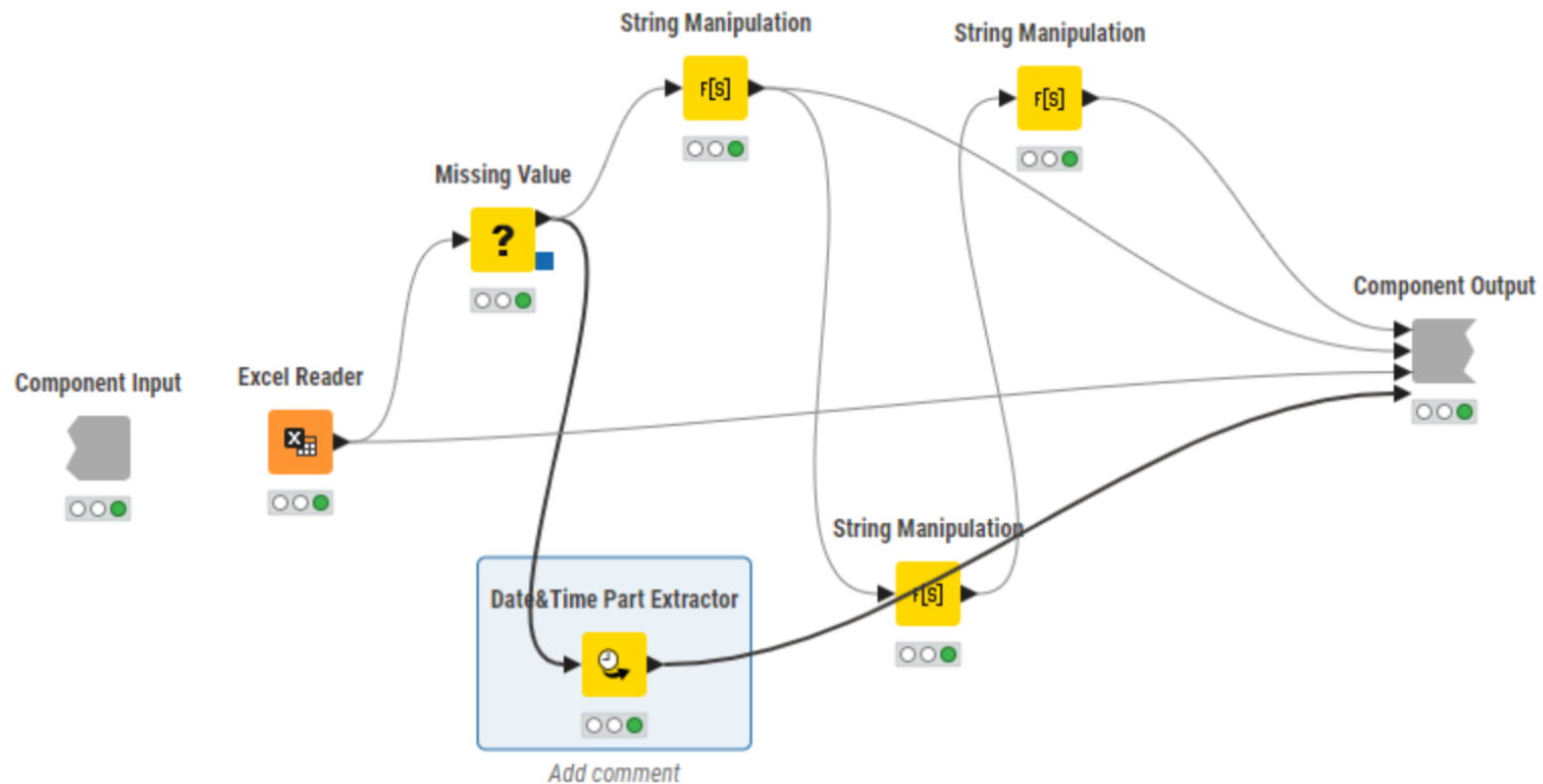# MRA EXTENDED PROJECT

DATA CLEANING AND PRE-PROCESSING

To prepare the dataset for analysis, missing values were addressed by filling Local Date and Local Time using the Previous Value method, replacing strings with [NA], and filling integer and double columns with mean values. String manipulation steps included standardizing time to a 24-hour format, merging the Item Description and Category columns for minimization, and removing unnecessary spaces for uniform formatting. Additionally, the Date column was processed to extract Day, Month, and Year fields to enable time-based analysis.

