Ottawa bike theft analysis

Locating Hotspots and Trends

DONE BY:

SRISHTI (041097754)

SAKSHIT(041098605)

Agenda

INTRODUCTION

BUSINESS UNDERSTANDING

DATA UNDERSTANDING

DATA PREPRATION

MODELING

EVALUATION

DEPLOYMENT

INTRODUCTION

Bike theft remains a prevalent concern in Ottawa's bustling urban environment, posing challenges to residents and visitors alike. To tackle this issue effectively, our project leverages the power of data science. At the heart of our analysis lies a comprehensive dataset sourced from Ottawa's official records. This dataset encapsulates information, including the spatial and temporal dimensions of bike theft incidents, as well as attributes detailing the characteristics of stolen bicycles and the locations where thefts occur. By delving into this rich dataset, we aim to unravel the intricate patterns and trends underlying bike theft in Ottawa. Through our analysis, we seek to provide stakeholders with invaluable insights that can inform targeted interventions to enhance bike security across the city.

BUSINESS UNDERSTANDING

Project Question: Our primary focus revolves around identifying the locations in Ottawa with the highest frequency of bike thefts and understanding the patterns associated with the types of bikes stolen in these areas. This central question guides our exploration and analysis throughout the project.

BUSINESS UNDERSTANDING

Business Goals

Identify High-Risk Areas: Our aim is to pinpoint geographical locations in Ottawa where bike thefts occur most frequently, enabling stakeholders to prioritize resources and interventions effectively.

Understand Bike Theft Patterns: We seek to uncover patterns and trends regarding the types of bikes targeted in high-risk areas. This understanding will inform targeted strategies for bike theft prevention and security enhancement.

BUSINESS UNDERSTANDING

Project Plan: We adhere to the CRISP-DM methodology, focusing particularly on the initial phases: Business Understanding, Data Understanding, and Data Preparation. This structured approach ensures that our analysis is systematic and aligned with the project objectives.

DATA UNDERSTANDING

Source of Data: Our dataset originates from Ottawa's official website, ensuring reliability and authenticity in our analysis. This data source provides us with access to a comprehensive repository of bike theft incidents recorded within the city.

Data Collection Process: Obtaining the dataset, and navigating through Ottawa's official website to locate and access the relevant data. We ensured adherence to data usage policies and obtained the dataset in a format compatible with Rapid Miner and Excel.

