COFFEE SHOP PRODUCT AND CUSTOMER ANALYSIS

WIE3007 DATA MINING AND WAREHOUSING

GROUP MEMBERS:

Muhammad Rafid Ikhwan Bin Samsuri	17204331
Amrin Hafiz Bin Eddy Rosyadie	17205136
Azrul Haikal Bin Uhaidi	17201828
Wan Suraya Binti Wan Mohd Lotfi	U2005345
Nurul Filzah Binti Abdul Hadi	17205112

TABLE OF CONTENTS

SELECT DATASET

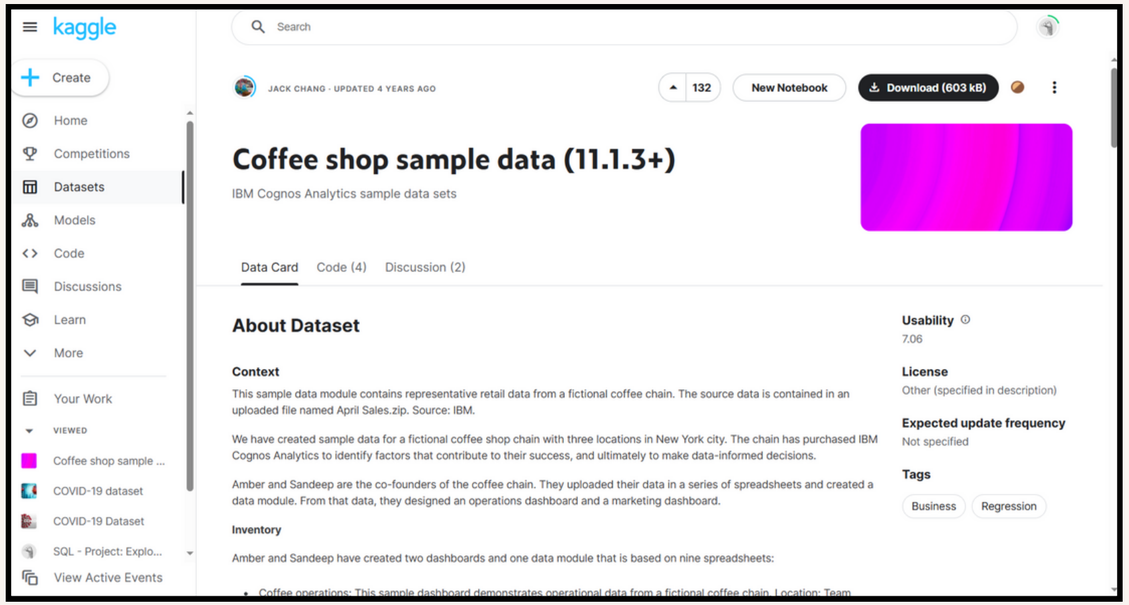
Coffee Shop Sample Data

from: Kaggle

TOOLS:

SAS Enterprise Miner

Missing Values:
quantity: 14



Dataset from Kaggle

ATTRIBUTE

Sales Receipt

sales_id	Unique ID for every sales
transaction_id	Unique ID for every transaction
transaction_date	Date of transaction
transaction_time	Time of transaction
customer_id	Unique ID for every customer
instore_yn	Method of buying i.e. instore or online
product_id	Unique ID for every product
quantity	Quantity purchased for every transaction
unit_price	Price for every unit of product

Customer

customer_id	Unique ID for every customer
customer_firstName	Name of every customer
customer_since	Date of customer's first purchase
gender	Gender of customer
birthyear	Birth year of every customer
generation	Generation of every customer based on birth year

Product

product_id	Unique ID for every product
product_group	Category of product
product	Product name
current_retail_price	Current product price
promo_yn	Product promotion availability
new_product_yn	New product availability

Business Objectives

Understanding Customer Behavior

Research demands advanced analytical techniques to process diverse customer behavior data, providing actionable insights.

Assessing Product Popularity and Demand

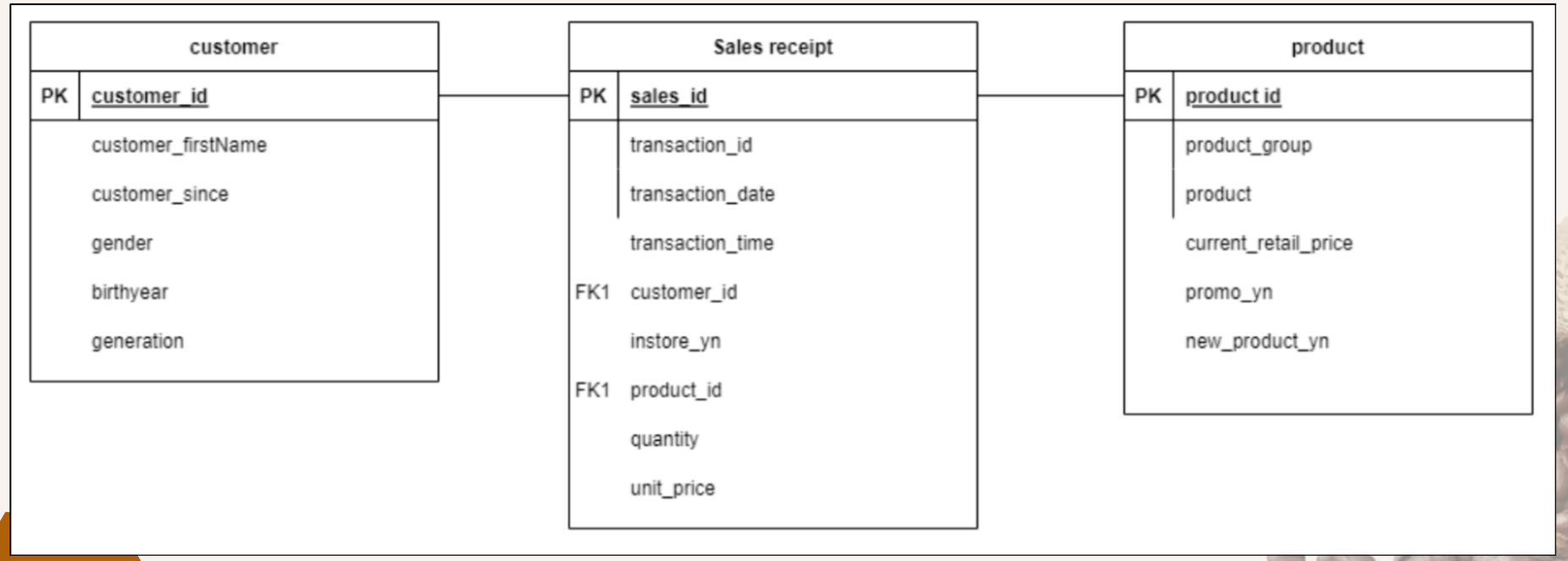
Accurate forecasting aids inventory management, cost reduction, sales, and customer satisfaction.

Uncovering Seasonal Buying Trends

eCommerce companies prioritize seasonal demand forecasting to optimize stock levels and product selection.

Star Schema

To understand the relationship between tables as well as to organize the data so that it is easy to understand and analyze.



Featuretools

TOOLS:

Pyhton

Python Code:

```
import featuretools as ft
import pandas as pd

sales_df = pd.read_csv("sales_receipts.csv")
customer_df = pd.read_csv("customer.csv")
product_df = pd.read_csv("product.csv")

es = ft.EntitySet(id='coffee_data')
es.add_dataframe(dataframe_name='customer', dataframe=customer_df, index='
customer_id')
es.add_dataframe(dataframe_name='product', dataframe=product_df, index='
product_id')
es.add_dataframe(dataframe_name='sales', dataframe=sales_df, index=' sales_id')

relationships = [
   ('customer', ' customer_id', 'sales', ' customer_id'),
   ('product', ' product_id', 'sales', ' product_id')]

for relationship in relationships:
   es = es.add_relationship(parent_dataframe_name=relationship[0],
                            parent_column_name=relationship[1],
                            child_dataframe_name=relationship[2],
                            child_column_name=relationship[3])

feature_matrix, feature_defs = ft.dfs(entityset=es, target_dataframe_name='sales',
                                      agg_primitives=['sum', 'mean', 'count', 'mode'],
                                      trans_primitives=['month', 'weekday', 'day'],
                                      max_depth=2)
```

Result: 9 features created

1. **customer.COUNT(sales)**
   - The quantity of products every consumer has bought.
2. **customer.MEAN(sales. IMP__quantity)**
   - Average number of products a customer has bought..
3. **customer.MEAN(sales.unit_price)**
   - Average amount of money spent with each customer
4. **customer.MODE(sales. IMP__instore_yn)**
   - Highest frequency of the customer's approach to purchase
5. **customer.SUM(sales. IMP__quantity)**
   - Total amount of goods purchased in quantity
6. **customer.SUM(sales.unit_price)**
   - Total amount of money spent
7. **product.COUNT(sales)**
   - Total sales of a product assigned by popularity.
8. **product.SUM(sales. IMP__quantity)**
   - Total amount of highly desired products that have been sold.
9. **product.SUM(sales.unit_price)**
   - The total sales for every product

Data Integration

We use **Talend Data Integration** to integrate three datasets

- sales receipt
- customer
- product

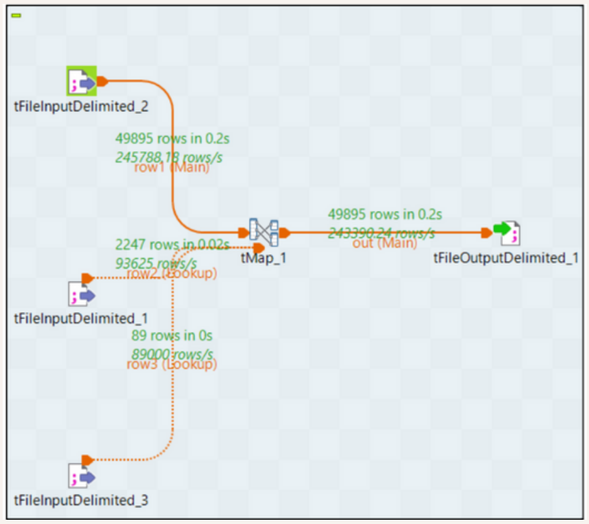
into a centralized dataset.

Centralising data, increasing accuracy, streamlining analysis, and promoting cross-functional insights are all made possible by integrating datasets, which also promotes improved decision-making and teamwork.

