

**DEPARTMENT OF COMPUTER SCIENCE &
ENGINEERING (AIML, CYS, IoT)**



(REGULATION 2022)

LABMANUAL

22CS701-DATA AND INFORMATION VISUALIZATION

Laboratory In-charges

Dr.SUMA SIRA JACOB

Associate Professor &
Programme Coordinator – CSE(AIML,CYS,IoT)

Mr SHANMUGA RAJU S

Assistant Professor- CSE (IoT)

PREFACE

The Data and Information Visualization laboratory is designed for final-year B.E. students of Computer Science and Engineering (CYS, IoT). This lab course focuses on practical applications of data preprocessing, analysis, mining, and visualization using Orange software, a visual programming tool that simplifies data science workflows. In today's data-driven world, understanding how to process and visualize data is critical. This lab provides students with hands-on experience in the entire data analysis pipeline—from cleaning and transforming raw datasets to mining patterns and visualizing insights. The experiments are structured to reinforce theoretical concepts through real-world application.

The manual comprises ten well-organized experiments that guide students through key processes such as data preprocessing, data cube construction, noise and missing value handling, exploratory clustering techniques (like k-means and MST), association rule mining, hypothesis generation, and dimensionality reduction techniques like Haar Wavelet and PCA. It also includes advanced topics like scalable algorithms and cluster assessment with visualization.

Through these experiments, students gain exposure to building data cubes, generating strong association rules using support and confidence measures, and visualizing results through histograms and regression plots. Emphasis is placed on developing both technical competence and analytical thinking. By completing this lab, students will be equipped with practical skills that are essential in fields like data science, business intelligence, and AI model development.

**FACULTY OF COMPUTER SCIENCE & ENGINEERING
(AIML, CYS, IoT)**

SRI KRISHNA COLLEGE OF TECHNOLOGY

COIMBATORE-641042

Prepared by
Shanmuga Raju S
Assistant Professor
/ CSE(IoT)

Verified by
Dr. Suma Sira Jacob
Associate Professor
PC/ CSE (AIML, CYS, IoT)

Approved by
Dr. M.G.Sumithra
Principal

PROFILE OF THE INSTITUTION

Nestled at the foothills of the Western Ghats, located in a sprawling 52-acre campus in Kovaipudur, Coimbatore, Sri Krishna College of Technology (SKCT) is a vibrant institute of higher education established in 1985 promoted by Sri Krishna Institutions. An extraordinary freedom of opportunity—to explore, to collaborate, and to challenge oneself—is the hallmark of the Institute. Being an autonomous institute, affiliated with Anna University, Chennai, and approved by AICTE, New Delhi, SKCT lays strong emphasis on collaborative research and stands apart from other institutes by its participatory work culture, student care programmes, and high industry interaction.

In a span of 38 years, it has emerged as one of the premier engineering colleges for learning, discovery, and innovation due to the dynamic leadership of the Chairperson and Managing Trustee Smt. S. Malarvizhi. Being an acclaimed educationalist, she continues to contribute profusely for the glory and happiness of advancing generations. The college is accredited with A Grade by NAAC, and eligible undergraduate programs are accredited by the National Board of Accreditation (NBA), New Delhi. The college offers 11 undergraduate programmes, 6 postgraduate programmes, and 5 doctoral programmes in Engineering, Technology, and Management Studies.

VISION:

Sri Krishna College of Technology aspires to be recognized as one of the pioneers in imparting world class technical education through technology enabled innovative teaching learning processes with a focus on research activities to cater, to the societal needs.

MISSION:

- To be recognized as centre of excellence in science, engineering and technology through effective teaching and learning processes by providing a conducive learning environment.
- To foster research and development with creative and entrepreneurial skills by means of innovative applications of technology.
- Accomplish expectations of the society and industry by nurturing the students to be competent professionals with integrity.

COURSES OFFERED

UNDERGRADUATE PROGRAMMES (Four Years B.E/B.Tech)

- B.E-Civil Engineering
- B.E-Computer Science and Engineering
- B.E-Computer Science and Engineering (Cyber Security)
- B.E-Computer Science and Engineering(Internet of Things)
- B.E - Computer Science and Engineering (Artificial Intelligence and Machine Learning)
- B.E-Electronics and Communication Engineering
- B.E-Electrical and Electronics Engineering
- B.E-Instrumentation and Control Engineering
- B.E-Mechanical Engineering
- B.Tech - Artificial Intelligence and DataScience
- B.Tech – Information Technology

POST GRADUATE PROGRAMMES (Two Years)

- Master of Business Administration
- M.E-Applied Electronics
- M.E–Computer Science Engineering
- M.E–Engineering Design
- M.E-PowerSystem Engineering
- M.E-Structural Engineering

DOCTORAL PROGRAMMES (Ph.D.)

- Civil Engineering
- Computer Science and Engineering
- Electronics and Communication Engineering
- Electrical and Electronics Engineering
- Mechanical Engineering

DEPARTMENT OF COMPUTER SCIENCE & ENGINEERING (AIML, CYS, IoT)

The Department of Computer Science and Engineering (Artificial Intelligence & Machine Learning, Cyber Security, Internet of Things) at the esteemed institution, SKCT, which was established in 2022. At our institution, we are committed to providing a 4-year Bachelor of Engineering (B.E.) degree with a specific focus on Internet of Things. The eminent team of faculty is dedicated in delivering a high-quality education to equip students with the necessary skills to navigate the dynamic and ever-changing domains of AI & ML, Cyber Security, and IoT. The program has been strategically developed to cultivate creativity, critical thinking, and problem-solving aptitudes, equipping our graduates with the necessary capabilities to contribute sustainable solutions to industrial and society problems.

VISION:

The department of CSE fosters a conducive ambience to meet the global standards by equipping the students with modern techniques in the area of Computer Science and relevant research to address the societal needs.

MISSION:

- To provide positive working environment that would help the students perform to their highest abilities in various fields of computer science.
- To enable students and faculty with the best of technologies and knowledge emerging in the domain of Computer Science and Engineering.
- To establish nationally and internationally recognized research centers and expose the students to broad research experience.

PROGRAM EDUCATIONAL OBJECTIVES:

PEO1: Apply the acquired engineering knowledge to solve economic, social, ethical, and environmental issues related to Internet of Things.

PEO2: Adapt the emerging Information and Communication Technologies to innovate and to cater the Industrial and Societal needs.

PEO3: Contribute Internet of Things expertise to research and development and create novel products that benefit society.

PROGRAMOUTCOMES:

PO 1: Engineering knowledge: Apply the knowledge of mathematics, science, engineering fundamentals, and an engineering specialization to the solution of complex engineering problems.

PO 2: Problem analysis: Identify, formulate, review research literature, and analyze complex engineering problems reaching substantiated conclusions using first principles of mathematics, natural sciences, and engineering sciences.

PO 3: Design/development of solutions: Design solutions for complex engineering problems and design system components or processes that meet the specified needs with appropriate consideration for the public health and safety, and the cultural, societal, and environmental considerations.

PO 4: Conduct investigations of complex problems: Use research-based knowledge and research methods including design of experiments, analysis and interpretation of data, and synthesis of the Information to provide valid conclusions.

PO 5: Modern tool usage: Create, select, and apply appropriate techniques, resources, and modern engineering and IT tools including prediction and modeling to complex engineering activities with an understanding of the limitations.

PO 6: The engineer and society: Apply reasoning informed by the contextual knowledge to assess societal, health, safety, legal and cultural issues and the consequent responsibilities relevant to the professional engineering practice.

PO 7: Environment and sustainability: Understand the impact of the professional engineering solutions in societal and environmental contexts, and demonstrate the knowledge of, and need for sustainable development.

PO 8: Ethics: Apply ethical principles and commit to professional ethics and responsibilities and norms of the engineering practice.

PO 9: Individual and team work: Function effectively as an individual, and as a member or leader in diverse teams, and in multidisciplinary settings.

PO10: Communication: Communicate effectively on complex engineering activities with the engineering community and with society at large, such as, being able to comprehend and write effective reports and design documentation, make effective presentations, and give and receive clear instructions.

PO 11: Project management and finance: Demonstrate knowledge and understanding of the engineering and management principles and apply these to one's own work, as a member and leader in a team, to manage projects and in multidisciplinary environments.

PO 12: Life-long learning: Recognize the need for, and have the preparation and ability to engage in independent and life-long learning in the broadest context of technological change.

PROGRAMSPECIFICOUTCOMES

PSO1: Acquire and apply standard Software Engineering practices and strategies in IoT project development to deliver a quality product for industry success.

PSO2: Analyze connected sensors, devices and equipment for transferring data over a network.

SYLLABUS

SNo	NAME OF THE EXPERIMENT	CO Mapping	RBT
1	DATA PRE-PROCESSING AND DATA CUBE Data preprocessing methods on student and labor datasets Implement data cube for data warehouse on 3-dimensional data	CO1,CO2	AP
2	DATA CLEANING Implement various missing handling mechanisms, Implement various noisy handling mechanisms	CO1,CO2	AP
3	EXPLORATORY ANALYSIS Develop k-means and MST based clustering techniques, Develop the methodology for assessment of clusters for given dataset	CO2	AP
4	ASSOCIATION ANALYSIS Design algorithms for association rule mining algorithms	CO2	AP
5	HYPOTHESIS GENERATION Derive the hypothesis for association rules to discovery of strong association rules; Use confidence and support thresholds	CO3	AP
6	TRANSFORMATION TECHNIQUES Construct Haar wavelet transformation for numerical data, Construct principal component analysis (PCA) for 5-dimensional data	CO3	AP
7	DATA VISUALIZATION Implement binning visualizations for any real time dataset, Implement linear regression techniques	CO4	AP
8	CLUSTERS ASSESSMENT Visualize the clusters for any synthetic dataset, Implement the program for converting the clusters into histograms	CO4	AP
9	HIERARCHICAL CLUSTERING Write a program to implement agglomerative clustering technique ,Write a program to implement divisive hierarchical clustering technique	CO4	AP
10	SCALABILITY ALGORITHMS Develop scalable clustering algorithms ,Develop scalable apriori algorithm	CO4	AP
Total Hours (P):			30

TABLE OF CONTENT

Experiment No:	NAME OF THE EXPERIMENT	Page No.
1	DATA PRE-PROCESSING AND DATA CUBE	1
2	DATA CLEANING	6
3	EXPLORATORY ANALYSIS	12
4	ASSOCIATION ANALYSIS	18
5	HYPOTHESIS GENERATION	22
6	TRANSFORMATION TECHNIQUES	26
7	DATA VISUALIZATION	31
8	CLUSTERS ASSESSMENT	35
9	HIERARCHICAL CLUSTERING	42
10	SCALABILITY ALGORITHMS	46

GENERAL SAFETY INSTRUCTIONS

The following Safety Instructions should be strictly followed by the students in Automation Laboratory.

- Students should wear durable clothing that covers the arms, legs, torso and feet. (Note: sandals, shorts, tank tops etc. have no place in the lab. Students inappropriately dressed for lab, at the instructor's discretion, be denied access)
- To protect clothing from chemical damage or other dirt, wear a lab apron or lab coat. Long hair should be tied back to keep it from coming into contact with lab chemicals or flames.
- In case of injury (cut, burn, fire etc.) notify the instructor immediately.
- In case of a fire or imminently dangerous situation, notify everyone who may be affected immediately; be sure the lab instructor is also notified.
- In case of a serious cut, stop blood flow using direct pressure using a clean towel, notify the lab instructor immediately.
- Eating, drinking and smoking are prohibited in the laboratory at all times.
- Never work in the laboratory without proper supervision by an instructor.
- Never carry out unauthorized experiments. Come to the laboratory prepared. If you are unsure about what to do, please ask the instructor.

LABORATORY CLASS INSTRUCTIONS

The following Laboratory class instructions should be strictly followed by the students in the Laboratory.

- Students must attend the lab classes with ID cards and in the prescribed uniform.
- Students must follow the dress code prescribed in the institution guidelines.
- Students must check if the computers and components are in working condition before setting up the experiment.
- Any damage to any of the computer/components caused due to carelessness, the cost will be fully recovered from the individual (or) group of students.
- Students may contact the lab in charge immediately for any unexpected incidents (like system restart) and emergency.
- Make sure, while leaving the lab after the stipulated time, that all the power connections are switched off.

[illegible]



AIM:

To achieve data preprocessing on student and labor datasets and to implement a data cube for data warehousing on 3-dimensional data in Orange, follow these steps:

TOOLS AND SOFTWARE USED:

Orange Data Mining Software – For visual data preprocessing and data cube creation.

Datasets: Publicly available Student and Labor datasets (in .csv format).

System Requirements: Any desktop OS with internet access, Orange 3.x or later.

DESCRIPTION:**Data Preprocessing:**

Data preprocessing, a component of data preparation, describes any type of processing performed on raw data to prepare it for another data processing procedure.

Data Cube:

A data cube is a data structure that, contrary to tables and spreadsheets, can store data in more than 2 dimensions. They are mainly used for fast retrieval of aggregated data.

Dataset:

Datasets involve a large amount of data points grouped into one table.

Pre-Viva Questions

1. What are the key steps involved in data preprocessing?
2. Why is data cleaning important before visualization?
3. What is a data cube and how does it differ from a simple table?
4. Define discretization and its relevance in preprocessing.
5. How does imputation help with missing values?

PROBLEM DEFINITION:

In this experiment, the goal is to process raw data, remove any unwanted or irrelevant information, and organize the cleaned dataset into a structured format suitable for analysis using a data cube approach.

PROCEDURE:**Steps for Data Preprocessing**

1. **Load Data:**
 - Use the 'File' widget to load your student and labor datasets.
2. **Handle Missing Values:**
 - Use the 'Edit Domain' widget to impute missing values or remove instances with missing values.
3. **Normalize Data:**
 - Use the 'Normalize' widget to scale the data to a standard range.
4. **Discretize Data:**
 - Use the 'Discretize' widget to convert continuous variables into categorical bins if

necessary.

5. Feature Selection:

- Use the 'Select Columns' widget to choose the relevant features for analysis.

Example Workflow for Student Dataset

1. Load Data:

- Drag the 'File' widget to the canvas.
- Load your student dataset file (e.g., `students.csv`).

2. Handle Missing Values:

- Drag the 'Edit Domain' widget to the canvas.
- Connect the 'File' widget to the 'Edit Domain' widget.
- Configure the 'Edit Domain' widget to handle missing values (e.g., impute with mean/median).

3. Normalize Data:

- Drag the 'Normalize' widget to the canvas.
- Connect the 'Edit Domain' widget to the 'Normalize' widget.
- Configure the 'Normalize' widget to standardize the data.

4. Discretize Data (if necessary):

- Drag the 'Discretize' widget to the canvas.
- Connect the 'Normalize' widget to the 'Discretize' widget.
- Configure the 'Discretize' widget to bin continuous variables.

5. Feature Selection:

- Drag the 'Select Columns' widget to the canvas.
- Connect the final preprocessing widget to the 'Select Columns' widget.
- Select the relevant features for analysis.

Implementing a Data Cube for Data Warehousing

Steps to Implement a Data Cube

1. Load Data:

- Use the 'File' widget to load your dataset with at least three dimensions.

2. Pivot Table:

- Use the 'Pivot Table' widget to create and visualize the data cube.

Example Workflow for a 3-Dimensional Dataset

1. Load Data:

- Drag the 'File' widget to the canvas.
- Load your 3-dimensional dataset file (e.g., `3d_dataset.csv`).

2. Create Data Cube:

- Drag the 'Pivot Table' widget to the canvas.
- Connect the 'File' widget to the 'Pivot Table' widget.
- Configure the 'Pivot Table' widget to set rows, columns, and layer attributes to the desired dimensions.

Detailed Steps for Implementing the Data Cube

1. Load and Preprocess Data:

- Use the steps described in the data preprocessing section to load and preprocess your dataset.

