MIS 545: Data Mining for Business Intelligence

CUSTOMER SEGMENTATION & CUSTOMER TREND ANALYSIS

Group 13

MEET OUR TEAM

NEHA

PRANIT

RISHABH

TEJASWINI

YOGITA

TABLE OF CONTENTS

PROBLEM STATEMENT

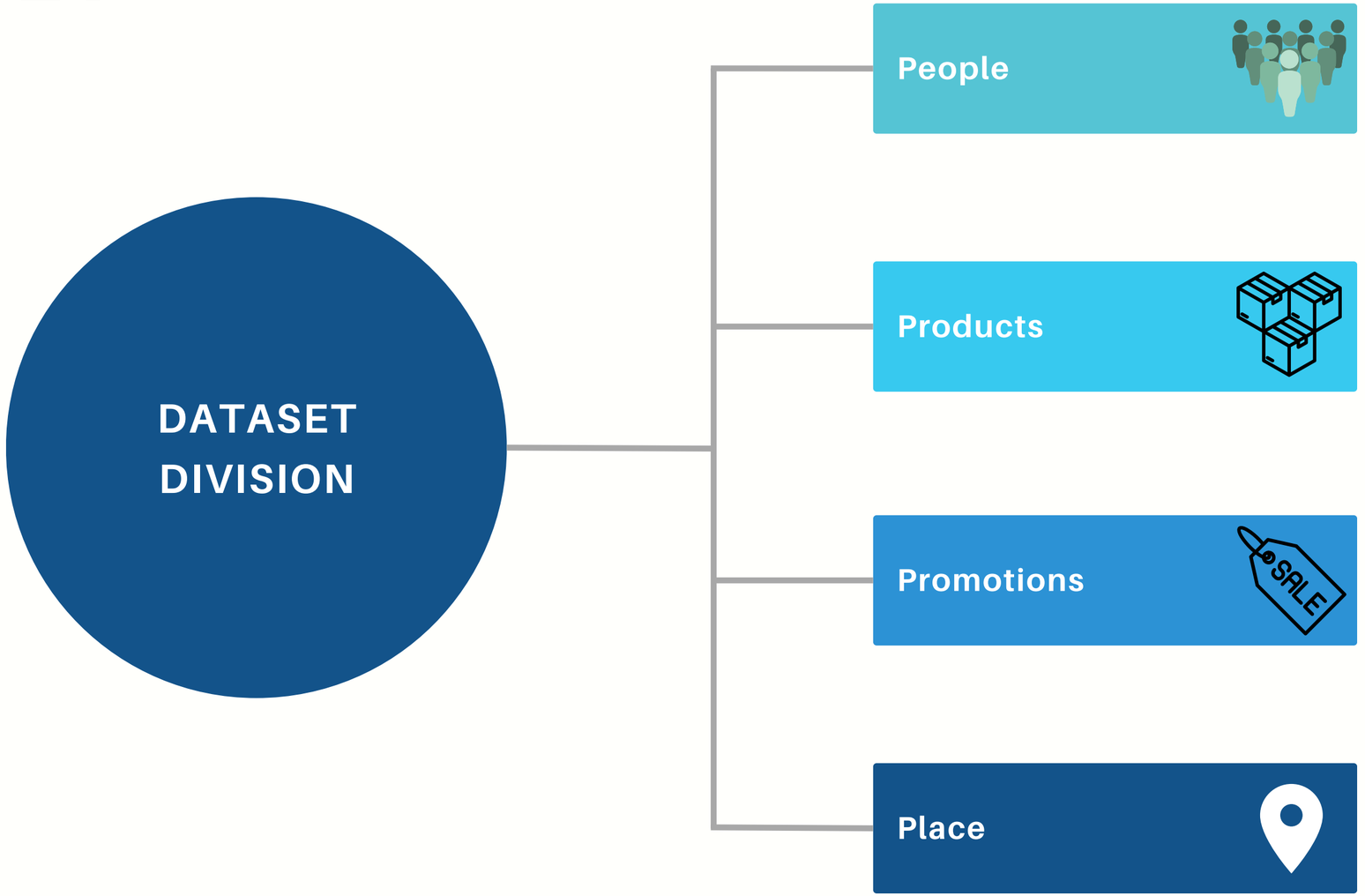
Challenges:

- Dynamic consumer adaptation poses significant challenge.
- Mass marketing hampers current market dynamics.
- Lack of concise Trend Analysis hinders adaptability.
- Struggle in targeting specific customer segments efficiently.

Recommendations:

- Develop real-time insights for Trend Analysis.
- Implement precise segmentation for targeting customers.
- Utilize Customer Personality Analysis for ideal understanding.
- Drive success by tailoring strategies with customer insights.

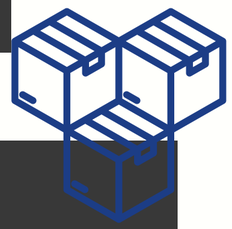
DATASET



DATASET

People:

- `ID`: Customer's unique identifier
- `Year_Birth`: Customer's birth year
- `Education`: Customer's education level
- `Marital_Status`: Customer's marital status
- `Income`: Customer's yearly household income
- `Kidhome`: Number of children in customer's household
- `Teenhome`: Number of teenagers in customer's household
- `Dt_Customer`: Date of customer's enrollment with the company
- `Recency`: Number of days since customer's last purchase
- `Complain`: 1 if the customer complained in the last 2 years, 0 otherwise



Products:

- `MntWines`: Amount spent on wine in last 2 years
- `MntFruits`: Amount spent on fruits in last 2 years
- `MntMeatProducts`: Amount spent on meat in last 2 years
- `MntFishProducts`: Amount spent on fish in last 2 years
- `MntSweetProducts`: Amount spent on sweets in last 2 years
- `MntGoldProds`: Amount spent on gold in last 2 years

Promotion:

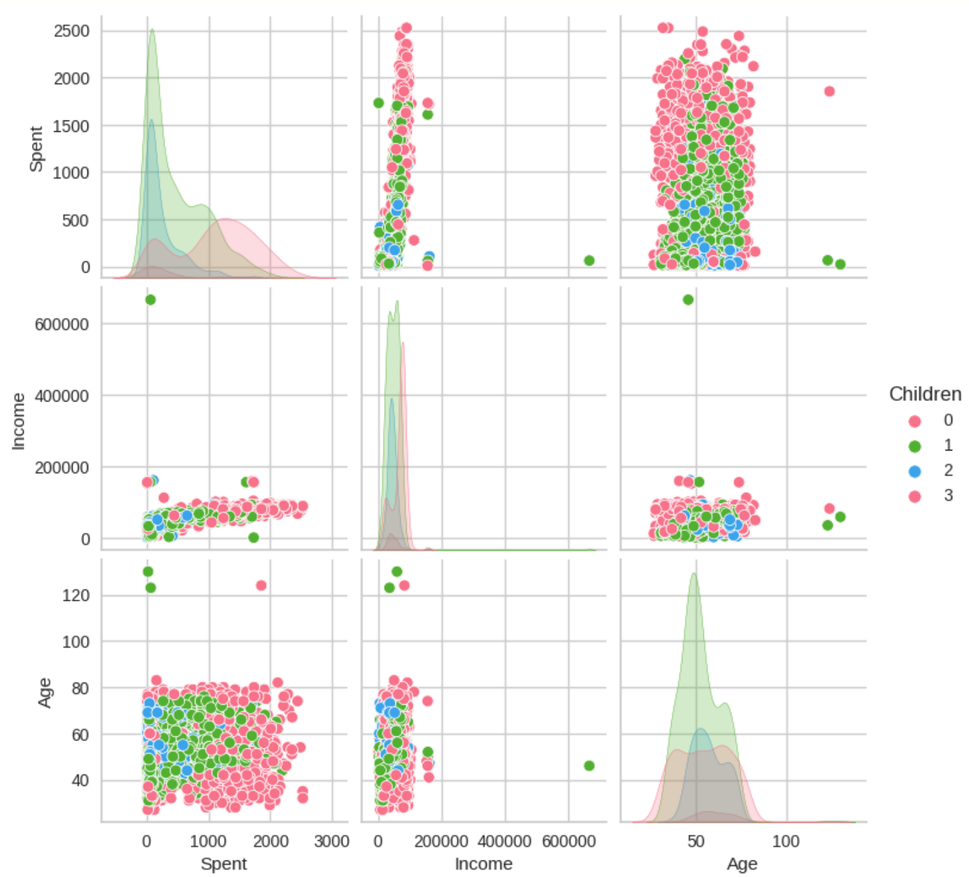
- `NumDealsPurchases`: Number of purchases made with a discount
- `AcceptedCmp1`: 1 if customer accepted the offer in the 1st campaign, 0 otherwise
- `AcceptedCmp2`: 1 if customer accepted the offer in the 2nd campaign, 0 otherwise
- `AcceptedCmp3`: 1 if customer accepted the offer in the 3rd campaign, 0 otherwise
- `AcceptedCmp4`: 1 if customer accepted the offer in the 4th campaign, 0 otherwise
- `AcceptedCmp5`: 1 if customer accepted the offer in the 5th campaign, 0 otherwise
- `Response`: 1 if customer accepted the offer in the last campaign, 0 otherwise

