##### Description

Dear Participants,

The project has two parts of 45 marks each. Please go through the problem statements and datasets attached below.

#### **Part A: 45 Marks**

**Problem Statement:**

An automobile parts manufacturing company has collected data on transactions for 3 years. They do not have any in-house data science team, thus they have hired you as their consultant. Your job is to use your data science skills to find the underlying buying patterns of the customers, provide the company with suitable insights about their customers, and recommend customized marketing strategies for different segments of customers.

**Dataset:**

Auto Sales Data: [Sales\_Data.xlsx](https://olympus.mygreatlearning.com/courses/85968/files/8669170/download?verifier=ldsABCtOj0DURL92ZZPmrbqcQjhBuiitGTrG1stf&wrap=1)

Please note that the data dictionary is available in Sheet 2 of the provided dataset.

#### **Part B: 45 Marks**

**Problem Statement:**

A grocery store shared the transactional data with you. Your job is to conduct a thorough analysis of Point of Sale (POS) data, identify the most commonly occurring sets of items in the customer orders, and provide recommendations through which a grocery store can increase its revenue by popular combo offers & discounts for customers.

**Dataset:**

Grocery Store Data: [dataset\_group.csv](https://olympus.mygreatlearning.com/courses/85968/files/8669171/download?verifier=ei5fUYVTCpaful1Q4whJZvTS2vYOTyDI4Y5gkd9k&wrap=1)

#### **Submission Guidelines**

Please go through the following guidelines before you proceed to make your submission for the MRA Project.

* There are two parts to the submission for the project
  + **Business reports**: These files will be the primary criteria for evaluation
  + **Supporting files**: These files will be used to validate the content of the business reports
  + The**EDA** section of the project can be done using **one or more of the following tools - Python, Tableau, KNIME**
  + In case multiple tools are used, all the files must be submitted for optimal evaluation
  + The**RFM and Market Basket Analysis** sections of the project **must be done in KNIME**
  + Any other tool will NOT be accepted for these sections
  + The file formats for the submissions of different sections of the project are mentioned below:
  + Business Report - pptx, pdf (presentation)
  + EDA
    - Python - .ipynb (Python notebook)
    - KNIME - .knwf (KNIME workflow)
    - Tableau - .docx, .pdf (the file should contain the link to the Tableau Public dashboard)
  + RFM - .knwf (KNIME workflow)
  + Market Basket Analysis - .knwf (KNIME workflow)
  + Please note that in case the business reports are not submitted, the assessment will be graded zero
  + In case the business report for one part of the project is submitted, the evaluation will be done for that part only
* As the business reports are the primary criteria for evaluation, kindly make sure that all the required information asked in the rubric are included in the business reports.
* The **requirements for the business report** are as follows:
  + Minimum 14 slides each for Part A and Part B, which includes
    - minimum 8 slides on EDA and inferences from EDA
    - minimum 2 slides on RFM / Market Basket Analysis
    - minimum 4 slides on inferences from RFM / Market Basket Analysis
    - Note: 1.  The plots used in the PPT must be readable, else marks will be deducted for those plots. 2. KNIME Workflow Image is a must that needs to be put in the PPT. Also, make sure that you are putting the correct image, else marks will be deducted.
* **Submitting the supporting files is mandatory** as without them, the business reports cannot be validated
  + In case one or more supporting files are not submitted, 50% of the grades associated with the corresponding sections of the project will be deducted
* Please ensure timely submission as a post-deadline assignment will not be accepted.

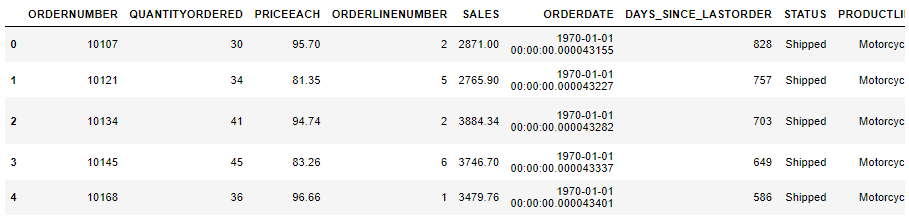
Thanks

Program Office

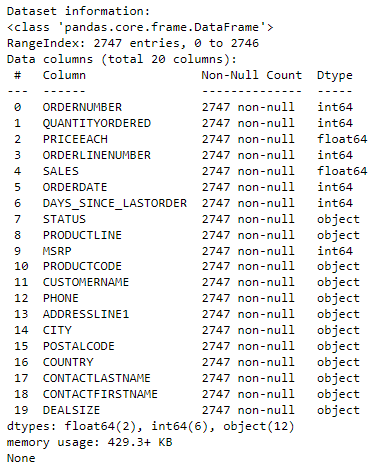
**Scoring guide (Rubric) - DSBA MRA Project**