2. Create Pivot Table for Data Cube:

- Drag the 'Pivot Table' widget to the canvas.

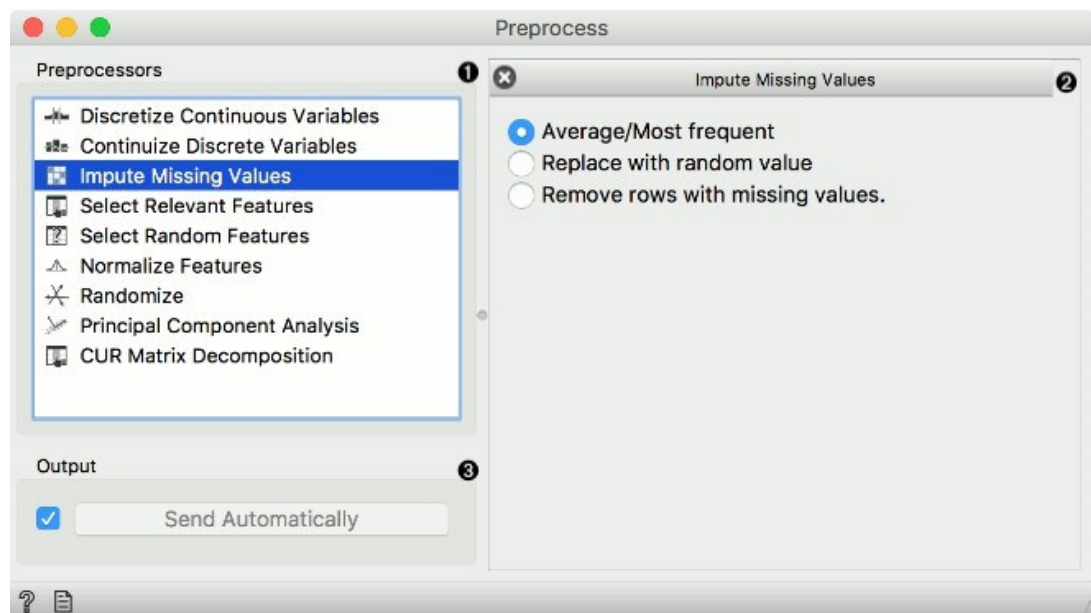
- Connect the preprocessed data widget to the 'Pivot Table' widget.
 - Configure the rows, columns, and layers to represent the three dimensions of your data.
3. **Analyze Data Cube:**
- Use the pivot table to analyze data across the selected dimensions, such as aggregating by sum, average, count, etc.

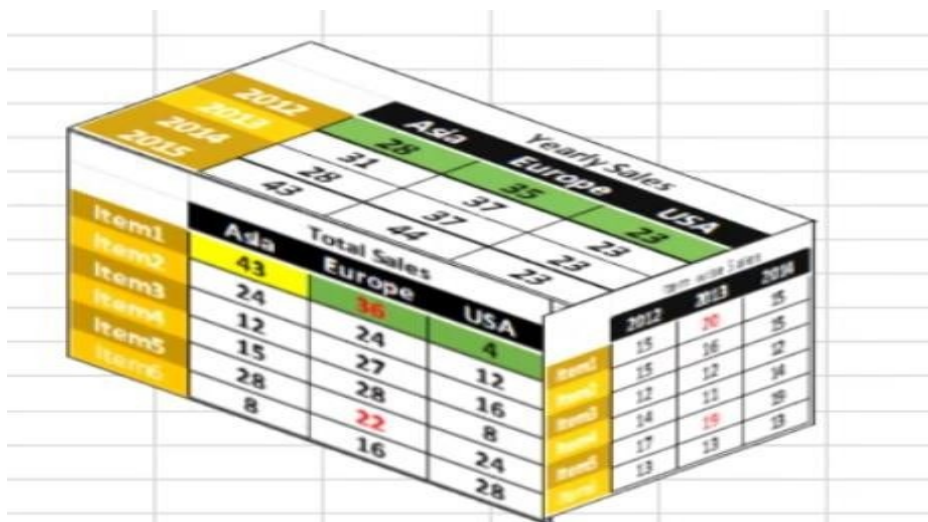
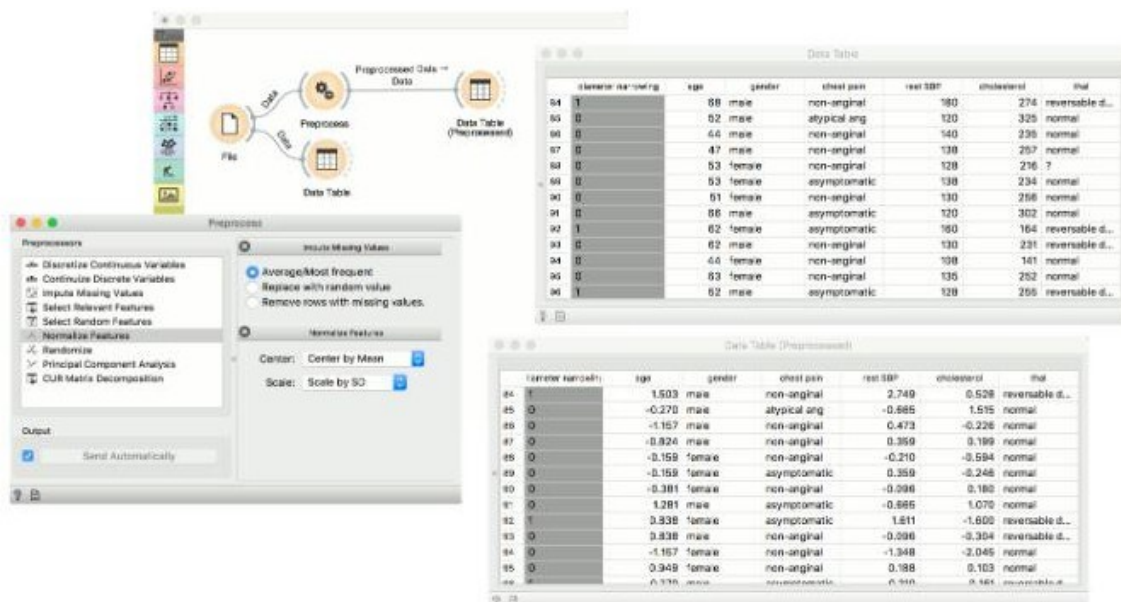
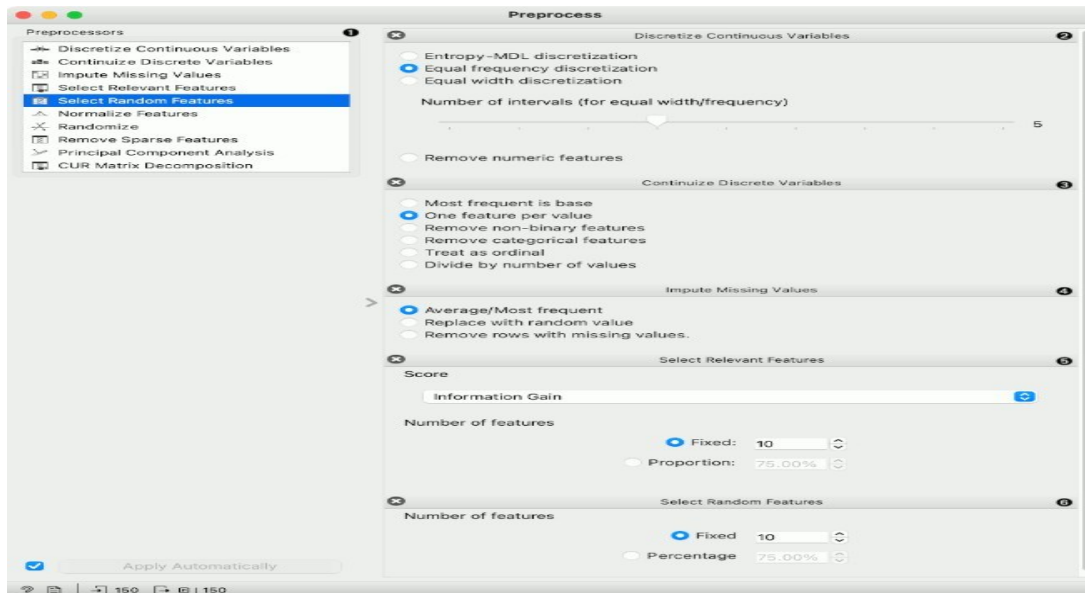
Example Implementation Preprocessing Student Dataset

1. **Load Data:**
 - Drag the 'File' widget to the canvas and load `students.csv`.
2. **Handle Missing Values:**
 - Connect 'File' to 'Edit Domain'.
 - Impute missing values with the median.
3. **Normalize Data:**
 - Connect 'Edit Domain' to 'Normalize'.
 - Standardize the data.
4. **Feature Selection:**
 - Connect 'Normalize' to 'Select Columns'.
 - Select relevant features like Age, Grade, Hours of Study.

Creating a Data Cube

1. **Load Data:**
 - Drag the 'File' widget to the canvas and load `3d_dataset.csv`.
2. **Create Pivot Table:**
 - Connect 'File' to 'Pivot Table'.
 - Configure rows to `Product`, columns to `Region`, and layers to `Time`.
3. **Analyze:**
 - Use the pivot table to aggregate data (e.g., sum of sales).





Post-Viva Questions:

1. What challenges did you face during the preprocessing step?
2. How did the data cube improve your understanding of the dataset?
3. Can you think of a real-time application where such preprocessing is critical?
4. What preprocessing techniques would you use for textual data?
5. How does the Orange tool visually aid in preprocessing better than manual scripting?

Innovative Approaches:

1. Use of Orange's Rank Widget to prioritize relevant features post-preprocessing.
2. Comparative preprocessing: Apply different strategies (mean, median, mode imputation) and visualize how these impact subsequent data analysis.
3. Real-time datasets: Replace static datasets with institution-specific academic or labor records to localize insights.

RESULT:

By following these steps, you can perform data preprocessing on student and labor datasets and implement a data cube for data warehousing in Orange. Preprocessing ensures your data is clean and ready for analysis, while the pivot table simulates a data cube, allowing multi-dimensional analysis.

AIM:

Data cleaning, particularly handling missing and noisy data, is essential for ensuring accurate analysis and results. Here's how you can implement various mechanisms for handling missing and noisy data using Orange.

TOOLS AND SOFTWARE USED:

1. Orange Data Mining Software
2. Dataset: Loan dataset with missing and noisy entries
3. Widgets Used: File, Data Table, Feature Statistics, Impute, Save Data

DESCRIPTION:**Data Cleaning:**

Data cleaning is the process of fixing or removing incorrect, corrupted, incorrectly formatted, duplicate or incomplete data within a dataset.

Dataset:

Datasets involve a large amount of data points grouped into one table.

PROBLEM DEFINITION:

In this experiment, the goal is to implement various mechanisms for handling missing values and addressing noisy data in a Loan dataset using the Orange software.

PRE-VIVA QUESTIONS:

1. What is the difference between missing and noisy data?
2. Explain two methods of handling missing values.
3. Why is data cleaning essential before visualization?
4. What are the types of noise that may occur in a dataset?
5. What is the purpose of the Impute widget in Orange?

PROCEDURE:

Handling Missing Data

1. **Load Data:**
 - Use the 'File' widget to load your dataset.
2. **Data Imputation:**
 - Use the 'Impute' widget to handle missing values.
 - Connect the 'File' widget to the 'Impute' widget.

Options in the Impute Widget:

- **Mean/Median Imputation:** Replace missing values with the mean or median of the column.
- **Most Frequent:** Replace missing values with the most frequent value (mode) in the column.
- **Model-Based Imputation:** Use a machine learning model to predict and replace

- missing values based on other features.
- **User-Defined:** Specify a custom value to replace missing values.
- 3. **Removing Rows/Columns with Missing Data:**
 - Use the 'Select Columns' widget to remove columns with missing values if they are not essential.
 - Use the 'Edit Domain' widget to remove rows with missing values.

Handling Noisy Data

1. **Load Data:**
 - Use the 'File' widget to load your dataset.
2. **Data Smoothing:**
 - Use the 'Smoothing' widget to handle noisy data.
 - Connect the 'File' widget to the 'Smoothing' widget.

Options in the Smoothing Widget:

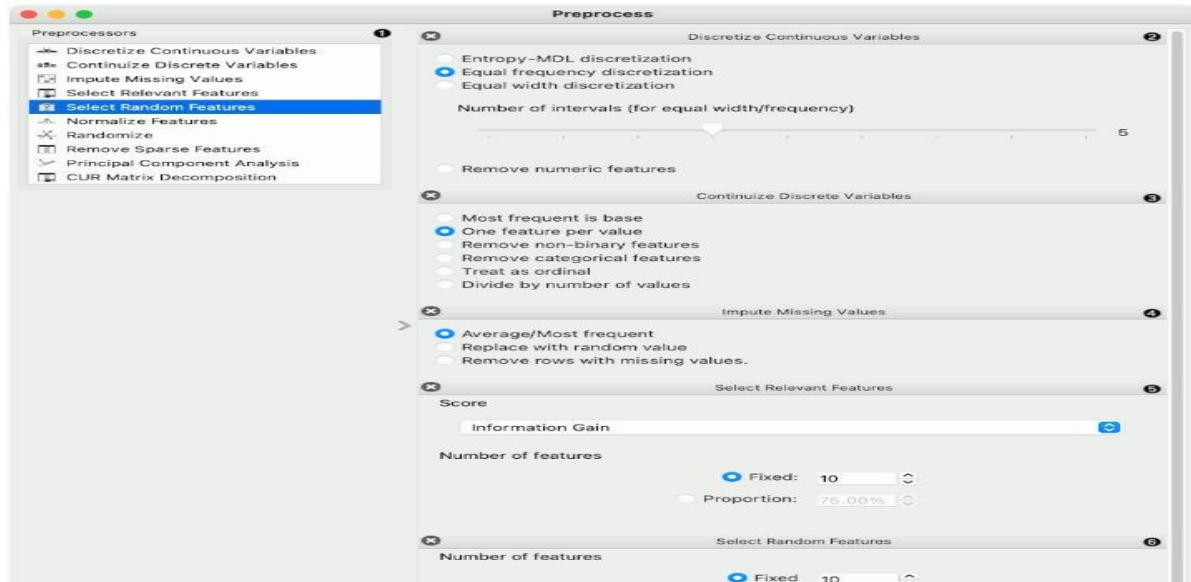
 - **Binning:** Group continuous data into bins to reduce noise.
 - **Moving Average:** Use moving averages to smooth time-series data.
 - **Gaussian Smoothing:** Apply Gaussian smoothing to reduce noise.
3. **Outlier Detection and Removal:**
 - Use the 'Outliers' widget to detect and handle outliers.
 - Connect the 'File' widget to the 'Outliers' widget.

Options in the Outliers Widget:

 - **Z-Score:** Detect outliers based on the Z-score. Values above or below a certain threshold (e.g., 3 standard deviations from the mean) are considered outliers.
 - **IQR Method:** Detect outliers using the Interquartile Range (IQR). Values outside 1.5 times the IQR above the third quartile or below the first quartile are considered outliers.
 - **Isolation Forest:** Use an Isolation Forest model to detect outliers.