Place:

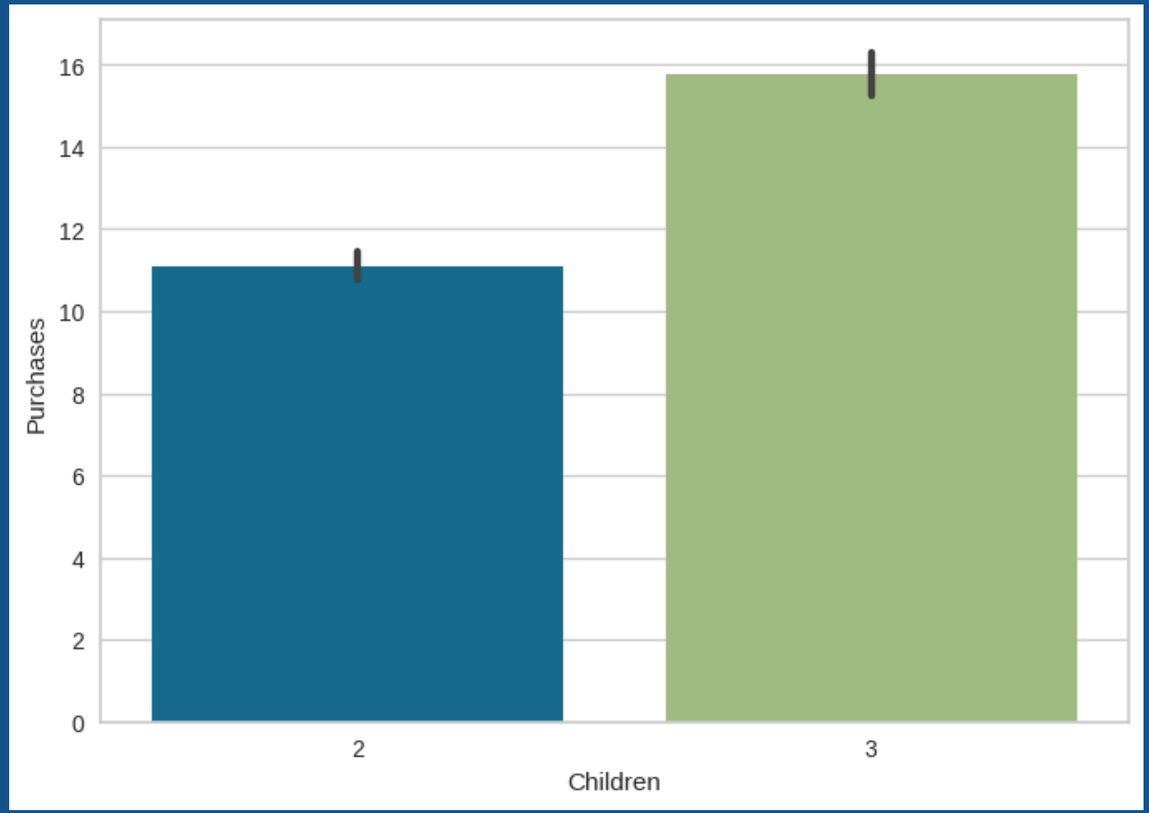
- `NumWebPurchases`: Number of purchases made through the company's website
- `NumCatalogPurchases`: Number of purchases made using a catalogue
- `NumStorePurchases`: Number of purchases made directly in stores
- `NumWebVisitsMonth`: Number of visits to company's website in the last month