# Criteria

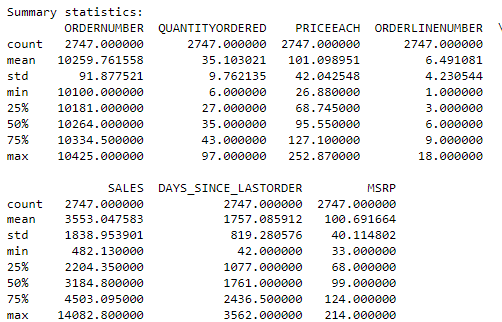
# Agenda & Executive Summary of the data -> Contents of the ppt -> Problem statement -> About Data (Info, Shape, Summary Stats, your assumptions about data



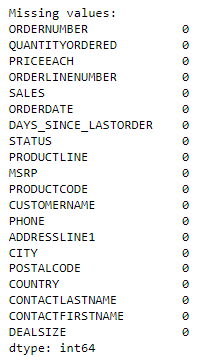
Snippet 1. 1 Head of the dataset



Snippet 1. 2 Information of the dataset



Snippet 1. 3 Statistical Summary of the Numerical Attributes



Snippet 1. 4 Missing Values

PART A: Exploratory Analysis and Inferences -> Univariate, Bivariate, and multivariate analysis using data visualization (Weekly, Monthly, Quarterly, Yearly Trends in Sales and Sales Across different Categories of different features in the given data) -> Summarise the inferences

