

FACULTY OF ENGINEERING AND TECHNOLOGY BACHELOR OF TECHNOLOGY

Data Mining and Business Intelligence

(DMBI) (203105454)

VI SEMESTER

Computer Science & Engineering Department





**CERTIFICATE**

*This is to certify that*

*Mr.*  **LUCKY PATHAN** *with Enrollment No.* **210303105790** has *successfully completed his laboratory experiments in the subject (with Code)* **Data Mining and Business Intelligence (203105454)** *from the department of* **Computer Science and Engineering** *during the academic year* ***2022-2023.***



**Date of Submission …..…………… Staff In charge …..……………**

**Head of Department …..……………**

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| **Sr. No** | **Experiment Title** | **Page No** | | **Date of Start** | **Date of Completing** | **Sign** | **Marks**  **(Out of 10)** |
| **From** | **To** |
| 1 | Design and Create cube by identifying measures and dimensions for Star Schema, Snowflake schema and fact Constellation Schema. | 1 | 3 |  |  |  |  |
| 2 | Make an OLAP cube and perform Roll Up and Drill Down operations on it. Show the Apex and Base cuboid for the same. Draw Star-net query model for the cube. | 4 | 5 |  |  |  |  |
| 3 | Create calculated member using arithmetic operators and member property of dimension member. | 6 |  |  |  |  |  |
| 4 | Design and Create cube by identifying measures and dimensions for Design storage using storage mode MOLAP, ROLAP and HOLAP. |  |  |  |  |  |  |
| 5 | Perform Pre-processing on a dataset. Apply various Filters and discuss the effect of each filter applied.   1. Handle Missing Values 2. Handle Infrequent Nominal Values 3. Derive an attribute from the existing attribute 4. Sampling 5. Discretization 6. Use Weka Too 7. Use XL Miner Tool |  |  |  |  |  |  |
| 6 | Perform different binning techniques to smooth out the noise in the dataset. Make sure that the user should have the choice to apply all the possible techniques. Show the output of different bins. Use histogram to partition the dataset into groups. |  |  |  |  |  |  |
| 7 | Perform regression on the data set using R programming. |  |  |  |  |  |  |
| 8 | Perform Association rule mining using WEKA tool. |  |  |  |  |  |  |
| 9 | Perform classification with WEKA tool   1. using Decision Tree Classifier 2. using Naïve Bayes Classifier 3. using Multilayer Perceptron. |  |  |  |  |  |  |
| 10 | Perform Clustering using WEKA too. |  |  |  |  |  |  |

# Practical 1

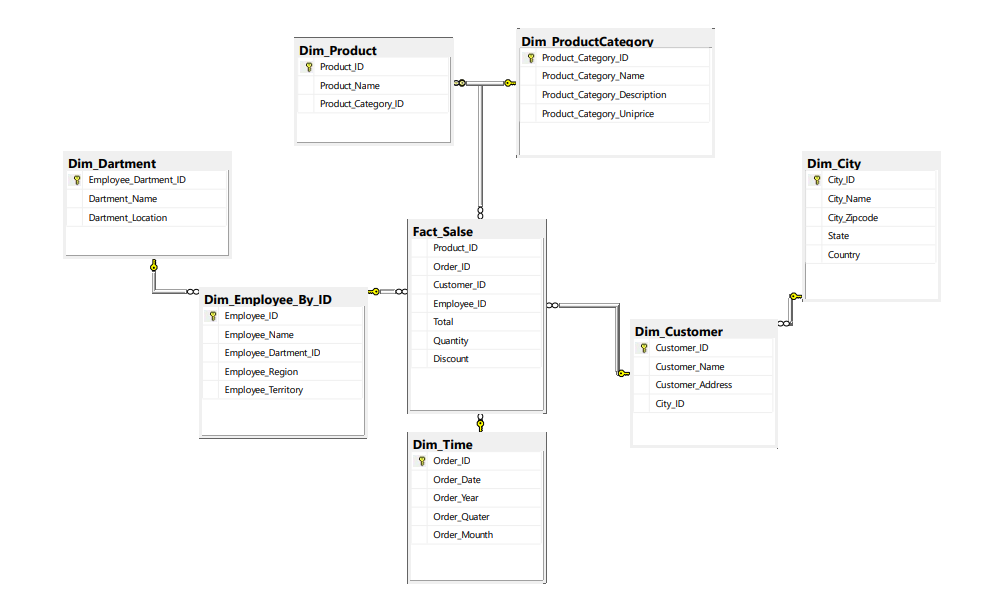
**Aim** : Design and Create cube by identifying measures and dimensions for Star Schema, Snowflake schema and fact Constellation Schema.