ANALYSIS & VISUALIZATION

Demographic Analysis of Spending, Income, and Age



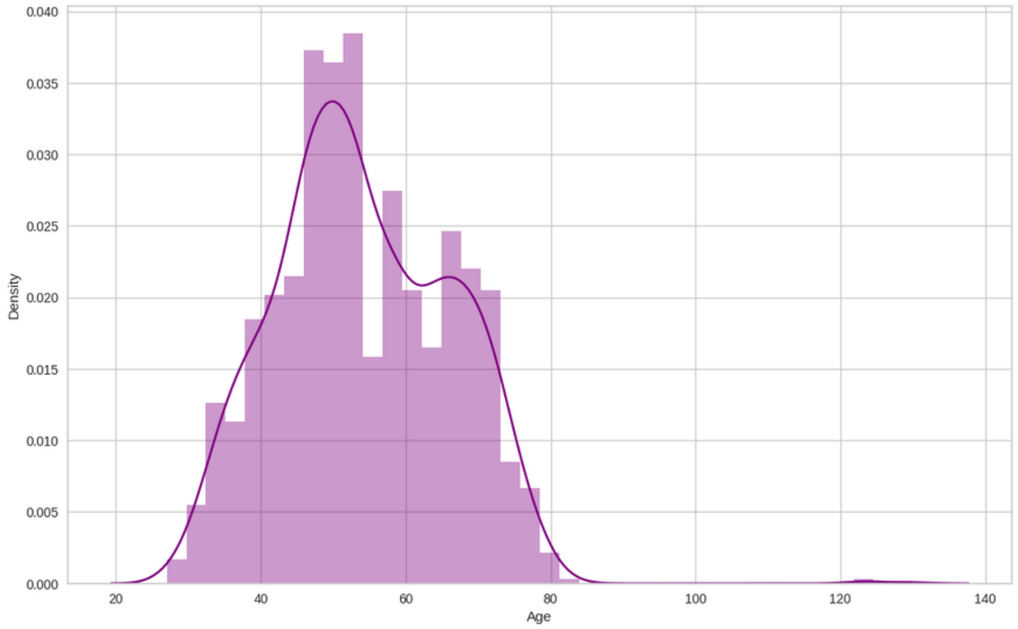
Shopping Habits Based on Family Size

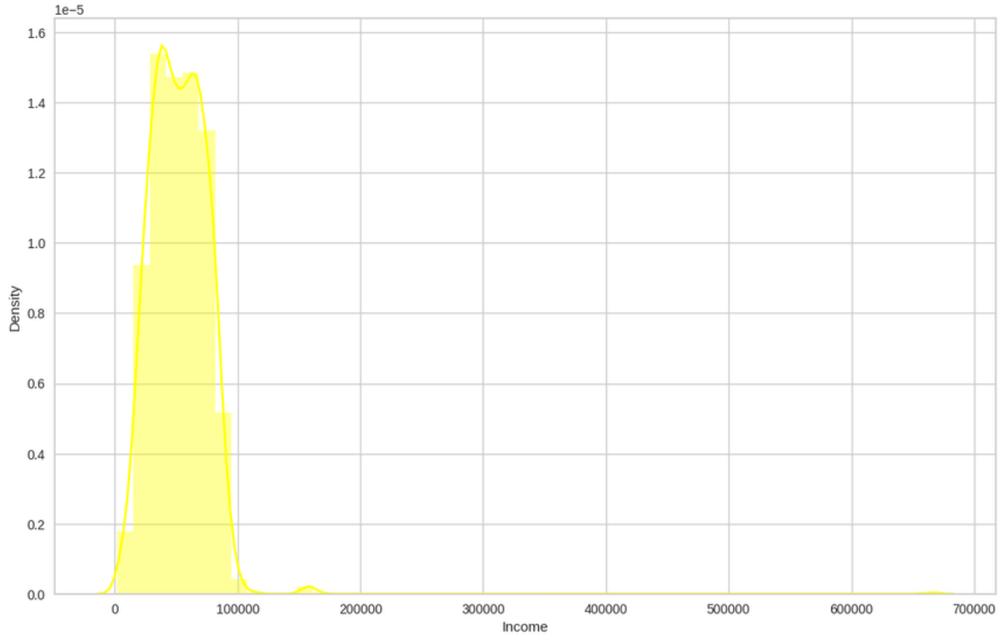


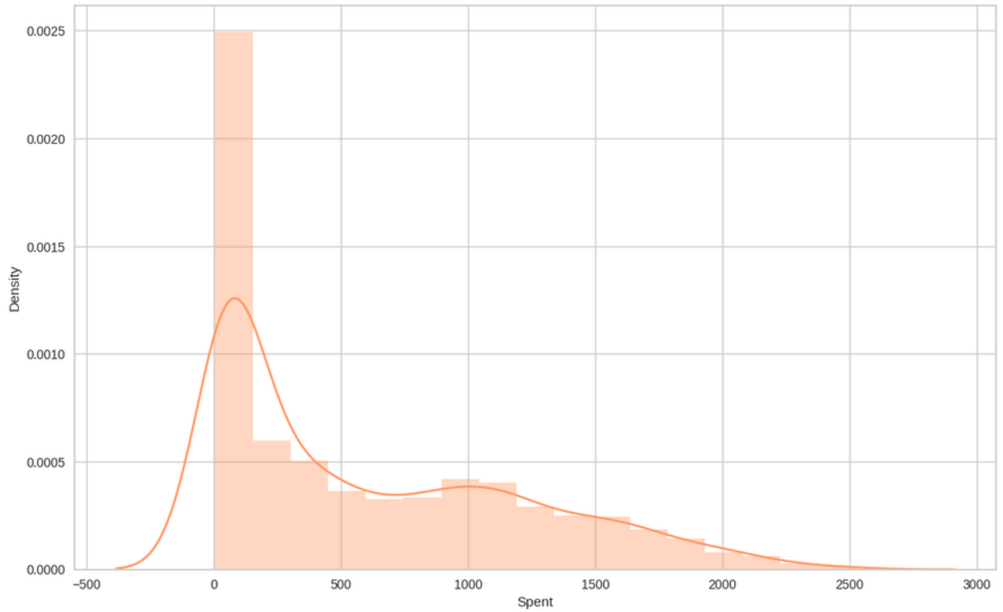
ANALYSIS & VISUALIZATION

Detecting the Presence of Outliers:

- Density Distributions by Age, Income, and Spent

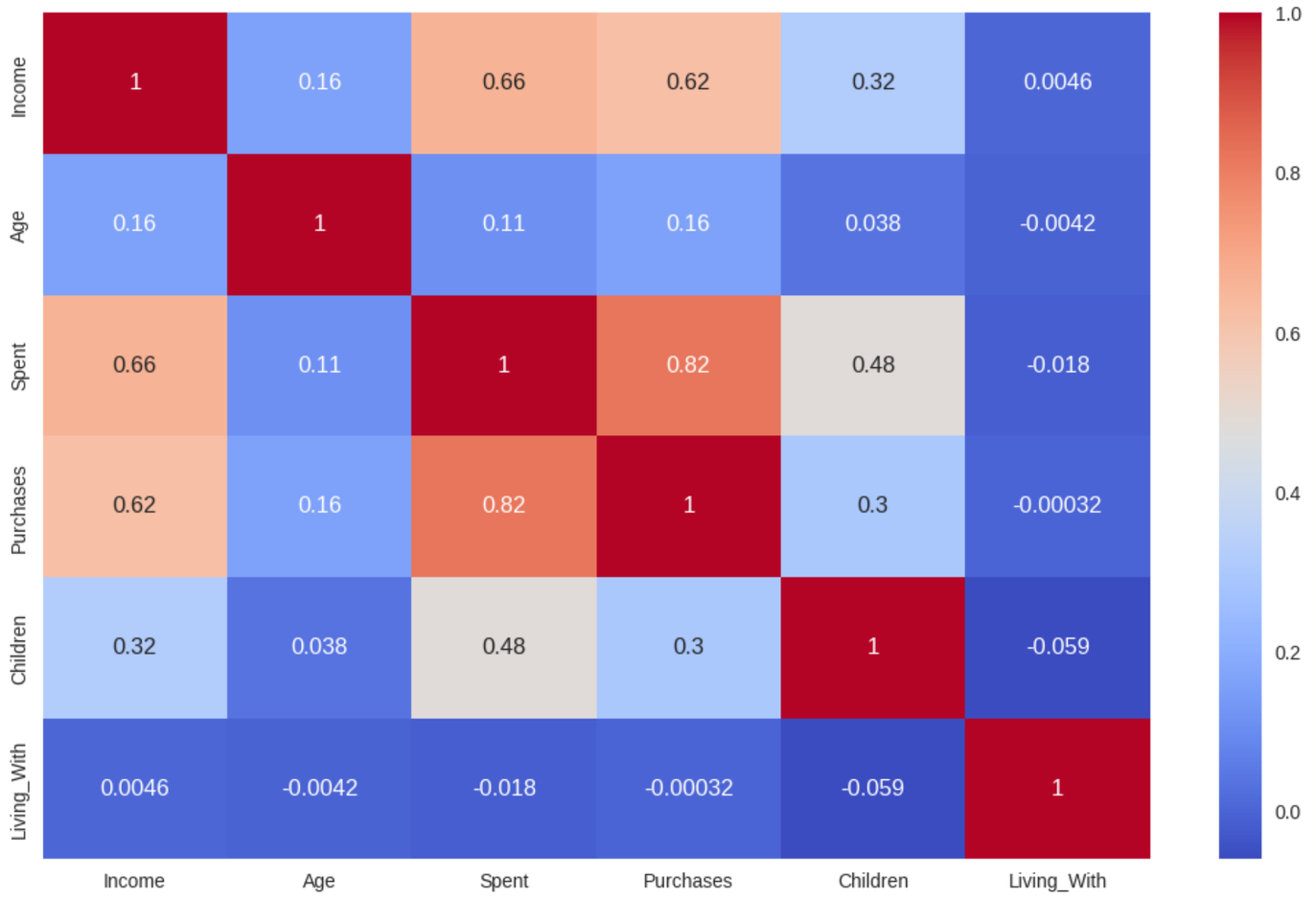






ANALYSIS & VISUALIZATION

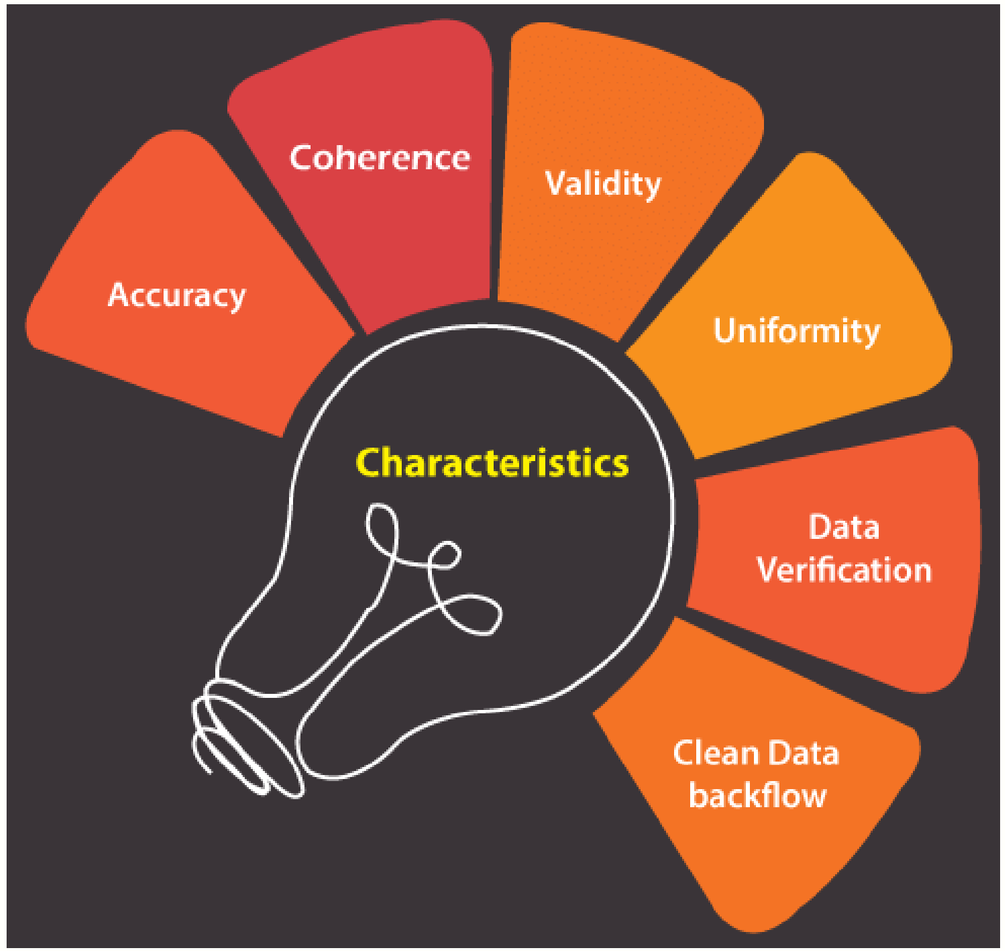
Correlation Matrix



Data Cleaning

Data cleaning is the process of correcting or deleting inaccurate, damaged, improperly formatted, duplicated, or insufficient data from a dataset. Even if results and algorithms appear to be correct, they are unreliable if the data is inaccurate. There are numerous ways for data to be duplicated or incorrectly labeled when merging multiple data sources.

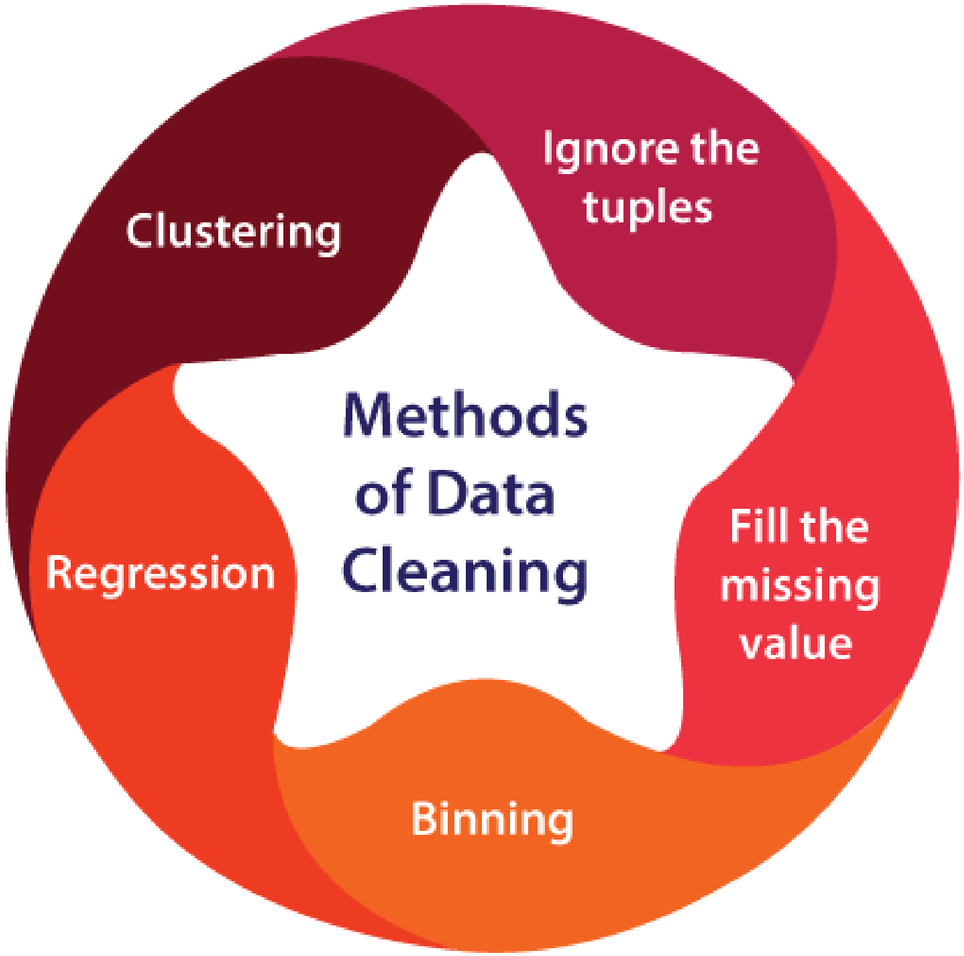
Data Cleaning Characteristics



To ensure the correctness, integrity, and security of corporate data, data cleaning is a requirement. These may be of varying quality depending on the properties or attributes of the data. The key components of data cleansing in data mining are as follows:

- Accuracy
- Coherence
- Validity
- Uniformity
- Data Verification
- Clean Data Backflow

