Ex No: 01

Date:

DATA PRE-PROCESSING AND DATA CUBE

AIM:

Implement Data Preprocessing methods on Student and Labor Dataset implement data cube for data warehouse on 3-dimensional data.

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.

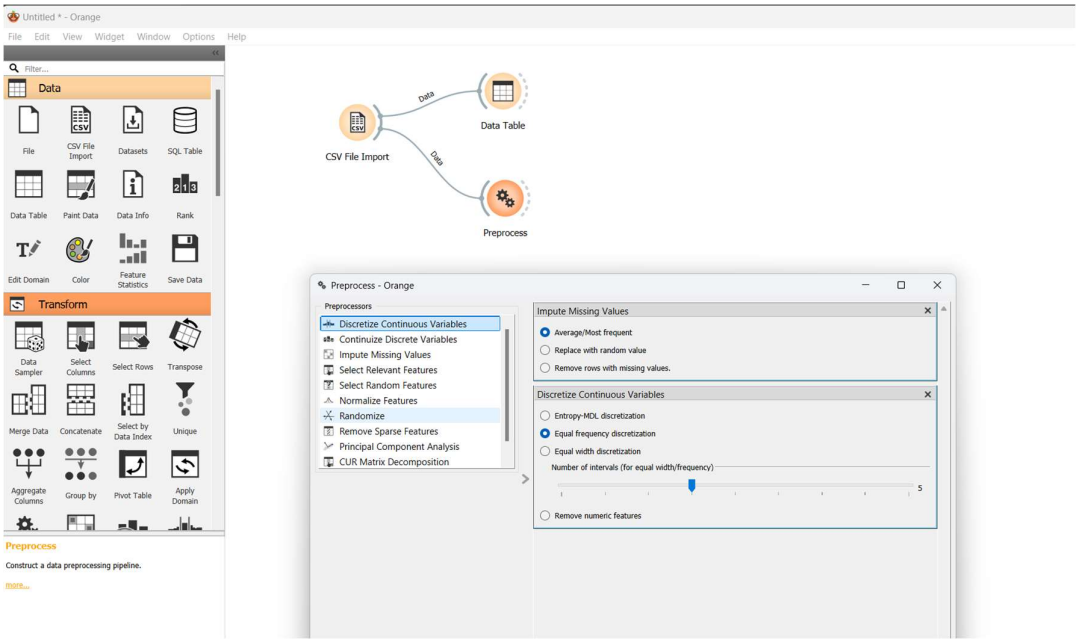
PROBLEM DEFINITION:

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.

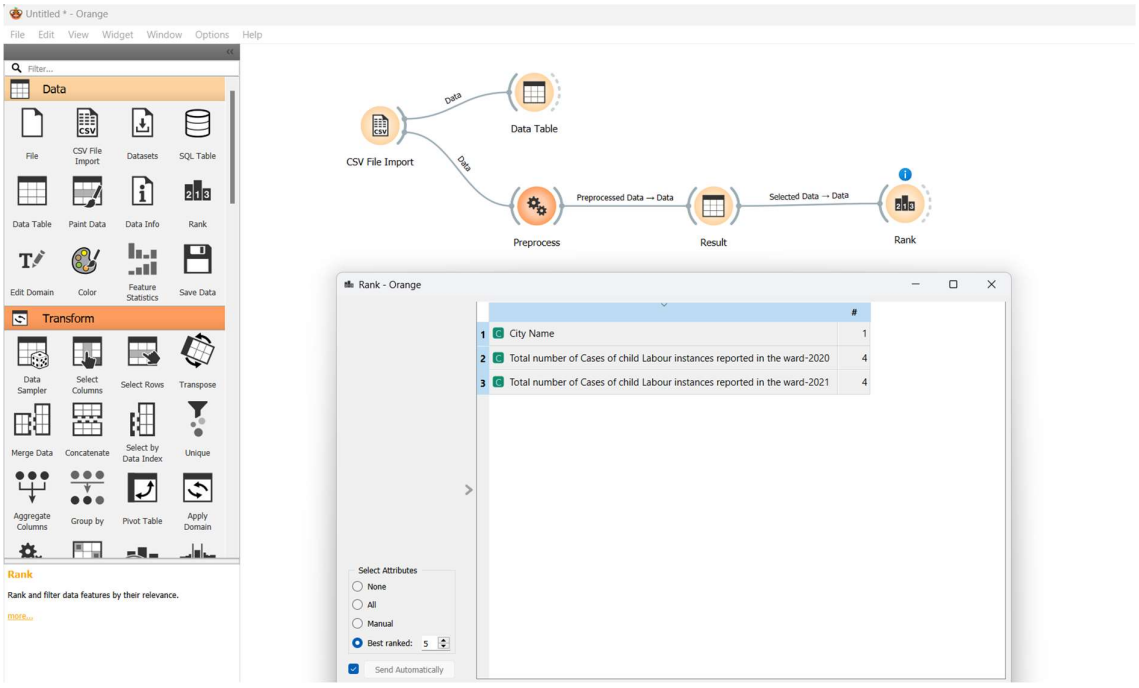
PROCEDURE:

1. Download and install Orange software from https://orangedatamining.com/.
2. Download Labor and Student Dataset from Internet.
3. Open Orange software and click on "New". Then, click on the "File" widget.
4. Double-click on the "File" widget to import your dataset. Next, add a "Data Table" widget and connect it with the "File" widget to display the raw data.

5. Right-click on the canvas, search for the "Preprocess" widget, and connect it with the "File" widget. Double-click on the "Preprocess" widget and choose options like "Impute Missing Values" and "Discretize Continuous Variables" to modify the raw data as needed.



6. Add another "Data Table" widget, rename it as "Result", and connect it with the "Preprocess" widget. This new table will display the preprocessed data (modified data).
7. Finally, add a "Rank" widget and connect it with the "Result" widget. Double-click on the "Rank" widget to view the classification of the dataset.



RESULT:

The implementation of Data Preprocessing methods on Student and Labor Dataset implement data cube for data warehouse on 3-dimensional data is completed and verified successfully.

Ex No: 02

Date:

DATA CLEANING

AIM:

Implement various missing handling mechanisms, Implement various noisy handling mechanisms.

DESCRIPTION:

Data Cleaning:

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

Dataset:

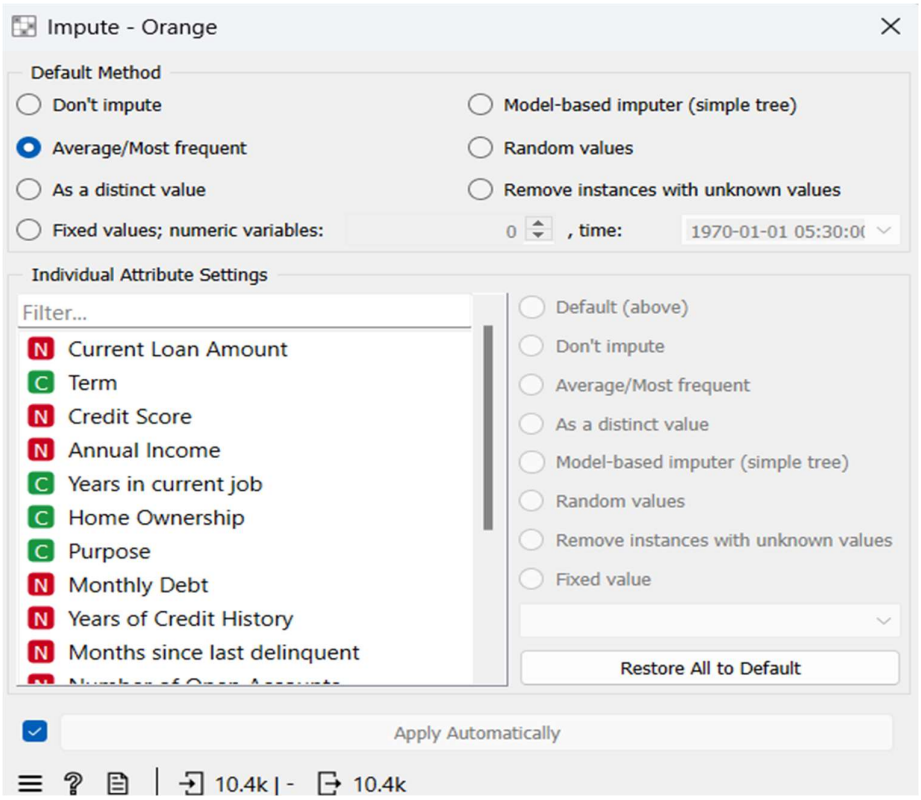
Datasets involve a large amount of data points grouped into one table.

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.

PROCEDURE:

1. Download and install Orange software from https://orangedatamining.com/.
2. Download Loan Dataset from Internet.
3. Open Orange software and click on "New". Then, click on the "File" widget.
4. Double-click on the "File" widget to import your dataset. Next, add a "Data Table" widget and connect it with the "File" widget to display the raw data.
5. Search for "Feature Statistics" in the search bar and Click on it and connect it with “File”. (The Feature Statistics widget provides a quick way to inspect and find interesting features in a given data set).
6. Search for "Impute" in Search bar and connect it with “File” (Impute – Replaces unknown value in data). Double click on Impute and Select the following option,



7. Connect the "Data Table" and "Feature Statistics" widgets to the "Impute" widget.
8. Search for "Save Data" in the search bar and connect it with "Save Data" widget.

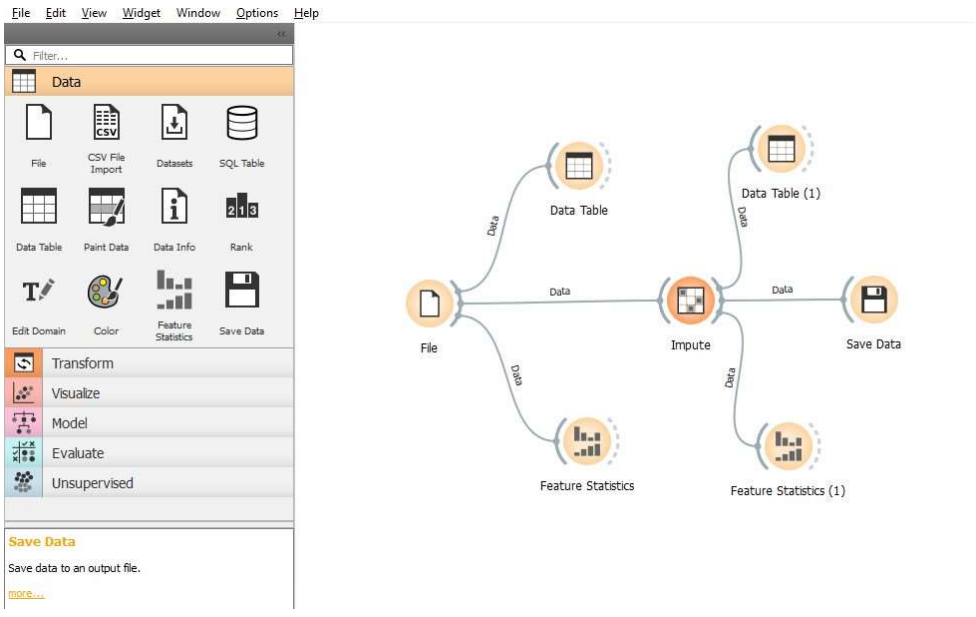
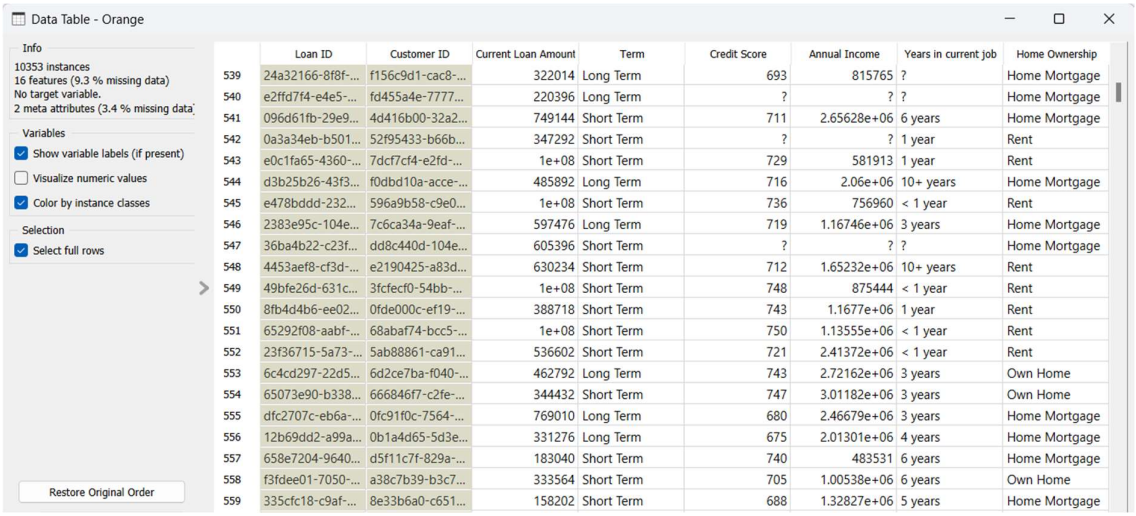
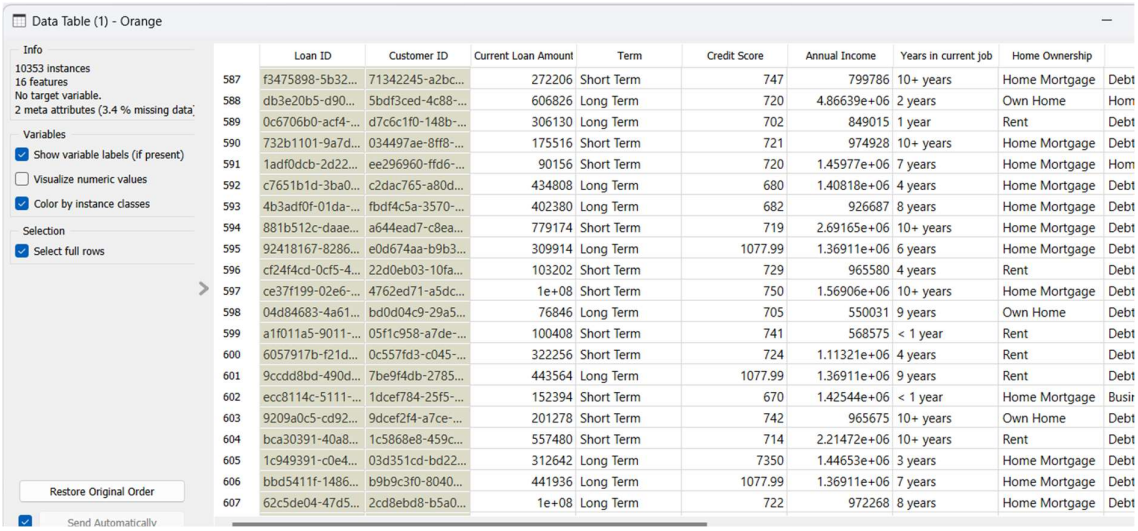


