAF Imports	2276	49642.05	2	Lost Customers	LLL
Cambridge Collectables Co.	2226	36163.62	2	Lost Customers	LLL

Insight: These customers can be **excluded from high-cost campaigns** but considered for **low-cost bulk promotions or feedback collection**.

Conclusion of RFM analysis

This segmentation provides a clear path for personalized marketing strategies:

- Best customers need loyalty programs
- Loyal customers need continued engagement
- Customers on the verge need reactivation strategies
- Lost customers require either feedback or budget-friendly campaigns