EXPLORATORY DATA ANALYSIS(EDA) FOR CAFE CHAIN

Rule Engine node was utilized to categorize the day into Morning, Afternoon, and Evening based on time. This was followed by using a Row Aggregator node, configured to calculate the total sum for each category. These steps provided insights into the time-of-day performance, aiding in identifying sales trends and peak hours for the café these are visualize using Bar chart.
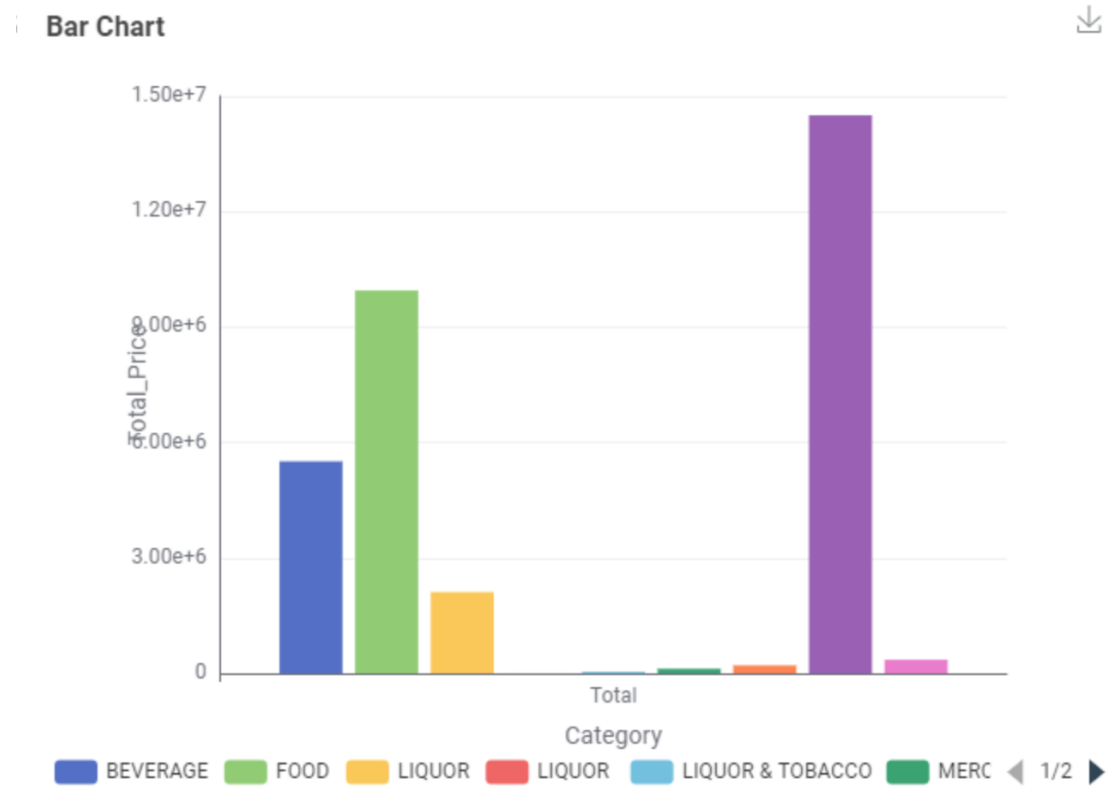
CATEGORY-WISE SALES ANALYSIS

Pie Chart
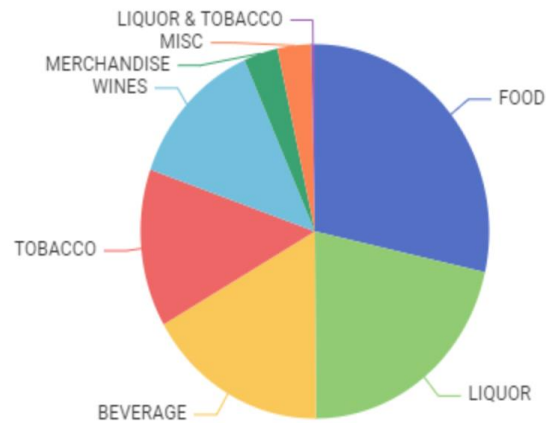
The analysis began by filtering Morning transactions using the Row Filter node to focus on specific time-based data. A GroupBy node was applied to aggregate the data based on Total, Prediction, and Item Description. Additionally, a second GroupBy node was used to summarize sales by Category and Total. The resulting insights were visualized using a pie chart, providing a clear depiction of the proportionate sales contributions of different categories during the Morning period. This aids in identifying top-performing categories and predicting future trends.