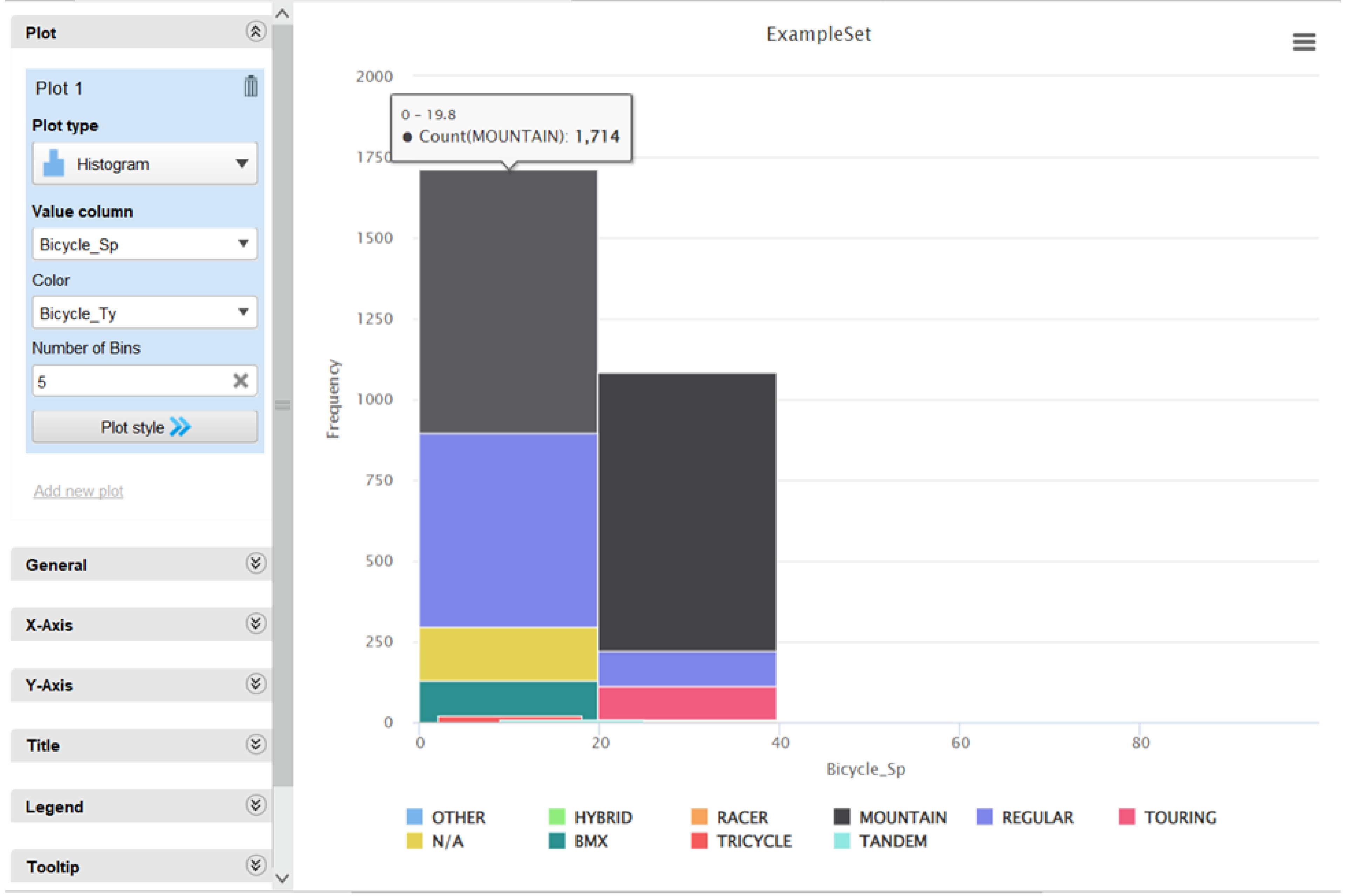
Description of Attributes: The dataset comprises various attributes that capture essential information about bike theft incidents. These attributes include:

Name: The name or label of each attribute.

Description: A brief description of the attribute's significance and relevance to our analysis.

Data Type: The data type associated with each attribute, indicating whether it is numerical, categorical, spatial, or temporal in nature.