TOOLS:

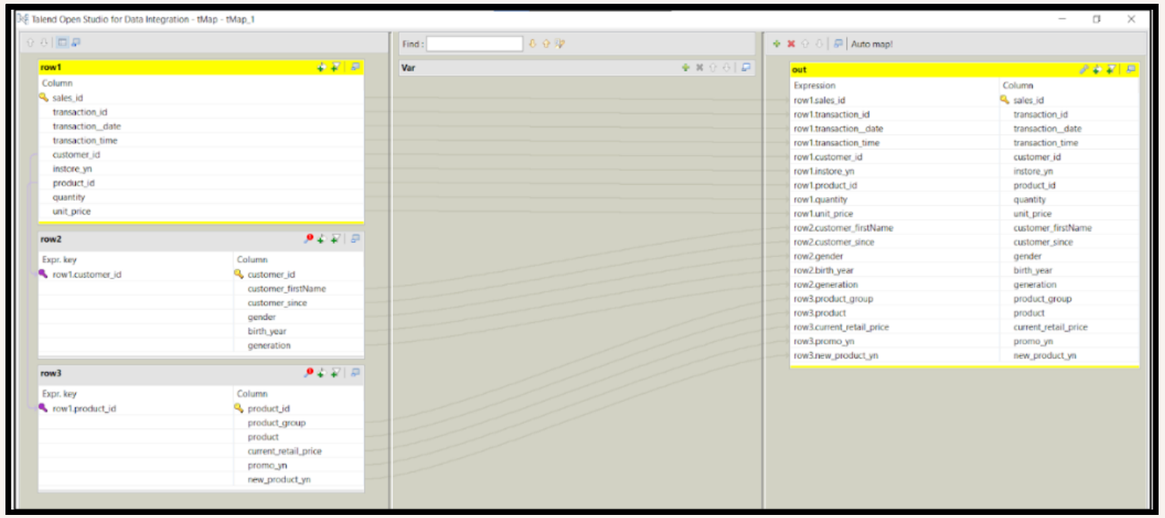
Talend Data Integration

01



Join three dataset using tMap

02



Define output structure

SAMPLE

A representative subset of data from the entire dataset is selected. This subset is used in the initial phases of the data mining process to make the computations more manageable and efficient.

RANDOM SAMPLING METHOD

- There are 49894 rows in the original dataset.
- A random sampling configuration has been applied with a set percentage of 1%.
- Consequently, the resulting subset consists of 499 rows.

Variable Summary

Role	Measurement Level	Frequency Count
ID	NOMINAL	4
INPUT	INTERVAL	4
INPUT	NOMINAL	8
TIMEID	INTERVAL	3

Sampling Summary

Type	Data Set	Number of Observations
DATA	EMWS3.FIMPORT_train	49894
SAMPLE	EMWS3.Smpl_DATA	499

Result of random sampling

General	
Node ID	Smpl
Imported Data	
Exported Data	
Notes	
Train	
Variables	
Output Type	Data
Sample Method	Random
Random Seed	12345
Size	
Type	Percentage
Observations	.
Percentage	1.0
Alpha	0.01
PValue	0.01
Cluster Method	Random

The settings for random sampling

TOOLS:

SAS Enterprise Miner

Target Variable

Generation

Comprehending the "Generation" variable is vital in customising marketing and product strategies to the unique interests and behaviours of every age group.

Product Group

Understand beverage popularity requires a grasp of the "Product Group" variable. We hope to accurately predict product group popularity, align marketing, and optimise resource allocation by leveraging it along with other dataset elements.

EXPLORE

Use the "Explore" node in SAS Enterprise Miner for potential text mining, detecting missing values, variable transformations, correlation analysis, summary statistics, data visualisation, outlier detection, and documenting discoveries for analysis and communication.

Summary Explore Method

- Explore the sample dataset using three types of techniques (descriptive statistics, data visualisation and correlation analysis).
- StatExplore, GraphExplore and Python on Jupyter Notebook is used to obtain the data and visualization

Descriptive Statistics

- Involve the computation and analysis of measures of central tendency (mean, median) and variability (standard deviation, minimum, maximum) to provide a clear overview of dataset features.
- Fourteen missing values are identified under the quantity column.

TOOLS:

SAS Enterprise Miner

Variable Summary

Role	Measurement Level	Frequency Count
ID	NOMINAL	4
INPUT	INTERVAL	5
INPUT	NOMINAL	9

Variable Levels Summary
(maximum 500 observations printed)

Variable	Role	Frequency Count
_customer_id	ID	233
_product_id	ID	70
_transaction_id	ID	431
sales_id	ID	499

Class Variable Summary Statistics
(maximum 500 observations printed)

Data Role=TRAIN

Data Role	Variable Name	Role	Number of Levels	Missing	Mode	Mode Percentage	Mode2	Mode2 Percentage
TRAIN	VAR21	INPUT	1	499		100.0		0.00
TRAIN	_customer_firstName	INPUT	222	0	Alexa Hancock	4.81	Giacomo Luna	3.01
TRAIN	_gender	INPUT	3	0	M	42.89	F	42.08
TRAIN	_generation	INPUT	5	0	Baby Boomers	36.67	Gen X	20.44
TRAIN	_instore_yn	INPUT	3	3	N	49.90	Y	49.50
TRAIN	_new_product_yn	INPUT	1	0	N	100.0		0.00
TRAIN	_product	INPUT	70	0	Morning Sunrise Chai Rg	3.21	Earl Grey Rg	3.01
TRAIN	_product_group	INPUT	5	0	Beverages	76.35	Food	13.83
TRAIN	_promo_yn	INPUT	1	0	N	100.0		0.00

Interval Variable Summary Statistics
(maximum 500 observations printed)

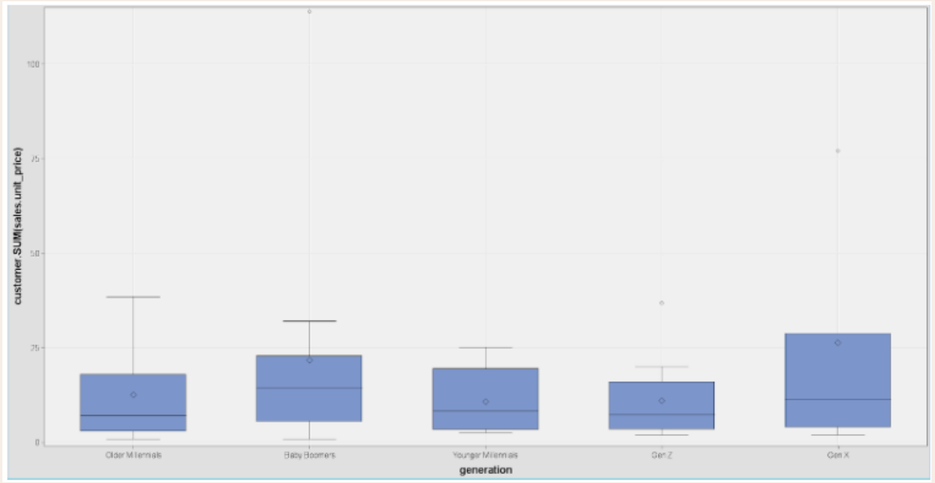
Data Role=TRAIN

Variable	Role	Mean	Standard Deviation	Non Missing	Missing	Minimum	Median	Maximum	Skewness	Kurtosis
__dataobs_	INPUT	25735.74	14818.52	499	0	151	24850	49883	-0.02326	-1.28605
_birth_year	INPUT	1974.922	15.54072	499	0	1950	1972	2001	0.132874	-1.39266
_current_retail_price	INPUT	3.739479	3.060348	499	0	1	3	28	5.127938	30.13735
_quantity	INPUT	1.447423	0.556534	485	14	1	1	3	0.750323	-0.49067
_unit_price	INPUT	3.567154	3.068847	499	0	0.8	3	28	5.158455	30.32614

Result of StatExplore

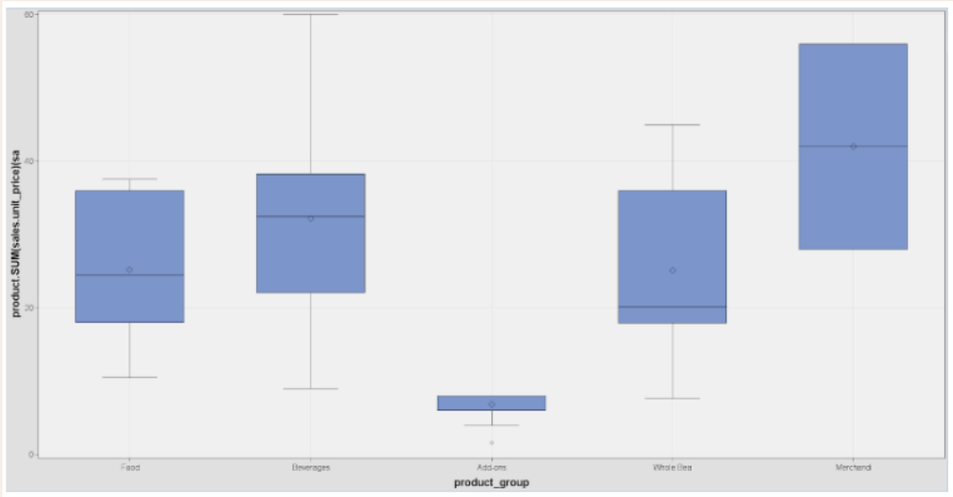
VISUALIZATION

Two-Variable Visualization (Bivariate)



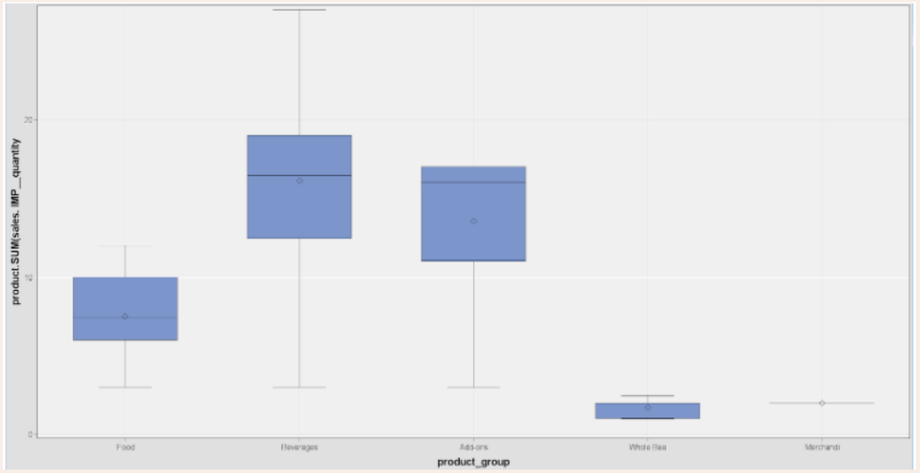
customer.SUM(sales.unit_price) and generation.