TIME-BASED QUANTITY ANALYSIS

The Rule Engine node was employed to segment the day into Morning, Afternoon, and Evening based on time. Subsequently, a Row Aggregator node was configured to compute the total quantity sum for each time category. This approach helped in identifying the quantity distribution across different times of the day, providing insights into customer purchasing patterns using Bar chart.
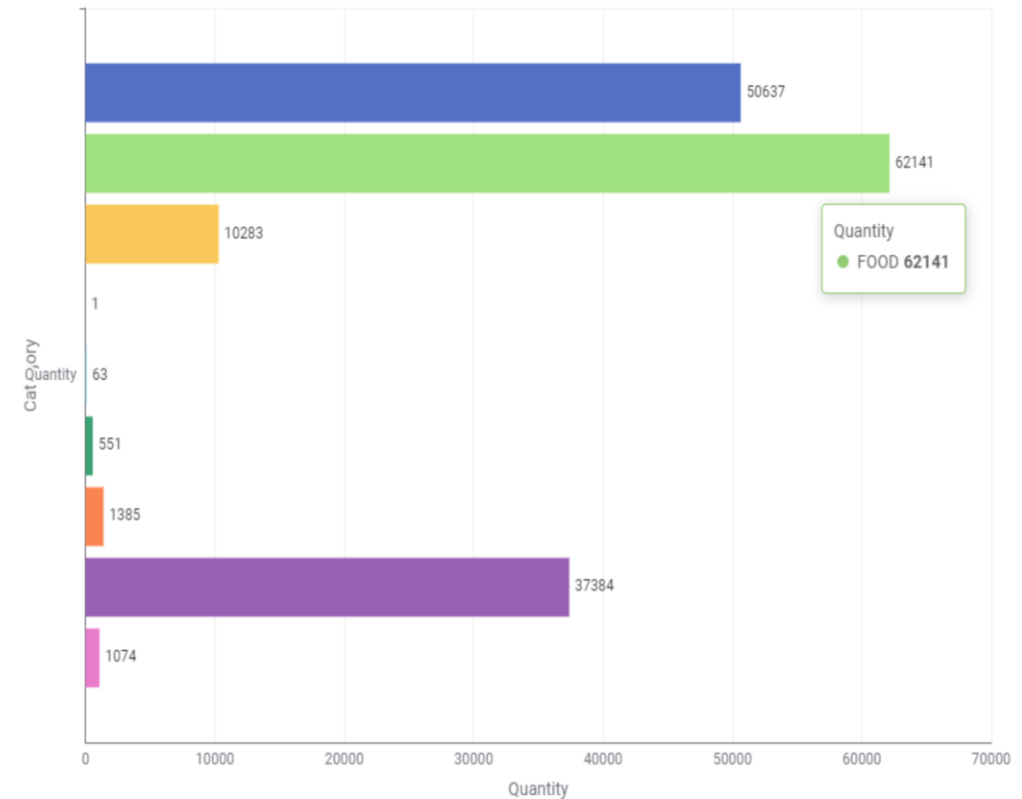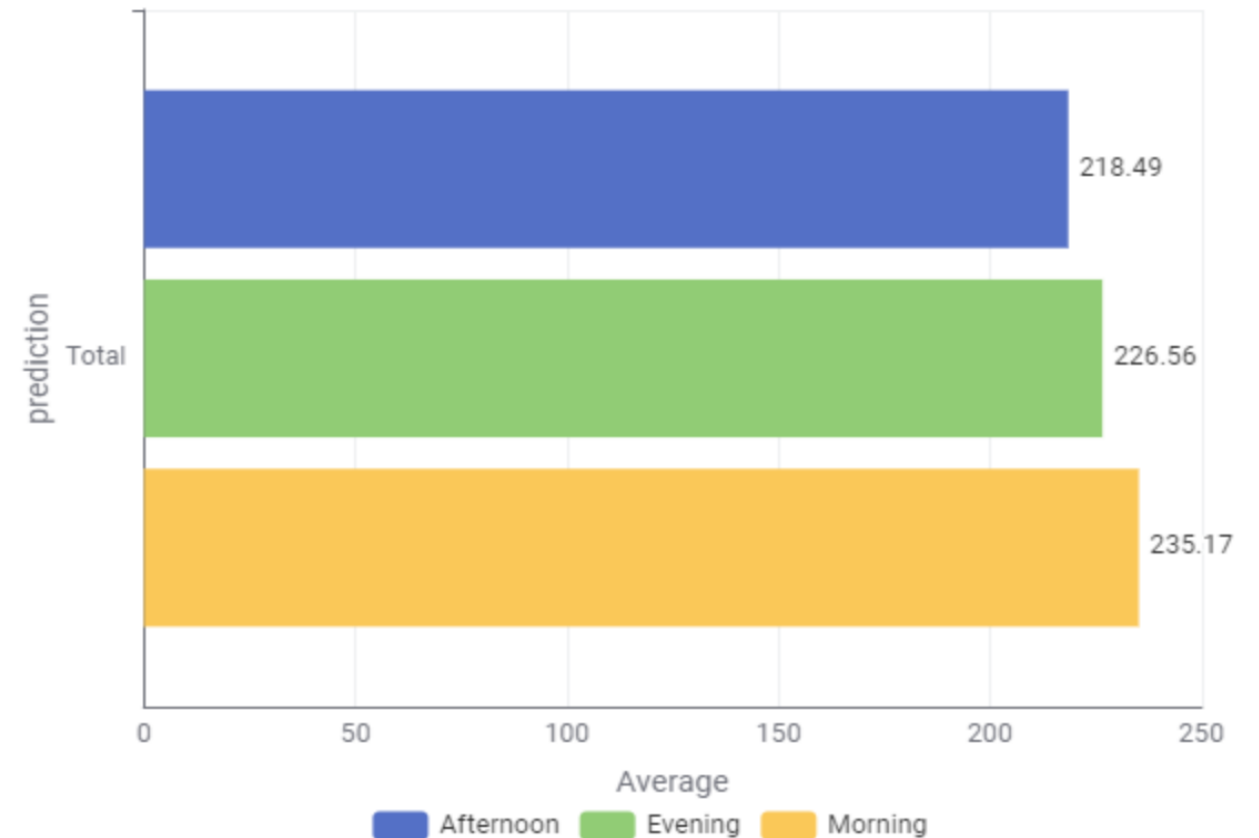
# TIME-BASED SALES DISTRIBUTION