**OUTPUT:-**

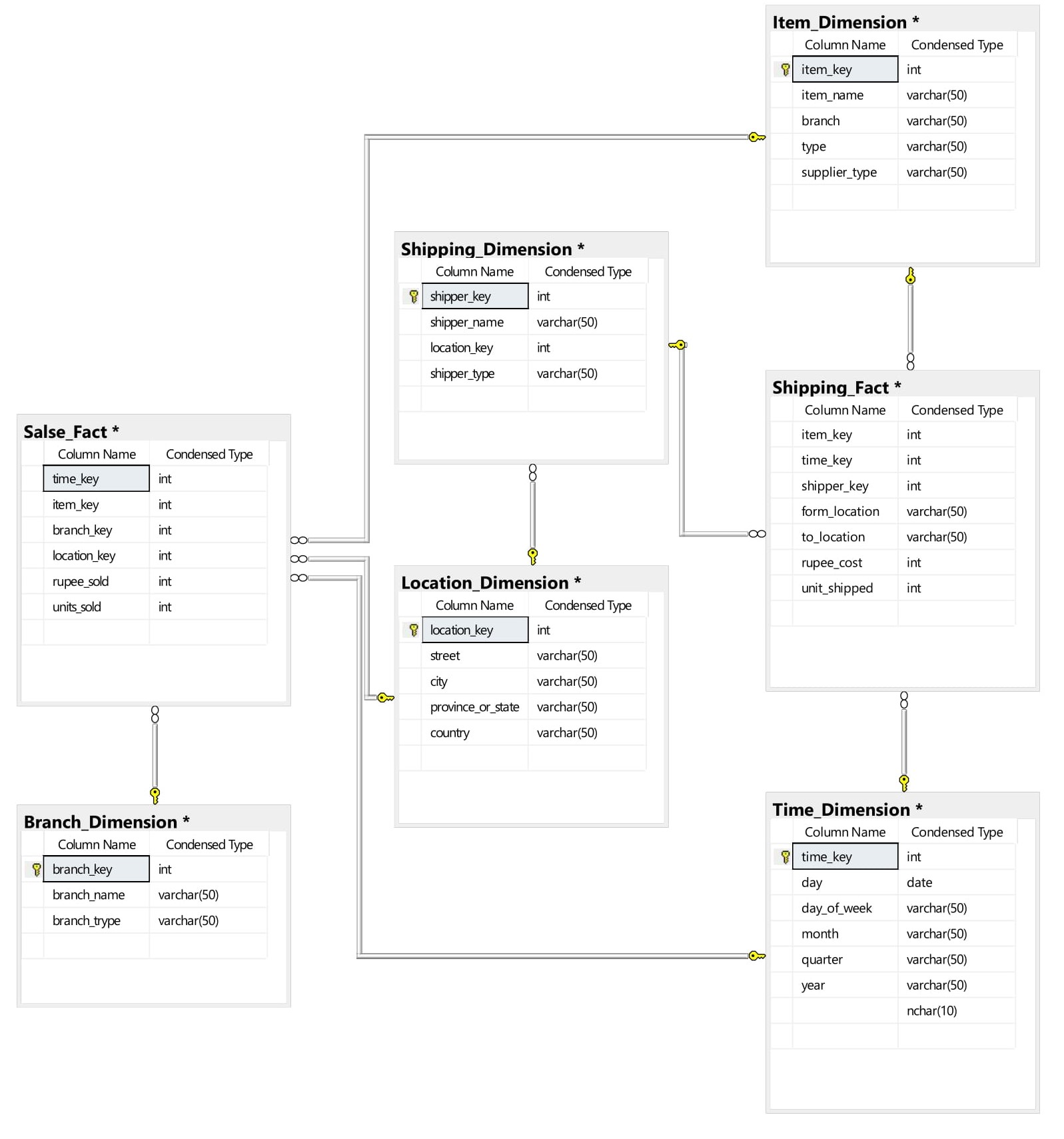
Star Schema -



SnowFlack Schema –



Constellation Schema -



# Practical 2

**Aim** : Make an OLAP cube and perform Roll Up and Drill Down operations on it. Show the Apex and Base cuboid for the same. Draw Star-net query model for the cube.

**CODE:-**

CREATE TABLE STUDENT(ID INT,NAME VARCHAR(50),CITY VARCHAR(30));

INSERT INTO STUDENT VALUES(1,'Lucky','Goghamba');

INSERT INTO STUDENT VALUES(2,'Swet','virpur');

INSERT INTO STUDENT VALUES(3,'Dhaval','Rampura');

INSERT INTO STUDENT VALUES(4,'Vasu','Rajcort');

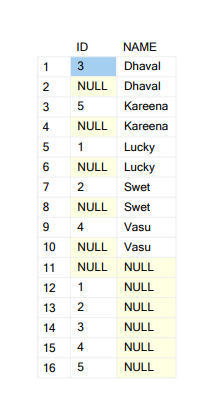
INSERT INTO STUDENT VALUES(5,'Kareena','Mumbai');

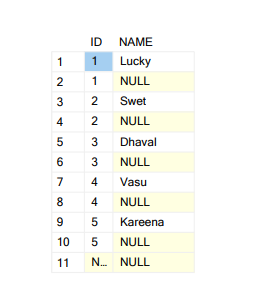
select ID, NAME from STUDENT group by cube(ID,NAME);

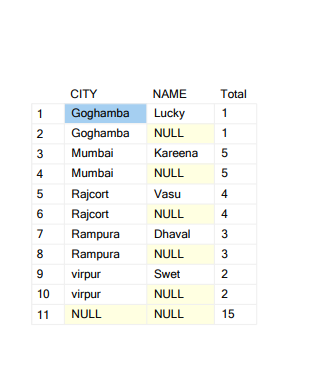
select ID, NAME from STUDENT group by rollup(ID,NAME);

select STUDENT.CITY,STUDENT.NAME ,SUM(STUDENT.ID) Total from STUDENT GROUP BY GROUPING SETS( ([CITY], [NAME]), ([CITY]), ());