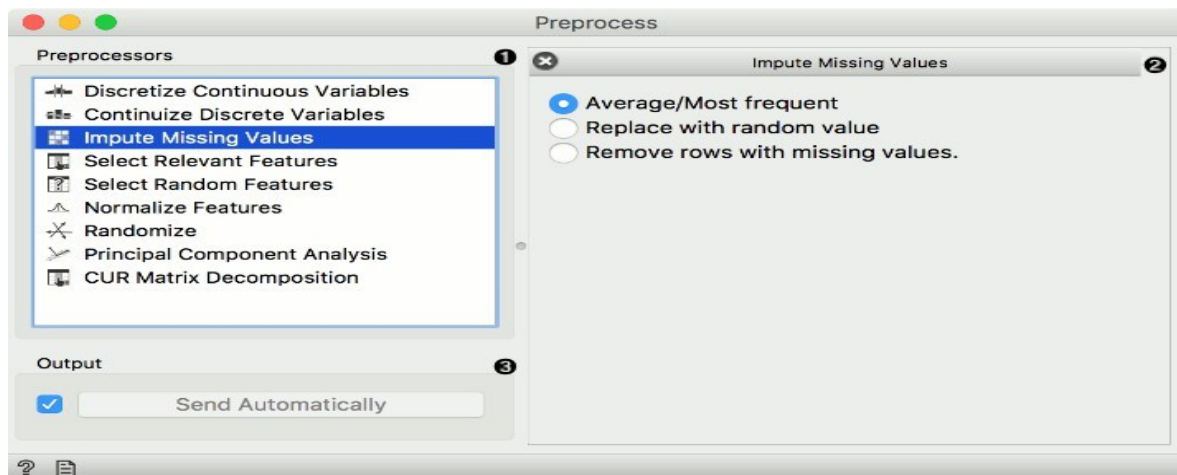
Example Workflow in Orange

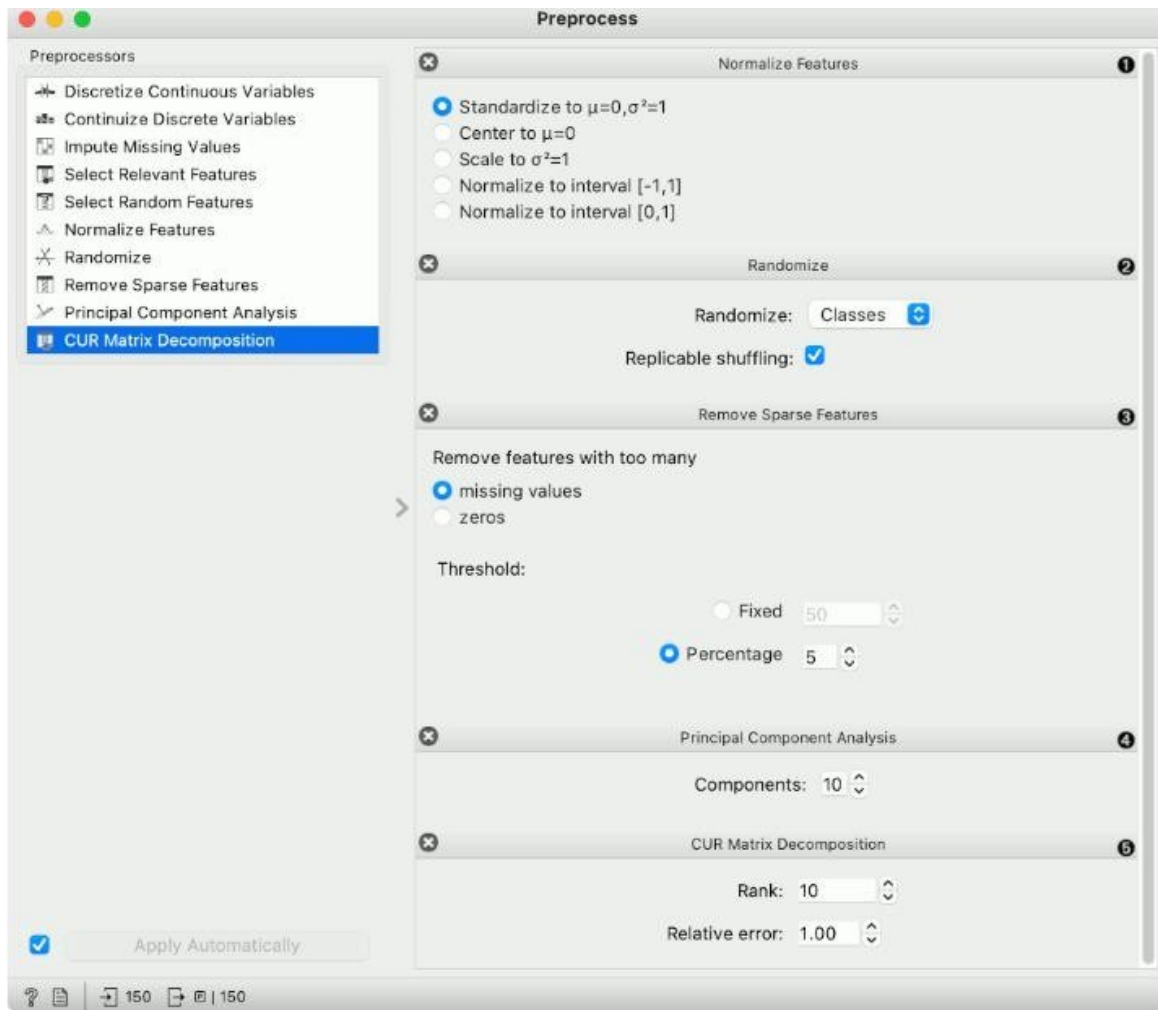
1. **Load and Inspect Data:**
 - Use the 'File' widget to load `dataset.csv`.
 - Connect it to the 'Data Table' widget to inspect the data.
2. **Handle Missing Data:**
 - Connect the 'File' widget to the 'Impute' widget.
 - In the 'Impute' widget, choose 'Mean/Median' for numeric features and 'Most Frequent' for categorical features.
 - Connect the 'Impute' widget to a 'Data Table' widget to inspect the imputed data.
3. **Remove Columns with Excessive Missing Data:**
 - Connect the 'Impute' widget to the 'Select Columns' widget.
 - In the 'Select Columns' widget, manually remove columns with a high percentage of missing values.
4. **Handle Noisy Data:**
 - Connect the 'Select Columns' widget to the 'Smoothing' widget.
 - In the 'Smoothing' widget, apply 'Binning' for continuous variables.
 - Connect the 'Smoothing' widget to a 'Data Table' widget to inspect the smoothed data.



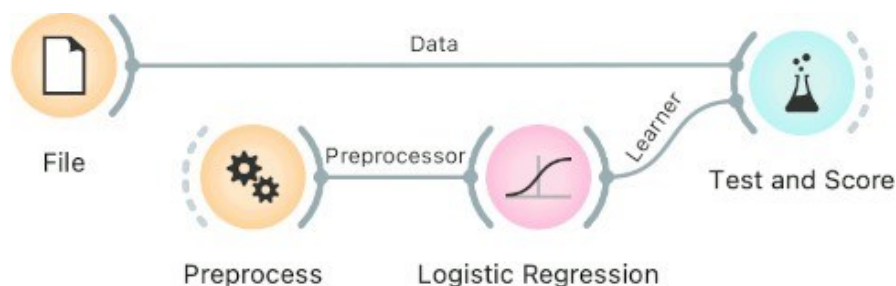
5. Detect and Remove Outliers:

- Connect the 'Smoothing' widget to the 'Outliers' widget.
- In the 'Outliers' widget, choose the 'Z-Score' method and set a threshold (e.g., 3).
- Connect the 'Outliers' widget to a 'Data Table' widget to inspect the data with outliers removed.

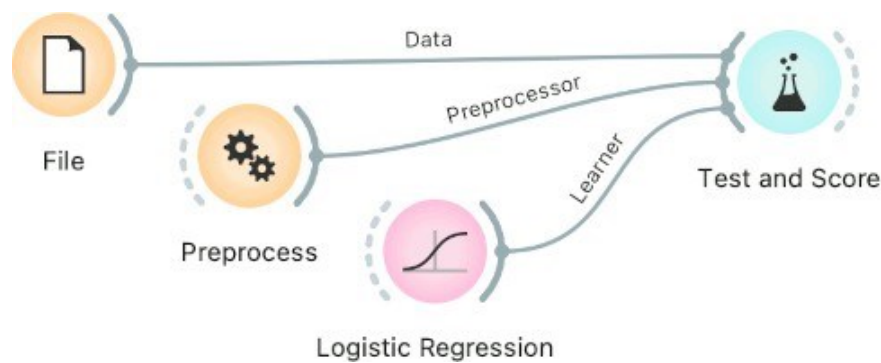




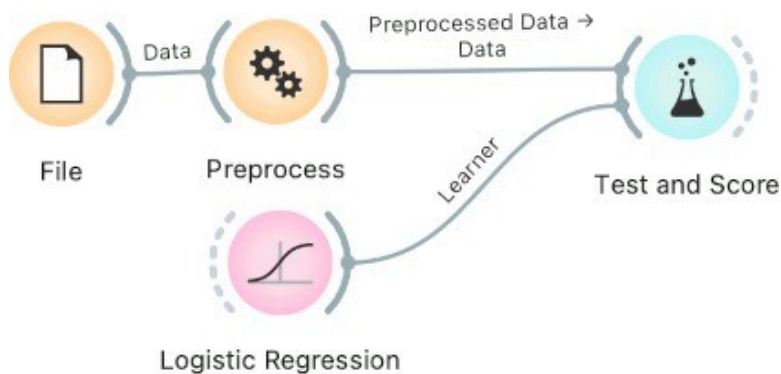
1. Connect **Preprocess** to the learner. This will override the default preprocessing pipeline for the learner and apply only custom preprocessing pipeline (default preprocessing steps are described in each learner's documentation).



2. Connect **Preprocess** to Test and Score. This will apply the preprocessors to each batch within cross-validation. Then the learner's preprocessors will be applied to the preprocessed subset.



Finally, there's a wrong way to do it. Connecting **Preprocess** directly to the original data and outputting preprocessed data set will likely overfit the model. Don't do it.



Examples

In the first example, we have used the heart_disease.tab dataset available in the dropdown menu of the File widget. then we used **Preprocess** to impute missing values and normalize features. We can observe the changes in the Data Table and compare it to the non-processed data.

The screenshot shows the Orange3 software interface. On the left, the 'Preprocess' widget is configured with the following settings:

- Preprocessors:**
 - Discretize Continuous Variables
 - Continue Discrete Variables
 - Impute Missing Values
 - Select Relevant Features
 - Select Random Features
 - Normalize Features
 - Randomize
 - Principal Component Analysis
 - CUR Matrix Decomposition
- Impute Missing Values:**
 - ☒ Average/Most frequent
 - ☐ Replace with random value
 - ☐ Remove rows with missing values
- Normalize Features:**
 - Center:** Center by Mean
 - Scale:** Scale by SD
- Output:**
 - ☒ Send Automatically

On the right, two data tables are shown for comparison:

Data Table

	diameter normalizing	age	gender	chest pain	rest SGP	cholesterol	thal
84	1	68	male	non-anginal	180	274	reversible d...
85	0	52	male	atypical ang	120	325	normal
86	0	44	male	non-anginal	140	235	normal
87	0	47	male	non-anginal	138	267	normal
88	0	53	female	non-anginal	128	218	?
89	0	53	female	asymptomatic	138	234	normal
90	0	51	female	non-anginal	130	258	normal
91	0	88	male	asymptomatic	120	302	normal
92	1	82	female	asymptomatic	160	184	reversible d...
93	0	82	male	non-anginal	130	231	reversible d...
94	0	44	female	non-anginal	108	141	normal
95	0	83	female	non-anginal	135	252	normal
96	1	52	male	asymptomatic	128	255	reversible d...

Data Table (Preprocessed)

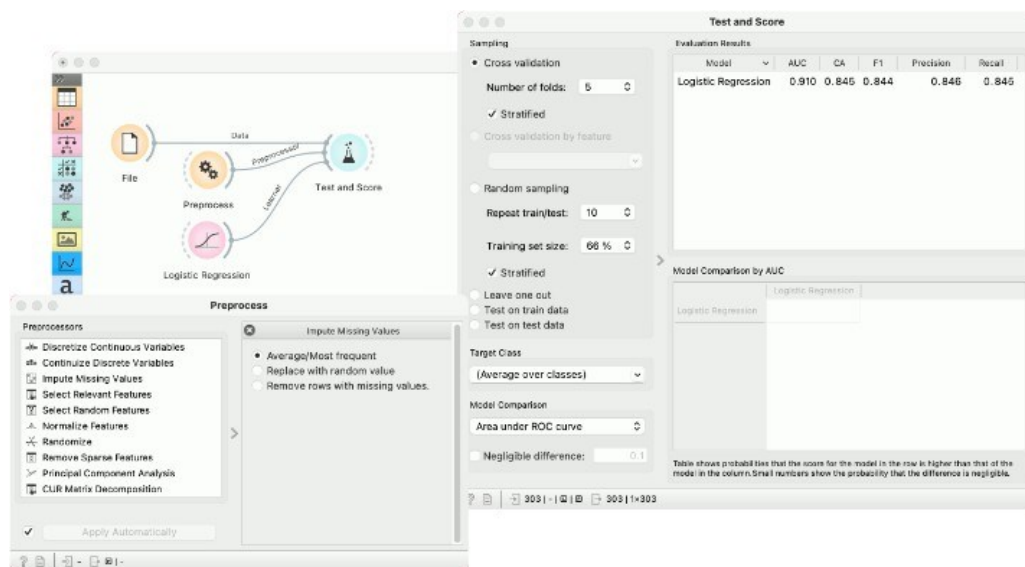
	diameter normalizing	age	gender	chest pain	rest SGP	cholesterol	thal
84	1	1.503	male	non-anginal	2.749	0.528	reversible d...
85	0	-0.270	male	atypical ang	-0.685	1.518	normal
86	0	-1.157	male	non-anginal	0.473	-0.228	normal
87	0	-0.824	male	non-anginal	0.359	0.199	normal
88	0	-0.159	female	non-anginal	-0.210	-0.594	normal
89	0	-0.159	female	asymptomatic	0.359	-0.246	normal
90	0	-0.381	female	non-anginal	-0.096	0.180	normal
91	0	1.281	male	asymptomatic	-0.665	1.070	normal
92	1	0.838	female	asymptomatic	1.611	-1.600	reversible d...
93	0	0.838	male	non-anginal	-0.096	-0.304	reversible d...
94	0	-1.157	female	non-anginal	-1.348	-2.045	normal
95	0	0.949	female	non-anginal	0.188	0.103	normal
96	1	0.735	male	asymptomatic	0.749	0.182	reversible d...

Preprocess for predictive modeling:

This time we are using the heart_disease.tab data from the File widget. You can access the data in the dropdown menu. This is a dataset with 303 patients that came to the doctor suffering from a chest pain. After the tests were done, some patients were found to have diameter narrowing and others did not (this is our class variable).

Some values are missing in our data set, so we would like to impute missing values before evaluating the model. We do this by passing a preprocessor directly to Test and Score. In **Preprocess**, we set the correct preprocessing pipeline (in our example only a single preprocessor with Impute missing values), then connect it to the Preprocessor input of Test and Score.

We also pass the data and the learner (in this case, a Logistic Regression). This is the correct way to pass a preprocessor to cross-validation as each fold will independently get preprocessed in the training phase. This is particularly important for feature



POST-VIVA QUESTIONS:

1. How did the cleaning methods improve the data quality?
2. Did imputation change the data distribution significantly?
3. How would you handle categorical missing values differently?
4. Can noise removal ever remove useful information?
5. Suggest improvements for large-scale dataset cleaning.

INNOVATIVE APPROACHES:

1. Statistical Summary Integration: Use the Feature Statistics widget to dynamically prioritize fields for cleaning.
2. Comparative Evaluation: Perform imputation using mean, median, and random assignment—compare results visually.
3. Noise Filtering: Apply binning and outlier detection techniques before imputation.

RESULT:

By following these steps, you can effectively clean your data in Orange, handling both missing and noisy data. These preprocessing steps are crucial for ensuring that your data is of high quality, leading to more accurate and reliable analysis and modeling results.