The Rule Engine node was employed to segment the day into Morning, Afternoon, and Evening based on time . This bar chart visualizes the total quantity sold across Morning, Afternoon, and Evening time segments, offering a clear representation of peak and low sales periods throughout the day.

Bar Chart

CUSTOMER SEGMENTATION ANALYSIS

Pie Chart
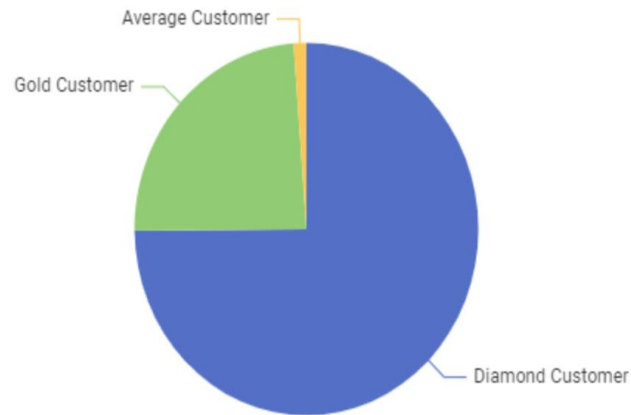
The dataset was grouped by Bill Number and Total using the GroupBy node to analyze individual customer spending patterns. A Rule Engine node was then employed to classify customers based on their total spending:

• Diamond Customer : Those with a total bill greater than ₹14,000

• Gold Customer : Those with a total bill between ₹4,001 and ₹14,000.

• Average Customer : Those with a total bill ranging from ₹0 to ₹4,000.

PRIORITY-BASED CUSTOMER ANALYSIS

The dataset was grouped by Priority Based Customer and Bill number to count the occurrences of each priority category. A bar chart was then created to visualize the distribution of customer priorities, offering a clear breakdown of customer segments and their frequency, assisting in identifying the most and least common spending categories.
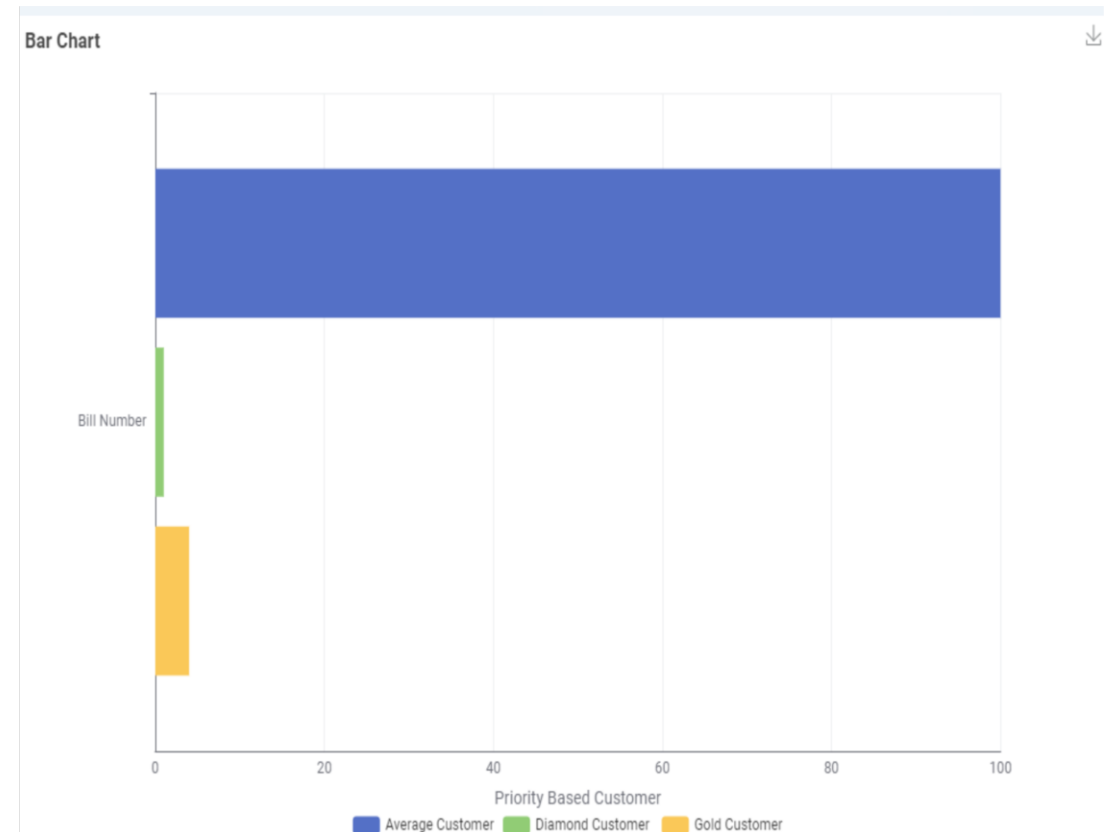
MONTHLY SALES DISTRIBUTION ANALYSIS

