UNIVERSITI
MALAYA

WIE3007

DATA MINING AND WAREHOUSING

SEMESTER 1 2023/2024

GROUP ASSIGNMENT

NAME	MATRIC NO
NURUL AIN BINTI KHAIRUL ANWAR	U2005370
IFFAH SORFINA BINTI MOHAMAD NAIM	17203173
LUBNA BINTI AHMAD NIZAM	17204040
NURAUFA NATASHA BINTI AZROL	U2102739
EZYAN MUNIRAH BINTI ZAINUDDIN	U2000643

INSTRUCTOR: PROFESSOR DR. TEH YING WAH

1.0 Introduction

Retail, as an industry, thrives on the ability to decipher the intricate patterns woven by consumers as they navigate choices, make purchasing decisions, and shape the market dynamics. In this report, we delve into the realm of Superstore's purchase history, recognizing the fundamental significance of unraveling these patterns through data mining techniques.

At its essence, retail is a dynamic interplay between supply and demand, with customer preferences acting as the guiding force. The choices consumers make reflect not only their immediate needs but also broader trends that influence product offerings, marketing strategies, and inventory management. In this context, the strategic interpretation of customer data becomes an indispensable tool for retailers seeking not just to meet demand but to anticipate it.

The importance of understanding customer preferences becomes even more pronounced in a consumer-driven era. With an abundance of information and an array of choices at their fingertips, customers wield significant influence over market trends. In this environment, retailers need to go beyond merely providing products; they must anticipate and cater to evolving preferences to stay relevant and competitive.

Data mining emerges as a powerful ally in this endeavor, allowing retailers to extract meaningful insights from the vast troves of transactional data. It is not merely about understanding what customers bought in the past, but about distilling actionable intelligence that illuminates the path forward.

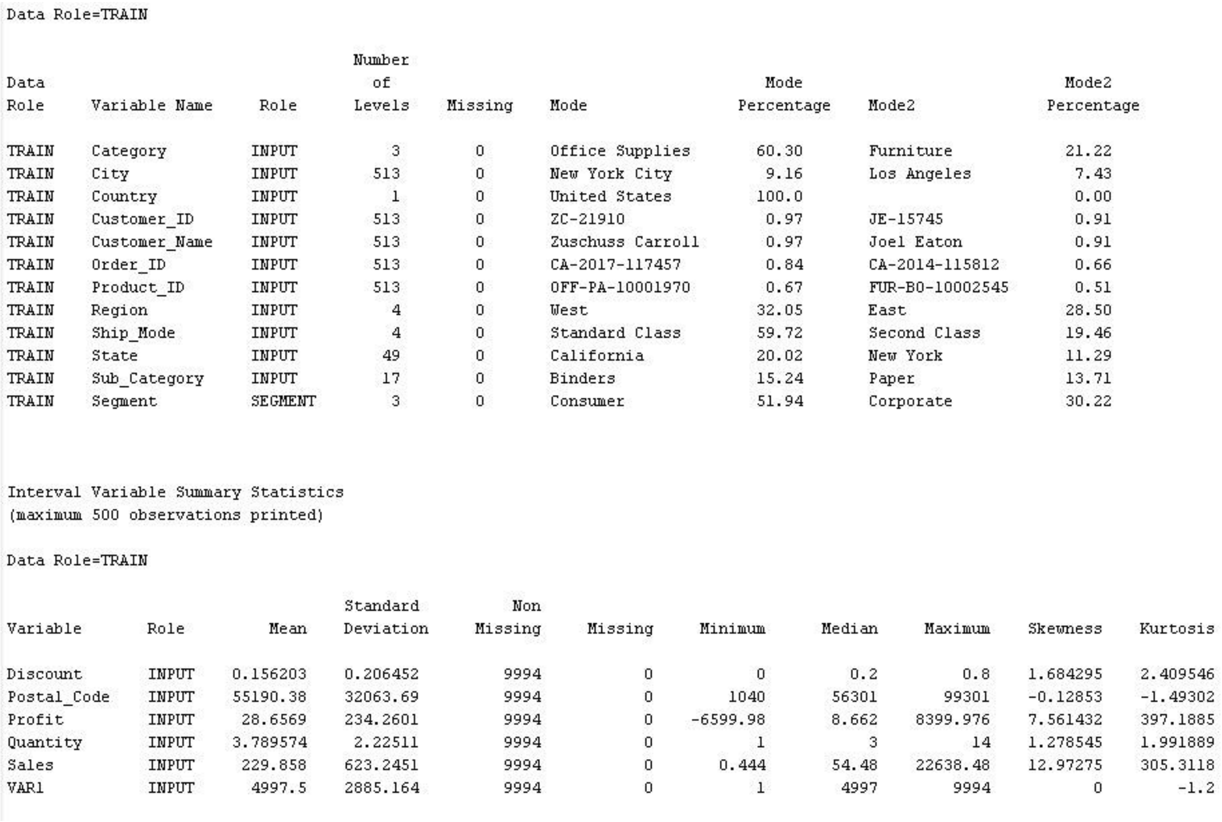
1.1 Problem Statements

Understanding customer buying habits is key for a business to thrive. This report explores the purchase history of Superstore, a major retail player, using data mining techniques. By diving into the data, we aim to discover patterns and trends that can help the organization to make smarter decisions about what products to offer and improve the overall shopping experience.

1.2 Objectives

- To understand the preferences of customers and their purchasing behavior
- To understand the performance of the product sales

1.3 Dataset Description



```
Data Role=TRAIN

                                       Number
Data                                     of                                       Mode                          Mode2
Role     Variable Name     Role        Levels    Missing    Mode                Percentage    Mode2             Percentage

TRAIN    Category          INPUT          3         0       Office Supplies       60.30       Furniture            21.22
TRAIN    City              INPUT        513         0       New York City          9.16       Los Angeles           7.43
TRAIN    Country           INPUT          1         0       United States        100.0                              0.00
TRAIN    Customer_ID       INPUT        513         0       ZC-21910               0.97       JE-15745              0.91
TRAIN    Customer_Name     INPUT        513         0       Zuschuss Carroll       0.97       Joel Eaton            0.91
TRAIN    Order_ID          INPUT        513         0       CA-2017-117457         0.84       CA-2014-115812        0.66
TRAIN    Product_ID        INPUT        513         0       OFF-PA-10001970        0.67       FUR-BO-10002545       0.51
TRAIN    Region            INPUT          4         0       West                  32.05       East                 28.50
TRAIN    Ship_Mode         INPUT          4         0       Standard Class        59.72       Second Class         19.46
TRAIN    State             INPUT         49         0       California            20.02       New York             11.29
TRAIN    Sub_Category      INPUT         17         0       Binders               15.24       Paper                13.71
TRAIN    Segment           SEGMENT        3         0       Consumer              51.94       Corporate            30.22


Interval Variable Summary Statistics
(maximum 500 observations printed)

Data Role=TRAIN

                                      Standard        Non
Variable        Role        Mean      Deviation     Missing     Missing     Minimum      Median      Maximum    Skewness     Kurtosis

Discount        INPUT    0.156203     0.206452        9994           0           0         0.2          0.8    1.684295     2.409546
Postal_Code     INPUT    55190.38     32063.69        9994           0        1040       56301        99301    -0.12853     -1.49302
Profit          INPUT     28.6569     234.2601        9994           0    -6599.98       8.662     8399.976    7.561432     397.1885
Quantity        INPUT    3.789574      2.22511        9994           0           1           3           14    1.278545     1.991889
Sales           INPUT     229.858     623.2451        9994           0       0.444       54.48     22638.48    12.97275     305.3118
VAR1            INPUT      4997.5     2885.164        9994           0           1        4997         9994           0         -1.2
```

The dataset used for this assignment is retail sales data from Superstore, “Superstore.csv”. This dataset appears to be covering various aspects of sales transactions including customer information, order details, product details, and financial figures like sales, quantity, discount, and profit. The data types are a mix of integers, floats, and strings, indicating numerical, categorical, and textual data.

The "Superstore.csv" dataset comprises 9994 entries with 21 columns, each representing different attributes. Here's a brief description of each column:

1. Row ID: An integer identifier for each row in the dataset.

2. Order ID: A string identifier for each order.

3. Order Date: The date when the order was placed (as a string).

4. Ship Date: The date when the order was shipped (as a string).

5. Ship Mode: The mode of shipping used for the order (as a string).

6. Customer ID: A string identifier for each customer.

7. Customer Name: The name of the customer (as a string)

8. Segment: The market segment to which the custome belongs as a string).

9. Country: The country where the order was placed (as astring).

10. City: The city where the order was placed (as a string).

11. State: The state where the order was placed (as a string).

12. Postal Code: The postal code for the order's location (as an integer).

13. Region: The region where the order was placed (as a string).

14. Product ID: A string identifier for each product.

15. Category: The category of the product (as a string).

16. Sub-Category: The sub-category of the product (as a string).

17. Product Name: The name of the product (as a string).

18. Sales: The sales amount for each order (as a float).

19. Quantity: The quantity of items in each order (as an integer).

20. Discount: The discount given on each order (as a float).

21. Profit: The profit made on each order (as a float).

2.0 Methodology

2.1 Tools

- SAS Enterprise Miner
- Talend Data Preparation
- Featuretools

2.2 Approach

For this assignment, SEMMA methodology was utilized to perform a data mining process upon the dataset to gain valuable insights. SEMMA is an acronym representing Sample, Explore, Modify, Model and Assess. The SEMMA methodology was popularized by SAS (Statistical Analysis System) Institute as a structured approach for solving business problems through data mining and predictive modeling.

Sample: The first step involves selecting a representative sample of the data. This sample is used for exploration and model development. Understanding the characteristics of the data is crucial in subsequent stages.

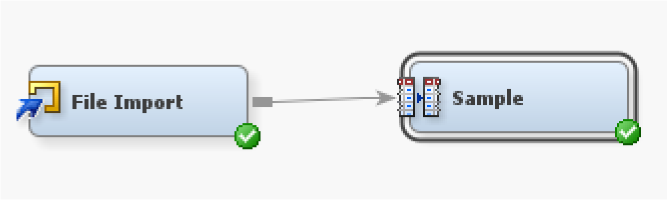
Explore: In the exploration phase, analysts examine and visualize the data to gain insights into its structure, patterns, and potential relationships. Exploratory data analysis helps identify important variables and understand the distribution of data.

Modify: The modification phase focuses on data preparation and preprocessing. This includes handling missing values, transforming variables, dealing with outliers, and other steps to ensure the data is suitable for modeling.

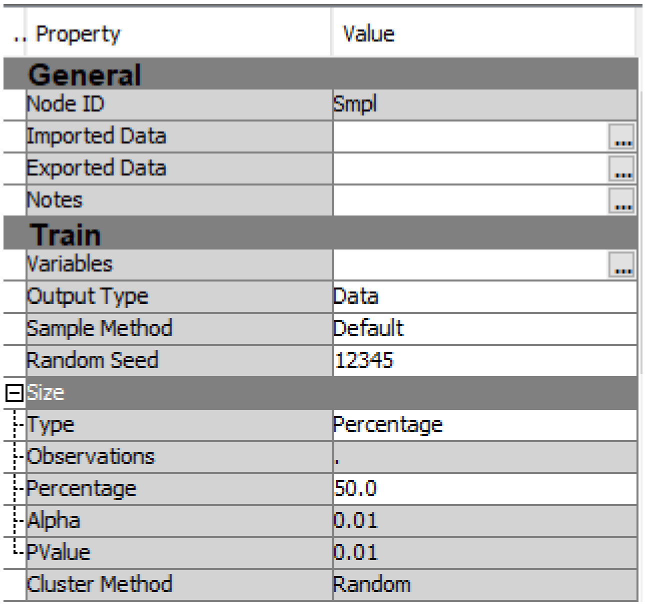
Model: The modeling phase involves the development of predictive models. Different algorithms and techniques are applied to build models that can make accurate predictions or classifications based on the data. This is a crucial step in deriving actionable insights.

Assess: In the assessment phase, the performance of the models is evaluated. This includes assessing the accuracy, reliability, and generalization capability of the models using validation techniques. Model assessment helps ensure that the models perform well on new, unseen data.

3.0 Sample

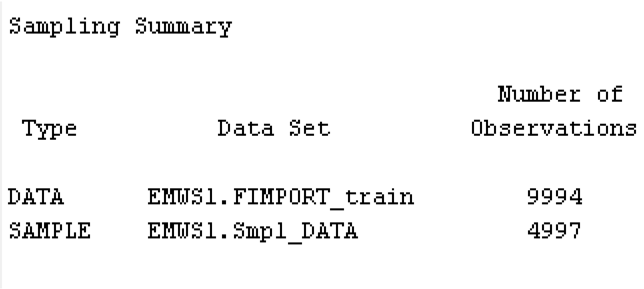


It is a process of choosing a sample from the data to be representative of the whole dataset. It can minimize the size of the dataset while still preserving data integrity and essential information. It can be quite challenging to do analysis when working with a huge volume of data.



.. Property	Value
General	
Node ID	Smpl
Imported Data	...
Exported Data	...
Notes	...
Train	
Variables	...
Output Type	Data
Sample Method	Default
Random Seed	12345
⊟ Size	
Type	Percentage
Observations	.
Percentage	50.0
Alpha	0.01
PValue	0.01
Cluster Method	Random

The sample node was used in sampling 50% of the data to be used. We reduced the size of the data by random sampling where each row has an equal chance to be chosen in the final dataset.

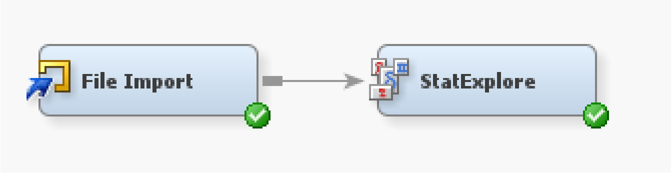


```
Sampling Summary

                                            Number of
 Type             Data Set               Observations

DATA        EMWS1.FIMPORT_train             9994
SAMPLE      EMWS1.Smpl_DATA                 4997
```

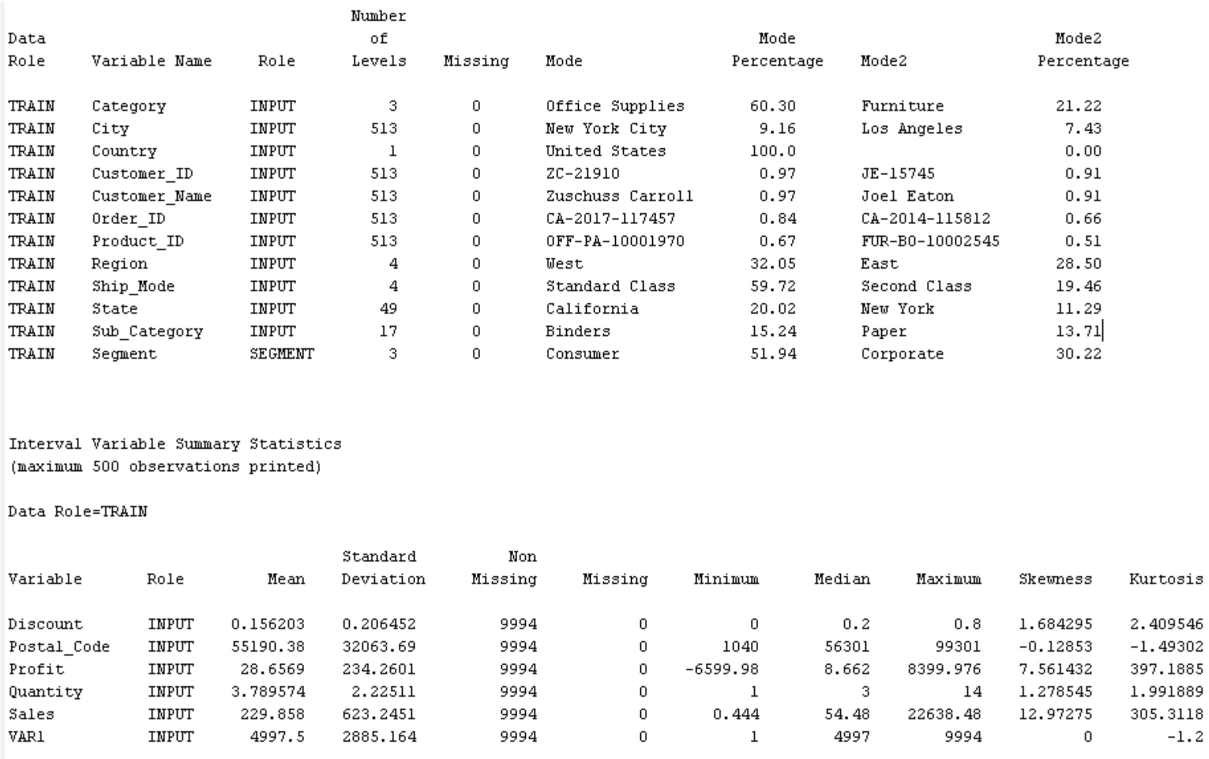
4.0 Explore

Exploratory data analysis is a crucial step in the data mining process in order for us to get an overall understanding of the dataset used before making any assumptions. It helps to detect any abnormalities in the data and to find interesting relationships between variables. SAS Enterprise Miner is used to explore data.