AIM:

To perform exploratory analysis using k-means and MST (Minimum Spanning Tree) based clustering techniques in Orange, and to develop a methodology for assessing clusters, follow these steps:

TOOLS AND SOFTWARE USED:

1. Orange Data Mining Software
2. Dataset: Any publicly available multidimensional dataset
3. Widgets Used: File, Data Table, Select Columns, K-Means, Hierarchical Clustering, Scatter Plot, Silhouette Plot

DESCRIPTION:**K-means clustering:**

K-means clustering is primarily an exploratory technique to discover the structure of the data that you might not have notice before and as a prelude to more focused analysis or decision processes.

MST:

The minimum spanning tree- (MST-) based clustering method can identify clusters of arbitrary shape by removing inconsistent edges. The definition of the inconsistent edges is a major issue that has to be addressed in all MST-based clustering algorithms.

PROBLEM DEFINITION:

In this experiment, the goal is to implement exploratory analysis, develop k-means and MST based clustering techniques and the methodology for assessment of clusters for dataset using Orange software.

Pre-Viva Questions

1. What is clustering, and how is it useful in EDA?
2. Differentiate between K-means and MST-based clustering.
3. What does the term 'k' represent in K-means?
4. Why is visual representation of clusters important?
5. What metrics can be used to validate cluster quality?

PROCEDURE:**K-Means Clustering in Orange**

1. **Load Data:**
 - Use the 'File' widget to load your dataset.
2. **Data Preprocessing:**
 - Use the 'Preprocess' widget to handle any missing values or normalize/standardize the data.
3. **K-Means Clustering:**
 - Drag the 'K-Means' widget into the workspace.

- Connect the 'File' widget to the 'K-Means' widget.
- Configure the 'K-Means' widget by setting the number of clusters (k).
- 4. **Inspect Clusters:**
 - Connect the 'K-Means' widget to the 'Scatter Plot' widget to visualize the clusters.
 - Optionally, connect the 'K-Means' widget to the 'Data Table' widget to inspect the clustered data.

MST (Minimum Spanning Tree) Based Clustering in Orange

Orange does not directly provide an MST-based clustering widget, but you can implement a similar approach using hierarchical clustering or custom scripts.

1. **Load Data:**
 - Use the 'File' widget to load your dataset.
2. **Data Preprocessing:**
 - Use the 'Preprocess' widget to handle any missing values or normalize/standardize the data.
3. **Hierarchical Clustering** (as a proxy for MST-based clustering):
 - Drag the 'Hierarchical Clustering' widget into the workspace.
 - Connect the 'File' widget to the 'Hierarchical Clustering' widget.
 - Configure the 'Hierarchical Clustering' widget by selecting an appropriate linkage method (e.g., single linkage for an MST-like effect).
4. **Inspect Dendrogram:**
 - Connect the 'Hierarchical Clustering' widget to the 'Dendrogram' widget to visualize the hierarchical structure.
 - Use the 'Silhouette Plot' widget to evaluate the clustering quality.
 -

Methodology for Assessment of Clusters

1. **Load and Cluster Data:**
 - Perform clustering using the 'K-Means' and 'Hierarchical Clustering' widgets as described above.
2. **Cluster Evaluation Metrics:**
 - **Silhouette Score:**
 - Drag the 'Silhouette Plot' widget into the workspace.
 - Connect it to the 'K-Means' or 'Hierarchical Clustering' widget to evaluate the silhouette score of each cluster.
 - **Elbow Method:**
 - Use the 'Elbow Plot' to determine the optimal number of clusters for k-means. This can be done by manually setting different values of k and plotting the within-cluster sum of squares.
 - **Dunn Index and Davies-Bouldin Index:**
 - Use custom scripts or external tools to compute these indices if not directly available in Orange.
3. **Visualization:**
 - Use the 'Scatter Plot', 'Silhouette Plot', and 'Dendrogram' widgets to visualize the clusters and their quality.
 - For multi-dimensional data, use the 't-SNE' or 'PCA' widgets to reduce dimensions and visualize clusters.

Example Workflow in Orange

1. **Load and Preprocess Data:**
 - Use the 'File' widget to load `dataset.csv`.
 - Connect it to the 'Preprocess' widget to handle missing values and normalize data.
 - Connect the 'Preprocess' widget to a 'Data Table' widget to inspect the preprocessed data.
2. **Perform K-Means Clustering:**

- Connect the 'Preprocess' widget to the 'K-Means' widget.
 - Set the number of clusters (e.g., $k=3$).
 - Connect the 'K-Means' widget to the 'Scatter Plot' widget to visualize clusters.
3. **Evaluate Clusters:**
- Connect the 'K-Means' widget to the 'Silhouette Plot' widget to evaluate the silhouette scores.
 - Manually run k-means for different k values and plot the within-cluster sum of squares to create an Elbow Plot.
4. ****Perform Hier**

Example Workflow in Orange (continued)

5. **Perform Hierarchical Clustering:**
- Connect the 'Preprocess' widget to the 'Hierarchical Clustering' widget.
 - Configure the widget with the desired linkage method (e.g., single linkage for MST-like clustering).
 - Connect the 'Hierarchical Clustering' widget to the 'Dendrogram' widget to visualize the dendrogram.
 - Optionally, connect to the 'Silhouette Plot' widget to evaluate cluster quality.

Detailed Steps K-Means Clustering

1. **Load Data:**
 - Drag the 'File' widget to the canvas.
 - Select your dataset file (e.g., `dataset.csv`).
2. **Preprocess Data:**
 - Drag the 'Preprocess' widget to the canvas.
 - Connect the 'File' widget to the 'Preprocess' widget.
 - Configure preprocessing steps such as imputing missing values, normalizing, or standardizing data.
3. **K-Means Clustering:**
 - Drag the 'K-Means' widget to the canvas.
 - Connect the 'Preprocess' widget to the 'K-Means' widget.
 - Set the number of clusters (k) in the 'K-Means' widget.
 - Optionally, connect to the 'Data Table' widget to inspect clustered data.
4. **Inspect Clusters:**
 - Drag the 'Scatter Plot' widget to the canvas.
 - Connect the 'K-Means' widget to the 'Scatter Plot' widget to visualize clusters.
5. **Cluster Evaluation:**
 - Drag the 'Silhouette Plot' widget to the canvas.
 - Connect the 'K-Means' widget to the 'Silhouette Plot' widget to evaluate silhouette scores.
 - For the Elbow Method, manually adjust k and observe the within-cluster sum of squares.

Hierarchical Clustering (MST-like)

1. **Load Data:**
 - Follow the same steps as in K-Means to load and preprocess the data.
2. **Hierarchical Clustering:**
 - Drag the 'Hierarchical Clustering' widget to the canvas.
 - Connect the 'Preprocess' widget to the 'Hierarchical Clustering' widget.
 - Configure linkage method, such as single linkage.
3. **Inspect Dendrogram:**
 - Drag the 'Dendrogram' widget to the canvas.
 - Connect the 'Hierarchical Clustering' widget to the 'Dendrogram' widget.
4. **Cluster Evaluation:**
 - Drag the 'Silhouette Plot' widget to the canvas.
 - Connect the 'Hierarchical Clustering' widget to the 'Silhouette Plot' widget to evaluate clusters.

Assessment Methodology

1. Silhouette Score:

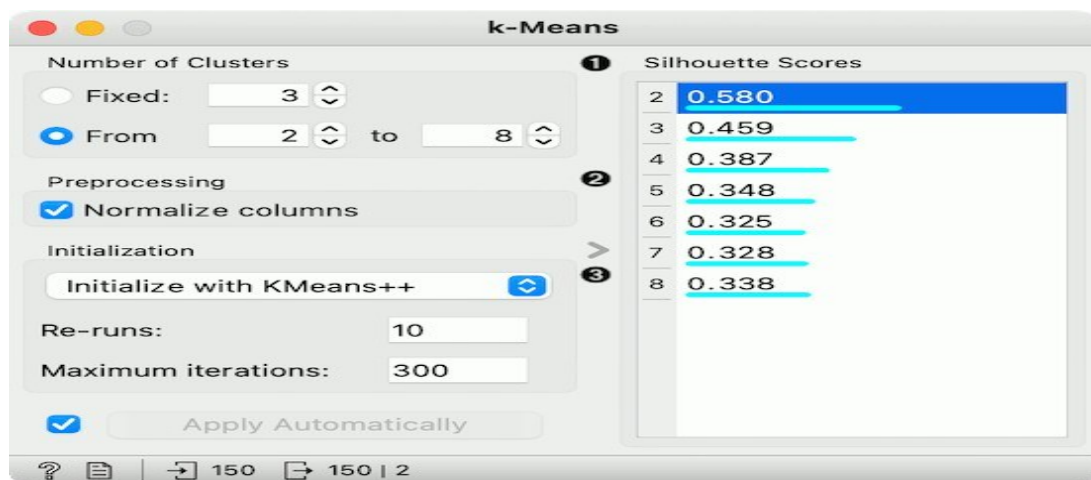
- Measures how similar an object is to its own cluster compared to other clusters.
- Use the 'Silhouette Plot' widget to assess cluster quality.

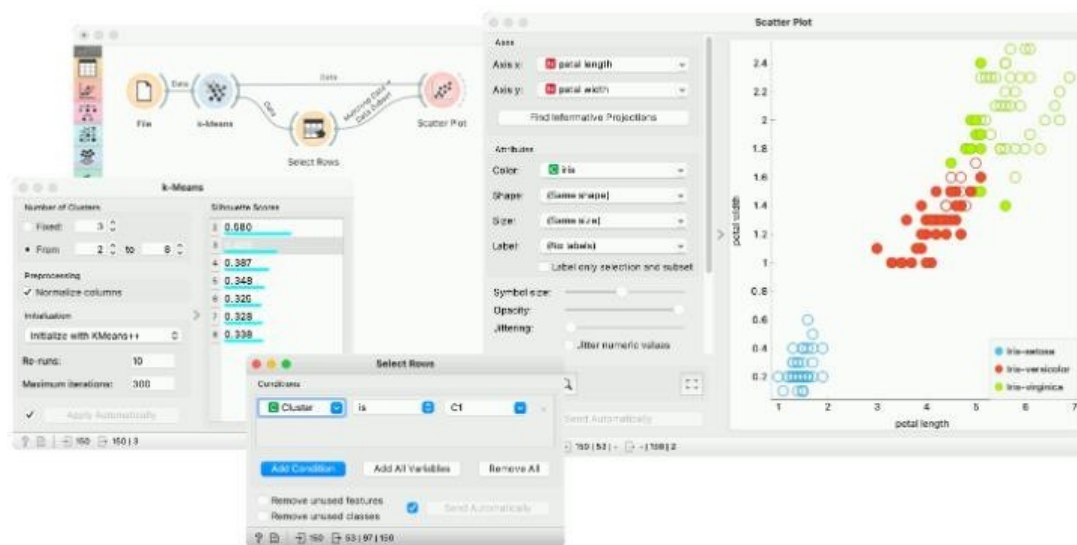
2. Elbow Method:

- Plot within-cluster sum of squares for different values of k.
- Look for an "elbow point" where the decrease in sum of squares becomes linear.

3. Additional Metrics:

- **Dunn Index:** Measures compactness and separation of clusters.
- **Davies-Bouldin Index:** Measures average similarity ratio of each cluster with its most similar cluster.
- Use custom scripts or external tools for these metrics if necessary.





POST-VIVA QUESTIONS:

1. Which clustering method performed better and why?
2. How did you decide on the number of clusters (k)?
3. What was the distribution of points among the clusters?
4. Could you identify any outliers?
5. How would clustering help in anomaly detection?

INNOVATIVE APPROACHES:

1. Silhouette Score Evaluation: Assess cluster quality using silhouette plots.
2. Dual Clustering Comparison: Compare results from K-means and MST visually.
3. Cluster Label Overlay: Annotate clusters to understand feature distribution.

RESULT:

By following these steps in Orange, you can effectively perform k-means and MST-based clustering, and assess the quality of your clusters using various evaluation metrics. This structured approach ensures thorough exploratory analysis and meaningful insights from your data.

AIM:

Designing algorithms for association rule mining involves several key steps: preparing the data, generating frequent itemsets, and creating association rules. Orange uses established algorithms like Apriori or FP-Growth for these tasks. Below is a detailed description of how you can implement and understand these algorithms within the Orange framework.

TOOLS AND SOFTWARE USED:

1. Orange Data Mining Software
2. Dataset: Transactional dataset
3. Widgets Used: File, Edit Domain, Association Rules, Data Table

PROBLEM DEFINITION:

In this experiment, the goal is to design associate rule mining algorithm using the Orange software.

PROCEDURE:**Step 1: Data Preparation**

1. **Import Dataset:** Start by loading your dataset into Orange using the 'File' widget. Ensure that the dataset is in the correct format (transactions in rows and items in columns).
2. **Preprocessing:** If necessary, preprocess your data using the 'Preprocess' widget. This might include discretizing continuous variables, handling missing values, or encoding categorical variables.

Step 2: Generating Frequent Itemsets

Orange typically uses the Apriori algorithm to generate frequent itemsets. Here's how it works:

1. **Set Support Threshold:** Define the minimum support threshold. This is the minimum frequency with which an itemset must appear in the transactions to be considered frequent.
2. **Initial Pass:**
 - Count the frequency of each item in the dataset.
 - Identify individual items that meet the minimum support threshold.
3. **Subsequent Passes:**
 - Generate candidate itemsets of length k from frequent itemsets of length $k-1$.
 - Count the frequencies of these candidate itemsets in the dataset.
 - Prune itemsets that do not meet the minimum support threshold.

- Repeat until no more frequent itemsets can be generated.

Step 3: Generating Association Rules

Once frequent itemsets are identified, association rules can be generated. Orange uses the following steps:

- 1. Generate Rules:**
 - For each frequent itemset, generate all non-empty subsets.
 - For each non-empty subset, create a rule of the form $\text{antecedent} \rightarrow \text{consequent}$, where the antecedent and consequent are disjoint subsets of the itemset.
- 2. Calculate Confidence:**
 - For each rule $A \rightarrow B$, calculate the confidence as $\text{Confidence}(A \rightarrow B) = \frac{\text{Support}(A \cup B)}{\text{Support}(A)}$.
- 3. Set Confidence Threshold:**
 - Retain rules that meet or exceed the minimum confidence threshold.
- 4. Additional Metrics (optional):**
 - Calculate additional metrics such as lift, leverage, and conviction to evaluate the strength of the rules.

Implementation in Orange

- 1. Load Data:**
 - Use the 'File' widget to load your dataset.
 - Connect the 'File' widget to a 'Data Table' widget to inspect the data.
- 2. Preprocess Data (if necessary):**
 - Use the 'Preprocess' widget for any required data transformations.
- 3. Generate Frequent Itemsets:**
 - Drag the 'Association Rules' widget to the canvas.
 - Connect the 'File' widget to the 'Association Rules' widget.
 - Set the support threshold in the 'Association Rules' widget.
- 4. Generate Association Rules:**
 - In the 'Association Rules' widget, set the confidence threshold.
 - Run the algorithm to generate rules.
- 5. Inspect Results:**
 - Connect the 'Association Rules' widget to a 'Data Table' widget to inspect the generated rules.
 - Optionally, use visualization widgets like 'Scatter Plot' or 'Heat Map' to explore the rules.