Table with Missing Values:



	Loan ID	Customer ID	Current Loan Amount	Term	Credit Score	Annual Income	Years in current job	Home Ownership
539	24a32166-8f8f-...	f156c9d1-cac8-...	322014	Long Term	693	815765	?	Home Mortgage
540	e2ffd7f4-e4e5-...	fd455a4e-7777...	220396	Long Term	?	?	?	Home Mortgage
541	096d61fb-29e9...	4d416b00-32a2...	749144	Short Term	711	2.65628e+06	6 years	Home Mortgage
542	0a3a34eb-b501...	52f95433-b66b...	347292	Short Term	?	?	1 year	Rent
543	e0c1fa65-4360-...	7dcf7cf4-e2fd-...	1e+08	Short Term	729	581913	1 year	Rent
544	d3b25b26-43f3...	f0dbd10a-acce-...	485892	Long Term	716	2.06e+06	10+ years	Home Mortgage
545	e478bddd-232...	596a9b58-c9e0...	1e+08	Short Term	736	756960	< 1 year	Rent
546	2383e95c-104e...	7c6ca34a-9eaf-...	597476	Long Term	719	1.16746e+06	3 years	Home Mortgage
547	36ba4b22-c23f...	dd8c440d-104e...	605396	Short Term	?	?	?	Home Mortgage
548	4453aef8-cf3d-...	e2190425-a83d...	630234	Short Term	712	1.65232e+06	10+ years	Rent
549	49bfe26d-631c...	3fcfecf0-54bb-...	1e+08	Short Term	748	875444	< 1 year	Rent
550	8fb4d4b6-ee02...	0fde000c-ef19-...	388718	Short Term	743	1.1677e+06	1 year	Rent
551	65292f08-aabf-...	68abaf74-bcc5-...	1e+08	Short Term	750	1.13555e+06	< 1 year	Rent
552	23f36715-5a73-...	5ab88861-ca91...	536602	Short Term	721	2.41372e+06	< 1 year	Rent
553	6c4cd297-22d5...	6d2ce7ba-f040-...	462792	Long Term	743	2.72162e+06	3 years	Own Home
554	65073e90-b338...	666846f7-c2fe-...	344432	Short Term	747	3.01182e+06	3 years	Own Home
555	dfc2707c-eb6a-...	0fc91f0c-7564-...	769010	Long Term	680	2.46679e+06	3 years	Home Mortgage
556	12b69dd2-a99a...	0b1a4d65-5d3e...	331276	Long Term	675	2.01301e+06	4 years	Home Mortgage
557	658e7204-9640...	d5f11c7f-829a-...	183040	Short Term	740	483531	6 years	Home Mortgage
558	f3fdee01-7050-...	a38c7b39-b3c7...	333564	Short Term	705	1.00538e+06	6 years	Own Home
559	335cfc18-c9af-...	8e33b6a0-c651...	158202	Short Term	688	1.32827e+06	5 years	Home Mortgage

Final Output:



	Loan ID	Customer ID	Current Loan Amount	Term	Credit Score	Annual Income	Years in current job	Home Ownership	
587	f3475898-5b32...	71342245-a2bc...	272206	Short Term	747	799786	10+ years	Home Mortgage	Debt
588	db3e20b5-d90...	5bdf3ced-4c88-...	606826	Long Term	720	4.86639e+06	2 years	Own Home	Hom
589	0c6706b0-acf4-...	d7c6c1f0-148b-...	306130	Long Term	702	849015	1 year	Rent	Debt
590	732b1101-9a7d...	034497ae-8ff8-...	175516	Short Term	721	974928	10+ years	Home Mortgage	Debt
591	1adf0dcb-2d22...	ee296960-ffd6-...	90156	Short Term	720	1.45977e+06	7 years	Home Mortgage	Hom
592	c7651b1d-3ba0...	c2dac765-a80d...	434808	Long Term	680	1.40818e+06	4 years	Home Mortgage	Debt
593	4b3adf0f-01da-...	fbdf4c5a-3570-...	402380	Long Term	682	926687	8 years	Home Mortgage	Debt
594	881b512c-daae...	a644ead7-c8ea...	779174	Short Term	719	2.69165e+06	10+ years	Home Mortgage	Debt
595	92418167-8286...	e0d674aa-b9b3...	309914	Long Term	1077.99	1.36911e+06	6 years	Home Mortgage	Debt
596	cf24f4cd-0cf5-4...	22d0eb03-10fa...	103202	Short Term	729	965580	4 years	Rent	Debt
597	ce37f199-02e6-...	4762ed71-a5dc...	1e+08	Short Term	750	1.56906e+06	10+ years	Home Mortgage	Debt
598	04d84683-4a61...	bd0d04c9-29a5...	76846	Long Term	705	550031	9 years	Own Home	Debt
599	a1f011a5-9011-...	05f1c958-a7de-...	100408	Short Term	741	568575	< 1 year	Rent	Debt
600	6057917b-f21d...	0c557fd3-c045-...	322256	Short Term	724	1.11321e+06	4 years	Rent	Debt
601	9ccdd8bd-490d...	7be9f4db-2785...	443564	Long Term	1077.99	1.36911e+06	9 years	Rent	Debt
602	ecc8114c-5111-...	1dcef784-25f5-...	152394	Short Term	670	1.42544e+06	< 1 year	Home Mortgage	Busi
603	9209a0c5-cd92...	9dcef2f4-a7ce-...	201278	Short Term	742	965675	10+ years	Own Home	Debt
604	bca30391-40a8...	1c5868e8-459c...	557480	Short Term	714	2.21472e+06	10+ years	Rent	Debt
605	1c949391-c0e4...	03d351cd-bd22...	312642	Long Term	7350	1.44653e+06	3 years	Home Mortgage	Debt
606	bbd5411f-1486...	b9b9c3f0-8040...	441936	Long Term	1077.99	1.36911e+06	7 years	Home Mortgage	Debt
607	62c5de04-47d5...	2cd8ebd8-b5a0...	1e+08	Long Term	722	972268	8 years	Home Mortgage	Debt

RESULT:

The Implementation of various missing handling mechanisms and various noisy handling mechanisms is completed and verified successfully.

Ex No: 03

Date:

EXPLORATORY ANALYSIS DEVELOP K-MEANS CLUSTERING

AIM:

Implement exploratory analysis develop k-means and MST based clustering techniques and Develop the methodology for assessment of clusters for dataset.

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.

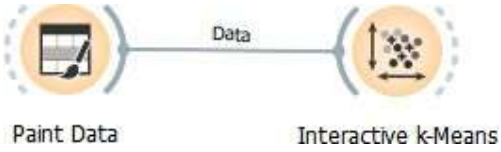
PROBLEM DEFINITION:

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

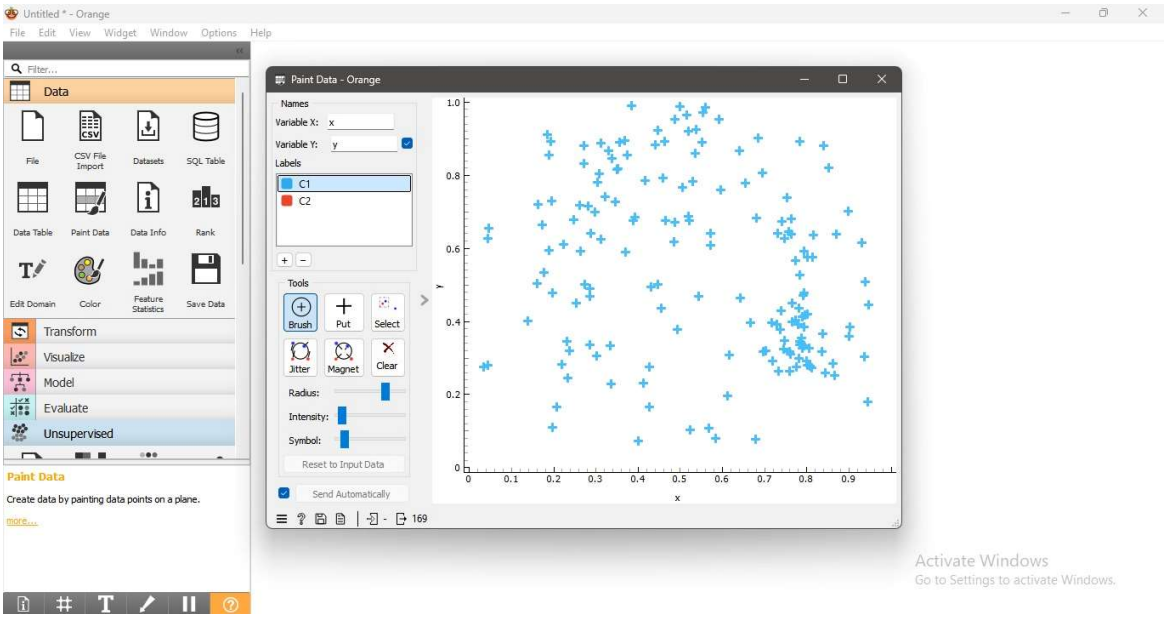
PROCEDURE:

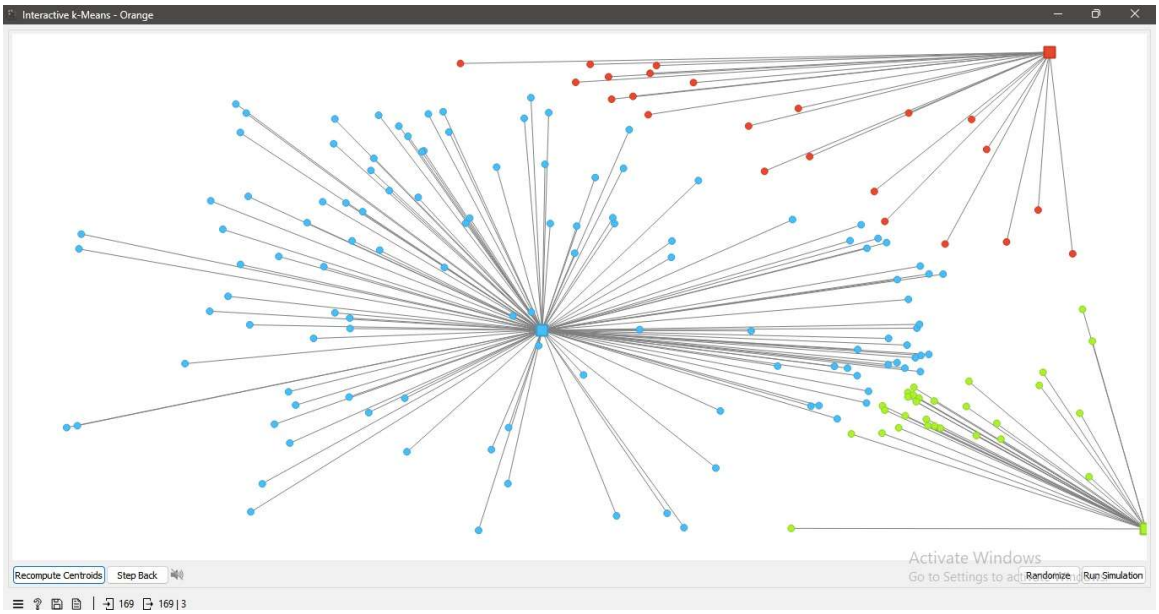
1. Download and install Orange software from https://orangedatamining.com/.
2. Download dataset from Internet.
3. Open Orange software and click on "New". Then, click on the "File" widget.
4. Double-click on the "File" widget to import your dataset. Next, add a "Data Table" widget and connect it with the "File" widget to display the raw data.