4.1 StatExplore Node

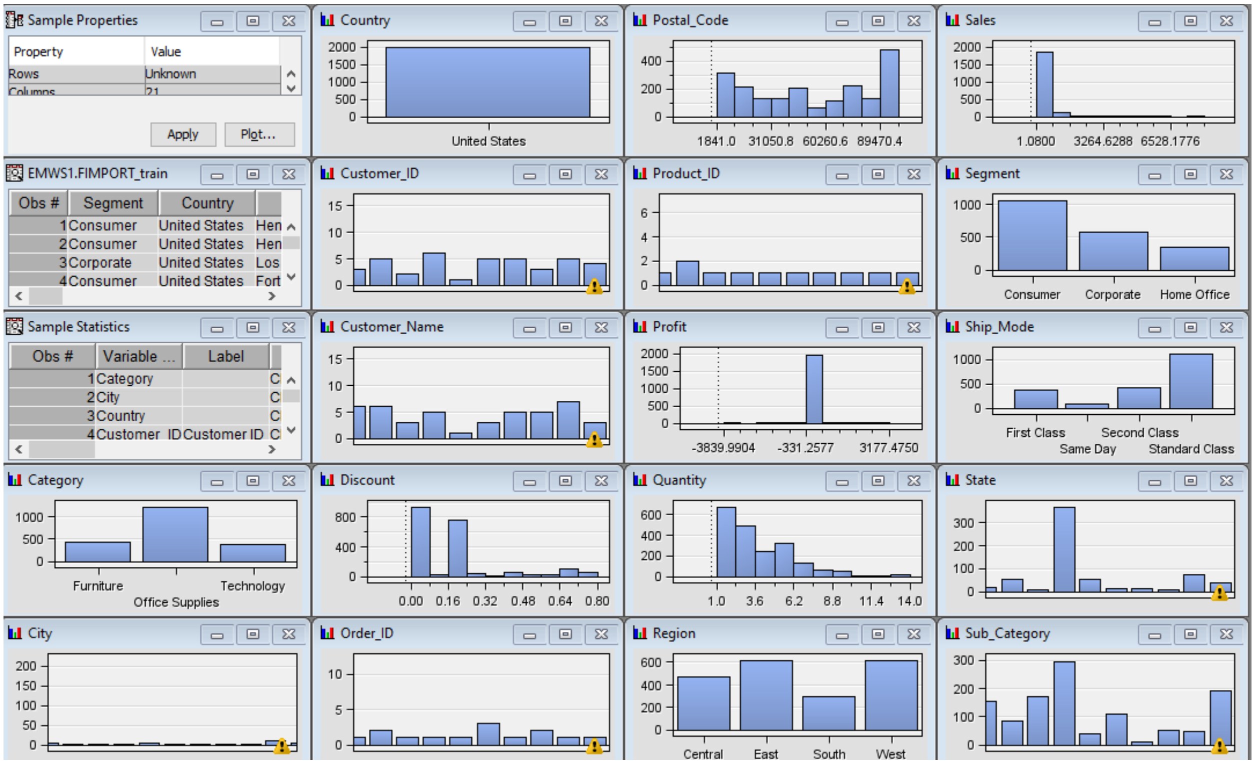


The StatExplore node was leveraged which allowed us to gain valuable insights into the underlying patterns and distributions within the dataset. It provides a comprehensive statistical analysis for each variable in the dataset.



```
                                     Number
Data                                   of                                         Mode                            Mode2
Role     Variable Name     Role      Levels    Missing    Mode                 Percentage    Mode2              Percentage

TRAIN    Category         INPUT           3         0      Office Supplies        60.30       Furniture            21.22
TRAIN    City             INPUT         513         0      New York City           9.16       Los Angeles           7.43
TRAIN    Country          INPUT           1         0      United States          100.0                             0.00
TRAIN    Customer_ID      INPUT         513         0      ZC-21910                0.97       JE-15745              0.91
TRAIN    Customer_Name    INPUT         513         0      Zuschuss Carroll        0.97       Joel Eaton            0.91
TRAIN    Order_ID         INPUT         513         0      CA-2017-117457          0.84       CA-2014-115812        0.66
TRAIN    Product_ID       INPUT         513         0      OFF-PA-10001970         0.67       FUR-BO-10002545       0.51
TRAIN    Region           INPUT           4         0      West                   32.05       East                 28.50
TRAIN    Ship_Mode        INPUT           4         0      Standard Class         59.72       Second Class         19.46
TRAIN    State            INPUT          49         0      California             20.02       New York             11.29
TRAIN    Sub_Category     INPUT          17         0      Binders                15.24       Paper                13.71
TRAIN    Segment          SEGMENT         3         0      Consumer               51.94       Corporate            30.22



Interval Variable Summary Statistics
(maximum 500 observations printed)

Data Role=TRAIN

                                    Standard           Non
Variable       Role          Mean   Deviation      Missing     Missing     Minimum      Median     Maximum    Skewness    Kurtosis

Discount       INPUT     0.156203    0.206452         9994           0           0         0.2         0.8    1.684295    2.409546
Postal_Code    INPUT     55190.38    32063.69         9994           0        1040       56301       99301    -0.12853    -1.49302
Profit         INPUT      28.6569    234.2601         9994           0    -6599.98       8.662    8399.976    7.561432    397.1885
Quantity       INPUT     3.789574     2.22511         9994           0           1           3          14    1.278545    1.991889
Sales          INPUT      229.858    623.2451         9994           0       0.444       54.48    22638.48    12.97275    305.3118
VAR1           INPUT       4997.5    2885.164         9994           0           1        4997        9994           0        -1.2
```

Based on the figure above, there are no missing data for each of the column. We are also able to know the mean, standard deviation, minimum, median, maximum, skewness, and kurtosis for each of the numeric variables. As for the non-numeric variables, it also provides the number of levels, mode, mode percentage, mode 2, and mode 2 percentage.



Sample Properties

Property	Value
Rows	Unknown
Columns	21

EMWS1.FIMPORT_train

Obs #	Segment	Country	
1	Consumer	United States	Hen
2	Consumer	United States	Hen
3	Corporate	United States	Los
4	Consumer	United States	Fort

Sample Statistics

Obs #	Variable ...	Label	
1	Category		C
2	City		C
3	Country		C
4	Customer_ID	Customer ID	C

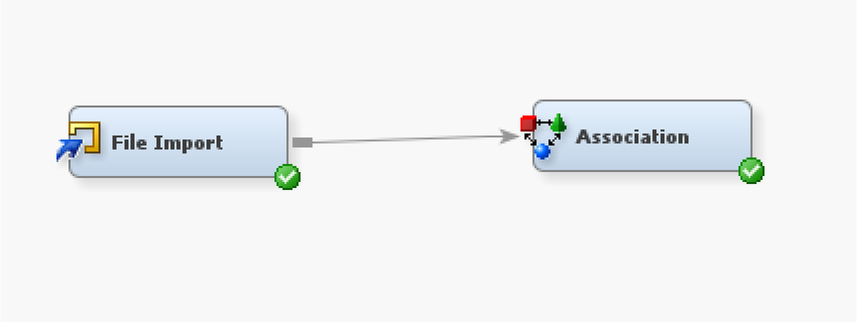
The figure above shows bar charts for each of the variables to provide an overview of the most frequent value and any abnormalities for the value in each variable.

4.2 Association Node

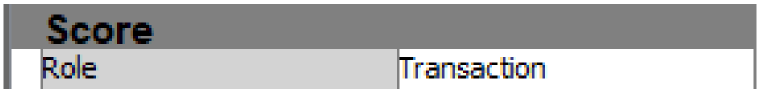
In this project, association rule mining was incorporated in the Explore stage. Association rule mining is generally used to identify relationships among a set of items in a database. It is implemented in order to extract interesting associations, frequent patterns, casual structures or correlations among sets of items, mainly in transactional databases.

SAS Enterprise Miner provided a node feature for Associaton in the Explore section, allowing users to carry out association rules mining directly upon desired dataset.

For our dataset from Superstore, association rule mining was applied to find interesting patterns between pair of items based on the rules generated by the node itself .



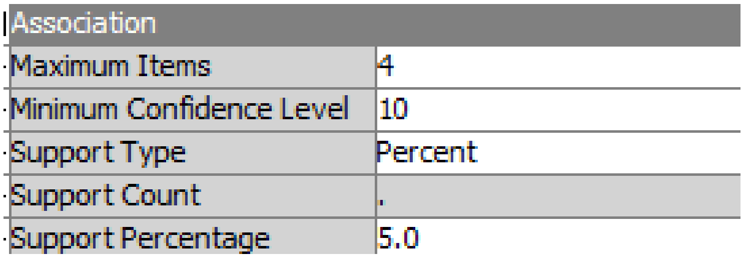
Association node was connected to data source, in this case the dataset imported using File Import node, for it to gain variables and their assigned roles to execute association rule mining.



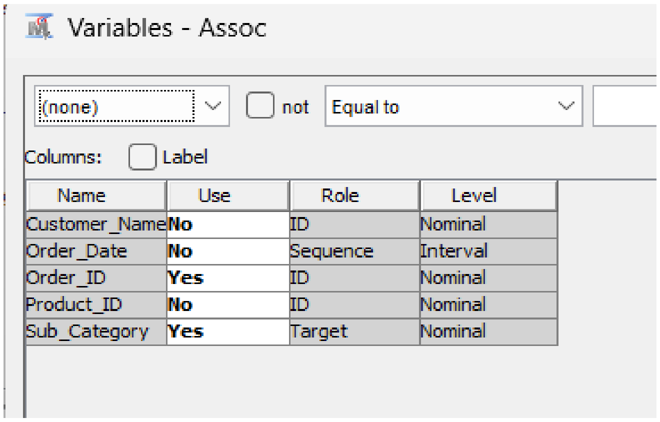
Score

Role	Transaction

In order for the Association node to run successfully, the role of the dataset should be configured as Transaction.



Association	
Maximum Items	4
Minimum Confidence Level	10
Support Type	Percent
Support Count	.
Support Percentage	5.0



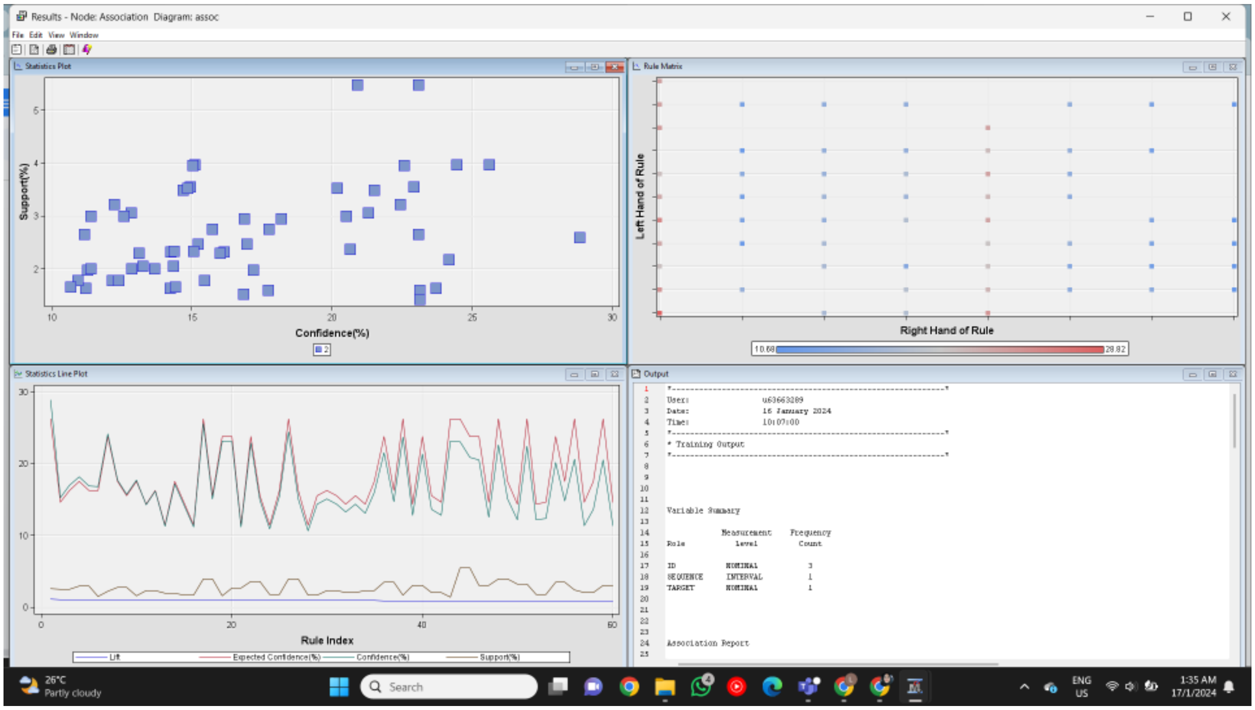
Variables - Assoc

(none) □ not Equal to

Columns: □ Label

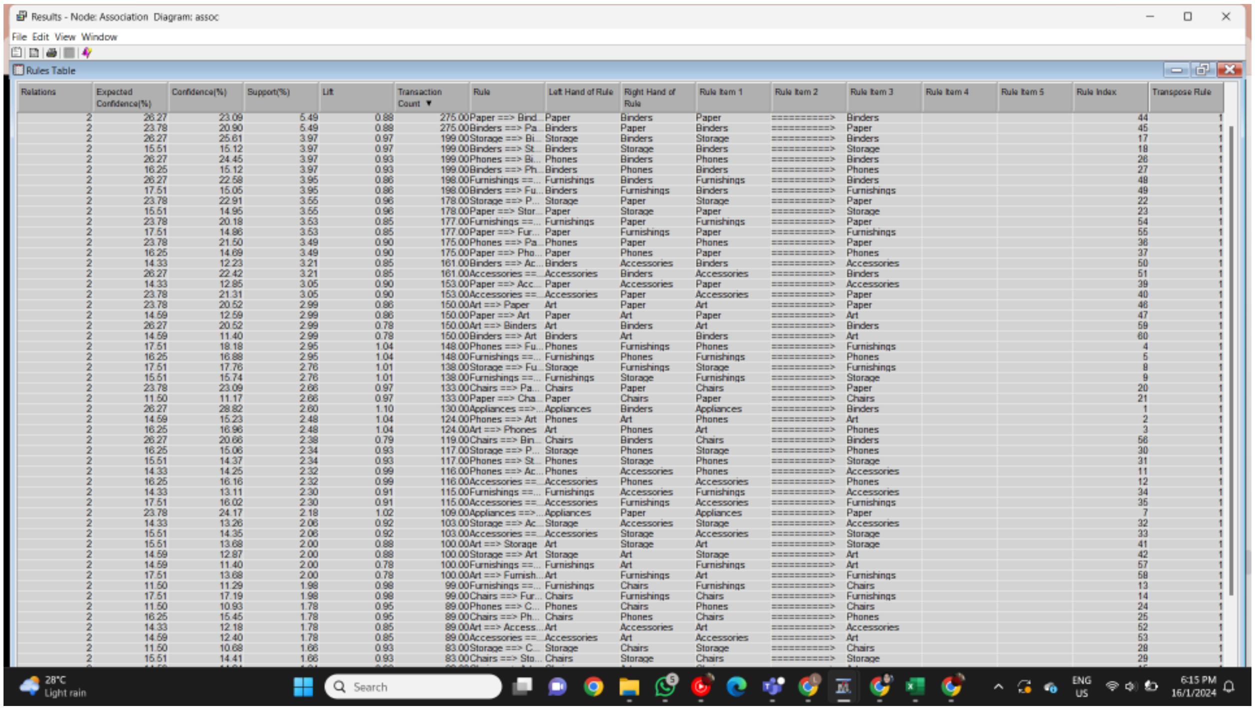
Name	Use	Role	Level
Customer_Name	**No**	ID	Nominal
Order_Date	**No**	Sequence	Interval
Order_ID	**Yes**	ID	Nominal
Product_ID	**No**	ID	Nominal
Sub_Category	**Yes**	Target	Nominal

Other than that, it is required for one ID role variable and one Target role variable to be used. Hence, Order_ID column was selected as ID role because it is unique identifier for every order in the dataset while Sub_Category was selected as target due to the large dataset and it represents the item that we want to focus on gaining inputs.

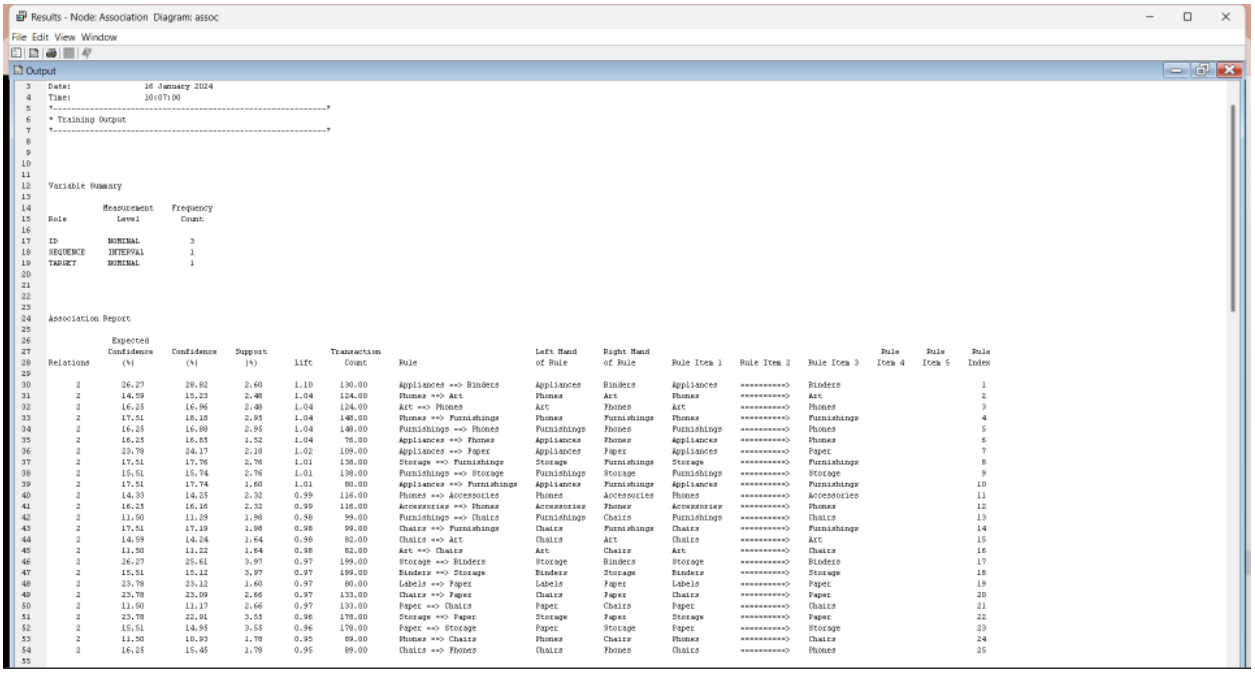


```
*------------------------------------------------------------*
User:                 u63663289
Date:                 16 January 2024
Time:                 10:07:00
*------------------------------------------------------------*
* Training Output
*------------------------------------------------------------*




Variable Summary

              Measurement    Frequency
Role             Level         Count

ID              NOMINAL          3
SEQUENCE        INTERVAL         1
TARGET          NOMINAL          1




Association Report

```

After running the node, these results were obtained, including a statistics plot for Confidence (%) and Support(%), and also Rule Matrix for RIght Hand of Rule and Left Hand of Rule.