**OUTPUT:-**

CUBE:-

ROLLUP:-

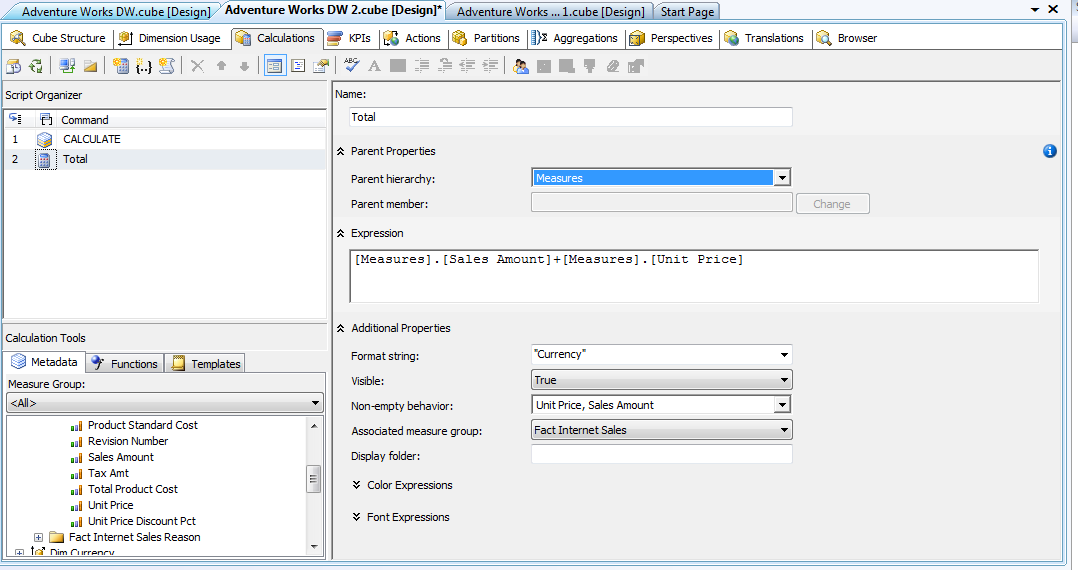
DRILL DOWN :-

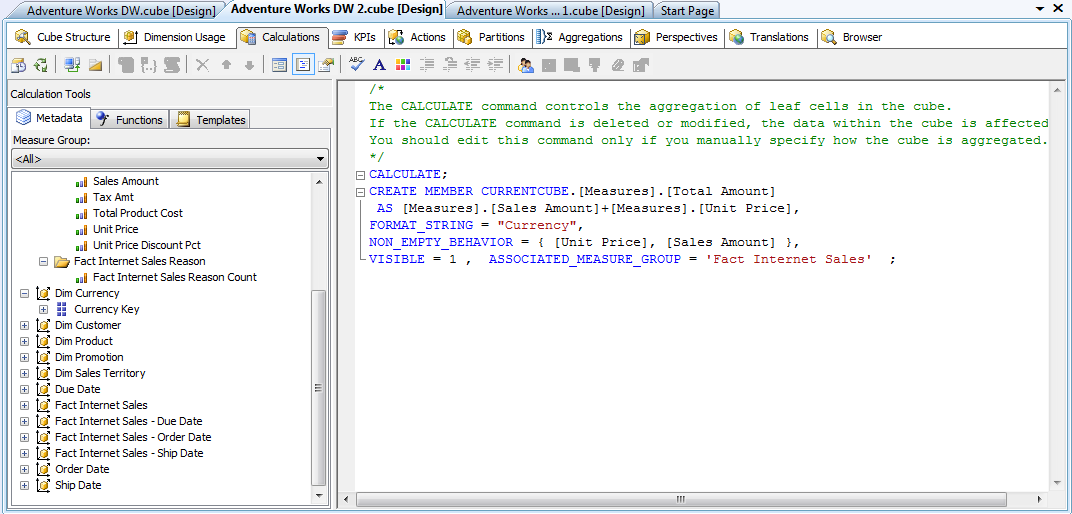
# Practical 3

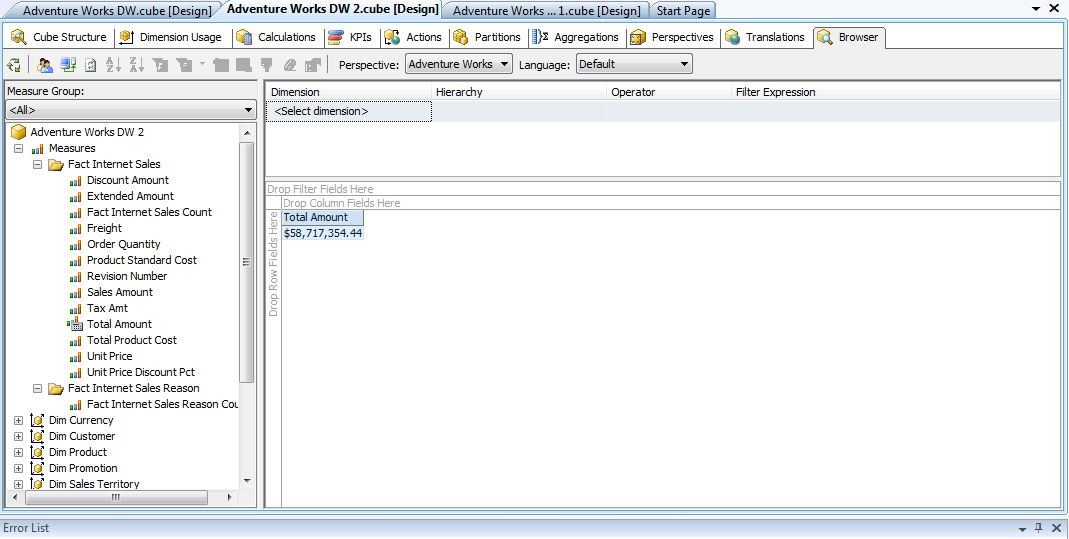
**Aim** : Create calculated member using arithmetic operators and member property of dimension member.

**CODE:-**

1. Open Cube Designer for the Analysis Services Tutorial cube, and then click the Calculations tab.
2. On the toolbar of the Calculations tab, click New Calculated Member. A new form appears in the Calculation Expressions pane within which you define the properties of this new calculated member. The new member also appears in the Script Organizer pane.
3. In the Name box, change the name of the calculated measure to [Total Amount].
4. On the Metadata tab in the Calculation Tools pane of the Calculations tab, expand Measures and then expand Fact Internet Sales to view the metadata for the Fact Internet Sales measure group.
5. Drag SalesAmount from the Metadata tab in the Calculation Tools pane into the Expression box in the Calculation Expressions pane.
6. In the Expression box, type a plus sign (+) after [Measures]. [Sales Amount].
7. In the Format string list, select "Currency".
8. In the Non‐empty behavior list, select the check boxes for Unit Price and Sales Amount then click OK.
9. On the toolbar of the Calculations tab, click Script View, and then review the calculation script in the Calculation Expressions pane.

**OUTPUT:-**

****



# Practical 4

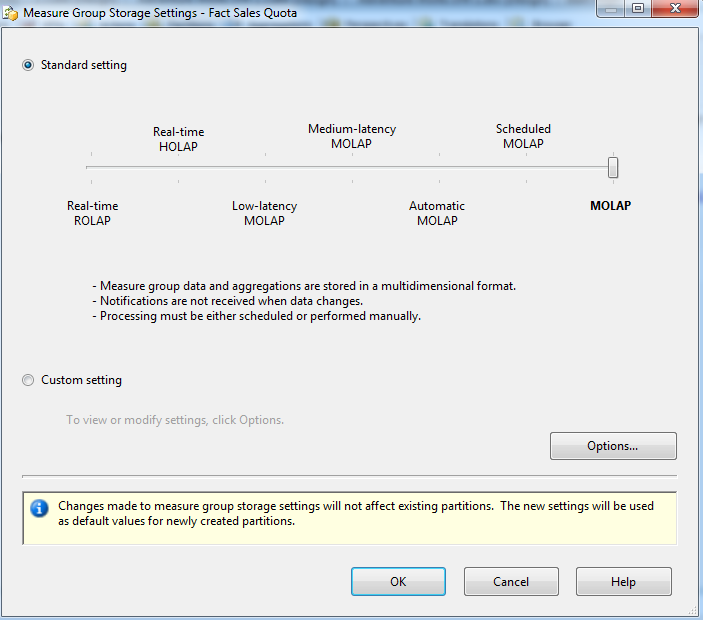
**Aim** : Design and Create cube by identifying measures and dimensions for Design storage using storage mode MOLAP, ROLAP and HOLAP.

**CODE:-**

Microsoft SQL Server Analysis Services provides several standard storage configurations for storage modes and caching options. These provide commonly used configurations for update notification, latency, and rebuilding data.Storage modes are ways to physically store the data in cubes and dimensions.

SQL Server offers three cube types-MOLAP, HOLAP and ROLAP.

User can select the storage mode from the settings.



All the settings are discussed below:

1. Multidimensional OLAP (MOLAP):

MOLAP storage mode uses array-based multidimensional storage engine for multidimensional views of data. With Multidimensional data stores, the storage utilization is low if the data set is sparse. Many MOLAP servers use two levels of data storage representation to handle dense and sparse data sets.

1. Scheduled MOLAP:

Scheduled MOLAP used for a data source when only daily updates are required. Queries are always against data in the MOLAP cache, which is not discarded until a new cache is built and its objects are processed.

