Thomas Koutsidis
Final Project
07/29/22

This project implements the k-means algorithm for the Wisconsin Breast Cancer data set while using Python. The project fills in missing values, adds a predicted class and then applies an error statement to the predicted class.

Phase 1 consisted of downloading the breast cancer data into Python. Certain values were missing in column A7 and needed to be replaced to further analyze. To replace the missing values, the mean imputation method was used. This was done by using the fillna() function. Once the missing values were replaced, the mean, median, standard deviation and variance of each of the attributes A2 to A10 were found using built in Python functions. Each column's results were rounded to one decimal place and printed.

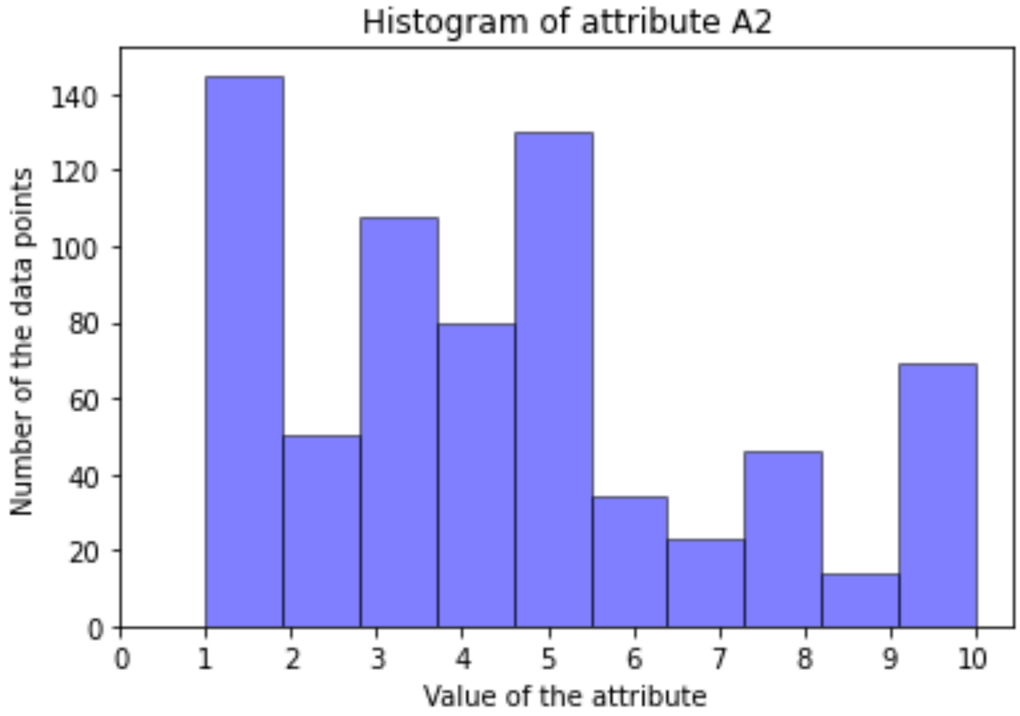
In Phase 2, k-means algorithm was implemented. Using the dataset used and adjusted in Phase 1, we used k-means computation on columns A2 to A10. In this phase, two initial centroids were to be chosen at random, which is reflected by the programming. Once the centroids were chosen, they are displayed with their values from column A2 to A10. A new column is created called the Predicted Class. Each of the 699 data points were computed for their Euclidian distance from the initial centroids. Each point would fall into one of the two predicted clusters (or class). If the distance of the data point was closer from mu2 to mu4, it would be assigned to Predicted Class = 2, and vice versa. The phase would assign each data point to a cluster, then update the centroids. This would happen until the centroids did not change from their previous iteration or until the steps were iterated 50 times. The results were then printed.