Pie Chart
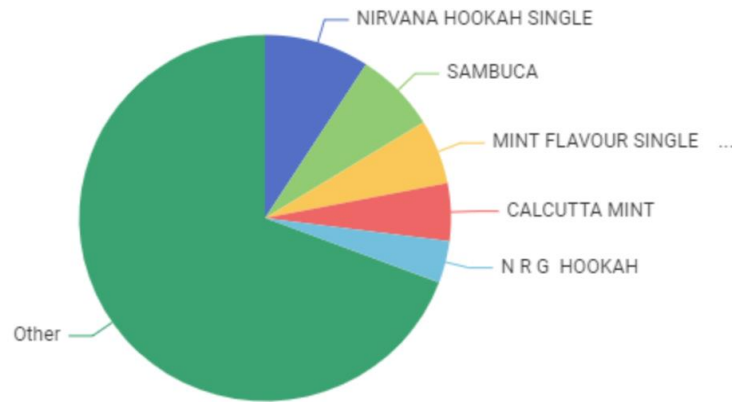
The dataset was grouped by month , year , week , and day of week(number) to analyze trends across different time periods. A Manual aggregation was applied using the GroupBy node to compute the sum of sales metrics for each time-based grouping.

To streamline the insights, a Row aggregator node was used to calculate the overall total sales for these categories. This comprehensive aggregation facilitated a direct comparison of monthly sales performance.

Finally, a Pie chart was generated to visualize the contribution of each month to the total sales, highlighting the peak months and periods of low activity. This representation supports strategic decision-making, such as timing promotions and optimizing operational resources.

DATA COLUMN ANALYSIS USING EXTRACT TABLE SPECIFIC NODE

The Extract Table Specific Node was used to analyze each column's lower and upper bounds, and column types. This helped identify the range for numerical data, detect outliers, and ensure correct data types (e.g., String, Integer, Date) were applied. This analysis streamlined the data cleaning and preparation process for further analysis and modeling.

Table View

Rows: 10 | Columns: 5

	RowID	Column ... String	Column T... String	Column I... Number (inte...	Lower Bo... ?	Upper Bo... ?
☐	Date	Date	Local Date	0	2010-04-01	2011-03-31
☐	Bill Nui	Bill Number	String	1	⦿	⦿
☐	Item Dı	Item Desc	String	2	⦿	⦿
☐	Time	Time	Local Time	3	00:00:01	23:59:57
☐	Quantii	Quantity	Number (double	4	1	30
☐	Rate	Rate	Number (double	5	0.01	2100.0
☐	Tax	Tax	Number (double	6	0.0	2731.25
☐	Discou	Discount	Number (double	7	0.0	825.0