Relations	Expected Confidence(%)	Confidence(%)	Support(%)	Lift	Transaction Count ▼	Rule	Left Hand of Rule	Right Hand of Rule	Rule Item 1	Rule Item 2	Rule Item 3	Rule Item 4	Rule Item 5	Rule Index	Transpose Rule
2	26.27	23.09	5.49	0.88	275.00	Paper ==> Bind...	Paper	Binders	Paper	==========>	Binders			44	1
2	23.78	20.90	5.49	0.88	275.00	Binders ==> Pa...	Binders	Paper	Binders	==========>	Paper			45	1
2	26.27	25.61	3.97	0.97	199.00	Storage ==> Bi...	Storage	Binders	Storage	==========>	Binders			17	1
2	15.51	15.12	3.97	0.97	199.00	Binders ==> St...	Binders	Storage	Binders	==========>	Storage			18	1
2	26.27	24.45	3.97	0.93	199.00	Phones ==> Bi...	Phones	Binders	Phones	==========>	Binders			26	1
2	16.25	15.12	3.97	0.93	199.00	Binders ==> Ph...	Binders	Phones	Binders	==========>	Phones			27	1
2	26.27	22.58	3.95	0.86	198.00	Furnishings ==...	Furnishings	Binders	Furnishings	==========>	Binders			48	1
2	17.51	15.05	3.95	0.86	198.00	Binders ==> Fu...	Binders	Furnishings	Binders	==========>	Furnishings			49	1
2	23.78	22.91	3.55	0.96	178.00	Storage ==> P...	Storage	Paper	Storage	==========>	Paper			22	1
2	15.51	14.95	3.55	0.96	178.00	Paper ==> Stor...	Paper	Storage	Paper	==========>	Storage			23	1
2	23.78	20.18	3.53	0.85	177.00	Furnishings ==...	Furnishings	Paper	Furnishings	==========>	Paper			54	1
2	17.51	14.86	3.53	0.85	177.00	Paper ==> Fur...	Paper	Furnishings	Paper	==========>	Furnishings			55	1
2	23.78	21.50	3.49	0.90	175.00	Phones ==> Pa...	Phones	Paper	Phones	==========>	Paper			36	1
2	16.25	14.69	3.49	0.90	175.00	Paper ==> Pho...	Paper	Phones	Paper	==========>	Phones			37	1
2	14.33	12.23	3.21	0.85	161.00	Binders ==> Ac...	Binders	Accessories	Binders	==========>	Accessories			50	1
2	26.27	22.42	3.21	0.85	161.00	Accessories ==...	Accessories	Binders	Accessories	==========>	Binders			51	1
2	14.33	12.85	3.05	0.90	153.00	Paper ==> Acc...	Paper	Accessories	Paper	==========>	Accessories			39	1
2	23.78	21.31	3.05	0.90	153.00	Accessories ==...	Accessories	Paper	Accessories	==========>	Paper			40	1
2	23.78	20.52	2.99	0.86	150.00	Art ==> Paper	Art	Paper	Art	==========>	Paper			46	1
2	14.59	12.59	2.99	0.86	150.00	Paper ==> Art	Paper	Art	Paper	==========>	Art			47	1
2	26.27	20.52	2.99	0.78	150.00	Art ==> Binders	Art	Binders	Art	==========>	Binders			59	1
2	14.59	11.40	2.99	0.78	150.00	Binders ==> Art	Binders	Art	Binders	==========>	Art			60	1
2	17.51	18.18	2.95	1.04	148.00	Phones ==> Fu...	Phones	Furnishings	Phones	==========>	Furnishings			4	1
2	16.25	16.88	2.95	1.04	148.00	Furnishings ==...	Furnishings	Phones	Furnishings	==========>	Phones			5	1
2	17.51	17.76	2.76	1.01	138.00	Storage ==> Fu...	Storage	Furnishings	Storage	==========>	Furnishings			8	1
2	15.51	15.74	2.76	1.01	138.00	Furnishings ==...	Furnishings	Storage	Furnishings	==========>	Storage			9	1
2	23.78	23.09	2.66	0.97	133.00	Chairs ==> Pa...	Chairs	Paper	Chairs	==========>	Paper			20	1
2	11.50	11.17	2.66	0.97	133.00	Paper ==> Cha...	Paper	Chairs	Paper	==========>	Chairs			21	1
2	26.27	28.82	2.60	1.10	130.00	Appliances ==>...	Appliances	Binders	Appliances	==========>	Binders			1	1
2	14.59	15.23	2.48	1.04	124.00	Phones ==> Art	Phones	Art	Phones	==========>	Art			2	1
2	16.25	16.96	2.48	1.04	124.00	Art ==> Phones	Art	Phones	Art	==========>	Phones			3	1
2	26.27	20.66	2.38	0.79	119.00	Chairs ==> Bin...	Chairs	Binders	Chairs	==========>	Binders			56	1
2	16.25	15.06	2.34	0.93	117.00	Storage ==> P...	Storage	Phones	Storage	==========>	Phones			30	1
2	15.51	14.37	2.34	0.93	117.00	Phones ==> St...	Phones	Storage	Phones	==========>	Storage			31	1
2	14.33	14.25	2.32	0.99	116.00	Phones ==> Ac...	Phones	Accessories	Phones	==========>	Accessories			11	1
2	16.25	16.16	2.32	0.99	116.00	Accessories ==...	Accessories	Phones	Accessories	==========>	Phones			12	1
2	14.33	13.11	2.30	0.91	115.00	Furnishings ==...	Furnishings	Accessories	Furnishings	==========>	Accessories			34	1
2	17.51	16.02	2.30	0.91	115.00	Accessories ==...	Accessories	Furnishings	Accessories	==========>	Furnishings			35	1
2	23.78	24.17	2.18	1.02	109.00	Appliances ==>...	Appliances	Paper	Appliances	==========>	Paper			7	1
2	14.33	13.26	2.06	0.92	103.00	Storage ==> Ac...	Storage	Accessories	Storage	==========>	Accessories			32	1
2	15.51	14.35	2.06	0.92	103.00	Accessories ==...	Accessories	Storage	Accessories	==========>	Storage			33	1
2	15.51	13.68	2.00	0.88	100.00	Art ==> Storage	Art	Storage	Art	==========>	Storage			41	1
2	14.59	12.87	2.00	0.88	100.00	Storage ==> Art	Storage	Art	Storage	==========>	Art			42	1
2	14.59	11.40	2.00	0.78	100.00	Furnishings ==...	Furnishings	Art	Furnishings	==========>	Art			57	1
2	17.51	13.68	2.00	0.78	100.00	Art ==> Furnish...	Art	Furnishings	Art	==========>	Furnishings			58	1
2	11.50	11.29	1.98	0.98	99.00	Furnishings ==...	Furnishings	Chairs	Furnishings	==========>	Chairs			13	1
2	17.51	17.19	1.98	0.98	99.00	Chairs ==> Fur...	Chairs	Furnishings	Chairs	==========>	Furnishings			14	1
2	11.50	10.93	1.78	0.95	89.00	Phones ==> C...	Phones	Chairs	Phones	==========>	Chairs			24	1
2	16.25	15.45	1.78	0.95	89.00	Chairs ==> Ph...	Chairs	Phones	Chairs	==========>	Phones			25	1
2	14.33	12.18	1.78	0.85	89.00	Art ==> Access...	Art	Accessories	Art	==========>	Accessories			52	1
2	14.59	12.40	1.78	0.85	89.00	Accessories ==...	Accessories	Art	Accessories	==========>	Art			53	1
2	11.50	10.68	1.66	0.93	83.00	Storage ==> C...	Storage	Chairs	Storage	==========>	Chairs			28	1
2	15.51	14.41	1.66	0.93	83.00	Chairs ==> Sto...	Chairs	Storage	Chairs	==========>	Storage			29	1



```
Date:              16 January 2024
Time:              10:07:00
*------------------------------------------------------------*
* Training Output
*------------------------------------------------------------*




Variable Summary

            Measurement    Frequency
Role           Level         Count

ID           NOMINAL           3
SEQUENCE     INTERVAL          1
TARGET       NOMINAL           1




Association Report

               Expected
              Confidence   Confidence   Support              Transaction                               Left Hand     Right Hand                                                    Rule     Rule     Rule
Relations        (%)          (%)         (%)        Lift       Count       Rule                        of Rule       of Rule       Rule Item 1   Rule Item 2   Rule Item 3     Item 4   Item 5   Index

    2           26.27        28.82        2.60       1.10      130.00       Appliances ==> Binders      Appliances    Binders       Appliances    ==========>   Binders                             1
    2           14.59        15.23        2.48       1.04      124.00       Phones ==> Art              Phones        Art           Phones        ==========>   Art                                 2
    2           16.25        16.96        2.48       1.04      124.00       Art ==> Phones              Art           Phones        Art           ==========>   Phones                              3
    2           17.51        18.18        2.95       1.04      148.00       Phones ==> Furnishings      Phones        Furnishings   Phones        ==========>   Furnishings                         4
    2           16.25        16.88        2.95       1.04      148.00       Furnishings ==> Phones      Furnishings   Phones        Furnishings   ==========>   Phones                              5
    2           16.25        16.85        1.52       1.04       76.00       Appliances ==> Phones       Appliances    Phones        Appliances    ==========>   Phones                              6
    2           23.78        24.17        2.18       1.02      109.00       Appliances ==> Paper        Appliances    Paper         Appliances    ==========>   Paper                               7
    2           17.51        17.76        2.76       1.01      138.00       Storage ==> Furnishings     Storage       Furnishings   Storage       ==========>   Furnishings                         8
    2           15.51        15.74        2.76       1.01      138.00       Furnishings ==> Storage     Furnishings   Storage       Furnishings   ==========>   Storage                             9
    2           17.51        17.74        1.60       1.01       80.00       Appliances ==> Furnishings  Appliances    Furnishings   Appliances    ==========>   Furnishings                        10
    2           14.33        14.25        2.32       0.99      116.00       Phones ==> Accessories      Phones        Accessories   Phones        ==========>   Accessories                        11
    2           16.25        16.16        2.32       0.99      116.00       Accessories ==> Phones      Accessories   Phones        Accessories   ==========>   Phones                             12
    2           11.50        11.29        1.98       0.98       99.00       Furnishings ==> Chairs      Furnishings   Chairs        Furnishings   ==========>   Chairs                             13
    2           17.51        17.19        1.98       0.98       99.00       Chairs ==> Furnishings      Chairs        Furnishings   Chairs        ==========>   Furnishings                        14
    2           14.59        14.24        1.64       0.98       82.00       Chairs ==> Art              Chairs        Art           Chairs        ==========>   Art                                15
    2           11.50        11.22        1.64       0.98       82.00       Art ==> Chairs              Art           Chairs        Art           ==========>   Chairs                             16
    2           26.27        25.61        3.97       0.97      199.00       Storage ==> Binders         Storage       Binders       Storage       ==========>   Binders                            17
    2           15.51        15.12        3.97       0.97      199.00       Binders ==> Storage         Binders       Storage       Binders       ==========>   Storage                            18
    2           23.78        23.12        1.60       0.97       80.00       Labels ==> Paper            Labels        Paper         Labels        ==========>   Paper                              19
    2           23.78        23.09        2.66       0.97      133.00       Chairs ==> Paper            Chairs        Paper         Chairs        ==========>   Paper                              20
    2           11.50        11.17        2.66       0.97      133.00       Paper ==> Chairs            Paper         Chairs        Paper         ==========>   Chairs                             21
    2           23.78        22.91        3.55       0.96      178.00       Storage ==> Paper           Storage       Paper         Storage       ==========>   Paper                              22
    2           15.51        14.95        3.55       0.96      178.00       Paper ==> Storage           Paper         Storage       Paper         ==========>   Storage                            23
    2           11.50        10.93        1.78       0.95       89.00       Phones ==> Chairs           Phones        Chairs        Phones        ==========>   Chairs                             24
    2           16.25        15.45        1.78       0.95       89.00       Chairs ==> Phones           Chairs        Phones        Chairs        ==========>   Phones                             25
```

Relations: The rule index or identifier for each association rule.

Expected Confidence (%): The expected confidence represents the likelihood that the rule is correct based on the data. It ranges from 11.50% to 26.27%.

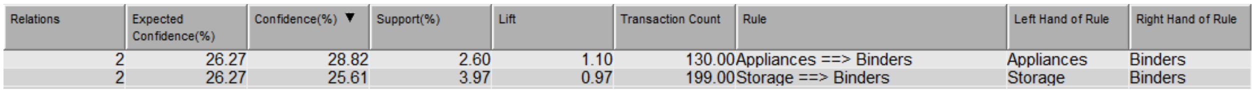
Confidence (%): The actual observed confidence in the data for each rule. It indicates the probability of the right-hand side item occurring given the left-hand side items.

Support (%): The percentage of transactions that contain the items on the left-hand side of the rule.

Lift: Lift is a measure of how much more likely the right-hand side item is to be purchased when the left-hand side items are purchased, compared to when it is purchased on its own.

Transaction Count: The number of transactions that support the rule.

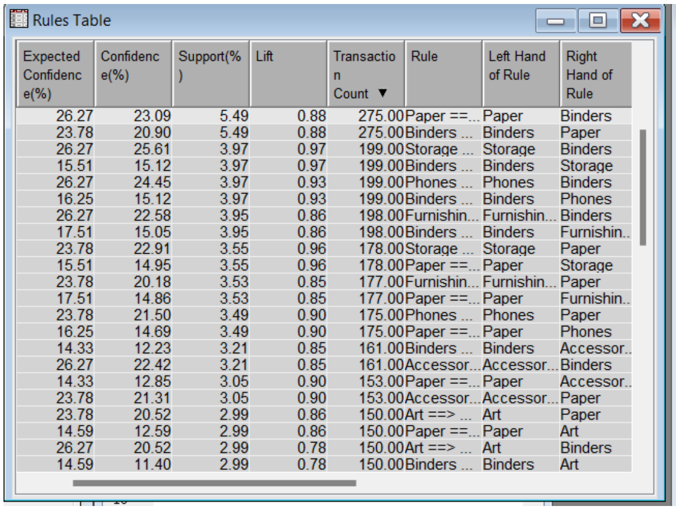
Rule: The association rule in terms of items. It shows the left-hand side items, the arrow (==>) indicating the association, and the right-hand side item.



Relations	Expected Confidence(%)	Confidence(%) ▼	Support(%)	Lift	Transaction Count	Rule	Left Hand of Rule	Right Hand of Rule
2	26.27	28.82	2.60	1.10	130.00	Appliances ==> Binders	Appliances	Binders
2	26.27	25.61	3.97	0.97	199.00	Storage ==> Binders	Storage	Binders

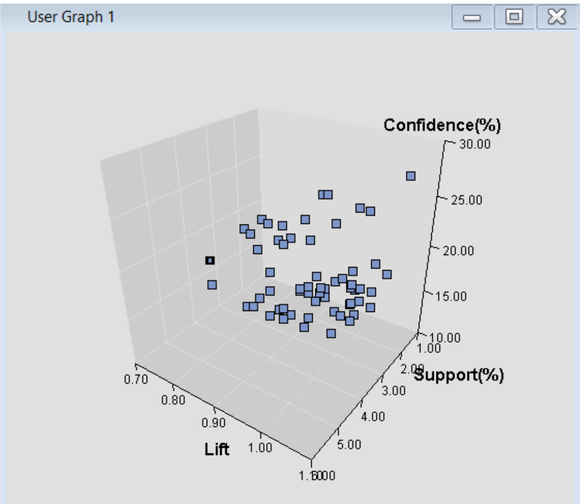
For Rule Appliances==>Binders, there is an observed 28.82% chance that Binders will also be purchased when Appliances being bought, and the likelihood they would always being bought together is at 1.10.

For Rule Storage==>Binders, there is an observed 25.61% chance that Binders will also be purchased when Storage being bought, and the likelihood they would always being bought together is at 0.97.