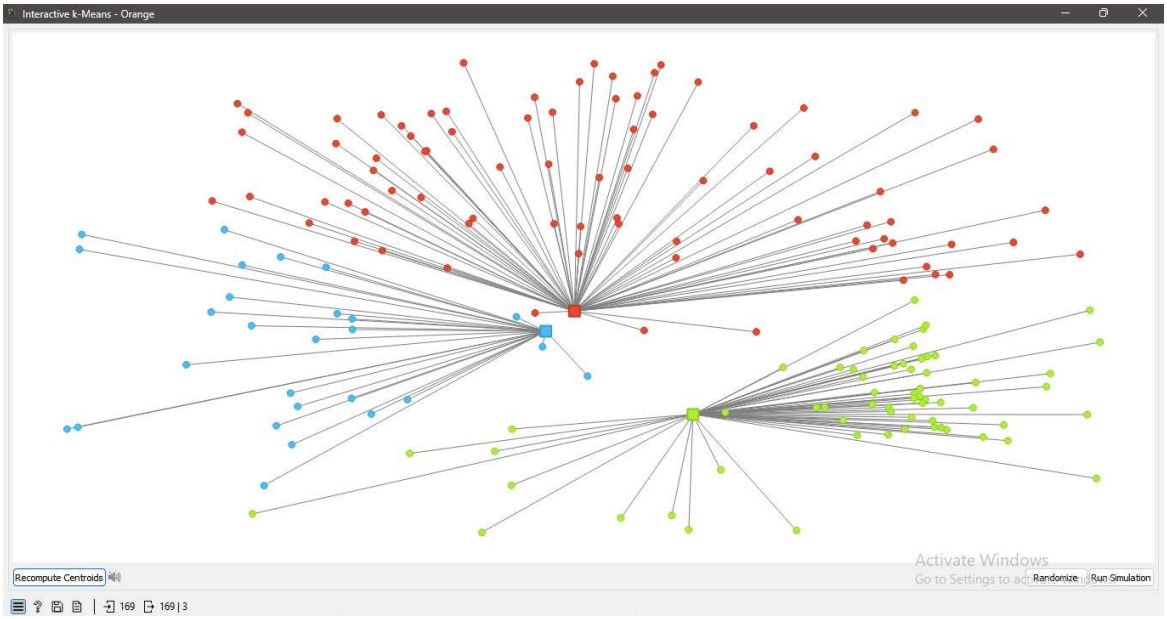
5. Drag the K-Means widget onto the canvas. Connect the Select Columns widget to the K-Means widget. Double-click the K-Means widget to set the number of clusters (k). Connect the K-Means widget to the Scatter Plot widget to visualize the clusters.



6. Drag the Hierarchical Clustering widget onto the canvas. Connect the Select Columns widget to the Hierarchical Clustering widget. Double-click the Hierarchical Clustering widget to configure clustering settings.
7. Double-click the Scatter Plot or Dendrogram widgets to explore and interpret clusters. Use the Silhouette Plot to assess cluster cohesion and separation.







RESULT:

The implementation of exploratory analysis develop k-means and MST based clustering techniques and Develop the methodology for assessment of clusters for dataset is completed and verified successfully.

Ex No: 04

Date:

ASSOCIATION ANALYSIS

AIM:

To Design algorithm for association rule mining algorithm.

DESCRIPTION:

Association rule:

Association rule mining is a popular data mining technique to find interesting relationships (associations) among a large set of data items. This technique is commonly used in market basket analysis to identify sets of products that frequently co-occur in transactions.

Association rule mining involves two main steps:

1. **Frequent Itemset Generation**
2. **Rule Generation**

Dataset:

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

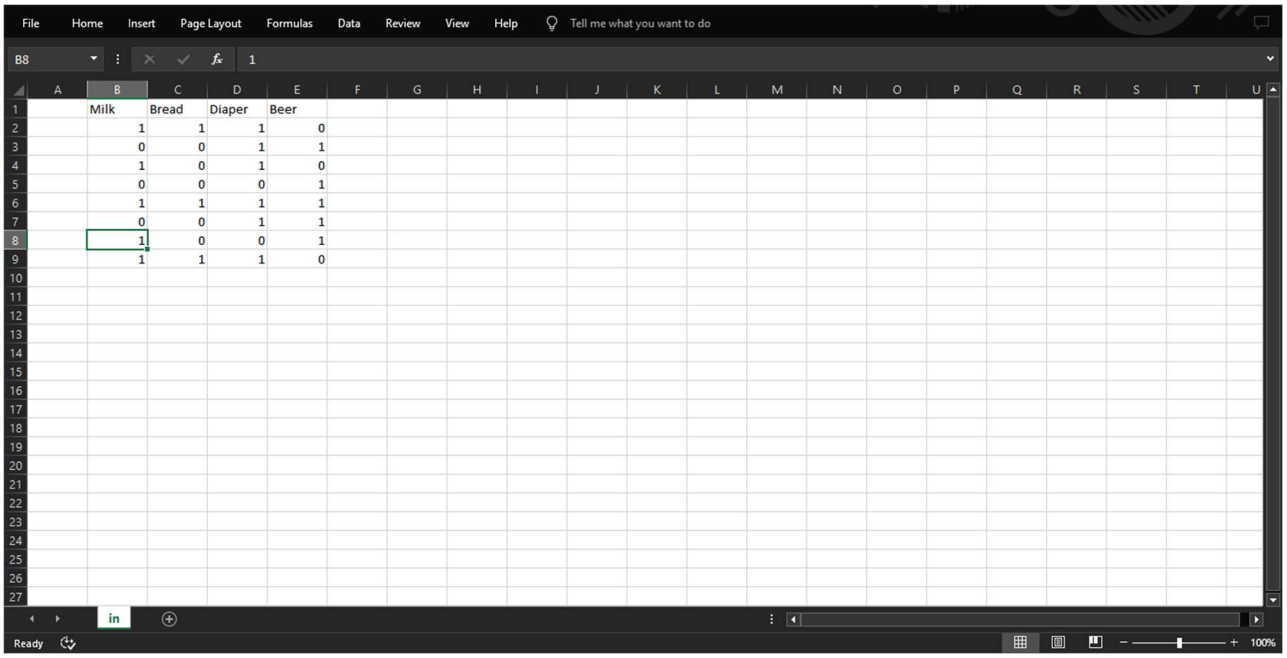
PROBLEM DEFINITION:

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

PROCEDURE:

1. Download and install Orange software from https://orangedatamining.com/.

2. Open Excel and import the data or Open Notepad , import data and save with extension ".csv"



	Milk	Bread	Diaper	Beer
1	1	1	1	0
2	0	0	1	1
3	1	0	1	0
4	0	0	0	1
5	1	1	1	1
6	0	0	1	1
7	1	0	0	1
8	1	1	1	0

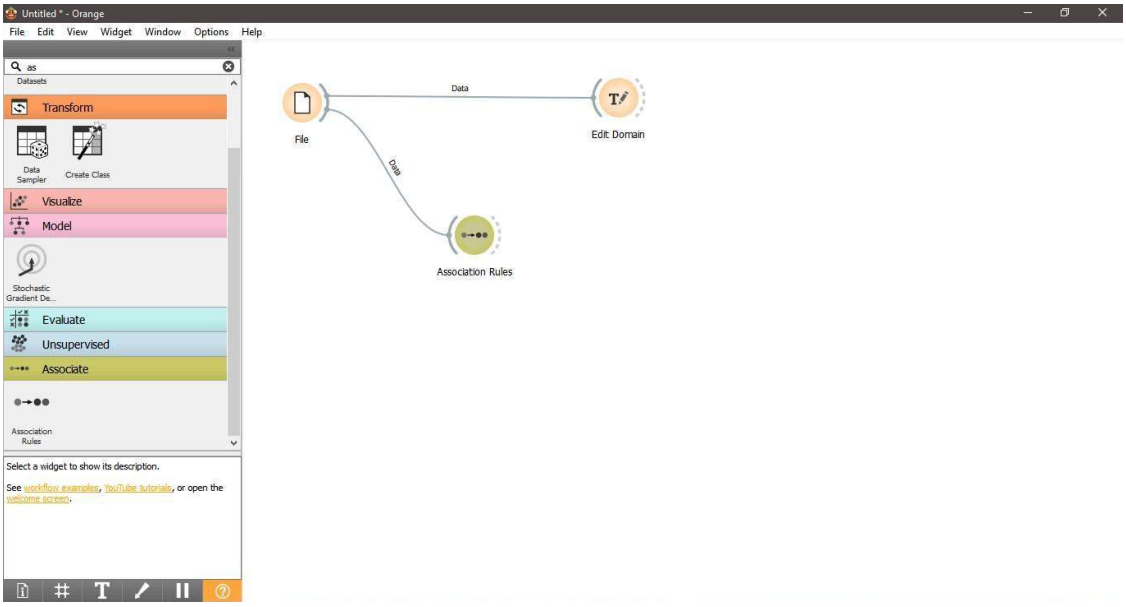
3. Open Orange software and click on "New". Then, click on the "File" widget.

4. Double-click on the "File" widget to import your dataset. Next, add a "Data Table" widget and connect it with the "File" widget to display the raw data(.

5. Next, add a"Edit Domain" widget and connect to the "File". It is used to check the data is in binary format in rows and columns.

6. Next, add an "Associate rule" widget, if not found Click on "Widget" , click on "Add-ons",install "Associate" add-ons.

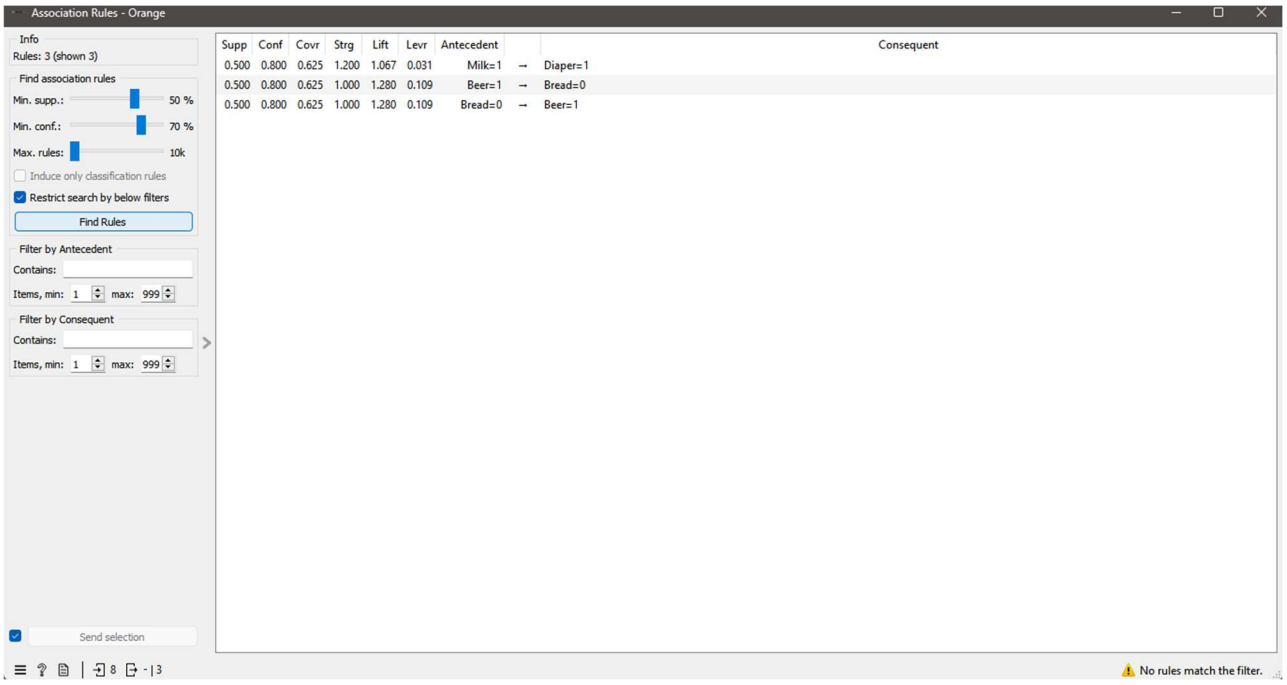
7. Connect "Associate rule" widget with "File" widget.