Box Plot



product.SUM(sales.unit_price) and product_group.

Bivariate visualisation examines two variables' relationship, correlation, or interaction. Bivariate representation allows examination of links and patterns between two variables. Scatter plots are useful bivariate visualisation tools.



product.SUM(sales. IMP__quantity) and product_group.

TOOLS:

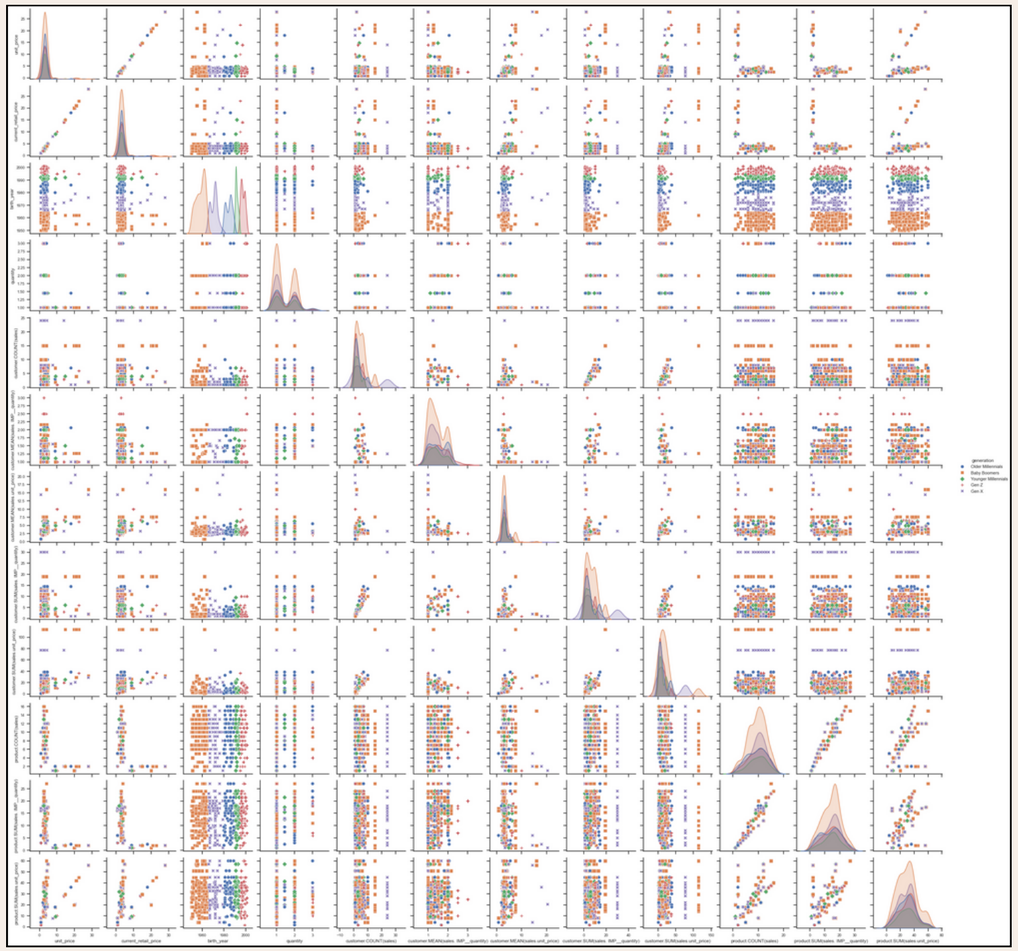
SAS Enterprise Miner

Three-Variable Visualization (Trivariate)

Scatter Plot

Trivariate visualisation seeks to understand three variables' interactions and correlations. This strategy improves understanding of these factors' linkages and influences.

We use a scatter plot in this group project, where the x- and y-axes reflect numerical attributes and the colour represents generation.



Scatter plot showing the relationship between attributes in the sample dataset.

TOOLS:

Jupyter Notebook (Python)

TOOLS:

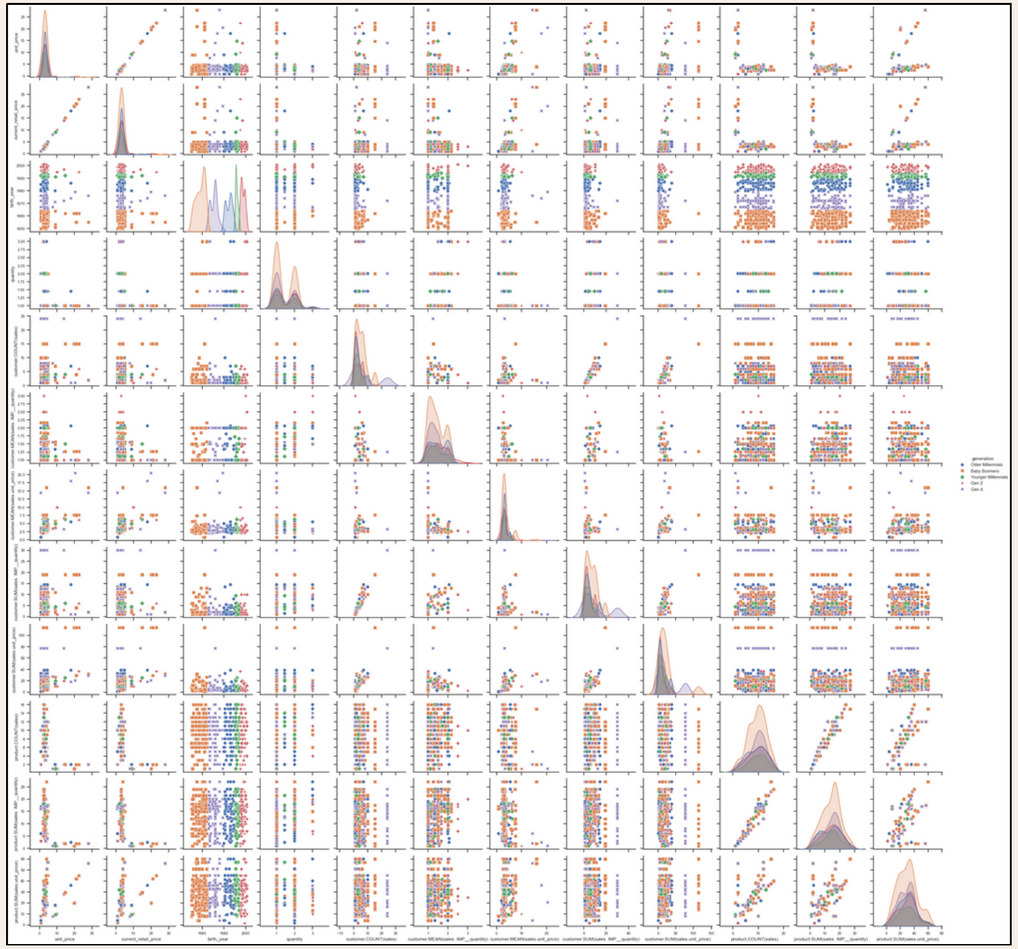
Jupyter Notebook (Python)

01

Correlation Analysis

A statistical method called correlation analysis evaluates the link and dependencies between variables. The project's objectives are to ascertain the distribution of generation groups and investigate relationships between numerical variables.

- Generation (colour) significantly influences customer.COUNT(sales), customer.SUM(sales.quantity), and customer.SUM(sales.unit_price).
  - **Gen X leads** in customer.COUNT(sales) and customer.SUM(sales.quantity).
  - **Baby Boomers excel** in customer.SUM(sales.unit_price).
- Highly correlated variables:
  - **unit_price and current_retail_price.**
  - **product.COUNT(sales) and product.SUM(sales.IMP_quantity).**
- One variable from each pair needs to be dropped before fitting the data into the model.

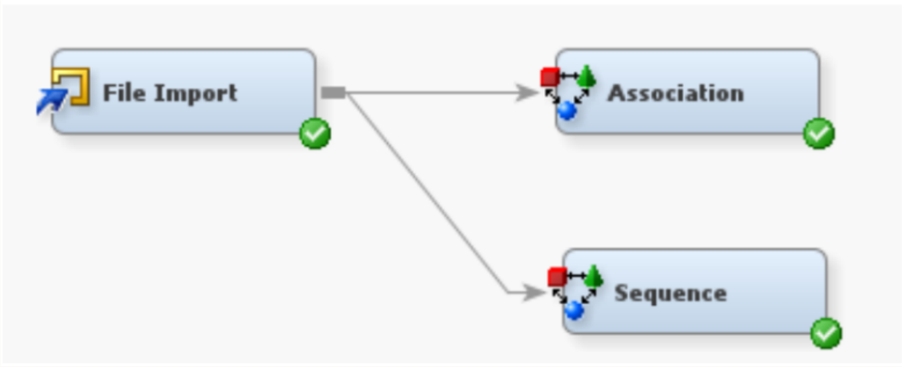


Scatter plot showing the relationship between attributes in the sample dataset.

02

Association Analysis

Similar to correlation analysis in statistical methods, association analysis becomes an important tool during the Explore phase of data warehouse development. This approach explores large datasets to find hidden connections, trends, and insights.