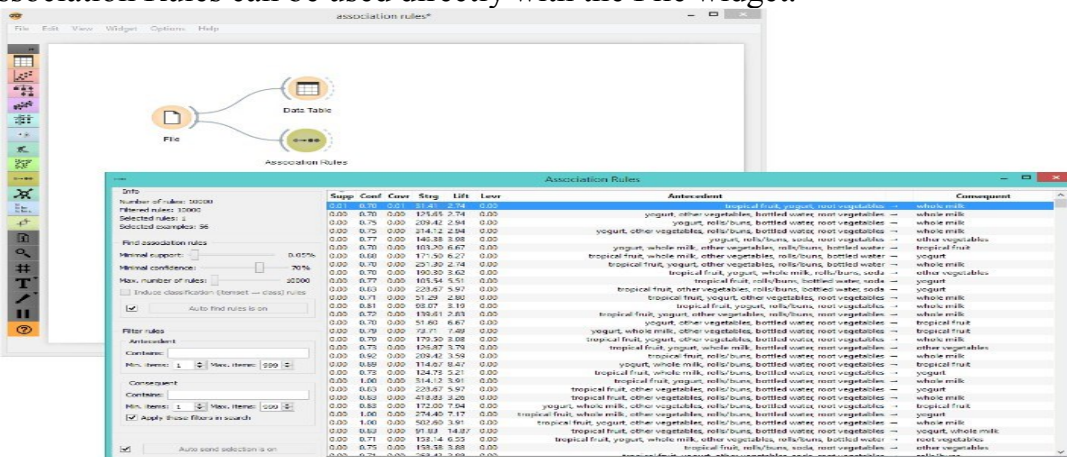
Example

Let's say you have a dataset of retail transactions, and you want to find associations between products:

- 1. Load Data:**
 - Use the 'File' widget to load the dataset (e.g., `transactions.csv`).
- 2. Set Parameters:**
 - In the 'Association Rules' widget, set a minimum support threshold of 0.01 (1%).
 - Set a minimum confidence threshold of 0.6 (60%).
- 3. Generate Rules:**
 - The widget will output rules such as:
 - $\{\text{bread}\} \rightarrow \{\text{butter}\}$ with confidence 0.8 and support 0.05.
 - $\{\text{milk}, \text{bread}\} \rightarrow \{\text{eggs}\}$ with confidence 0.7 and support 0.03.

Association Rules									
Info	Supp	Conf	Covr	Strg	Lift	Levr	Antecedent	Consequent	
Number of rules: 10000	0.01	0.70	0.01	31.41	2.74	0.00	tropical fruit, yogurt, root vegetables	whole milk	
Filtered rules: 2329	0.00	0.70	0.00	125.65	2.74	0.00	yogurt, other vegetables, bottled water, root vegetables	whole milk	
Selected rules: 1	0.00	0.75	0.00	209.42	2.94	0.00	yogurt, rolls/buns, bottled water, root vegetables	whole milk	
Selected examples: 56	0.00	0.75	0.00	314.12	2.94	0.00	yogurt, other vegetables, rolls/buns, bottled water, root vegetables	whole milk	
Find association rules	0.00	0.70	0.00	251.30	2.74	0.00	tropical fruit, yogurt, other vegetables, rolls/buns, bottled water	whole milk	
Minimal support: 0.05%	0.00	0.71	0.00	51.79	2.80	0.00	tropical fruit, yogurt, other vegetables, root vegetables	whole milk	
Minimal confidence: 70%	0.00	0.81	0.00	93.07	3.19	0.00	tropical fruit, yogurt, rolls/buns, root vegetables	whole milk	
Max. number of rules: 10000	0.00	0.72	0.00	139.61	2.83	0.00	tropical fruit, yogurt, other vegetables, rolls/buns, root vegetables	whole milk	
<input type="checkbox"/> Induce classification (blossom → class) rules	0.00	0.79	0.00	179.50	3.08	0.00	tropical fruit, yogurt, other vegetables, bottled water, root vegetables	whole milk	
<input checked="" type="checkbox"/> Auto find rules is on	0.00	0.92	0.00	209.42	3.59	0.00	tropical fruit, rolls/buns, bottled water, root vegetables	whole milk	
Filter rules	0.00	1.00	0.00	314.12	3.91	0.00	tropical fruit, yogurt, rolls/buns, bottled water, root vegetables	whole milk	
Antecedent	0.00	0.83	0.00	418.63	3.26	0.00	tropical fruit, yogurt, other vegetables, rolls/buns, root vegetables	whole milk	
Contains: root vegetables	0.00	1.00	0.00	302.80	3.91	0.00	tropical fruit, yogurt, other vegetables, bottled water, root vegetables	whole milk	
Min. items: 1 Max. items: 999	0.00	0.83	0.00	61.83	14.87	0.00	tropical fruit, other vegetables, rolls/buns, bottled water, root vegetables	yogurt, whole milk	
Consequent	0.00	0.75	0.00	314.12	2.94	0.00	yogurt, rolls/buns, root vegetables, shopping bags	whole milk	
Contains: whole milk	0.00	0.71	0.00	179.50	2.80	0.00	other vegetables, soda, root vegetables, shopping bags	whole milk	
Min. items: 1 Max. items: 999	0.00	0.77	0.00	314.12	2.94	0.00	tropical fruit, yogurt, other vegetables, rolls/buns, sausage	whole milk	
<input checked="" type="checkbox"/> Apply these filters in search	0.00	0.77	0.00	71.80	3.02	0.00	tropical fruit, root vegetables, sausage	whole milk	
<input checked="" type="checkbox"/> Auto send selection is on	0.00	0.94	0.00	157.06	3.67	0.00	tropical fruit, yogurt, root vegetables, sausage	whole milk	
	0.00	0.71	0.00	147.82	2.76	0.00	tropical fruit, other vegetables, root vegetables, sausage	whole milk	
	0.00	0.86	0.00	359.00	3.35	0.00	tropical fruit, yogurt, other vegetables, root vegetables, sausage	whole milk	
	0.00	1.00	0.00	251.30	3.91	0.00	tropical fruit, rolls/buns, root vegetables, sausage	whole milk	
	0.00	1.00	0.00	359.00	3.91	0.00	tropical fruit, yogurt, rolls/buns, root vegetables, sausage	whole milk	
	0.00	0.70	0.00	55.10	12.49	0.00	tropical fruit, rolls/buns, root vegetables, sausage	yogurt, whole milk	
	0.00	0.73	0.00	167.53	2.87	0.00	yogurt, other vegetables, soda, pastry	whole milk	
	0.00	0.91	0.00	228.45	3.56	0.00	tropical fruit, yogurt, pastry, root vegetables	whole milk	
	0.00	0.73	0.00	114.73	2.85	0.00	yogurt, other vegetables, pastry, root vegetables	whole milk	
	0.00	0.83	0.00	418.63	3.26	0.00	tropical fruit, yogurt, other vegetables, pastry, root vegetables	whole milk	
	0.00	0.75	0.00	209.42	2.94	0.00	yogurt, rolls/buns, pastry, root vegetables	whole milk	
	0.00	1.00	0.00	412.62	3.01	0.00	tropical fruit, yogurt, rolls/buns, pastry, root vegetables	whole milk	

Example
Association Rules can be used directly with the File widget.



POST-VIVA QUESTIONS:

1. Which rules were the strongest based on confidence?
2. How does support affect the number of rules generated?
3. Can association rules lead to false positives?
4. Suggest one domain-specific application for this technique.
5. How would you optimize performance for large datasets?

INNOVATIVE APPROACHES:

1. Frequent Itemset Visualization: Graphical view of itemsets using network plots.
2. Python Integration: Extend Orange with embedded Python scripts for deeper rule filtering.
3. Threshold Tuning: Experiment with various support-confidence levels.

RESULT:

By following these steps, you can effectively design and implement association rule mining algorithms in Orange. Adjusting the support and confidence thresholds allows you to fine-tune the analysis to discover the most significant and interesting rules in your dataset.

AIM:

To derive the hypothesis for association rules to discovery of strong association rules; use confidence and support thresholds.

TOOLS AND SOFTWARE USED:

1. Orange Data Mining Software
2. Python Script for advanced filtering
3. Widgets Used: File, Edit Domain, Association Rules, Python Script, Data Table

DESCRIPTION:**ASSOCIATION RULES:**

- An association rule is typically in the form "If A, then B," where A and B are sets of items.
- Support for a rule $A \rightarrow B$ is the percentage of transactions that contain both A and B.
- Confidence for a rule $A \rightarrow B$ is the percentage of transactions containing A that also contain B.

HYPOTHESIS GENERATION:

- The hypothesis involves creating rules like "If A, then B" and testing these rules to see if they meet predefined thresholds for support and confidence.

SUPPORT THRESHOLD:

- The minimum support level set for the rules to be considered valid.

CONFIDENCE THRESHOLD:

- The minimum confidence level required for a rule to be accepted.

PRE-VIVA QUESTIONS:

1. What constitutes a valid hypothesis in rule mining?
2. How are confidence and support used in rule testing?
3. Explain the purpose of hypothesis generation.
4. What makes a rule statistically significant?
5. What is a frequent itemset?

PROCEDURE:**Hypothesis Generation**

1. **Define the Objective:** Determine what you aim to find with association rule mining. This could be identifying products frequently bought together, common sequences of events, etc.
2. **Select Data:** Choose the dataset that contains the items or events you want to analyze.

This dataset should be formatted appropriately for association rule mining, typically in a transactional format.

3. **Identify Items and Transactions:** Define what constitutes an item and a transaction in your dataset. An item could be a product, an event, etc., and a transaction could be a purchase instance, a session, etc.

Discovering Strong Association Rules

To discover strong association rules using Orange, you need to set up and apply association rule mining algorithms, adjusting the confidence and support thresholds to filter out the most significant rules.

Steps in Orange:

1. **Load Data:** Import your dataset into Orange. You can do this by dragging the 'File' widget and connecting it to the 'Data Table' widget to view the data.
2. **Preprocess Data:** Ensure your data is in the correct format. If needed, use the 'Preprocess' widget to transform your data.
3. **Apply Association Rule Mining:**
 - Drag the 'Association Rules' widget into the workspace.
 - Connect it to the 'File' widget to feed your dataset into the Association Rules widget.
4. **Set Parameters:** In the 'Association Rules' widget:
 - **Support Threshold:** This parameter determines the minimum frequency at which an itemset appears in the dataset. Set a value that is suitable for your dataset size and the rarity of the combinations you are interested in. For example, a support threshold of 0.01 means that the itemset must appear in at least 1% of all transactions.
 - **Confidence Threshold:** This parameter measures the likelihood that a rule is true for a transaction. Set a confidence level that reflects the strength of the rules you are interested in. For example, a confidence threshold of 0.7 means the rule must be true in at least 70% of the cases where the antecedent occurs.
5. **Generate Rules:** Run the algorithm to generate association rules. The output will be a list of rules that meet the specified support and confidence thresholds.
6. **Evaluate and Filter Rules:** Examine the generated rules. You can further filter them by adjusting the thresholds or by using additional metrics such as lift or leverage.

Example of Hypothesis and Rules:

- **Hypothesis:** Customers who buy bread also buy butter.
 - **Support:** The proportion of transactions that include both bread and butter.
 - **Confidence:** The proportion of transactions that include butter among those that include bread.
- **Rule:**
 - **IF** {bread} **THEN** {butter}
 - Support: 0.05 (5% of all transactions include both bread and butter)
- Confidence: 0.8 (80% of transactions that include bread also include butter) The widget finds frequent items in a data set based on a measure of support for the rule.

Frequent Itemsets

Info

Number of itemsets: 122
 Selected itemsets: 1
 Selected examples: 2513

Expand all Collapse all

Find itemsets

Minimal support: 2%
 Max. number of itemsets: 10000

☒ Auto find itemsets is on

Filter itemsets

Contains:
 Min. items: 1 Max. items: 999

☒ Apply these filters in search

☒ Auto send selection is on

Itemsets	Support	%
whole milk	2513	25.55
other vegetables	736	7.483
root vegetables	228	2.318
rolls/buns	557	5.663
soda	394	4.006
bottled water	338	3.437
root vegetables	481	4.891
shopping bags	241	2.45
sausage	294	2.989
pastry	327	3.325
bottled beer	201	2.044
newspapers	269	2.735
pip fruit	296	3.01
fruit/vegetable juice	262	2.664
whipped/sour cream	317	3.223
brown bread	248	2.522
domestic eggs	295	2.999
frankfurter	202	2.054
pork	218	2.217
butter	271	2.755
curd	257	2.613

frequent itemsets*

File Edit View Widget Options Help

File Data Table Frequent Itemsets

Data Table

Info
 9933 instances
 169 features (tags, density 2.61%)
 No target variable.
 No meta attributes

Variables
☒ Show variable labels (if present)
☐ Visualize continuous values
☒ Color by instance classes

Selection
☐ Select full rows

4	yogurt, pip fruit, cream cheese, meat spreads
5	whole milk, other vegetables, condensed milk, long life bakery product
6	yogurt, whole milk, butter, rice, abrasive cleaner
7	rolls/buns
8	other vegetables, rolls/buns, UHT-milk, bottled beer, liquor (appetizer)
9	pot plants
10	whole milk, cereals
11	tropical fruit, other vegetables, white bread, bottled water, chocolate
12	citrus fruit, tropical fruit, yogurt, whole milk, butter, bottled water, curd, flour, dis...
13	beef
	rolls/buns, frankfurter, soda
	tropical fruit, chicken
	butter, sugar, fruit/vegetable juice, newspapers
	fruit/vegetable juice
	packaged fruit/vegetables
	chocolate

Frequent Itemsets

Info
 Number of itemsets: 122
 Selected itemsets: 2
 Selected examples: 3680

Expand all Collapse all

Find itemsets

Minimal support: 2%
 Max. number of itemsets: 10000

☒ Auto find itemsets is on

Filter itemsets

Contains: soda
 Min. items: 1 Max. items: 5

☒ Apply these filters in search

☐ Send selection

Itemsets	Support	%
whole milk	2513	25.55
soda	394	4.006
other vegetables	1903	19.35
soda	322	3.274
rolls/buns	1809	18.39
soda	377	3.833
soda	1715	17.44
shopping bags	242	2.461
sausage	239	2.43
pastry	297	2.105
yogurt	1372	13.95
soda	269	2.735
bottled water	1087	11.05
soda	285	2.898
tropical fruit	1082	10.49
soda	203	2.084
whole milk	2513	25.55
soda	394	4.006
other vegetables	1903	19.35
soda	322	3.274
rolls/buns	1809	18.39

POST-VIVA QUESTIONS:

1. Which rules validated your hypothesis?
2. What if support was too low—how did it impact the output?
3. How could false positives be minimized?
4. Suggest a real-world cybersecurity hypothesis.
5. How does rule generalization affect performance?

INNOVATIVE APPROACHES:

1. Dynamic Rule Hypothesizing: Use a rule template bank and filter based on frequency.
2. Python Automation: Validate rules with mlxtend in a Python widget.
3. Interactive Threshold Testing: Create sliders for support and confidence.

RESULT:

Using Orange for association rule mining involves setting appropriate support and confidence thresholds to discover meaningful patterns in your data. By adjusting these thresholds and examining the generated rules, you can derive valuable insights and strong association rules from your dataset.

AIM:

To construct haarwavelet transformation for numerical data, construct principal component analysis (pca) for 5-dimensional data.

TOOLS REQUIRED:

1. Orange Data Mining Software
2. Python Script for advanced filtering

PRE VIVA QUESTION:

1. What is the purpose of data transformation in data analysis?
2. Define Haar Wavelet Transformation. How is it useful in data compression?
3. What is Principal Component Analysis (PCA)?
4. Why is PCA preferred for high-dimensional data?
5. How do you determine the number of principal components to retain in PCA?

PROCEDURE:

a) Implementing the Haar wavelet transformation for numerical data involves the following steps:

- **Divide the data into pairs.**
- **Calculate the averages and differences** for each pair.
- **Repeat the process** on the averages until you are left with a single value (this process can be stopped earlier if needed).
- Step-by-Step Haar Wavelet Transformation
- **Divide Data into Pairs:**

Suppose you have a list of data points $[x_1, x_2, x_3, x_4, \dots, x_n]$ where n is a power of 2. If n is not a power of 2, zero-padding can be applied to the data.