DATA UNDERSTANDING



Plot type: Histogram

Value column: Bicycle_Sp

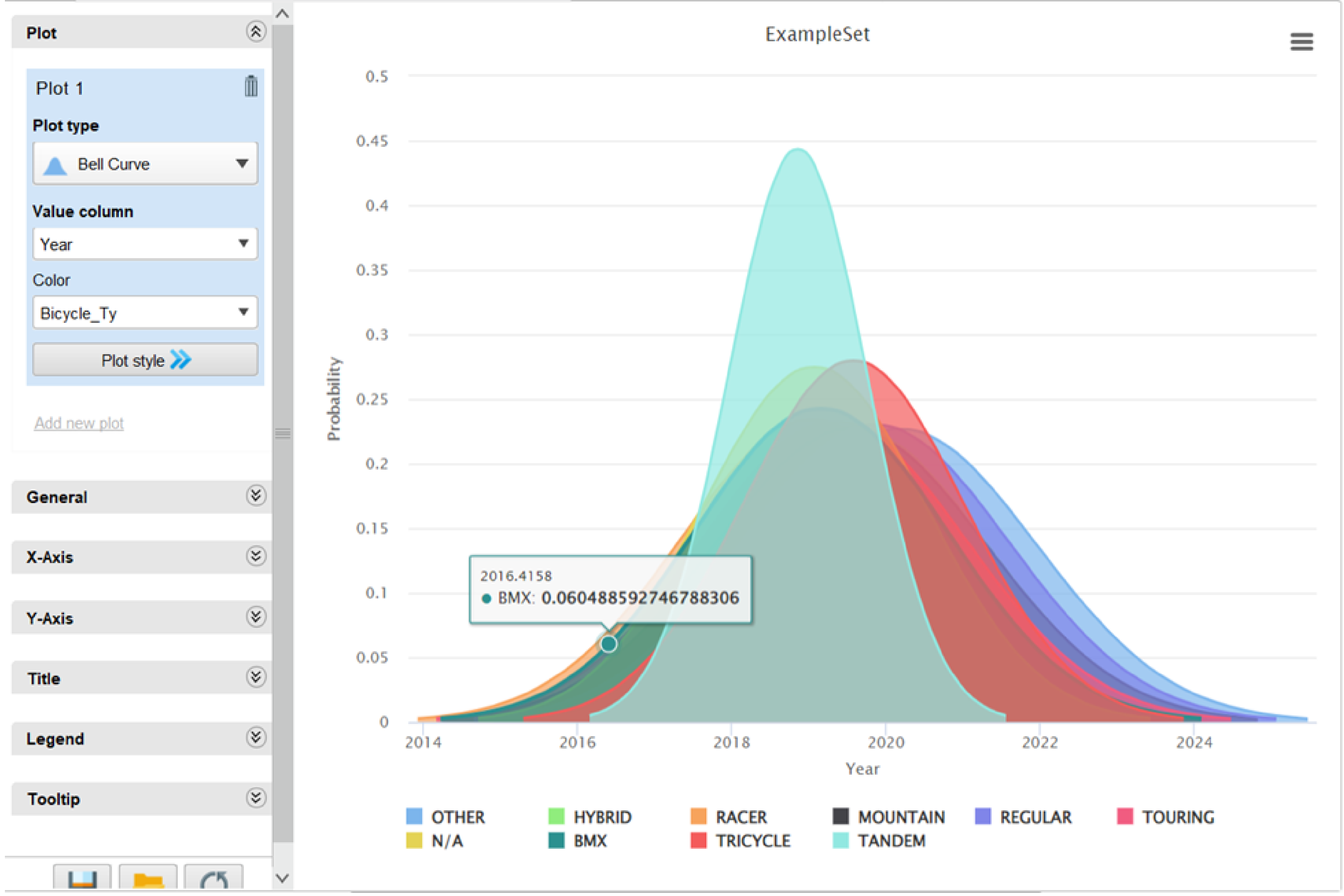
Color: Bicycle_Ty

Number of Bins: 5

Plot style

Histogram Plot: The histogram visualizes the frequency distribution of different types of bicycles based on their speed. The X-axis represents the speed of the bicycle, with each point corresponding to a specific speed interval. The Y-axis indicates the frequency or count of bicycle occurrences. Each bar's height reflects how many times a particular type of bicycle appears in the dataset.