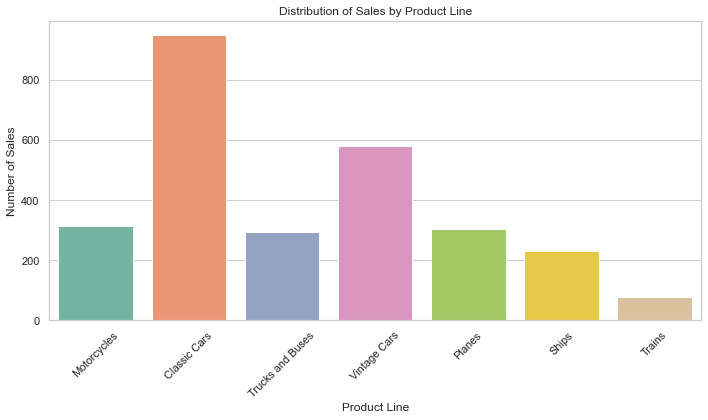


Figure 1. 1 Distribution of Sales by Product Line

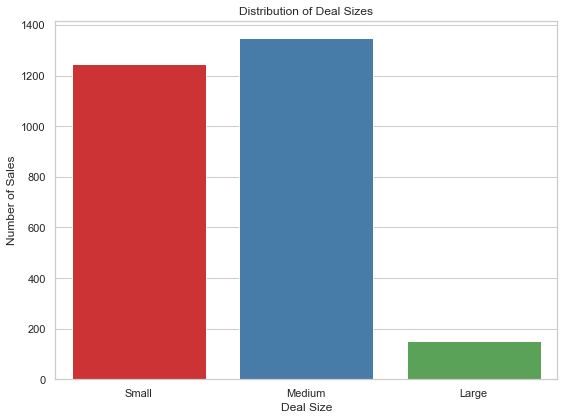


Figure 1. 2 Distribution of Deal Sizes

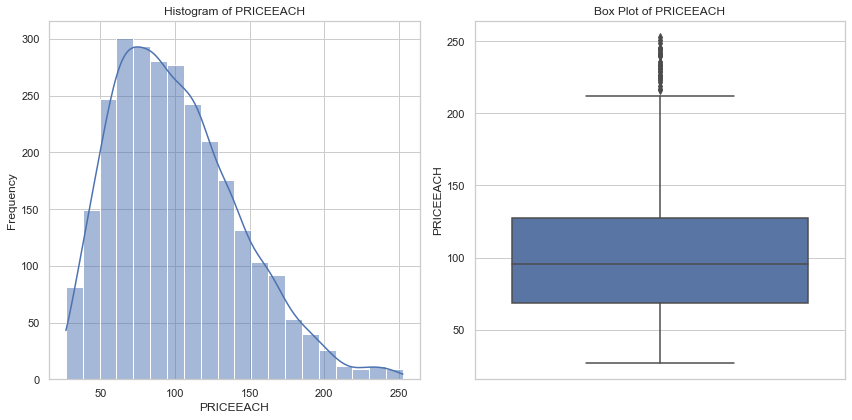


Figure 1. 3 Univariate Analysis of PriceEach

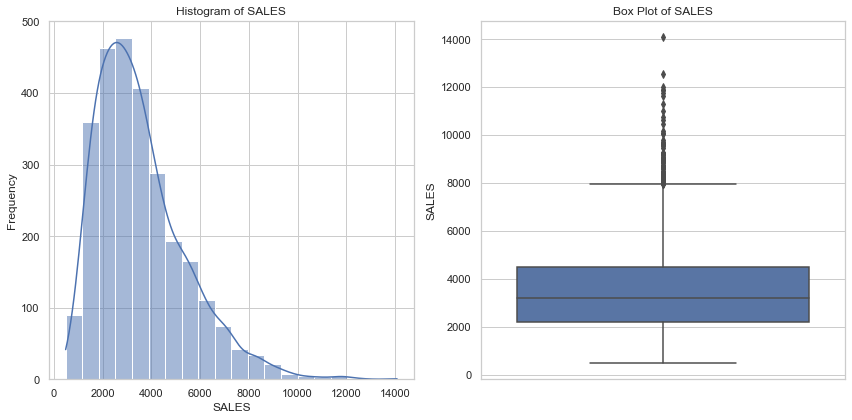


Figure 1. 4 Univariate Analysis of Sales

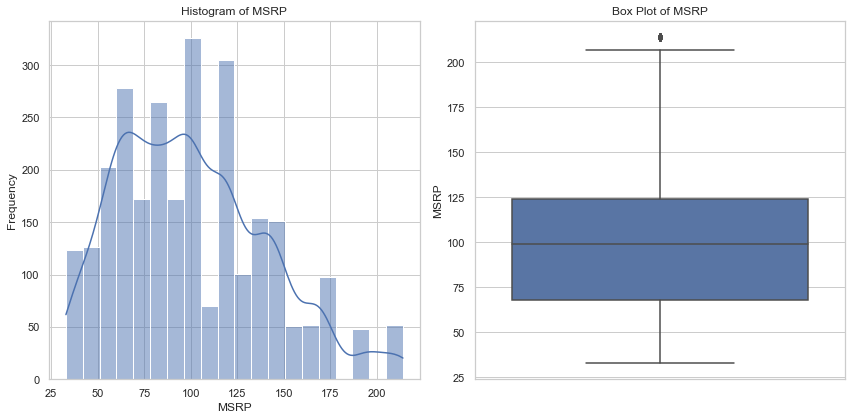


Figure 1. 5 Univariate Analysis of MSRP

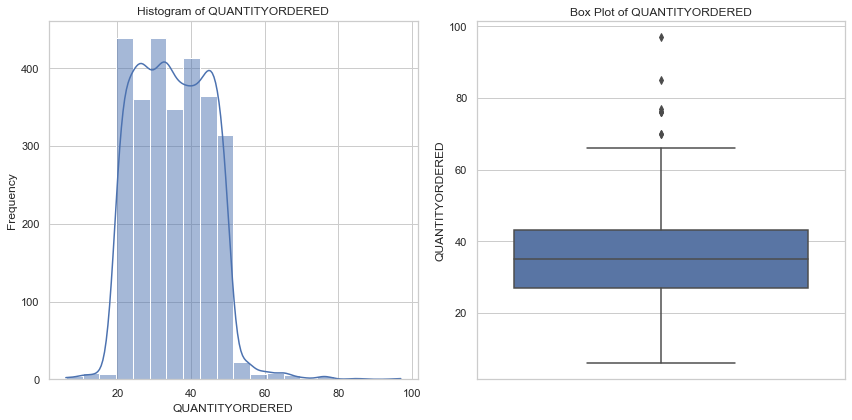


Figure 1. 6 Univariate Analysis of Quantity Order

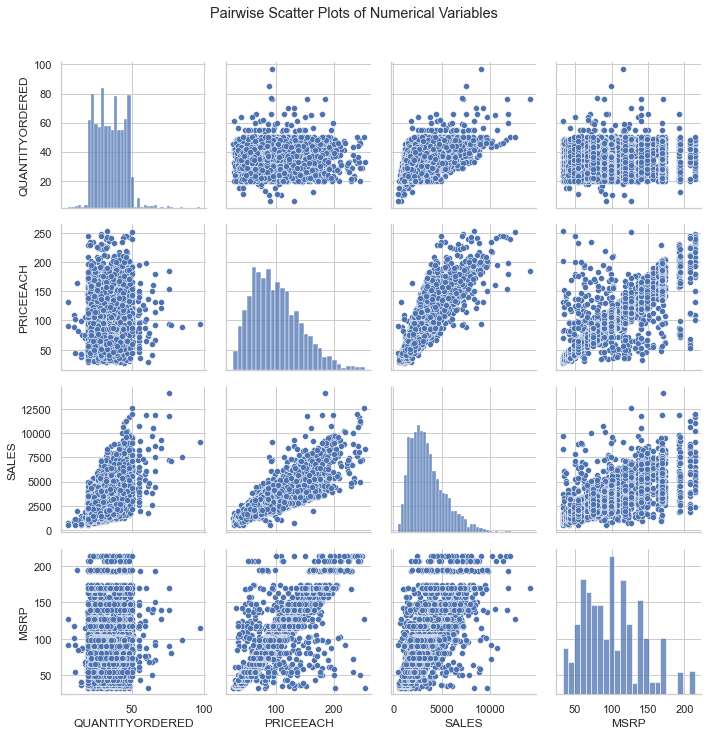


Figure 1. 7 Pair wise Scatter Plots of Numerical Variables

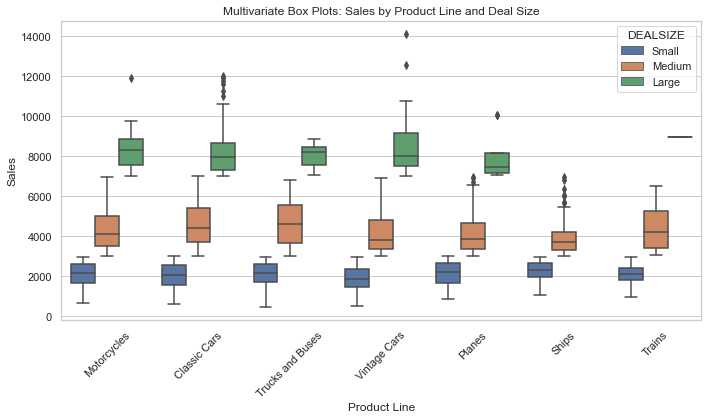


Figure 1. 8 Multivariate Analysis: Sales by Product Line and Deal Size

# Customer Segmentation using RFM analysis (4 segments) -> What is RFM? -> What all parameters used and assumptions made? -> Showcase the KNIME workflow image -> What results are there in the output table head?

##### Customer Segmentation using RFM Analysis:

RFM stands for Recency, Frequency, and Monetary. It's a technique used to segment customers based on their purchasing behaviour. Each letter of RFM represents a key parameter:

* Recency (R): How recently a customer made a purchase.
* Frequency (F): How often a customer makes a purchase.
* Monetary (M): How much money a customer spends on purchases.

The RFM analysis involves assigning a score to each customer for each parameter, typically on a scale of 1 to 5 (with 5 being the highest). Then, customers are segmented based on these scores to create groups with similar behaviours. Common segmentation schemes include "Best Customers," "Loyal Customers," "Churn Risk," and "Low-Value Customers."