We do Association Rules and Sequence Rules using the node of ‘Association’ in SAS Enterprise Miner

TOOLS:

SAS Enterprise Miner

Association Rules

Map	Rule
RULE1	Sustainably Grown Organic Rg & Sugar Free Vanilla syrup ==> Brazilian Rg & Almond Croissant
RULE2	Sustainably Grown Organic Rg & Lemon Grass Rg ==> Brazilian Rg & Almond Croissant
RULE3	Sustainably Grown Organic Rg & Jumbo Savory Scone ==> Brazilian Rg & Almond Croissant
RULE4	Sustainably Grown Organic Rg & Ethiopia Lg ==> Brazilian Rg & Almond Croissant
RULE5	Sustainably Grown Organic Rg & Dark chocolate Lg ==> Brazilian Rg & Almond Croissant
RULE6	Sustainably Grown Organic Rg & Croissant ==> Brazilian Rg & Almond Croissant
RULE7	Sugar Free Vanilla syrup & Lemon Grass Rg ==> Brazilian Rg & Almond Croissant
RULE8	Sugar Free Vanilla syrup & Jumbo Savory Scone ==> Brazilian Rg & Almond Croissant
RULE9	Sugar Free Vanilla syrup & Jamaican Coffee River Lg ==> Brazilian Rg & Almond Croissant
RULE10	Sugar Free Vanilla syrup & Ethiopia Lg ==> Brazilian Rg & Almond Croissant
RULE11	Sugar Free Vanilla syrup & Dark chocolate Lg ==> Brazilian Rg & Almond Croissant
RULE12	Lemon Grass Rg & Jumbo Savory Scone ==> Brazilian Rg & Almond Croissant
RULE13	Lemon Grass Rg & Jamaican Coffee River Lg ==> Brazilian Rg & Almond Croissant
RULE14	Lemon Grass Rg & Croissant ==> Brazilian Rg & Almond Croissant
RULE15	Jumbo Savory Scone & Jamaican Coffee River Lg ==> Brazilian Rg & Almond Croissant
RULE16	Jumbo Savory Scone & Ethiopia Lg ==> Brazilian Rg & Almond Croissant
RULE17	Jumbo Savory Scone & Croissant ==> Brazilian Rg & Almond Croissant
RULE18	Jamaican Coffee River Lg & Ethiopia Lg ==> Brazilian Rg & Almond Croissant
RULE19	Jamaican Coffee River Lg & Dark chocolate Lg ==> Brazilian Rg & Almond Croissant
RULE20	Jamaican Coffee River Lg & Croissant ==> Brazilian Rg & Almond Croissant
RULE21	Ethiopia Lg & Croissant ==> Brazilian Rg & Almond Croissant
RULE22	Dark chocolate Lg & Croissant ==> Brazilian Rg & Almond Croissant
RULE23	Traditional Blend Chai Lg & Peppermint Rg ==> Cappuccino & Brazilian Sm
RULE24	Traditional Blend Chai Lg & Ethiopia Sm ==> Cappuccino & Brazilian Sm
RULE25	Traditional Blend Chai Lg & Carmel syrup ==> Cappuccino & Brazilian Sm
RULE26	Peppermint Rg & Jamaican Coffee River Rg ==> Cappuccino & Brazilian Sm
RULE27	Peppermint Rg & Carmel syrup ==> Cappuccino & Brazilian Sm
RULE28	Jamaican Coffee River Rg & Ethiopia Sm ==> Cappuccino & Brazilian Sm
RULE29	Ethiopia Sm & Carmel syrup ==> Cappuccino & Brazilian Sm
RULE30	Jumbo Savory Scone & Ginger Scone ==> Cappuccino Lg & Almond Croissant
RULE31	Jumbo Savory Scone & Columbian Medium Roast Sm ==> Cappuccino Lg & Almond Croissant
RULE32	Jumbo Savory Scone & Chocolate Croissant ==> Cappuccino Lg & Almond Croissant
RULE33	Jumbo Savory Scone & Carmel syrup ==> Cappuccino Lg & Almond Croissant
RULE34	Ginger Scone & Columbian Medium Roast Sm ==> Cappuccino Lg & Almond Croissant
RULE35	Ginger Scone & Chocolate Croissant ==> Cappuccino Lg & Almond Croissant
RULE36	Ginger Scone & Carmel syrup ==> Cappuccino Lg & Almond Croissant
RULE37	Sustainably Grown Organic Rg & Earl Grey Rg ==> Cappuccino Lg & Brazilian Sm
RULE38	Sustainably Grown Organic Rg & Carmel syrup ==> Cappuccino Lg & Brazilian Sm
RULE39	Jumbo Savory Scone & Ginger Scone ==> Carmel syrup & Almond Croissant
RULE40	Jumbo Savory Scone & Columbian Medium Roast Sm ==> Carmel syrup & Almond Croissant
RULE41	Jumbo Savory Scone & Chocolate Croissant ==> Carmel syrup & Almond Croissant
RULE42	Jumbo Savory Scone & Cappuccino Lg ==> Carmel syrup & Almond Croissant
RULE43	Ginger Scone & Columbian Medium Roast Sm ==> Carmel syrup & Almond Croissant
RULE44	Ginger Scone & Chocolate Croissant ==> Carmel syrup & Almond Croissant
RULE45	Ginger Scone & Cappuccino Lg ==> Carmel syrup & Almond Croissant
RULE46	Columbian Medium Roast Sm & Cappuccino Lg ==> Carmel syrup & Almond Croissant
RULE47	Chocolate Croissant & Cappuccino Lg ==> Carmel syrup & Almond Croissant
RULE48	Sustainably Grown Organic Lg & Sugar Free Vanilla syrup ==> Chocolate Chip Biscotti & Brazilian Rg
RULE49	Sustainably Grown Organic Lg & Spicy Eye Opener Chai Rg ==> Chocolate Chip Biscotti & Brazilian Rg
RULE50	Sustainably Grown Organic Lg & Dark chocolate Rg ==> Chocolate Chip Biscotti & Brazilian Rg

Association rules are patterns or links found in datasets utilising association analysis or mining.
Finding interesting links:

- helps understand consumer behaviour
- gain insights into the data.

In the association node:

- ID variable = customer_id
- Target variable = product.

Sequence Rules

Sequence rules describe the temporal ordering and sequential relationships of dataset events.
focus on:

- co-occurrence,
- capture for patterns based on order.

In the sequence node:

- ID variable = customer_id,
- Target variable = product,
- Sequence variable = transaction_time.

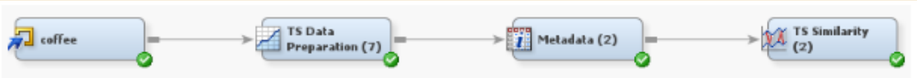
Rule Description

Map	Rule
RULE1	Jamaican Coffee River Rg ==> Brazilian Sm
RULE2	Brazilian Sm ==> Carmel syrup
RULE3	Jamaican Coffee River Rg ==> Carmel syrup
RULE4	Jamaican Coffee River Rg ==> Jamaican Coffee River Sm
RULE5	Morning Sunrise Chai Rg ==> Brazilian Lg
RULE6	Spicy Eye Opener Chai Rg ==> Brazilian Lg
RULE7	Sugar Free Vanilla syrup ==> Brazilian Rg
RULE8	Cappuccino ==> Carmel syrup
RULE9	Chocolate Croissant ==> Carmel syrup
RULE10	Columbian Medium Roast Sm ==> Carmel syrup
RULE11	Cappuccino Lg ==> Columbian Medium Roast Lg
RULE12	Ginger Biscotti ==> Cranberry Scone
RULE13	Sugar Free Vanilla syrup ==> Croissant
RULE14	Ethiopia Lg ==> Dark chocolate Lg
RULE15	Carmel syrup ==> Earl Grey Rg
RULE16	Hazelnut Biscotti ==> Earl Grey Rg
RULE17	Chocolate syrup ==> Ethiopia Lg
RULE18	Jamaican Coffee River Sm ==> Hazelnut Biscotti
RULE19	Jamaican Coffee River Rg ==> Latte Rg
RULE20	Peppermint Lg ==> Latte Rg
RULE21	English Breakfast Rg ==> Morning Sunrise Chai Lg
RULE22	Brazilian Lg ==> Oatmeal Scone
RULE23	Morning Sunrise Chai Rg ==> Oatmeal Scone
RULE24	Sustainably Grown Organic Rg ==> Ouro Brasileiro shot
RULE25	Chocolate Croissant ==> Peppermint Lg
RULE26	Morning Sunrise Chai Rg ==> Peppermint Lg
RULE27	Ethiopia Sm ==> Peppermint Rg
RULE28	Morning Sunrise Chai Lg ==> Peppermint Rg
RULE29	Earl Grey Rg ==> Spicy Eye Opener Chai Rg
RULE30	Morning Sunrise Chai Rg ==> Spicy Eye Opener Chai Rg
RULE31	Oatmeal Scone ==> Spicy Eye Opener Chai Rg
RULE32	Jamaican Coffee River Lg ==> Sustainably Grown Organic Rg
RULE33	Morning Sunrise Chai Rg ==> Sustainably Grown Organic Rg
RULE34	Cappuccino ==> Traditional Blend Chai Lg
RULE35	Jamaican Coffee River Rg ==> Traditional Blend Chai Lg
RULE36	Jamaican Coffee River Rg ==> Brazilian Sm ==> Carmel syrup
RULE37	Morning Sunrise Chai Rg ==> Oatmeal Scone ==> Spicy Eye Opener Chai Rg
RULE38	Brazilian Rg ==> Almond Croissant
RULE39	Croissant ==> Almond Croissant
RULE40	Ethiopia Lg ==> Almond Croissant
RULE41	Jamaican Coffee River Lg ==> Almond Croissant
RULE42	Jamaican Coffee River Sm ==> Almond Croissant
RULE43	Jumbo Savory Scone ==> Almond Croissant
RULE44	Serenity Green Tea Lg ==> Almond Croissant
RULE45	Sugar Free Vanilla syrup ==> Almond Croissant
RULE46	Sustainably Grown Organic Rg ==> Almond Croissant
RULE47	Chocolate Croissant ==> Brazilian Lg
RULE48	Chocolate syrup ==> Brazilian Lg
RULE49	Dark chocolate Lg ==> Brazilian Lg
RULE50	Dark chocolate Rg ==> Brazilian Lg

03

Time-series Analysis

We looked to identify trends in consumer behavior and product appeal by using generations and product categories as Cross IDs



- The insight provides understanding on each product group's popularity within the given time interval.
- When it comes to spending, Baby Boomers, Gen Z, X, and older Millennials all show less than younger Millennials.
- Gaining insight into the purchasing habits of different generations is essential for making strategic decisions and focusing marketing efforts.

