Market Retail Analysis – Automobile Parts sales Coded Project

Part A

Business Report

DSBA – Course

Created by – Rishabh Gupta

Foreword

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.

Your role as a Business Analyst is to use the provided dataset to achieve these goals and present findings in a manner that can guide the company's decision-making.

-

Contents

List of Tables

List of Figures

Objective

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.

For this assignment, we will analyze data do EDA using python and RFM using Knime.

After EDA, I have removed not so important columns and have only below 4 columns required for RFM analysis –

- ORDERNUMBER
- CUSTOMERNAME
- ORDERDATE
- SALES

Data Analysis of problem statement –

The dataset contains data corresponding to polished and unpolished stones.

Sheet name –

1. *Sales_Data.xlsx*

Part A –

Data Dictionary –

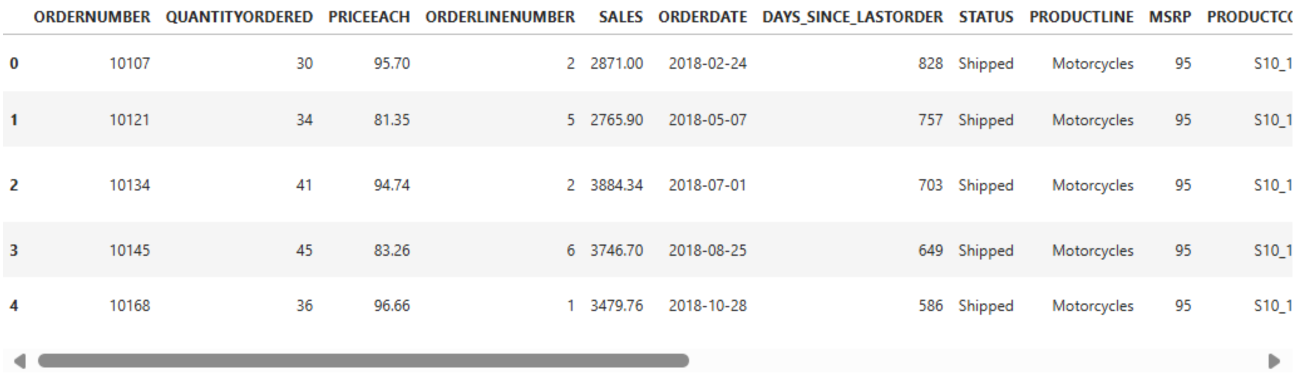
The dataset provided contains three years of transactional data from the company, with each row representing a unique order. Below is an explanation of the key attributes:

- 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).

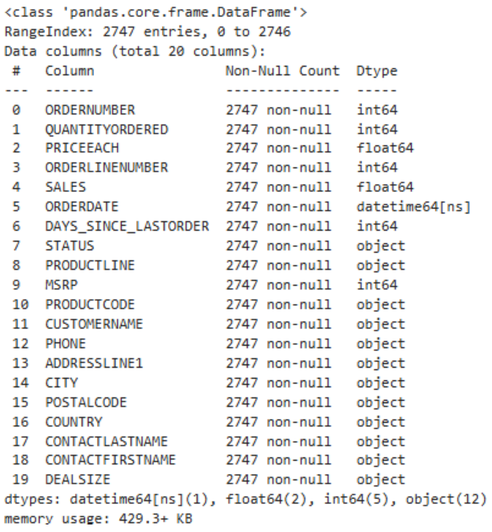
Data Overview –

To start the data analysis, we need to import the necessary libraries, specify the working directory, and load the dataset. Then, we will view the first five rows using head () function to get a preliminary understanding of the data. The Dataset has 2747 number of rows with 20 columns.



	ORDERNUMBER	QUANTITYORDERED	PRICEEACH	ORDERLINENUMBER	SALES	ORDERDATE	DAYS_SINCE_LASTORDER	STATUS	PRODUCTLINE	MSRP	PRODUCTC
0	10107	30	95.70	2	2871.00	2018-02-24	828	Shipped	Motorcycles	95	S10_1
1	10121	34	81.35	5	2765.90	2018-05-07	757	Shipped	Motorcycles	95	S10_1
2	10134	41	94.74	2	3884.34	2018-07-01	703	Shipped	Motorcycles	95	S10_1
3	10145	45	83.26	6	3746.70	2018-08-25	649	Shipped	Motorcycles	95	S10_1
4	10168	36	96.66	1	3479.76	2018-10-28	586	Shipped	Motorcycles	95	S10_1

TABLE 1 - TOP 5 ROWS OF DATASET



```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 2747 entries, 0 to 2746
Data columns (total 20 columns):
 #   Column                Non-Null Count  Dtype
---  ------                --------------  -----
 0   ORDERNUMBER           2747 non-null   int64
 1   QUANTITYORDERED       2747 non-null   int64
 2   PRICEEACH             2747 non-null   float64
 3   ORDERLINENUMBER       2747 non-null   int64
 4   SALES                 2747 non-null   float64
 5   ORDERDATE             2747 non-null   datetime64[ns]
 6   DAYS_SINCE_LASTORDER  2747 non-null   int64
 7   STATUS                2747 non-null   object
 8   PRODUCTLINE           2747 non-null   object
 9   MSRP                  2747 non-null   int64
 10  PRODUCTCODE           2747 non-null   object
 11  CUSTOMERNAME          2747 non-null   object
 12  PHONE                 2747 non-null   object
 13  ADDRESSLINE1          2747 non-null   object
 14  CITY                  2747 non-null   object
 15  POSTALCODE            2747 non-null   object
 16  COUNTRY               2747 non-null   object
 17  CONTACTLASTNAME       2747 non-null   object
 18  CONTACTFIRSTNAME      2747 non-null   object
 19  DEALSIZE              2747 non-null   object
dtypes: datetime64[ns](1), float64(2), int64(5), object(12)
memory usage: 429.3+ KB
```

TABLE 2 - BASIC INFO. OF DATASET

Regarding the datatype of the columns in the dataset, we can confirm that-

• We can observe that there is 1 datetime variable, 7 numerical variables and 12 object variables

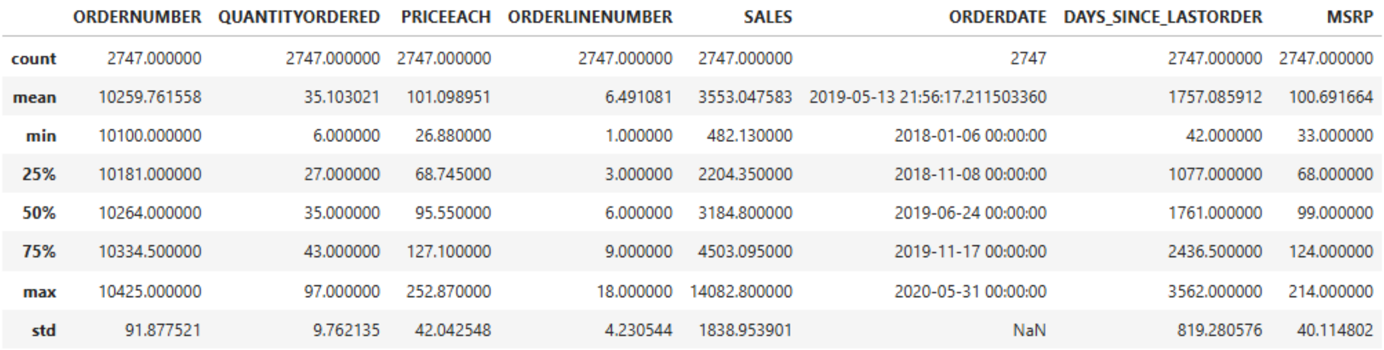
Data Pre-processing-

Missing value treatment and Analysis-

- On analysis, we can observe there are no missing values.

Statistical Summary –

Using Describe () function, we can analyses the summary statistics of the dataset –



	ORDERNUMBER	QUANTITYORDERED	PRICEEACH	ORDERLINENUMBER	SALES	ORDERDATE	DAYS_SINCE_LASTORDER	MSRP
count	2747.000000	2747.000000	2747.000000	2747.000000	2747.000000	2747	2747.000000	2747.000000
mean	10259.761558	35.103021	101.098951	6.491081	3553.047583	2019-05-13 21:56:17.211503360	1757.085912	100.691664
min	10100.000000	6.000000	26.880000	1.000000	482.130000	2018-01-06 00:00:00	42.000000	33.000000
25%	10181.000000	27.000000	68.745000	3.000000	2204.350000	2018-11-08 00:00:00	1077.000000	68.000000
50%	10264.000000	35.000000	95.550000	6.000000	3184.800000	2019-06-24 00:00:00	1761.000000	99.000000
75%	10334.500000	43.000000	127.100000	9.000000	4503.095000	2019-11-17 00:00:00	2436.500000	124.000000
max	10425.000000	97.000000	252.870000	18.000000	14082.800000	2020-05-31 00:00:00	3562.000000	214.000000
std	91.877521	9.762135	42.042548	4.230544	1838.953901	NaN	819.280576	40.114802

Table 3 - Statistical summary of dataset

Observations-

- **General:** One missing value in DAYS_SINCE_LASTORDER.
- **ORDERNUMBER:** Range 10100-10425, avg ~10259 (clustered).
- **QUANTITYORDERED:** Range 6-97, avg ~35 (moderate spread).
- **PRICEEACH:** Range $26.88-$252.87, avg ~$101.10 (wide range).
- **ORDERLINENUMBER:** Range 1-18, avg ~6.5 (moderate variability).
- **SALES:** Range $482.13-$14082.80, avg ~$3553.05 (significant variation).
- **ORDERDATE:** Jan 2018 - May 2020, avg May 2019.
- **DAYS_SINCE_LASTORDER:** Range 42-3562, avg ~1757 (high disparity in purchase frequency).
- **MSRP:** Range $33-$214, avg ~$100.69 (moderate spread).

Exploratory Data Analysis

Let’s analyze Histograms for Numerical Variables (e.g., QUANTITYORDERED, PRICEEACH, SALES, DAYS_SINCE_LASTORDER, MSRP):

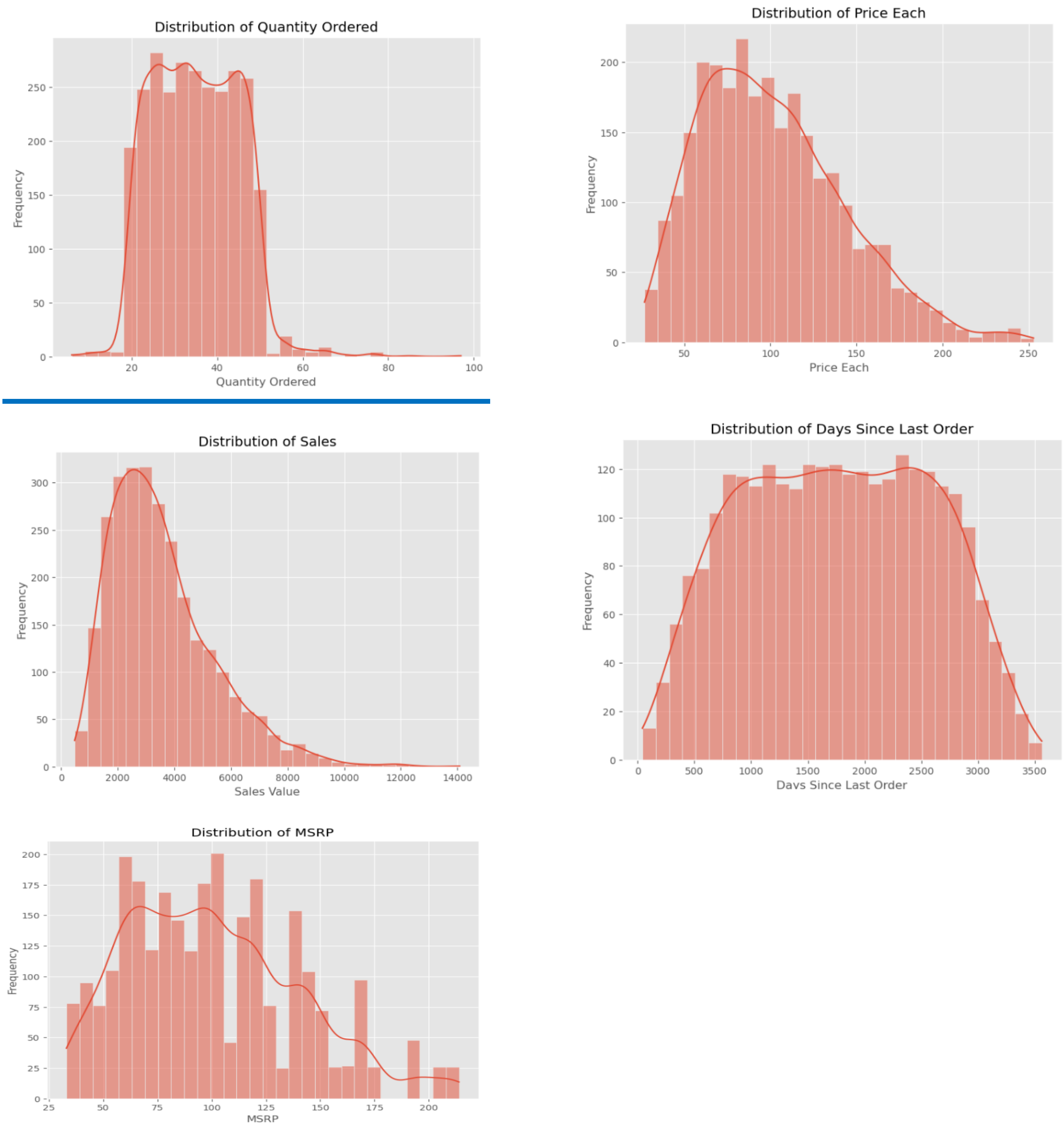


Figure 1 - Histograms for Numerical variables

Observations:

- **Quantity Ordered:** Distribution appears multimodal, with a primary peak around 30 and a secondary peak near 45. Right-skewed with a tail extending to higher quantities.
- **Price Each:** Right-skewed distribution, peaking around $75-$100, with a long tail towards higher prices.
- **Sales:** Strongly right-skewed, peaking at lower sales values ($0-$2000), with a significant tail indicating some high-value transactions.

