Experiment 5

Aim: Implementation of Data Discretization (any one) & Visualization (any one).

Theory:

Data discretization: Data discretization refers to a method of converting a huge number of data values into smaller ones so that the evaluation and management of data become easy. In other words, data discretization is a method of converting attributes values of continuous data into a finite set of intervals with minimum data loss. There are two forms of data discretization first is supervised discretization, and the second is unsupervised discretization. Supervised discretization refers to a method in which the class data is used. Unsupervised discretization refers to a method depending upon the way which operation proceeds. It means it works on the top-down splitting strategy and bottom-up merging strategy.

Some of the data discretization techniques are:

- **Histogram analysis**
  Histogram refers to a plot used to represent the underlying frequency distribution of a continuous data set. Histogram assists the data inspection for data distribution. For example, Outliers, skewness representation, normal distribution representation, etc.
- **Binning**
  Binning refers to a data smoothing technique that helps to group a huge number of continuous values into smaller values. For data discretization and the development of idea hierarchy, this technique can also be used.
- **Cluster Analysis**
  Cluster analysis is a form of data discretization. A clustering algorithm is executed by dividing the values of x numbers into clusters to isolate a computational feature of x.
- **Data discretization using decision tree analysis**
  Data discretization refers to a decision tree analysis in which a top-down slicing technique is used. It is done through a supervised procedure. In a numeric attribute discretization, first, you need to select the attribute that has the least entropy, and then you need to run it with the help of a recursive process. The recursive process divides it into various discretized disjoint intervals, from top to bottom, using the same splitting criterion.
- **Data discretization using correlation analysis**
  Discretizing data by linear regression technique, you can get the best neighboring interval, and then the large intervals are combined to develop a larger overlap to form the final 20 overlapping intervals. It is a supervised procedure.

Data Visualization: Data Visualization is used to communicate information clearly and efficiently to users by the usage of information graphics such as tables and charts. It helps users in analyzing a large amount of data in a simpler way. It makes complex data more accessible, understandable, and usable.

Some of the data visualization techniques are:

- Box plots
  A box plot or box and whisker plot give a visual outline of information through its quartiles.
- Histograms
  A histogram is a graphical presentation of information using bars of various heights and in a histogram, each bar groups numbers into ranges.
- Heat maps
  A heatmap has a very different concept of representing the data. It is a graphical portrayal of data that uses different colors to address different values. This difference in color representation makes it easy for the viewers to understand the trend more quickly.
- Charts
  The easiest way to show the development of one or several data sets is a chart. Charts vary from bar and line charts that show the relationship between elements over time to pie charts that demonstrate the components or proportions between the elements of one whole.
  There are several types of charts:
  1. Bar Charts
  2. Line Charts
  3. Pie Charts
  4. Scatter Charts

Implementation (Binning):

```
import numpy as np
from matplotlib import pyplot
from sklearn.datasets import load_iris

# load iris data set
dataset = load_iris()
a = dataset.data
b = np.zeros(150)

# take 1st column among 4 column of data set
for i in range(150):
    b[i] = a[i, 1]
```

```
b = np.sort(b)  # sort the array

# create bins
bin1 = np.zeros((30, 5))
bin2 = np.zeros((30, 5))
bin3 = np.zeros((30, 5))

# Bin mean
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)

# Bin boundaries
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)

# Bin median
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)

# Show histogram for bin1
pyplot.hist(bin1, bins=10)
pyplot.show()

# Show histogram for bin2
pyplot.hist(bin2, bins=10)
pyplot.show()

# Show histogram for bin3
pyplot.hist(bin3, bins=10)
pyplot.show()
```

Output:

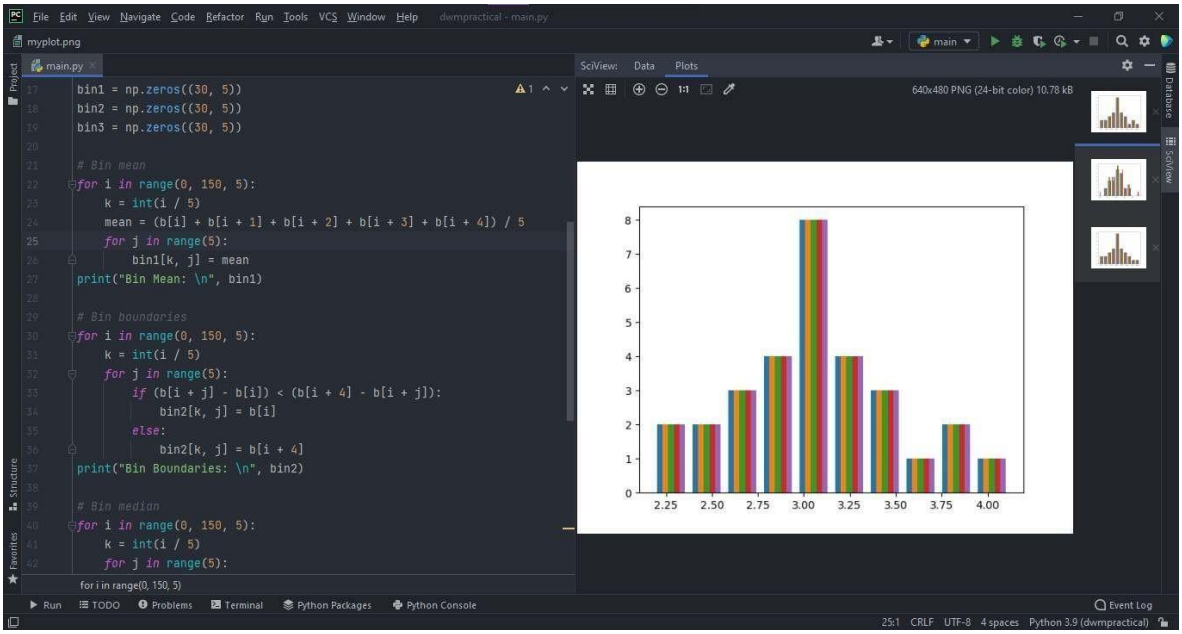
```
Bin Mean:
 [[2.18 2.18 2.18 2.18 2.18]
 [2.34 2.34 2.34 2.34 2.34]
 [2.48 2.48 2.48 2.48 2.48]
 [2.52 2.52 2.52 2.52 2.52]
 [2.62 2.62 2.62 2.62 2.62]
 [2.7  2.7  2.7  2.7  2.7 ]
 [2.74 2.74 2.74 2.74 2.74]
```

```
 [2.8  2.8  2.8  2.8  2.8 ]
 [2.8  2.8  2.8  2.8  2.8 ]
 [2.86 2.86 2.86 2.86 2.86]
 [2.9  2.9  2.9  2.9  2.9 ]
 [2.96 2.96 2.96 2.96 2.96]
 [3.   3.   3.   3.   3.  ]
 [3.   3.   3.   3.   3.  ]
 [3.   3.   3.   3.   3.  ]
 [3.   3.   3.   3.   3.  ]
 [3.04 3.04 3.04 3.04 3.04]
 [3.1  3.1  3.1  3.1  3.1 ]
 [3.12 3.12 3.12 3.12 3.12]
 [3.2  3.2  3.2  3.2  3.2 ]
 [3.2  3.2  3.2  3.2  3.2 ]
 [3.26 3.26 3.26 3.26 3.26]
 [3.34 3.34 3.34 3.34 3.34]
 [3.4  3.4  3.4  3.4  3.4 ]
 [3.4  3.4  3.4  3.4  3.4 ]
 [3.5  3.5  3.5  3.5  3.5 ]
 [3.58 3.58 3.58 3.58 3.58]
 [3.74 3.74 3.74 3.74 3.74]
 [3.82 3.82 3.82 3.82 3.82]
 [4.12 4.12 4.12 4.12 4.12]]
Bin Boundaries:
 [[2.  2.3 2.3 2.3 2.3]
 [2.3 2.3 2.3 2.4 2.4]
 [2.4 2.5 2.5 2.5 2.5]
 [2.5 2.5 2.5 2.5 2.6]
 [2.6 2.6 2.6 2.6 2.7]
 [2.7 2.7 2.7 2.7 2.7]
 [2.7 2.7 2.7 2.8 2.8]
 [2.8 2.8 2.8 2.8 2.8]
 [2.8 2.8 2.8 2.8 2.8]
 [2.8 2.8 2.9 2.9 2.9]
 [2.9 2.9 2.9 2.9 2.9]
 [2.9 2.9 3.  3.  3. ]
 [3.  3.  3.  3.  3. ]
 [3.  3.  3.  3.  3. ]
 [3.  3.  3.  3.  3. ]
 [3.  3.  3.  3.  3. ]
 [3.  3.  3.  3.1 3.1]
 [3.1 3.1 3.1 3.1 3.1]
 [3.1 3.1 3.1 3.1 3.2]
 [3.2 3.2 3.2 3.2 3.2]
 [3.2 3.2 3.2 3.2 3.2]
 [3.2 3.2 3.3 3.3 3.3]
 [3.3 3.3 3.3 3.4 3.4]
 [3.4 3.4 3.4 3.4 3.4]
 [3.4 3.4 3.4 3.4 3.4]
 [3.5 3.5 3.5 3.5 3.5]
 [3.5 3.6 3.6 3.6 3.6]
 [3.7 3.7 3.7 3.8 3.8]
 [3.8 3.8 3.8 3.8 3.9]
 [3.9 3.9 3.9 4.4 4.4]]
Bin Median:
 [[2.2 2.2 2.2 2.2 2.2]
 [2.3 2.3 2.3 2.3 2.3]
```

```
[2.5 2.5 2.5 2.5 2.5]
[2.5 2.5 2.5 2.5 2.5]
[2.6 2.6 2.6 2.6 2.6]
[2.7 2.7 2.7 2.7 2.7]
[2.7 2.7 2.7 2.7 2.7]
[2.8 2.8 2.8 2.8 2.8]
[2.8 2.8 2.8 2.8 2.8]
[2.9 2.9 2.9 2.9 2.9]
[2.9 2.9 2.9 2.9 2.9]
[3.  3.  3.  3.  3. ]
[3.  3.  3.  3.  3. ]
[3.  3.  3.  3.  3. ]
[3.  3.  3.  3.  3. ]
[3.  3.  3.  3.  3. ]
[3.  3.  3.  3.  3. ]
[3.1 3.1 3.1 3.1 3.1]
[3.1 3.1 3.1 3.1 3.1]
[3.2 3.2 3.2 3.2 3.2]
[3.2 3.2 3.2 3.2 3.2]
[3.3 3.3 3.3 3.3 3.3]
[3.3 3.3 3.3 3.3 3.3]
[3.4 3.4 3.4 3.4 3.4]
[3.4 3.4 3.4 3.4 3.4]
[3.5 3.5 3.5 3.5 3.5]
[3.6 3.6 3.6 3.6 3.6]
[3.7 3.7 3.7 3.7 3.7]
[3.8 3.8 3.8 3.8 3.8]
[4.1 4.1 4.1 4.1 4.1]]
```