##### Assumptions:

* Recency is measured from the last transaction date to the current date.
* Frequency is calculated as the total number of transactions for a customer.
* Monetary is calculated as the total amount spent by a customer.

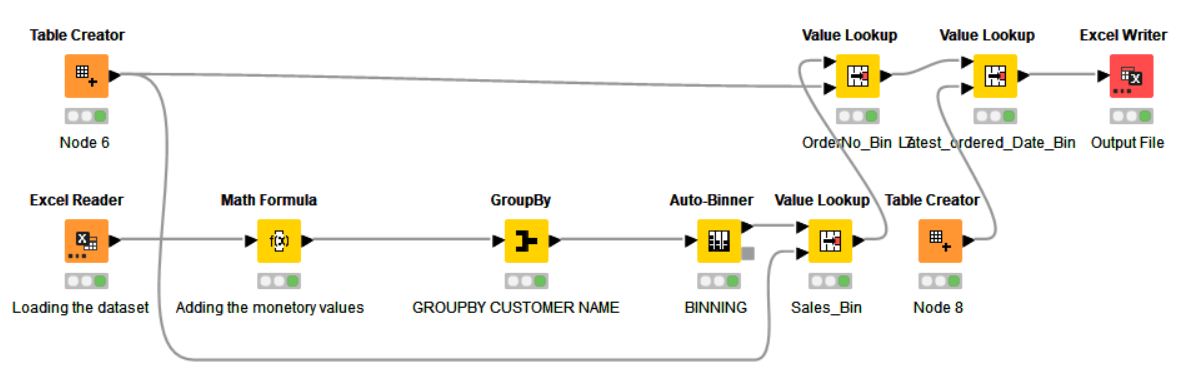


Figure 1. 9 KNIME Workflow Image

|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **CUSTOMERNAME** | **QUANTITYORDERED** | **PRICEEACH** | **COUNTRY** | **SALES** | **DAYS\_SINCE\_LASTORDER** | **R\_LABEL** | **F\_LABEL** | **M\_LABEL** |
| AV Stores, Co. | 34.8627 | 91.085 | UK | 157807.8 | 421 | H | M | M |
| Alpha Cognac | 34.35 | 101.16 | France | 70488.44 | 675 | L | M | L |
| Amica Models & Co. | 32.4231 | 110.85 | Italy | 94117.26 | 328 | M | L | M |
| Anna's Decorations, Ltd | 31.9348 | 106.42 | Australia | 153996.1 | 131 | H | H | H |
| Atelier graphique | 38.5714 | 92.239 | France | 24179.96 | 312 | L | M | M |
| Australian Collectables, Ltd | 30.6522 | 90.042 | Australia | 64591.46 | 1018 | L | M | L |
| Australian Collectors, Co. | 35.0182 | 104.59 | Australia | 200995.4 | 229 | H | H | H |
| Australian Gift Network, Co | 36.3333 | 110.55 | Australia | 59469.12 | 190 | L | M | H |
| Auto Assoc. & Cie. | 35.3889 | 99.488 | France | 64834.32 | 275 | L | L | M |
| Auto Canal Petit | 37.0741 | 94.255 | France | 93170.66 | 127 | M | M | H |
| Auto-Moto Classics Inc. | 35.875 | 92.8 | USA | 26479.26 | 1353 | L | M | L |