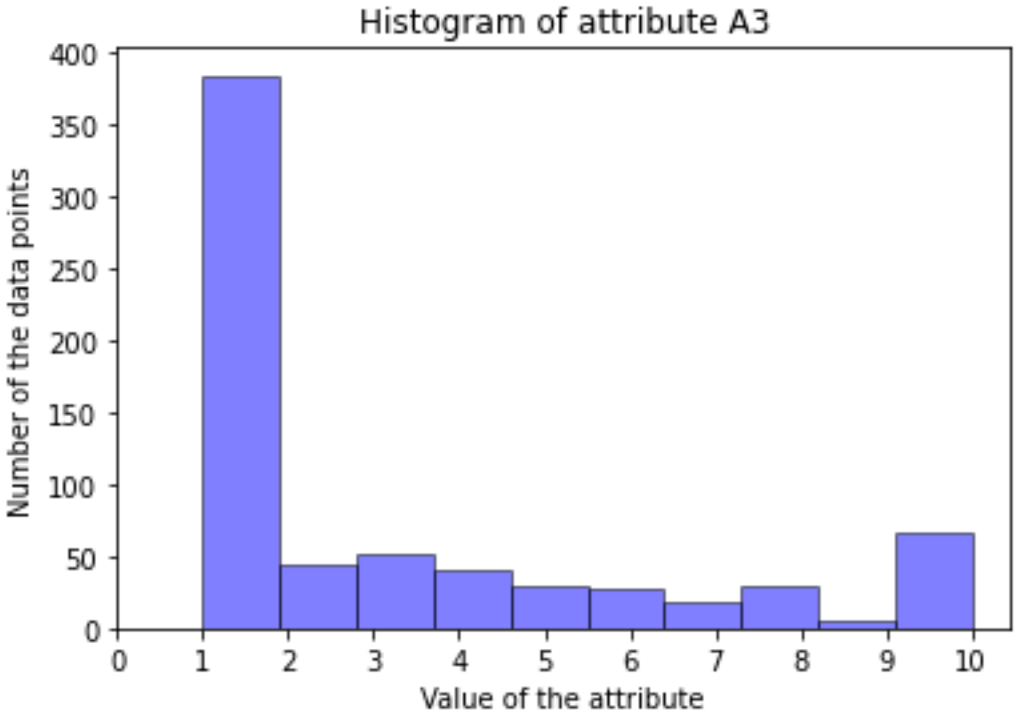
In Phase 3, the quality of the clustering was analyzed. This was done by calculating the error rate of the clusters. There were two clusters, benign and malign cells. This phase found the error rate for the benign cells, malign cells, and total error rate. Using the definitions and formulae stated in the instructions, we can calculate the error rates. The results for each phase are shown below.