- **Days Since Last Order:** Relatively uniform distribution between approximately 500 and 3000 days, suggesting varied customer return intervals.
- **MSRP:** Bimodal distribution with peaks around $50-$75 and $100-$125, indicating two common price points for products.

Now lets analyze countplot for other numerical variables-

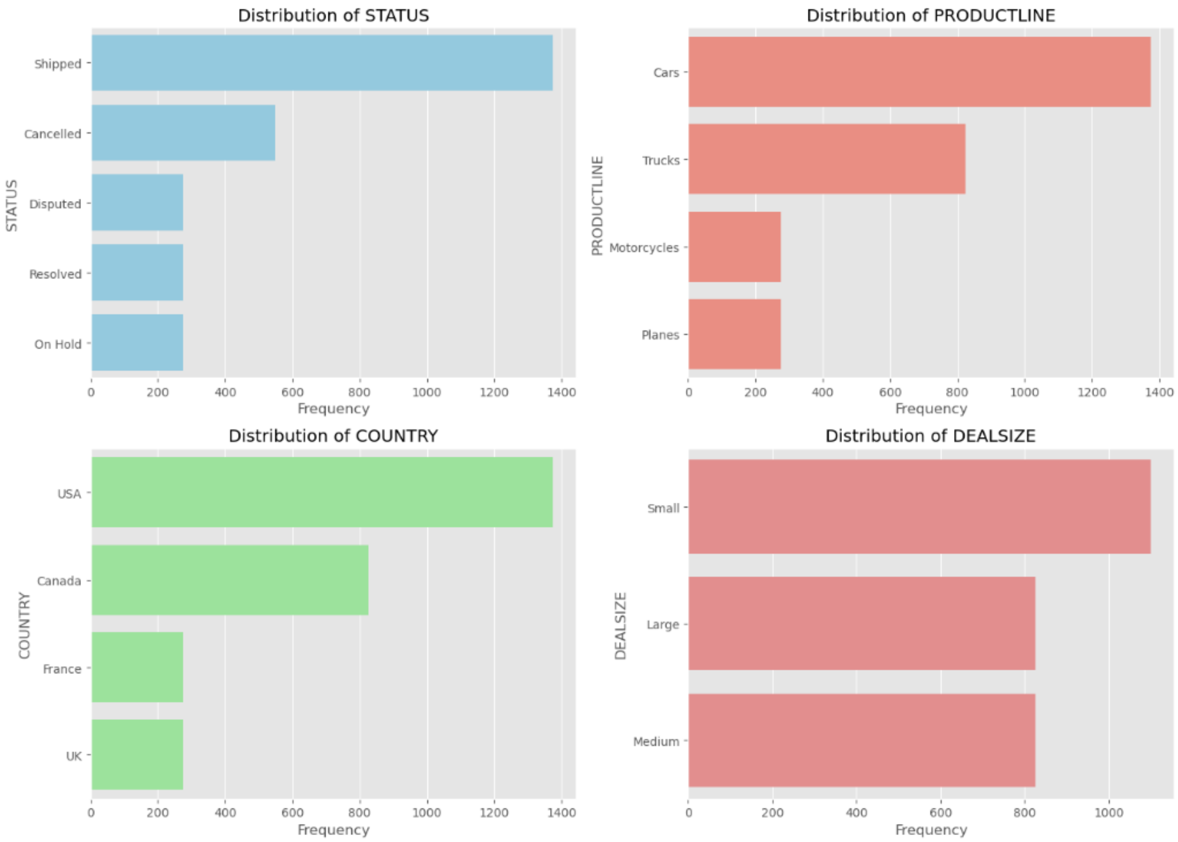


Figure 2 - Countplot for Numerical analysis

Observations:

- **STATUS:** "Shipped" dominates. "Disputed," "Cancelled," "On Hold," "Resolved" are less frequent, indicating potential areas for process review.

- **PRODUCTLINE:** "Cars" are the top seller, followed by "Trucks." "Motorcycles" and "Planes" are niche. Focus on "Cars" and explore growth for others.

- **COUNTRY:** USA is the primary market, Canada secondary. France and UK have a smaller footprint, suggesting international growth potential.

- **DEALSIZE:** "Small" and "Large" are prevalent, "Medium" less so. Analyze drivers for "Small" and "Large" deals.

- **Overall:** Clear categorical distributions with significant imbalances, highlighting strengths and potential areas for investigation and targeted strategies.

Now lets analyze barplot for high count in numerical columns-

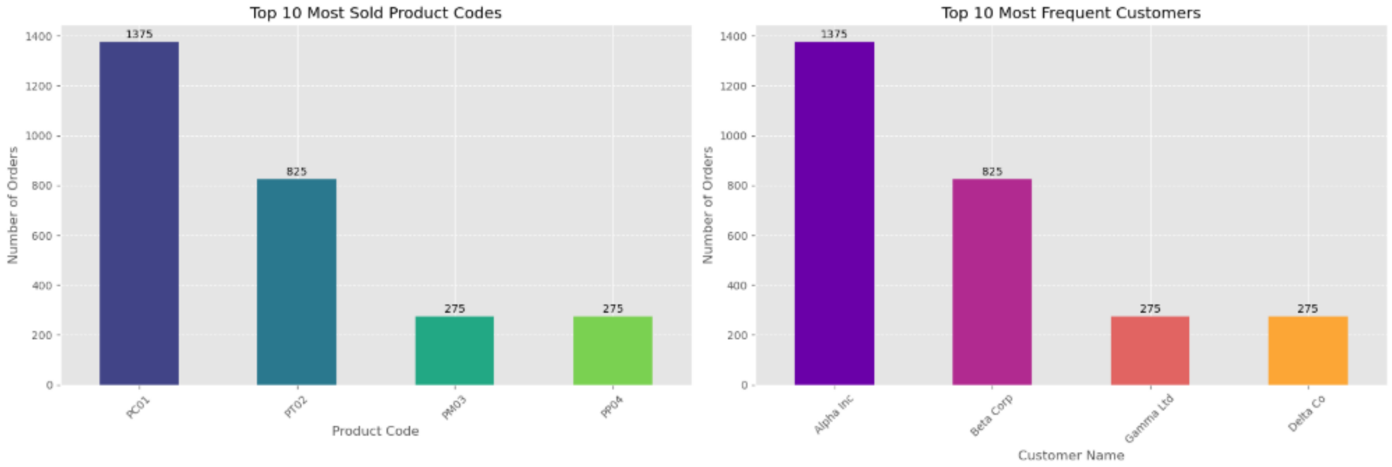


FIGURE 3 - BARPLOT FOR MOST SOLD AND FREQUENT CUSTOMERS

- **Top Products:** Product code 'PC01' is the clear top seller, significantly outperforming other top product codes in order frequency.
- **Second Tier Products:** 'PT02' represents the second most frequently ordered product, with a noticeable drop in order count compared to 'PC01'.
- **Less Frequent Top Products:** 'PM03' and 'PP04' have identical, and considerably lower, order frequencies compared to the top two.
- **Top Customers:** 'Alpha Inc' is the most frequent customer, showing a similar dominance in order count as 'PC01' among products.
- **Frequent Customer Groups:** 'Beta Corp' is the second most frequent customer, while 'Gamma Ltd' and 'Delta Co' exhibit the same, lower order frequency among the top customers.

Now let's analyze barplot for Total Sales and Avg Sales for numerical columns-

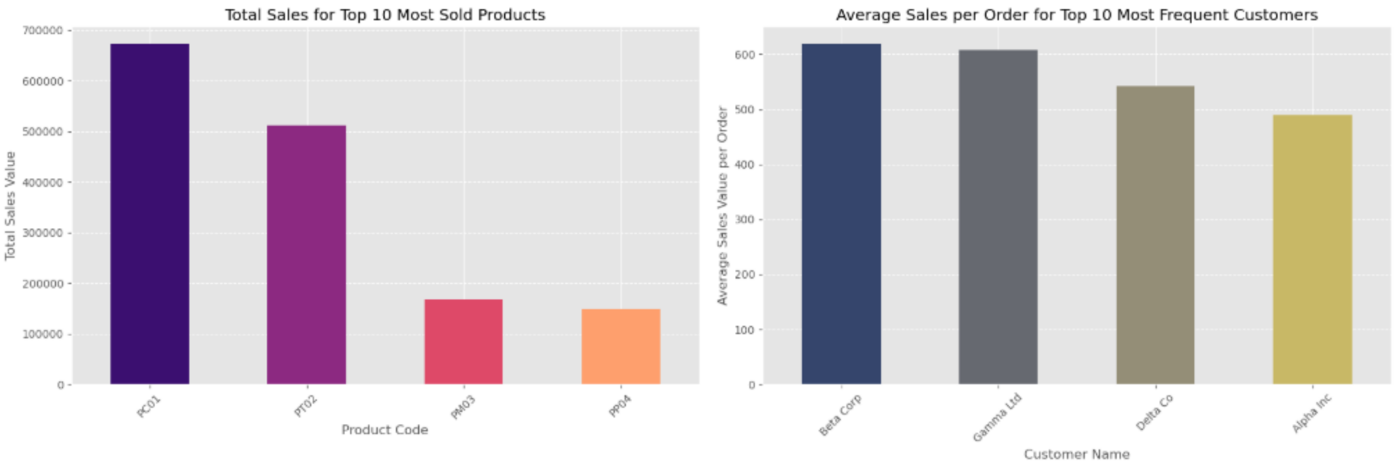


Figure 4 - Total Sales and Avg Sales

- **Top Product Sales:** While 'PC01' leads in order frequency, it also generates the highest total sales value, significantly surpassing other top products. 'PT02' contributes the second-highest total sales, maintaining its relative importance.
- **Lower Sales Contribution:** 'PM03' and 'PP04', despite being in the top ordered products, contribute considerably less to the overall sales revenue compared to 'PC01' and 'PT02'.
- **Average Customer Spend:** 'Beta Corp' exhibits the highest average sales value per order among the top frequent customers. 'Gamma Ltd' and 'Delta Co' also show relatively high average spending.
- **Frequent but Lower Average Spend:** 'Alpha Inc', despite being the most frequent customer, has the lowest average sales value per order among the top frequent buyers.

Multivariate Analysis –

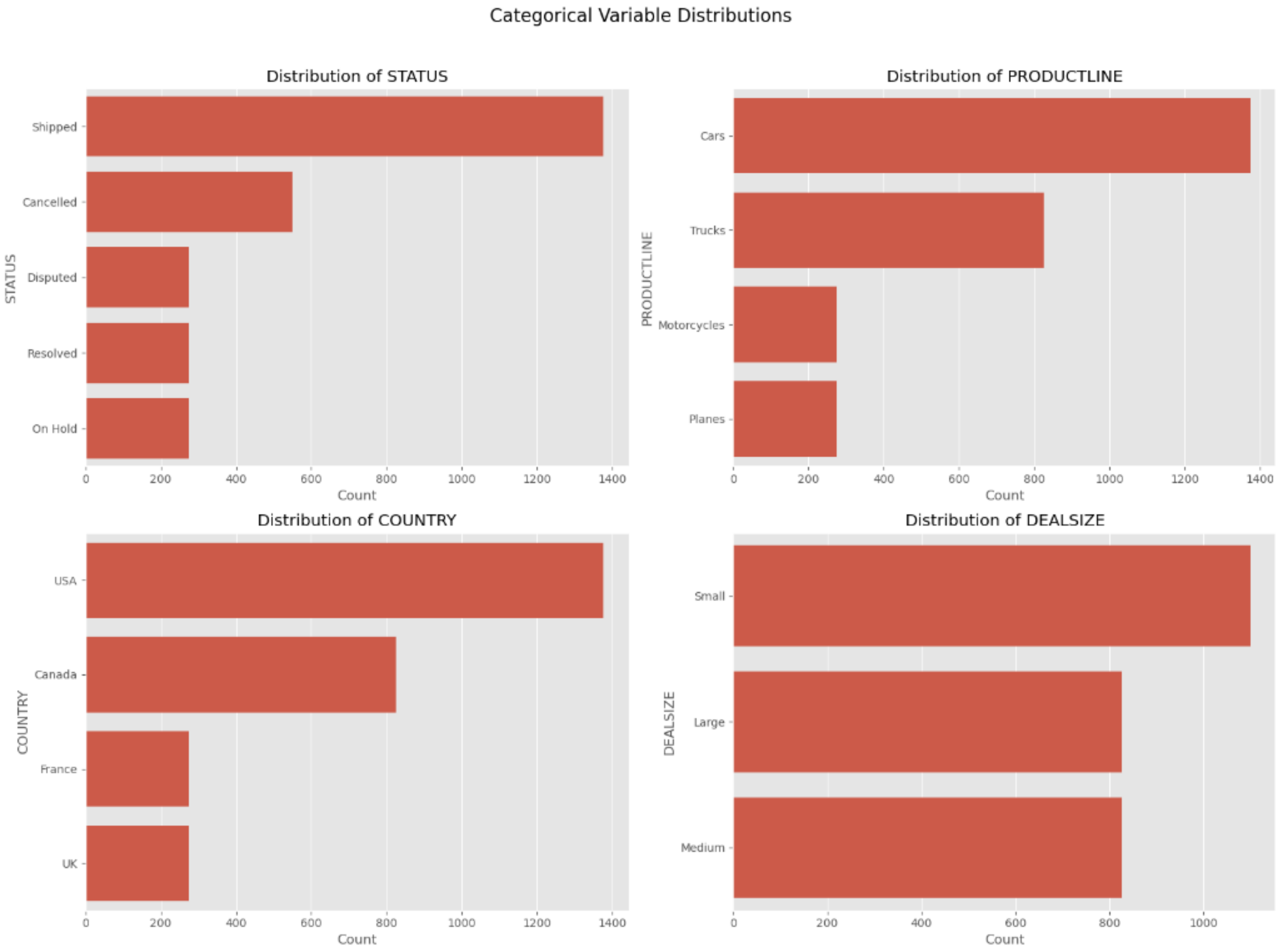


FIGURE 5 - CATEGORICAL VARIABLES DISTRIBUTION

- STATUS: Most orders are "Shipped," while "Cancelled" is the second most frequent status, suggesting potential fulfilment issues.