8. Double Click “Associate rule” widget.

9. Set “Minimum Support” to “50%” and “Minimum Configuration” to “70%”.

10. Click on “Find Rules”. You will find the Associate rules in the screen



RESULT:

The Associate rules mining algorithm using Orange is completed and verified successfully.

EX.NO:5 **HYPOTHESIS GENERATION**

DATE:

Aim:

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

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 → B is the percentage of transactions that contain both A and B.
- Confidence for a rule A → 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:

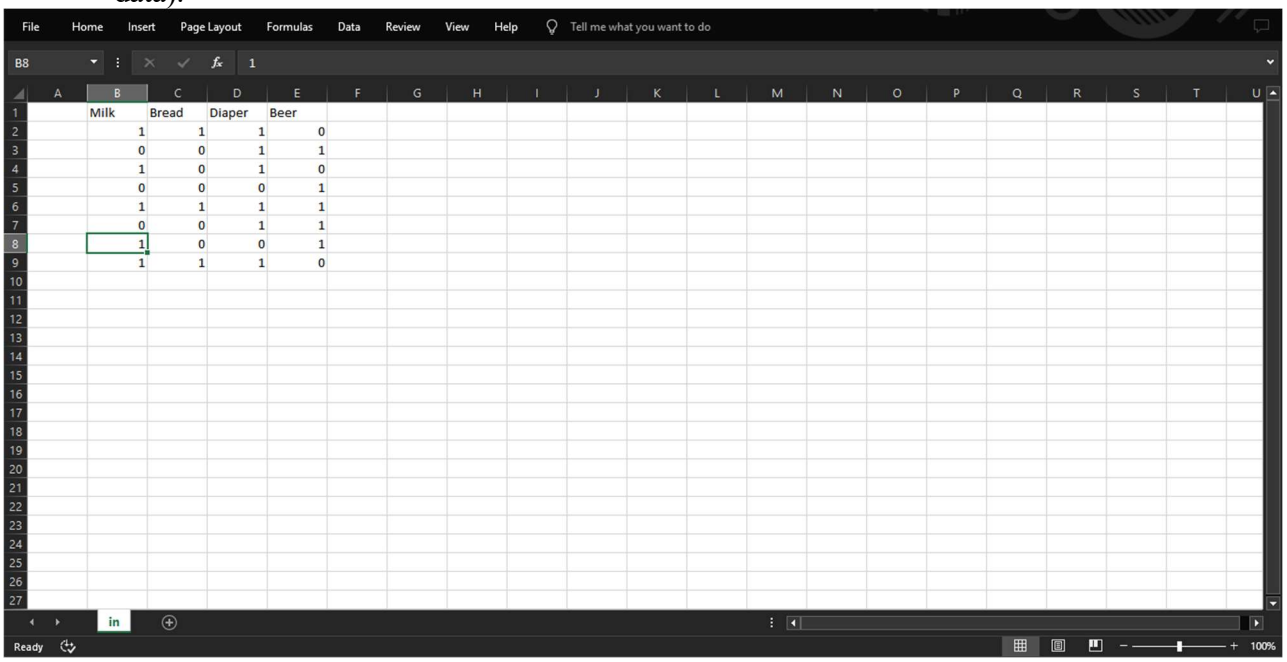
- The minimum support level set for the rules to be considered valid.

Confidence Threshold:

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

Procedure:

1. Download and install Orange software from https://orangedatamining.com/.
2. Drag and drop the **"File"** widget onto the canvas.
3. Double-click the "File" widget, and load your dataset (e.g., a .csv file containing transaction data).

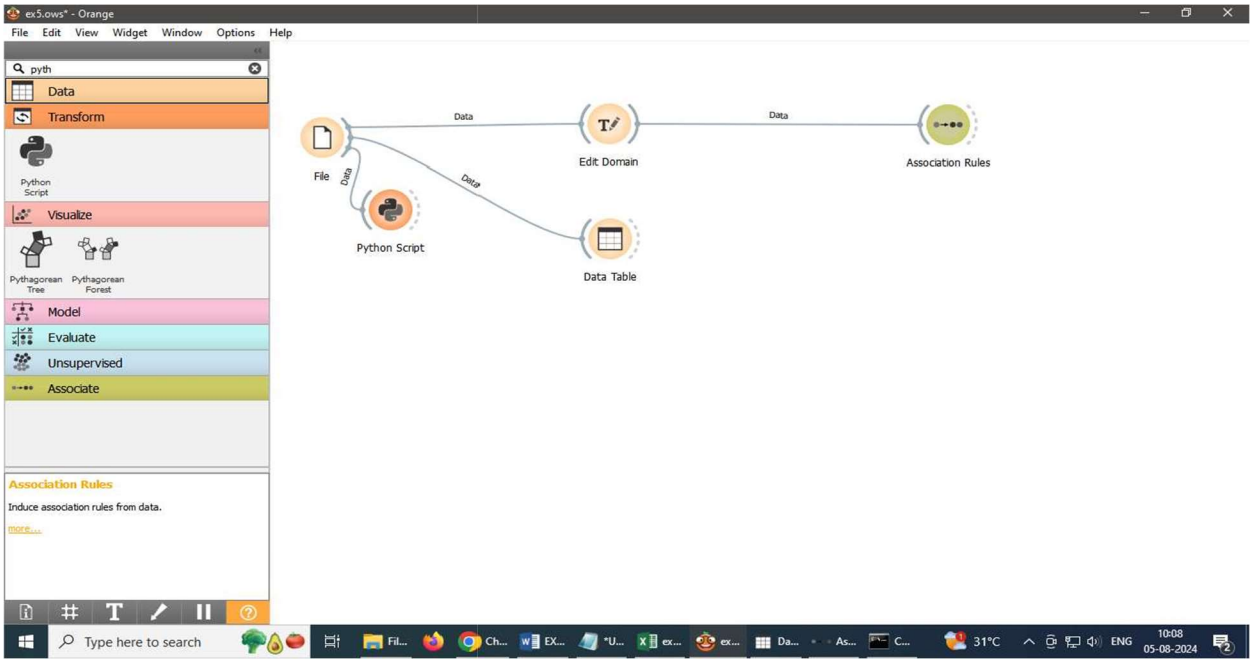


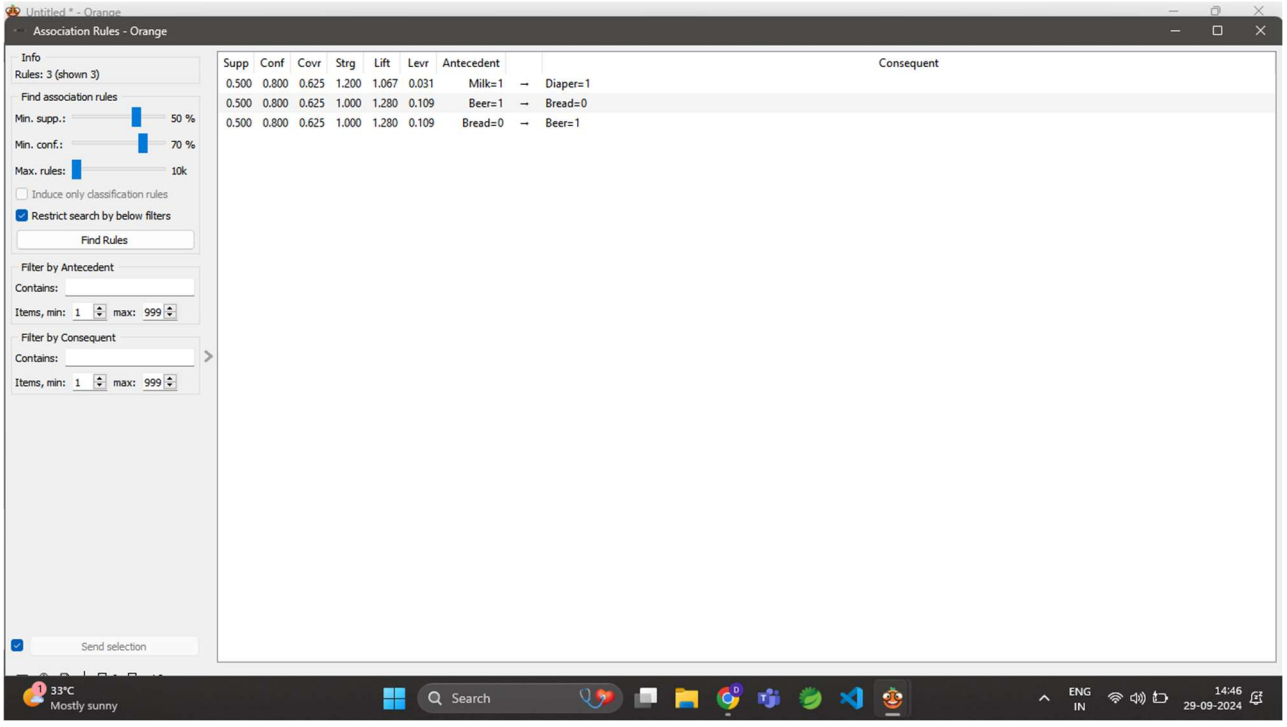
4. Use the "Edit Domain" widget to modify or select the relevant features for analysis to check if the dataset binary format.
5. Connect the data from "Edit Domain" to the "Association Rules" widget.
6. Double-click the "Association Rules" widget to set the support and confidence thresholds.
7. Set the minimum **support threshold** (for 50%).
8. Set the minimum **confidence threshold** (for 70%).
9. Connect the "Associate" widget to the "Data Table" widget to view the generated association rules.
10. Simply look at the rules in the **Data Table** widget to see if your hypothesized rules appear and if they meet your support and confidence criteria. If you want to automate the validation process, you can use a Python script in Orange to filter and validate rules directly.

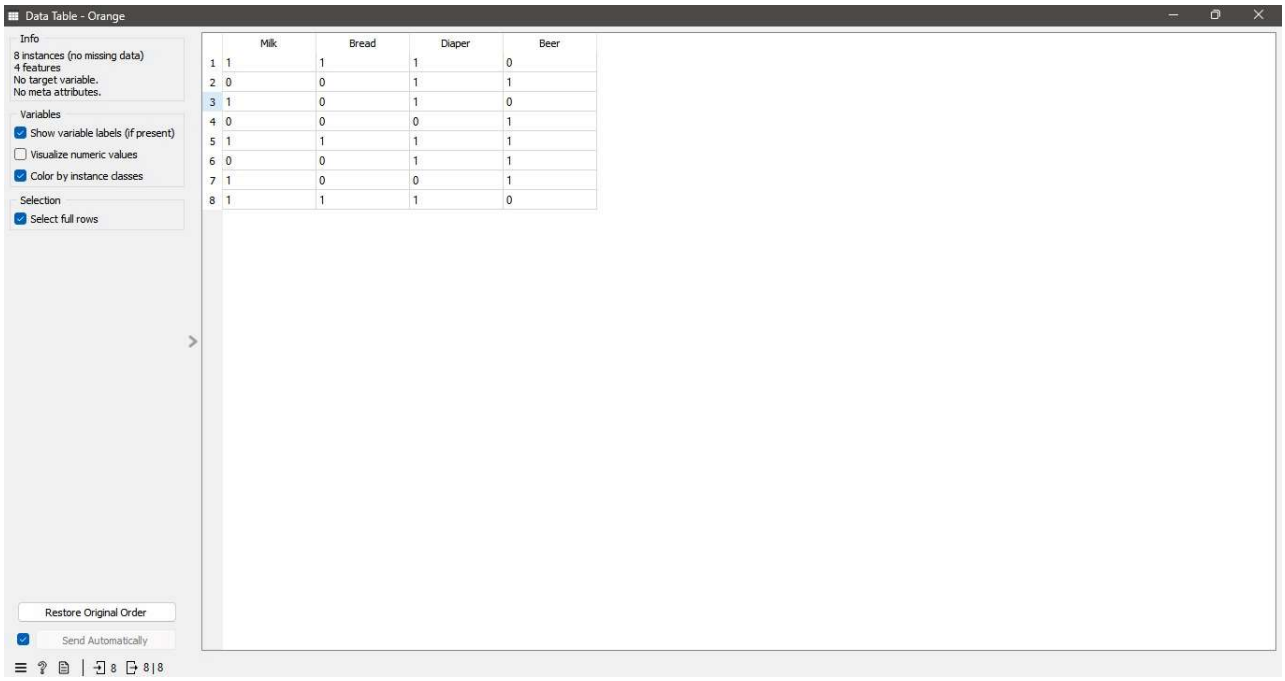
    Code:

```
import pandas as pd
from mlxtend.frequent_patterns import apriori, association_rules
def orange_to_dataframe(table):
    feature_names = table.domain.attributes
    data = [list(row) for row in table]
    df = pd.DataFrame(data, columns=[str(f) for f in feature_names])
    return df

data = orange_to_dataframe(in_data)

def create_transaction_df(df):
    items = df.columns
    df = pd.DataFrame(df.values, columns=items)
    df = df.astype(bool).astype(int)
    return df

transaction_df = create_transaction_df(data)
frequent_itemsets = apriori(transaction_df, min_support=0.5, use_colnames=True)
rules = association_rules(frequent_itemsets, metric="confidence", min_threshold=0.7)

print("Frequent Itemsets:")
print(frequent_itemsets)

print("\nAssociation Rules:")
print(rules)
```