Cross ID: Product Group

TSID Map Table

Time Series ID	Original Variable Name	Role	Variable Label	group
1	TS_1	TARGET	SALES 1	Add-ons
2	TS_2	TARGET	SALES 2	Beverages
3	TS_3	TARGET	SALES 3	Food
4	TS_4	TARGET	SALES 4	WholeBea

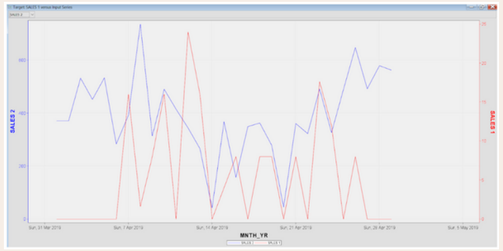
TSID Map Table for Product Group

Cross ID: Generation

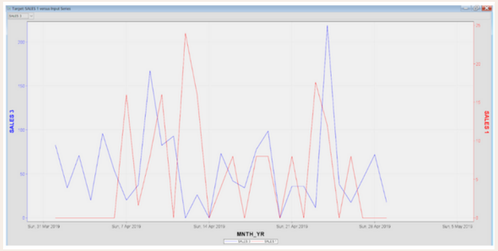
TSID Map Table

Time Series ID	Original Variable Name	Role	Variable Label	group
1	TS_1	TARGET	SALES 1	BabyBoomers
2	TS_2	TARGET	SALES 2	GenX
3	TS_3	TARGET	SALES 3	GenZ
4	TS_4	TARGET	SALES 4	OlderMillennials
5	TS_5	TARGET	SALES 5	YoungerMillennials

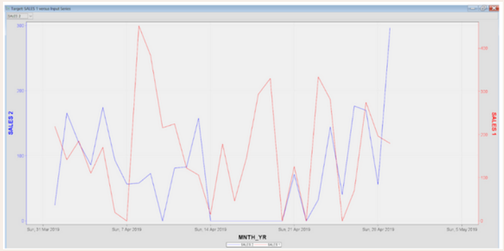
TSID Map Table for Product Group



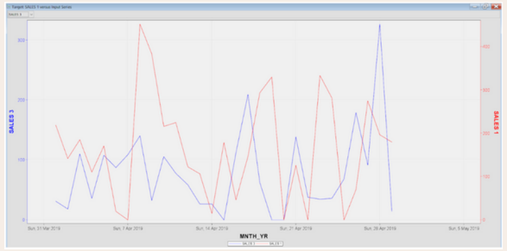
Time Series Analysis of Sales1 vs Sales2



Time Series Analysis of Sales1 vs Sales3



Time Series Analysis of Sales1 vs Sales2



Time Series Analysis of Sales1 vs Sales3

TOOLS:

SAS Enterprise Miner

MODIFY

Objective

Preprocess and transform the data to prepare it for modelling.

Importance

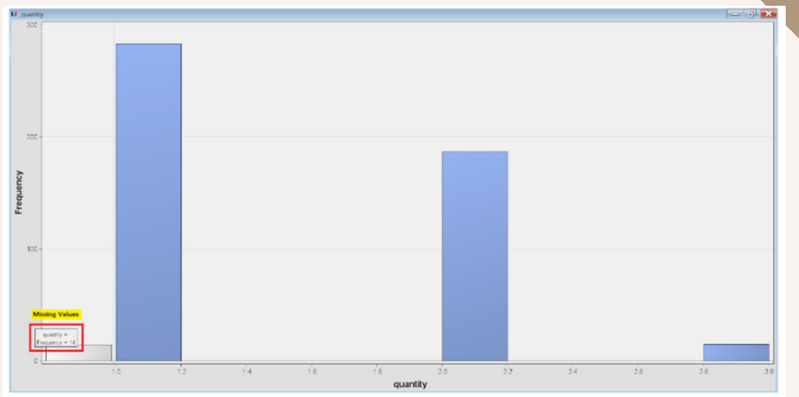
Modification ensures that the data is clean, relevant, and properly formatted, improving the accuracy and effectiveness of the subsequent modelling step.

Techniques

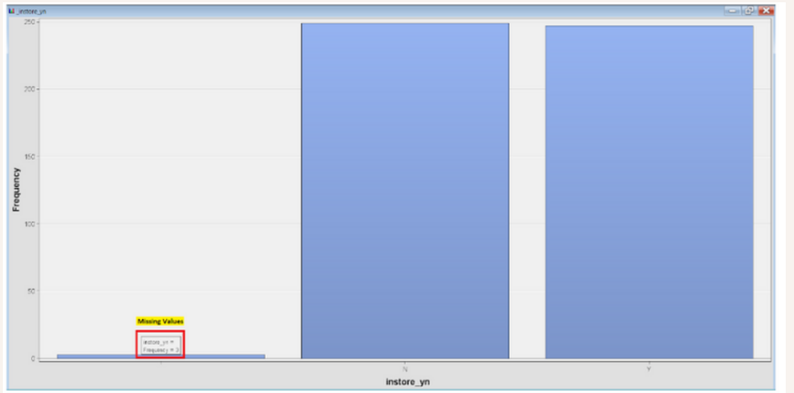
Data Cleaning, Feature Engineering, Data Scaling and Normalization

Column	Data Type	Number of Missing Values
sales_id	Integer	0
transaction_id	Integer	0
transaction_date	Date	0
transaction_time	Time	0
customer_id	Integer	0
customer_since	Date	0
product_id	Integer	0
unit_price	Integer	0
product_group	String	0
product	String	0
current_retail_price	Integer	0
promo_yn	String	0
new_product_yn	String	0
customer_firstName	String	0
generation	String	0
gender	String	0
birth_year	Integer	0
quantity	Integer	14
instore_yn	String	3

Missing values for each column



14 missing values for quantity



3 missing values for height

TOOLS:

SAS Enterprise Miner

Impute Missing Data

The Process

- Address missing values in datasets.
- Common in real-world datasets due to incomplete records, data collection errors, and system failures.

The Approach

01 Imputation Techniques

- Methods to replace missing values for a more complete dataset.
- Enhance dataset usefulness and completeness.

02 Mean Imputation Method:

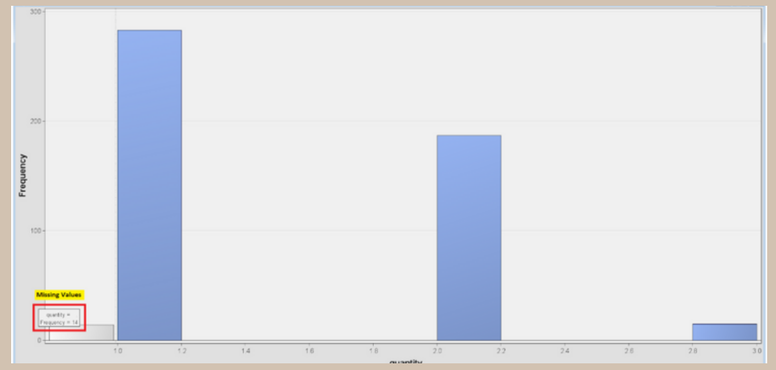
- Simple and easy to implement.
- Suitable for continuous numeric variables (e.g., height, age).
- Maintains original distribution by substituting the mean of observed values for missing values.
- Preserves the central tendency of the dataset.

TOOLS:

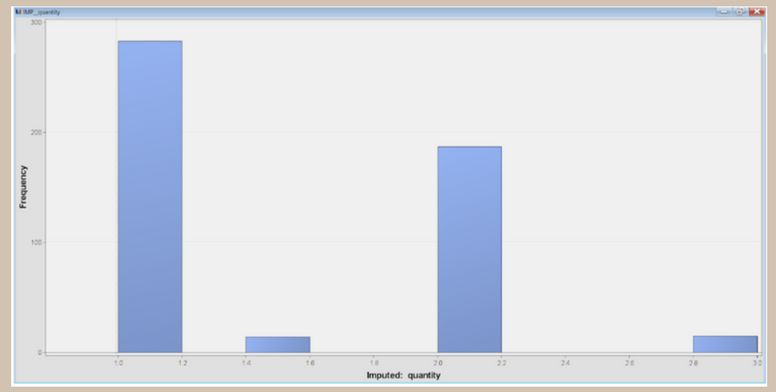
SAS Enterprise Miner

Impute Missing Data

Variable: Quantity

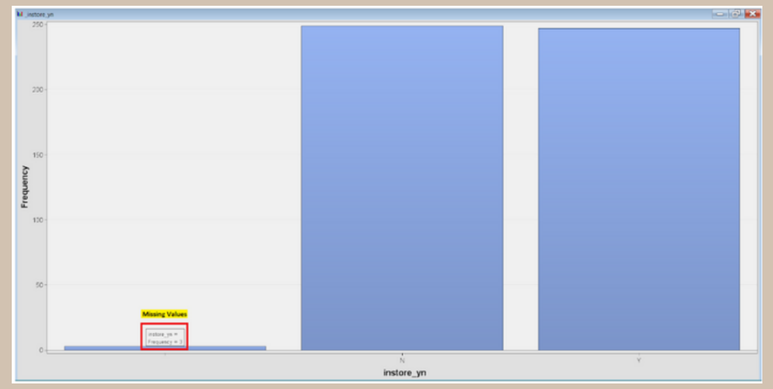


Before: 14 missing values before imputed

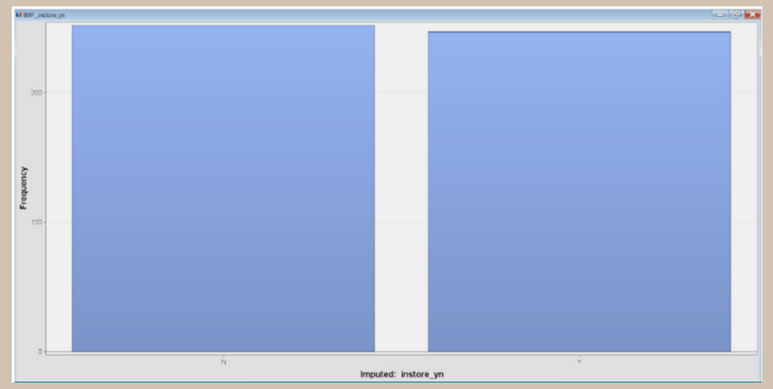


After: Solved 14 missing values after imputed

Variable: instore_yn



Before: 3 missing values before imputed



After: Solved 3 missing values after imputed

Filtering Outliers

Filter the outlier

Results:

Filter Limits for Interval Variables
(maximum 500 observations printed)