STATISTICAL OVERVIEW OF THE DATASET

A detailed statistical analysis was conducted to understand the dataset's structure and key metrics. Central tendencies like mean, median, and standard deviation were calculated for numeric fields, along with range analysis to identify outliers and extremes. Missing values and inconsistencies were also assessed to ensure data quality. This foundational analysis provided valuable insights for trends, anomalies, and preparation for deeper (EDA).
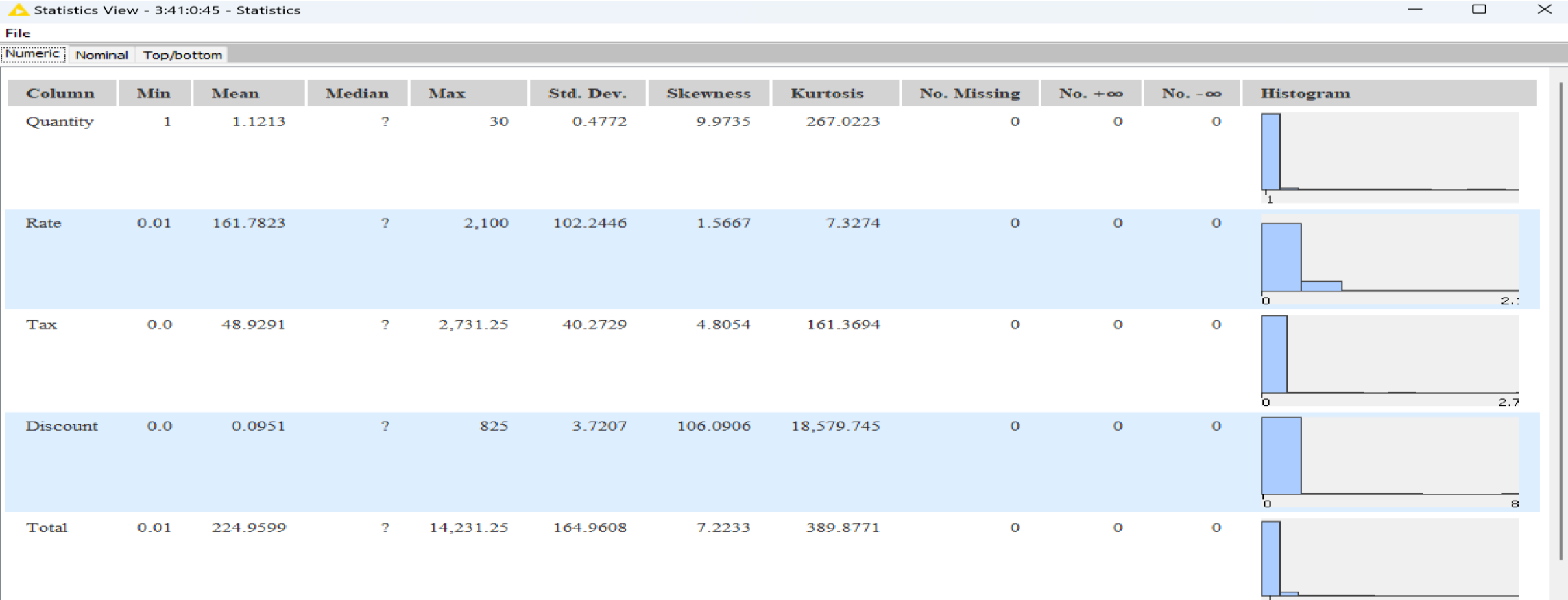
Statistics View - 3:41:0:45 - Statistics

File

Numeric | Nominal | Top/bottom

Column	Min	Mean	Median	Max	Std. Dev.	Skewness	Kurtosis	No. Missing	No. +∞	No. -∞	Histogram
Quantity	1	1.1213	?	30	0.4772	9.9735	267.0223	0	0	0	
Rate	0.01	161.7823	?	2,100	102.2446	1.5667	7.3274	0	0	0	
Tax	0.0	48.9291	?	2,731.25	40.2729	4.8054	161.3694	0	0	0	
Discount	0.0	0.0951	?	825	3.7207	106.0906	18,579.745	0	0	0	
Total	0.01	224.9599	?	14,231.25	164.9608	7.2233	389.8771	0	0	0	