DATA UNDERSTANDING



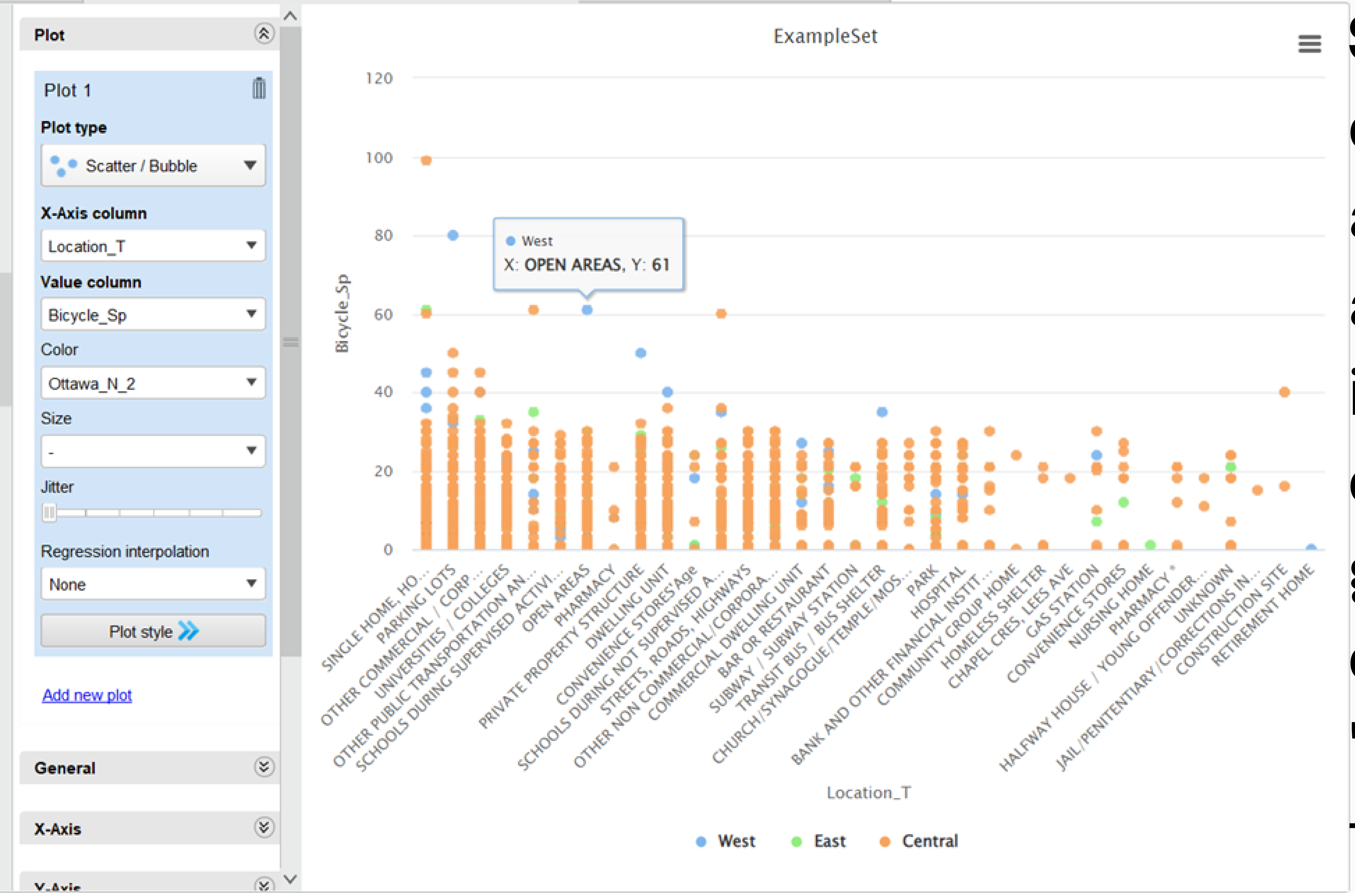
Plot type: Bell Curve

Value column: Year

Color: Bicycle_Ty

Probability Distribution Over Time: This plot illustrates the probability distribution of specific bicycle types over the years from 2014 to 2024. The X-axis denotes the years, while the Y-axis represents the probability of a particular bicycle type occurrence. The colorful bell-shaped curve showcases how the probability of different bicycle types changes over time.

DATA UNDERSTANDING



Plot

Plot 1

Plot type: Scatter / Bubble

X-Axis column: Location_T

Value column: Bicycle_Sp

Color: Ottawa_N_2

Size: -

Jitter

Regression interpolation: None

Scatter Plot: Data points are categorized by the "Location_T" attribute on the X-axis and "Bicycle_Sp" attribute on the Y-axis. Each data point is represented by a dot, with different colors indicating various categories or groups of data points. The X-axis categories include "West," "East," and "Central," while the Y-axis represents the speed of the bicycles, ranging from 0 to 120.

DATA UNDERSTANDING

Data Type Matching: We ensured the data type of each attribute matched its respective type, maintaining consistency and accuracy in our analysis.