11. Review the output rules, checking for those that meet your thresholds.


ex5.ows* - Orange
File Edit View Widget Window Options Help
pyth
Data
Transform
Python Script
Visualize
Pythagorean Tree
Pythagorean Forest
Model
Evaluate
Unsupervised
Associate
Association Rules
Induce association rules from data.
more...
File
Edit Domain
Association Rules
Python Script
Data Table
Data
Type here to search
31°C
ENG
10:08
05-08-2024


Association Rules - Orange
Info
Rules: 3 (shown 3)
Find association rules
Min. supp.: 50 %
Min. conf.: 70 %
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
0.500 0.800 0.625 1.200 1.067 0.031 Milk=1 → Diaper=1
0.500 0.800 0.625 1.000 1.280 0.109 Beer=1 → Bread=0
0.500 0.800 0.625 1.000 1.280 0.109 Bread=0 → Beer=1
33°C
Mostly sunny
Search
ENG
IN
14:46
29-09-2024



Conclusion:

The output shows the frequent itemsets and the association rules derived from the dataset. Frequent itemsets are those that meet the minimum support threshold, while association rules are evaluated based on the minimum confidence threshold. By analyzing the output, you can identify which rules have strong associations, meaning they have high confidence and support.

EX.No: 6
Date:

TRANSFORMATION TECHNIQUES - Construct Haar wavelet transformation for numerical data, Construct principal componentanalysis (PCA) for 5-dimensional data

Aim:

To perform Haar wavelet transformation and Principal Component Analysis (PCA) on numerical data using Orange software and analyze the results.

Experiment 1: Haar Wavelet TransformationIntroduction:

The Haar wavelet transformation decomposes data into different frequency components using Haar wavelets. It simplifies analysis by representing data with a hierarchical structure of averages and differences, making it useful for tasks like signal compression and feature extraction.

Materials Required:

1. Google Colab or Local Python Environment
2. Python installed with packages: PyWavelets, numpy, pandas
3. Orange Software
4. CSV File (generated from Google Colab)

Procedure:

Part 1: Haar Wavelet Transformation using Google Colab

1. Setup Google Colab:
   - Open Google Colab.
   - Click on the '+ Code' icon at top of the page.

2. Install Required Libraries:

   !pip install PyWavelets numpy pandas

   Run the cell to install the libraries.

3. Write the Code for Haar Wavelet Transformation:

```
import pywt
import numpy as npimport
pandas as pd

data = np.random.rand(100) # Sample data
coeffs = pywt.wavedec(data, 'haar') # Perform Haarwavelet transformation
df = pd.DataFrame(coeffs).T # Convert to DataFrame

df.to_csv('haar_transformed_data.csv', index=False)
```

4. Run the Code:
   Execute the cell to perform the Haar wavelet transformation and save the data to a CSV file**.**

5. Download the CSV File:

```
from google.colab import files files.download('haar_transformed_data.csv')
```

Run this cell to download the CSV file to your local machine.

Part 2: Analyzing the Data using Orange Software

1. Open Orange Software:

Launch Orange on your local machine or access Orange3 through the web interface.

2. Add a File Widget:

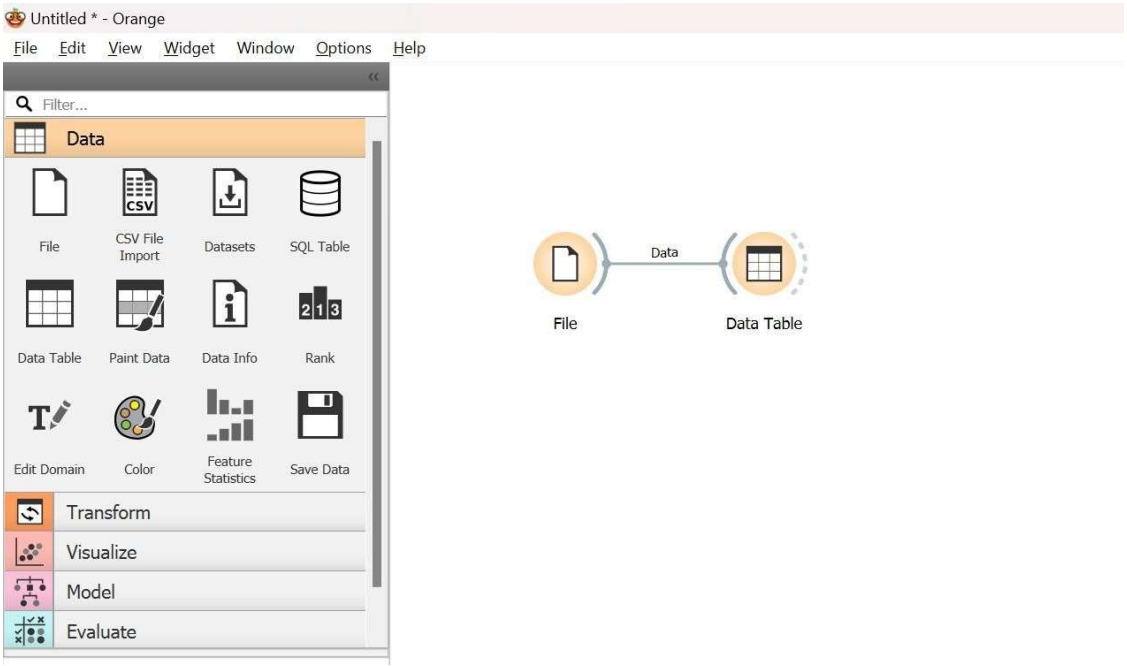
- Drag and drop the ‘File’ widget from the ‘Data’ section onto the Orange canvas.

3. Load the CSV File:

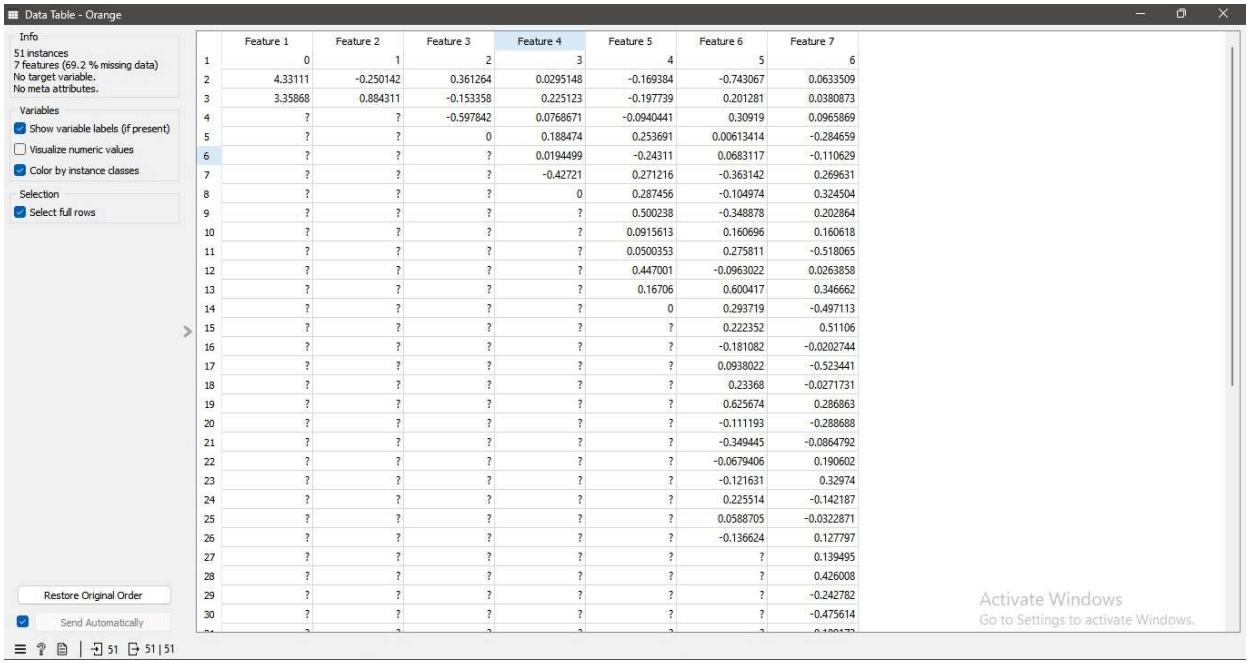
- Double-click the ‘File’ widget to open its settings.
- Click the Browse button and upload the haar_transformed_data.csv file that you downloaded fromGoogle Colab.

4. View Data:

- After loading the file, connect a ‘Data Table’ widget to the‘File’ widget to view the data in a tabular format.



Output:



	Feature 1	Feature 2	Feature 3	Feature 4	Feature 5	Feature 6	Feature 7
1	0	1	2	3	4	5	6
2	4.33111	-0.250142	0.361264	0.0295148	-0.169384	-0.743067	0.0633509
3	3.35868	0.884311	-0.153358	0.225123	-0.197739	0.201281	0.0380873
4	?	?	-0.597842	0.0768671	-0.0940441	0.30919	0.0965869
5	?	?	0	0.188474	0.253691	0.00613414	-0.284659
6	?	?	?	0.0194499	-0.24311	0.0683117	-0.110629
7	?	?	?	-0.42721	0.271216	-0.363142	0.269631
8	?	?	?	0	0.287456	-0.104974	0.324504
9	?	?	?	?	0.500238	-0.348878	0.202864
10	?	?	?	?	0.0915613	0.160696	0.160618
11	?	?	?	?	0.0500353	0.275811	-0.518065
12	?	?	?	?	0.447001	-0.0963022	0.0263858
13	?	?	?	?	0.16706	0.600417	0.346662
14	?	?	?	?	0	0.293719	-0.497113
15	?	?	?	?	?	0.222352	0.51106
16	?	?	?	?	?	-0.181082	-0.0202744
17	?	?	?	?	?	0.0938022	-0.523441
18	?	?	?	?	?	0.23368	-0.0271731
19	?	?	?	?	?	0.625674	0.286863
20	?	?	?	?	?	-0.111193	-0.288688
21	?	?	?	?	?	-0.349445	-0.0864792
22	?	?	?	?	?	-0.0679406	0.190602
23	?	?	?	?	?	-0.121631	0.32974
24	?	?	?	?	?	0.225514	-0.142187
25	?	?	?	?	?	0.0588705	-0.0322871
26	?	?	?	?	?	-0.136624	0.127797
27	?	?	?	?	?	?	0.139495
28	?	?	?	?	?	?	0.426008
29	?	?	?	?	?	?	-0.242782
30	?	?	?	?	?	?	-0.475614

Experiment 2: Principal Component Analysis (PCA)Introduction:

Principal Component Analysis (PCA) is a technique for reducing the dimensionality of data while preserving as much variance as possible. It transforms data into a set of orthogonal components.

Materials Required:

1. Orange software installed on your computer
2. A dataset with numerical data (CSV file or similarformat)

Procedure:

Step 1: Load Data

1. Open Orange software.
2. Drag and drop the File widget from the left pane to thecanvas.
3. Click on the File widget and load your dataset.