Rules Table

Expected Confidence(%)	Confidence(%)	Support(%)	Lift	Transaction Count ▼	Rule	Left Hand of Rule	Right Hand of Rule
26.27	23.09	5.49	0.88	275.00	Paper ==...	Paper	Binders
23.78	20.90	5.49	0.88	275.00	Binders ...	Binders	Paper
26.27	25.61	3.97	0.97	199.00	Storage ...	Storage	Binders
15.51	15.12	3.97	0.97	199.00	Binders ...	Binders	Storage
26.27	24.45	3.97	0.93	199.00	Phones ...	Phones	Binders
16.25	15.12	3.97	0.93	199.00	Binders ...	Binders	Phones
26.27	22.58	3.95	0.86	198.00	Furnishin...	Furnishin...	Binders
17.51	15.05	3.95	0.86	198.00	Binders ...	Binders	Furnishin..
23.78	22.91	3.55	0.96	178.00	Storage ...	Storage	Paper
15.51	14.95	3.55	0.96	178.00	Paper ==...	Paper	Storage
23.78	20.18	3.53	0.85	177.00	Furnishin...	Furnishin...	Paper
17.51	14.86	3.53	0.85	177.00	Paper ==...	Paper	Furnishin..
23.78	21.50	3.49	0.90	175.00	Phones ...	Phones	Paper
16.25	14.69	3.49	0.90	175.00	Paper ==...	Paper	Phones
14.33	12.23	3.21	0.85	161.00	Binders ...	Binders	Accessor..
26.27	22.42	3.21	0.85	161.00	Accessor...	Accessor...	Binders
14.33	12.85	3.05	0.90	153.00	Paper ==...	Paper	Accessor..
23.78	21.31	3.05	0.90	153.00	Accessor...	Accessor...	Paper
23.78	20.52	2.99	0.86	150.00	Art ==> ...	Art	Paper
14.59	12.59	2.99	0.86	150.00	Paper ==...	Paper	Art
26.27	20.52	2.99	0.78	150.00	Art ==> ...	Art	Binders
14.59	11.40	2.99	0.78	150.00	Binders ...	Binders	Art

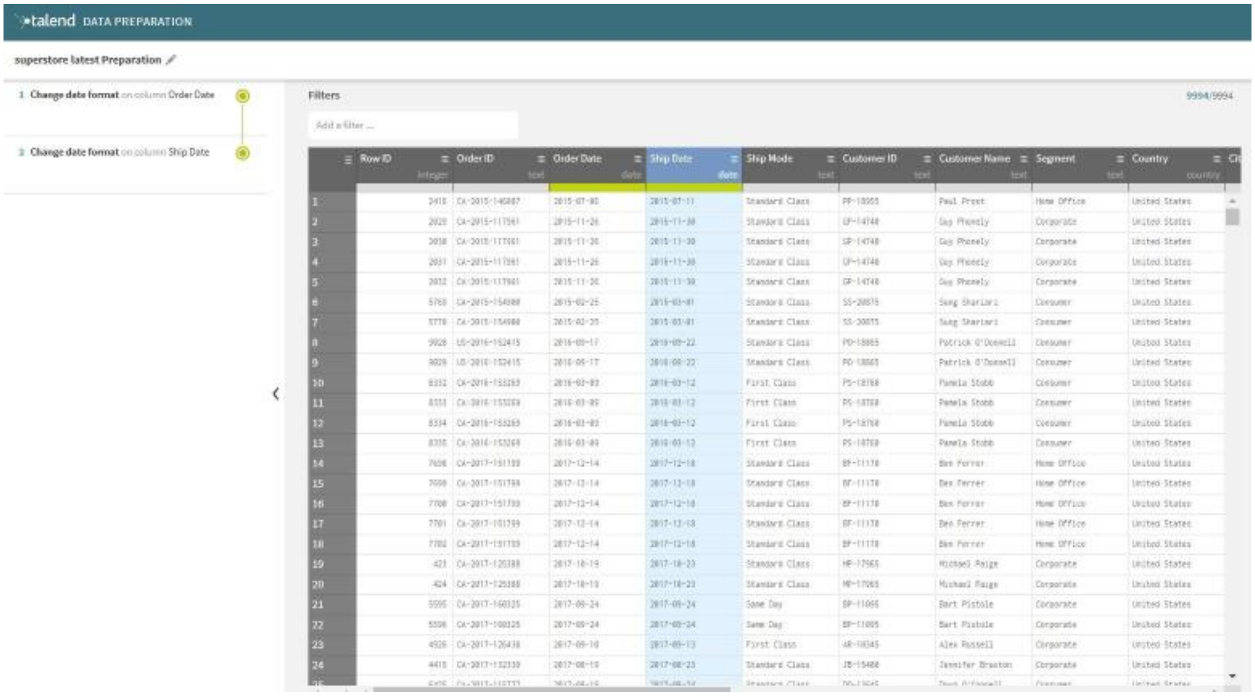


One of the most interesting findings is the insight gain from Rule Paper==>Binders, where the lift is 0.88 which is not the highest in the dataset, but it is considered great as the number of transactions is 275, far above other pairs of items. The similar thing occurs to Rule Binders==>Paper, although the confidence is a bit lower compared to the previous rule at 20.90%.

5.0 Modify

5.1 Data cleaning

5.1.1 Date formatting



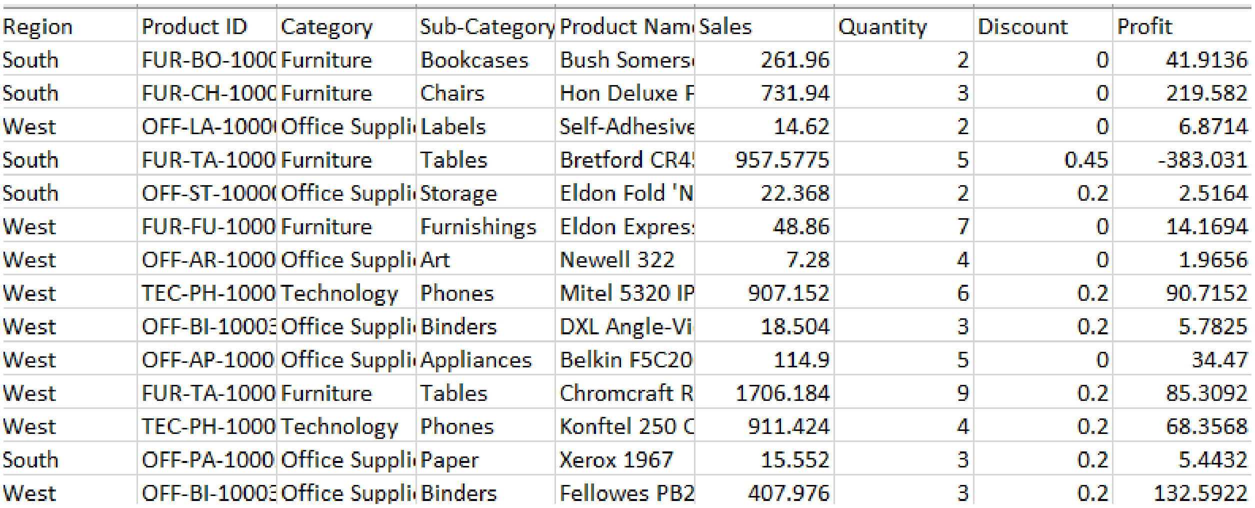
1 Change date format on column Order Date

2 Change date format on column Ship Date

	Row ID	Order ID	Order Date	Ship Date	Ship Mode	Customer ID	Customer Name	Segment	Country
	integer	text	date	date	text	text	text	text	country
1	2416	CA-2015-140887	2015-07-08	2015-07-11	Standard Class	PP-18955	Paul Prost	Home Office	United States
2	2029	CA-2015-117961	2015-11-26	2015-11-30	Standard Class	GP-14740	Guy Phonely	Corporate	United States
3	2030	CA-2015-117961	2015-11-26	2015-11-30	Standard Class	GP-14740	Guy Phonely	Corporate	United States
4	2031	CA-2015-117961	2015-11-26	2015-11-30	Standard Class	GP-14740	Guy Phonely	Corporate	United States
5	2032	CA-2015-117961	2015-11-26	2015-11-30	Standard Class	GP-14740	Guy Phonely	Corporate	United States
6	5769	CA-2015-154900	2015-02-25	2015-03-01	Standard Class	SS-20875	Sung Shariari	Consumer	United States
7	5770	CA-2015-154900	2015-02-25	2015-03-01	Standard Class	SS-20875	Sung Shariari	Consumer	United States
8	9928	US-2016-152415	2016-09-17	2016-09-22	Standard Class	PO-18865	Patrick O'Donnell	Consumer	United States
9	9929	US-2016-152415	2016-09-17	2016-09-22	Standard Class	PO-18865	Patrick O'Donnell	Consumer	United States
10	8332	CA-2016-153269	2016-03-09	2016-03-12	First Class	PS-18760	Pamela Stobb	Consumer	United States
11	8333	CA-2016-153269	2016-03-09	2016-03-12	First Class	PS-18760	Pamela Stobb	Consumer	United States
12	8334	CA-2016-153269	2016-03-09	2016-03-12	First Class	PS-18760	Pamela Stobb	Consumer	United States
13	8335	CA-2016-153269	2016-03-09	2016-03-12	First Class	PS-18760	Pamela Stobb	Consumer	United States
14	7698	CA-2017-151799	2017-12-14	2017-12-18	Standard Class	BF-11170	Ben Ferrer	Home Office	United States
15	7699	CA-2017-151799	2017-12-14	2017-12-18	Standard Class	BF-11170	Ben Ferrer	Home Office	United States
16	7700	CA-2017-151799	2017-12-14	2017-12-18	Standard Class	BF-11170	Ben Ferrer	Home Office	United States
17	7701	CA-2017-151799	2017-12-14	2017-12-18	Standard Class	BF-11170	Ben Ferrer	Home Office	United States
18	7702	CA-2017-151799	2017-12-14	2017-12-18	Standard Class	BF-11170	Ben Ferrer	Home Office	United States
19	423	CA-2017-125388	2017-10-19	2017-10-23	Standard Class	MP-17965	Michael Paige	Corporate	United States
20	424	CA-2017-125388	2017-10-19	2017-10-23	Standard Class	MP-17965	Michael Paige	Corporate	United States
21	5595	CA-2017-160325	2017-09-24	2017-09-24	Same Day	BP-11095	Bart Pistole	Corporate	United States
22	5596	CA-2017-160325	2017-09-24	2017-09-24	Same Day	BP-11095	Bart Pistole	Corporate	United States
23	4926	CA-2017-126438	2017-09-10	2017-09-13	First Class	AR-10345	Alex Russell	Corporate	United States
24	4415	CA-2017-132339	2017-08-19	2017-08-23	Standard Class	JB-15400	Jennifer Braxton	Corporate	United States
25	[illegible]	[illegible]	[illegible]	[illegible]	Standard Class	[illegible]	Doug O'Connell	Consumer	United States

Columns 'Ship Date' and 'Order Date' were originally in string format even though they were dates in context. Using Talend Data Preparation, after loading the dataset, the current date format of the 'Ship Date' column is identified to inform subsequent actions. Using the "Change Type" or "Convert Type" option, the 'Ship Date' column is converted to the desired date type, with the appropriate format selected. This process is repeated for the 'Order Date' column. Validation and adjustment are crucial, ensuring that both columns display the correct date format.

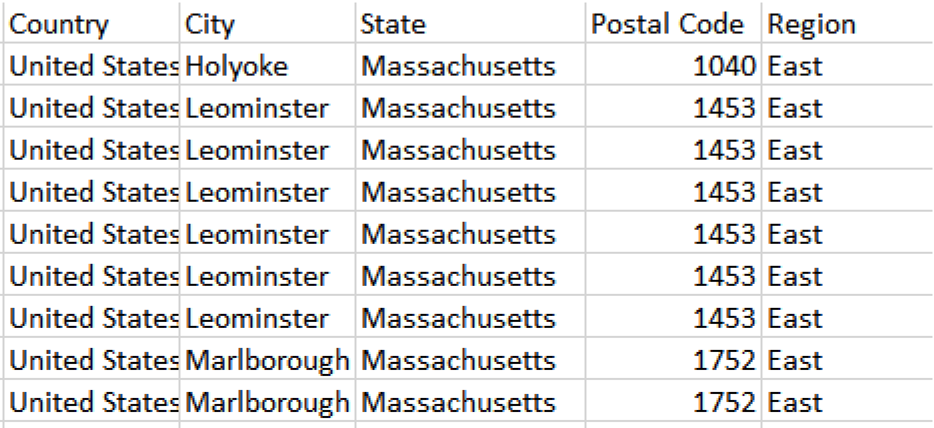
5.1.2 Profit and sales columns formatting



Region	Product ID	Category	Sub-Category	Product Nam	Sales	Quantity	Discount	Profit
South	FUR-BO-100(	Furniture	Bookcases	Bush Somers	261.96	2	0	41.9136
South	FUR-CH-100(	Furniture	Chairs	Hon Deluxe F	731.94	3	0	219.582
West	OFF-LA-1000	Office Suppli	Labels	Self-Adhesive	14.62	2	0	6.8714
South	FUR-TA-1000	Furniture	Tables	Bretford CR4	957.5775	5	0.45	-383.031
South	OFF-ST-1000	Office Suppli	Storage	Eldon Fold 'N	22.368	2	0.2	2.5164
West	FUR-FU-1000	Furniture	Furnishings	Eldon Expres	48.86	7	0	14.1694
West	OFF-AR-1000	Office Suppli	Art	Newell 322	7.28	4	0	1.9656
West	TEC-PH-1000	Technology	Phones	Mitel 5320 IP	907.152	6	0.2	90.7152
West	OFF-BI-1000З	Office Suppli	Binders	DXL Angle-Vi	18.504	3	0.2	5.7825
West	OFF-AP-1000	Office Suppli	Appliances	Belkin F5C20	114.9	5	0	34.47
West	FUR-TA-1000	Furniture	Tables	Chromcraft R	1706.184	9	0.2	85.3092
West	TEC-PH-1000	Technology	Phones	Konftel 250 C	911.424	4	0.2	68.3568
South	OFF-PA-1000	Office Suppli	Paper	Xerox 1967	15.552	3	0.2	5.4432
West	OFF-BI-1000З	Office Suppli	Binders	Fellowes PB2	407.976	3	0.2	132.5922

To standardize the numerical value of profit and sales, both columns were formatted to keep only up to 2 decimal points using Talend Data Preparation.

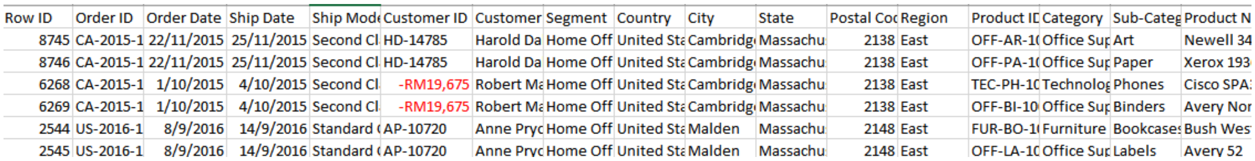
5.1.3 Postal code formatting



Country	City	State	Postal Code	Region
United States	Holyoke	Massachusetts	1040	East
United States	Leominster	Massachusetts	1453	East
United States	Leominster	Massachusetts	1453	East
United States	Leominster	Massachusetts	1453	East
United States	Leominster	Massachusetts	1453	East
United States	Leominster	Massachusetts	1453	East
United States	Leominster	Massachusetts	1453	East
United States	Marlborough	Massachusetts	1752	East
United States	Marlborough	Massachusetts	1752	East

There were 438 rows of data that consisted of invalid postal codes with only 4 digits. After further inspection, the leading zero was removed from the postal code causing it to have inaccurate values. Therefore, zero was added to every row with 4 digits of postal code and formatted to not automatically remove the leading zero.

5.1.4 Customer ID formatting



Row ID	Order ID	Order Date	Ship Date	Ship Mode	Customer ID	Customer	Segment	Country	City	State	Postal Coc	Region	Product ID	Category	Sub-Categ	Product N
8745	CA-2015-1	22/11/2015	25/11/2015	Second Cl	HD-14785	Harold Da	Home Off	United Sta	Cambridg	Massachu	2138	East	OFF-AR-1	Office Sup	Art	Newell 34
8746	CA-2015-1	22/11/2015	25/11/2015	Second Cl	HD-14785	Harold Da	Home Off	United Sta	Cambridg	Massachu	2138	East	OFF-PA-1	Office Sup	Paper	Xerox 193
6268	CA-2015-1	1/10/2015	4/10/2015	Second Cl	-RM19,675	Robert Ma	Home Off	United Sta	Cambridg	Massachu	2138	East	TEC-PH-10	Technolog	Phones	Cisco SPA
6269	CA-2015-1	1/10/2015	4/10/2015	Second Cl	-RM19,675	Robert Ma	Home Off	United Sta	Cambridg	Massachu	2138	East	OFF-BI-10	Office Sup	Binders	Avery Nor
2544	US-2016-1	8/9/2016	14/9/2016	Standard	AP-10720	Anne Pryc	Home Off	United Sta	Malden	Massachu	2148	East	FUR-BO-1	Furniture	Bookcases	Bush Wes
2545	US-2016-1	8/9/2016	14/9/2016	Standard	AP-10720	Anne Pryc	Home Off	United Sta	Malden	Massachu	2148	East	OFF-LA-10	Office Sup	Labels	Avery 52

After exporting the data from Talend Data Preparation, some of the Customer ID row's format is converted into currency automatically. The format for each of the affected rows were change back to numerical.