- **Calculate Averages and Differences:**
- For each pair (x_{2i-1}, x_{2i}) , calculate:
 - Average: $a_i = \frac{x_{2i-1} + x_{2i}}{2}$
 - Difference: $d_i = \frac{x_{2i-1} - x_{2i}}{2}$
- The averages will form a new sequence of length $n/2$, and the differences will form another sequence of length $n/2$.
- **Repeat the Process:**
- Apply the same process to the new sequence of averages until only one value remains.

import numpy as np

```
def haar_wavelet_transform(data):
    n = len(data)
    output = np.copy(data)
    step = 1
```

```

while step < n:
    for i in range(0, n, 2 * step):
        for j in range(step):
            avg = (output[i + j] + output[i + j + step]) / 2
            diff = (output[i + j] - output[i + j + step]) / 2
            output[i + j] = avg
            output[i + j + step] = diff
        step *= 2

return output

```

```

# Example usage
data = [4, 6, 10, 12, 14, 16, 18, 20]
transformed_data = haar_wavelet_transform(data)
print("Transformed Data:", transformed_data)

```

- output:
Transformed Data: [12 -1 -2 0 -4 0 0 0]

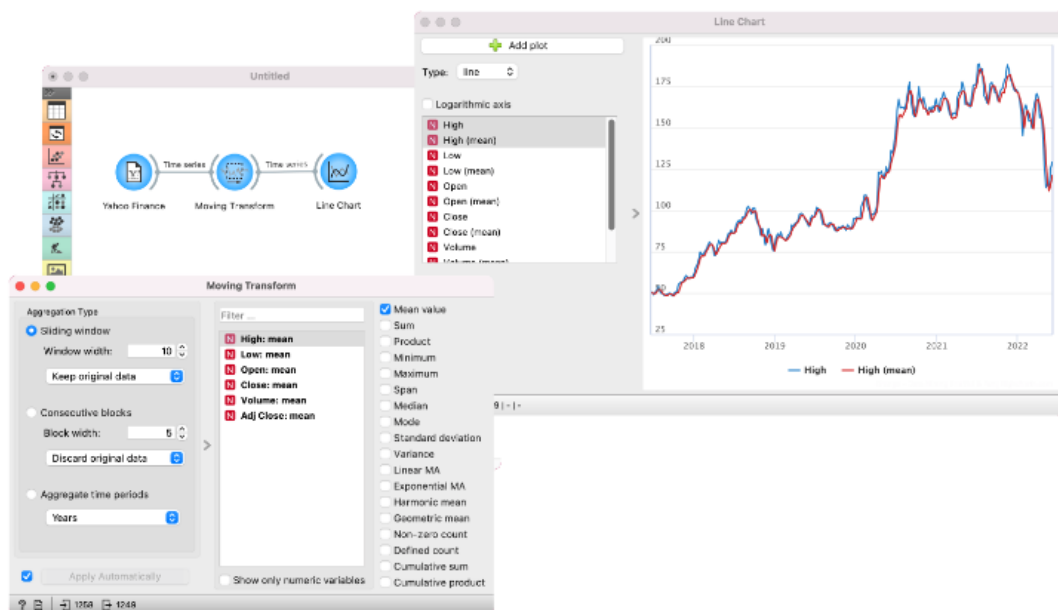
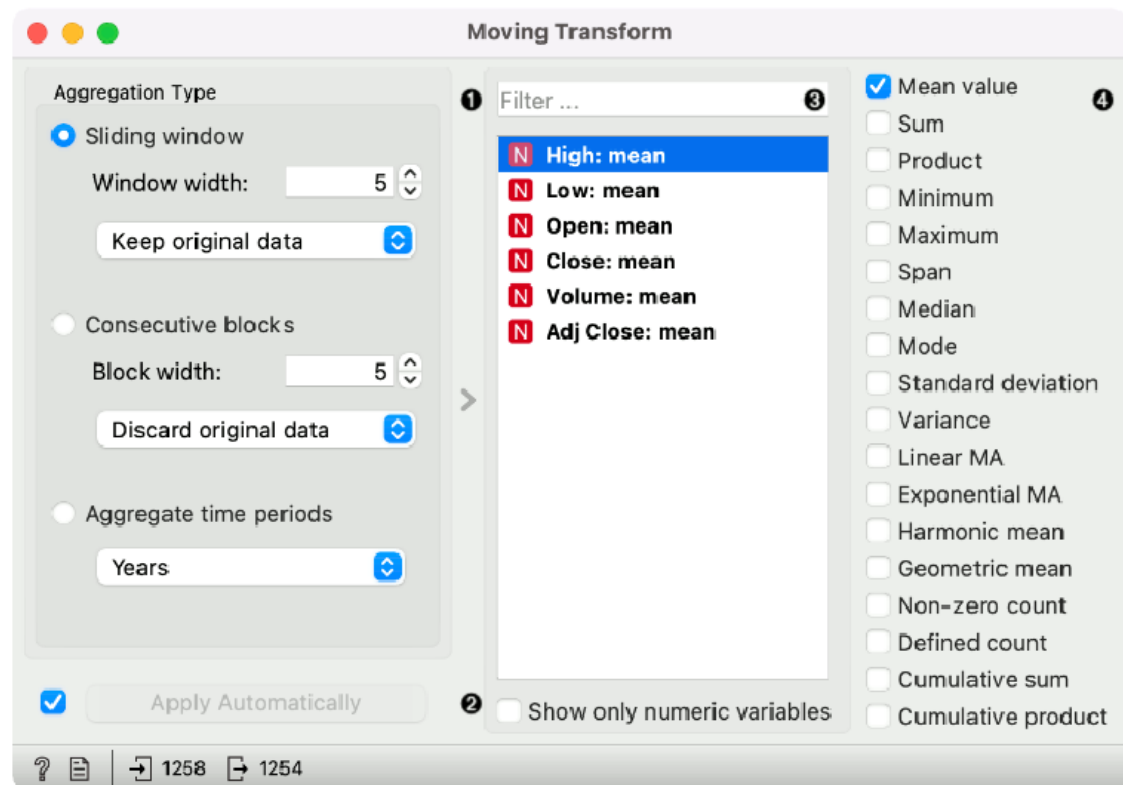
Constructing a Principal Component Analysis (PCA) for 5-dimensional data using the Orange tool can be accomplished through the following steps:

- **Install Orange:** Ensure you have Orange installed. If not, you can install it via pip:
- **Load Orange:**
- Open the Orange application.
- **Import Data:**
- Drag and drop the '**File**' widget to the canvas.
- Double-click the '**File**' widget and load your dataset (ensure it has 5 dimensions).
- **Data Table:**
- Connect the '**File**' widget to the '**Data Table**' widget to inspect your data and ensure it is loaded correctly.
- **PCA:**
- Drag and drop the '**PCA**' widget to the canvas.
- Connect the '**File**' widget to the '**PCA**' widget.
- **PCA Settings:**
- Double-click the '**PCA**' widget to configure it.
- You can choose how many components you want to retain. For visualization purposes, retaining 2 or 3 components is often useful.
- Click '**Apply**' to perform PCA.
- **Visualize:**
- To visualize the results, you can connect the '**PCA**' widget to the '**Scatter Plot**' widget.
- Double-click the '**Scatter Plot**' widget to configure and visualize your data in the new principal component space.

b) Principal Component Analysis (PCA) in Orange

- **Open Orange:** Launch the Orange application.
- **Load Your Data:**
- Drag and drop the '**File**' widget to the canvas.
- Load your dataset.
- **PCA Widget:**

- Drag and drop the '**PCA**' widget to the canvas.
- Connect the '**File**' widget to the '**PCA**' widget.
- Double-click the '**PCA**' widget to configure it.
- **Configure PCA:**
- In the PCA widget, you can specify the number of components you want to retain. For 5-dimensional data, you might want to reduce it to 2 or 3 components for visualization purposes.
- Apply the transformation.
- **Visualize PCA Results:**
- Connect the '**PCA**' widget to the '**Scatter Plot**' widget.
- Double-click the '**Scatter Plot**' widget to configure and visualize the PCA results.



POST-VIVA QUESTIONS

1. How many principal components were required for 95% variance?
2. What insights did the transformed data reveal?
3. What limitations did you observe in the Haar method?
4. How would you handle non-numerical features?
5. Can transformation alter semantic meaning?

INNOVATIVE APPROACHES

1. Hybrid Workflow: Use Colab for transformation and Orange for visualization.
2. Compression Analysis: Evaluate compression loss post-Haar transform.
3. Variance Plotting: Use PCA to visualize explained variance ratio.

RESULT:

By following these steps, you can effectively Construct Haarwavelet transformation for numerical data and principal component analysis (PCA) for 5-dimensional data

AIM:

To implement binning visualizations for any real time dataset, implement linear regression techniques.

TOOLS AND SOFTWARE USED:

1. Orange Data Mining Software
2. Dataset: Boston Housing dataset (or any real-time data)
3. Widgets Used: File, Discretize, Box Plot, Linear Regression, Scatter Plot

PRE-VIVA QUESTIONS:

1. Why is binning necessary in data visualization?
2. What does linear regression model?
3. What is the meaning of the R^2 value?
4. How are residuals interpreted in regression?
5. What is the difference between classification and regression?

PROCEDURE:

Binning is a data preprocessing step where continuous data is divided into intervals or "bins." This can be useful for visualizing data distributions or simplifying models. In Orange, you can implement binning visualizations with the following steps:

Steps to Implement Binning Visualization in Orange

1. **Load Data:**
 - Use the 'File' widget to load your dataset.
2. **Binning:**
 - Use the 'Continuize' widget to bin continuous data.
 - Alternatively, use the 'Discretize' widget for more control over the binning process.
3. **Visualize Binned Data:**
 - Use the 'Box Plot', 'Histogram', or 'Distributions' widget to visualize the binned data.

Example Workflow

1. **Load Data:**
 - Drag the 'File' widget to the canvas.
 - Select your dataset file (e.g., `real_time_data.csv`).
2. **Binning Data:**
 - Drag the 'Discretize' widget to the canvas.
 - Connect the 'File' widget to the 'Discretize' widget.
 - Configure the 'Discretize' widget to specify the number of bins or the binning method (e.g., equal-width, equal-frequency).
3. **Visualize Binned Data:**
 - Drag the 'Histogram' widget to the canvas.
 - Connect the 'Discretize' widget to the 'Histogram' widget.
 - Configure the 'Histogram' widget to select the binned variable and visualize its

distribution.

Visualizing Binned Data with Box Plot

1. **Load Data:**
 - Follow the steps to load the data using the 'File' widget.
2. **Binning Data:**
 - Follow the steps to bin the data using the 'Discretize' widget.
3. **Box Plot Visualization:**
 - Drag the 'Box Plot' widget to the canvas.
 - Connect the 'Discretize' widget to the 'Box Plot' widget.
 - Configure the 'Box Plot' widget to visualize the distribution of the binned variable across different categories or bins.

Implementing Linear Regression in Orange

Linear regression is a statistical method for modeling the relationship between a dependent variable and one or more independent variables. In Orange, you can implement linear regression as follows:

Steps to Implement Linear Regression in Orange

1. **Load Data:**
 - Use the 'File' widget to load your dataset.
2. **Data Preprocessing:**
 - Use the 'Select Columns' widget to select the dependent (target) and independent (predictor) variables.
 - Optionally, use the 'Preprocess' widget to handle missing values, normalize data, or encode categorical variables.
3. **Linear Regression:**
 - Use the 'Linear Regression' widget to perform linear regression analysis.
 - Connect the 'Select Columns' widget to the 'Linear Regression' widget.
4. **Evaluate Model:**
 - Use the 'Test & Score' widget to evaluate the linear regression model.
 - Connect the 'Linear Regression' widget to the 'Test & Score' widget.
 - Connect the 'File' widget (with the data split if necessary) to the 'Test & Score' widget.
5. **Inspect Results:**
 - Use the 'Scatter Plot' widget to visualize the regression line.
 - Connect the 'Linear Regression' widget to the 'Scatter Plot' widget.
 - Configure the 'Scatter Plot' widget to display the regression line and the data points.

Example Workflow

1. **Load Data:**
 - Drag the 'File' widget to the canvas.
 - Select your dataset file (e.g., `real_time_data.csv`).
2. **Select Columns:**
 - Drag the 'Select Columns' widget to the canvas.
 - Connect the 'File' widget to the 'Select Columns' widget.
 - Select the dependent variable (e.g., `y`) and independent variables (e.g., `x1`, `x2`).
3. **Linear Regression:**
 - Drag the 'Linear Regression' widget to the canvas.
 - Connect the 'Select Columns' widget to the 'Linear Regression' widget.
4. **Evaluate Model:**
 - Drag the 'Test & Score' widget to the canvas.
 - Connect the 'Linear Regression' widget to the 'Test & Score' widget.
 - Connect the 'File' widget to the 'Test & Score' widget for evaluation.
5. **Visualize Results:**
 - Drag the 'Scatter Plot' widget to the canvas.

- Connect the 'Linear Regression' widget to the 'Scatter Plot' widget.
- Configure the 'Scatter Plot' widget to visualize the regression line and the data points.



	name	coef
6	NOX	-17.7666
9	DIS	-1.47557
12	PTRATIO	-0.952747
14	LSTAT	-0.524758
2	CRIM	-0.108011
11	TAX	-0.0123346
8	AGE	0.000692225
13	B	0.00931168
4	INDUS	0.0205586
3	ZN	0.0464205
10	RAD	0.306049
5	CHAS	2.68673
7	RM	3.80987
1	intercept	36.4595

Linear Regression

Name: Linear Regression

Regularization:

- ☒ No regularization
- ☐ Ridge regression (L2)
- ☐ Lasso regression (L1)
- ☐ Elastic net regression

Regularization strength: Alpha: 0.0001

Elastic net mixing: L1 0.50 : 0.50 L2

Report ☒ Apply Automatically

Test & Score

Sampling:

- ☒ Cross validation
 - Number of folds: 10
 - ☒ Stratified
 - ☐ Random sampling
 - Repeat train/test: 10
 - Training set size: 66 %
 - ☒ Stratified
 - ☐ Leave one out
 - ☐ Test on train data
 - ☐ Test on test data

Report

Evaluation Results

Method	MSE	RMSE	MAE	R2
Linear Regression	23.370	4.834	3.376	0.723
Random Forest	11.313	3.364	2.317	0.866

POST-VIVA QUESTIONS

1. Did binning affect the visual clarity of data?
2. What was the regression accuracy?
3. Which feature most influenced the output?
4. What outliers were observed?
5. How can regression be applied in cybersecurity?

INNOVATIVE APPROACHES

- Real-Time Data Feeds: Import up-to-date CSV from APIs.
- Interactive Regression Tuning: Modify inputs and observe impact on prediction.
- Bin Width Optimization: Test fixed vs equal-frequency binning visually.

RESULT:

By following these steps in Orange, you can effectively implement binning visualizations and linear regression analysis. These workflows allow you to preprocess your data, perform exploratory analysis, and build predictive models, ensuring a comprehensive data analysis process.

AIM:

To visualize the clusters for any synthetic dataset, implement the program for converting the clusters into histograms

TOOLS AND SOFTWARE USED:

1. Orange Data Mining Software
2. Synthetic or real-world dataset (e.g., Iris)
3. Widgets Used: File, K-Means, Scatter Plot, Distributions, Data Table

PRE-VIVA QUESTIONS:

1. What is the significance of visualizing clusters?
2. How do histograms represent the density of clusters?
3. What types of datasets are suitable for clustering?
4. What does intra-cluster similarity imply?
5. How does the Distributions widget enhance interpretation?

PROCEDURE:

The histogram works by organizing and visualizing the distribution of data into intervals or bins along a continuous scale.