Data Cleaning Methods

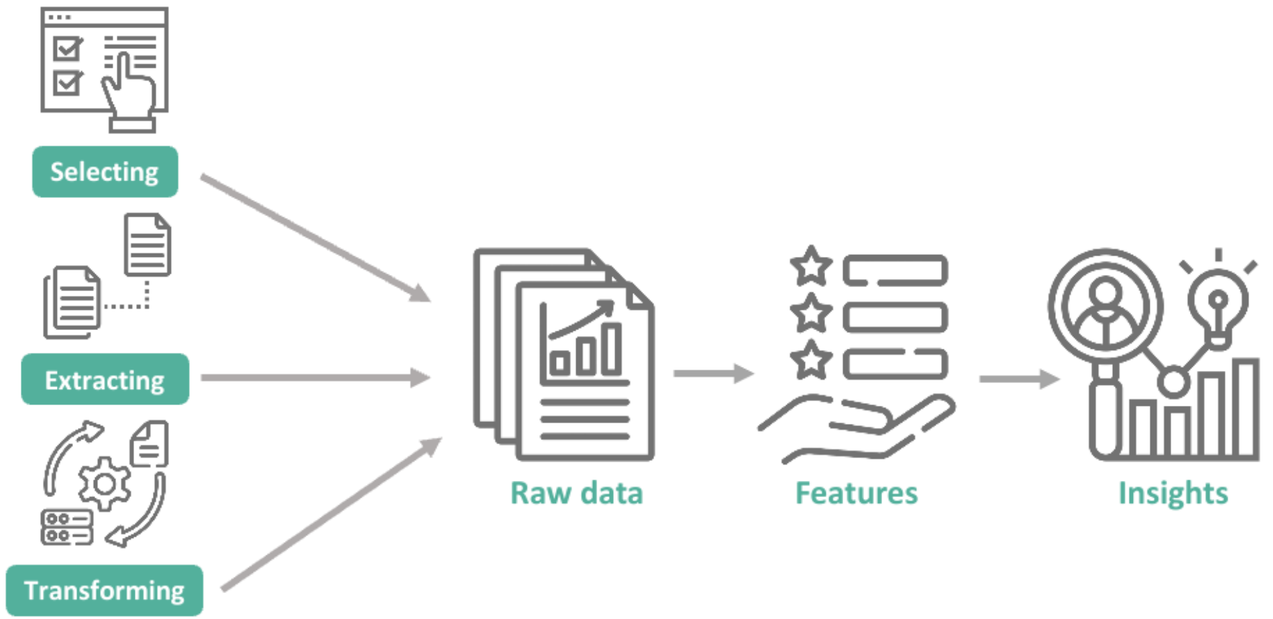


The data should be passed through one of the various data-cleaning procedures available. The procedures are explained below:

- Fill in the missing value
- Binning method
- Regression
- Clustering
- Ignore the tuples

Feature Engineering

Feature engineering is the pre-processing step of machine learning, which extracts features from raw data. It helps to represent an underlying problem to predictive models in a better way, which as a result, improve the accuracy of the model for unseen data. The predictive model contains predictor variables and an outcome variable, and while the feature engineering process selects the most useful predictor variables for the model.



Feature Engineering Techniques

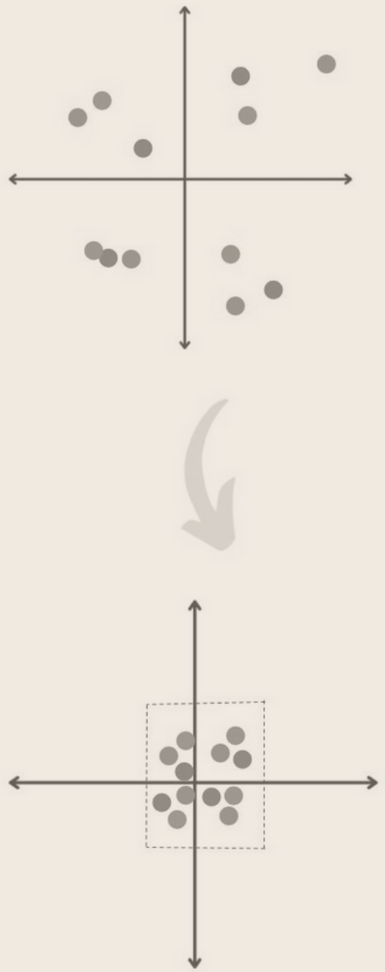
The following techniques used in feature engineering are as follows –

- **Feature Encoding:** This step involves encoding categorical data into a format that can be used by the machine learning algorithm.
- **Feature Scaling:** This step involves scaling the features so that they are on the same scale.
- **One-Hot Encoding:** This is a technique used to convert categorical variables into numerical values by creating a binary column for each category.
- **Discretization:** Discretization is a technique used to convert continuous variables into discrete values to simplify the model.
- **Binning:** Binning is a technique used to group continuous variables into bins based on specific intervals.
- **Imputation:** Imputation is a technique used to fill in missing values in a dataset. Various imputation techniques are available like mean imputation, median imputation, and mode imputation.

Feature Scaling

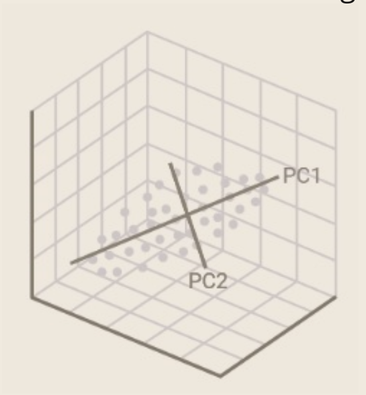
Feature scaling is a preprocessing step in ML where numerical features are standardized to a common scale. This ensures that all features contribute equally to model training. It enhances model performance by facilitating more accurate comparisons and efficient convergence during training. We used StandardScaler where,

- mean (μ) = 0
- standard deviation = 1



Dimensionality

In addressing the complexity of consumer behavior analysis, dimensionality reduction becomes pivotal. Principal Component Analysis (PCA) was employed to streamline the dataset, retaining essential information while reducing the number of features. By distilling the data to its core components, businesses gain a more concise yet insightful perspective, facilitating accurate Consumer Trend Analysis and categorical segmentation.



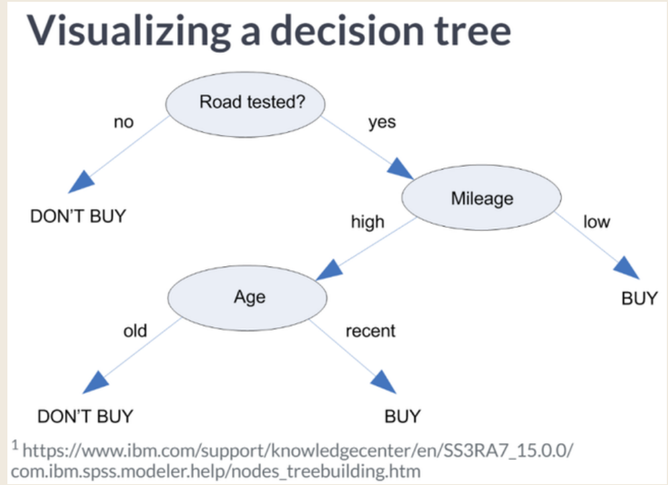
Machine Learning Models

REGRESSION

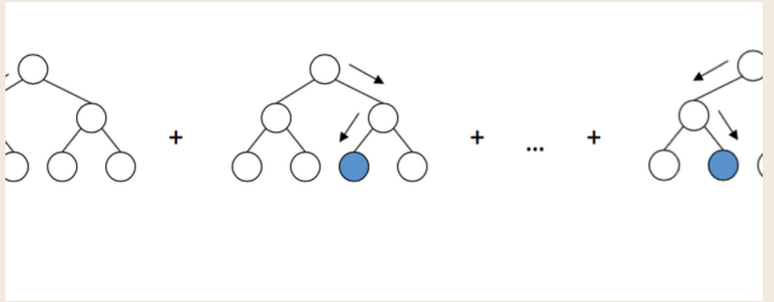
MODELS

XG Boost

XGBoost classifier is a Machine learning algorithm that is applied for structured and tabular data. XGBoost is an implementation of gradient boosted decision trees designed for speed and performance. XGBoost is an extreme gradient boost algorithm. And that means it's a big Machine learning algorithm with lots of parts. XGBoost works with large, complicated datasets. XGBoost is an ensemble modelling technique.