Variable	Role	Minimum	Maximum	Filter Method	Keep Missing Values	Label
birth_year	INPUT	1928.30	2021.54	STDDEV	Y	
current_retail_price	INPUT	-5.44	12.92	STDDEV	Y	
customer_COUNT_sales_	INPUT	-11.14	20.95	STDDEV	Y	customer.COUNT(sales)
customer_MEAN_sales__IMP__quanti	INPUT	0.27	2.62	STDDEV	Y	customer.MEAN(sales. IMP__quanti
customer_MEAN_sales_unit_price_	INPUT	-2.15	9.28	STDDEV	Y	customer.MEAN(sales.unit_price)
customer_SUM_sales__IMP__quantit	INPUT	-13.69	27.25	STDDEV	Y	customer.SUM(sales. IMP__quantit
customer_SUM_sales_unit_price_	INPUT	-52.77	89.42	STDDEV	Y	customer.SUM(sales.unit_price)
product_COUNT_sales_	INPUT	-1.21	20.28	STDDEV	Y	product.COUNT(sales)
product_SUM_sales__IMP__quantity	INPUT	-4.45	32.78	STDDEV	Y	product.SUM(sales. IMP__quantity
product_SUM_sales_unit_price__sa	INPUT	-9.24	68.55	STDDEV	Y	product.SUM(sales.unit_price)(sa
quantity	INPUT	-0.20	3.09	STDDEV	Y	
unit_price	INPUT	-5.64	12.77	STDDEV	Y	

Excluded Class Values
(maximum 500 observations printed)

Variable	Role	Level	Train Count	Train Percent	Label	Filter Method
product_group	INPUT	MERCHANDI	4	0.80160		MINPCT

Number Of Observations

Data Role

Role	Filtered	Excluded	DATA
TRAIN	447	52	499

Outliers, representing extreme or unusual values in a dataset, pose challenges in data analysis and machine learning. Since algorithms frequently need numerical inputs, handling outliers correctly is essential for making accurate decisions.

Tool: SAS Enterprise Miner

1. The dataset undergoes filtration using the 'Filter' node, employing the 'Standard Deviations from the Mean' method in order to filter the outliers.
2. 52 values have been recognized as outliers, and 447 values is selected.

TOOLS:

SAS Enterprise Miner

Encoding Categorical Variables

Data analysis and machine learning methods require numerical inputs, making categorical variables difficult. It's crucial to encode categorical variables appropriately in machine learning tasks so algorithms can accurately and meaningfully use them in decision-making.

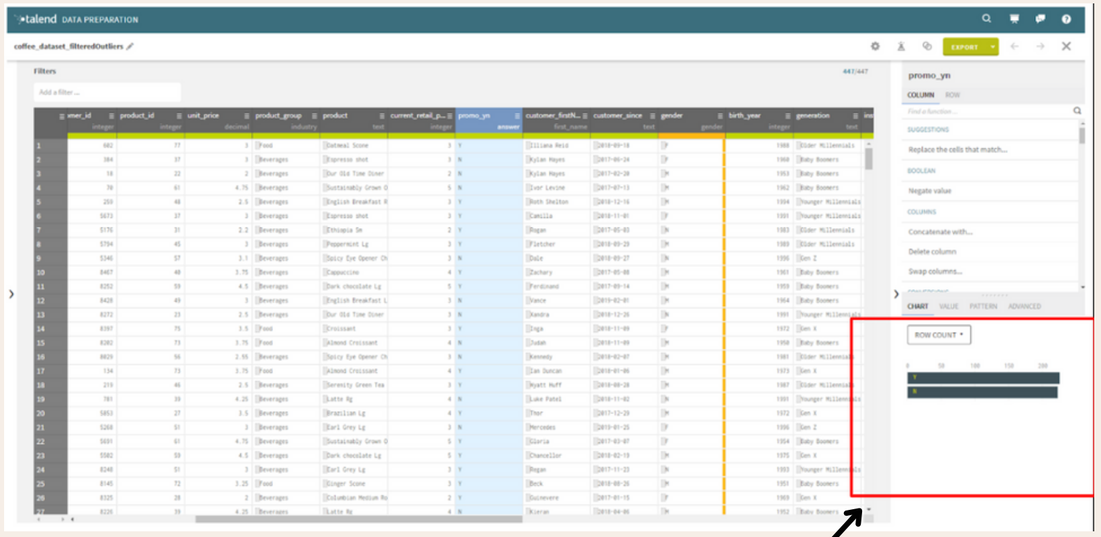
TOOLS

Talend Preperation

Tool: Talend Data Preparation

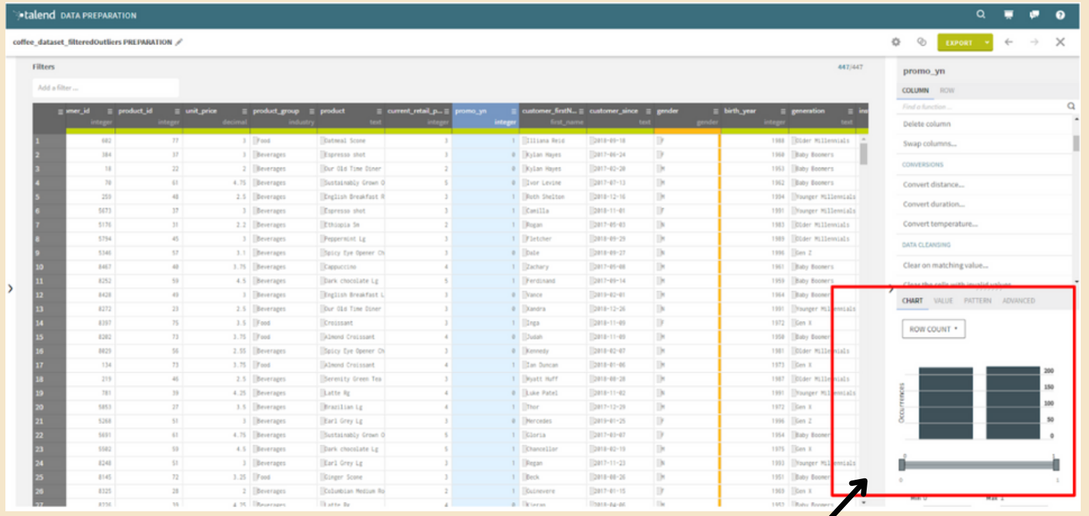
Column Involved:
Variable: promo_yn, transaction_date, birthyear current_retail_price, promo_yn, new_product_yn, quantity, instore_yn, customer_COUNT_sales, customer_MEAN_sales_IMP_quantity, customer_MEAN_sales_unit_price, customer_MODE_sales_IMP_instore_yn

Before



Before encoding:
(data type: answer)

After



After encoding:
(data type: Integer)

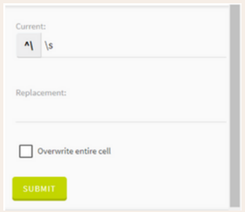
Standardize Data Types and Format

Data types and formats must be standardized for consistency, correctness, and reliability. Transforming different attributes entails formatting variables to match standardized data formats like dates, decimals, and integers. Standardized data types and formats improve data quality by making missing values and outliers easier to spot.

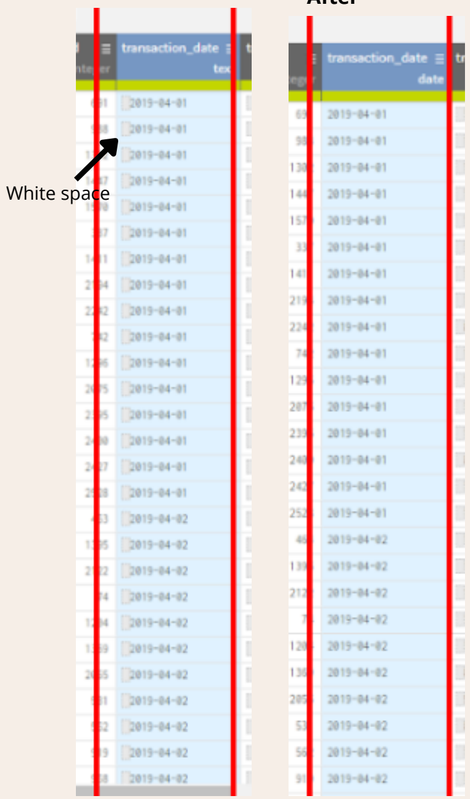
Column Involved:
Age, Height, Weight, Variable_FCVC, Variable_CH20, Variable_FAF, Variable_TUE

The Process

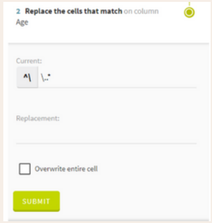
01 Most data have a white space in front, thus we use RegEx on each column to remove the string.



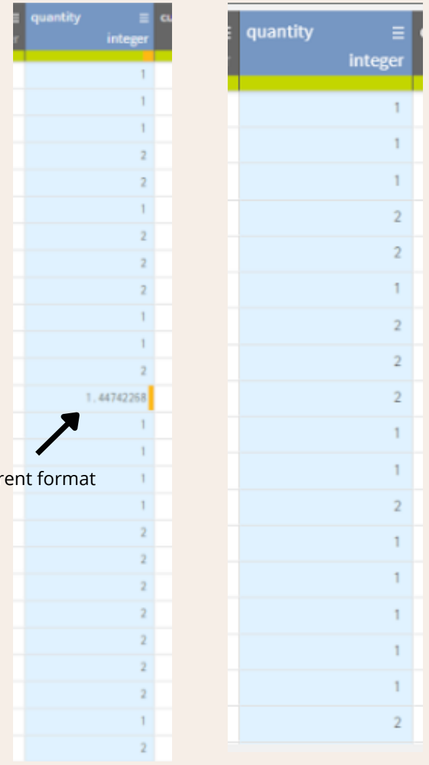
Before **After**



02 Standardised the data type for each column with suitable data type.



Before **After**



TOOLS

Talend Preperation

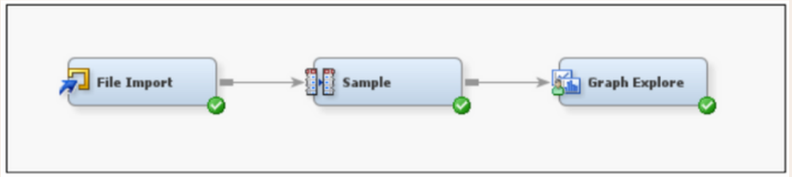
MODEL

Unbalanced Class

- generation
- product_group

If the training data comprises unequal target variable classes, the model can favor the majority class and perform badly on a more balanced testing set.