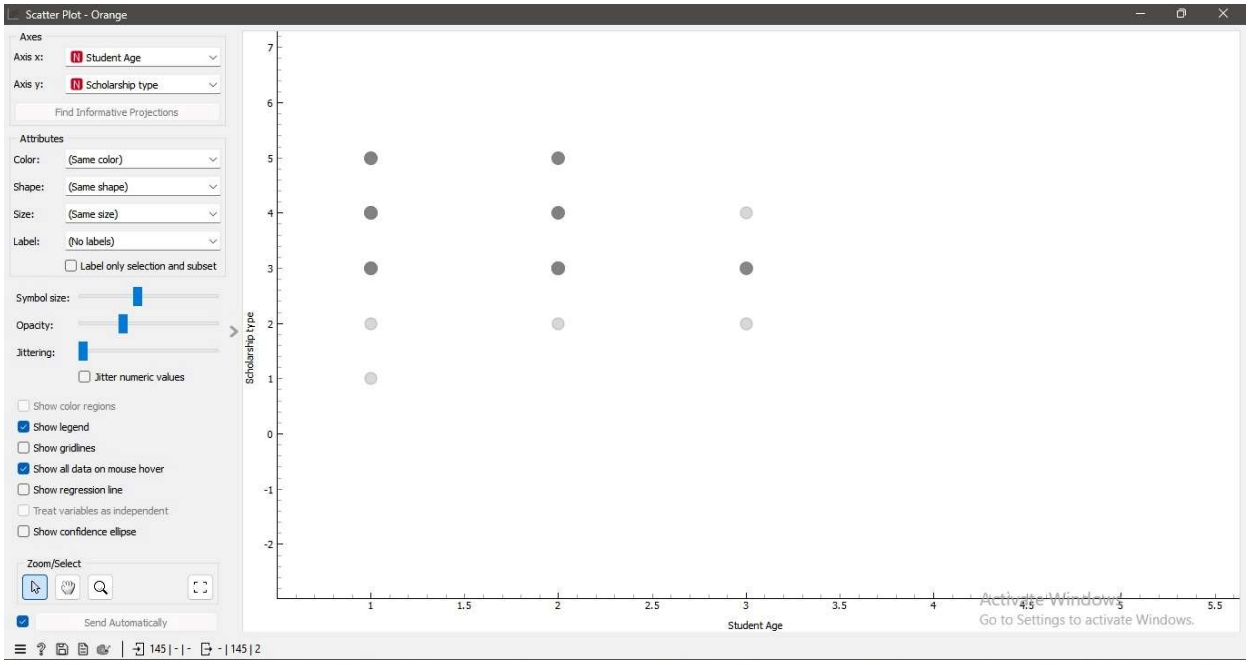
Step 2: Perform PCA

1. Drag and drop the 'PCA' widget from the left pane to thecanvas.
2. Connect the 'File' widget to the 'PCA' widget.
3. Open the 'PCA' widget to configure the parameters ifneeded.

Step 3: Visualize PCA Results

1. Drag and drop the ‘Scatter Plot’ widget to the canvas.
2. Connect the ‘PCA’ widget to the ‘Scatter Plot’ widget.
3. Open the ‘Scatter Plot’ widget to visualize the PCAresults.

Output :



Conclusion:

Haar Wavelet Transformation and Principal ComponentAnalysis has been successfully constructed using Orange Software.

Ex no:7

Date:

Data visualization implement binning visualisations for any real time dataset, implement linear regression techniques

Aim

To implement binning visualizations for a real-time dataset and perform linear regression using Orange software, with the goal of better understanding the data distribution and identifying patterns and relationships between variables.

Materials Required

- Orange Data Mining software (installed on your system)
- Real-time dataset (CSV or Excel file format)
- Computer with sufficient memory and processing power
- Internet connection (to download real-time datasets, if necessary)

Description

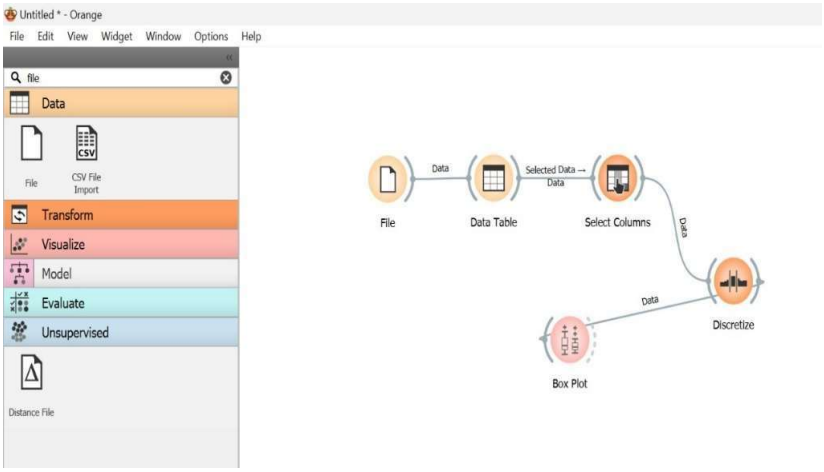
Data visualization and linear regression are essential techniques in data analytics that help in identifying patterns, trends, and relationships within datasets. The goal of this project is to use Orange software to implement binning visualizations for data distribution and apply linear regression techniques to model relationships between dependent and independent variables. Binning allows for better understanding of continuous data by grouping it into intervals or bins. Linear regression is used to predict the dependent variable based on one or more independent variables.

Problem Definition

The goal of this project is to visualize the distribution of continuous variables through binning and build a linear regression model to predict housing prices in Boston. This will help in understanding how different housing and environmental factors affect property values.

Procedure

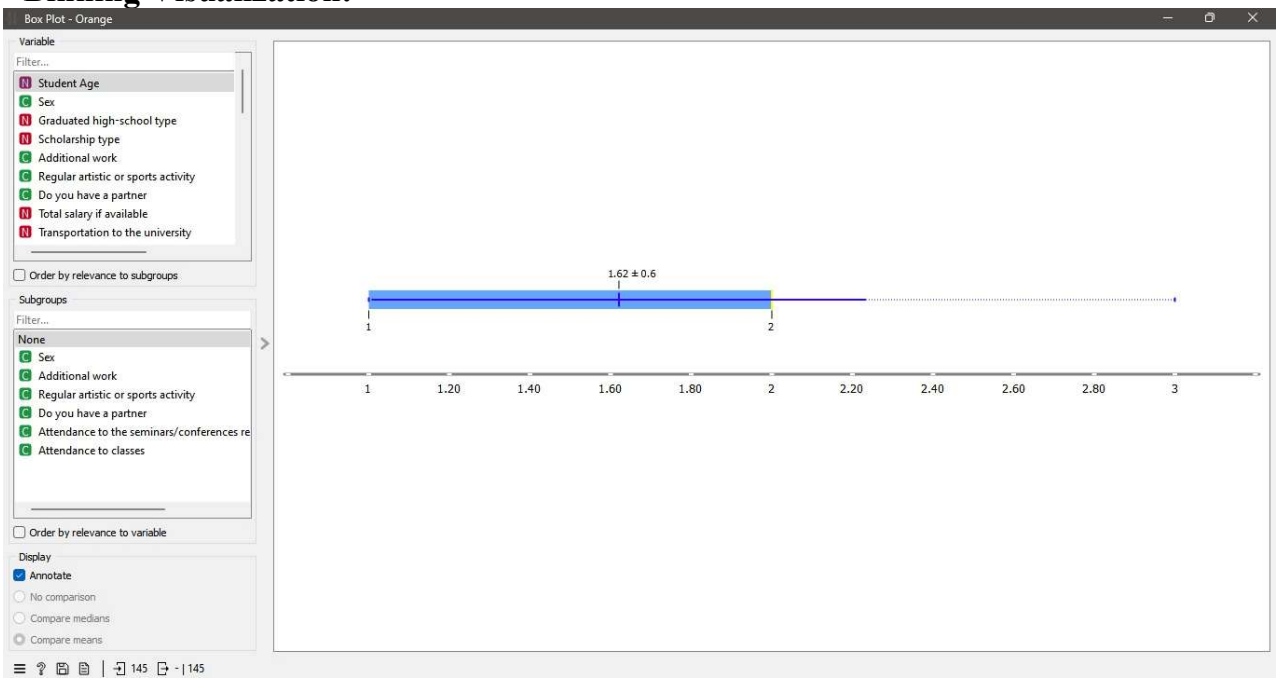
1. **Install and Launch Orange Software**
    - Download and install Orange from the official website.
    - Open Orange and create a new workflow.
2. **Import the Dataset**
    - Obtain a real-time dataset (such as from Kaggle, government databases, or stock prices).
    - In Orange, use the “File” widget to import the dataset in CSV or Excel format.
3. **Data Preprocessing**
    - Drag the “Data Table” widget to explore the dataset.
    - Use “Select Columns” widgets to clean and filter the data as needed.
    - Apply the “Discretize” widget for binning. This will transform continuous variables into bins.
4. **Binning Visualization**
    - Connect the “Discretize” widget to the “Box Plot” widget.
    - Visualize the distribution of the data using these tools.
    - Adjust the binning strategy (e.g., fixed width, equal frequency) to observe how it affects the distribution.



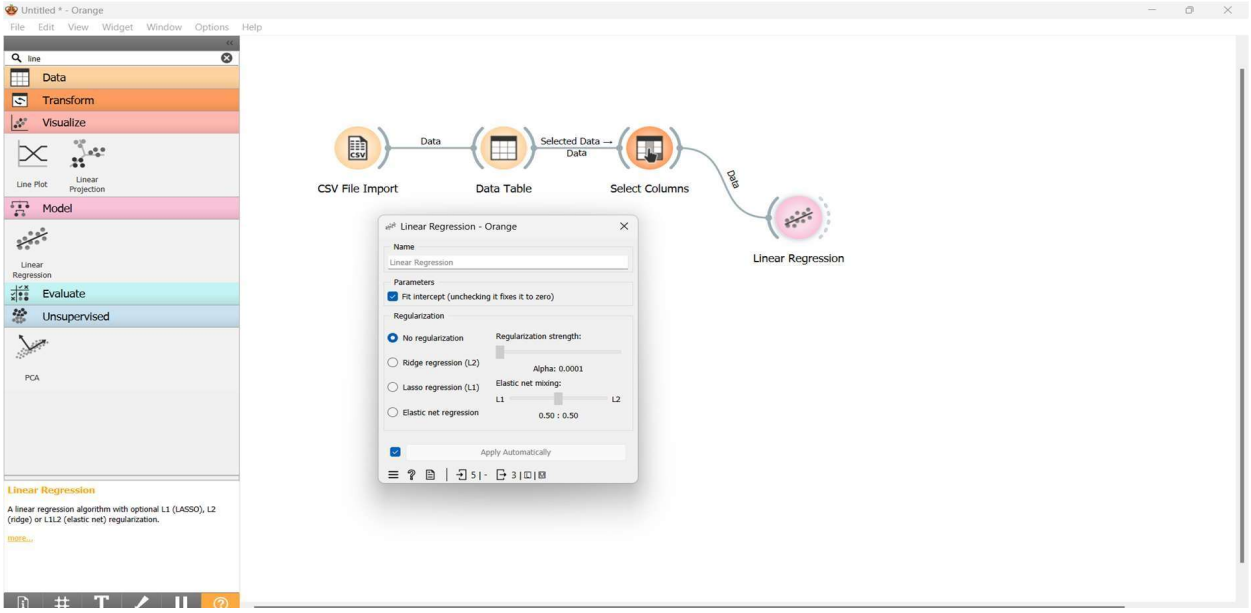
5. **Linear Regression Implementation**
    - Add the “Linear Regression” widget to your workflow.
    - Connect the pre-processed dataset to the “Linear Regression” widget.
    - Choose the independent (input) variables and the dependent (output) variable.
    - Run the regression to get the model parameters.
6. **Model Evaluation**
    - Use the “Test & Score” widget to evaluate the performance of the linear regression model.
    - Connect the "Linear Regression" widget to the “Test & Score” widget and run the evaluation.
    - Analyse the regression outputs such as R-squared value, Mean Absolute Error (MAE), and Root Mean Square Error (RMSE).
7. **Output Visualization**
    - Use the “Scatter Plot” widget to visualize the linear regression results.
    - Add the “Predictions” widget to compare the predicted values with actual data.
8. **Save Results**
    - You can export your results and visualizations using the “Save” or “Report” widget in Orange.

Output

Binning Visualization:



Linear Regression:



Result

Thus, the implementation of binning visualizations the dataset was completed and verified successfully using Orange software.

Ex no:8

Date:

Visualize the clusters for any synthetic dataset, Implement the program for converting the clusters into histograms

AIM:

To visualize the clusters for a synthetic dataset and convert these clusters into histograms using Orange software.

MATERIAL REQUIRED:

- Orange Data Mining Software
- Synthetic dataset (generated or imported)
- Computer with internet access (to download the software and dataset)

DESCRIPTION:

Clustering is an unsupervised learning technique used to group similar data points based on specific characteristics. In this experiment, we'll use Orange software to visualize clusters of a synthetic dataset and convert these clusters into histograms to understand their distribution. Orange is a powerful data mining tool that offers interactive workflows for data visualization, machine learning, and data analysis.