Handling Missing Values: The dataset underwent a thorough check for missing values, and we utilized the replace missing value operator to address any gaps in the data effectively.

Duplicate Removal: To maintain data integrity, we identified and removed duplicates from the dataset using the remove duplicates operator, ensuring that each observation is unique and reliable for analysis.

DATA UNDERSTANDING

DATA PREPARATION

Selection of Relevant Attributes: We carefully selected attributes deemed crucial for our analysis, focusing on those that provide meaningful insights into bike theft patterns. These attributes were chosen based on their relevance to the project question and business goals.

Data Cleaning Process: Our data cleaning process involved handling missing values and duplicates to ensure the integrity and accuracy of our dataset. We utilized appropriate operators to address missing values, replacing them with suitable alternatives, and removed duplicates to eliminate redundancy and streamline our analysis.

DATA PREPARATION

Construction of New Attributes: To enhance our analysis and gain deeper insights, we constructed new attributes tailored to our specific objectives. These new attributes were designed to capture additional dimensions of bike theft incidents, facilitating a more comprehensive understanding of the underlying patterns. Through this process, we aimed to enrich our dataset and enable more robust analysis for informed decision-making.

DATA PREPARATION

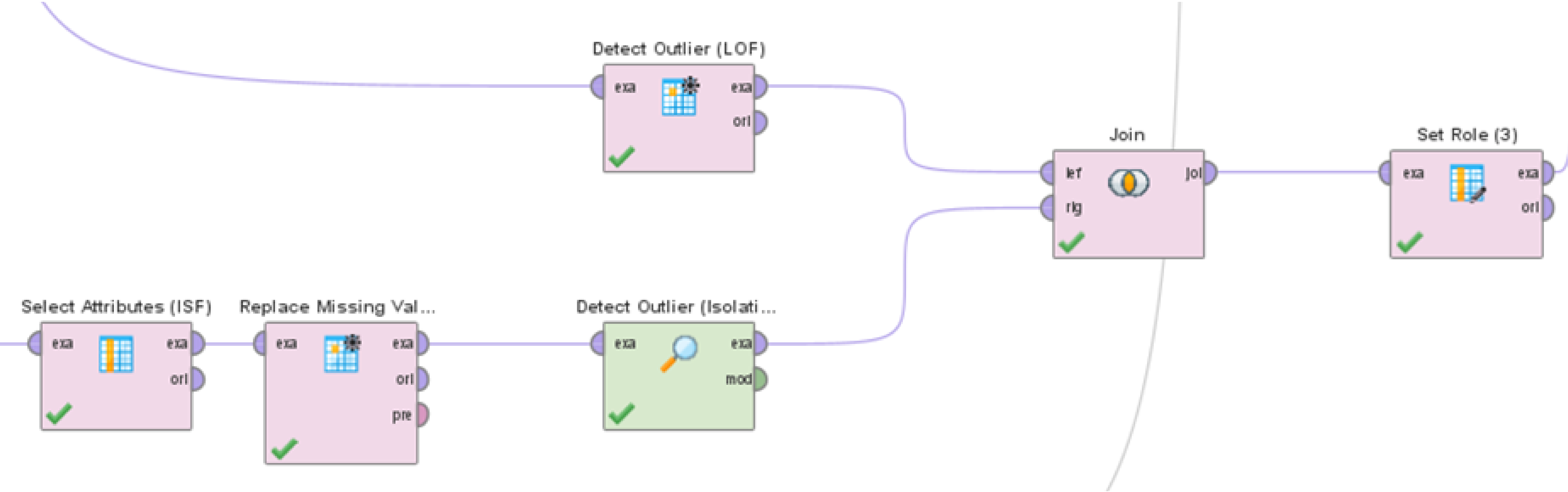
FORMAT DATA:

We used an operator called "nominal to numerical" operator,in which the format of the data has been transformed from a categorical or nominal format to a numerical format. This means that the previously categorical variables, have been encoded into numerical values.