Table 1 Results in the output head of the data

# Inferences from RFM Analysis and identified segments -> Who are your best customers? (give at least 5) -> Which customers are on the verge of churning? (give at least 5) -> Who are your lost customers? (give at least 5) -> Who are your loyal customers? (give at least 5)

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
|  |  |  | **Monetary** | | |
| **Segmentation** | **Recency** | **Frequency** | **H** | **M** | **L** |
| **Loyal** | **H** | **H** | **11** | **1** | **0** |
| **M** | **1** | **9** | **1** |
| **L** | **0** | **0** | **0** |
| **At Risk** | **M** | **H** | **3** | **3** | **0** |
| **M** | **5** | **11** | **1** |
| **L** | **0** | **11** | **10** |
| **Churn** | **L** | **H** | **1** | **1** | **1** |
| **M** | **1** | **6** | **5** |
| **L** | **0** | **2** | **5** |

Table 2 Customer Segmentation with RFM Analysis

## Let's interpret the segments:

1. HHH Segment (High Recency, High Frequency, High Monetary):

* There are 11 customers who have recently made frequent purchases with high monetary spending.
* These are your most valuable and active customers.

2. HMH Segment (High Recency, Medium Frequency, High Monetary):

* There is 1 customer who has recently made moderate-frequency purchases with high monetary spending.
* This customer might be someone who has made several high-value purchases recently but is not as frequent as the HHH segment.

3. MMH Segment (Medium Recency, Medium Frequency, High Monetary):

* There are 11 customers who have made purchases with medium frequency and high monetary spending.
* These customers might be consistently spending good amounts.

4. MLL Segment (Medium Recency, Low Frequency, Low Monetary):

* There are 10 customers who have made purchases with medium recency and low frequency and monetary spending.
* These customers might have made a few small purchases and haven't returned recently.

5. LLL Segment (Low Recency, Low Frequency, Low Monetary):

* There are 5 customers who have low recency, low frequency, and low monetary spending.
* These are potentially inactive customers.

Best Customers (Top 5):

* **Anna's Decorations, Ltd** - High Monetary (M\_LABEL), High Frequency (F\_LABEL), High Recency (R\_LABEL)
* **Australian Collectors, Co.** - High Monetary (M\_LABEL), High Frequency (F\_LABEL), High Recency (R\_LABEL)
* **Euro Shopping Channel** - High Monetary (M\_LABEL), High Frequency (F\_LABEL), High Recency (R\_LABEL)
* **Mini Gifts Distributors Ltd**. - High Monetary (M\_LABEL), High Frequency (F\_LABEL), High Recency (R\_LABEL)
* **Rovelli Gifts** - High Monetary (M\_LABEL), High Frequency (F\_LABEL), High Recency (R\_LABEL)

Customers on the Verge of Churning (Top 5):

* **Clover Collections, Co.** - Low Monetary (M\_LABEL), Low Frequency (F\_LABEL), Moderate Recency (R\_LABEL)
* **Gift Ideas Corp.** - Low Monetary (M\_LABEL), Low Frequency (F\_LABEL), Low Recency (R\_LABEL)
* **Muscle Machine Inc** - High Monetary (M\_LABEL), High Frequency (F\_LABEL), Moderate Recency (R\_LABEL)
* **Quebec Home Shopping Network** - High Monetary (M\_LABEL), High Frequency (F\_LABEL), Low Recency (R\_LABEL)
* **Signal Collectibles Ltd.** - Low Monetary (M\_LABEL), Low Frequency (F\_LABEL), Low Recency (R\_LABEL)

Lost Customers (Top 5):

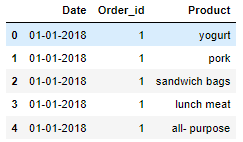
* **AV Stores, Co.** - High Monetary (M\_LABEL), Low Frequency (F\_LABEL), High Recency (R\_LABEL)
* **Auto Assoc. & Cie**. - Low Monetary (M\_LABEL), Low Frequency (F\_LABEL), Low Recency (R\_LABEL)
* **Herkku Gifts** - High Monetary (M\_LABEL), High Frequency (F\_LABEL), Moderate Recency (R\_LABEL)
* **Super Scale Inc.** - Low Monetary (M\_LABEL), Low Frequency (F\_LABEL), Moderate Recency (R\_LABEL)
* **Tokyo Collectables, Ltd** - High Monetary (M\_LABEL), High Frequency (F\_LABEL), Moderate Recency (R\_LABEL)