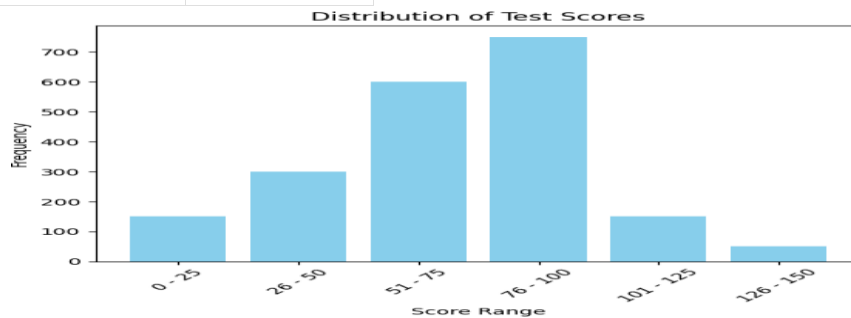
- The range of data values is divided into intervals called “bins.” The number of bins and their widths can be predefined or determined algorithmically based on the range and distribution of the data.
- Each data point in the dataset is assigned to a corresponding bin based on its value. As data points are assigned to bins, the frequency or count of data points falling within each bin is calculated.
- The histogram is constructed by plotting the bins along the x-axis and the frequencies (or densities) along the y-axis. Each bin is represented by a bar, and the height of the bar corresponds to the frequency of data points in that bin.

By examining the histogram, you can gain insights into the distribution of the data. You can identify patterns, trends, central tendencies, variability, outliers, and other characteristics of the dataset. For example, a **symmetric bell-shaped histogram** suggests a , while skewed histograms indicate asymmetry in the data.

Suppose you’re analyzing the distribution of scores on a standardized test. You have data for 2000 students, and you want to visualize how many students scored within different score ranges. For this you can create a histogram using the following data.

Score Range	Frequency
0-25	150
26-50	300

51-75	600
76-100	750
101-125	150
126-150	50



Histogram

The histogram show that the data is normally distributed, and the students have mostly score between 76-100. This histogram displays the frequency of students falling within different score ranges on the standardized test. Each bar represents a score range, and the height of the bar represents the frequency of students in that range. By customizing the x-axis intervals and the labels, you can effectively visualize the distribution of test scores. Additionally, you can further customize the histogram by changing the y-axis to display percentages or density if needed.

Histogram and its Interpretation

Normal Histogram

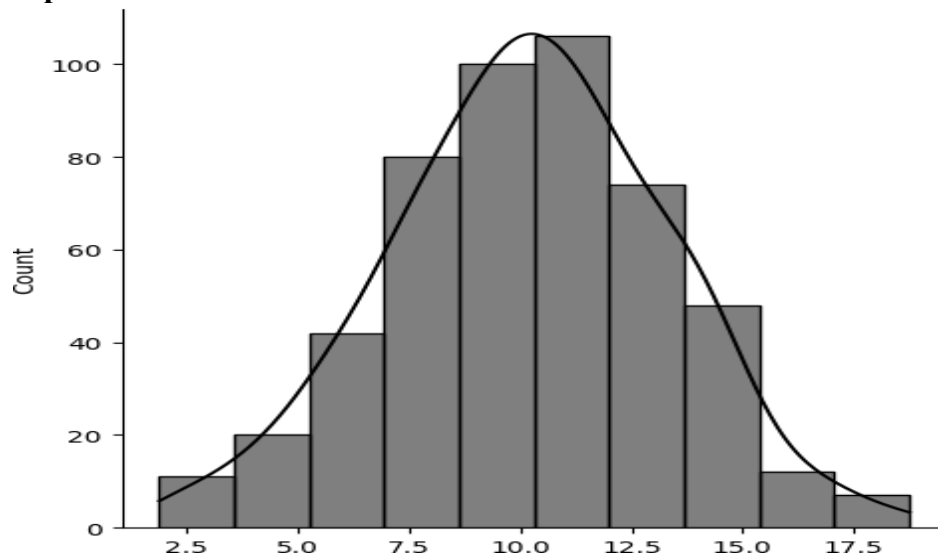
Normal histogram is a classical bell-shaped histogram with most of the frequency counts focused on the middle with diminishing tails and there is symmetry with respect to the median. Since the normal distribution is most commonly observed in real-world scenarios, you are most likely to find these. In Normally distributed histogram mean is almost equal to median.

- Python3

```
import numpy as np
import matplotlib.pyplot as plt
import seaborn as sns

# Normal histogram plot
data = np.random.normal(10.0, 3, 500)
sns.displot(data, kde=True, bins=10, color='black')
```

Output:



Normal Distribution Graph

We have plotted a normal distribution graph.

- The peak of the curve represents the mean of the dataset.
- The normal distribution graph is symmetric.

Non-normal Short-tailed/ long-tailed histogram

In short-tailed distribution tail approaches 0 very fast, as we move from the median of data, In the long-tailed histogram, the tail approaches 0 slowly as we move far from the . Here, we refer tail as the extreme regions in the histogram where most of the data is not concentrated and this is on both sides of the peak.

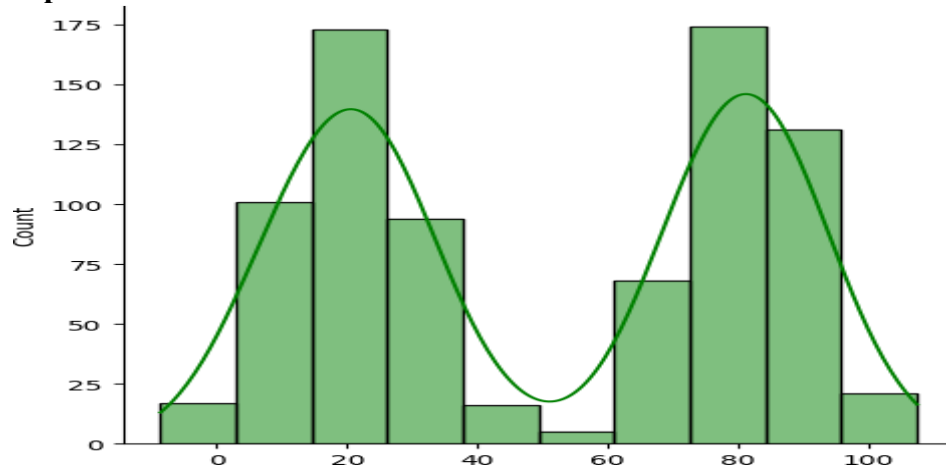
Bimodal Histogram

Aof data represents the most common values in the histogram (i.e. peak of the histogram. A bimodal histogram represents that there are two peaks in the histogram. The histogram can be used to test the unimodality of data. The bimodality (or for instance non-unimodality) in the dataset represents that there is something wrong with the process. Bimodal histogram many one or both of two characters: Bimodal normal distribution and symmetric distribution.

- Python3

```
# Bi-modal histogram
N=400
mu_1, sigma_1 = 80, 10
mu_2, sigma_2 = 20, 10
# Generate two normal distributions of given mean sband concatenate
X_1 = np.random.normal(mu_1, sigma_1, N)
X_2 = np.random.normal(mu_2, sigma_2, N)
X = np.concatenate([X_1, X_2])
sns.displot(X, bins=10, kde=True, color='green')
```


Output:



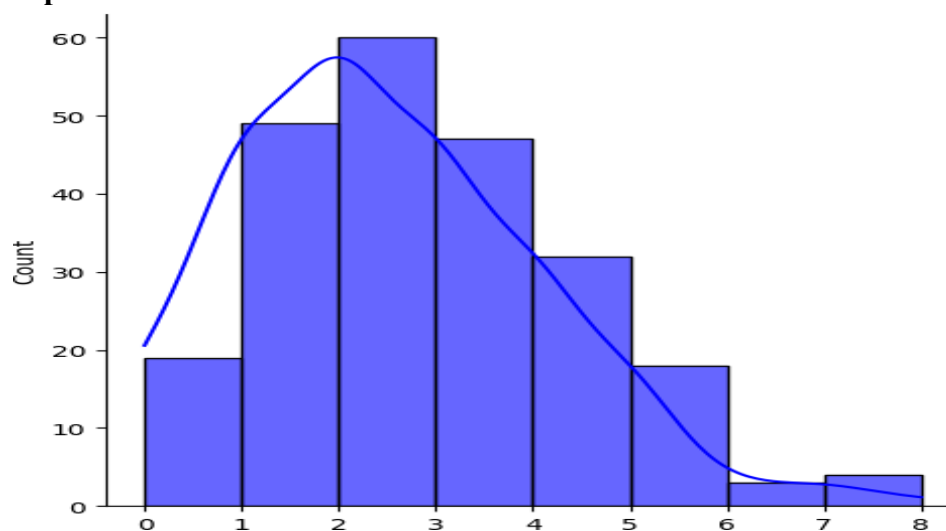
Skewed Left/Right Histogram

is those where the one-side tail is quite clearly longer than the other-side tail. A right-skewed histogram means that the right-sided tail of the peak is more stretched than its left and vice-versa for the left-sided. In a left-skewed histogram, the mean is always lesser than the median, while in a right-skewed histogram mean is greater than the histogram.

Right-skewed Histogram

```
# Right-skewed Histogram
rdata = [0] * 19 + [1]*49 + [2]*60 + [3] * \
    47 + [4]*32 + [5] * 18 + [6]*3 + [7]*3 + [8]
sns.displot(rdata, bins=8, kde=True, alpha=0.6, color='blue')
```

Output:

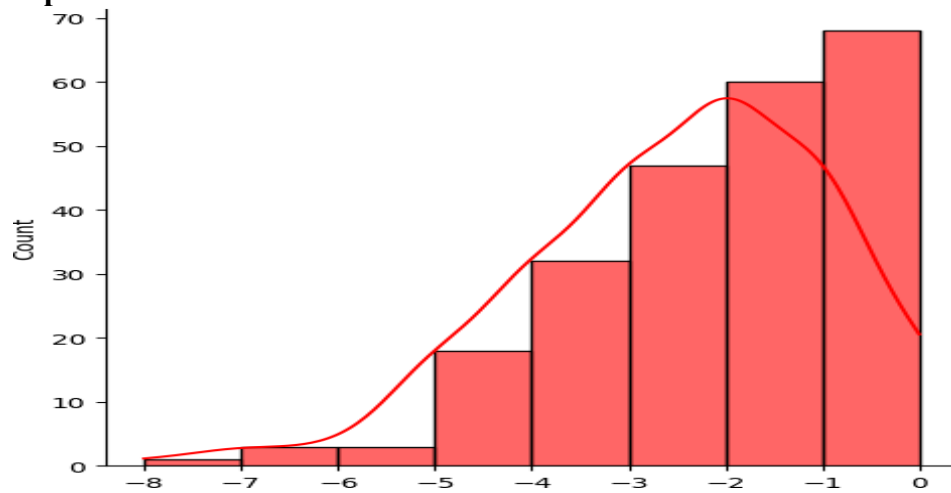


Right Skewed Histogram
Left-skewed Histogram

- Python3

```
# Left-skewed Histogram
ldata = [0]*19 + [-1]*49 + [-2]*60 + [-3]*47 + [-4]*32 + [-5]*18 +
[-6]*3 + [-7]*3 + [-8]
sns.displot(ldata, kde = True, bins=8, alpha=0.6, color='red')
```

Output:



Left Skewed Histogram

Uniform Histogram

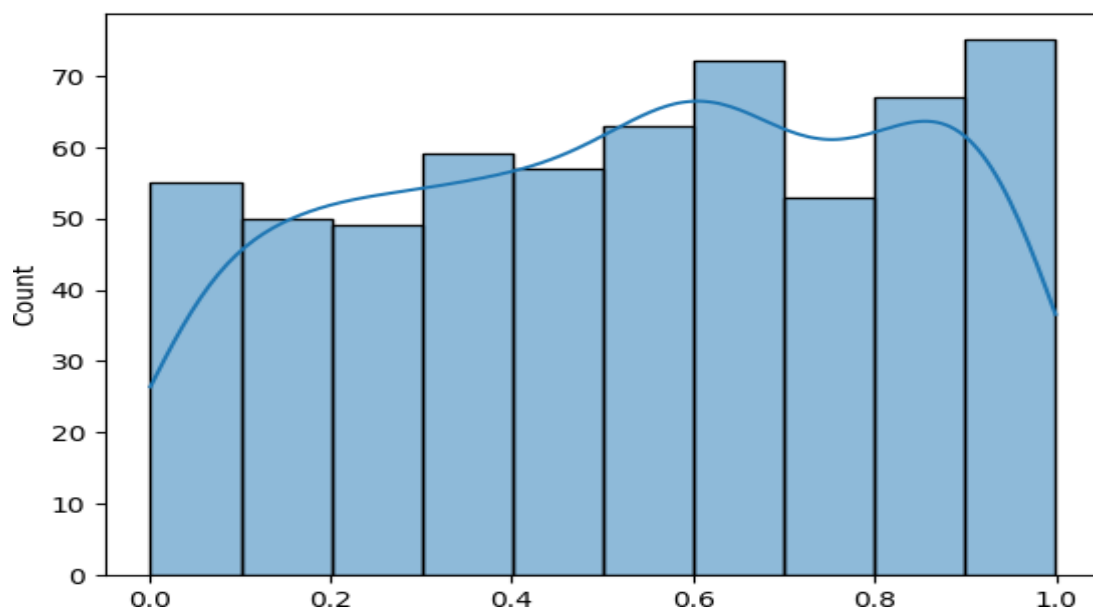
In uniform histogram, each bin contains approximately the same number of counts (frequency). The example of uniform histogram is such as a die is rolled n ($n \gg 30$) number of times and record the frequency of different outcomes.

- Python3

```
# Generate random data following a uniform distribution
data = np.random.uniform(low=0, high=1, size=600)
sns.histplot(data, kde =True, bins =10 )
plt.show()
```

Output:

Uniform Distribution



POST-VIVA QUESTIONS

1. Which attributes were most discriminative in the histogram?
2. How many clusters were formed and how evenly distributed were they?
3. What did the histogram reveal about cluster spread?
4. Can you identify noise or outliers from the plots?
5. What would change if a different number of clusters were chosen?

INNOVATIVE APPROACHES

1. Dynamic Histogram Generation: Binning clusters by attributes and visualizing their feature-wise spread.
2. Custom Cluster Coloring: Differentiate clusters using custom color mappings.
3. Cluster Comparison Matrix: Use Distributions widget to compare attribute-wise frequencies across clusters.

RESULT:

By following these steps, you can effectively converting the clusters into histograms

AIM:

To write a program to implement agglomerative clustering technique and to divisive hierarchical clustering technique

TOOLS AND SOFTWARE USED:

1. Orange Data Mining Software
2. Python Scripting (for divisive clustering using scipy)
3. Widgets Used: File, Hierarchical Clustering, Scatter Plot, Python Script

PRE-VIVA QUESTIONS:

1. What is the main difference between agglomerative and divisive clustering?
2. What is a dendrogram and how do you interpret it?
3. Which linkage methods are used in hierarchical clustering?
4. Why is hierarchical clustering useful for exploratory data analysis?
5. How do distance metrics impact clustering?

PROCEDURE:**Agglomerative Clustering in Orange**

Agglomerative clustering is a bottom-up approach where each observation starts in its own cluster, and pairs of clusters are merged as one moves up the hierarchy.

Steps to Implement Agglomerative Clustering

1. **Load Data:**
 - Use the 'File' widget to load your dataset.
2. **Agglomerative Clustering:**
 - Use the 'Hierarchical Clustering' widget to perform agglomerative clustering.
 - Visualize the clustering using the 'Dendrogram' widget.

Example Workflow

1. **Load Data:**
 - Drag the 'File' widget to the canvas.
 - Load your dataset file (e.g., `dataset.csv`).
2. **Agglomerative Clustering:**
 - Drag the 'Hierarchical Clustering' widget to the canvas.
 - Connect the 'File' widget to the 'Hierarchical Clustering' widget.
 - Configure the widget to use agglomerative clustering (default behavior).
3. **Visualize Clustering:**
 - Drag the 'Dendrogram' widget to the canvas.
 - Connect the 'Hierarchical Clustering' widget to the 'Dendrogram' widget.

Divisive Hierarchical Clustering in Orange

Divisive clustering is a top-down approach where all observations start in one cluster, and splits are performed recursively as one moves down the hierarchy. Orange does not provide a built-in widget for divisive clustering, but you can implement it using a Python script.