We renamed the attributes so they are better understandable according to our question based on bike thefts and is focused on organizing and clarifying the structure of your data, rather than integrating new data sources. Also changed the data types of some of the attributes like the X and Y coordinates from real to polynominal.

MODELLING

Outlier Detection



MODELLING

Outlier Detection

Techniques Used: Local Outlier Factor (LOF) and Isolation Forest (ISF).

LOF: Density-based algorithm measuring local deviation, identifying less dense outliers.

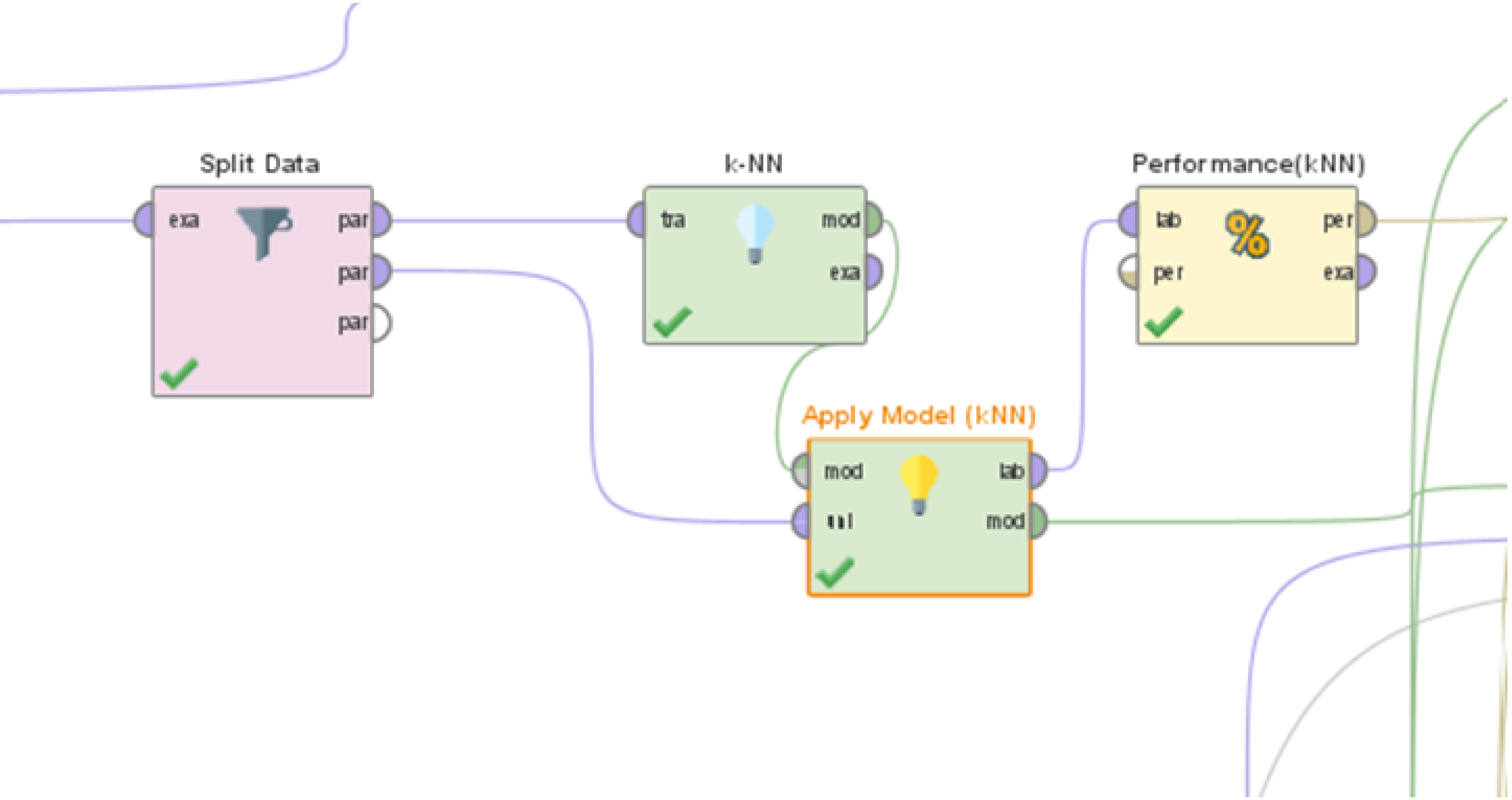
ISF: Algorithm isolating outliers by partitioning data, effective for high-dimensional data