→ Resampling techniques using **stratify sampling** on generation and product_group



Resampling method

Train	
Variables	
Output Type	Data
Sample Method	Stratify
Random Seed	12345
Size	
Type	Percentage
Observations	.
Percentage	100.0
Alpha	0.01
PValue	0.01
Cluster Method	Random
Stratified	
Criterion	Equal
Ignore Small Strata	No
Minimum Strata Size	5

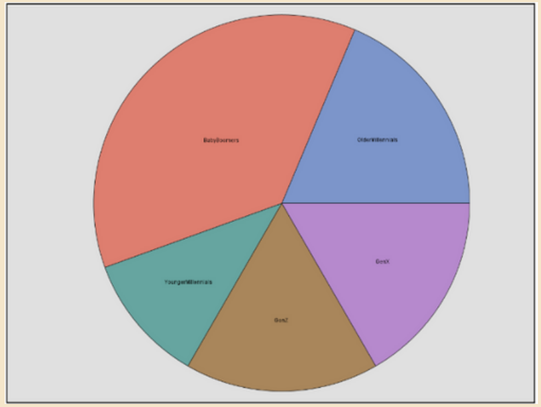
Setting to stratify sampling

TOOLS:

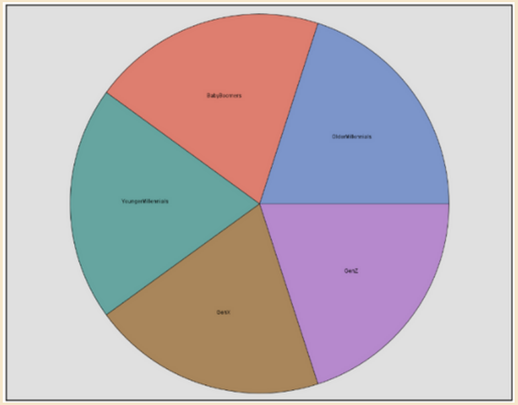
SAS Enterprise Miner

Result

Target Variable: generation

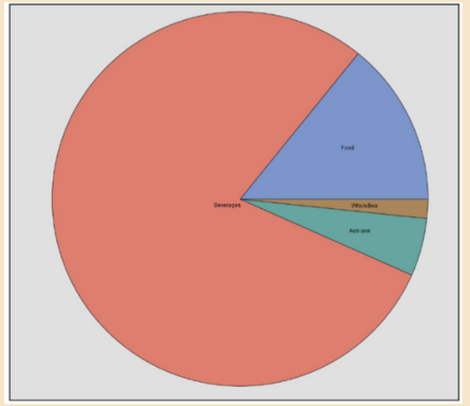


Before

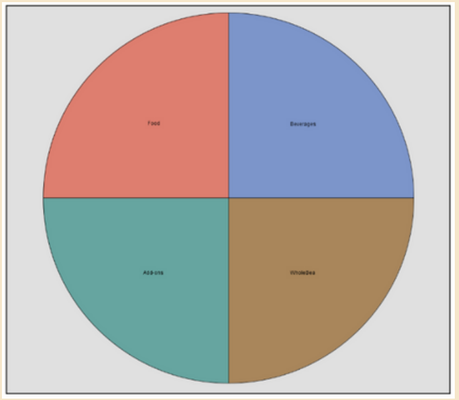


After

Target Variable: product_group



Before

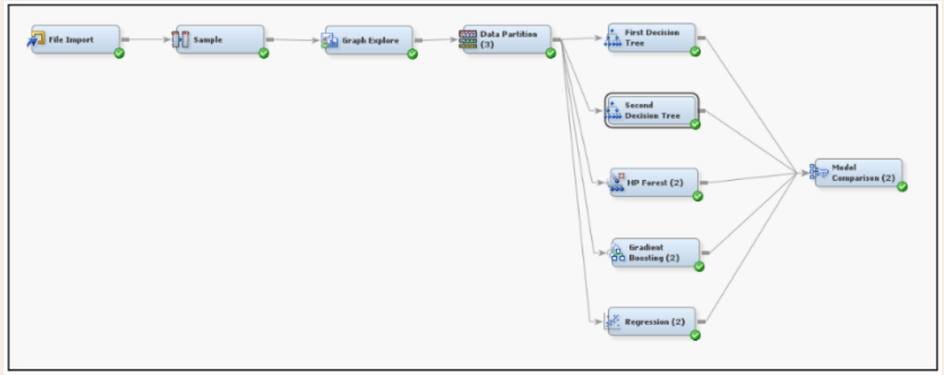


After

Modelling

To interpret data patterns, statistical or machine learning models are created. This step involves developing and testing prediction models to find dataset correlations, trends, and linkages.

- Decision Tree
- Random Forest
- Gradient Boosting



Whole Modeling SAS Diagram

Data Set Allocations	
Training	70.0
Validation	30.0
Test	0.0

Setting for Data Partition Node

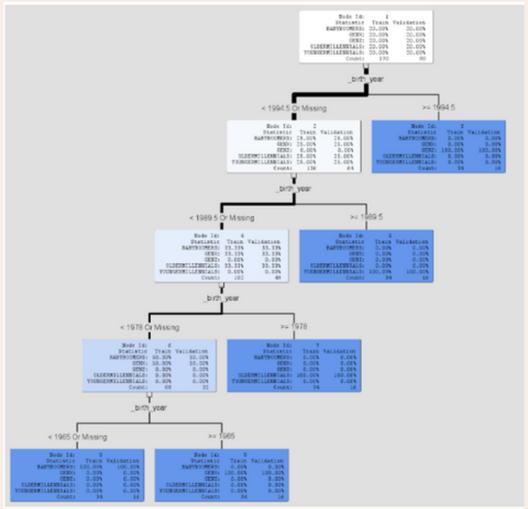
TOOLS:

SAS Enterprise Miner

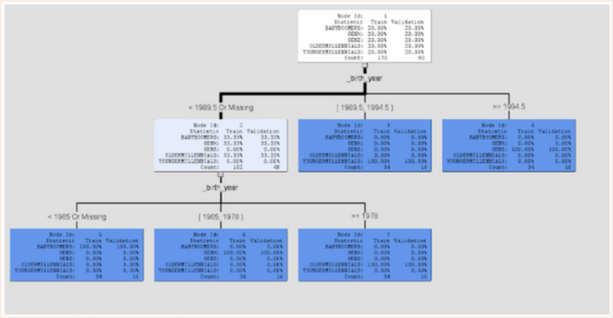
Decision Tree

Decision trees serve diverse functions: classification, regression, feature selection, interpretability, ensemble methods, data exploration, handling missing values, informing strategic decisions, pattern recognition, and scalability for large datasets.

Target Variable: generation



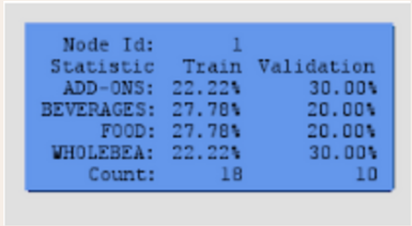
First Decision Tree



Second Decision Tree

- Classify into generational categories based on relevant features
- Identify key features into distinct generational groups.
- Determine the behavioral and preference pattern for each generation

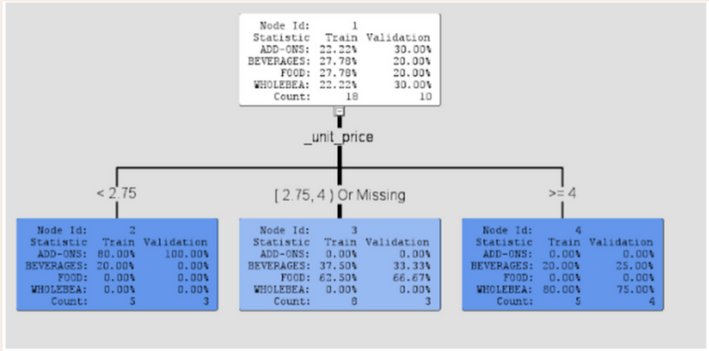
Target Variable: product_group



Node Id:	1	
Statistic	Train	Validation
ADD-ONS:	22.22%	30.00%
BEVERAGES:	27.78%	20.00%
FOOD:	27.78%	20.00%
WHOLEBEA:	22.22%	30.00%
Count:	18	10

First Decision Tree

- Categorize into different product groups based on relevant features
- Determine the feature for product group that focus on key characteristics.
- Get insights into potential correlations in customer choices



Second Decision Tree

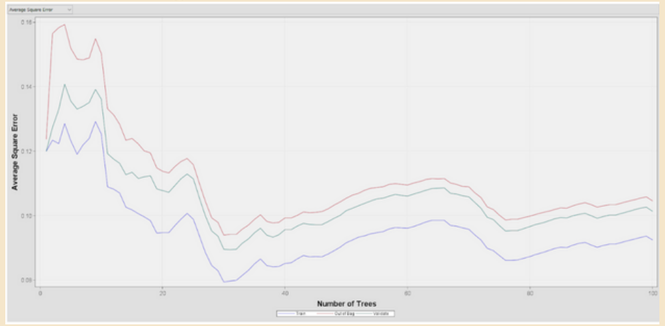
Modelling

TOOLS:

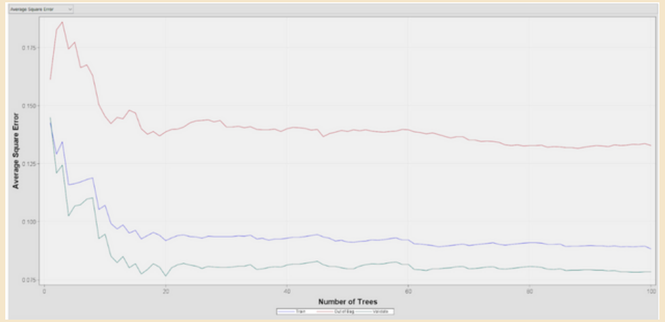
SAS Enterprise Miner

Random Forest

Target Variable: generation



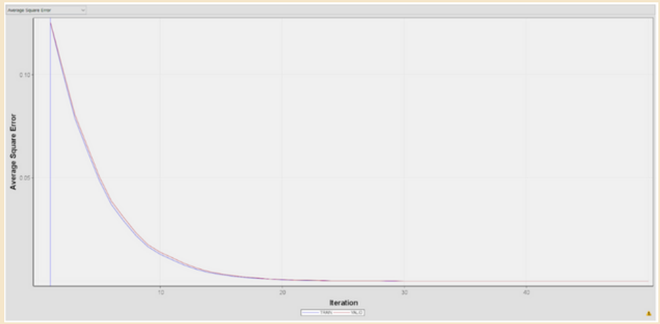
Target Variable: product_group



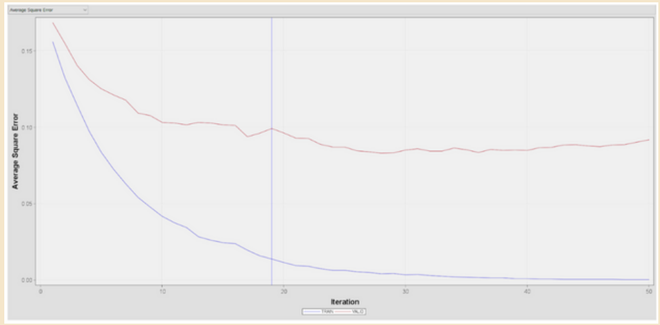
- At around 30 trees for 'generation' and 20 trees for 'product group,' the chart reveals minimum average square error
- However, beyond this point, error fluctuates with both increasing and decreasing patterns
- Suggesting that adding more trees may not consistently enhance predictive performance and could pose a risk of overfitting