Loyal Customers (Top 5):

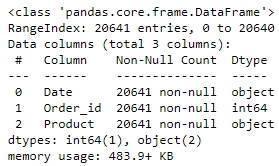
* **Oulu Toy Supplies, Inc.** - High Monetary (M\_LABEL), High Frequency (F\_LABEL), Moderate Recency (R\_LABEL)
* **Signal Gift Stores** - Moderate Monetary (M\_LABEL), Moderate Frequency (F\_LABEL), Moderate Recency (R\_LABEL)
* **Tekni Collectables Inc.** - High Monetary (M\_LABEL), High Frequency (F\_LABEL), High Recency (R\_LABEL)
* **The Sharp Gifts Warehouse** - High Monetary (M\_LABEL), High Frequency (F\_LABEL), High Recency (R\_LABEL)
* **Vitachrome Inc**. - High Monetary (M\_LABEL), High Frequency (F\_LABEL), High Recency (R\_LABEL)

# Part B

# Exploratory Analysis --> Exploratory Analysis of data & an executive summary (in PPT) of your top findings, supported by graphs. --> Are there trends across months/years/quarters/days etc. that you are able to notice?



Snippet 2. 1 Head of the Dataset



Snippet 2. 2 Information of the dataset

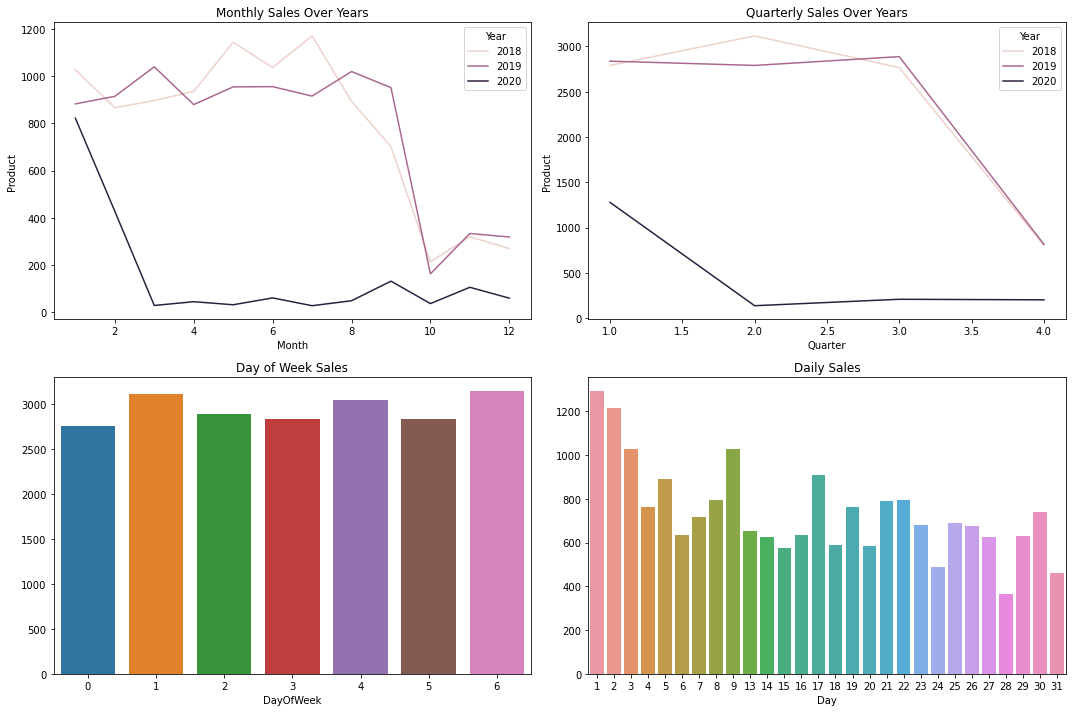


Figure 2. 1 Daily, Weekly, Quarterly and Yearly Sales Visualization

# PART B: Use of Market Basket Analysis (Association Rules) -->Write Something about the association rules and its relevance in this case -->Add KNIME workflow image -->Write about threshold values of Support and Confidence

##### Association Rules and Their Relevance:

* Association rules are a fundamental concept in data mining and analytics, commonly used in Market Basket Analysis.
* They reveal interesting relationships and patterns within large datasets by identifying co-occurring items in transactions.
* In the context of a grocery store's transactional data, association rules can provide insights into how different products are often purchased together.
* This information is invaluable for optimizing product placement, devising strategic cross-promotions, and enhancing customer experience.