Visualization (Histogram):

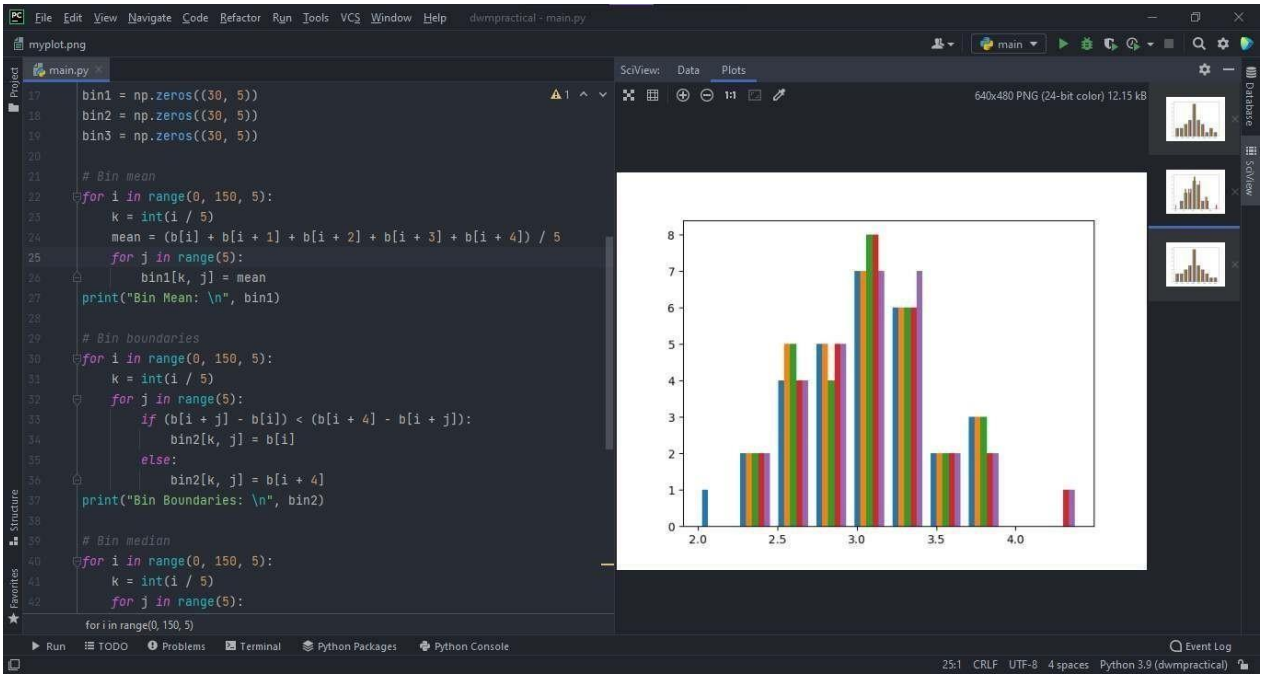
Histogram of Bin1:



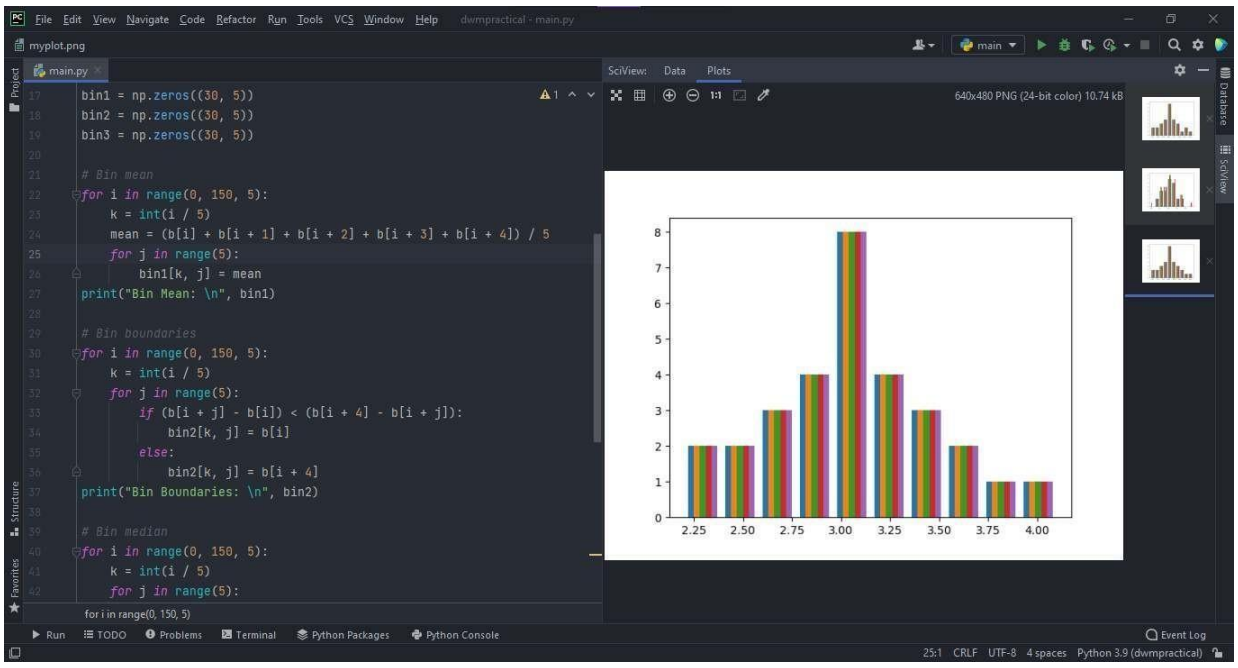
Conclusion:-

The experiment had successful implementation of Bayesian Theorem.

Histogram of Bin2:



Histogram of Bin3:



Conclusion:-

The experiment had successful Implementation of Data Discretization & Visualization.