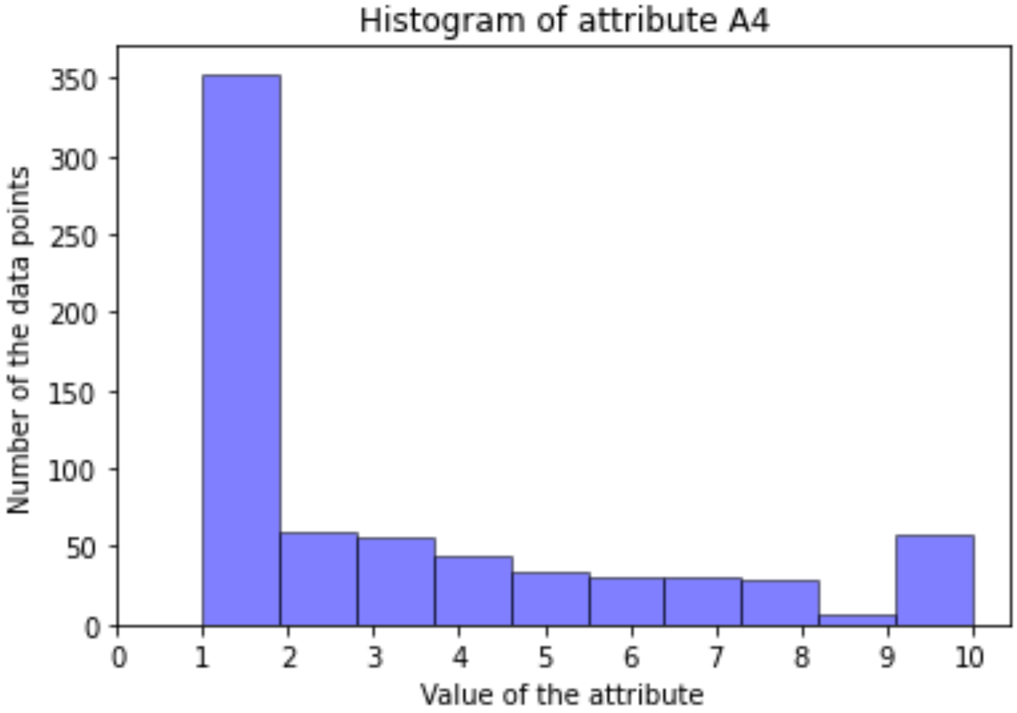
Phase 1

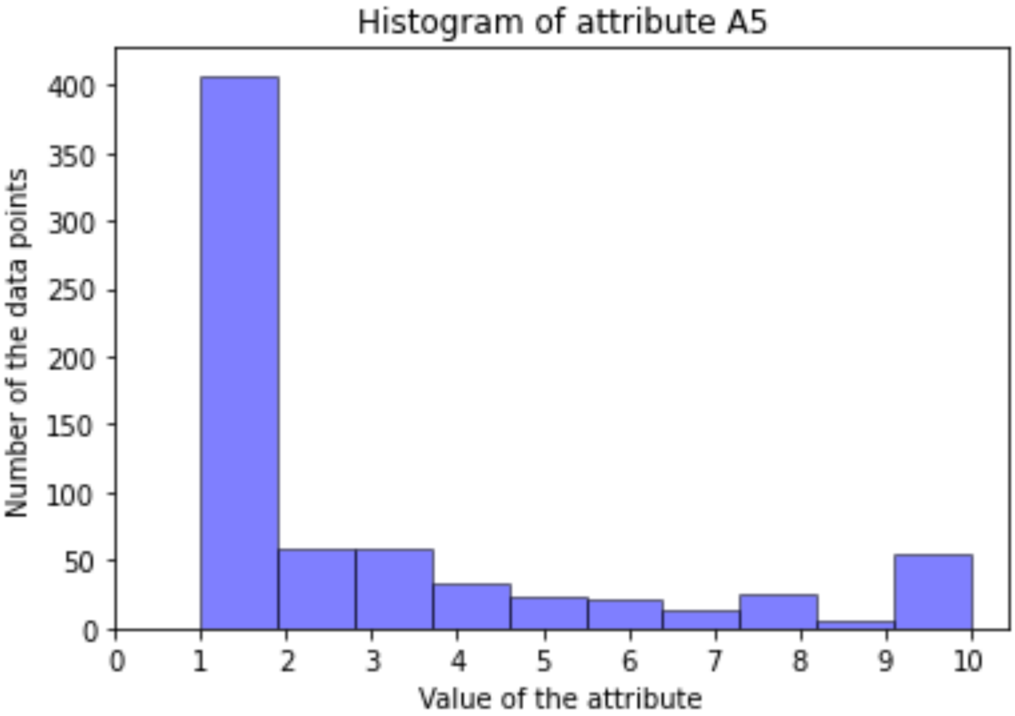
```
Attribute A2
------------
Mean: 4.4
Median: 4.0
Variance: 7.9
Standard Deviation: 2.8

Attribute A3
------------
Mean: 3.1
Median: 1.0
Variance: 9.3
Standard Deviation: 3.1

Attribute A4
------------
Mean: 3.2
Median: 1.0
Variance: 8.8
Standard Deviation: 3.0

Attribute A5
------------
Mean: 2.8
Median: 1.0
Variance: 8.2
Standard Deviation: 2.9

Attribute A6
------------
Mean: 3.2
Median: 2.0
Variance: 4.9
Standard Deviation: 2.2

Attribute A7
------------
Mean: 3.5
```