##### In the case of the grocery store, association rules can help in several ways:

* Product Bundling: Identifying items frequently purchased together can lead to bundling products into combo offers, enhancing customer value and boosting sales.
* Store Layout: Placing co-occurring products near each other can increase convenience and encourage additional purchases.
* Targeted Marketing: Leveraging associations can facilitate targeted marketing campaigns to specific customer segments.
* Inventory Management: Understanding product relationships aids in efficient stock management and supply chain optimization.

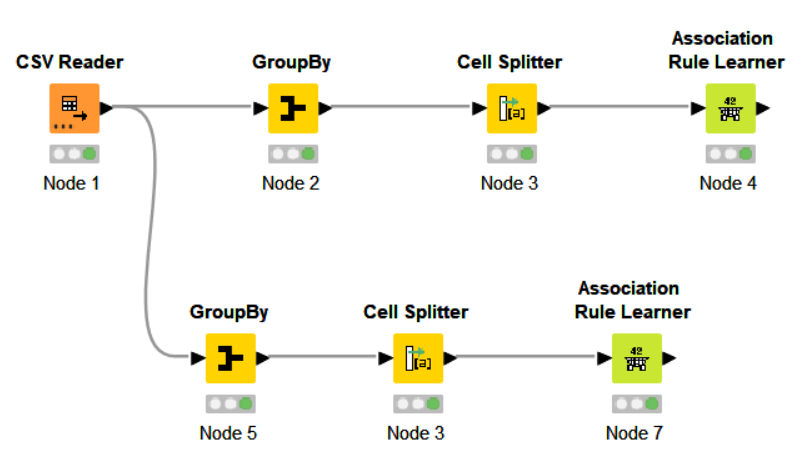


Figure 2. 2 KNIME workflow Image

##### Threshold Values of Support and Confidence:

* Support and confidence are critical parameters in association rule analysis. They define the significance of the discovered patterns:
* Support: These measures how frequently a set of items appears in the dataset. A higher support value implies that the rule is more common. A low support threshold might lead to rules that are too specific, while a high threshold might miss interesting associations.
* Confidence: This quantifies how often the consequent (right-hand side) of the rule occurs when the antecedent (left-hand side) is present. A high confidence suggests a strong association. Too low a confidence might lead to too many false-positive associations.
* Selecting appropriate threshold values involves a trade-off between the number of discovered rules and their significance. It's often a trial-and-error process. Start with higher thresholds and gradually lower them until you achieve a balance between relevant rules and manageable results.
* In the KNIME workflow, you can adjust these threshold values based on the nature of the data and the goals of your analysis. Iterative experimentation can help you identify the thresholds that provide the most valuable insights.

# Associations Identified --> Put the associations in a tabular manner --> Explain about support, confidence, & lift values that are calculated

##### Associations Identified and Their Significance:

* Below is the tabular representation of the identified associations along with their corresponding support, confidence, and lift values:

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **Support** | **Confidence** | **Lift** | **Consequent** | **Implies** | **Items** |
| 0.050 | 0.640 | 1.700 | juice | <--- | [yogurt, toilet paper, aluminium foil] |
| 0.050 | 0.620 | 1.645 | juice | <--- | [yogurt, poultry, aluminium foil] |
| 0.050 | 0.613 | 1.616 | coffee/tea | <--- | [yogurt, cheeses, cereals] |
| 0.050 | 0.600 | 1.424 | poultry | <--- | [dishwashing liquid/detergent, laundry detergent, mixes] |
| 0.051 | 0.630 | 1.678 | mixes | <--- | [yogurt, poultry, aluminium foil] |
| 0.051 | 0.611 | 1.660 | sandwich bags | <--- | [cheeses, bagels, cereals] |
| 0.051 | 0.674 | 1.726 | cheeses | <--- | [bagels, cereals, sandwich bags] |
| 0.051 | 0.617 | 1.558 | cereals | <--- | [cheeses, bagels, sandwich bags] |
| 0.051 | 0.630 | 1.621 | dinner rolls | <--- | [spaghetti sauce, poultry, cereals] |
| 0.051 | 0.637 | 1.512 | poultry | <--- | [dinner rolls, spaghetti sauce, cereals] |
| 0.051 | 0.604 | 1.589 | milk | <--- | [poultry, laundry detergent, cereals] |
| 0.052 | 0.628 | 1.610 | eggs | <--- | [dinner rolls, poultry, soda] |
| 0.052 | 0.641 | 1.649 | dinner rolls | <--- | [spaghetti sauce, poultry, ice cream] |
| 0.052 | 0.686 | 1.628 | poultry | <--- | [dinner rolls, spaghetti sauce, ice cream] |
| 0.052 | 0.628 | 1.614 | dinner rolls | <--- | [spaghetti sauce, poultry, juice] |
| 0.052 | 0.602 | 1.429 | poultry | <--- | [dinner rolls, spaghetti sauce, juice] |
| 0.052 | 0.634 | 1.627 | eggs | <--- | [paper towels, dinner rolls, pasta] |
| 0.052 | 0.602 | 1.621 | pasta | <--- | [paper towels, eggs, dinner rolls] |
| 0.054 | 0.642 | 1.651 | dinner rolls | <--- | [spaghetti sauce, poultry, laundry detergent] |
| 0.054 | 0.656 | 1.556 | poultry | <--- | [dinner rolls, spaghetti sauce, laundry detergent] |
| 0.055 | 0.624 | 1.565 | ice cream | <--- | [paper towels, eggs, pasta] |
| 0.055 | 0.630 | 1.616 | eggs | <--- | [paper towels, ice cream, pasta] |
| 0.055 | 0.643 | 1.731 | pasta | <--- | [paper towels, eggs, ice cream] |
| 0.055 | 0.649 | 1.791 | paper towels | <--- | [eggs, ice cream, pasta] |