1. Automatic MOLAP:

Automatic MOLAP used for a data source when query performance is of key importance. It is automatically processes MOLAP objects whenever required after the latency interval. Queries do not return the most recent data while the new cache is being built and processed.

1. Medium- Latency MOLAP:

Medium-Latency MOLAP used for a data source with frequent (or less frequent) updates when query performance is more important than always providing the most current data. It is automatically processes MOLAP objects whenever required after the latency interval. Performance is slower while the MOLAP objects are being reprocessed.

1. Low-Latency MOLAP:

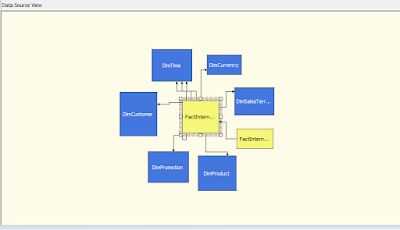
Low-Latency MOLAP used for a data source with frequent updates when query performance is somewhat more important than always providing the most current data. It is automatically processes MOLAP objects whenever required after the latency interval. Performance is slower while the MOLAP objects are being reprocessed.

1. Hybrid Online Analytical Processing (HOLAP):

Hybrid Online Analytical Processing used for a data source with frequent and continuous updates (but not so frequent as to require real-time ROLAP) and users always require the latest data. This method normally provides better overall performance than ROLAP storage. Users can get MOLAP performance from this setting if the data source stays silent long enough.

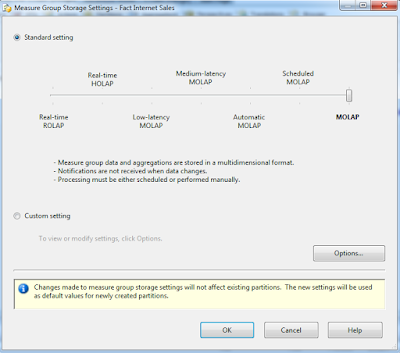
1. Real Time ROLAP (ROLAP):

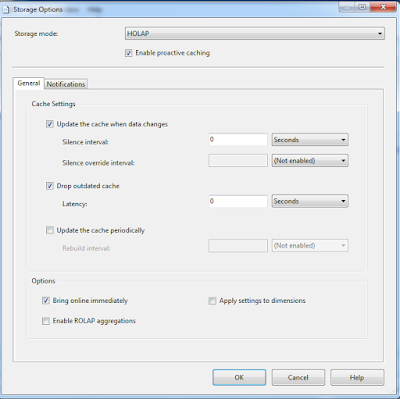
Real time ROLAP used for a data source with very frequent and continuous updates when the very latest data is always required by users. Depending on the types of queries generated by client applications, this method is liable to give the slowest response times.



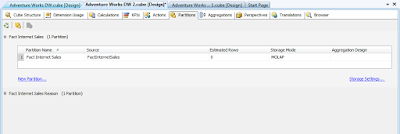
Cube Data Source View

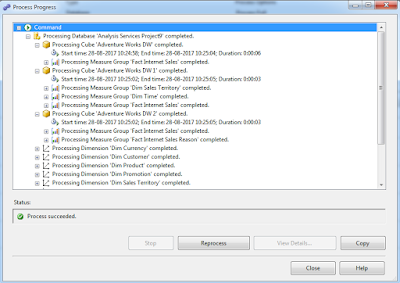
There are two types of storage settings available-

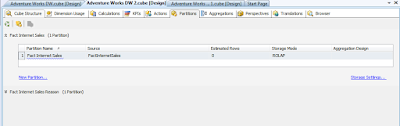
· Standard setting

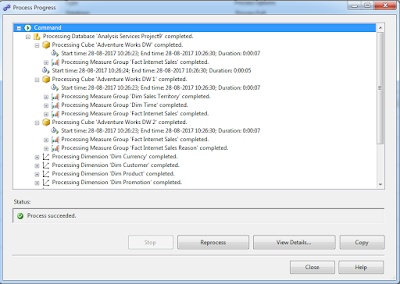
Custom setting

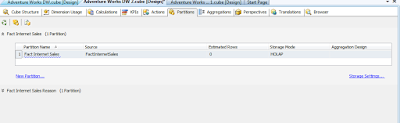
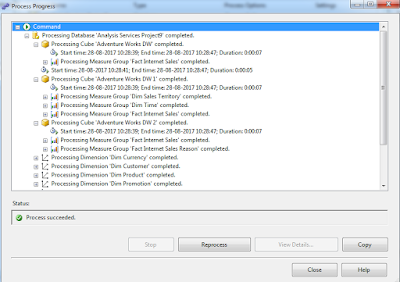
Now we will see the implementation of different Storage Modes on the above experimental Cube Data Source.

1. MOLAP (Multidimensional Online Analytical Processing)

[](https://1.bp.blogspot.com/-8D4B76c5N1o/WbT-AvgpJxI/AAAAAAAAEYo/6rOAoRJOlukUbzOhCaxdT-iW0d-CxwedgCLcBGAs/s1600/6.png)