TABLE VIEW

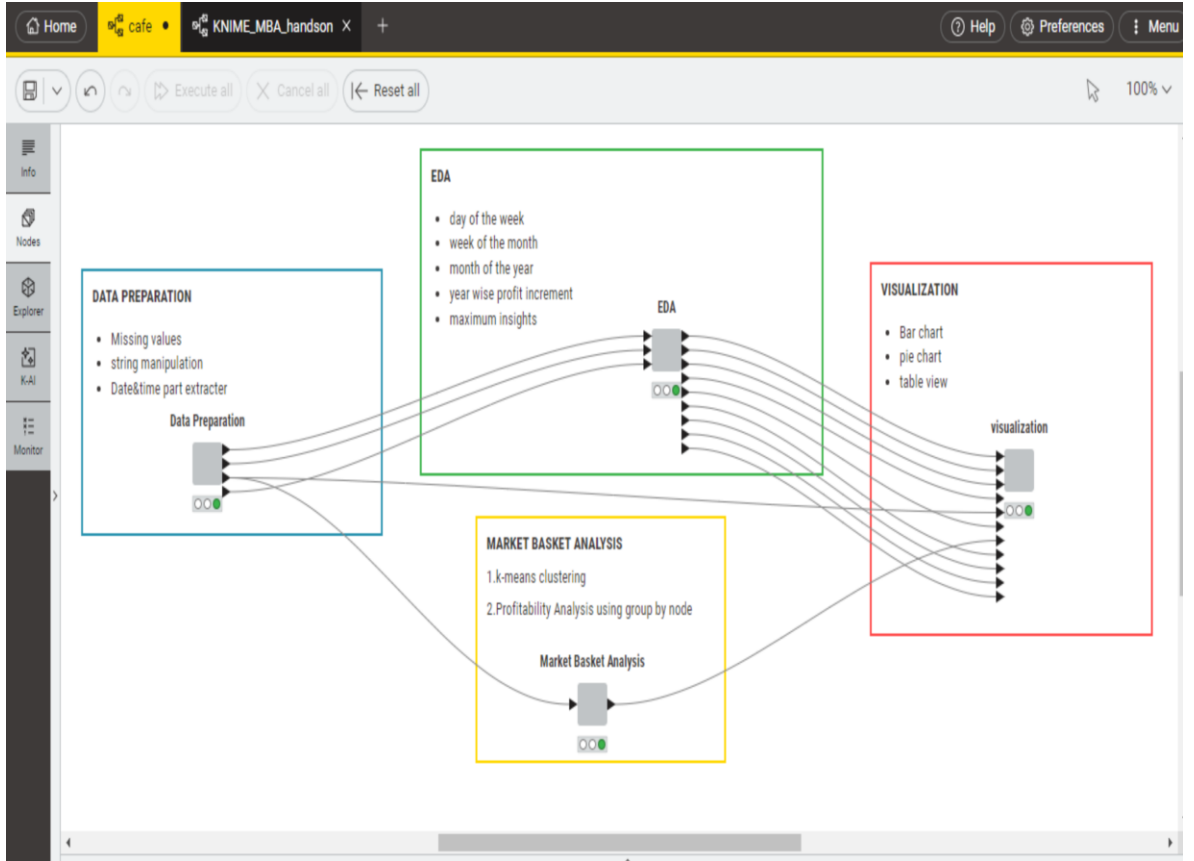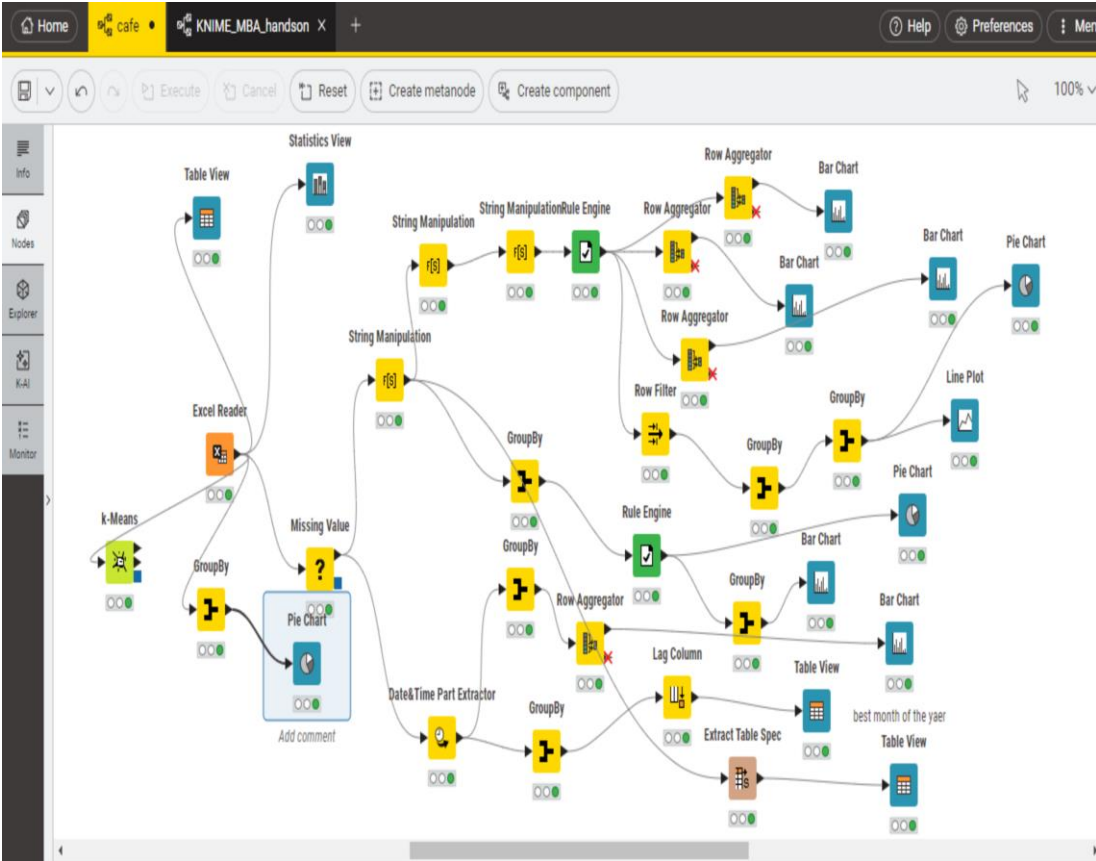**Table View**