[1] https://www.ibm.com/support/knowledgecenter/en/SS3RA7_15.0.0/com.ibm.spss.modeler.help/nodes_treebuilding.htm



Cache awareness and out-of-core computing

Regularization for avoiding overfitting

Tree pruning using depth-first approach

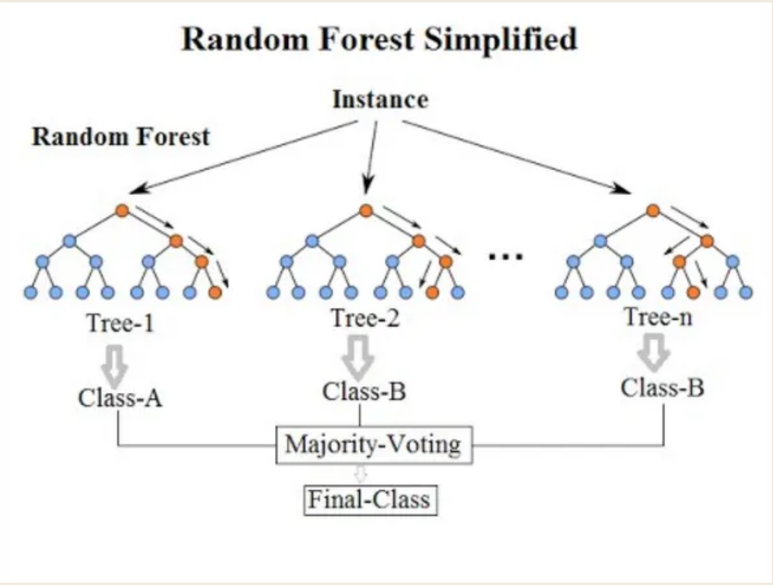
Efficient handling of missing data

Parallelized tree building

In-built cross-validation capability

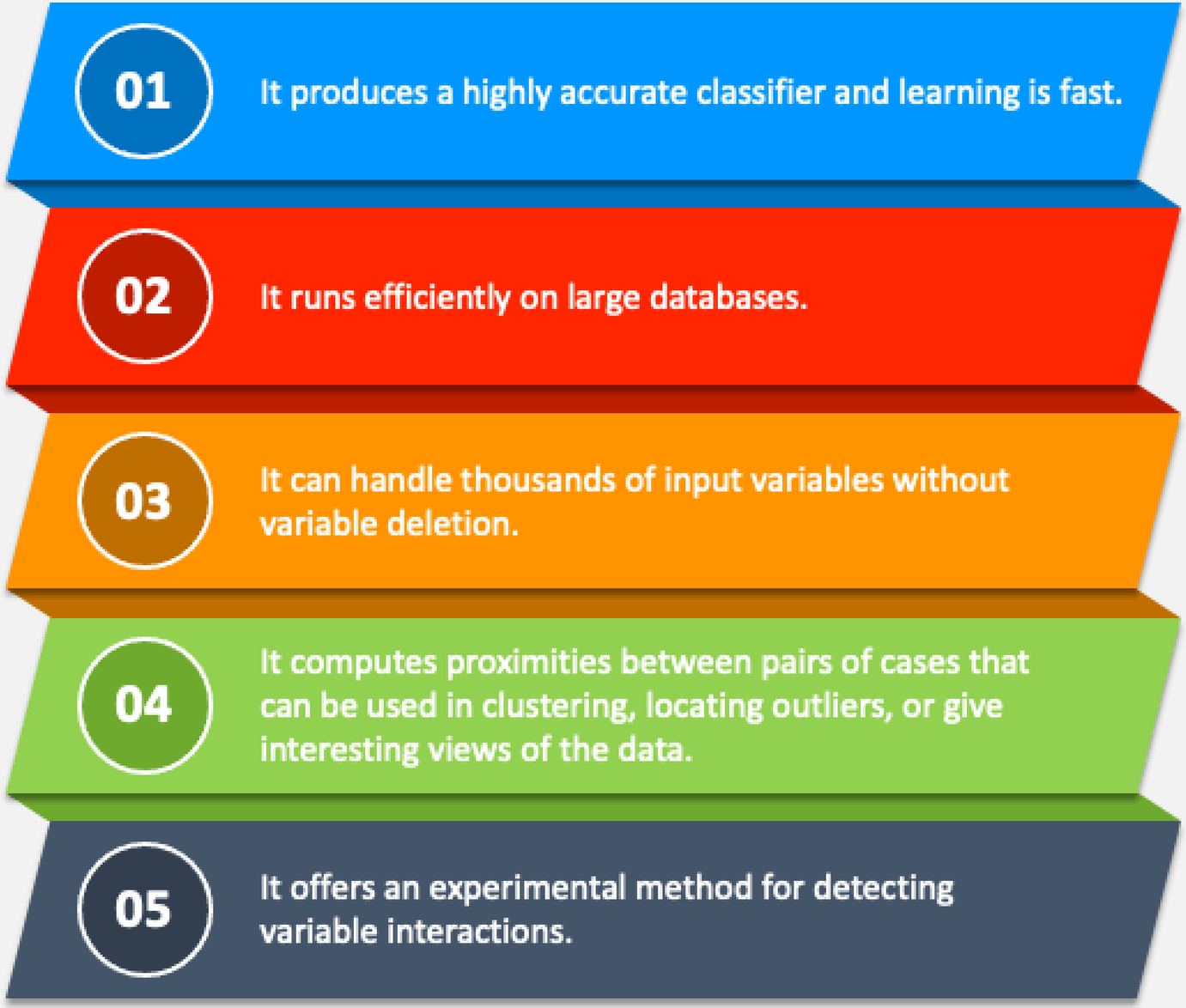
RANDOM FOREST MODEL

Random Forest Algorithm widespread popularity stems from its user-friendly nature and adaptability, enabling it to tackle both classification and regression problems effectively. The algorithm's strength lies in its ability to handle complex datasets and mitigate overfitting, making it a valuable tool for various predictive tasks in machine learning.