PROBLEM DEFINITION:

The goal is to:

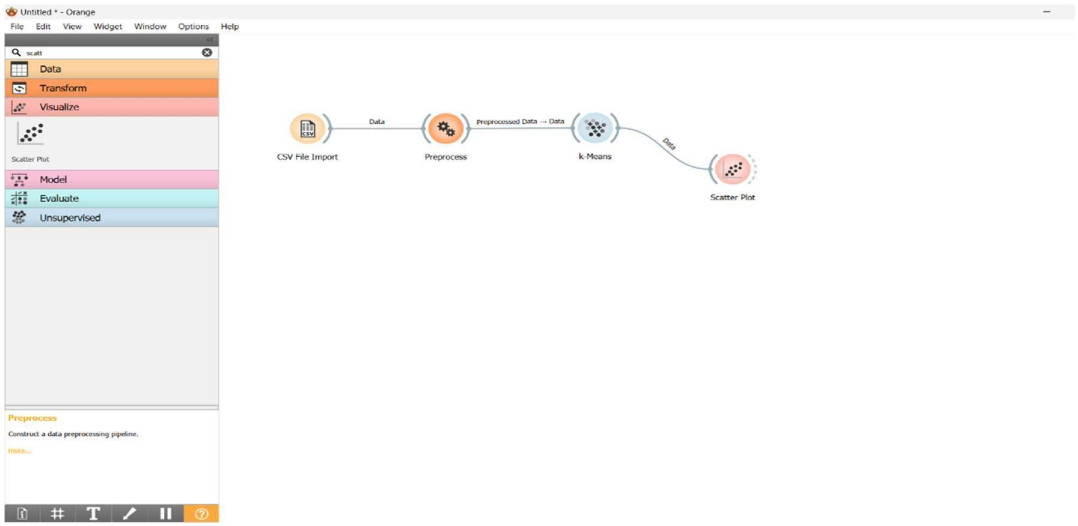
1. Generate or import a synthetic dataset in Orange.
2. Apply clustering algorithms to divide the data into clusters.
3. Visualize the clusters graphically.
4. Convert the clustered data into histograms to analyze the frequency distribution of the data points within each cluster.

PROCEDURE:

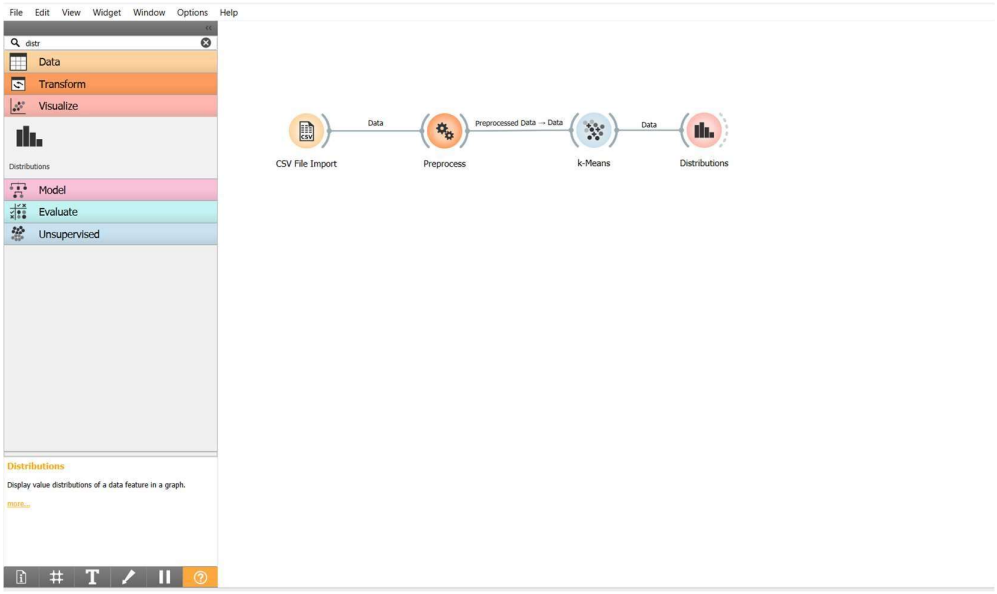
1. **Install Orange:**
    - Download and install Orange Data Mining Software from the official website.
2. **Load or Create a Dataset:**

- Use the "File" widget to load a synthetic dataset (either create your dataset or use an existing one, such as the Iris dataset or a randomly generated dataset).
- You can also use the "Data Table" widget to input synthetic data manually.

3. **Pre-process the Dataset:**
   - Ensure that the dataset is clean and suitable for clustering. Use the "Preprocess" widget for data cleaning if necessary.
4. **Apply Clustering Algorithm:**
   - Add the "K-Means" widget from the "Unsupervised Learning" category.
   - Connect the "File" widget to the "K-Means" widget.
   - Set the number of clusters (e.g., 3 or more, depending on your data).
   - Click on the K-Means widget to run the clustering algorithm.
5. **Visualize Clusters:**
   - Add a "Scatter Plot" widget from the "Visualize" category.
   - Connect the "K-Means" widget to the visualization widget to see the clusters plotted based on the features.



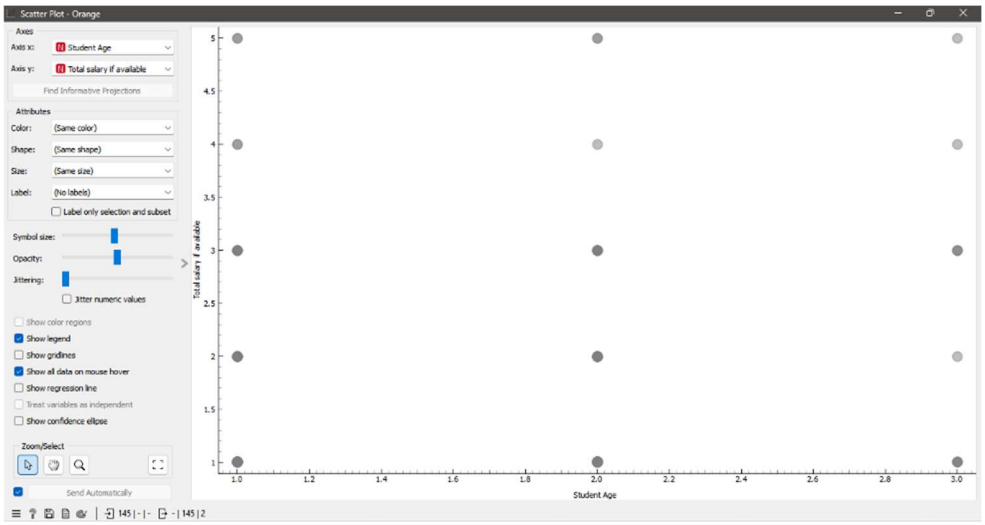
6. **Convert Clusters into Histograms:**
    - Add the "Distributions" widget.
    - Connect the "K-Means" widget to the "Distributions" widget.
    - Select the attribute you want to plot in histograms.
    - The output will show the histogram distribution of clusters.



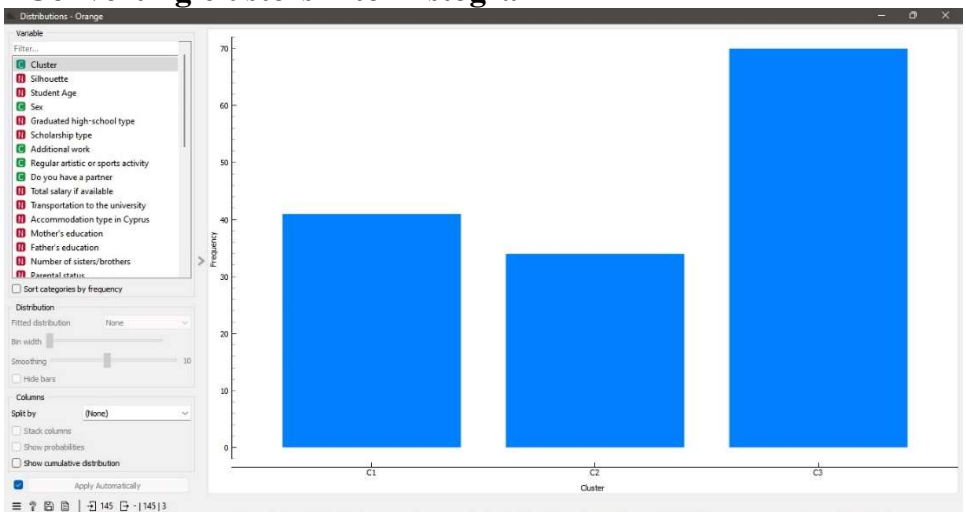
7. **Observe the Histograms:**
    - Each bar in the histogram will represent the frequency of data points in that cluster, providing an easy way to interpret the density of each cluster.

OUTPUT:

Visualize cluster



Converting clusters into Histogram



RESULT:

Thus visualizing the clusters from a synthetic dataset and converted the clusters into histograms using Orange software is successful and verified.

EX.NO : 9 HIERARCHICAL CLUSTERING

DATE :

AIM :

To perform Agglomerative Clustering and Divisive Hierarchical Clustering on numerical data using Orange software and analyze the results .

DESCRIPTION:

Hierarchical Clustering: It is a method of clustering that seeks to build a hierarchy of clusters. Two main types:

- **Agglomerative Clustering**: A bottom-up approach where each object starts as its own cluster, and pairs of clusters are merged as one moves up the hierarchy.
- **Divisive Clustering**: A top-down approach where all objects start in one cluster and splits are performed recursively as one moves down the hierarchy.

MATERIALS REQUIRED:

1. **Orange Software** installed on your local machine.
2. A dataset with numerical data
3. Python environment

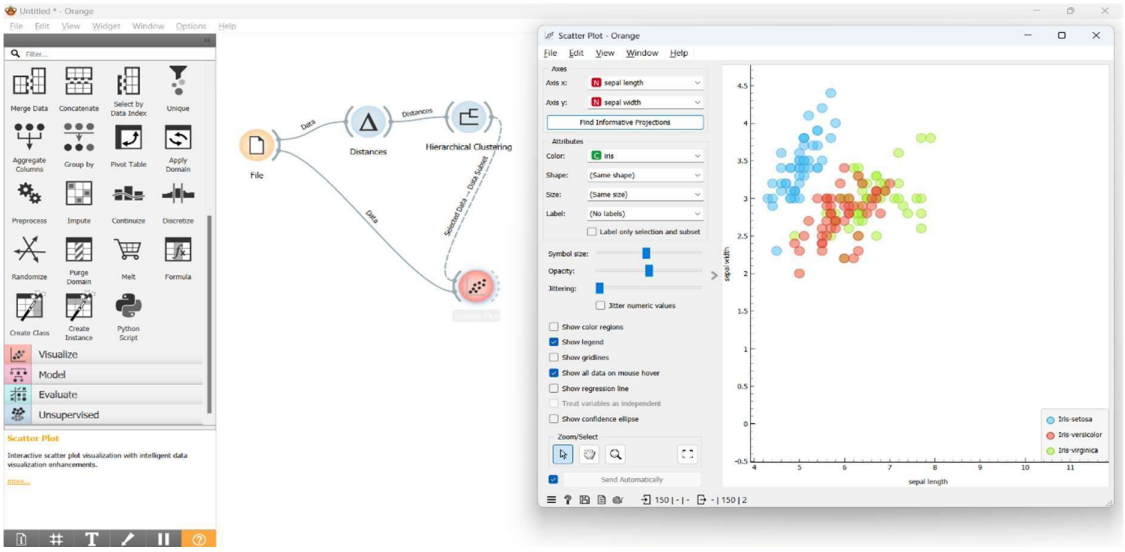
PROCEDURE:

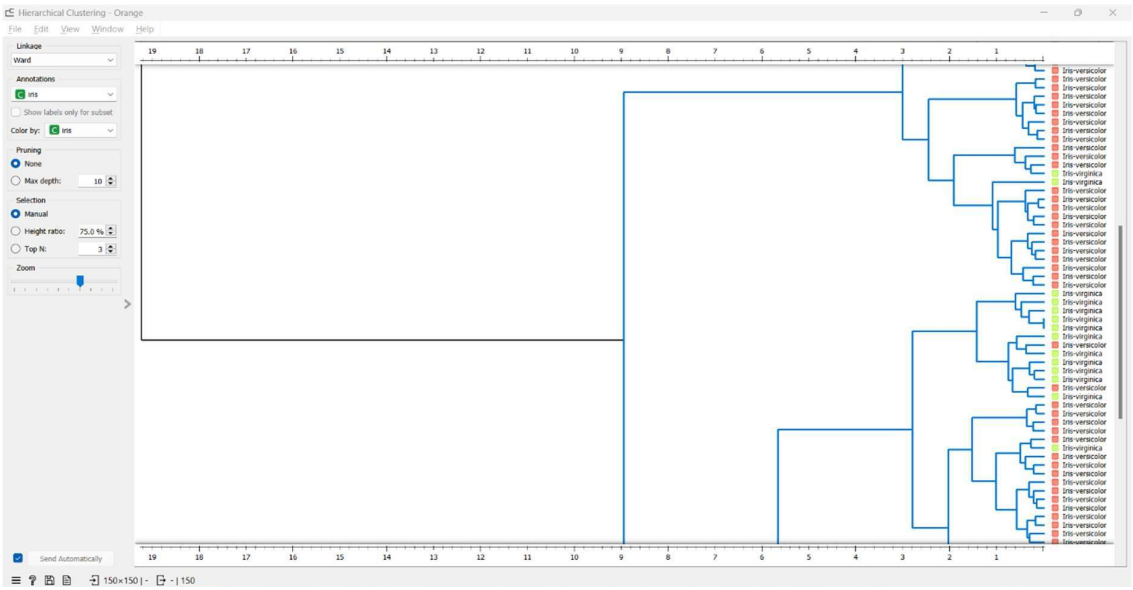
1. AGGLOMERATIVE HIERARCHICAL CLUSTERING:

1. **Download and install Orange software** from "https://orangedatamining.com/".
2. **Download dataset** from the Internet (in .csv format).
3. Open Orange software and click on **"New"** to create a new project. Then, click on the **"File"** widget.
4. Double-click on the **"File" widget** to import your dataset and click **"Apply"**.
5. **Add the Pre-process widget** to clean the data (optional) and configure it, if needed.
6. Locate the **Hierarchical Clustering** widget under the **"Unsupervised"** tab.