Gradient Boosting

Target Variable: generation



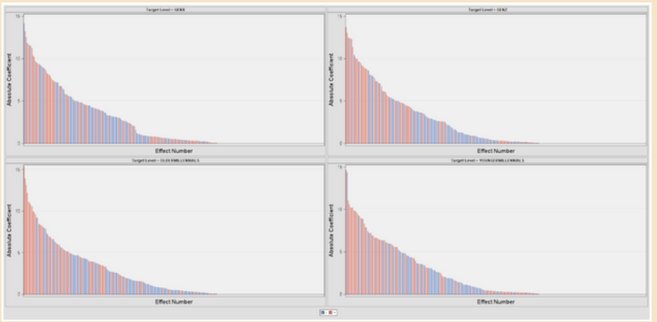
Target Variable: product_group



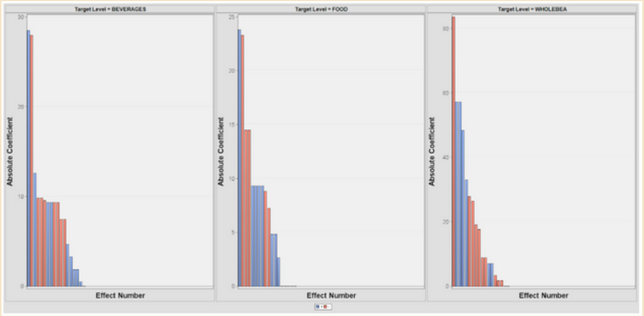
- After around 25 iterations for 'generation' and 35 iterations for 'product group,' the models reach peak accuracy, with subsequent iterations showing diminishing returns
- Understanding this is crucial for optimal model configuration
- For 'product group,' a refined pattern post-minimum error emphasizes the need for careful iteration balancing to capture meaningful patterns without overfitting.

Logistic Regression

Target Variable: generation



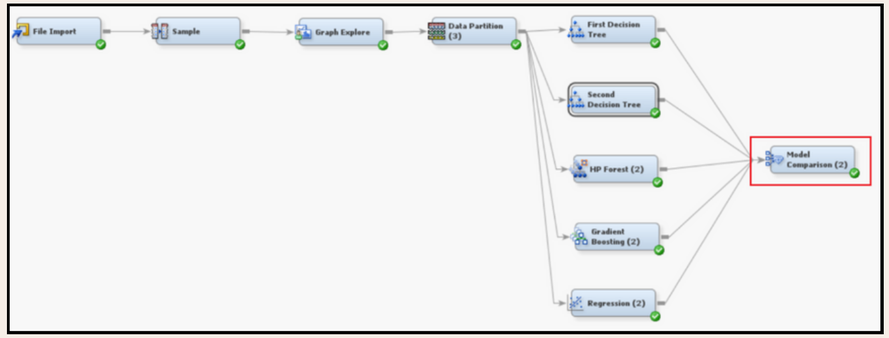
Target Variable: product_group



- The logistic regression chart highlights Older Millennials with the highest slope, indicating the greatest impact on 'generation.' Focusing on this demographic can enhance product popularity predictions and drive sales.
- For 'product group,' Whole Bread's steep slope suggests it has the highest effect, making it a strategic focus for logistics and promotions. Meanwhile, Beverages, with a higher coefficient than Food, warrants attention for tailored promotions and inventory predictions.

Assess

Model Comparison



Data Flow

TOOLS:

SAS Enterprise Miner

generation

Fit Statistics
Model Selection based on Valid: Misclassification Rate (_VMISC_)

Selected Model	Model Node	Model Description	Valid: Misclassification Rate	Train: Average Squared Error	Train: Misclassification Rate	Valid: Average Squared Error
Y	Tree1	Decision Tree (1)	0.0	0.00000	0.000000	0.00000
	Tree2	Decision Tree (2)	0.0	0.00000	0.000000	0.00000
	Boost2	Gradient Boosting (2)	0.0	0.12536	0.000000	0.12536
	HPDMForest2	HP Forest (2)	0.1	0.09249	0.017647	0.10136
	Reg2	Regression (2)	0.8	0.00000	0.000000	0.16000

Decision Tree (1) and Decision Tree (2):

- Perfect misclassification rate on validation.
- 0.00000 average squared error on validation.

Gradient Boosting (2) and HP Forest (2):

- Strong performance of misclassification rates, competitive average squared.
- Higher misclassification rate on validation, but low average squared error.

Logistic Regression (2):

- Higher misclassification rate of 0.8.
- Lower average squared error of 0.000.

Insights

Decision Tree (1) and Decision Tree (2) stand out for their flawless performance.

product_group

Fit Statistics
Model Selection based on Valid: Misclassification Rate (_VMISC_)

Selected Model	Model Node	Model Description	Valid: Misclassification Rate	Train: Average Squared Error	Train: Misclassification Rate	Valid: Average Squared Error
Y	HPDMForest3	HP Forest (2)	0.1	0.08844	0.16667	0.07844
	Tree2	Decision Tree (2)	0.2	0.09653	0.27778	0.07759
	Boost3	Gradient Boosting (2)	0.2	0.13274	0.05556	0.15530
	Reg3	Regression (2)	0.7	0.00000	0.00000	0.16312
	Tree1	Decision Tree (1)	0.8	0.18673	0.72222	0.19105

Decision Tree (1) and Decision Tree (2):

- 0.8 and 0.2 misclassification rate on validation.
- 0.19105 and 0.07759 average squared error on validation.

Gradient Boosting (2) and HP Forest (2):

- Misclassification rate of 0.2 and 0.1 respectively.
- A corresponding average squared error of 0.13274 and 0.08844 respectively.

Logistic Regression (2):

- Misclassification rate of 0.7.
- A corresponding average squared error of 0.16312.

Insights

HP Forest (2) stood out with a perfect misclassification rate and low average squared error on the validation dataset.

Conclusion and Future Works

Throughout the project, we concluded to suggest effective problem solving, insights from consumer behavior and optimization efficiency.

Key Findings

Effective Problem Solving

- Addressed crucial difficulties in the coffee shop business.
- Utilized modern analytical methodologies and technologies.

Operational Efficiency

- Studied daily coffee purchase patterns.
- Increased operational efficiency and resource distribution.

Insights from Consumer Behavior

- Investigated consumer behavior across generations.
- Customized marketing tactics and product offers based on insights.

Optimization Strategies

- Examined popularity of items across age groups.
- Optimized product selection and improved inventory control.

Opportunities for Improvement

Seasonal Purchasing Pattern

- Conduct more in-depth investigation into seasonal purchasing patterns.
- Optimize inventory management during specific time periods.

Flexibility and Adaptation

- Collaborate with other data sources for customer preference changes.
- Ensure ongoing monitoring of sector changes.

Project Impact

- We are able to provide a solid basis for continued improvements in coffee shop operations.
- We successfully manage to integrate all the tools that have been learnt during lecture which are SAS Enterprise Miner, Talend Data Integration, FeatureTools, Python, and Talend Data Preparation to ensure we fulfill the requirements of the business golas for this project.

References

Bhandari, P. (2023, June 21). Descriptive statistics: Definitions, types, examples. Scribbr. https://www.scribbr.com/statistics/descriptive-statistics/

Brownlee, J. (2020, August 14). A gentle introduction to the gradient boosting algorithm for machine learning. MachineLearningMastery.com. https://machinelearningmastery.com/gentle-introduction-gradient-boosting-algorithm-machine-learning/

Chang, J. (2019, November 8). Coffee shop sample data (11.1.3+). Kaggle. https://www.kaggle.com/datasets/ylchang/coffee-shop-sample-data-1113/data

Fontanella, C. (2022, August 18). A beginner's Guide to Customer Behavior Analysis. HubSpot Blog. https://blog.hubspot.com/service/customer-behavior-analysis

Insider Learning Machines. (2023, November 6). How to interpret decision trees with 1 simple example. Inside Learning Machines. https://insidelearningmachines.com/interpret_decision_trees/

Iqbal, J. (2022, August 16). What are the best statistical models to use for demand forecasting?. Causometrix. https://www.causometrix.com/what-are-the-best-statistical-models-to-use-for-demand-forecasting/

Javatpoint. (n.d.). Machine learning random forest algorithm - javatpoint. www.javatpoint.com. https://www.javatpoint.com/machine-learning-random-forest-algorithm

Lawton, G., Burns, E., & Rosencrance, L. (2022, January 20). What is logistic regression? - definition from Searchbusinessanalytics. Business Analytics. https://www.techtarget.com/searchbusinessanalytics/definition/logistic-regression

Mab, MabMab 4111 silver badge33 bronze badges, rolando2rolando2 12.2k11 gold badge4242 silver badges6161 bronze badges, RickyBRickyB 1, & AshOfFireAshOfFire 57033 silver badges1010 bronze badges. (1964, May 1). How to visualise coefficients of a binomial logistic regression?. Cross Validated. https://stats.stackexchange.com/questions/342627/how-to-visualise-coefficients-of-a-binomial-logistic-regression

Saini, A. (2024, January 5). Decision tree - a step-by-step guide. Analytics Vidhya. https://www.analyticsvidhya.com/blog/2021/08/decision-tree-algorithm/

THANK YOU