- PRODUCTLINE: "Cars" dominate the product line sales, followed by "Trucks," while "Planes" have the least representation.

- COUNTRY: The USA has the highest number of orders, followed by Canada, while France and the UK have significantly fewer.

- DEALSIZE: "Small" deals are the most common, while "Medium" and "Large" deals are almost equally distributed.

- Potential Insights: The dominance of "Cars" in product lines and "Shipped" in status suggests strong market demand but may warrant an analysis of cancellation reasons.

- Actionable Steps: Investigate cancellation rates, assess demand by country, and analyze deal sizes for potential revenue optimization.

Now lets compare Weekly and Monthly sales trend –

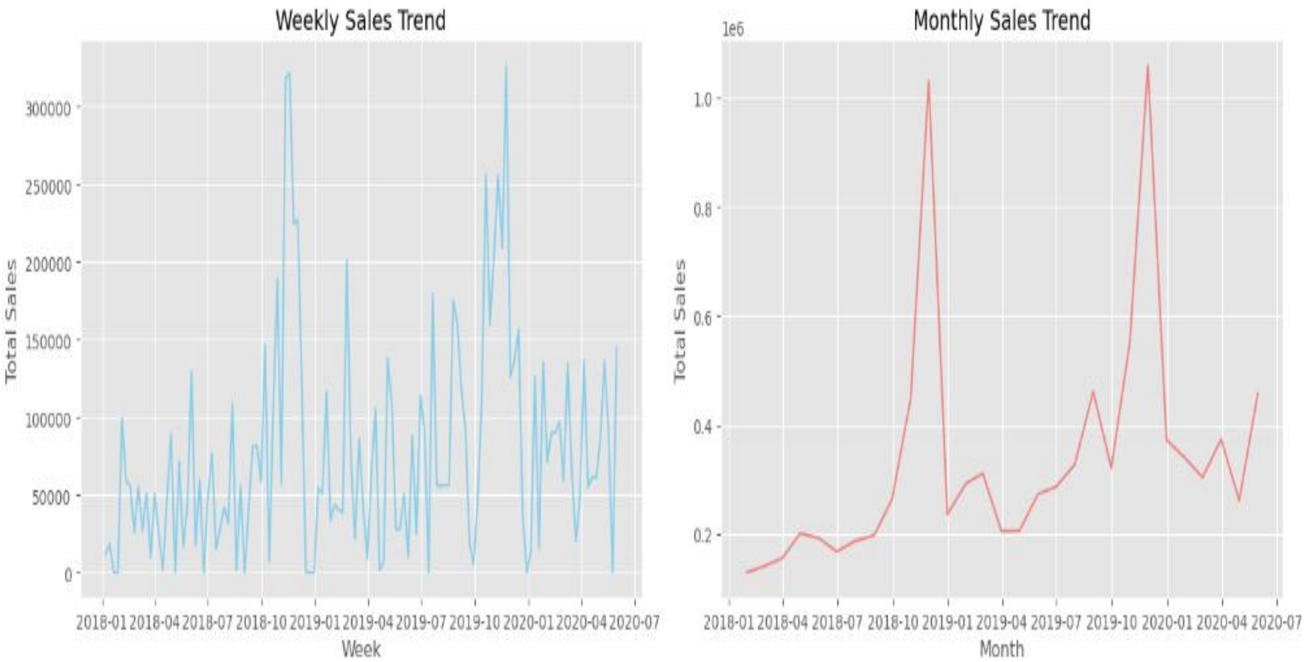


FIGURE 6 - WEEKLY AND MONTHLY SALES TREND

- **Seasonal Peaks**: Monthly sales show two major spikes — one around late 2018 and another in late 2019 — suggesting strong seasonal demand, possibly during holiday periods.
- **Weekly Variability**: The weekly trend is more volatile, with frequent sharp rises and drops, indicating short-term fluctuations in order volume.
- **Year-over-Year Pattern**: Sales activity increases noticeably in the second half of each year, aligning with peak months, reinforcing a seasonal sales cycle.

- **Growth Signals**: Despite fluctuations, there's a general upward trend in both weekly and monthly sales toward the end of 2019.
- **Potential Drops**: A few sudden dips to near-zero weekly sales may point to data gaps, holidays, or operational downtimes worth investigating.

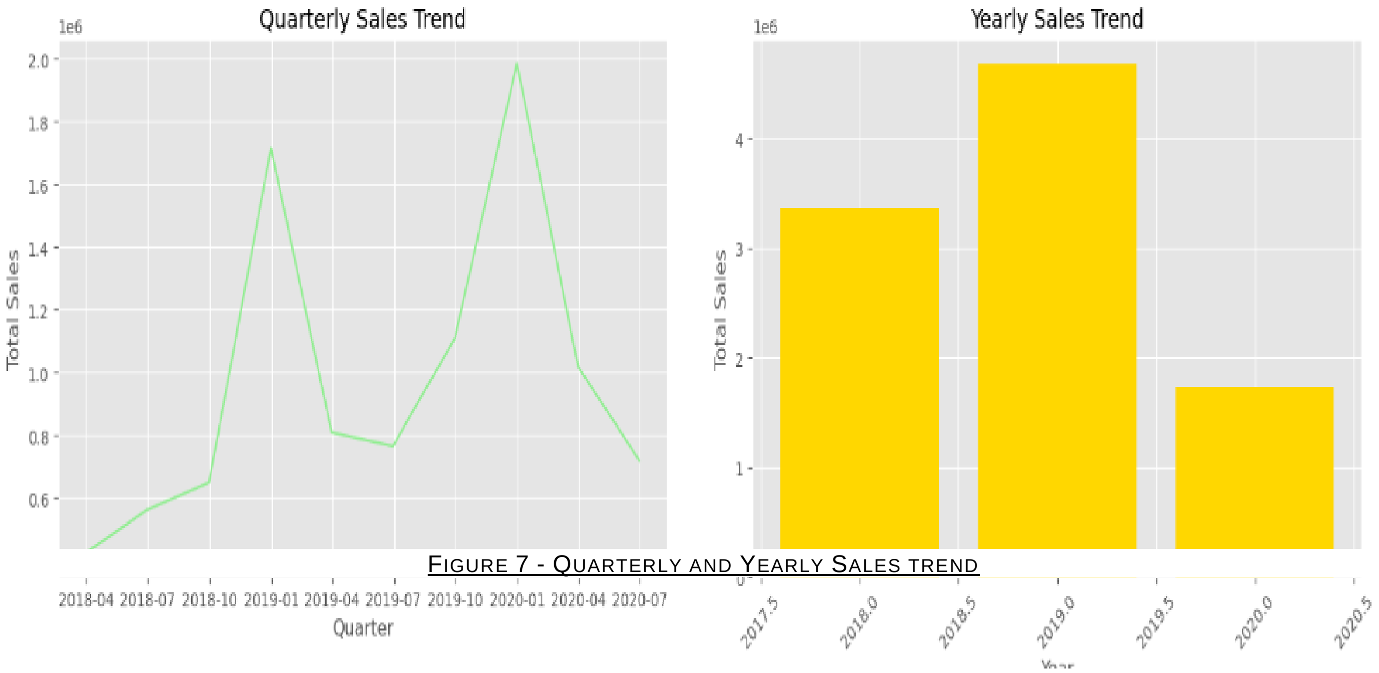


FIGURE 7 - QUARTERLY AND YEARLY SALES TREND

- **Quarterly Sales Peaks:** There are two noticeable spikes in sales—one in late 2018 and another, even higher, in late 2019—indicating strong seasonal trends.
- **Sales Decline in 2020:** Both quarterly and yearly trends show a **drop in sales** in early 2020, suggesting potential market slowdown, reduced demand, or external disruptions.
- **2019 Was the Best Year:** The **highest annual sales** occurred in 2019, surpassing both 2018 and 2020, reinforcing a growth trend before the decline.
- **Quarterly Fluctuations:** Sales do not remain steady across quarters; instead, there are periodic surges, likely due to seasonal demand or promotions.
- **2020 Underperformance:** Compared to previous years, **2020's sales are significantly lower**, possibly due to economic shifts, supply chain disruptions, or reduced customer demand.

Analysis Monthly sales w.r.t Product line -

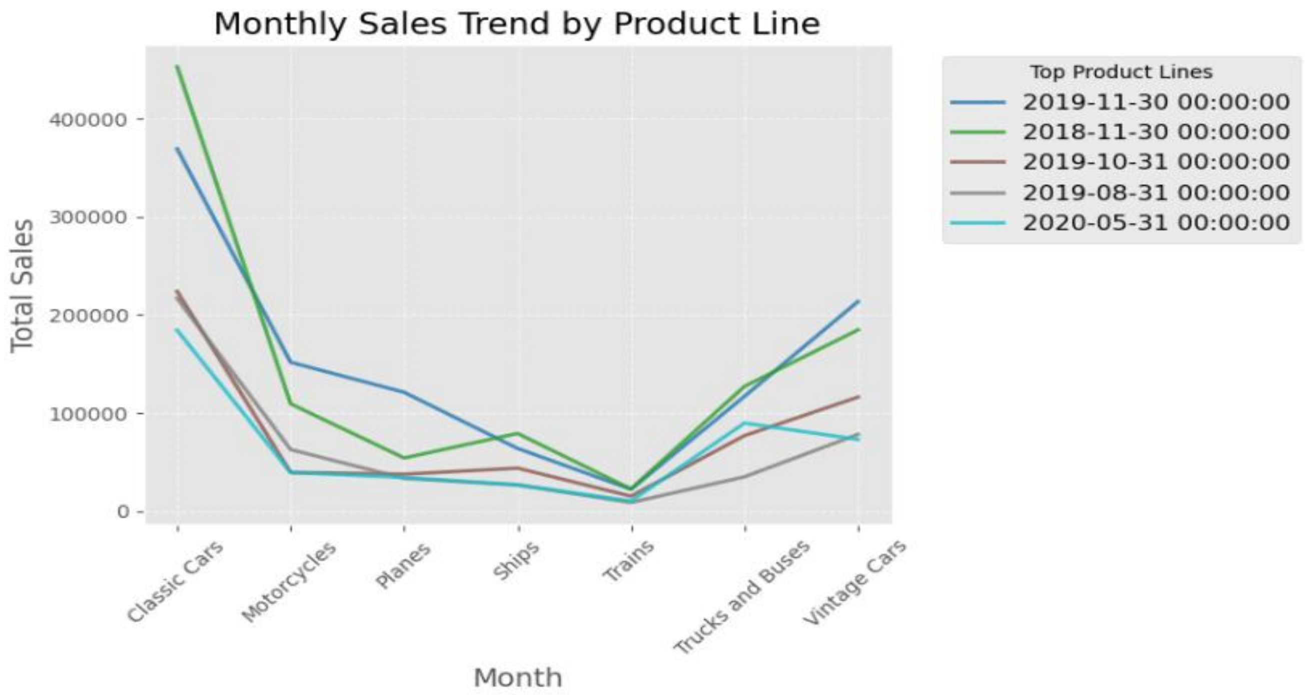


FIGURE 8 - MONTHLY SALES BY PRODUCT LINE

- **Classic Cars** consistently lead in sales, showing the highest figures across all months.
- **Motorcycles and Planes** have moderate sales but decline quickly after the initial months.
- **Ships and Trains** maintain relatively low sales throughout the timeline.
- A **noticeable dip** occurs in the middle of the timeline across all product lines, possibly due to seasonality.
- Sales of **Trucks and Buses and Vintage Cars** pick up towards the end, indicating a late surge.
- The **legend appears to show dates instead of product lines**, making interpretation difficult.
- The **color scheme helps differentiate trends**, but more spacing between lines could improve readability.
- Adjusting the **legend and labels** will provide better clarity for analysis.