Python Script for Divisive Hierarchical Clustering

Here is a Python script to implement divisive hierarchical clustering using Orange's scripting capabilities:

```

import numpy as np
from sklearn.cluster import AgglomerativeClustering
from Orange.data import Table, Domain
from Orange.clustering import hierarchical
from Orange.widgets.utils.plot import OWScatterPlotBase
import matplotlib.pyplot as plt

# Load data
data = Table("your_dataset.csv")

# Convert Orange Table to numpy array
X = np.array([list(row) for row in data])

# Function to perform divisive hierarchical clustering
def divisive_clustering(X, n_clusters=2):
    # Initially, all points are in one cluster
    clusters = [X]

    while len(clusters) < n_clusters:
        # Find the cluster to split
        cluster_to_split = max(clusters, key=len)
        clusters.remove(cluster_to_split)

        # Perform agglomerative clustering on the chosen cluster
        clustering = AgglomerativeClustering(n_clusters=2)
        labels = clustering.fit_predict(cluster_to_split)

        # Split the chosen cluster into two sub-clusters
        sub_cluster_1 = cluster_to_split[labels == 0]
        sub_cluster_2 = cluster_to_split[labels == 1]

        # Add the sub-clusters to the list of clusters
        clusters.append(sub_cluster_1)
        clusters.append(sub_cluster_2)

    return clusters

# Perform divisive clustering
clusters = divisive_clustering(X, n_clusters=3)

# Visualize the clusters
colors = ['red', 'green', 'blue']
for cluster, color in zip(clusters, colors):
    plt.scatter(cluster[:, 0], cluster[:, 1], c=color)

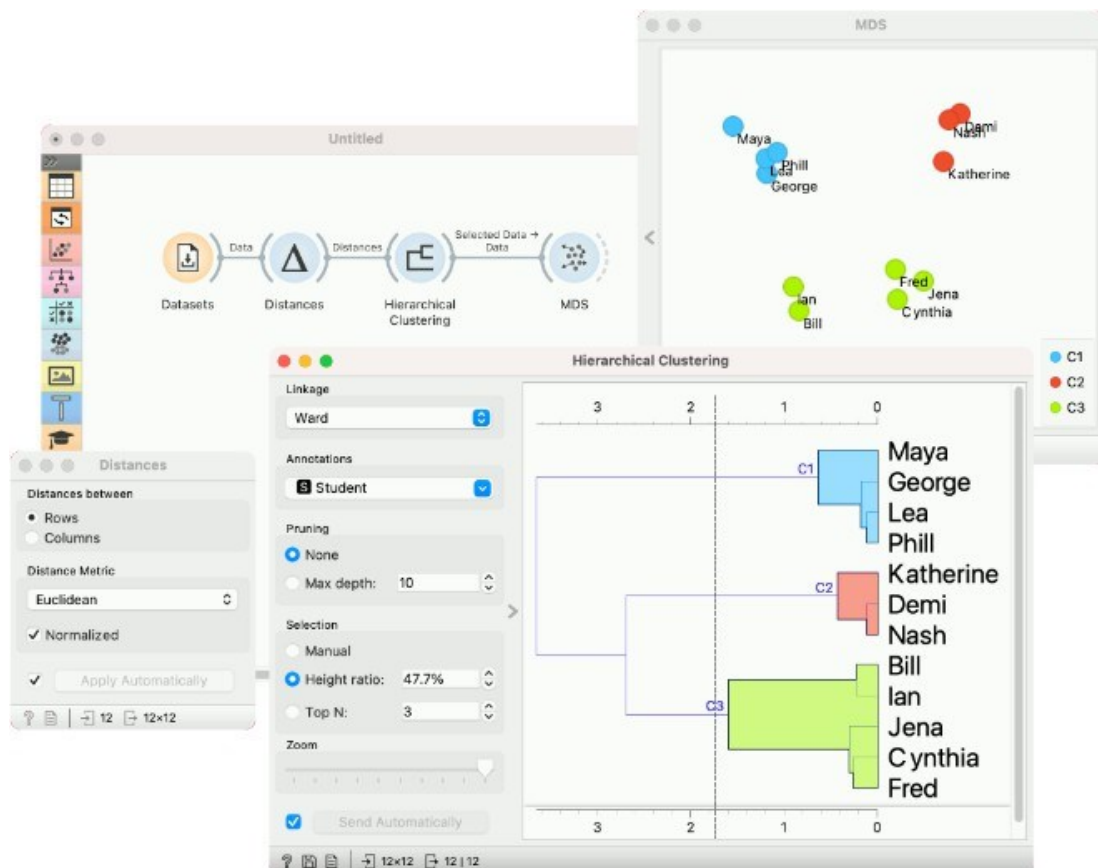
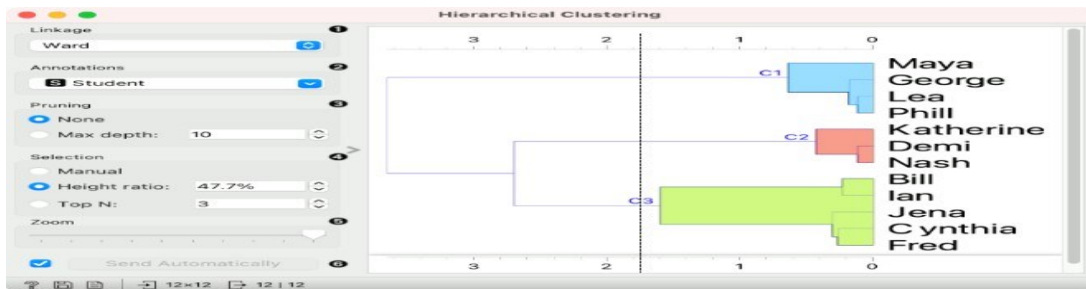
plt.xlabel('Feature 1')
plt.ylabel('Feature 2')

```

```
plt.title('Divisive Hierarchical Clustering')
plt.show()
```

How to Run the Script in Orange

1. **Load Data:**
 - Use the 'File' widget to load your dataset.
2. **Python Script:**
 - Drag the 'Python Script' widget to the canvas.
 - Connect the 'File' widget to the 'Python Script' widget.
 - Copy and paste the provided Python script into the 'Python Script' widget.
 - Ensure that the dataset file name in the script matches the file you loaded.
3. **Run Script:**
 - Execute the script to perform divisive hierarchical clustering and visualize the results.



POST-VIVA QUESTIONS

1. How many clusters did you infer from the dendrogram?
2. Which method—agglomerative or divisive—provided more meaningful clusters?
3. What challenges did you face while implementing divisive clustering?
4. How does hierarchical clustering compare to K-Means?
5. Suggest an application of this clustering in cybersecurity or IoT

INNOVATIVE APPROACHES:

1. Custom Dendrogram Plotting: Use Python to visualize dendrograms for divisive clustering.
2. Distance Metric Tuning: Try different metrics (Euclidean, Manhattan) and linkage methods (Ward, complete).
3. Threshold-Based Splitting: Define cut-off thresholds to extract clusters from dendrograms.

RESULT:

By following these steps, you can implement agglomerative and divisive hierarchical clustering techniques in Orange. Agglomerative clustering can be directly performed using the 'Hierarchical Clustering' and 'Dendrogram' widgets, while divisive clustering requires a custom Python script. These clustering techniques provide valuable insights into the structure and relationships within your data.

AIM:

To develop scalable clustering algorithms and a scalable Apriori algorithm in Orange, we'll focus on techniques and optimizations that can handle large datasets efficiently. While Orange's graphical interface is user-friendly for small to medium-sized datasets, scalable implementations often require custom scripting and leveraging optimized libraries.

TOOLS AND SOFTWARE USED

1. Orange Data Mining Software
2. Large-scale CSV datasets (synthetic or real-world)
3. Widgets Used: File, K-Means, Association Rules, Data Table

PRE-VIVA QUESTIONS

1. What is meant by scalability in data mining?
2. How does scalable K-Means differ from traditional K-Means?
3. What challenges arise in Apriori when datasets are large?
4. Define minimum support and its role in Apriori.
5. Why is clustering important for large datasets?

PROCEDURE:**Scalable Clustering Algorithms**

1. **MiniBatch K-Means:** This is a scalable version of the K-Means algorithm that processes mini-batches of the dataset to reduce computational load.
2. **DBSCAN with KD-Tree:** Using a KD-Tree for neighborhood searches can make DBSCAN more scalable.
3. **Hierarchical Clustering with Sparse Data Structures:** Optimizing hierarchical clustering using sparse data structures for large datasets.

Example: MiniBatch K-Means in Orange

Orange does not natively support MiniBatch K-Means via its GUI widgets, but you can use the `Python Script` widget to implement it.

```
import numpy as np
from sklearn.cluster import MiniBatchKMeans
from Orange.data import Table
```

```

# Load data
data = Table("your_large_dataset.csv")

# Convert Orange Table to numpy array
X = np.array([list(row) for row in data])

# Define MiniBatch K-Means
mbk = MiniBatchKMeans(init='k-means++', n_clusters=10, batch_size=100,
n_init=10, max_no_improvement=10, verbose=0)

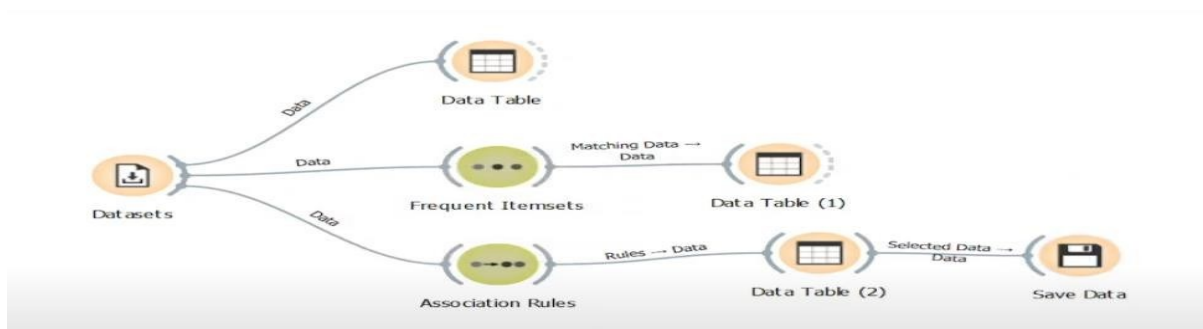
# Fit the model
mbk.fit(X)

# Get the cluster centers and labels
cluster_centers = mbk.cluster_centers_
labels = mbk.labels_

# Add the cluster labels back to the Orange Table
data[:, 'Cluster'] = labels

# Output the table with cluster labels
output_data = data

```



Scalable Apriori Algorithm

The Apriori algorithm can be made scalable by using data reduction techniques, transaction reduction, and parallel processing.

Example: Scalable Apriori in Orange

Orange does not provide a scalable version of Apriori directly. However, you can use a Python script with optimized libraries like `apyori` or `mlxtend`.

```

from mlxtend.frequent_patterns import apriori, association_rules
import pandas as pd
from Orange.data import Table, Domain, DiscreteVariable

# Load data
data = Table("your_large_dataset.csv")

# Convert Orange Table to pandas DataFrame
df = pd.DataFrame(data)

# Assume the dataset is in the transaction format with boolean values
# Apply Apriori
frequent_itemsets = apriori(df, min_support=0.01, use_colnames=True)

# Generate association rules
rules = association_rules(frequent_itemsets, metric="confidence", min_threshold=0.5)

# Convert results back to Orange Table
domain = Domain([DiscreteVariable.make("itemsets", values=[str(i) for i in
range(len(rules))])])
orange_table = Table(domain, np.array(rules))

# Output the association rules
output_data = orange_table

```

How to Run the Scripts in Orange

1. **Load Data:**
 - Use the 'File' widget to load your large dataset.
2. **Python Script:**
 - Drag the 'Python Script' widget to the canvas.
 - Connect the 'File' widget to the 'Python Script' widget.
 - Copy and paste the provided Python scripts into the 'Python Script' widget.
 - Ensure that the dataset file name in the script matches the file you loaded.

*** Association Rules - Orange

Info
Rules: 99 (shown 99)

Find association rules

Min. supp.: 1 %

Min. conf.: 12 %

Max. rules: 10k

☐ Induce only classification rules

☒ Restrict search by below filters

Find Rules

Filter by Antecedent

Contains:

Items, min: 1 max: 999

Filter by Consequent

Contains:

Items, min: 1 max: 999

☒ Send selection

Supp	Conf	Covr	Strg	Lift	Levr	Antecedent	Consequent
0.050	0.287	0.175	1.619	1.017	0.001	Fresh Fruit	Fresh Vegetables
0.050	0.178	0.283	0.618	1.017	0.001	Fresh Vegetables	Fresh Fruit
0.035	0.124	0.283	0.413	1.059	0.002	Fresh Vegetables	Dried Fruit
0.035	0.299	0.117	2.421	1.059	0.002	Dried Fruit	Fresh Vegetables
0.035	0.123	0.283	0.421	1.035	0.001	Fresh Vegetables	Soup
0.035	0.293	0.119	2.375	1.035	0.001	Soup	Fresh Vegetables
0.031	0.262	0.118	2.405	0.926	-0.002	Cheese	Fresh Vegetables
0.028	0.279	0.099	2.854	0.987	-0.000	STORE_ID_13	Fresh Vegetables
0.027	0.260	0.105	2.691	0.921	-0.002	Cookies	Fresh Vegetables
0.025	0.278	0.089	3.160	0.982	-0.000	STORE_ID_17	Fresh Vegetables
0.022	0.284	0.079	3.577	1.004	0.000	Paper Wipes	Fresh Vegetables
0.022	0.278	0.078	3.625	0.985	-0.000	Canned Vegetables	Fresh Vegetables
0.021	0.175	0.118	1.486	1.000	-0.000	Cheese	Fresh Fruit
0.020	0.171	0.119	1.467	0.982	-0.000	Soup	Fresh Fruit
0.020	0.253	0.080	3.523	0.894	-0.002	Wine	Fresh Vegetables
0.019	0.291	0.067	4.223	1.030	0.001	Nuts	Fresh Vegetables
0.019	0.282	0.068	4.136	0.999	-0.000	Frozen Vegetables	Fresh Vegetables
0.019	0.284	0.067	4.222	1.006	0.000	Chocolate Candy	Fresh Vegetables
0.019	0.290	0.065	4.341	1.026	0.000	STORE_ID_11	Fresh Vegetables
0.019	0.286	0.066	4.315	1.013	0.000	Preserves	Fresh Vegetables
0.018	0.175	0.105	1.663	1.005	0.000	Cookies	Fresh Fruit
0.018	0.156	0.117	1.496	0.895	-0.002	Dried Fruit	Fresh Fruit
0.018	0.277	0.066	4.314	0.981	-0.000	STORE_ID_15	Fresh Vegetables
0.018	0.279	0.065	4.366	0.988	-0.000	Chips	Fresh Vegetables

POST-VIVA QUESTIONS

1. What was the performance impact of low support in Apriori?
2. How were clusters affected as dataset size increased?
3. Which method scaled better—clustering or Apriori—and why?
4. Did any meaningful rules emerge at very low support values?
5. Suggest optimizations to improve scalability in Orange or via scripting.

INNOVATIVE APPROACHES

1. Apriori Parameter Tuning: Explore the effect of varying support from 0.00001 to 0.1 on performance and rule count.
2. Mini-Batch K-Means: Though Orange lacks this natively, discuss extending it via Python for real-world scalability.
3. Streaming Simulation: Simulate real-time ingestion by updating datasets iteratively.

RESULT:

Scalability in data mining involves using optimized algorithms and data structures to handle large datasets efficiently. In Orange, you can extend its capabilities by using custom Python scripts and leveraging powerful libraries such as `scikit-learn` and `mlxtend`. This approach allows you to implement scalable clustering algorithms and the Apriori algorithm, making it possible to work with large datasets effectively.