Implementation: LOF applied with "Detect Outlier (LOF)" operator, results combined with ISF using "Join" operator
ISF applied after excluding certain attributes, missing value replacement, and outer join

Purpose: Identifying anomalies in bike theft data to understand unusual patterns or occurrences

MODELLING

Classification



MODELLING

Classification

Algorithms Used: k-Nearest Neighbors (kNN), Decision Trees (DT), Random Forests (RF)

kNN Modeling:

Split data into training and testing sets

Build a model with varying k values, select optimal k=25

Test model performance and evaluate accuracy using the "Performance" operator

DT Modeling:

Select relevant features, split data, and build decision tree model

Predict on test data, evaluate performance metrics, visualize tree structure

RF Modeling:

Split data, train RF model, set parameters

Predict on test data, evaluate performance, visualize feature importance

Description of Results

Confusion Matrix Analysis:

Provides insights into the strength, reliability, and predictive power of association rules.

Each metric contributes to understanding the performance of rules in capturing patterns and relationships.

Description of Results

Performance Vector in kNN

kNN Algorithm Performance:

Precision and recall metrics analyzed for different classes.

Precision highest for "SINGLE HOME, HOUSE" class but lower for other categories.

Variability in performance across classes suggests influence of data distribution.

Further analysis and fine-tuning may be required for improved performance.

Description of Results

Performance Vector in Decision Trees

Decision Tree Model Performance:

Achieved high precision for some classes but lower recall.

Variability in precision and recall across classes.

Some classes show struggles in accurate prediction.

Further analysis needed to enhance model performance, especially for challenging classes.

Evaluation

Back to Agenda Page

Cross-Validation Process: Implementation of k-fold cross-validation for assessing model performance.

Dataset divided into 10 folds, model trained and tested iteratively.

Performance metrics averaged across folds for robust estimation.

Cross-Validation Results: Robust estimation of model performance provided by evaluating on multiple subsets.

Cross-validation performed for kNN, Decision Trees, and Random Forest models.

Evaluation

Back to Agenda Page

Evaluation of Association

Association Analysis Process:

Association algorithm applied to discover frequent item sets and rules.

Patterns and correlations between attributes of bike theft incidents identified

Association Analysis Results:

Interesting patterns and correlations revealed between different attributes.

High-confidence association rules provide actionable insights.

Conclusion

Objective and Approach:

Aimed to address bike theft in Ottawa using data science techniques.
Leveraged CRISP-DM methodology for comprehensive analysis.

Impact and Recommendations:

Provided valuable insights and recommendations for improving bike security.
Collaborative efforts aimed at reducing theft incidents in Ottawa.

Back to Agenda Page

Future Directions:

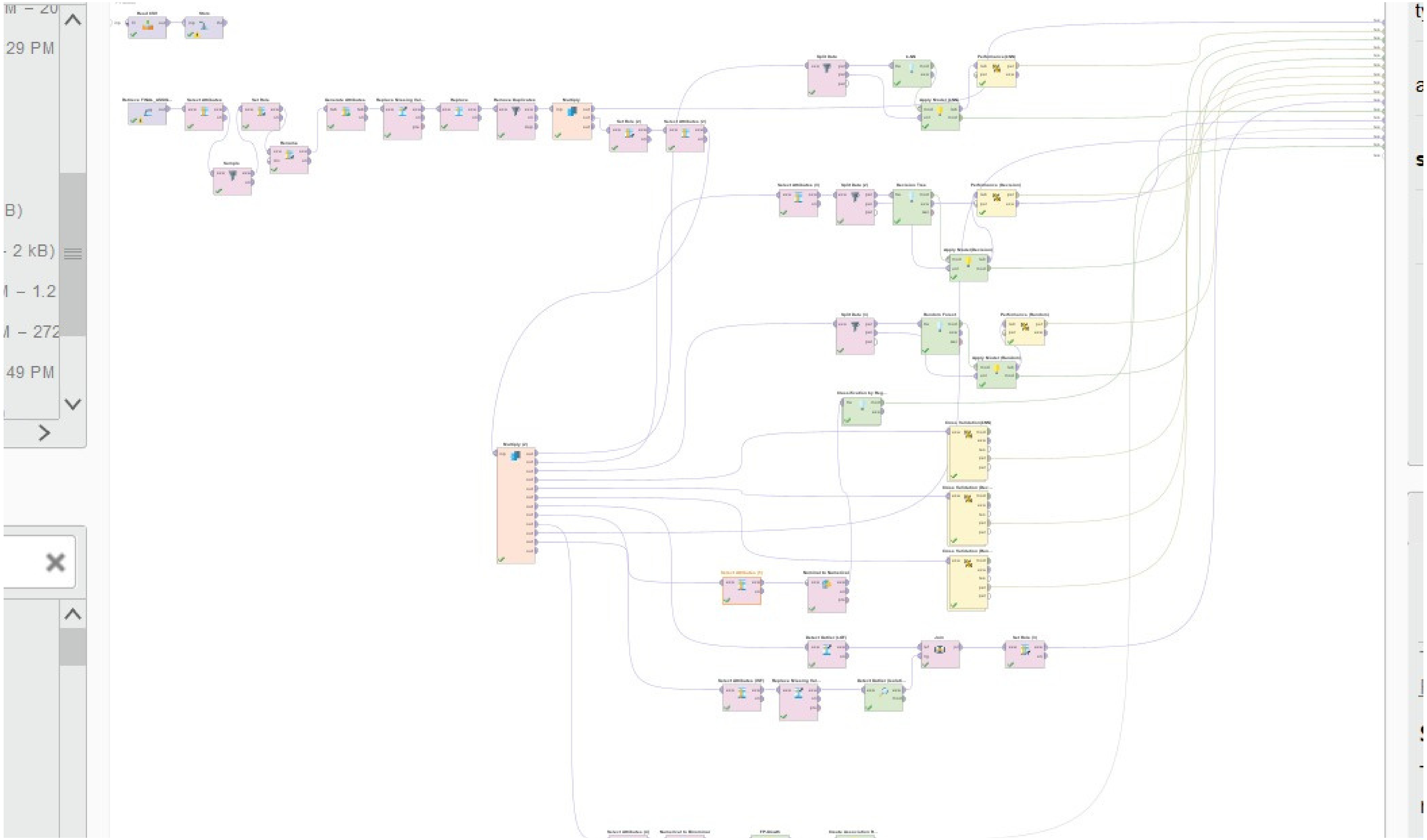
Aimed Further refinement and optimization of models for enhanced performance.

Continued monitoring and analysis of bike theft patterns for proactive measures.

Continued collaboration with stakeholders for effective implementation of recommendations.

Conclusion

Back to Agenda Page


M – 20
29 PM
B)
- 2 kB)
M – 1.2
M – 272
49 PM
Read CSV
Store
Retrieve TMdb_6000...
Select Attributes
Set Role
Sample
Rename
Generate Attributes
Replace Missing Val...
Replace
Remove Duplicates
Multiply
Set Role (2)
Select Attributes (2)
Split Data
k-NN
Performance (kNN)
Apply Model (kNN)
Select Attributes (3)
Split Data (2)
Decision Tree
Performance (Decision)
Apply Model(Decision)
Split Data (3)
Random Forest
Performance (Random)
Apply Model (Random)
Classification by Reg...
Cross Validation(kNN)
Cross Validation (Dec...
Cross Validation (Ran...
Multiply (2)
Select Attributes (5)
Nominal to Numerical
Join
Set Role (3)
Replace Missing Val...
FP-Growth
Create Association R...