Analysis Monthly sales w.r.t Deal Size –

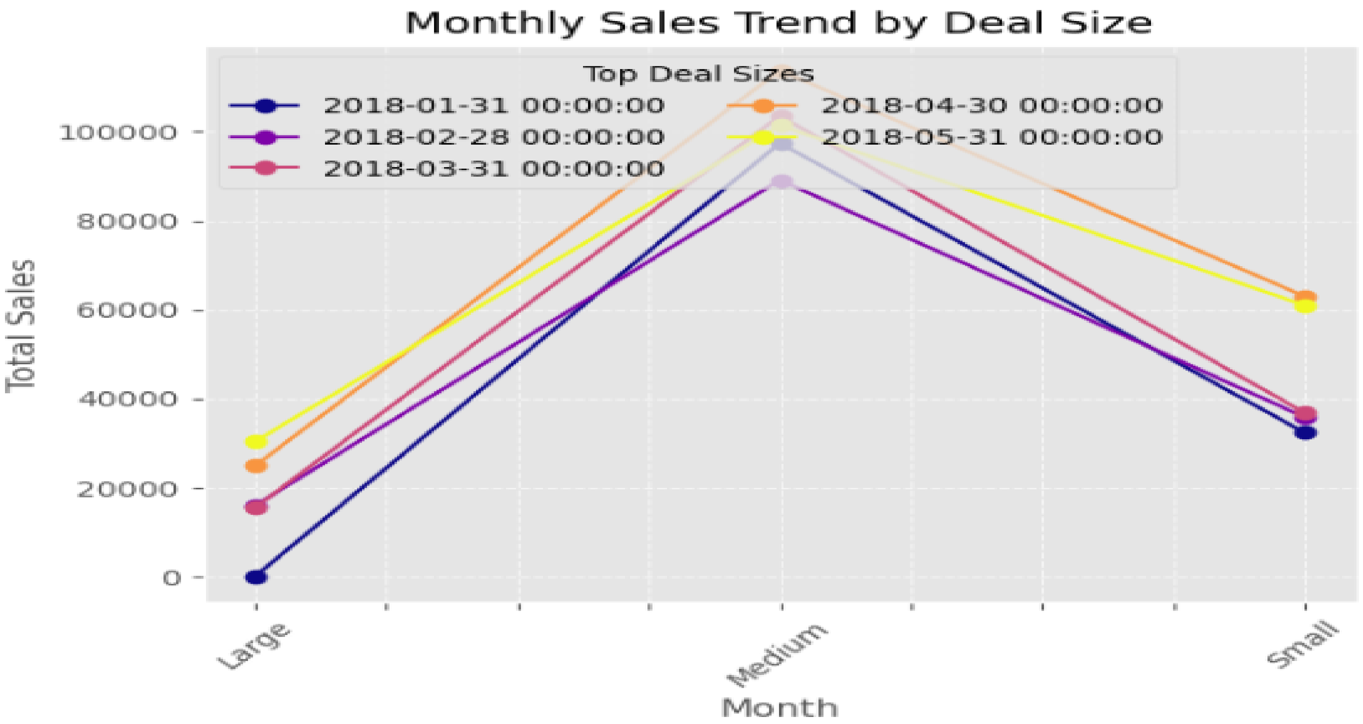


FIGURE 9 - MONTHLY SALES W.R.T DEAL SIZE

- **Medium Deal Size Dominates** → Across all selected months, the **Medium** deal size consistently shows the highest sales.
- **Large and Small Deal Sizes Show Similar Trends** → Sales for **Large** and **Small** deal sizes follow a similar pattern, staying lower than the Medium category.
- **Sales Peak at Medium Deal Size** → Each month exhibits a peak in sales for the **Medium** deal size, indicating it is the most common or profitable category.
- **April and May 2018 Show High Sales** → Sales for **April 30, 2018**, and **May 31, 2018**, are among the highest across deal sizes, especially in the Medium category.
- **General Upward Trend from Large to Medium, Then Decline** → Sales increase from **Large to Medium** and then decline when reaching **Small** deals.
- **Different Color Coding Helps Distinguish Trends** → The colormap helps in differentiating various months but may still need fine-tuning for better clarity.
- **Legend Placement Needs Adjustment** → The legend overlaps with the data points, making it slightly difficult to read. It would be better placed outside the plot.
- **Potential for Further Insights** → A deeper breakdown by individual months or year-over-year comparisons could provide more context into why certain months perform better.

Boxplots –

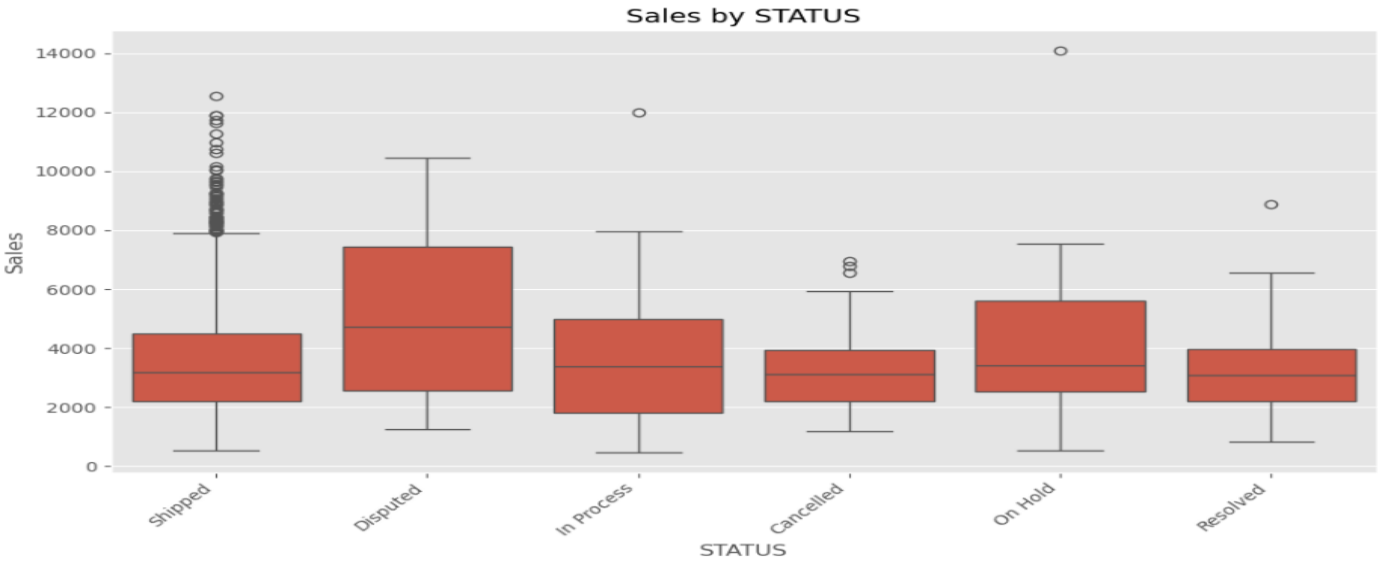


FIGURE 10 - BOXPLOT FOR SALES BY STATUS

- "Shipped" orders have a wide range of sales values and several high outliers: The box for 'Shipped' orders is relatively tall, indicating a significant variability in sales amounts. The presence of many outliers above the upper whisker suggests a number of very high-value shipped orders.
- "Disputed" orders tend to have higher median and upper quartile sales compared to "Shipped": While also having a considerable range, the central tendency of 'Disputed' order sales appears higher than 'Shipped', suggesting that disputes might often involve more valuable orders.
- "In Process" and "On Hold" orders show similar distributions with moderate sales values: Both of these statuses exhibit relatively similar box sizes and median sales, indicating a comparable range and central tendency of sales for orders currently in these states.
- "Cancelled" and "Resolved" orders have lower median sales: The median sales values for 'Cancelled' and 'Resolved' orders are noticeably lower than 'Shipped' and 'Disputed', suggesting that these order outcomes are often associated with less valuable transactions.
- "Cancelled" orders have a tighter interquartile range compared to "Resolved": The box for 'Cancelled' orders is shorter than 'Resolved', indicating less variability in the middle 50% of their sales values. 'Resolved' orders show a slightly wider spread in their central sales amounts.
- Outliers exist across various order statuses: While 'Shipped' has the most visible high-value outliers, other statuses like 'Disputed', 'In Process', and 'On Hold' also show some outlier transactions, indicating that high-value orders can end up in different stages of the order process.

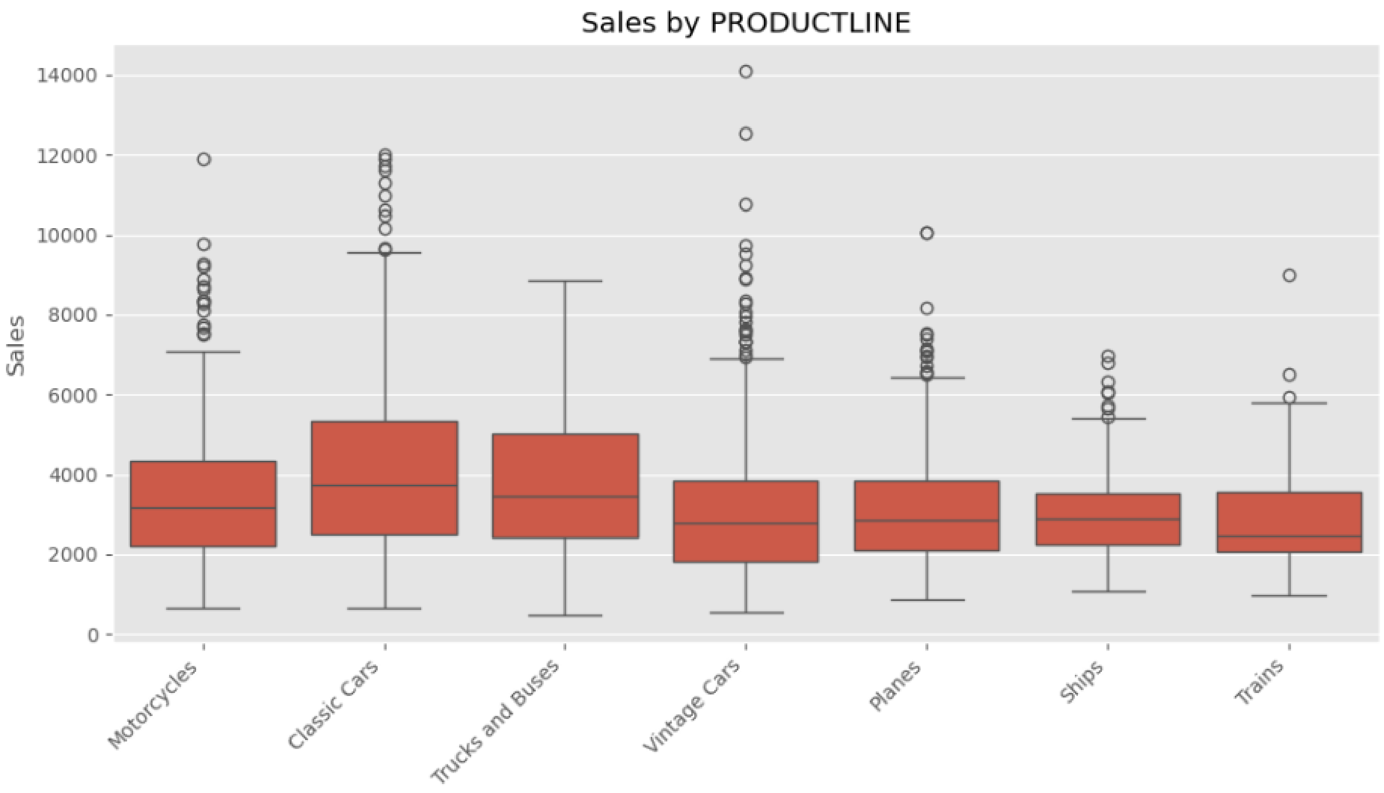


FIGURE 11 - BOXPLOT FOR SALES BY PRODUCTLINE

- **Classic Cars** exhibit the highest median sales: The median line within the box for 'Classic Cars' is the highest among all product lines, suggesting that, on average, orders for classic car parts tend to have the highest total sales value.
- **Classic Cars and Vintage Cars show the largest variability in sales**: The boxes for 'Classic Cars' and 'Vintage Cars' are the tallest, indicating the widest range of sales values within these product lines, including a significant number of high-value outliers.
- **Motorcycles, Planes, Ships, and Trains have lower median sales** and smaller interquartile ranges: These product lines generally show lower median sales values compared to 'Classic Cars', 'Vintage Cars', and 'Trucks and Buses'. Their shorter boxes also suggest less variability in the middle 50% of their sales data.
- **Trucks and Buses have a relatively high median sales value with moderate variability**: While not as high as 'Classic Cars', the median sales for 'Trucks and Buses' is still substantial, and the box size indicates a moderate spread in sales values.
- **Outliers are present across all product lines, indicating occasional high-value orders**: Every product line has some outliers above the upper whisker, signifying that even in product lines with generally lower sales, there are instances of significantly higher-value transactions.

- The distribution of sales differs significantly across product lines: The varying heights of the boxes, the positions of the median lines, and the number of outliers clearly demonstrate that the sales patterns and average transaction values differ considerably depending on the product category.