Rows: 145830 | Columns: 10   🔍   ⚙

☐	RowID	Date Local Date	Bill Numb... String	Item Desc String	Time Local Time	Quantity Number (inte...	
☐	Row0	2010-04-01	G0470115	QUA MINERAL	13:15:11	1	50
☐	Row1	2010-04-01	G0470115	MONSOON MA	13:15:11	1	10
☐	Row2	2010-04-01	G0470116	MASALA CHAI	13:17:35	1	40
☐	Row3	2010-04-01	G0470117	QUA MINERAL	13:19:55	1	50
☐	Row4	2010-04-01	G0470283	MOROCCAN MI	01:20:18	1	45
☐	Row5	2010-04-01	G0470283	QUA MINERAL	01:20:18	1	50
☐	Row6	2010-04-01	G0470118	MASALA CHAI	13:21:34	1	40
☐	Row7	2010-04-01	G0470118	MOROCCAN MI	13:21:34	1	45

OVERALL WORKFLOW

01 — Increasing Market Share

Behavior Analysis

02

03 — Optimization of in-store operations

Campaigns & Promotions

04

05 — Recommendations

MARKET BASKET ANALYSIS

K-MEANS CLUSTERING

Cluster View - 3:42:0:38 - k-Means

File   Hilite

📁 3 Clusters
└─ 📁 cluster_0 (coverage: 95125)
    ├─ ● Discount = 0.07844951379763347
    └─ ● Total = 133.10299826547677
└─ 📁 cluster_1 (coverage: 42256)
    ├─ ● Discount = 0.055622160166603554
    └─ ● Total = 351.59213295163556
└─ 📁 cluster_2 (coverage: 8449)
    ├─ ● Discount = 0.47964256124985205
    └─ ● Total = 625.824995857494

CATEGORY AND ITEM ANALYSIS
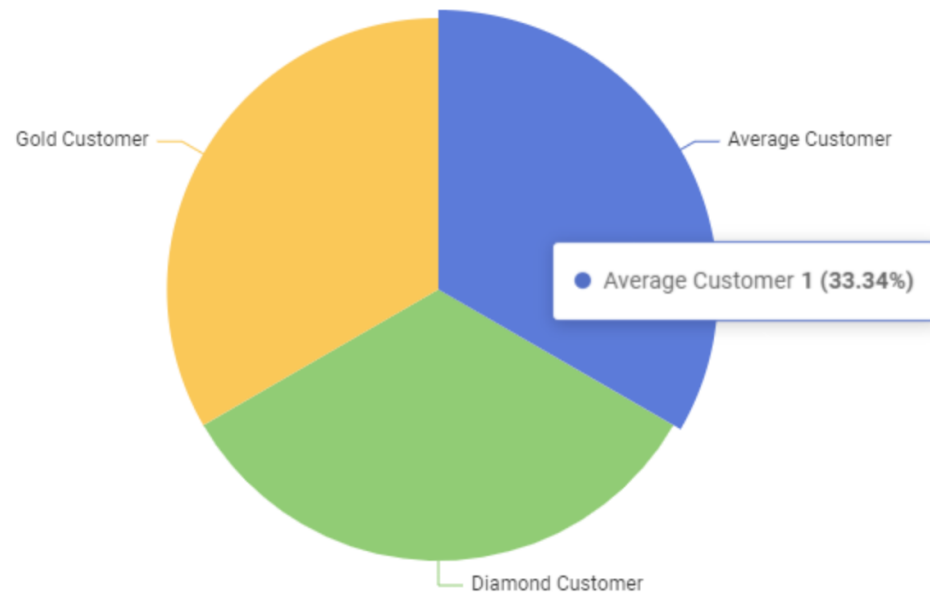
The GroupBy node aggregated sales data by Item Description and Category. A Pie Chart visualized the proportionate sales contribution of each, highlighting top-performing products and categories for strategic decisions.

YEARLY SALES ANALYSIS

Table View

Rows: 3 | Columns: 3

	RowID	year Number (integer)	Total Number (double)	Total(-1) Number (double)
☐	Row0	2010	24,054,138.14	⑦
☐	Row1	2011	8,751,757.13	24,054,138.14
☐	overflo ⑦		⑦	8,751,757.13

The GroupBy node aggregated total sales data by year using manual aggregation. This enabled the calculation of year-wise sales trends, highlighting increments or decrements in annual performance, providing valuable insights into long-term growth patterns

ASSOCIATION RULE ANALYSIS

Association Rule Mining was applied to uncover relationships between items frequently purchased together. This technique identified patterns within the dataset, such as product combinations commonly ordered by customers. These insights are particularly relevant for optimizing product placements, designing bundle offers, and enhancing customer experience by targeting popular item combinations. This analysis supports strategic decision-making to boost cross-selling opportunities and overall profitability.

THRESHOLD VALUES OF SUPPORT AND CONFIDENCE

In Association Rule Mining, Support measures the frequency of item combinations, while Confidence evaluates the likelihood of one product being purchased given another. Setting appropriate thresholds ensures the analysis focuses on significant, actionable patterns, balancing the trade-off between irrelevant combinations and valuable insights for strategic improvements

THANK YOU