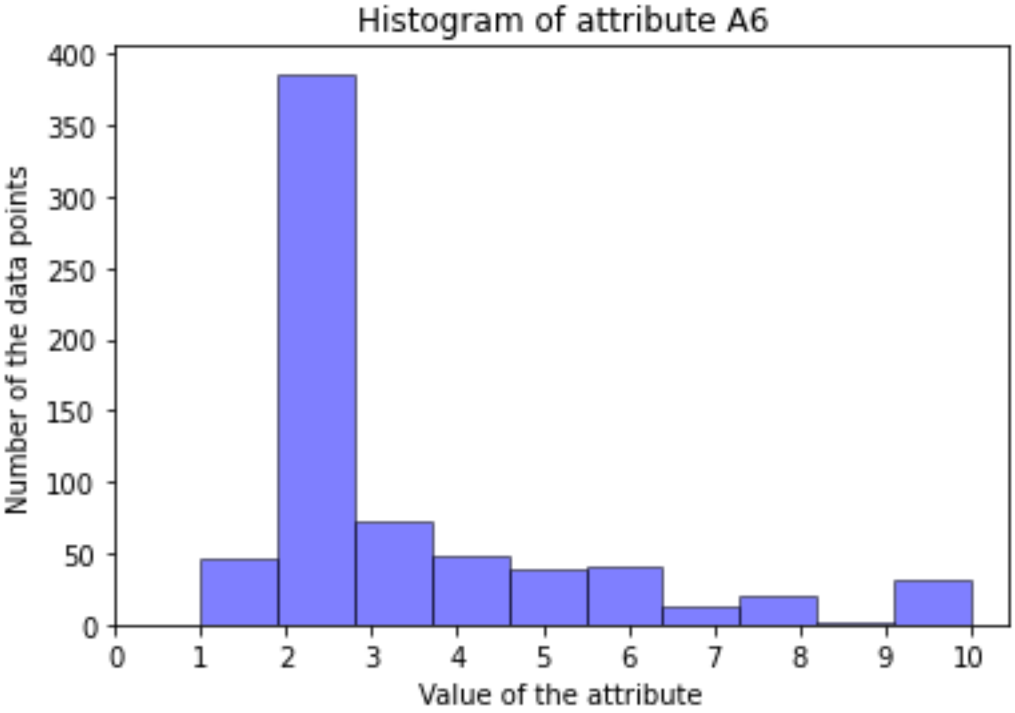
```
Median: 1.0
Variance: 13.0 Standard Deviation: 3.6

Attribute A8 ------------

Mean: 3.4
Median: 3.0
Variance: 5.9
Standard Deviation: 2.4

Attribute A9
------------
Mean: 2.9
Median: 1.0
Variance: 9.3
Standard Deviation: 3.1

Attribute A10 -------------
Mean: 1.6
Median: 1.0
Variance: 2.9
Standard Deviation: 1.7
```