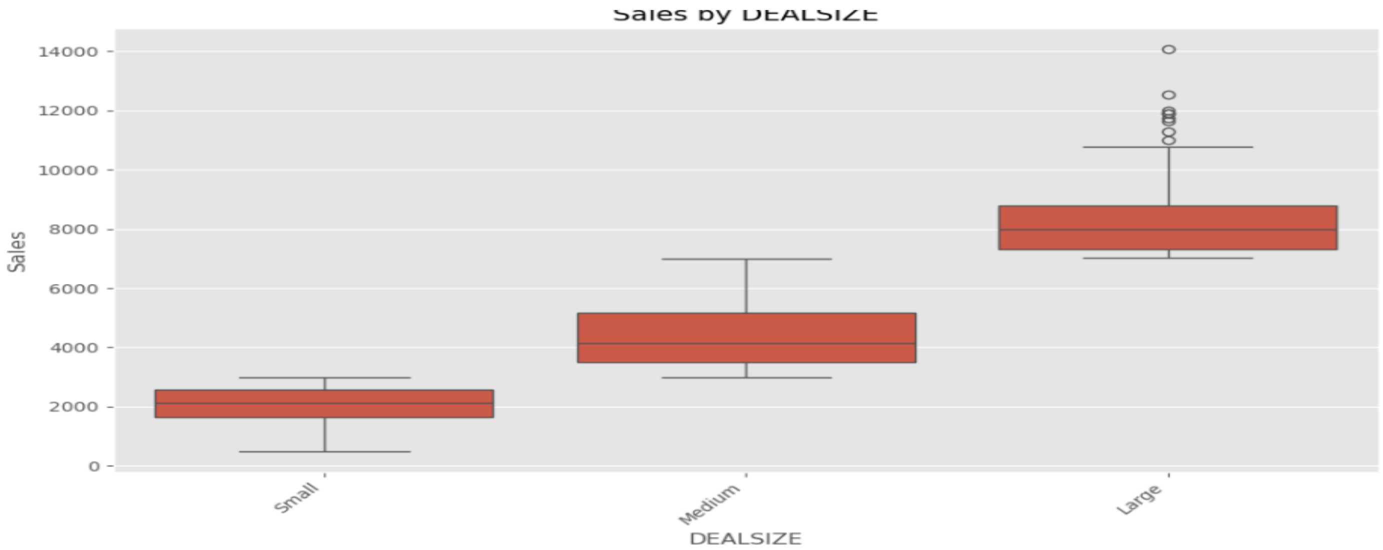


FIGURE 12 - BOXPLOT FOR SALES BY DEALSIZE

- Sales increase significantly with increasing deal size (Small < Medium < Large).

- Large deals exhibit the highest median sales and the widest range, including notable high-value outliers.

- Medium deals show a moderate increase in median sales and variability compared to Small deals.

- Small deals have the lowest median sales and the tightest interquartile range, indicating less sales value variation.

- The distinct separation of the boxes suggests deal size is a strong determinant of transaction value.

Correlation Matrix –

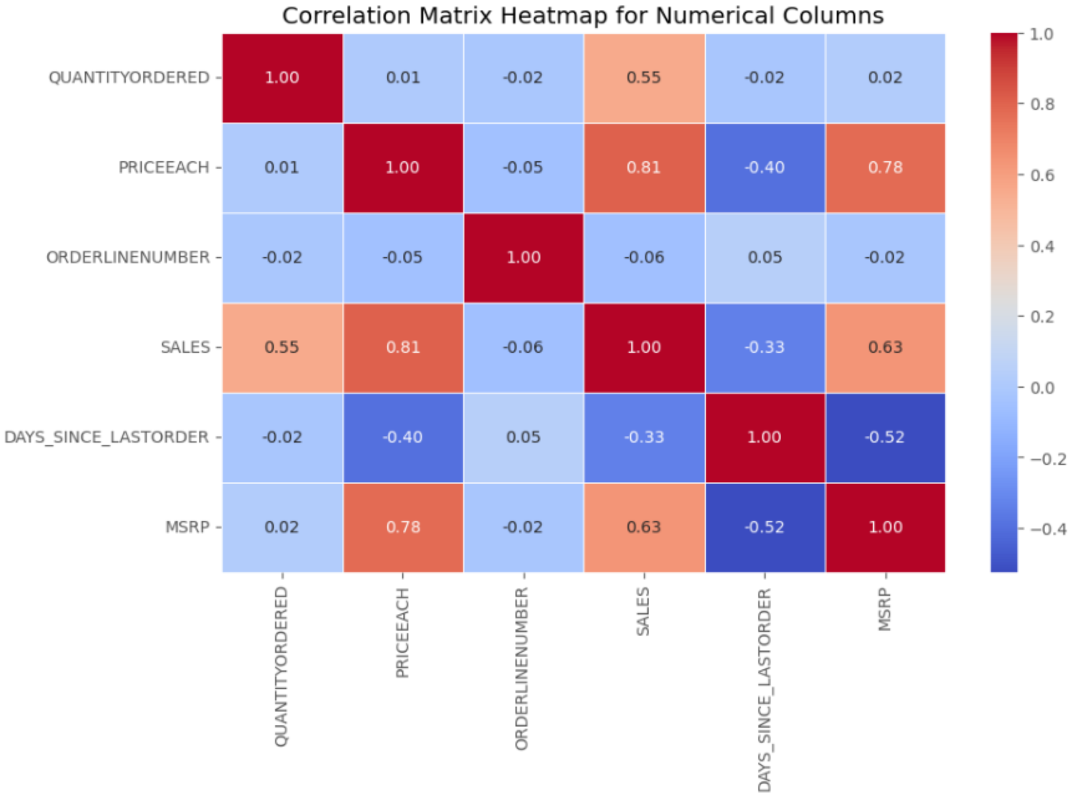


Figure 13 - Correlation matrix

- **Strong Positive Correlation between Price Each and Sales:** There is a strong positive correlation (0.81) between the price of each item and the total sales, indicating that higher priced items tend to contribute more to overall sales.
- **Moderate Positive Correlation between Quantity Ordered and Sales:** Quantity ordered shows a moderate positive correlation (0.55) with sales, suggesting that orders with more items generally result in higher sales figures.
- **Positive Correlation between Price Each and MSRP:** A notable positive correlation (0.78) exists between the price each and the Manufacturer's Suggested Retail Price (MSRP), implying that the selling price is often related to the MSRP.
- **Negative Correlation between Price Each and Days Since Last Order:** There is a moderate negative correlation (-0.40) between the price each and the days since the last order, suggesting that customers who ordered more recently might have purchased items with different price points compared to those with longer gaps between orders.
- **Negative Correlation between Days Since Last Order and MSRP:** A moderate negative correlation (-0.52) is observed between the days since the last order and the MSRP, indicating that customers with more recent orders might have purchased products with different MSRPs compared to those who haven't ordered in a while.
- **Weak Correlations for Order Line Number:** The 'ORDERLINENUMBER' shows very weak correlations with all other numerical variables, suggesting that the sequence of a product within an order has little linear relationship with these factors.

Scatter Plot –

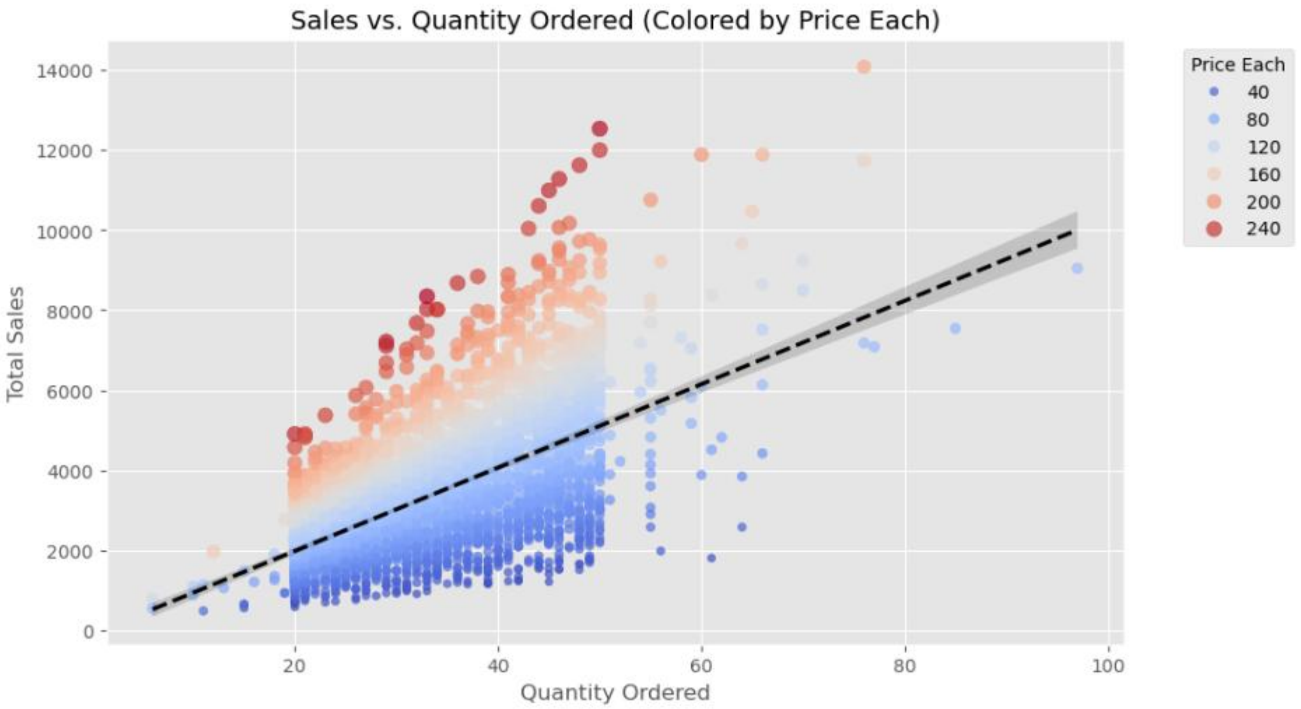


FIGURE 14 - SCATTERPLOT FOR SALES VS QUANTITY ORDERED

- **Positive Relationship between Quantity Ordered and Sales:** Generally, as the quantity of items ordered increases, the total sales tend to increase, as indicated by the upward-sloping trend line and the overall distribution of points.
- **Influence of Price Each on Sales:** For a given quantity ordered, the total sales vary significantly based on the price of each item. Points colored with warmer hues (representing higher price each) tend to be associated with higher total sales.
- **Lower Quantity Orders Dominated by Lower Priced Items:** Orders with smaller quantities (below approximately 40) are predominantly represented by cooler colors, suggesting that these orders often consist of lower-priced items.
- **Higher Sales Achieved Through Either High Quantity or High Price:** High total sales values can be achieved through either ordering a large quantity of items, even if they are lower priced, or by ordering a smaller quantity of significantly higher-priced items.
- **Spread Increases with Quantity:** As the quantity ordered increases, the spread of total sales also appears to widen. This suggests that for larger orders, the variation in the price of individual items has a more pronounced impact on the final sales amount.

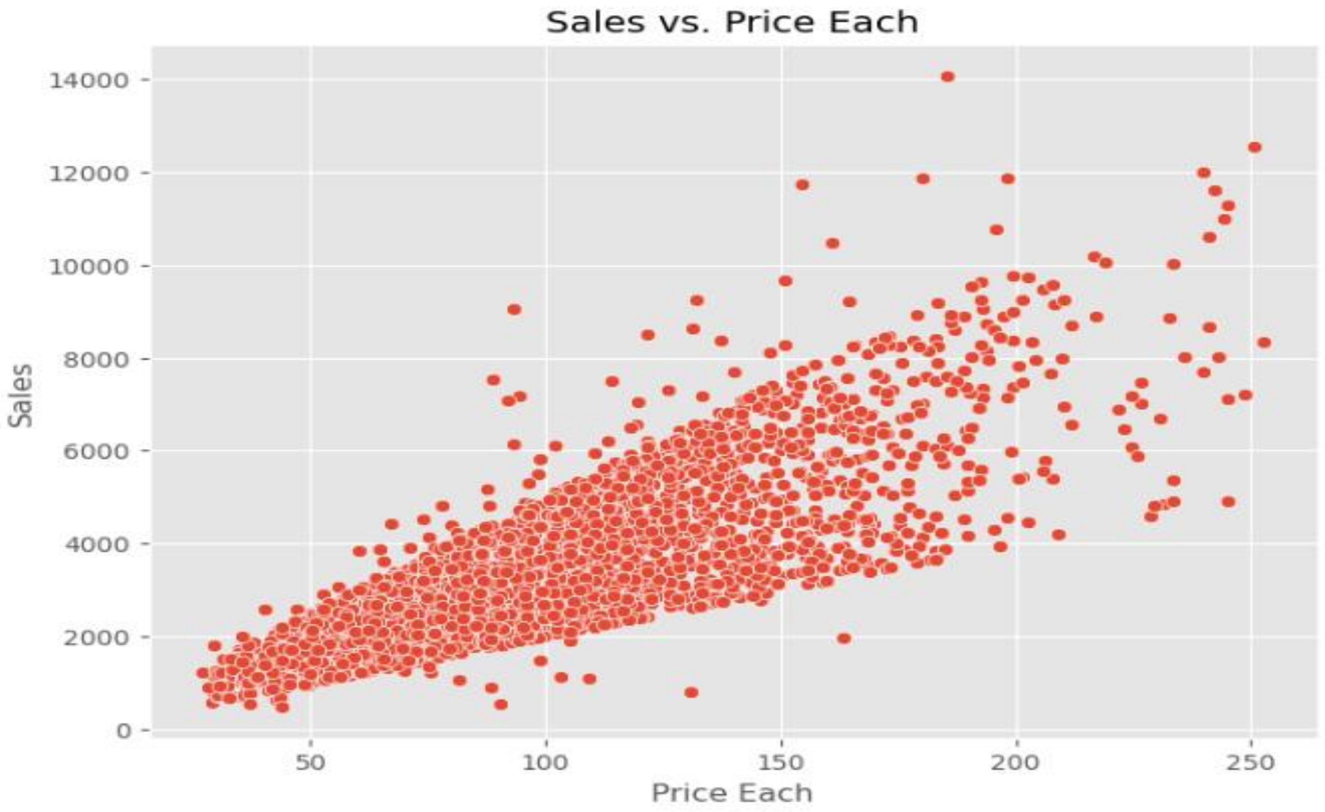


FIGURE 15 - SCATTER PLOT FOR SALES VS PRICE EACH

- There is a clear **positive trend** visible: as the Price Each of a product increases, the total Sales for that order tend to increase as well. This suggests that higher-priced items are significant contributors to the overall revenue.
- The relationship appears to be **non-linear**, possibly showing diminishing returns at higher price points. The density of points seems to decrease as the Price Each gets very high, and the increase in Sales for each unit increase in Price Each might become less pronounced.