RANDOM FOREST

Random Forest Advantages

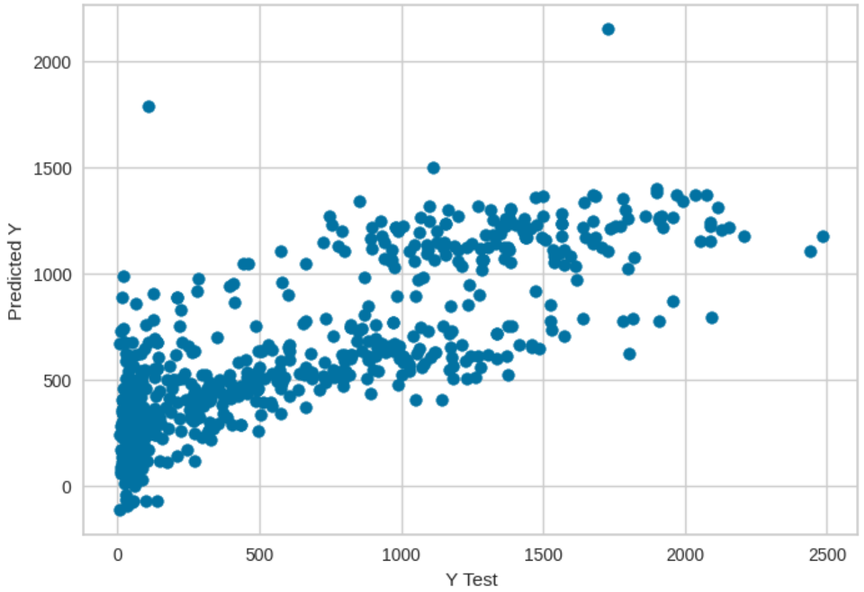


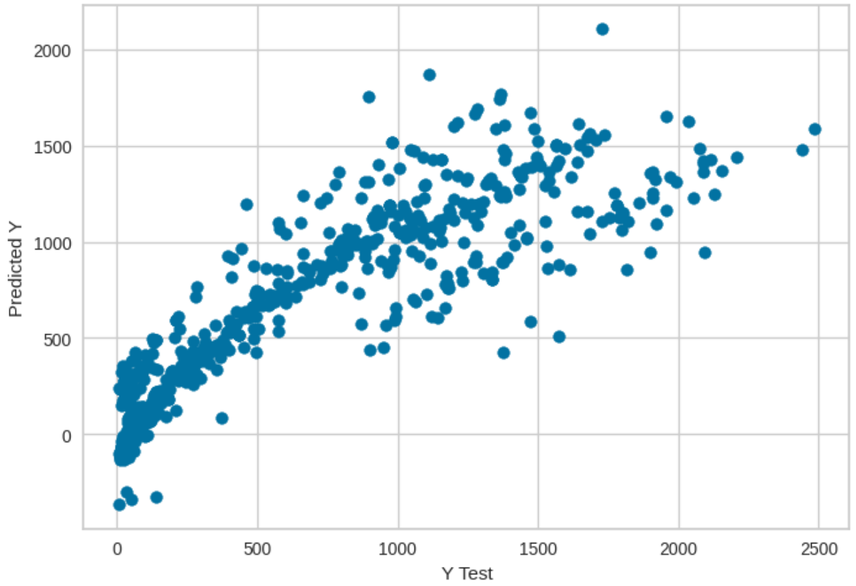
Multiple Linear Regression

Multiple linear regression refers to a statistical technique that is used to predict the outcome of a variable based on the value of two or more variables.
It extends the simple linear regression model, which involves only one independent variable, to accommodate the complexity of real-world scenarios where multiple factors may influence the outcome.

Dependent Variable (Y):
The variable that is being predicted or explained. It is the outcome or response variable.

Independent Variables (X1, X2, ..., Xn):
These are the variables that are believed to have an impact on the dependent variable. In multiple linear regression, there are two or more independent variables.



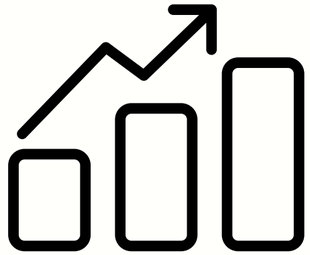


Multiple Linear Regression

Reasons for choosing this model:

Multifaceted Influences: The consumer behavior landscape is complex, with multiple variables influencing trends. Multiple linear regression allows us to analyze the impact of various factors simultaneously.

Quantitative Insights: This model provides quantitative insights into the relationship between consumer behaviors and business outcomes, aiding in data-driven decision-making.



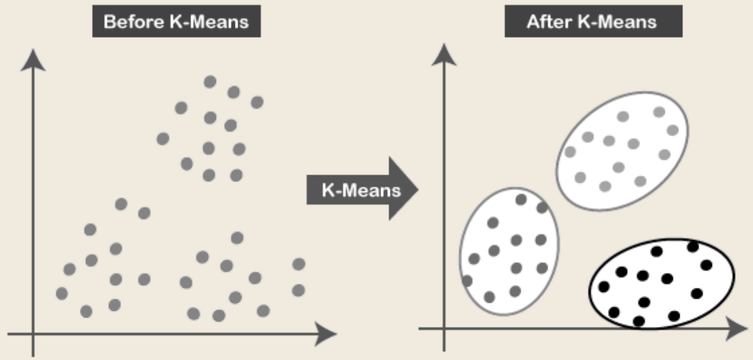
Customization: By incorporating multiple linear regression, businesses can customize product strategies based on a nuanced understanding of the diverse factors shaping consumer preferences.

CLUSTERING

MODELS

k-means Clustering

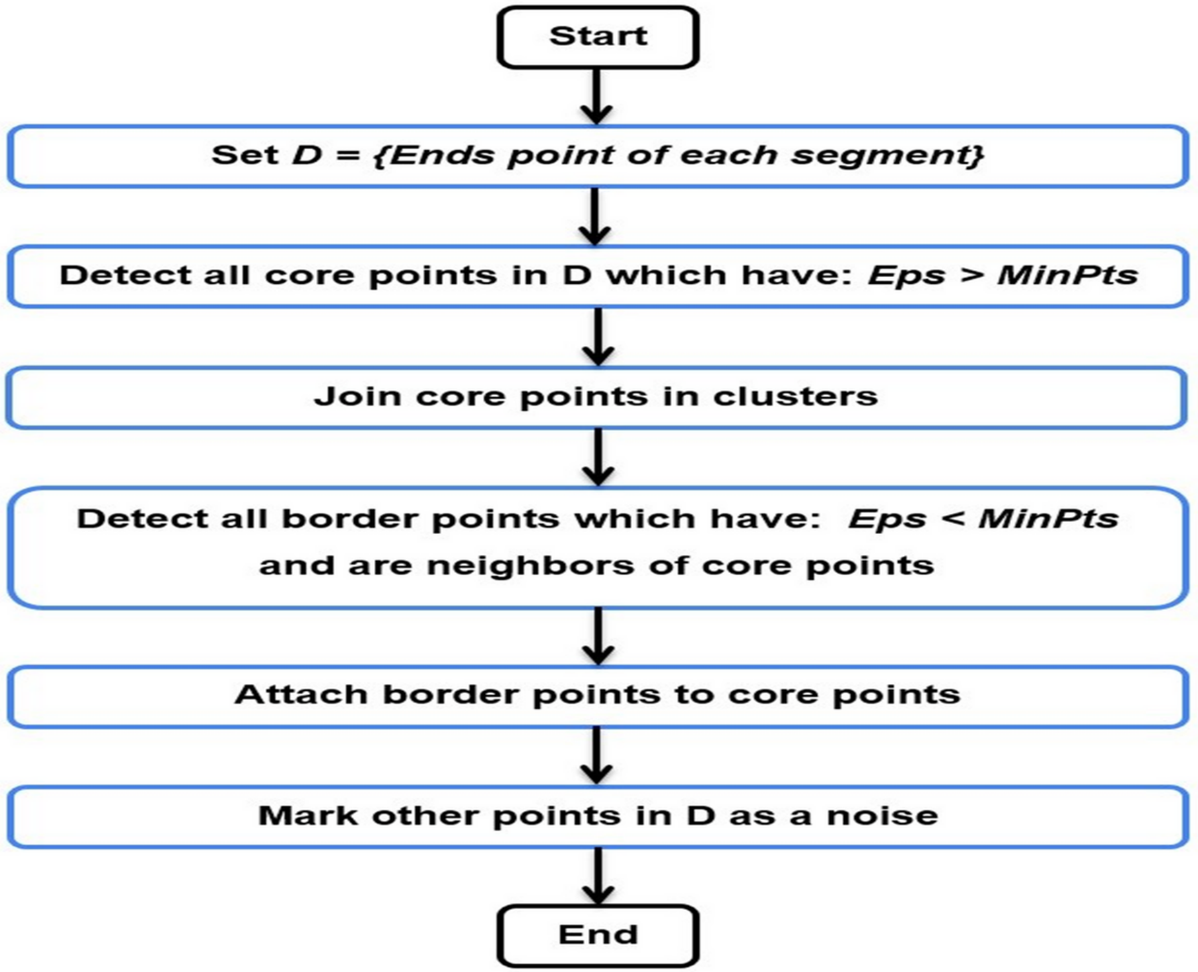
Leveraging the K-means clustering technique involves iteratively assigning customers to clusters and updating cluster centroids. By harnessing K-means clustering, businesses can unlock actionable insights, allowing for precise product modifications and targeted marketing strategies. This approach streamlines resource allocation, promoting efficiency in addressing the evolving dynamics of consumer preferences.



DB Scan

DBSCAN (Density-Based Spatial Clustering of Applications with Noise) proves invaluable in addressing this challenge. This technique efficiently captures irregularly shaped clusters and handles noise, providing a more nuanced segmentation of customers. By employing DBSCAN, businesses gain a robust method for Consumer Trend Analysis, empowering them to discern subtle variations within customer behavior.