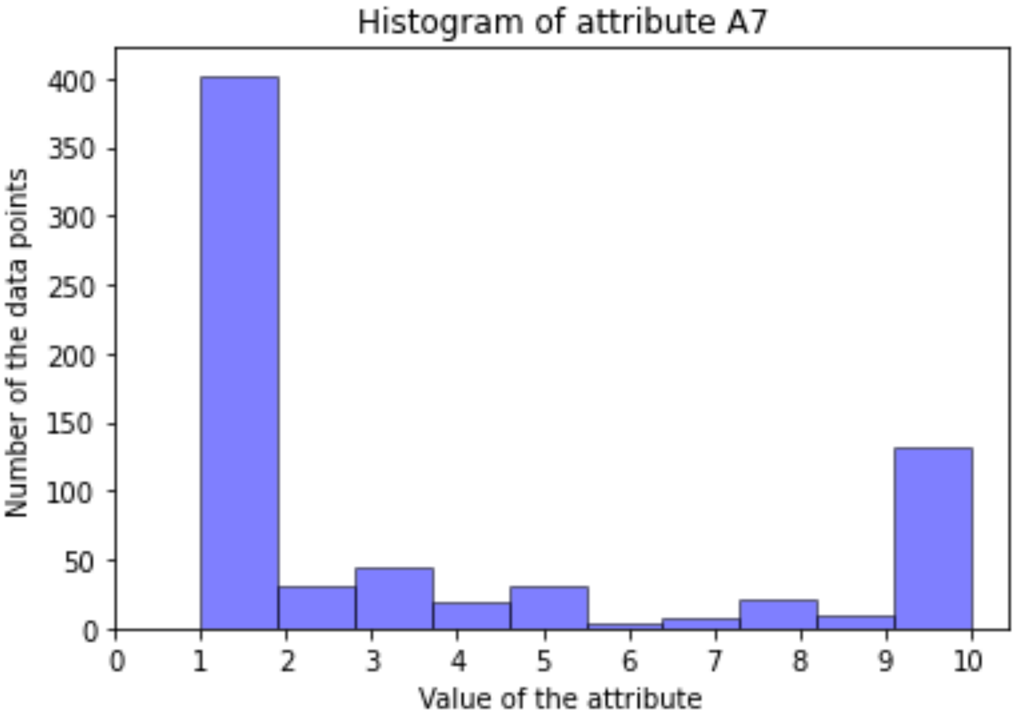


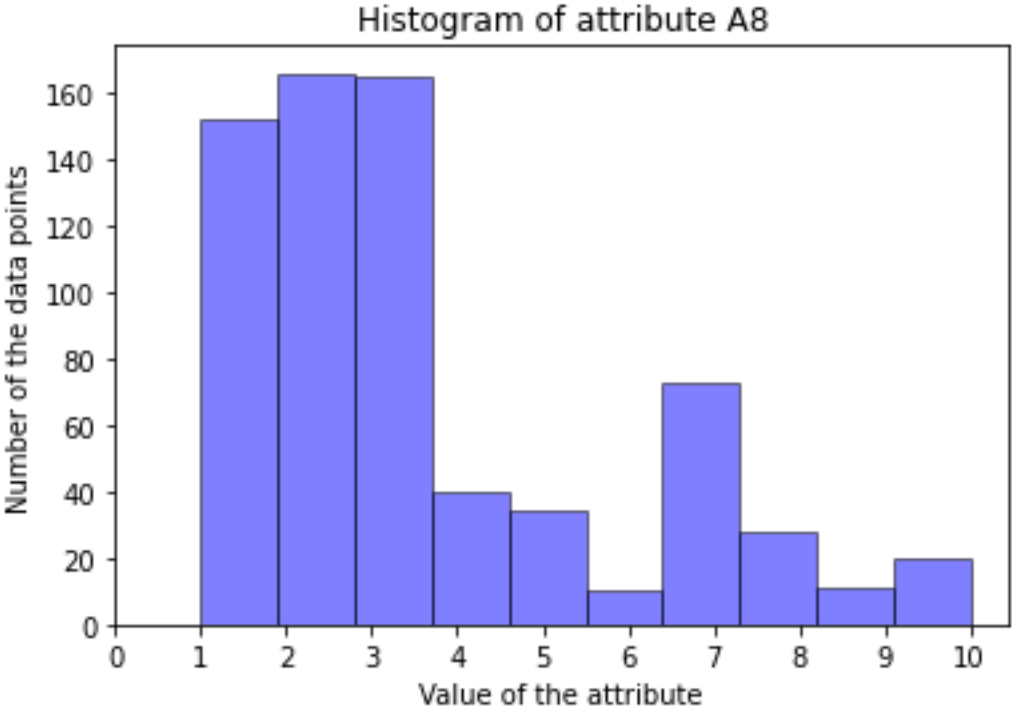

Histogram of attribute A3
Number of the data points
400
350
300
250
200
150
100
50
0
0
1
2
3
4
5
6
7
8
9
10
Value of the attribute