- There is a **considerable spread** in Sales for any given Price Each, especially in the mid-range of prices. This indicates that factors other than just the individual price (like the quantity ordered) also play a significant role in determining the total Sales.
- A **cluster of data points** is noticeable at lower Price Each values (below 100), indicating a higher frequency of transactions involving these lower-priced items.

- There are **fewer data points** at very high Price Each values (above 200), suggesting that transactions involving these expensive items are less common.
- The plot highlights the **importance of pricing strategy**, as higher prices generally correlate with higher sales per transaction, but the volume of transactions at different price points also needs consideration.

Cross tab plot–

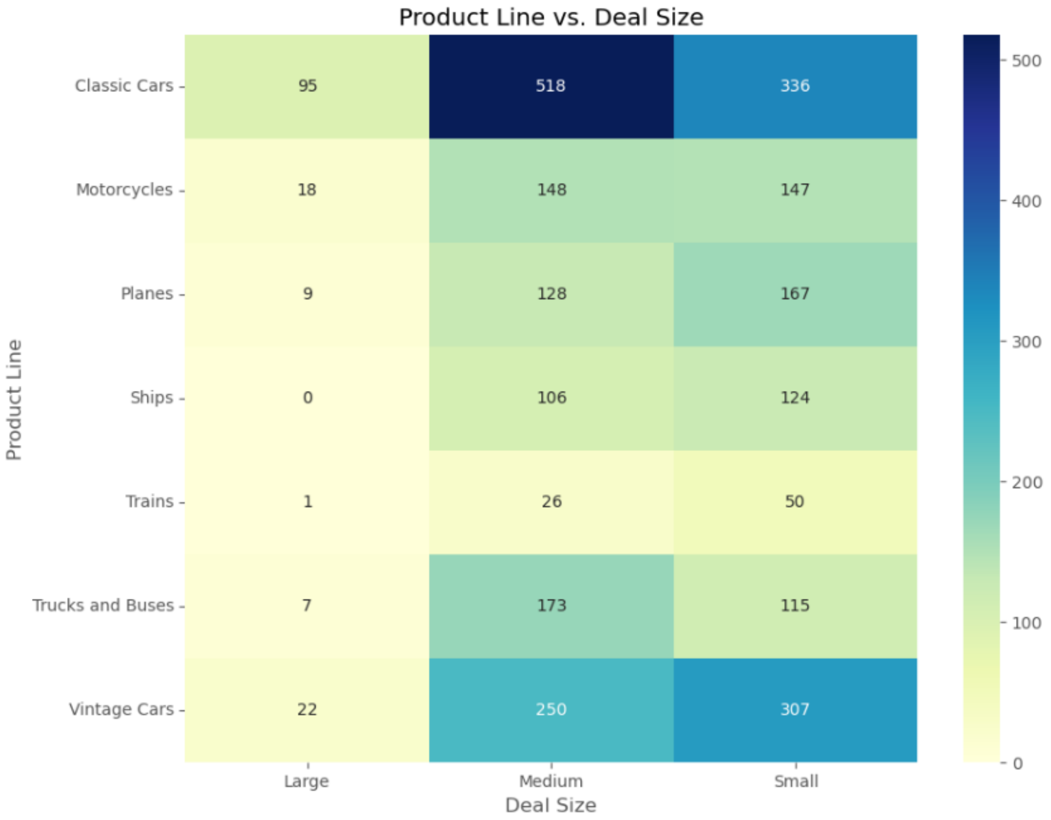


FIGURE 16 - CROSSTAB FOR PRODUCT LINE VS DEAL SIZE

- **Classic Cars dominate Medium and Small deal sizes:** The 'Classic Cars' product line has the highest number of transactions in both the Medium (518) and Small (336) deal size categories, indicating strong sales volume in these segments.
- **Motorcycles, Planes, and Vintage Cars show a preference for medium and small deals over large deals:** These product lines have significantly more transactions in the Medium and Small deal size categories compared to the Large deal size.
- **Large deals are relatively infrequent across most product lines:** With the exception of 'Classic Cars', the number of transactions classified as 'Large' is considerably lower for all other product lines. 'Ships' even have zero recorded large deals.
- **Trains and Trucks and Buses have a lower overall transaction volume:** The number of transactions across all deal sizes is notably lower for 'Trains' and 'Trucks and Buses' compared to other product lines, suggesting potentially niche markets or lower sales volume for these categories.

- **Deal size distribution varies by product line:** While Medium and Small deals are generally more common, the proportion of each deal size varies across different product lines. For instance, 'Vintage Cars' show a relatively higher number of Small deals compared to Medium deals than 'Classic Cars'.

Multivariate Analysis using Parallel Coordinates –

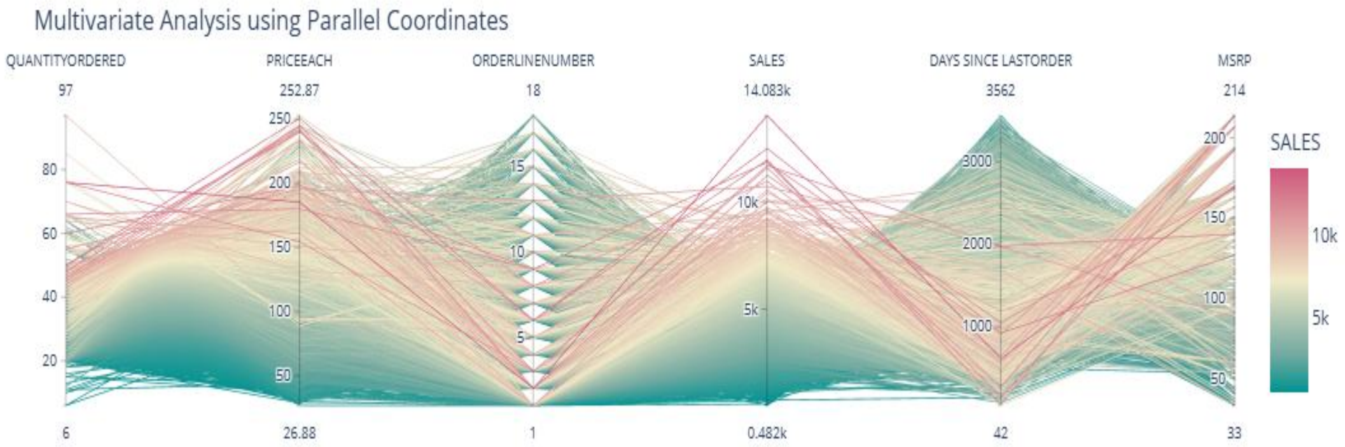


FIGURE 17 - MULTIVARIATE ANALYSIS USING PARALLEL COORDINATES

Observations -

- Higher 'PRICEEACH' and 'MSRP' values tend to align with orders having higher 'SALES'.

- 'QUANTITYORDERED' doesn't consistently dictate 'SALES'; high sales occur with both high and moderate quantities.
- 'DAYS_SINCE_LASTORDER' shows no clear pattern related to the 'SALES' amount.

- The position of an item in the order ('ORDERLINENUMBER') seems unrelated to the total 'SALES'.
- Lower 'PRICEEACH' and 'MSRP' are generally seen in orders with lower 'SALES'.

- 'SALES' are likely a result of interplay between 'PRICEEACH', 'MSRP', and 'QUANTITYORDERED'.

RFM Analysis

RFM analysis is a customer segmentation technique that uses past purchase behavior to divide customers into groups.

- **Recency**: How recently a customer made a purchase.
- **Frequency**: How often they purchase.
- **Monetary**: How much money they spend.

KNIME Final Workflow Overview -

1. Excel Reader
2. Date Conversion
3. GroupBy (Recency)
4. GroupBy (Frequency)
5. Groupby (Monetary)
6. Joiner (to form RFM table)
7. Binning/Rule Engine (Score R, F, M)
8. String Manipulation (RFM Score)
9. Rule Engine (Segment Label)
10. Visualizations + Top Customers

Also here in my analysis,

- R_Score - This will assign 5 to most recent buyers and 1 to oldest.

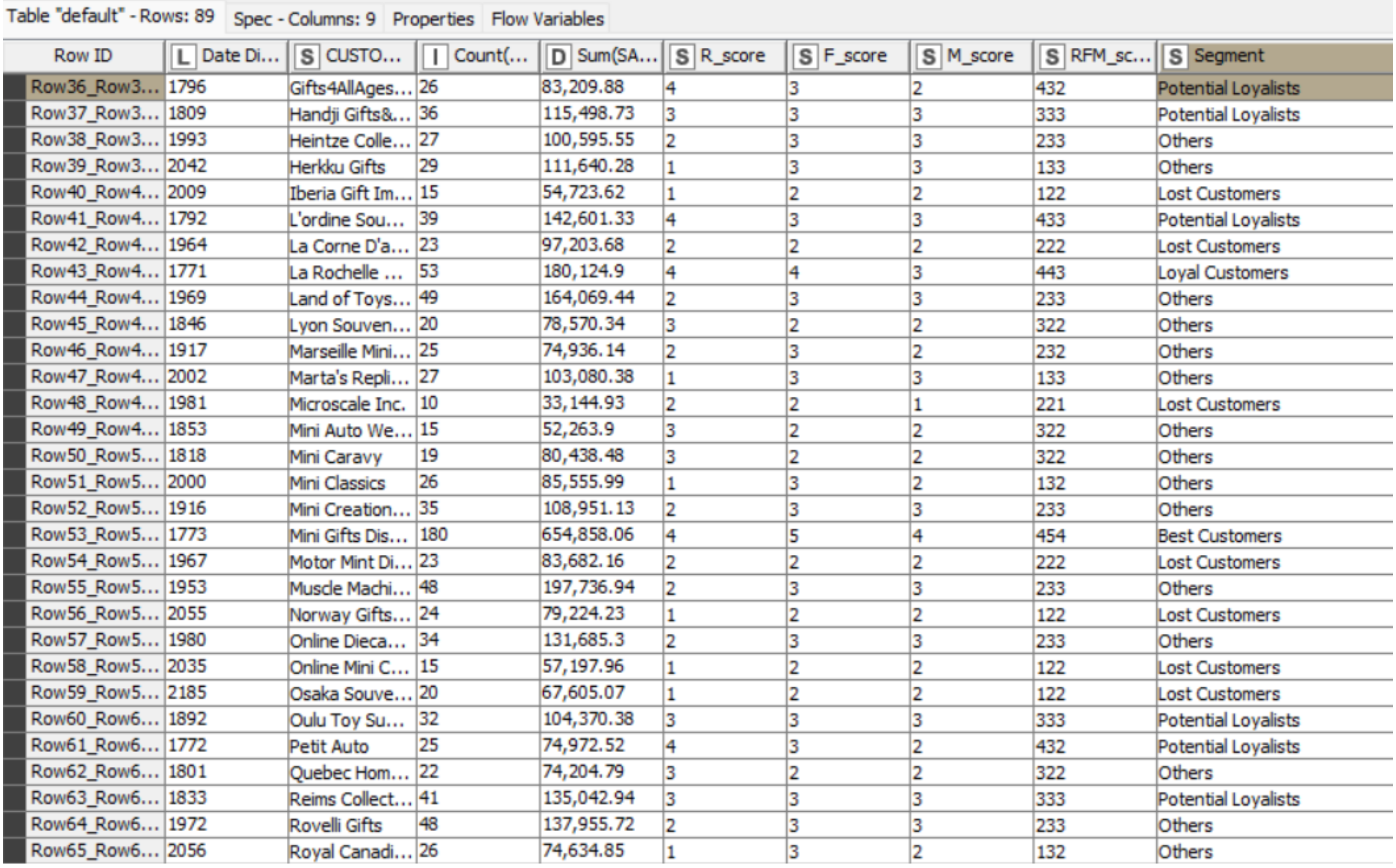
- F_Score - frequent customers get higher scores.