7. Connect the **Pre-process Widget** (or File Widget) to the **Hierarchical Clustering** widget.
8. Open the **Hierarchical Clustering Widget** and choose the appropriate linkage method (e.g., **Ward's method**, **Average linkage**, etc.) and the distance metric (e.g., **Euclidean distance**).
9. Visualize the clustering result using the **Scatter Plot Widget** .

OUTPUT:





2. DIVISIVE HIERARCHICAL CLUSTERING:

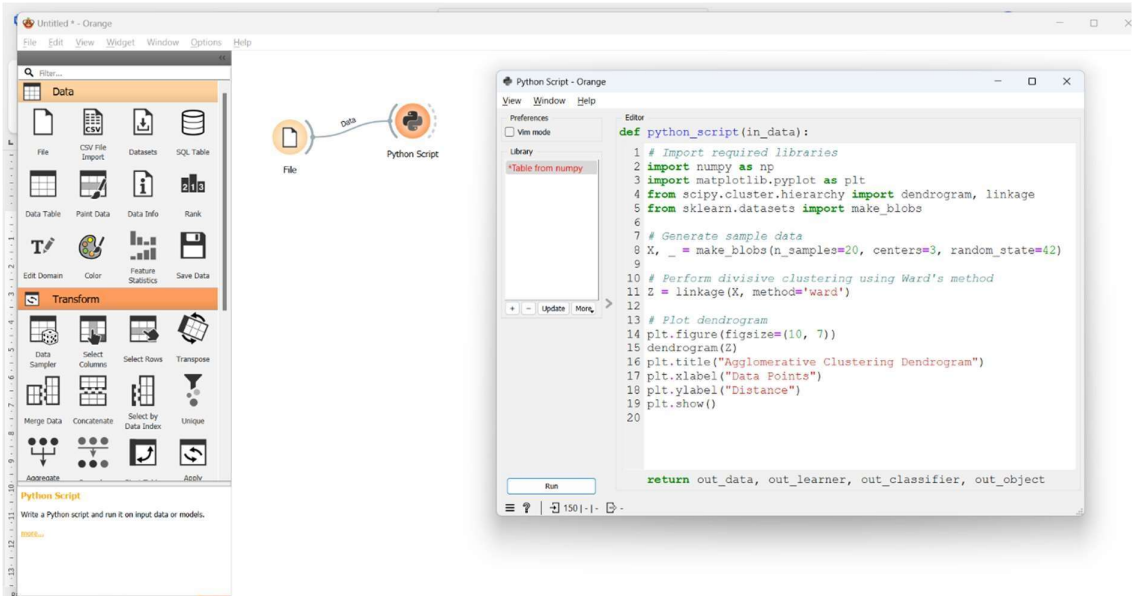
1. **Download and install Orange software** from "https://orangedatamining.com/".
2. **Download dataset** from the Internet (in .csv format).
3. Open Orange software and click on **"New"**. Then, click on the **"File" widget**.
4. Double-click on the **"File" widget** to import your dataset and click **"Apply"**.
5. Add the **Python Script** widget (as Orange does not natively support Divisive Hierarchical Clustering) and connect it to the **File Widget**.
6. Write or import a script that performs **Divisive Clustering** using libraries like **SciPy** or **scikit-learn**.

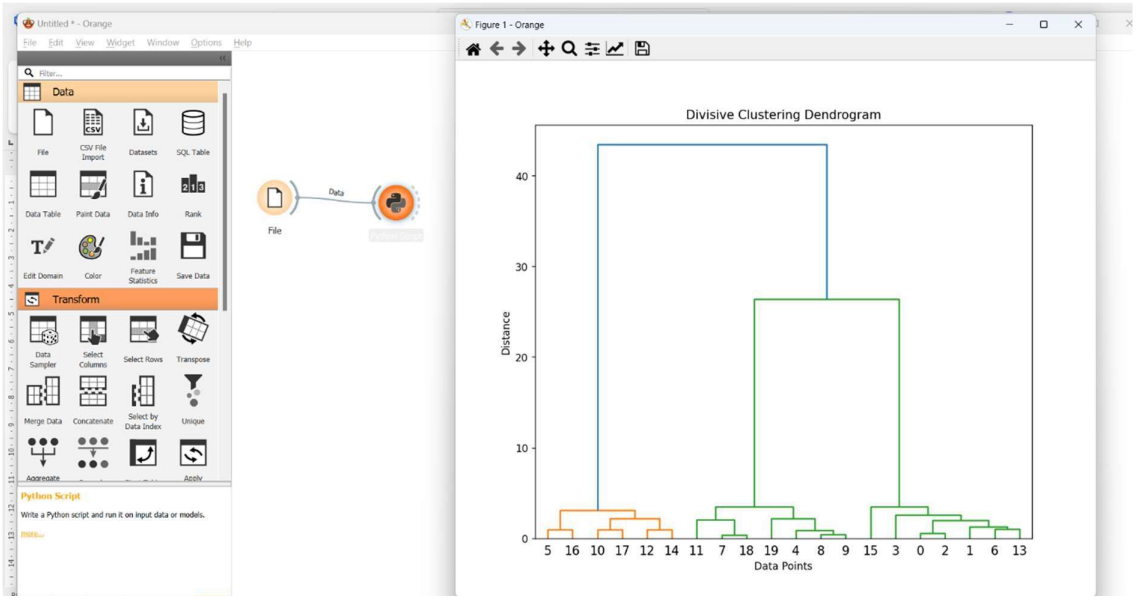
PYTHON CODE:

```
import numpy as np

import matplotlib.pyplot as plt

from scipy.cluster.hierarchy import dendrogram, linkage

from sklearn.datasets import make_blobs

X, _ = make_blobs(n_samples=20, centers=3, random_state=42)

Z = linkage(X, method='ward')

plt.figure(figsize=(10, 7))

dendrogram(Z)

plt.title("Divisive Clustering Dendrogram")

plt.xlabel("Data Points")

plt.ylabel("Distance")

plt.show()
```

7. Execute the above Python script, and visualize the result using the **Data Table** widget.

OUTPUT:





RESULT:

Thus the implementation of Agglomerative Clustering and Divisive Hierarchical Clustering using Orange Software was successfully completed and verified.

EX.NO:10 **SCALABILITY ALGORITHMS DEVELOP SCALABLE CLUSTERING ALGORITHMS, DEVELOP SCALABLE APRIORI ALGORITHM**

DATE:

AIM:

To develop scalable clustering algorithms and a scalable Apriori algorithm in Orange Software.

DESCRIPTION:

Scalable Clustering Algorithms

Overview of Clustering

- **Clustering**: A method of grouping a set of objects in such a way that objects in the same group (cluster) are more similar to each other than to those in other groups.
- **Common Algorithms**: K-Means, Hierarchical Clustering, DBSCAN, etc.

Overview of the Apriori Algorithm

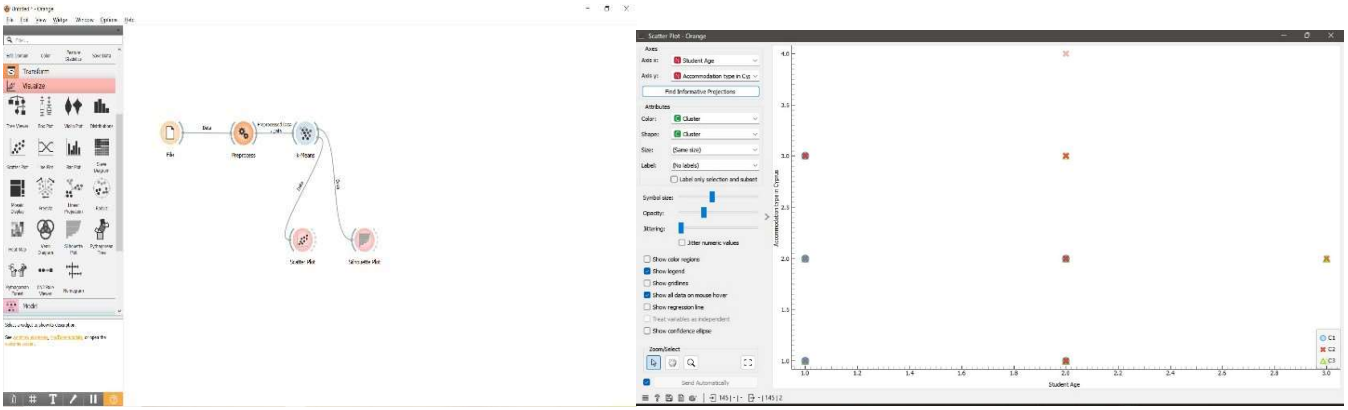
- **Association Rule Mining**: Association rule mining is a technique used to discover interesting relationships or associations between items in large datasets. It is widely used in market basket analysis, recommendation systems, and other applications where understanding item co-occurrence is valuable.

PROCEDURE:

1.K-means Clustering:

1. Download and install Orange software from “https://orangedatamining.com/”.
2. Download dataset from Internet(.csv file).
3. Open Orange software and click on "New". Then, click on the "File" widget.
4. Double-click on the "File" widget to import your dataset and click apply.
5. Add the Pre-process widget to clean the data(if optional) and configure it.
6. Locate the K-Means widget under the Unsupervised tab.
7. Connect the Pre-process Widget (or File Widget) to the K-Means Widget.
8. Open the K-Means Widget and Specify the Number of Clusters(i.e.,3 or 5) and the click apply.
9. Visualize the Clustering Result using any visualize widgets like scatter plot.

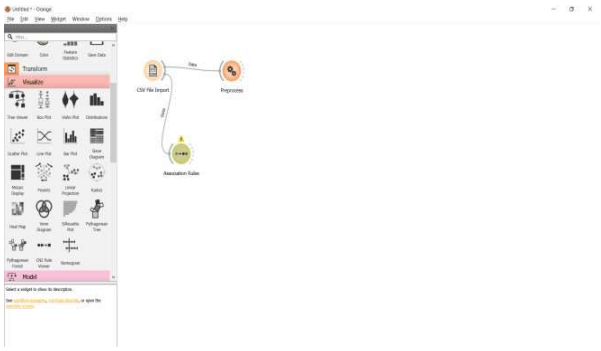
OUTPUT:

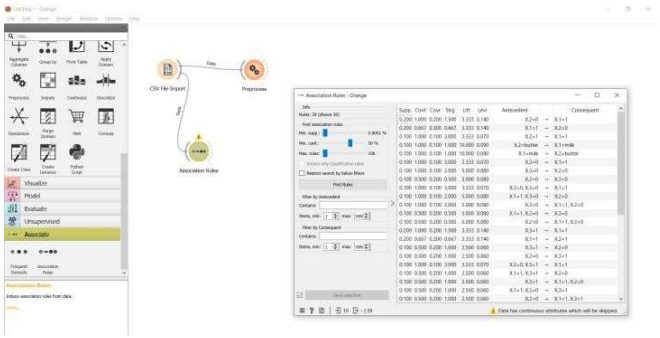


2. Apriori Algorithm:

1. Download and install Orange software from “https://orangedatamining.com/”.
2. Download dataset from Internet(.csv file).
3. Open Orange software and click on "New". Then, click on the "Import CSV File" widget.
4. Double-click on the "Import CSV File" widget to import your dataset and click apply.
5. Add the Pre-process widget to clean the data(if optional) and configure it.
6. Locate Association rule in Association Tab.
7. Connect the Pre-process Widget (or File Widget) to the Association widget.
8. Open the association widget and configure.
9. Give minimum support as 0.00001 and minimum Confidence as 0.5 and click apply.
10. Then click find rules it shows the association rules.
11. Visualize the Results by adding “Data Table” widget to it.

OUTPUT:





RESULT:

The implementation of Scalable Clustering algorithm and Apriori Algorithm using Orange software for data mining is successfully completed and verified.