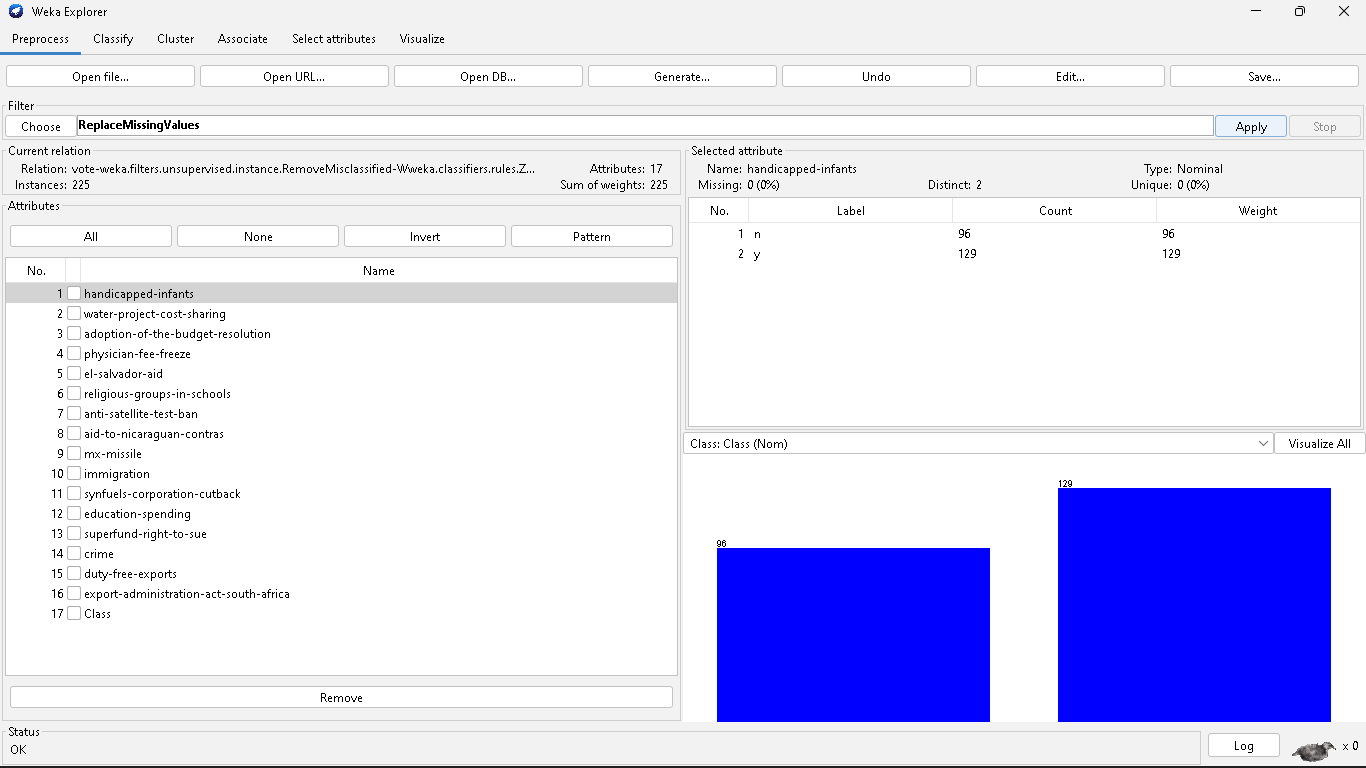
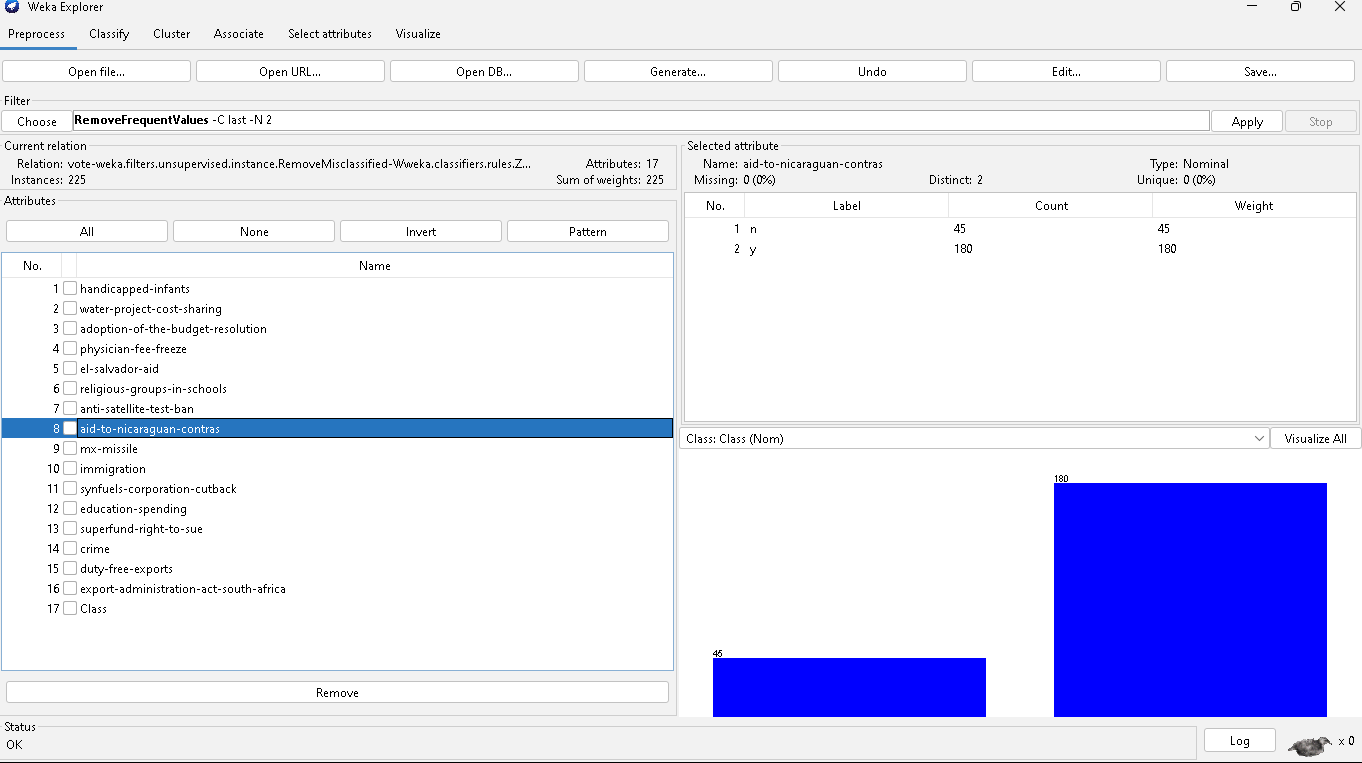
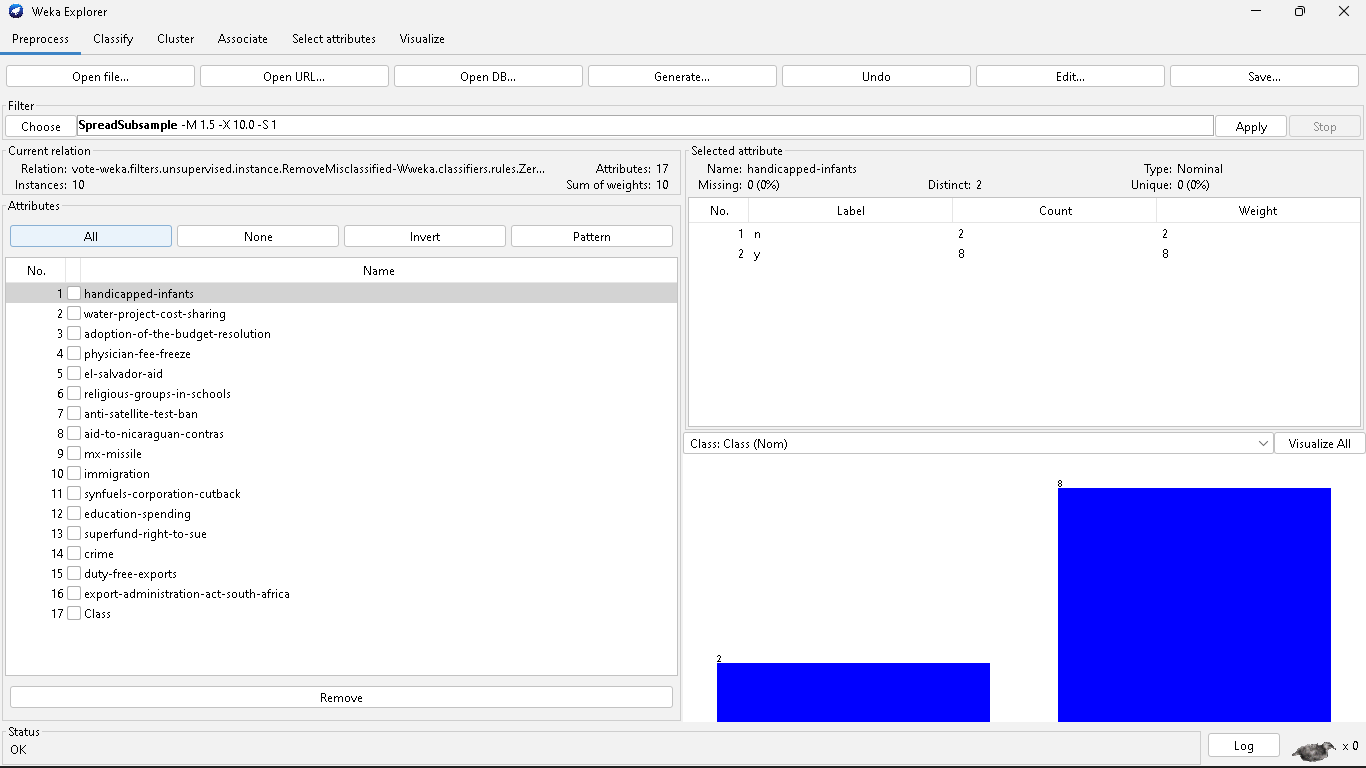
1. ROLAP (Relational Online Analytical Processing)