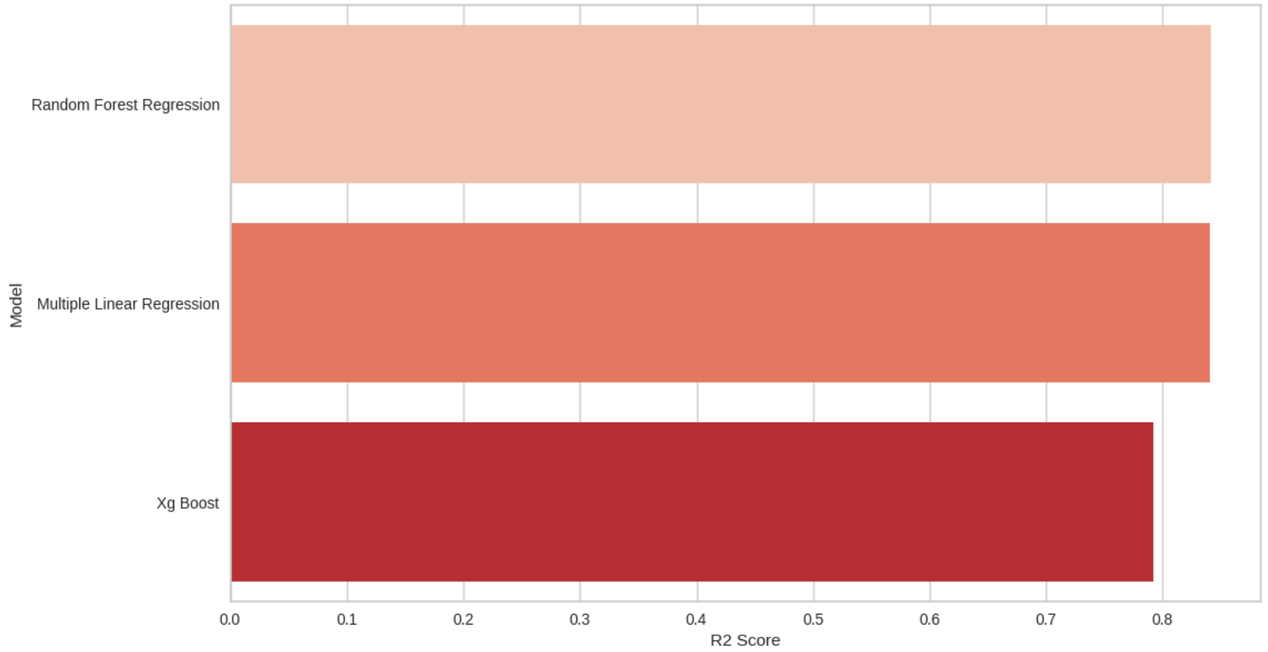
DB Scan Working Methodology



COMPARATIVE ANALYSIS

Regression Models

- Xg Boost
- Random Forest Regression
- Multiple Linear Regression

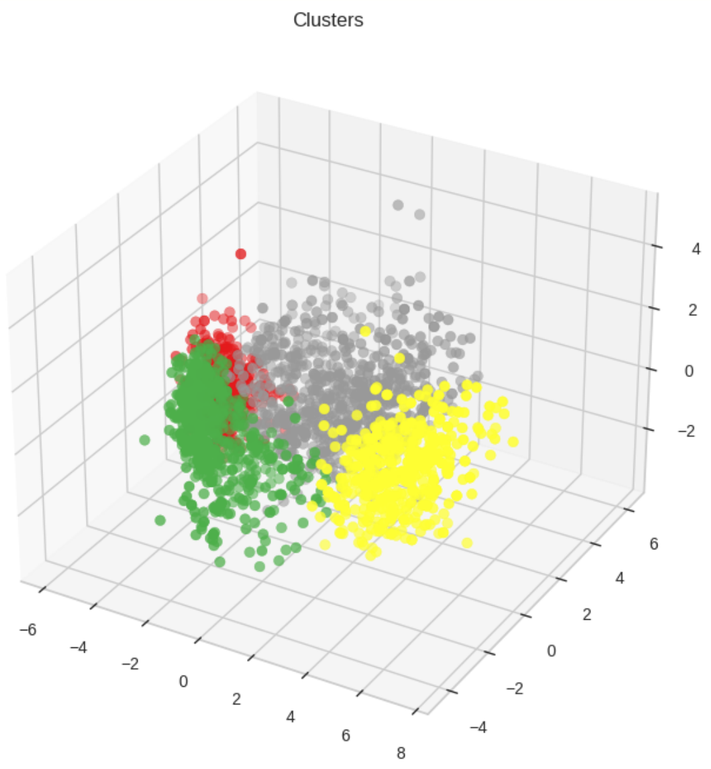


	Model	MAE	MSE	RSME	R2_Score(test)
0	Xg Boost	187.976783	74895.498972	273.670420	0.791904
1	Random Forest Regression	134.893479	57014.280687	238.776633	0.841586
2	Multiple Linear Regression	131.644077	57466.092686	239.720864	0.840331

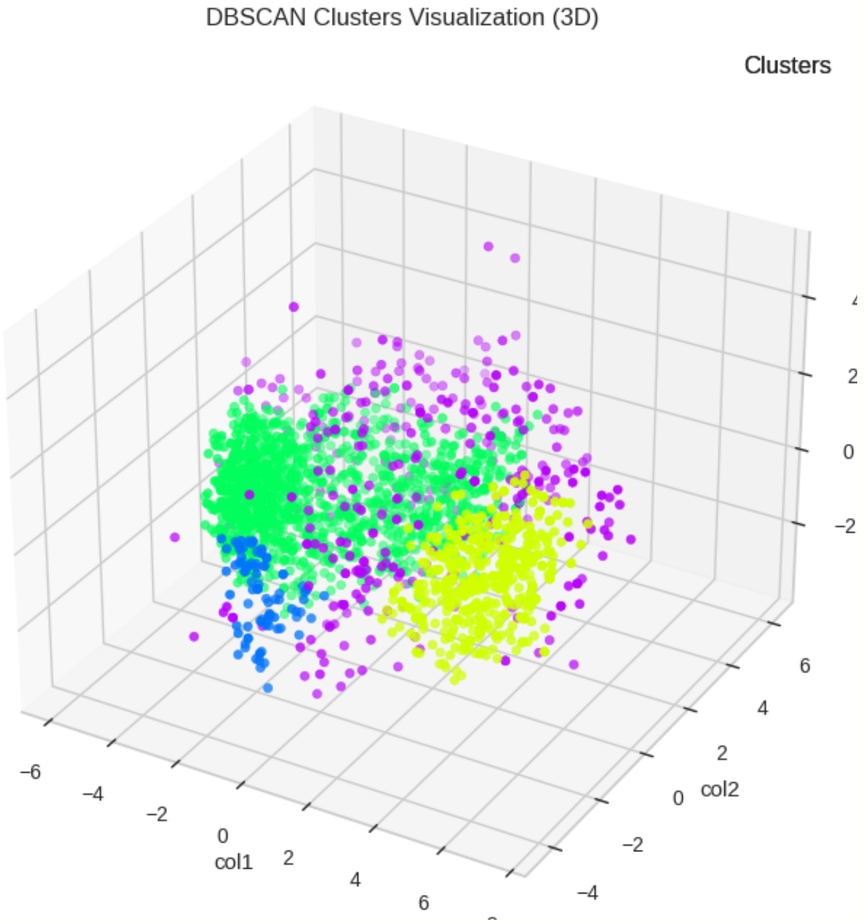
COMPARATIVE ANALYSIS

Clustering Algorithms

- K-means
- DB Scan



K-means



DB Scan

THANK YOU