- M_Score - higher score means higher sales

For customer segments, I have classified as below –

Segment	**Marketing Action**
Best Customers	VIP loyalty program, early product launches
Loyal Customers	Regular discounts, membership rewards
Potential Loyalists	Nurture with onboarding emails and small perks
Churn Risk	Special win-back offers, feedback surveys
Lost Customers	Email campaign or discounts to re-engage

As per Knime, the final output table is –



Row ID	L Date Di…	S CUSTO…	I Count(…	D Sum(SA…	S R_score	S F_score	S M_score	S RFM_sc…	S Segment
Row36_Row3…	1796	Gifts4AllAges…	26	83,209.88	4	3	2	432	Potential Loyalists
Row37_Row3…	1809	Handji Gifts&…	36	115,498.73	3	3	3	333	Potential Loyalists
Row38_Row3…	1993	Heintze Colle…	27	100,595.55	2	3	3	233	Others
Row39_Row3…	2042	Herkku Gifts	29	111,640.28	1	3	3	133	Others
Row40_Row4…	2009	Iberia Gift Im…	15	54,723.62	1	2	2	122	Lost Customers
Row41_Row4…	1792	L'ordine Sou…	39	142,601.33	4	3	3	433	Potential Loyalists
Row42_Row4…	1964	La Corne D'a…	23	97,203.68	2	2	2	222	Lost Customers
Row43_Row4…	1771	La Rochelle …	53	180,124.9	4	4	3	443	Loyal Customers
Row44_Row4…	1969	Land of Toys…	49	164,069.44	2	3	3	233	Others
Row45_Row4…	1846	Lyon Souven…	20	78,570.34	3	2	2	322	Others
Row46_Row4…	1917	Marseille Mini…	25	74,936.14	2	3	2	232	Others
Row47_Row4…	2002	Marta's Repli…	27	103,080.38	1	3	3	133	Others
Row48_Row4…	1981	Microscale Inc.	10	33,144.93	2	2	1	221	Lost Customers
Row49_Row4…	1853	Mini Auto We…	15	52,263.9	3	2	2	322	Others
Row50_Row5…	1818	Mini Caravy	19	80,438.48	3	2	2	322	Others
Row51_Row5…	2000	Mini Classics	26	85,555.99	1	3	2	132	Others
Row52_Row5…	1916	Mini Creation…	35	108,951.13	2	3	3	233	Others
Row53_Row5…	1773	Mini Gifts Dis…	180	654,858.06	4	5	4	454	Best Customers
Row54_Row5…	1967	Motor Mint Di…	23	83,682.16	2	2	2	222	Lost Customers
Row55_Row5…	1953	Muscle Machi…	48	197,736.94	2	3	3	233	Others
Row56_Row5…	2055	Norway Gifts…	24	79,224.23	1	2	2	122	Lost Customers
Row57_Row5…	1980	Online Dieca…	34	131,685.3	2	3	3	233	Others
Row58_Row5…	2035	Online Mini C…	15	57,197.96	1	2	2	122	Lost Customers
Row59_Row5…	2185	Osaka Souve…	20	67,605.07	1	2	2	122	Lost Customers
Row60_Row6…	1892	Oulu Toy Su…	32	104,370.38	3	3	3	333	Potential Loyalists
Row61_Row6…	1772	Petit Auto	25	74,972.52	4	3	2	432	Potential Loyalists
Row62_Row6…	1801	Quebec Hom…	22	74,204.79	3	2	2	322	Others
Row63_Row6…	1833	Reims Collect…	41	135,042.94	3	3	3	333	Potential Loyalists
Row64_Row6…	1972	Rovelli Gifts	48	137,955.72	2	3	3	233	Others
Row65_Row6…	2056	Royal Canadi…	26	74,634.85	1	3	2	132	Others

TABLE 4 – RFM SCORE AND FINAL SEGMENT CLASSIFICATION

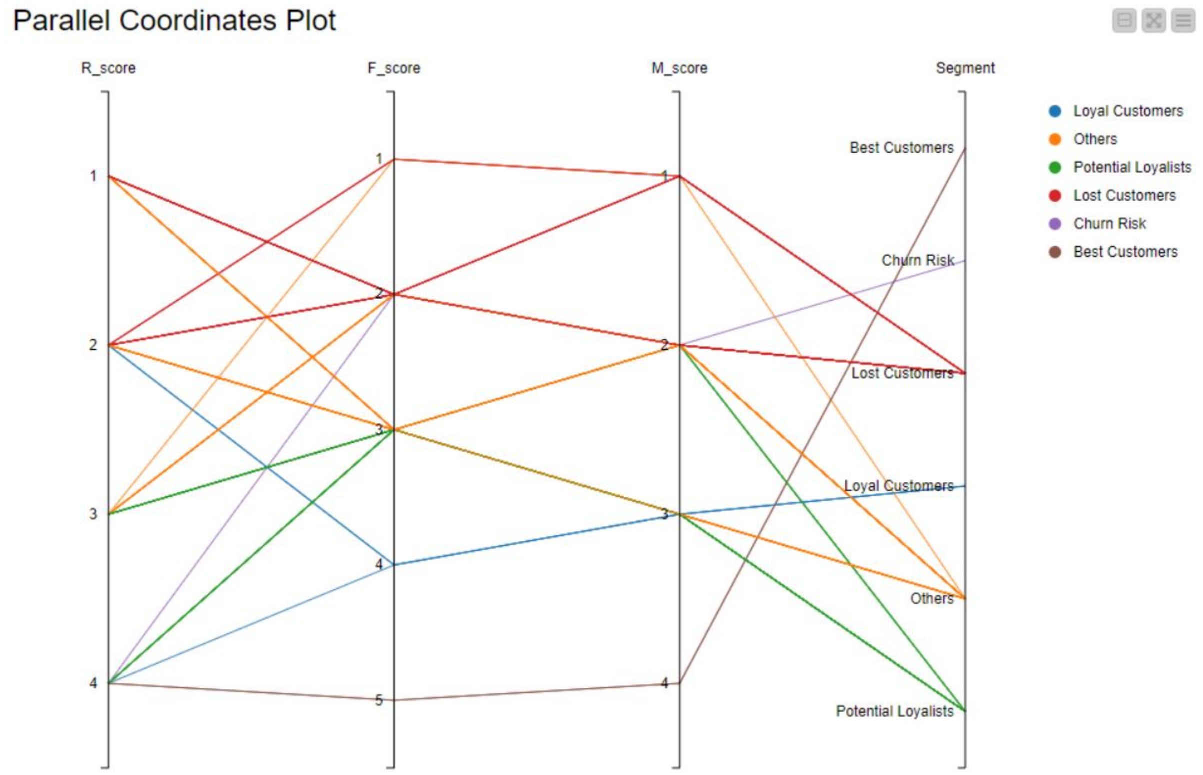


Figure 18 - Parallel Coordinates Plot using KNIME - Segments

Customer Segmentation Visualization: The plot effectively visualizes customer segments (Loyal Customers, Others, etc.) based on their R_score, F_score, and M_score.

- **"Best Customers" Profile:** The "Best Customers" segment (blue lines) generally exhibits high R_score (around 1), high F_score (around 1), and high M_score (around 1), aligning with the score definitions.
- **"Loyal Customers" Profile:** "Loyal Customers" (orange lines) show good Recency (around 2), moderate to high Frequency (2-3), and moderate Monetary value (2-3).
- **"Potential Loyalists" Traits:** "Potential Loyalists" (green lines) are recent buyers (high R_score around 1-2) with lower Frequency (around 4-5) but varying Monetary value.
- **"Lost Customers" Characteristics:** "Lost Customers" (red lines) are characterized by low Recency (high R_score around 1 but the scale is reversed, so actually low recency), lower Frequency (around 2-3), and varying Monetary value.
- **"Churn Risk" Indicators:** The "Churn Risk" segment (purple lines) shows low Recency (high R_score), relatively low Frequency (around 3-4), and moderate Monetary value.
- **Inter-Segment Differences:** The plot clearly distinguishes the behavioral patterns of different customer segments across the RFM dimensions, allowing for targeted marketing strategies.

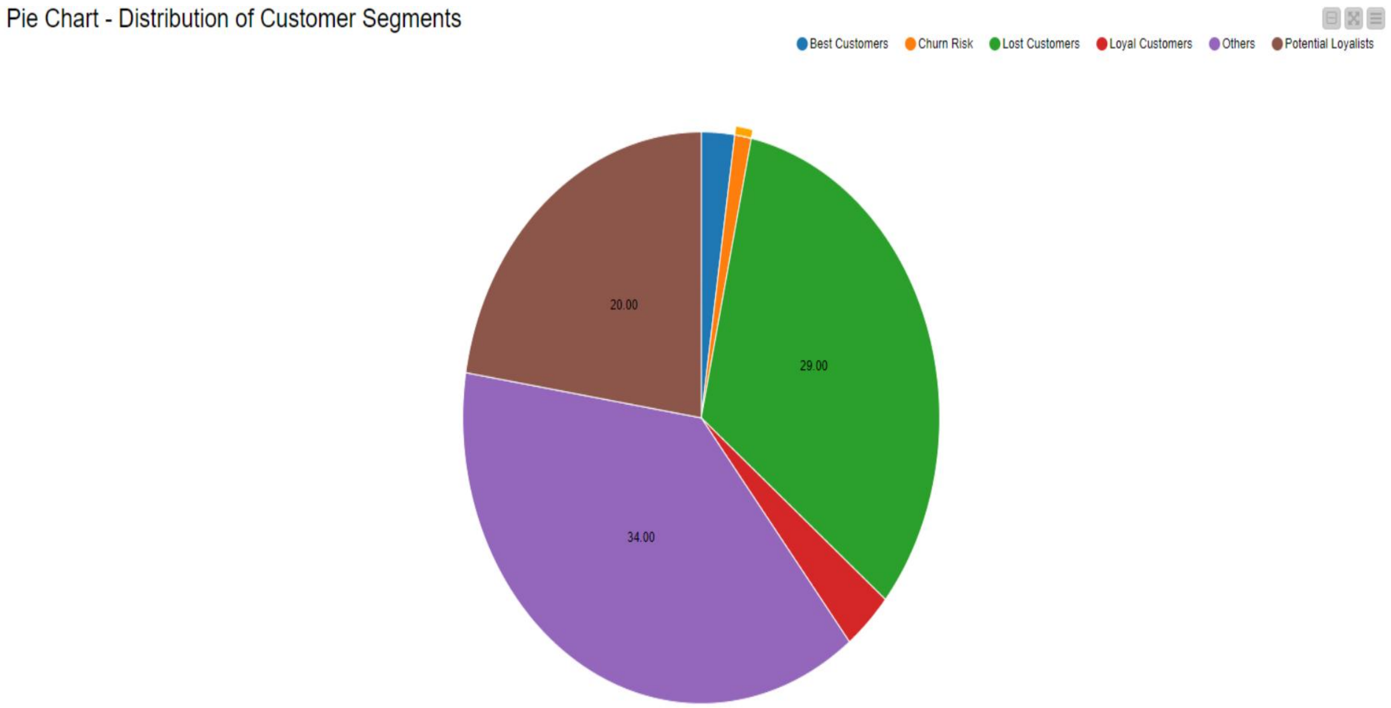


Figure 19 - Pie chart - Customer Distribution

- **Dominant Segments:** The purple segment represents the largest portion of customers at 34%, followed by the green segment at 29%, indicating these are the most prevalent customer groups.
- **Significant "Others" Category:** The brown segment, representing "Others," constitutes a substantial 20% of the customer base, suggesting a need to understand this heterogeneous group further.
- **Small "Best Customers" Proportion:** The blue segment, representing "Best Customers" (count 2), makes up a very small fraction of the total customer distribution.
- **Minimal "Churn Risk":** The orange segment, representing "Churn Risk" (count 1), is the smallest segment, indicating a low current proportion of customers identified as being at high risk of churning.
- **Relatively Low "Lost Customers":** The red segment, representing "Lost Customers" (count 3), also constitutes a small portion of the overall customer base.
- **Potential for Targeted Strategies:** The varying sizes of the segments highlight opportunities for tailored marketing and engagement strategies focused on the characteristics of each group.
- **Focus on Largest Segments:** The business might prioritize understanding and engaging the "Others" and the 34% and 29% segments due to their significant size.