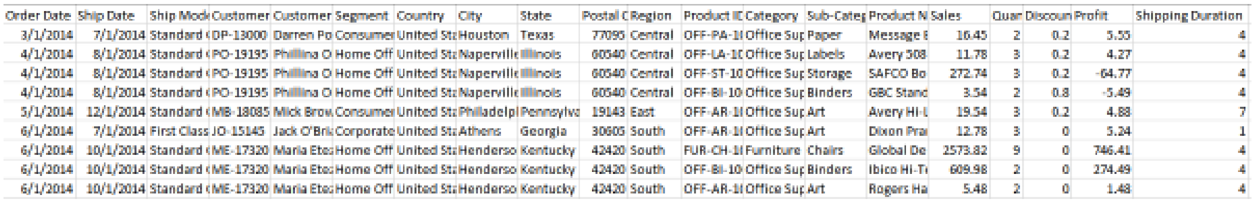
5.2 Generate New Columns

Based on the dataset, new columns were generated using Featuretools by Python. A new column for shipping duration was created by finding the difference between the ship date with the order date. The following is the Python code to generate the new column:

```
!pip install featuretools==0.27.0
!pip install pandas

import featuretools as ft
import pandas as pd

# Load the dataset
data = pd.read_csv('superstore.csv')

# Convert 'Order Date' and 'Ship Date' to datetime objects
data['Order Date'] = pd.to_datetime(data['Order Date'])
data['Ship Date'] = pd.to_datetime(data['Ship Date'])
# Calculate the shipping duration
data['Shipping Duration'] = (data['Ship Date'] - data['Order Date']).dt.days
```

```
data.to_csv('Superstore(20).csv', index=False)
```

The newly made column ‘Shipping duration”:

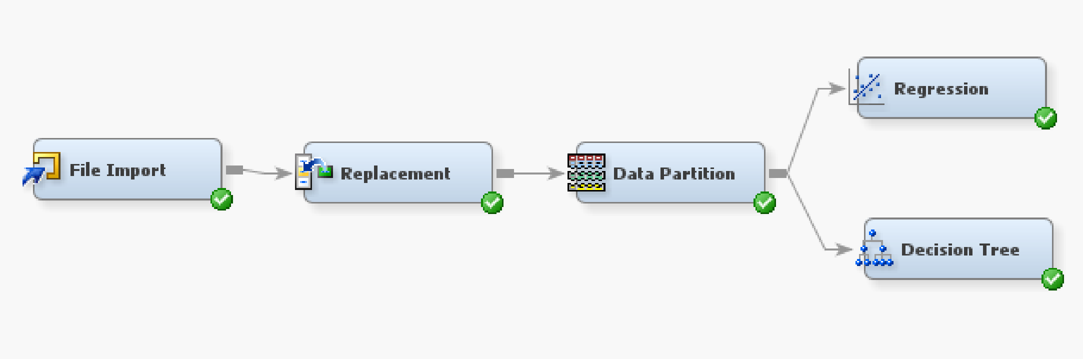


Order Date	Ship Date	Ship Mode	Customer	Customer	Segment	Country	City	State	Postal C	Region	Product ID	Category	Sub-Categ	Product N	Sales	Quar	Discoun	Profit	Shipping Duration
3/1/2014	7/1/2014	Standard (	DP-13000	Darren Po	Consumer	United Sta	Houston	Texas	77095	Central	OFF-PA-1(	Office Sup	Paper	Message E	16.45	2	0.2	5.55	4
4/1/2014	8/1/2014	Standard (	PO-19195	Phillina O	Home Off	United Sta	Naperville	Illinois	60540	Central	OFF-LA-10	Office Sup	Labels	Avery 508	11.78	3	0.2	4.27	4
4/1/2014	8/1/2014	Standard (	PO-19195	Phillina O	Home Off	United Sta	Naperville	Illinois	60540	Central	OFF-ST-10	Office Sup	Storage	SAFCO Bo	272.74	3	0.2	-64.77	4
4/1/2014	8/1/2014	Standard (	PO-19195	Phillina O	Home Off	United Sta	Naperville	Illinois	60540	Central	OFF-BI-10	Office Sup	Binders	GBC Stand	3.54	2	0.8	-5.49	4
5/1/2014	12/1/2014	Standard (	MB-18085	Mick Brow	Consumer	United Sta	Philadelpl	Pennsylva	19143	East	OFF-AR-1(	Office Sup	Art	Avery Hi-L	19.54	3	0.2	4.88	7
6/1/2014	7/1/2014	First Class	JO-15145	Jack O'Bri	Corporate	United Sta	Athens	Georgia	30605	South	OFF-AR-1(	Office Sup	Art	Dixon Pra	12.78	3	0	5.24	1
6/1/2014	10/1/2014	Standard (	ME-17320	Maria Ete:	Home Off	United Sta	Henderso	Kentucky	42420	South	FUR-CH-1(	Furniture	Chairs	Global De	2573.82	9	0	746.41	4
6/1/2014	10/1/2014	Standard (	ME-17320	Maria Ete:	Home Off	United Sta	Henderso	Kentucky	42420	South	OFF-BI-10	Office Sup	Binders	Ibico Hi-Te	609.98	2	0	274.49	4
6/1/2014	10/1/2014	Standard (	ME-17320	Maria Ete:	Home Off	United Sta	Henderso	Kentucky	42420	South	OFF-AR-1(	Office Sup	Art	Rogers Ha	5.48	2	0	1.48	4

6.0 Model

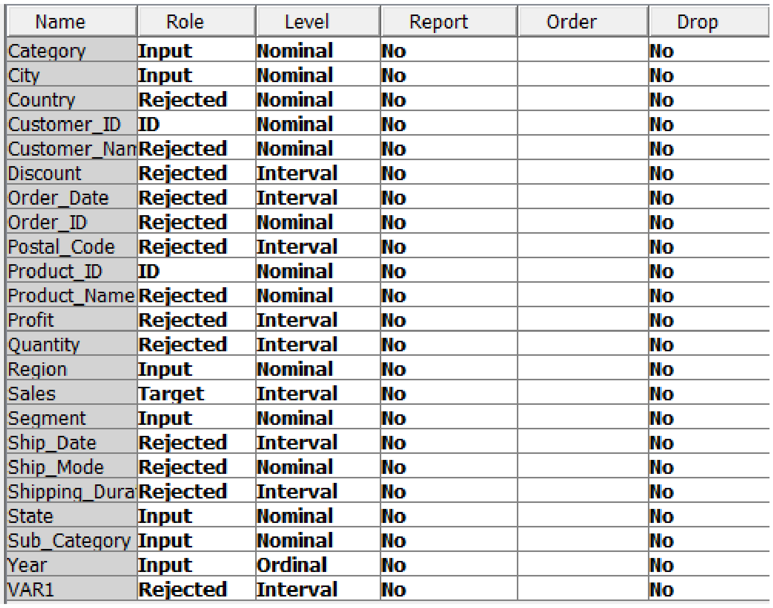
The data mining procedure for the retail sector in SAS Enterprise Miner is managed via a well-crafted process flow diagram. This schematic, created by selecting nodes from a special toolbar, follows the SEMMA (Sample, Explore, Modify, Model, and Assess) process. The investigation centers on a clean dataset reflecting retail transactions and consumer interactions within a retail business.

The two key methods regression analysis (supervised learning) and decision tree analysis (supervised learning) are incorporated into the process flow diagram. Every method is carefully used to glean insightful information on consumer behavior, preferences, and general market dynamics.



The cleaned dataset is sent to the SAS data sources. The file is then dragged into the diagram area after that. The imported file is attached to the Replacement node so that we can enter the replacement value or statistic that we want to use for specific levels of a class variable. By default, unknown levels are replaced with the value entered in the Unknown Levels field. Next, the Replacement node is linked to the Data Partition node. A method for evaluating the degree of model generalization quality is being developed in Data Partition. A portion of the data used to fit the model in its initial stages is called the training data set. The remainder is often split into test and validation data and reserved for empirical validation.

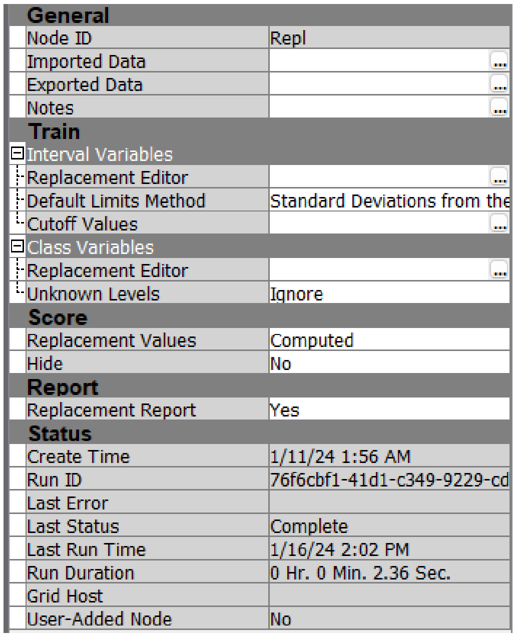
Sales are set as target:



Name	Role	Level	Report	Order	Drop
Category	Input	Nominal	No		No
City	Input	Nominal	No		No
Country	Rejected	Nominal	No		No
Customer_ID	ID	Nominal	No		No
Customer_Nan	Rejected	Nominal	No		No
Discount	Rejected	Interval	No		No
Order_Date	Rejected	Interval	No		No
Order_ID	Rejected	Nominal	No		No
Postal_Code	Rejected	Interval	No		No
Product_ID	ID	Nominal	No		No
Product_Name	Rejected	Nominal	No		No
Profit	Rejected	Interval	No		No
Quantity	Rejected	Interval	No		No
Region	Input	Nominal	No		No
Sales	Target	Interval	No		No
Segment	Input	Nominal	No		No
Ship_Date	Rejected	Interval	No		No
Ship_Mode	Rejected	Nominal	No		No
Shipping_Dura	Rejected	Interval	No		No
State	Input	Nominal	No		No
Sub_Category	Input	Nominal	No		No
Year	Input	Ordinal	No		No
VAR1	Rejected	Interval	No		No

6.1 Replacement node properties

In the Replacement node, we set the default limits method as Standard Deviations from the Mean method. The Standard Deviations from the Mean method filter values that are greater than or equal to n standard deviations from the mean.

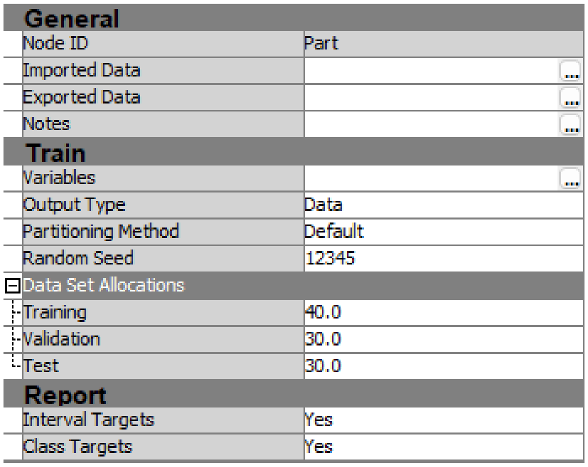


General	
Node ID	Repl
Imported Data	...
Exported Data	...
Notes	...
Train	
⊟Interval Variables	
Replacement Editor	...
Default Limits Method	Standard Deviations from the
Cutoff Values	...
⊟Class Variables	
Replacement Editor	...
Unknown Levels	Ignore
Score	
Replacement Values	Computed
Hide	No
Report	
Replacement Report	Yes
Status	
Create Time	1/11/24 1:56 AM
Run ID	76f6cbf1-41d1-c349-9229-cd
Last Error	
Last Status	Complete
Last Run Time	1/16/24 2:02 PM
Run Duration	0 Hr. 0 Min. 2.36 Sec.
Grid Host	
User-Added Node	No

6.2 Data Partition node properties

In the Data Partition node, there are the data set allocations in the Train properties. We set the value for

- Training: 40.0
- Validation: 30.0
- Test: 30.0

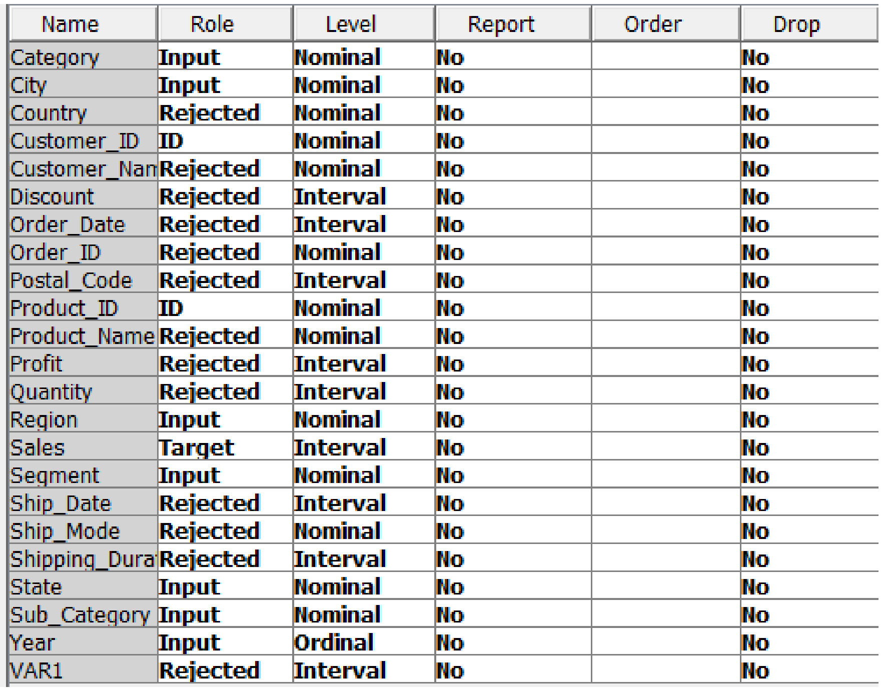


General	
Node ID	Part
Imported Data	...
Exported Data	...
Notes	...
Train	
Variables	...
Output Type	Data
Partitioning Method	Default
Random Seed	12345
⊟Data Set Allocations	
Training	40.0
Validation	30.0
Test	30.0
Report	
Interval Targets	Yes
Class Targets	Yes

These values specify the percentage of observations that want to allocate to the training data set.

6.3 Regression

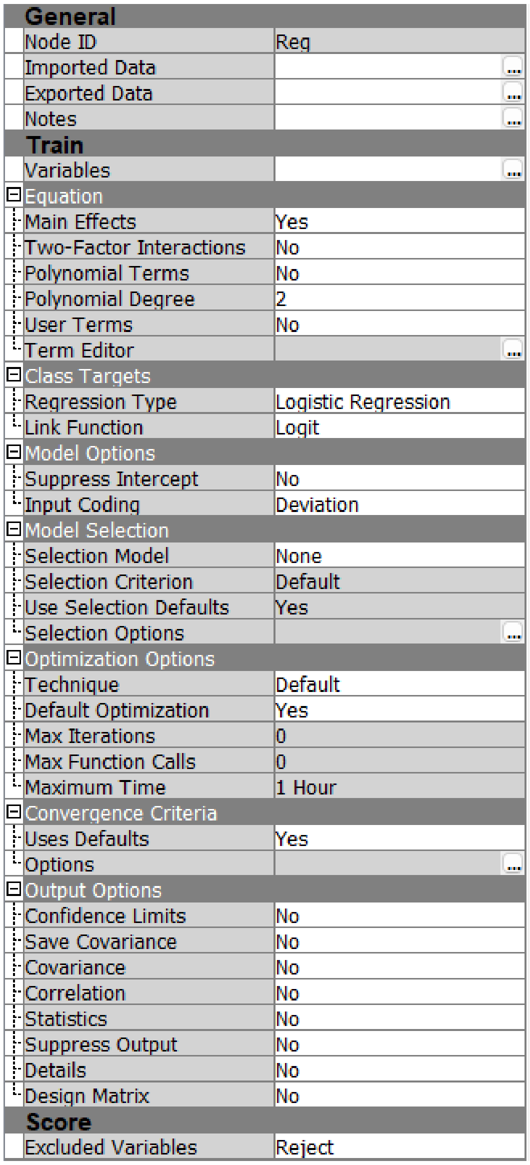
Logistic and linear regression models are fitted to a previous data set using the Regression node in the SAS Enterprise Miner process flow. To try and estimate the value of an interval target, linear regression uses a linear function of one or more independent inputs. Logistic regression is a statistical technique that estimates the probability of a binary or ordinal target experiencing a desired event, given one or more independent inputs.



Name	Role	Level	Report	Order	Drop
Category	Input	Nominal	No		No
City	Input	Nominal	No		No
Country	Rejected	Nominal	No		No
Customer_ID	ID	Nominal	No		No
Customer_Nan	Rejected	Nominal	No		No
Discount	Rejected	Interval	No		No
Order_Date	Rejected	Interval	No		No
Order_ID	Rejected	Nominal	No		No
Postal_Code	Rejected	Interval	No		No
Product_ID	ID	Nominal	No		No
Product_Name	Rejected	Nominal	No		No
Profit	Rejected	Interval	No		No
Quantity	Rejected	Interval	No		No
Region	Input	Nominal	No		No
Sales	Target	Interval	No		No
Segment	Input	Nominal	No		No
Ship_Date	Rejected	Interval	No		No
Ship_Mode	Rejected	Nominal	No		No
Shipping_Dura	Rejected	Interval	No		No
State	Input	Nominal	No		No
Sub_Category	Input	Nominal	No		No
Year	Input	Ordinal	No		No
VAR1	Rejected	Interval	No		No

Variables set up in File Import node

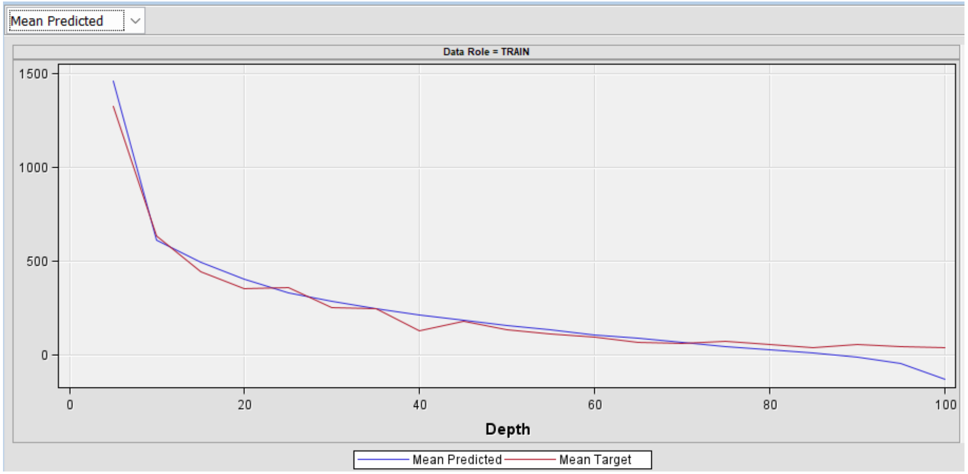
In the File Import node, changes in the variables have been made. For the imported data set, the role of Sales is set to Target. The Category, City, Region, Segment, State, Sub_Category, and Year are set as Input, while the Customer_ID, and Product_ID. The level of the Unit is in Intervals while the Category, City, Region, Segment, State, Sub_Category, Customer_ID, Product_ID, and Year as Ordinal.