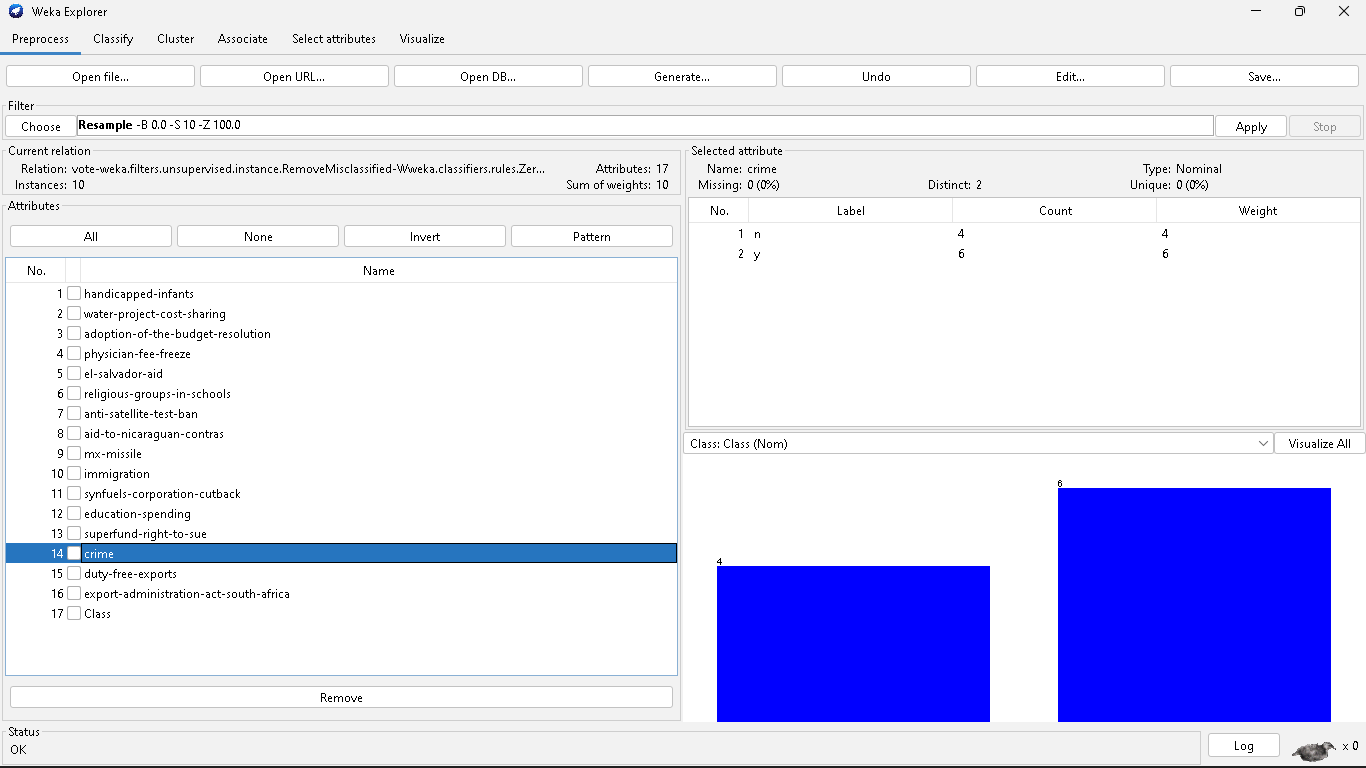


1. HOLAP (Hybrid Online Analytical Processing)

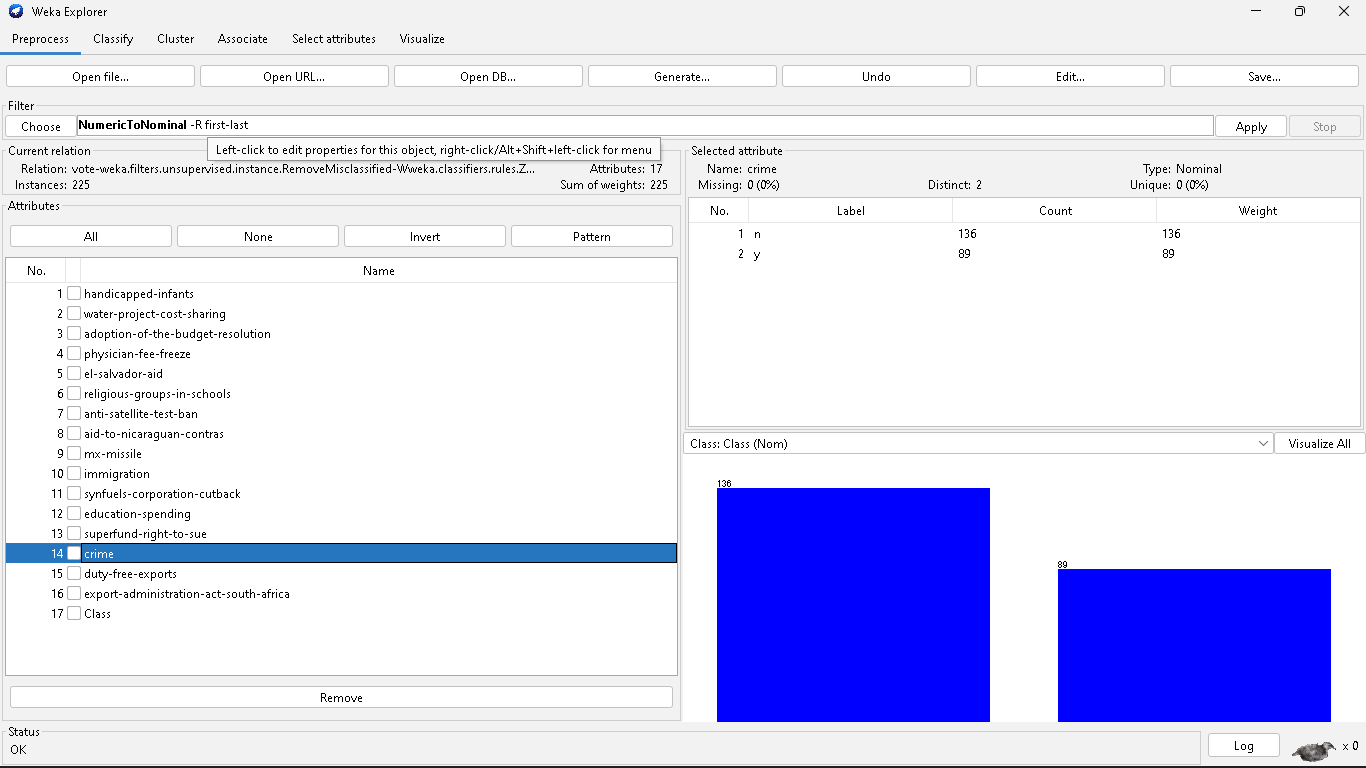
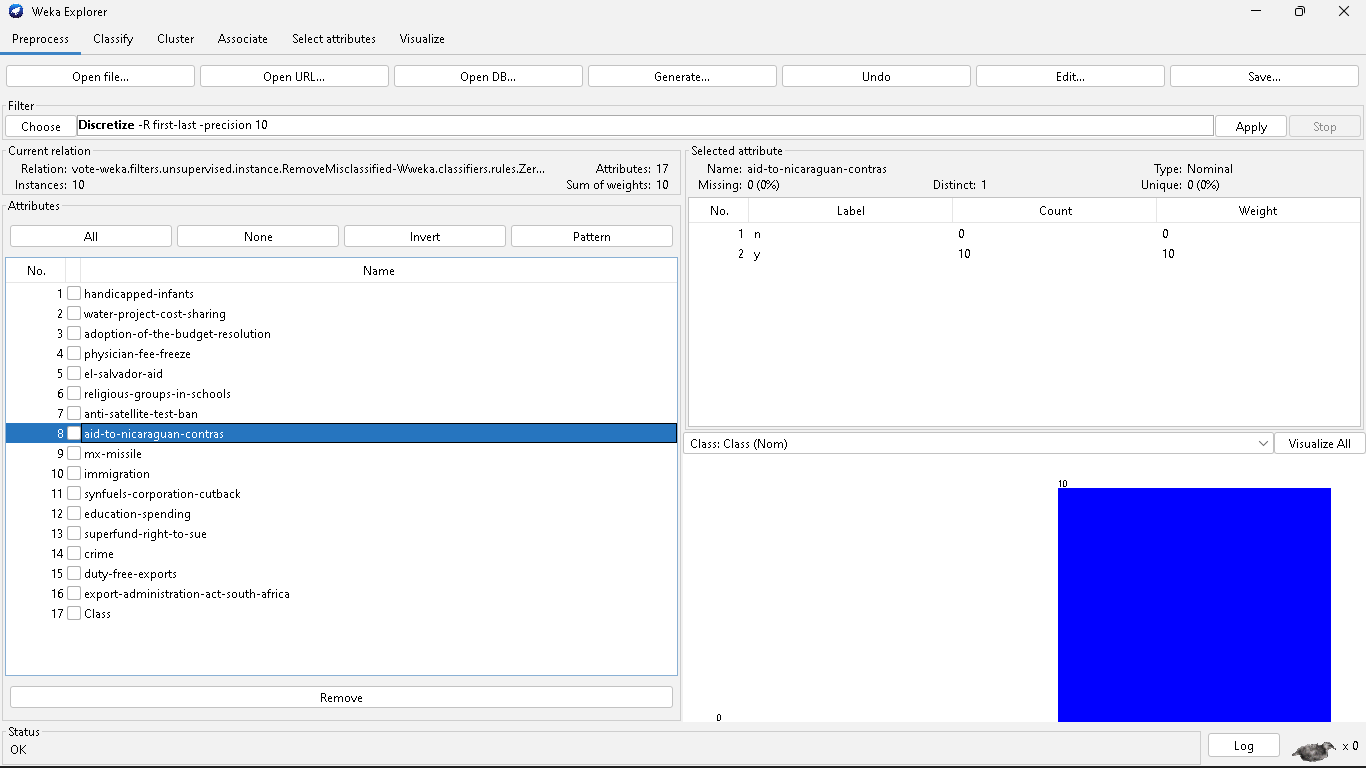
# Practical 5

**Aim** : Perform Pre-processing on a dataset. Apply various Filters and discuss the effect of each filter applied.

1. Handle Missing Values
2. Handle Infrequent Nominal Values
3. Derive an attribute from the existing attribute
4. Sampling



1. Discretization



# Practical 6

**Aim** : Perform different binning techniques to smooth out the noise in the dataset. Make sure that the user should have the choice to apply all the possible techniques. Show the output of different bins. Use histogram to partition the dataset into groups.

**Binning:**

* Binning method is used to smoothing data or to handle noisy data.
* In this method, the data is first sorted and then the sorted values are distributed into a

number of buckets or bins.

* As binning methods consult the neighborhood of values, they perform local

smoothing.

**Approaches to perform smoothing:**

1. Smoothing by bin means:

* In smoothing by bin means, each value in a bin is replaced by the mean value of the bin.

1. Smoothing by bin median:

* In this method each bin value is replaced by its bin median value.

1. Smoothing by bin boundary:

* In smoothing by bin boundaries, the minimum and maximum values in a given bin are identified as the bin boundaries. Each bin value is then replaced by the closest boundary value.