##### Understanding Support, Confidence, and Lift Values:

**Support:**

* Support indicates the frequency of an itemset in the dataset. For example, a support of 0.050 means that the itemset appears in approximately 5% of all transactions.
* Higher support values imply that the association is relatively common.

**Confidence:**

* Confidence measures how often the consequent occurs when the antecedent is present. For instance, a confidence of 0.640 means that juice is purchased along with the other items in about 64% of the cases where those items are present.
* Higher confidence values imply a stronger association between the antecedent and the consequent.

**Lift:**

* Lift quantifies how much more likely it is that the two items will be bought together compared to if they were bought independently. Lift values greater than 1 suggest a positive association between the items.
* A lift of 1 implies independence, while a lift greater than 1 implies a positive relationship.

**Interpretation:**

* For instance, the association rule "juice <--- [yogurt, toilet paper, aluminium foil]" has a support of 0.050, suggesting that this combination is present in 5% of transactions.
* The confidence value of 0.640 indicates that juice is purchased along with the other items in about 64% of cases where yogurt, toilet paper, and aluminium foil are purchased.
* A lift of 1.700 suggests that the purchase of juice is 1.7 times more likely when yogurt, toilet paper, and aluminium foil are purchased together than when they are bought independently.

# Suggestion of Possible Combos with Lucrative Offers --> Write recommendations --> Make discount offers or combos (or buy two get one free) based on the associations and your experience

##### Suggestions for Lucrative Combos with Discount Offers:

Based on the associations identified through the Association Rule analysis, we can recommend the following potential combos with lucrative offers for the grocery store. These offers are designed to capitalize on the relationships between products and encourage customers to make additional purchases, ultimately increasing revenue.

##### Combo: Breakfast Boost

* Offer: Buy a pack of yogurt, a box of cereal, and a carton of milk together at a discounted price.
* Rationale: The association rules show strong connections between yogurt, cereals, and milk. By bundling these items, customers are more likely to make a complete breakfast purchase.

##### Combo: Family Pasta Night

* Offer: Purchase a jar of spaghetti sauce, a pack of pasta, and a poultry product to get a special discount.
* Rationale: The analysis reveals associations between spaghetti sauce, poultry, and pasta. Offering a combo deal for a family pasta night can entice customers to buy all the necessary components.

##### Combo: Snack Delight

* Offer: Combine cheeses, bagels, and sandwich bags in a single purchase for a reduced price.
* Rationale: The strong associations between cheeses, bagels, and sandwich bags suggest that customers often buy these items together. Providing them as a combo can enhance convenience and save money for the shoppers.

##### Combo: Cleaning Essentials

* Offer: Bundle dishwashing liquid/detergent, laundry detergent, and cleaning mixes at a special offer price.
* Rationale: The association rule linking these cleaning items together can prompt customers to stock up on their household cleaning essentials.

##### Combo: Refreshing Beverage Pack

* Offer: Purchase juice, coffee/tea, and soda together with a discount.
* Rationale: The associations between these beverages indicate potential for cross-promotions, targeting customers who enjoy a variety of drinks.

##### Combo: Weekend Indulgence

* Offer: Buy a variety of products like ice cream, dinner rolls, and other treats as a weekend package at a lower cost.
* Rationale: Combining indulgent items associated with weekend enjoyment can drive increased sales, especially during leisure days.

##### Additional Strategies:

* **"Buy Two, Get One Free" Deals**: Implementing "buy two, get one free" or similar promotions on products that have strong associations can incentivize customers to purchase more and capitalize on the products' relationships.
* **Promotion Visibility:** Display products in these combos together, either physically in-store or digitally in an online platform, to remind customers of the enticing deals.
* **Loyalty Programs:** Introduce loyalty programs that offer exclusive discounts and rewards to customers who consistently purchase products from specific combos.
* **Limited-Time Offers:** Creating a sense of urgency with limited-time offers can encourage customers to make quicker purchasing decisions.

Remember that the success of these offers lies in effective communication, clear value propositions, and aligning the combos with customer preferences. Regularly analysing sales data and customer feedback will allow for fine-tuning these strategies over time.