General	
Node ID	Reg
Imported Data	...
Exported Data	...
Notes	...
Train	
Variables	...
⊟ Equation	
Main Effects	Yes
Two-Factor Interactions	No
Polynomial Terms	No
Polynomial Degree	2
User Terms	No
Term Editor	...
⊟ Class Targets	
Regression Type	Logistic Regression
Link Function	Logit
⊟ Model Options	
Suppress Intercept	No
Input Coding	Deviation
⊟ Model Selection	
Selection Model	None
Selection Criterion	Default
Use Selection Defaults	Yes
Selection Options	...
⊟ Optimization Options	
Technique	Default
Default Optimization	Yes
Max Iterations	0
Max Function Calls	0
Maximum Time	1 Hour
⊟ Convergence Criteria	
Uses Defaults	Yes
Options	...
⊟ Output Options	
Confidence Limits	No
Save Covariance	No
Covariance	No
Correlation	No
Statistics	No
Suppress Output	No
Details	No
Design Matrix	No
Score	
Excluded Variables	Reject

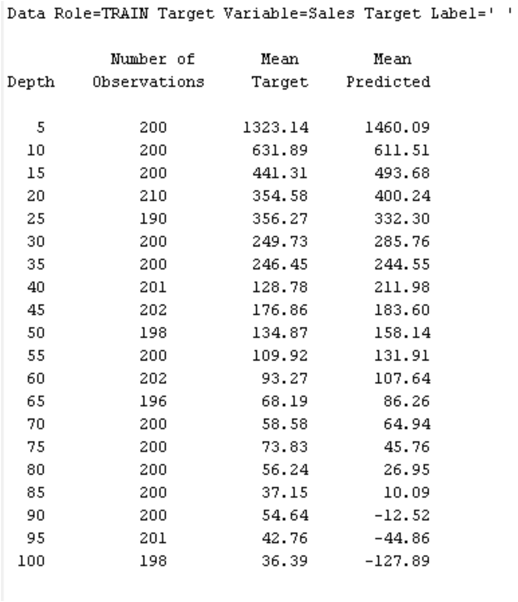
Regression node properties

In the Regression node, the Regression Type is set to Logistic Regression, while the rest of it remains by default.



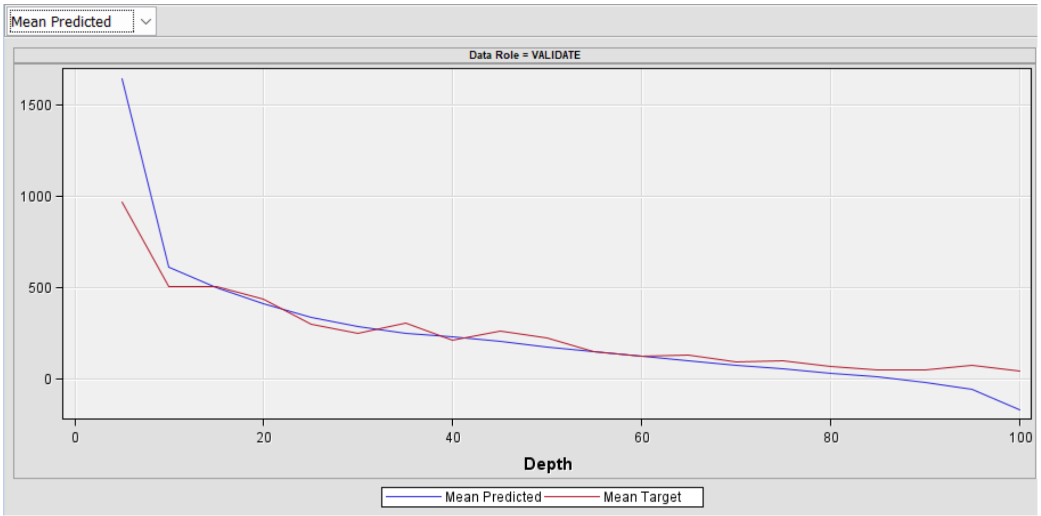
Score Rankings Matrix: Unit of TRAIN

The mean target and mean predicted value from depth of 5 to 100 with increment of 5 are found in the Assessment Score Rankings of Sales for Data Role = TRAIN. At a depth of five, the mean target and mean predicted are, respectively, 1323.14 and 1460.09. At a depth of 10, the mean target and mean predicted, with respective values of 631.89 and 611.51, significantly decrease. The average target was traveling erratically until it descended to 100 feet at 36.39. The expected mean value is falling until it reaches -127.89 at a depth of 100.



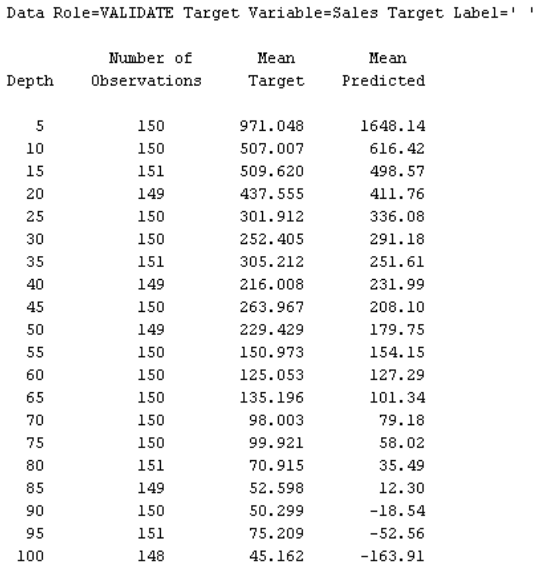
```
Data Role=TRAIN Target Variable=Sales Target Label=' '

           Number of        Mean        Mean
Depth    Observations     Target    Predicted

   5          200        1323.14     1460.09
  10          200         631.89      611.51
  15          200         441.31      493.68
  20          210         354.58      400.24
  25          190         356.27      332.30
  30          200         249.73      285.76
  35          200         246.45      244.55
  40          201         128.78      211.98
  45          202         176.86      183.60
  50          198         134.87      158.14
  55          200         109.92      131.91
  60          202          93.27      107.64
  65          196          68.19       86.26
  70          200          58.58       64.94
  75          200          73.83       45.76
  80          200          56.24       26.95
  85          200          37.15       10.09
  90          200          54.64      -12.52
  95          201          42.76      -44.86
 100          198          36.39     -127.89
```

Assessment Score Rankings for TRAIN



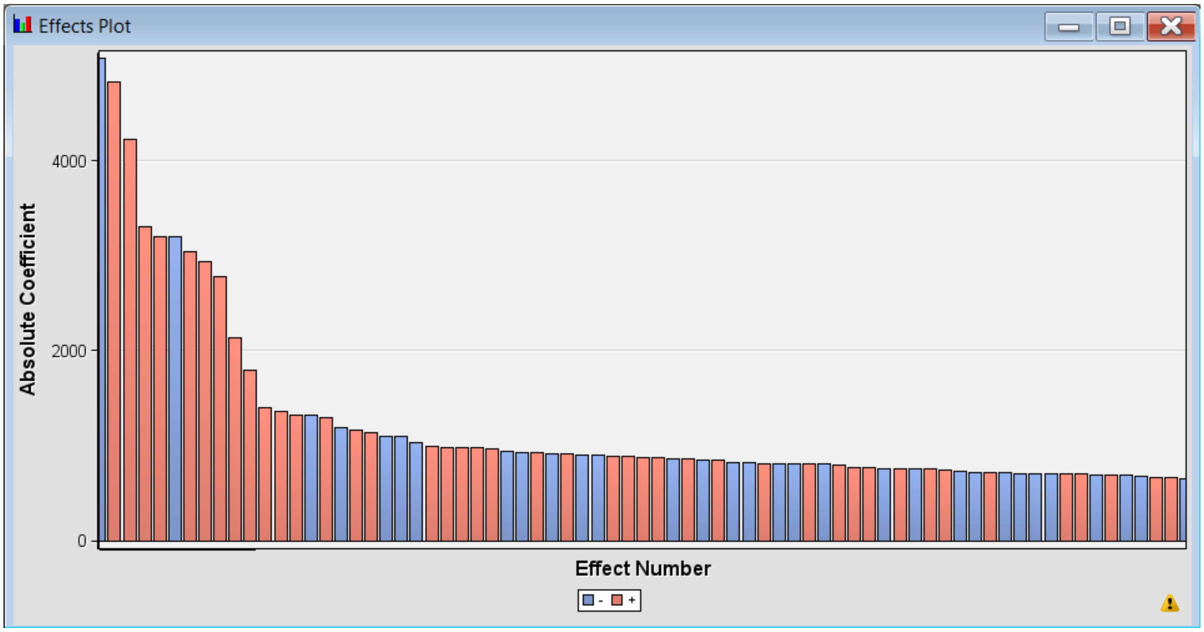
Score Rankings Matrix: Unit of VALIDATE

The mean target and mean predicted value from depth of 5 to 100 with increment of 5 are found in the Assessment Score Rankings of Unit for Data Role = VALIDATE. At a depth of 5, the mean target and mean predicted are, respectively, 971.048 and 1648.14. At a depth of 10, the mean target and mean predicted, with respective values of 507.007 and 616.42, significantly decrease. The average target was traveling erratically until it descended to 100 feet at 45.162. At a depth of 100, the expected mean value starts to decrease and eventually reaches -163.91.



```
Data Role=VALIDATE Target Variable=Sales Target Label=' '

              Number of        Mean         Mean
Depth       Observations      Target     Predicted

    5            150         971.048      1648.14
   10            150         507.007       616.42
   15            151         509.620       498.57
   20            149         437.555       411.76
   25            150         301.912       336.08
   30            150         252.405       291.18
   35            151         305.212       251.61
   40            149         216.008       231.99
   45            150         263.967       208.10
   50            149         229.429       179.75
   55            150         150.973       154.15
   60            150         125.053       127.29
   65            150         135.196       101.34
   70            150          98.003        79.18
   75            150          99.921        58.02
   80            151          70.915        35.49
   85            149          52.598        12.30
   90            150          50.299       -18.54
   95            151          75.209       -52.56
  100            148          45.162      -163.91
```

Assessment Score Rankings for VALIDATE

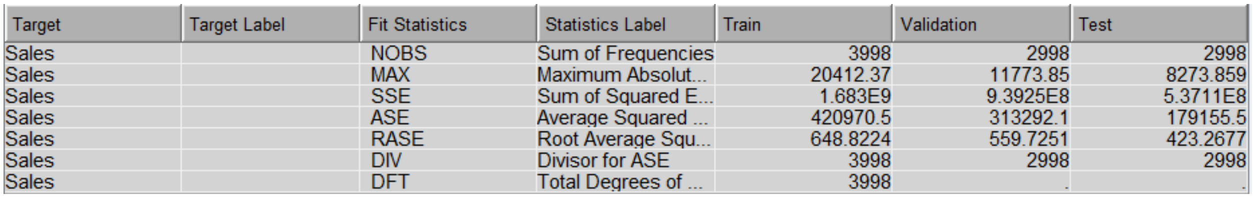


Effects Plot of parameters

The graph of Effects Plot is sorted out from the estimated value of Parameters. The positive value of the Absolute Coefficient is in blue, and the negative values are in red.

6.4 Decision Tree

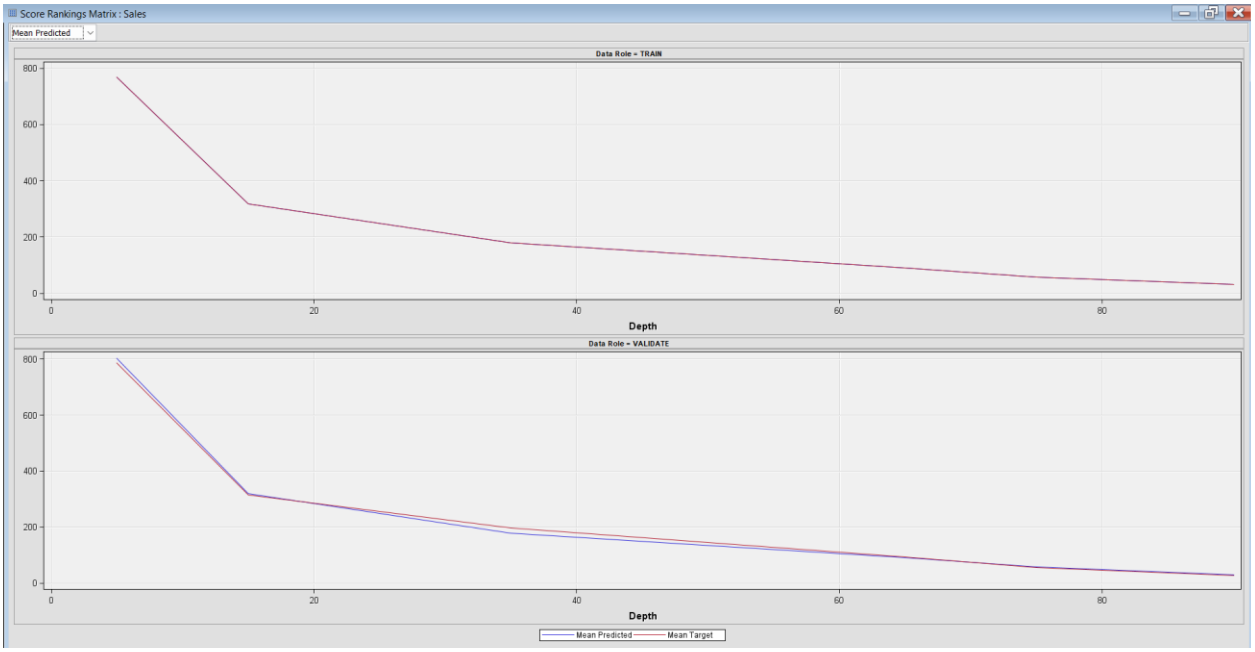
A decision tree is a visual aid for data analysis and machine learning that helps make decisions based on a model of potential outcomes that resembles a tree. The first decision or feature that best partitions the data into discrete groups is represented by the root node, which starts the process. Decision nodes represent test conditions based on features and are connected by branches to produce additional results. At the end of a branch, terminal nodes, also called leaf nodes, represent the final choice or anticipated result. In classification tasks, they represent class labels, and in regression tasks, they represent numerical values. To produce homogeneous groups and maximize purity within nodes, the algorithm chooses characteristics and conditions for data splitting. This recursive procedure keeps going until a predetermined stopping point—like a node purity threshold or predefined tree depth—is reached. After it is constructed, the decision tree uses feature values to traverse from the root to a leaf node, allowing for the classification or prediction of new data points.



Target	Target Label	Fit Statistics	Statistics Label	Train	Validation	Test
Sales		NOBS	Sum of Frequencies	3998	2998	2998
Sales		MAX	Maximum Absolut...	20412.37	11773.85	8273.859
Sales		SSE	Sum of Squared E...	1.683E9	9.3925E8	5.3711E8
Sales		ASE	Average Squared ...	420970.5	313292.1	179155.5
Sales		RASE	Root Average Squ...	648.8224	559.7251	423.2677
Sales		DIV	Divisor for ASE	3998	2998	2998
Sales		DFT	Total Degrees of ...	3998	.	.

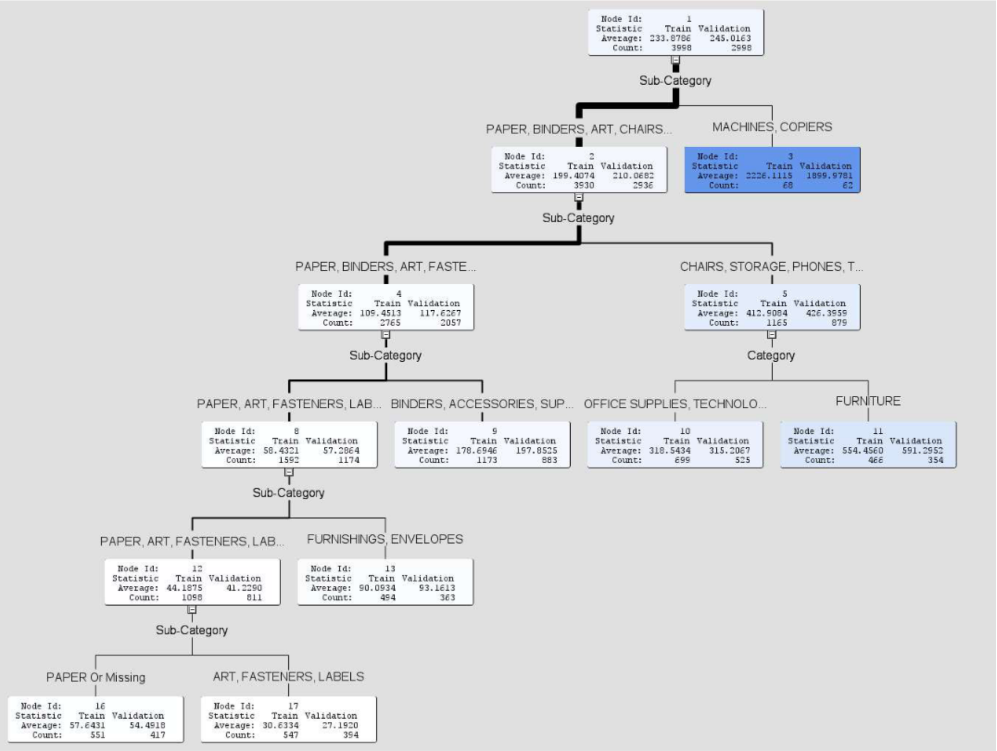
Fit Statistics

The Fit Statistics table offers important metrics for assessing how well a predictive model performs on various datasets, such as test, validation, and training sets, with a focus on the target variable "Unit." The sum of frequencies, or "_NOBS_" statistic, indicates the total number of observations in each dataset. "_MAX_" stands for the highest absolute error that was noted in the forecasts. The asterisk "_SSE_" represents the sum of squared errors, which is a measurement of the total squared discrepancies between the actual and predicted values. By dividing the total squared errors by a divisor, "_ASE_" stands for the average squared error. "_DIV_." By taking the average squared error's square root, one can determine the root average squared error, or "_RASE_". The total degrees of freedom in the model are indicated by "_DFT_". Collectively, these statistics provide information about the predictive model's precision, accuracy, and general fit across various datasets.