**Example:**

* Sorted data for price (in dollars): 4, 8, 9, 15, 21, 21, 24, 25, 26, 28, 29, 34
* Smoothing by bin means:

Bin 1: 9, 9, 9, 9 Bin 2: 23, 23, 23, 23 Bin 3: 29, 29, 29, 29

* Smoothing by bin boundaries:

Bin 1: 4, 4, 4, 15 Bin 2: 21, 21, 25, 25 Bin 3: 26, 26, 26, 34

* Smoothing by bin median:

Bin 1: 9 9, 9, 9 Bin 2: 24, 24, 24, 24 Bin 3: 29, 29, 29, 29

**Code:**

import numpy as np

import math

from sklearn.datasets import load\_iris

from sklearn import datasets, linear\_model, metrics

dataset = load\_iris()

a = dataset.data

b = np.zeros(150)

for i in range (150):

b[i]=a[i,1]

b=np.sort(b)

bin1=np.zeros((30,5))

bin2=np.zeros((30,5))

bin3=np.zeros((30,5))

for i in range (0,150,5):

k=int(i/5)

mean=(b[i] + b[i+1] + b[i+2] + b[i+3] + b[i+4])/5

for j in range(5):

bin1[k,j]=mean

print("Bin Mean: \n",bin1)

for i in range (0,150,5):

k=int(i/5)

for j in range (5):

if (b[i+j]-b[i]) < (b[i+4]-b[i+j]):

bin2[k,j]=b[i]

else:

bin2[k,j]=b[i+4]

print("Bin Boundaries: \n",bin2)

for i in range (0,150,5):

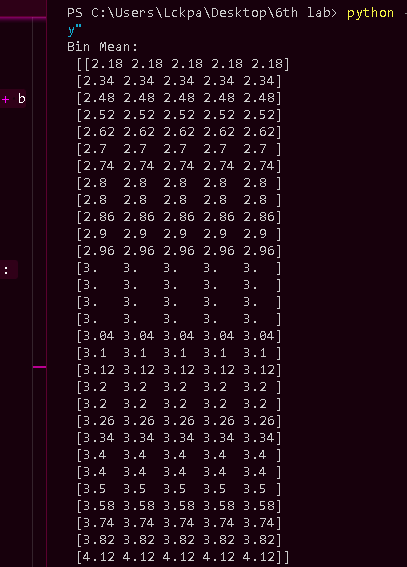
k=int(i/5)

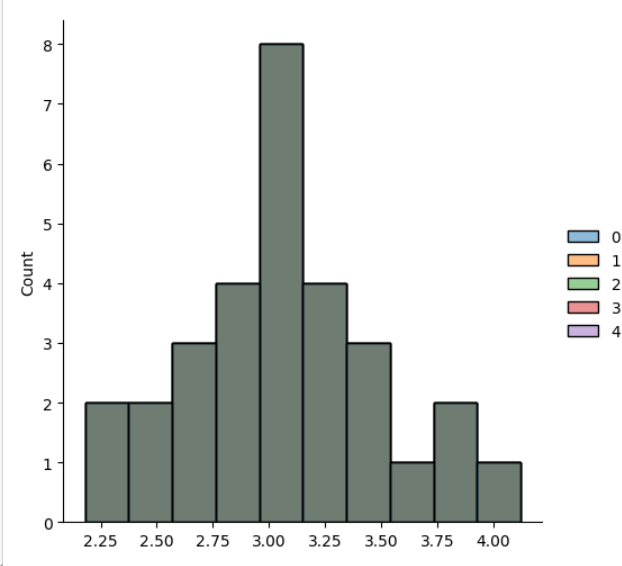
for j in range (5):

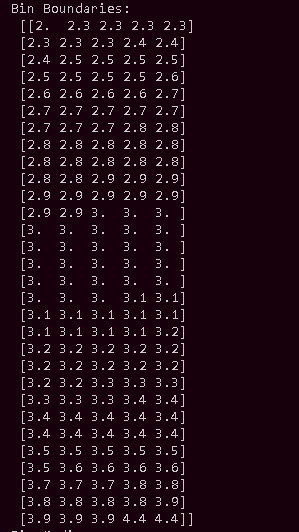
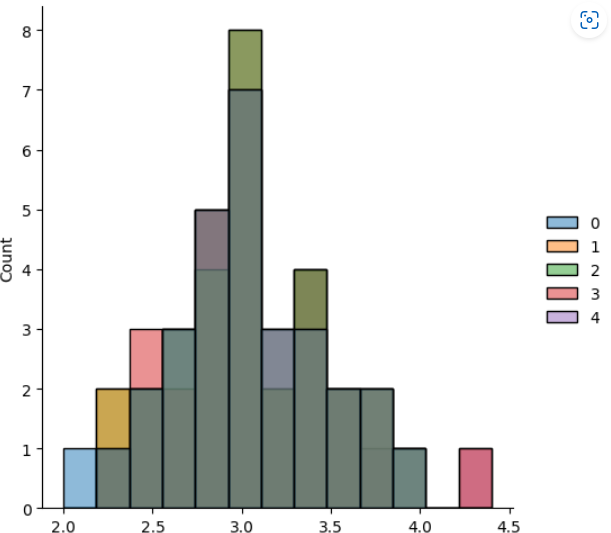
bin3[k,j]=b[i+2]

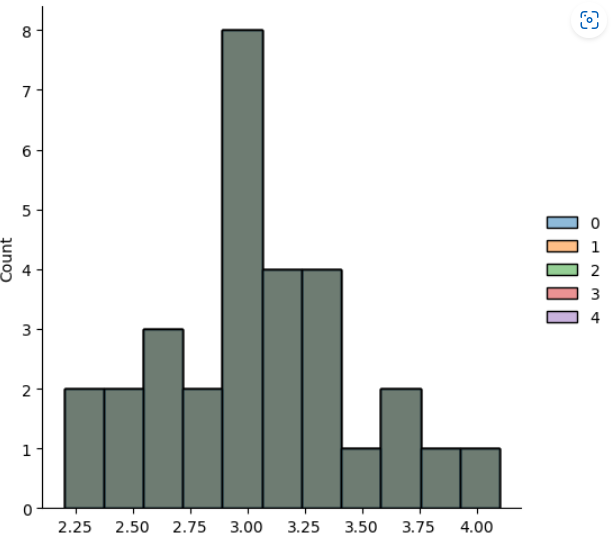
print("Bin Median: \n",bin3)

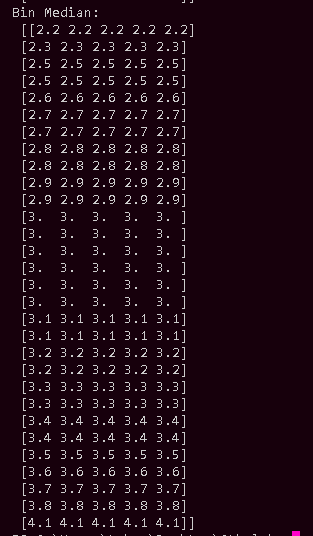
**OUTPUT:**

****



****



****