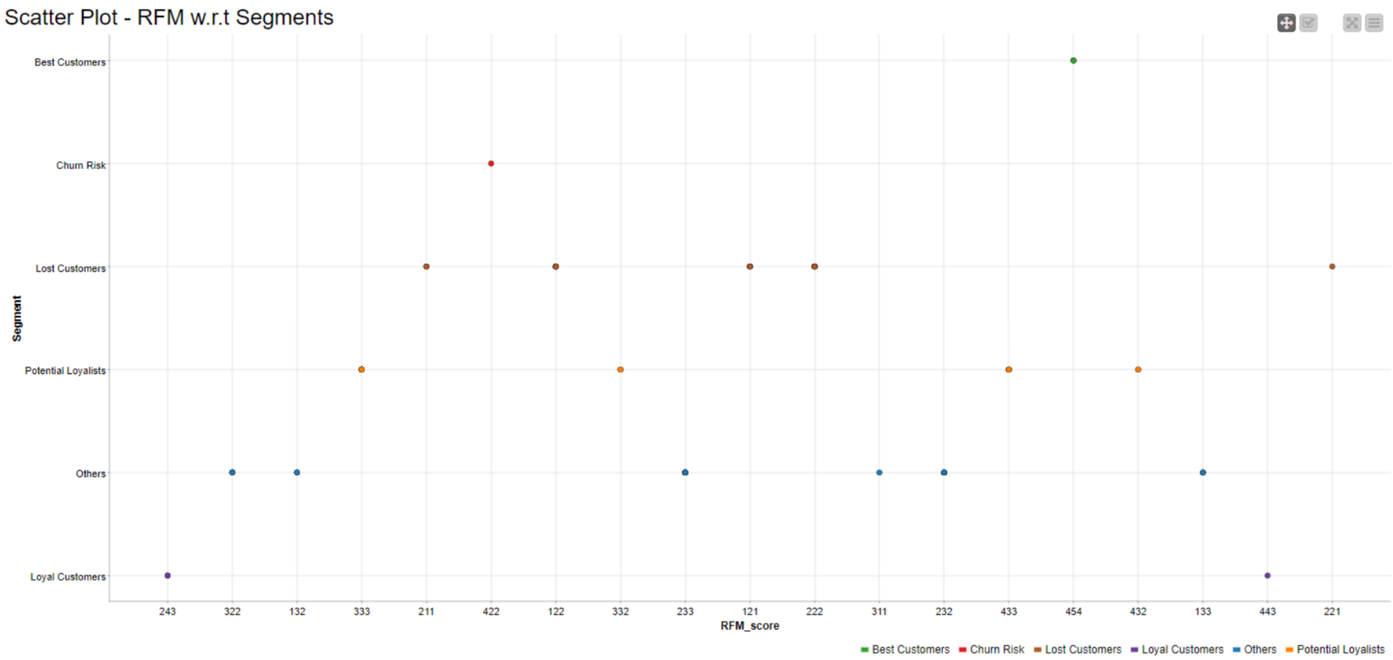


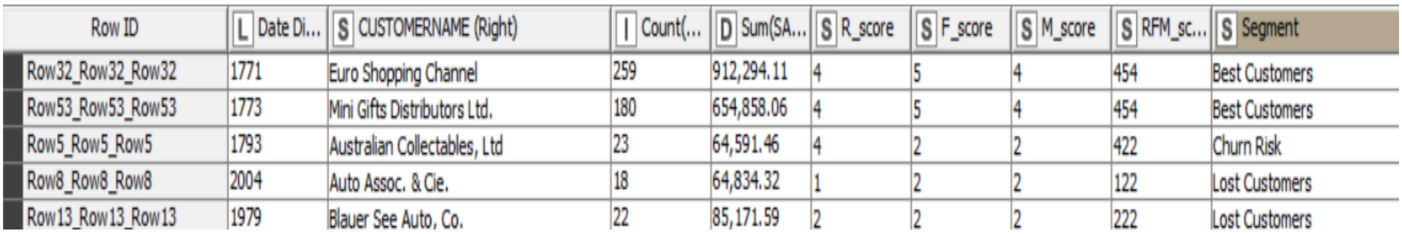
Figure 20 - Scatter Plot - RFM vs Segments

- **Segment Distribution by RFM Score:** The plot displays how different customer segments are distributed along the x-axis representing the RFM score. Each dot represents a customer, and its color indicates its assigned segment.
- **Vertical Clustering of Segments:** Notice that each customer segment tends to be grouped within a specific vertical range along the y-axis. This implies that the RFM segmentation process has resulted in distinct score ranges associated with each segment.
- **"Best Customers" at the High End:** The green dots representing "Best Customers" are primarily located towards the higher values of the RFM score on the x-axis, which aligns with the expectation that the best customers would have higher recency, frequency, and monetary value scores.
- **"Loyal Customers" at the Low End:** Conversely, the blue dots for "Loyal Customers" are concentrated at the lower end of the RFM score spectrum, suggesting they might have lower overall RFM scores compared to the "Best Customers."
- **Overlapping Score Ranges:** While there's a general separation, some segments show overlap in their RFM score ranges. For instance, you might see "Potential Loyalists" and "Others" occupying similar RFM score values, indicating that the segmentation considers more than just a single aggregated RFM score.
- **Segmentation Logic:** The plot suggests that the RFM segmentation logic likely involves thresholds or rules applied to the individual Recency, Frequency, and Monetary scores

(which are combined to form the RFM score on the x-axis) to define the boundaries of each segment.

- **Visual Assessment of Segmentation:** This visualization provides a quick way to assess how well the RFM segmentation has separated different customer groups based on their calculated RFM scores. Clear separation suggests a more effective segmentation strategy.

From our analysis in KNIME, our top 5 customers are -



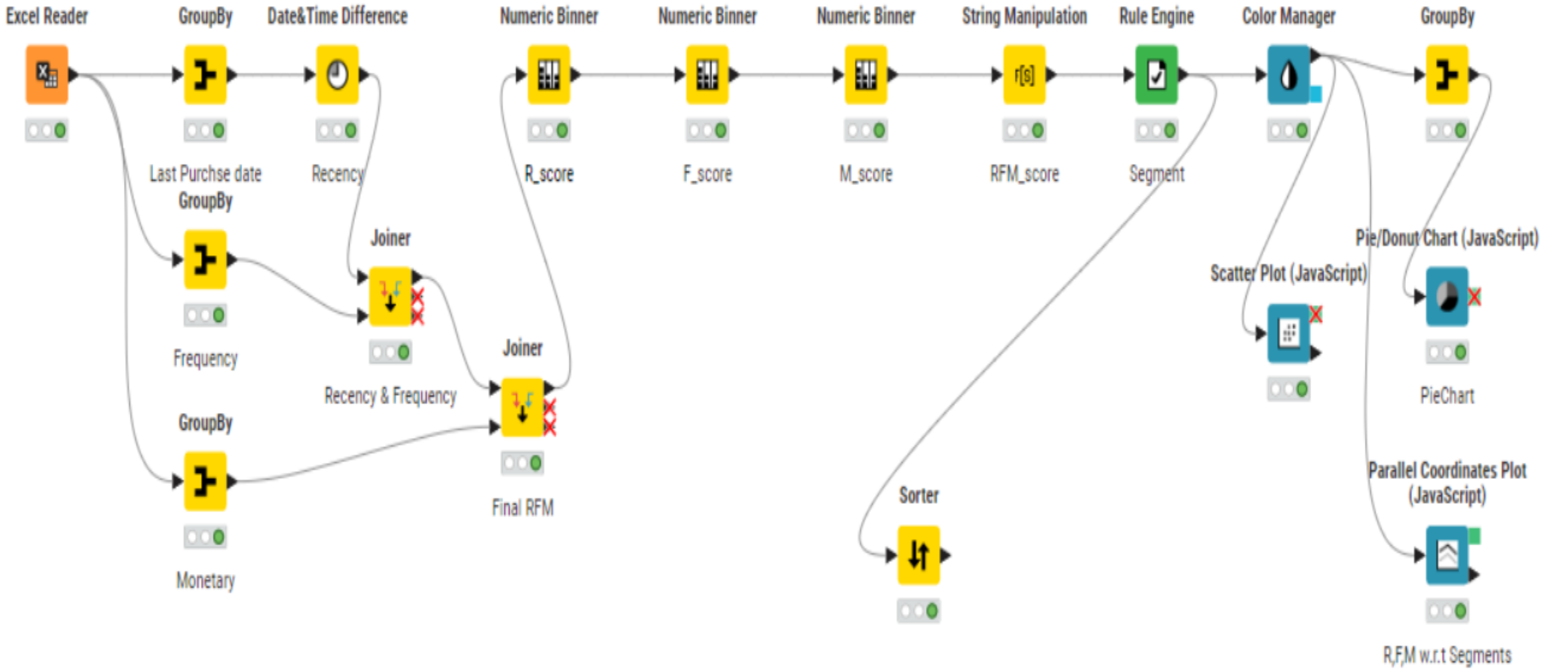
Row ID	L Date Di...	S CUSTOMERNAME (Right)	I Count(...	D Sum(SA...	S R_score	S F_score	S M_score	S RFM_sc...	S Segment
Row32_Row32_Row32	1771	Euro Shopping Channel	259	912,294.11	4	5	4	454	Best Customers
Row53_Row53_Row53	1773	Mini Gifts Distributors Ltd.	180	654,858.06	4	5	4	454	Best Customers
Row5_Row5_Row5	1793	Australian Collectables, Ltd	23	64,591.46	4	2	2	422	Churn Risk
Row8_Row8_Row8	2004	Auto Assoc. & Cie.	18	64,834.32	1	2	2	122	Lost Customers
Row13_Row13_Row13	1979	Blauer See Auto, Co.	22	85,171.59	2	2	2	222	Lost Customers

TABLE 5 – TOP 5 CUSTOMERS

Customer Name	Segment	RFM Score	Justification
Euro Shopping Channel	Best Customers	454	Very recent purchases (R=4), highest frequency (F=5), strong monetary value (M=4); this customer buys frequently and spends a lot, indicating high loyalty and value.
Mini Gifts Distributors Ltd.	Best Customers	454	Similar behavior to Euro Shopping Channel. High recency, highest frequency, and excellent spending – one of the most engaged customers.
Australian Collectables, Ltd.	Churn Risk	422	Previously strong (R=4, F=2, M=2) but dropping in frequency and value. Recency is still decent, so a well-timed reactivation campaign could prevent churn.
Auto Assoc. & Cie.	Lost Customers	122	Low recency, frequency, and monetary value. This customer has likely stopped purchasing. Consider surveying or offering incentives.
Blauer See Auto, Co.	Lost Customers	222	Slightly better than Auto Assoc. but still in “Lost” segment. Low engagement and spending. Could be targeted with a win-back campaign.

TABLE 6 – TOP 5 CUSTOMERS REASON

KNIME Workflow Diagram –



Inferences and Recommendations

RFM-Based Customer Categories:

1. Best Customers

Examples: Euro Shopping Channel, Mini Gifts Distributors Ltd.

Traits:

- High Recency (R=4) – recent purchasers.
- High Frequency (F=5) – buy often.
- High Monetary (M=4) – spend a lot.

Behaviour: Loyal, high-spending customers who frequently purchase and are likely to respond positively to personalized campaigns.

Strategy:

- Exclusive early-bird offers.
- Loyalty rewards and referral programs.
- Personalized recommendations based on purchase history.

2. Customers on the Verge of Churning (Churn Risk)

Examples - Australian Collectables, Ltd.

Traits:

- Recency is fair (R=4), but
- Low Frequency and Monetary (F=2, M=2).

Behaviour: They purchased recently but not often and don't spend much.

Strategy:

- Send re-engagement emails or limited-time offers.
- Incentivize with bundle discounts or free delivery.
- Ask for feedback to understand their drop-off reason.

3. Lost Customers

Examples: Auto Assoc. & Cie., Blauer See Auto, Co.

Traits:

- Low R, F, M scores (e.g., 122, 222).

Behaviour: Haven't purchased in a long time, bought infrequently, and low spenders.

Strategy:

- Send a win-back campaign (e.g., "We miss you" email).
- Offer a strong discount voucher or limited-time sale.
- Target with surveys to understand churn causes.

4. Loyal Customers

Traits:

- High Frequency (F ≥ 4), with stable Recency and good Monetary values.

Behaviour: Consistent, repeat buyers.

Strategy:

- Maintain strong relationships.
- Upsell or cross-sell based on purchase history.
- them with exclusive previews or offers.

Key Observations from EDA and RFM:

1. **Sales vs Quantity Ordered (colored by Price Each)**:
   - High-priced items generate higher revenue but may sell less often.
   - Mid-range priced items dominate in quantity and contribute significantly to total sales.
2. **Sales vs Price Each**:
   - Strong positive correlation between price and revenue up to a threshold (~200).
   - After that, diminishing returns likely appear.

3. **Monthly Sales Trend by Deal Size**:
    - **Medium deals** dominate sales over time.
    - Peaks in April-May signal strong seasonal trends or campaigns.
4. **Product Line vs Deal Size (Heatmap)**:
    - **Classic Cars** and **Vintage Cars** dominate across deal sizes—key revenue drivers.
    - Some product lines like *Trains* and *Ships* have low traction.
5. **Correlation Heatmap**:
    - Sales are positively correlated with both **Quantity Ordered** and **Price Each**.
    - Important for bundling and pricing strategy

- A **small subset of customers** (Best + Loyal) drive the **majority of revenue**.
- There's a noticeable **churn risk segment** that can be saved with proactive efforts.
- **Lost customers** may no longer be profitable unless reactivation cost is low.
- **Sales are product and season-dependent**, especially medium/large deal sizes.
- Strong opportunity lies in **price-based product bundling**.

Key Recommendations:

- ➢ **Retain High-Value Customers:**
    - Focus on **hyper-personalized** experiences for Best & Loyal segments.
    - Set up **tiered loyalty programs** and **VIP early access** benefits.
- ➢ **Reactivate Churn Risk:**
    - Offer **targeted coupons** or **limited-time discounts**.
    - Use recent product trends to recommend high-appeal bundles.
- ➢ **Re-engage Lost Customers:**
    - Use emotional campaigns (“We miss you!”) with exclusive deals.
    - Evaluate cost-effectiveness of recovery vs acquiring new customers.

- **Optimize Product Mix:**
  - Push mid-priced items with higher volume.
  - Focus promotions around **Classic and Vintage Cars** lines.

- **Leverage Seasonality:**
  - Plan **promotions around April-May** where historical peaks exist.
  - Use trend patterns for inventory and marketing alignment.