Score Ranking Matrix : Sales

Important details about the standard, baseline, and best models are displayed on a graph created using the training and validation data in the sales score rankings matrix plot. This graph provides a thorough analysis of the models' performance by displaying plots for the average predicted and actual values at various levels. Enhancing the analysis with plotting capabilities for report variables is a plus. In order to generate a validation sample for a robust evaluation, it is crucial to incorporate a Data Partition node in the process flow diagram for the purpose of creating these plots. The models' consistency and dependability can be seen in the very small difference between the mean predicted and mean target.


Node Id: 1
Statistic Train Validation
Average: 233.8786 245.0163
Count: 3998 2998
Sub-Category
PAPER, BINDERS, ART, CHAIRS...
MACHINES, COPIERS
Node Id: 2
Statistic Train Validation
Average: 199.4074 210.0682
Count: 3930 2936
Node Id: 3
Statistic Train Validation
Average: 2226.1115 1899.9781
Count: 68 62
Sub-Category
PAPER, BINDERS, ART, FASTE...
CHAIRS, STORAGE, PHONES, T...
Node Id: 4
Statistic Train Validation
Average: 109.4513 117.6267
Count: 2765 2057
Node Id: 5
Statistic Train Validation
Average: 412.9084 426.3959
Count: 1165 879
Sub-Category
Category
PAPER, ART, FASTENERS, LAB...
BINDERS, ACCESSORIES, SUP...
OFFICE SUPPLIES, TECHNOLO...
FURNITURE
Node Id: 8
Statistic Train Validation
Average: 58.4321 57.2864
Count: 1592 1174
Node Id: 9
Statistic Train Validation
Average: 178.6946 197.8525
Count: 1173 883
Node Id: 10
Statistic Train Validation
Average: 318.5434 315.2067
Count: 699 525
Node Id: 11
Statistic Train Validation
Average: 554.4560 591.2952
Count: 466 354
Sub-Category
PAPER, ART, FASTENERS, LAB...
FURNISHINGS, ENVELOPES
Node Id: 12
Statistic Train Validation
Average: 44.1875 41.2290
Count: 1098 811
Node Id: 13
Statistic Train Validation
Average: 90.0934 93.1613
Count: 494 363
Sub-Category
PAPER Or Missing
ART, FASTENERS, LABELS
Node Id: 16
Statistic Train Validation
Average: 57.6431 54.4918
Count: 551 417
Node Id: 17
Statistic Train Validation
Average: 30.6334 27.1920
Count: 547 394

Decision tree

The root node, the highest node in a vertical decision tree diagram, is the main category for the entire tree and includes all observations. Inside nodes are non-terminal nodes that contain the splitting rule; these include the root node. The final classification for a given set of observations is provided by leaf nodes, which are terminal nodes.

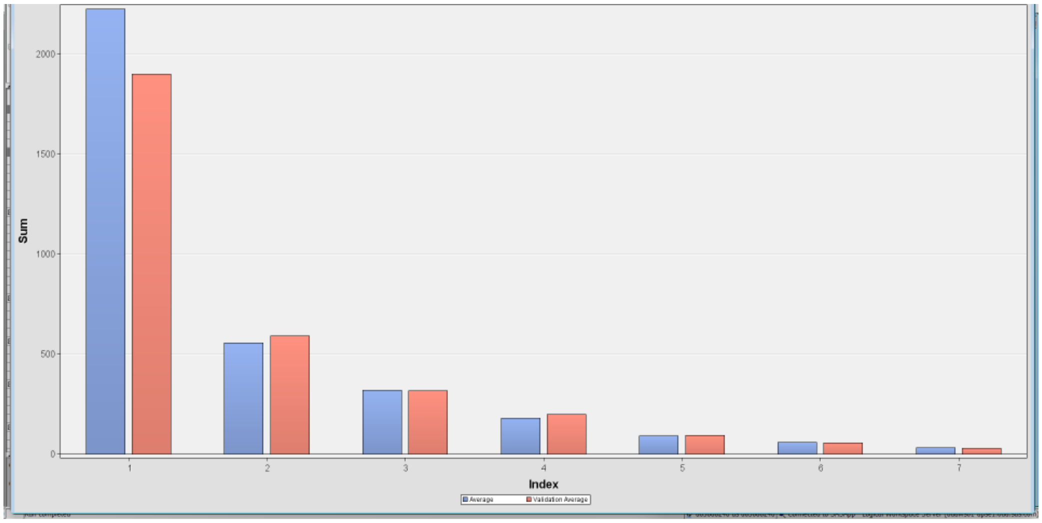
This decision tree's vertical orientation, in which the branches extend downward from the root node is one of its key characteristics. A visual representation of average values within various tree segments is provided by the colored nodes, which are based on the average of the "Sales" variable. The distribution of observations is shown by the width of the lines connecting nodes, which is proportionate to the number of observations in a branch divided by the total number of observations in the root node. Consistency in the visual representation of the decision tree structure is also ensured by the lines' consistent color throughout the tree.

The node text that is displayed in a tree depends on the Sales. Since the target is interval, the text box displays the number of observations in a node and the average value of the model assessment measure.

In interpreting this decision tree, we begin at the top level and focus on the results for the validation data. The root node reveals an average value for sub-category of approximately 245.016. The sub-category is identified as a crucial criterion, and a further distinction is made between the Machine and Copiers sub-category, which shows a higher average unit value of 1899.98, in contrast to other sub-categories such as Paper, Binder, Art, Chair, and more, which have a lower value of 210.07.

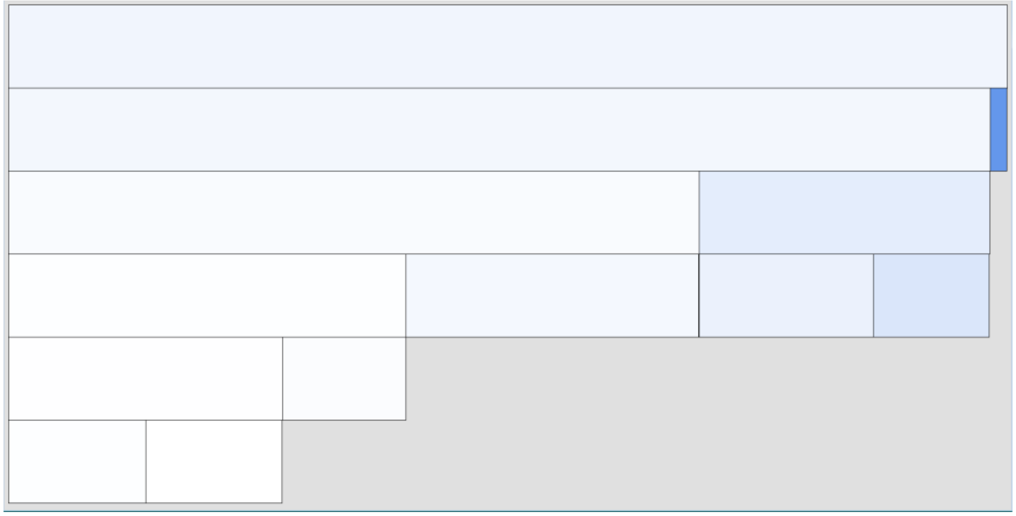
As we traverse down the tree, additional sub-categories like Chairs, Storage, Phone, etc., emerge, with a higher validation average unit value of 426.40 compared to other sub-categories with a lower value of 117.63. A notable observation is that the Furniture subcategory holds the second-highest average validation unit value, amounting to 592.30.

In summary, this decision tree highlights the importance of the sub-category in determining the average unit values. It distinguishes sub-categories based on their impact on the average unit values for validation data, with particular emphasis on the Machine and Copiers, as well as the Furniture sub-categories, which hold higher average unit values compared to other sub-categories. This tree provides valuable info about the factors influencing the unit values within different product sub-categories.



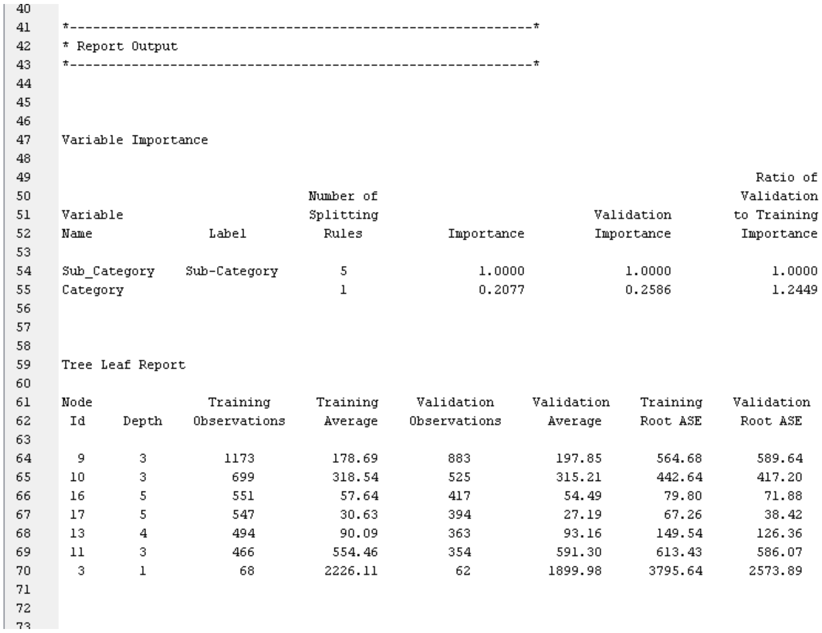
Leaf Statistic

The Leaf Statistics plot in the Decision Tree Results window shows a bar chart summarizing statistics for the leaves of the chosen subtree. To view detailed information, you can select the Leaf Statistics plot and then go to View → Table on the Results main menus to access the Leaf Statistics table. The data presented in this plot is derived from the average values of both the training and validation datasets for each leaf in the tree. This provides a quick visual representation and additional tabular details about the performance of individual leaves within the selected subtree.



Tree Map

The SAS Treemap plot is like a simpler version of tree structure.. It shows colors on the tree parts based on how much data they have from both training and validation sets. This coloring scheme provides visual information about the composition of nodes. The width of each tree part is decided by how many observations it has visually represents the distribution of data . When a node is selected in the Treemap window, the corresponding node in the Tree window is automatically selected, making it easy to look at both pictures together. This helps us quickly see how the data is spread in the tree.



```
*------------------------------------------------------------*
* Report Output
*------------------------------------------------------------*



Variable Importance

                                                                                     Ratio of
                                Number of                                           Validation
Variable                        Splitting                          Validation       to Training
Name              Label           Rules          Importance        Importance       Importance

Sub_Category    Sub-Category        5               1.0000            1.0000            1.0000
Category                            1               0.2077            0.2586            1.2449



Tree Leaf Report

Node                  Training      Training     Validation     Validation     Training     Validation
 Id      Depth      Observations     Average    Observations      Average      Root ASE      Root ASE

  9        3            1173          178.69        883            197.85        564.68        589.64
 10        3             699          318.54        525            315.21        442.64        417.20
 16        5             551           57.64        417             54.49         79.80         71.88
 17        5             547           30.63        394             27.19         67.26         38.42
 13        4             494           90.09        363             93.16        149.54        126.36
 11        3             466          554.46        354            591.30        613.43        586.07
  3        1              68         2226.11         62           1899.98       3795.64       2573.89
```

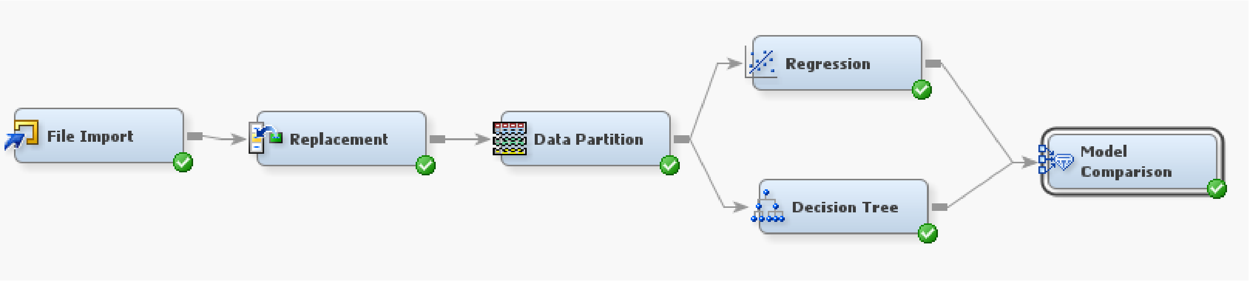
Output

The top section shows variable importance for a model’s predictive performance. The variables are named, along with the number of splitting rules used in the model, their validation importance, and their ratio of validation to training importance. The bottom section shows a tree leaf report, possibly from the same model. This section includes data about nodes in the tree, including their ID, depth, training and validation observations and averages, as well as root mean squared error (RMSE) for both training and validation.

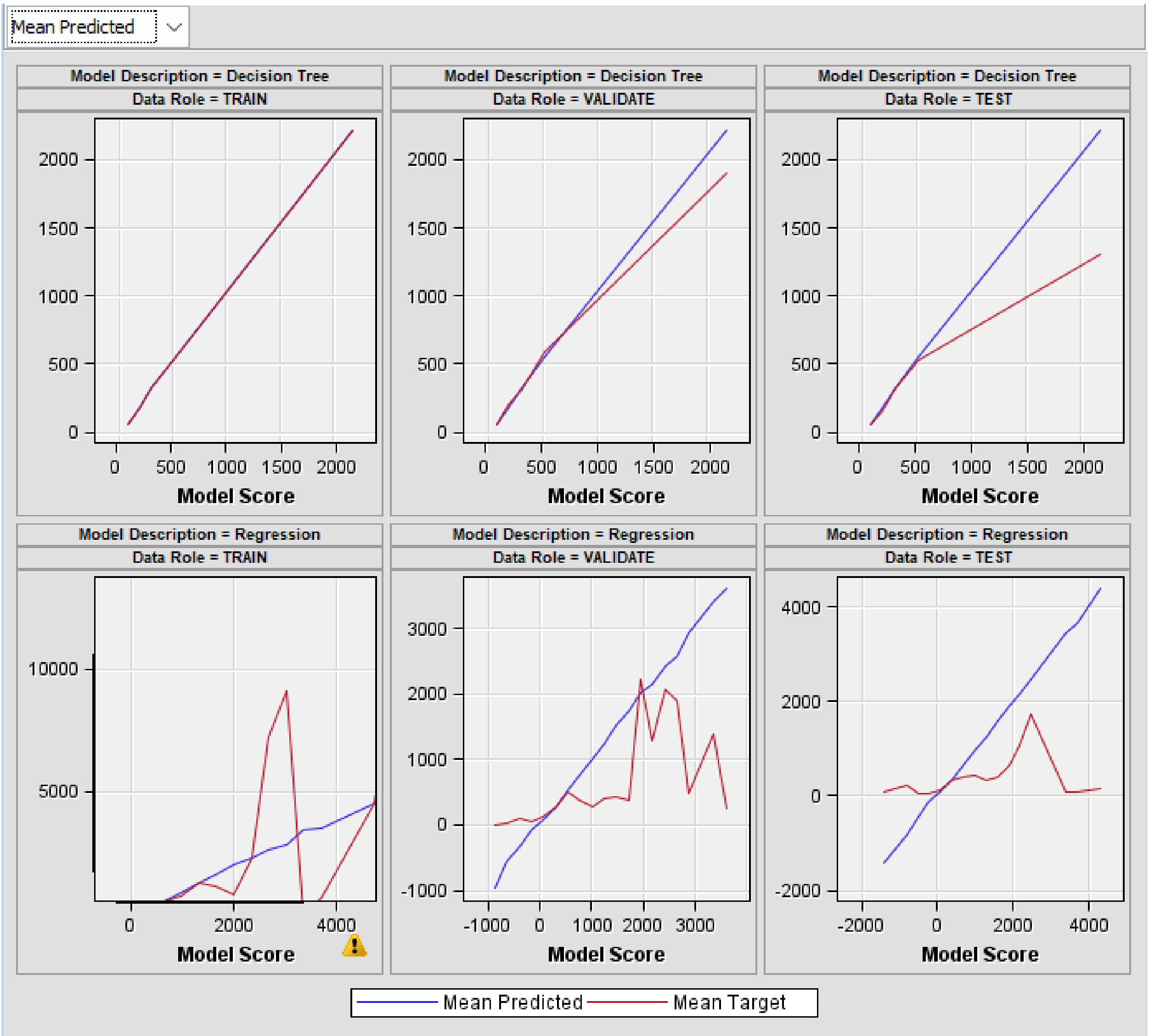
7.0 Assess

7.1 Model Comparison

For model comparison, we evaluated the performance of process flow diagrams utilizing one or more analytic modeling nodes available in the Model tab of the SAS Enterprise Miner toolbar. The nodes subjected to comparison were Regression and Decision Tree mo dels. The process and its components are visually represented in the diagram below:



Model Comparison

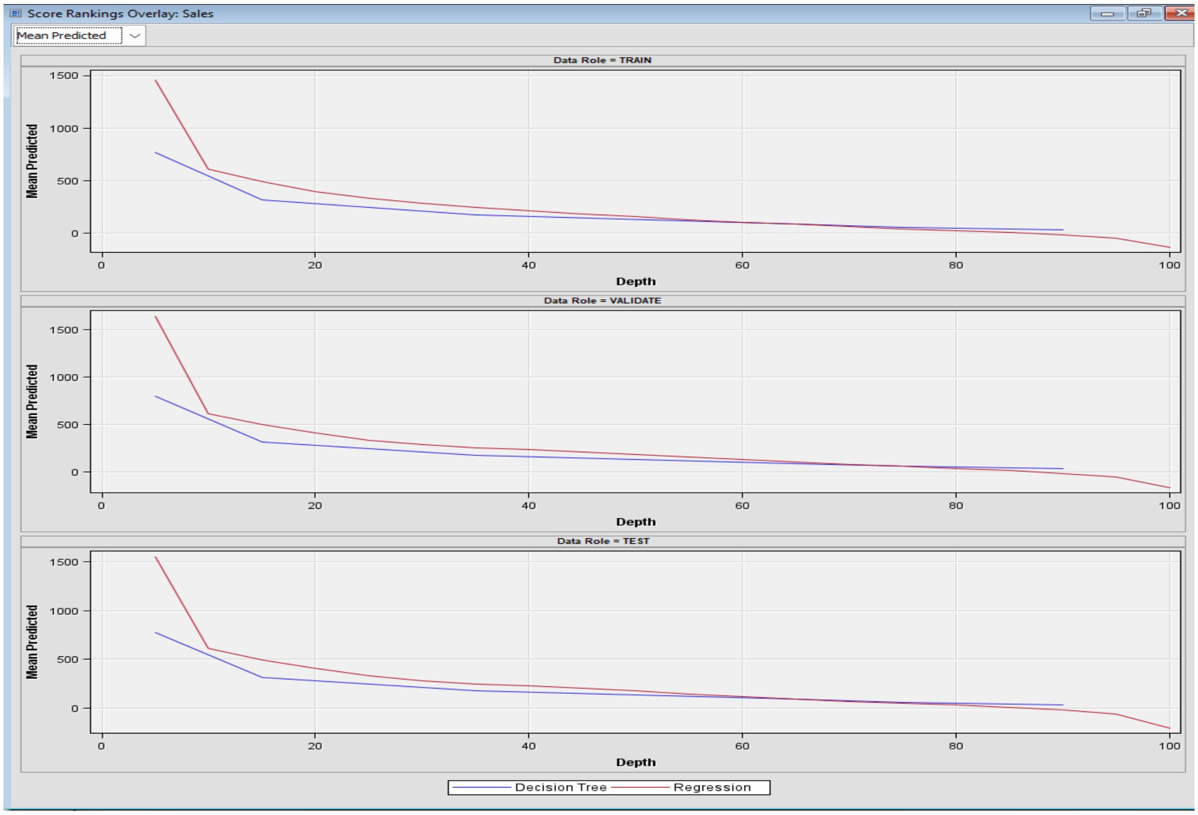


Score Distribution: Sales

The Score Distribution of Sales for the Decision Tree and Regression models is depicted in the above diagram. Three categories of data roles train, validate, and test are used in the analysis. The training data set is used to define the model, the testing data set is used to refine the model iteratively, and the validation data set is used to give the model a final evaluation. Depending on the profit and loss requirements in the data, misclassification is used in supervised learning with a defined outcome variable.The mean predict, mean target, and mean predict are all 58.43214 at a model score of 85.52035, according to the Decision Tree's score distribution of sales for Data Role = TRAIN. The model score of 2226.111 is reached at 2171.225 when the mean prediction and mean target grow linearly. The mean predict and mean target for Data Role = TRAIN in Regression are 30.34781 and 62.09261, respectively, with a model score of -30.0244. At a model score of 4714.144, the mean target has increased significantly to 4548.81, while the mean prediction increases linearly until it reaches 4548.81.

The mean prediction and mean target for the Decision Tree's Score Distribution of Unit for Data Role = VALIDATE are 58.61213 and 57.28643, respectively, with a model score of 85.52035. The mean target increases significantly until it reaches 591.2952 at the model score of 524.616, while the mean prediction increases linearly until it reaches 554.456. With a model score of 81.4343, the mean prediction and mean target for regression are 82.83242 and 118.1156, respectively. At a model score of 1954.425, the mean prediction rises linearly until it reaches 2029.662. The mean target rises steadily until it reaches 2216.299 at 1954.425, the model score, and then it falls to 242.16 at 3593.291, the model score.

The Decision Tree, with a model score of 85.52035, has a mean prediction and mean target of 57.40081 and 27.89585, respectively, in Data Role = TEST of Score Distribution of Unit. After reaching 554.456 at the model score of 524.616, the mean prediction increases linearly, and at 524.616, the mean target increases to 530.8136. In Regression, the model score is 83.56682, and the initial mean prediction and mean target are 84.68944 and 105.4677, respectively. After reaching 633.7958 at the model score of 681.761, the mean prediction keeps growing linearly. However, at the model score of 681.761, the mean target is 414.2715 and rises significantly to 1743.065 at the model score of 2476.344 then it falls sharply to 96.37333 at the model score of 3373.635.

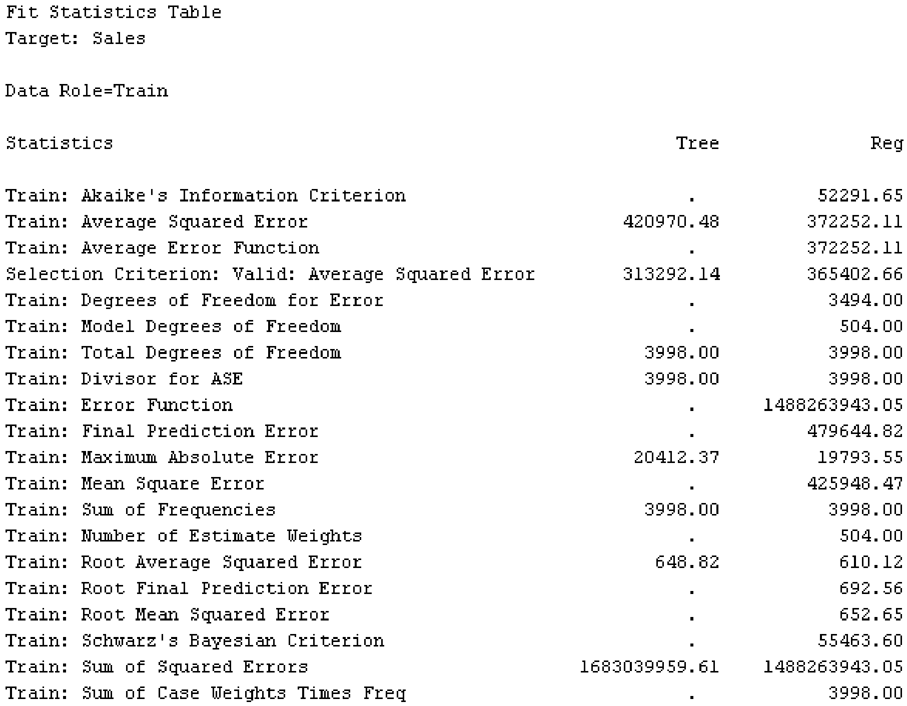


Score Rankings Overlay : Sales

The Score Ranking Overlay of the Sales diagram above compares the Decision Tree and Regression models in terms of training, validating, and testing. Based on Data Role = TRAIN, the mean predicted value of Decision Tree and Regression has depth ranges from 0 to 100 with increments of 20. The mean that Decision Tree and Regression predict at a depth of 5 are 767.326 and 1460.088, respectively. The mean predicted by the Decision Tree and the Regression both sharply declined to 318.5434 and 493.6832, respectively, at a depth of 15. The mean prediction in the Decision Tree decreases until it hits 30.6334 at a depth of 90. The mean prediction in the regression falls off rapidly, reaching -127.895 at a depth of 100.

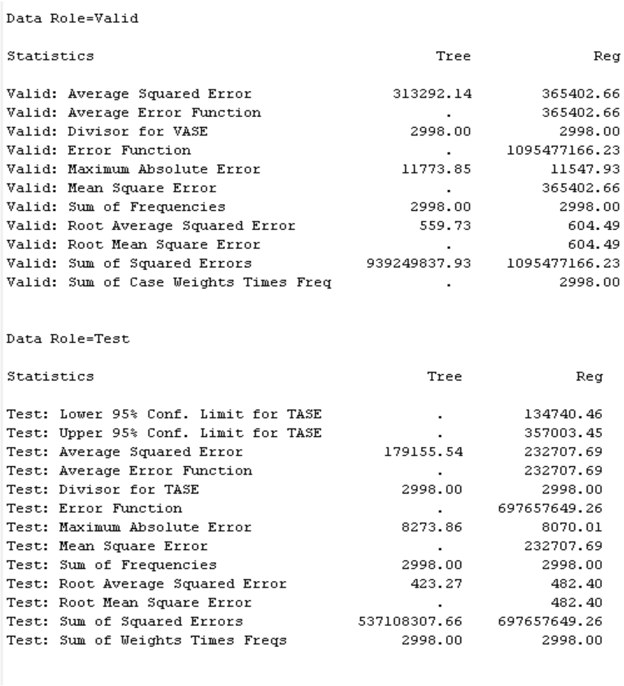
In Data Role = VALIDATE, the mean predicted for Regression and Decision Tree at a depth of 5 are 803.5969 and 1648.138, respectively. At a depth of 10, the Decision Tree method yields a significantly lower mean prediction, while at a depth of 15, Regression yields a significantly lower mean prediction. Following that, the Decision Tree's mean prediction continues to decrease until it reaches 30.6334 at a depth of 90. Regression analysis showed that the mean prediction dropped until it reached -18.5405, or a depth of 100.

Lastly, the mean prediction of the Decision Tree and Regression is displayed at a depth of 5 in the Data Role = TEST, with values of 5285.041 and 5048.821, respectively. The mean prediction dramatically drops at a depth of 20, with values for Regression and Decision Tree being 633.8932 and 323.2229, respectively. The Decision Tree mean prediction then drops to 1.588077 at a depth of 85, and the Regression mean predict value drops to -373.285 at a depth of 100.

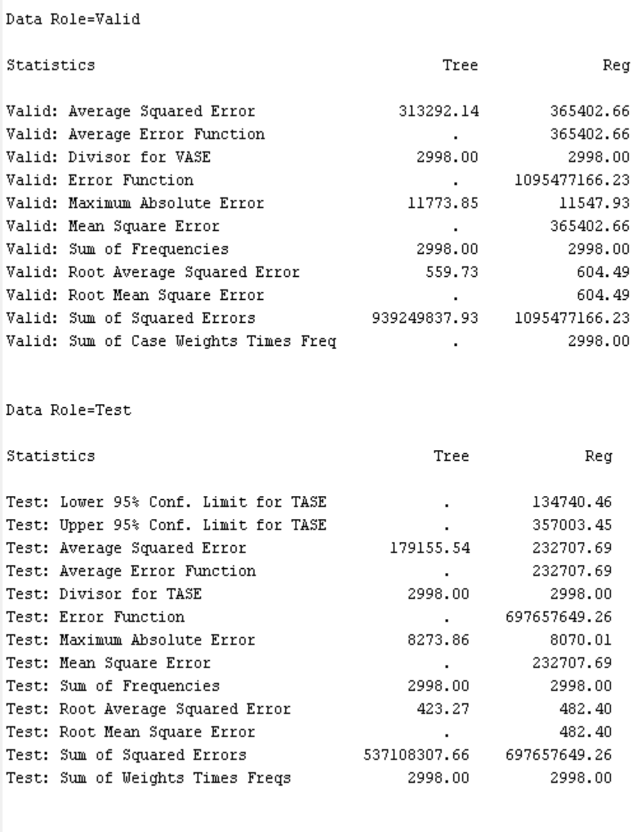


```
Fit Statistics Table
Target: Sales

Data Role=Train

Statistics                                                   Tree             Reg

Train: Akaike's Information Criterion                          .         52291.65
Train: Average Squared Error                           420970.48        372252.11
Train: Average Error Function                                  .        372252.11
Selection Criterion: Valid: Average Squared Error      313292.14        365402.66
Train: Degrees of Freedom for Error                            .          3494.00
Train: Model Degrees of Freedom                                .           504.00
Train: Total Degrees of Freedom                          3998.00          3998.00
Train: Divisor for ASE                                   3998.00          3998.00
Train: Error Function                                          .    1488263943.05
Train: Final Prediction Error                                  .        479644.82
Train: Maximum Absolute Error                           20412.37         19793.55
Train: Mean Square Error                                       .        425948.47
Train: Sum of Frequencies                                3998.00          3998.00
Train: Number of Estimate Weights                              .           504.00
Train: Root Average Squared Error                         648.82           610.12
Train: Root Final Prediction Error                             .           692.56
Train: Root Mean Squared Error                                 .           652.65
Train: Schwarz's Bayesian Criterion                            .         55463.60
Train: Sum of Squared Errors                       1683039959.61    1488263943.05
Train: Sum of Case Weights Times Freq                          .          3998.00
```

Fit Statistics



```
Data Role=Valid

Statistics                                      Tree             Reg

Valid: Average Squared Error               313292.14       365402.66
Valid: Average Error Function                      .       365402.66
Valid: Divisor for VASE                      2998.00         2998.00
Valid: Error Function                              .   1095477166.23
Valid: Maximum Absolute Error               11773.85        11547.93
Valid: Mean Square Error                           .       365402.66
Valid: Sum of Frequencies                    2998.00         2998.00
Valid: Root Average Squared Error             559.73          604.49
Valid: Root Mean Square Error                      .          604.49
Valid: Sum of Squared Errors            939249837.93   1095477166.23
Valid: Sum of Case Weights Times Freq              .         2998.00


Data Role=Test

Statistics                                    Tree             Reg

Test: Lower 95% Conf. Limit for TASE             .       134740.46
Test: Upper 95% Conf. Limit for TASE             .       357003.45
Test: Average Squared Error              179155.54       232707.69
Test: Average Error Function                     .       232707.69
Test: Divisor for TASE                     2998.00         2998.00
Test: Error Function                             .    697657649.26
Test: Maximum Absolute Error               8273.86         8070.01
Test: Mean Square Error                          .       232707.69
Test: Sum of Frequencies                   2998.00         2998.00
Test: Root Average Squared Error            423.27          482.40
Test: Root Mean Square Error                     .          482.40
Test: Sum of Squared Errors           537108307.66    697657649.26
Test: Sum of Weights Times Freqs           2998.00         2998.00
```



```
Data Role=Valid

Statistics                                      Tree             Reg

Valid: Average Squared Error               313292.14       365402.66
Valid: Average Error Function                      .       365402.66
Valid: Divisor for VASE                      2998.00         2998.00
Valid: Error Function                              .   1095477166.23
Valid: Maximum Absolute Error               11773.85        11547.93
Valid: Mean Square Error                           .       365402.66
Valid: Sum of Frequencies                    2998.00         2998.00
Valid: Root Average Squared Error             559.73          604.49
Valid: Root Mean Square Error                      .          604.49
Valid: Sum of Squared Errors            939249837.93   1095477166.23
Valid: Sum of Case Weights Times Freq              .         2998.00


Data Role=Test

Statistics                                    Tree             Reg

Test: Lower 95% Conf. Limit for TASE             .       134740.46
Test: Upper 95% Conf. Limit for TASE             .       357003.45
Test: Average Squared Error              179155.54       232707.69
Test: Average Error Function                     .       232707.69
Test: Divisor for TASE                     2998.00         2998.00
Test: Error Function                             .    697657649.26
Test: Maximum Absolute Error               8273.86         8070.01
Test: Mean Square Error                          .       232707.69
Test: Sum of Frequencies                   2998.00         2998.00
Test: Root Average Squared Error            423.27          482.40
Test: Root Mean Square Error                     .          482.40
Test: Sum of Squared Errors           537108307.66    697657649.26
Test: Sum of Weights Times Freqs           2998.00         2998.00
```

Fit Statistics Table (Data Role = Validate and Data Role = Test)

7.0 Conclusion

The data analysis and data mining of supervised and unsupervised learning approaches were reviewed in this report, and we then showed how to use and apply it in SAS Enterprise Miner tools. The depth and speed of analysis are both increased by data mining tools. Analysts can fully examine already existing datasets, spot relevant patterns and trends, and extract information by extracting the most important information through mining. The primary objectives were to understand customer preferences and purchasing behavior as well as the performance of product sales through the use of data reduction, cleaning, regression, association nodes, StatExplore nodes, decision trees, and model comparison. To improve data analysis results, we intend to continue refining the application of supervised learning techniques for classification within SAS Enterprise Miner.

8.0 Reflection

In the reflection, consider the evolving landscape of data analytics and the need for continuous adaptation and learning. Future work could involve exploring new data sources, such as social media or real-time data, to enrich the analysis. Additionally, integrating machine learning techniques could provide more nuanced insights into customer behavior. The report could also benefit from a comparative study of different analytical tools beyond SAS Enterprise Miner, to understand their relative strengths and applicability in various scenarios. Furthermore, the potential ethical implications of data mining should be considered, ensuring that customer privacy and data security are prioritized. Collaboration with experts in other fields, like marketing or consumer psychology, could also add valuable perspectives to the analysis, making it more holistic and impactful.