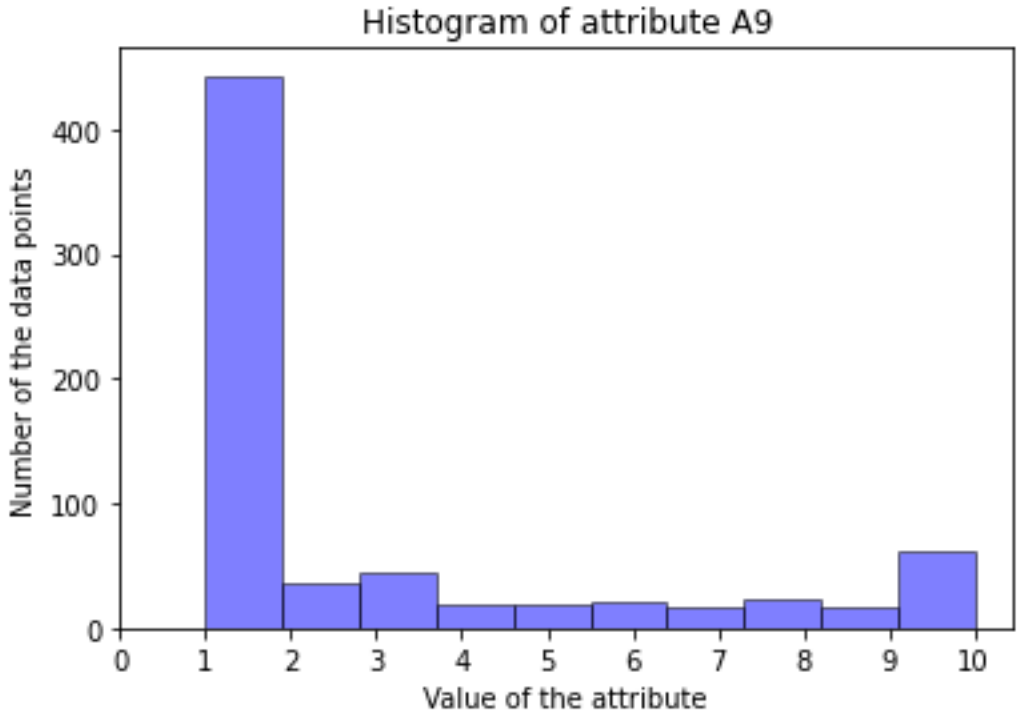

Histogram of attribute A4
Number of the data points
350
300
250
200
150
100
50
0
0
1
2
3
4
5
6
7
8
9
10
Value of the attribute

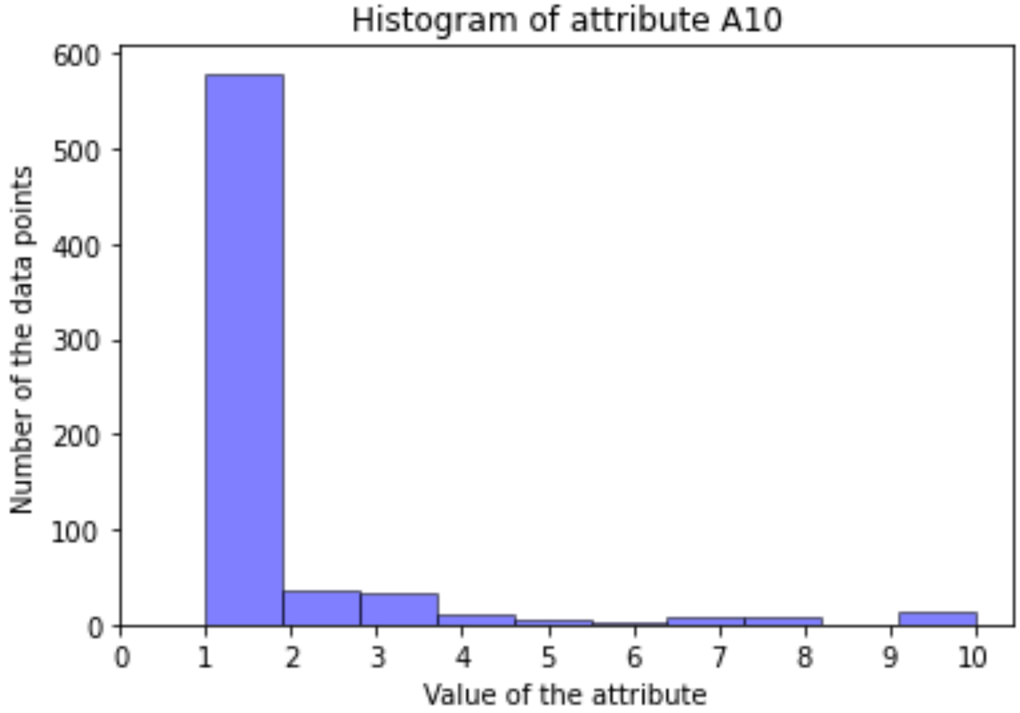

Histogram of attribute A5
Number of the data points
0
50
100
150
200
250
300
350
400
Value of the attribute
0
1
2
3
4
5
6
7
8
9
10


Histogram of attribute A6
Number of the data points
0
50
100
150
200
250
300
350
400
Value of the attribute
0
1
2
3
4
5
6
7
8
9
10


Histogram of attribute A7
Number of the data points
0
50
100
150
200
250
300
350
400
Value of the attribute
0
1
2
3
4
5
6
7
8
9
10


Histogram of attribute A8
Number of the data points
0
20
40
60
80
100
120
140
160
Value of the attribute
0
1
2
3
4
5
6
7
8
9
10


Histogram of attribute A9
Number of the data points
Value of the attribute
0
100
200
300
400
0
1
2
3
4
5
6
7
8
9
10


Histogram of attribute A10
Number of the data points
Value of the attribute
0
100
200
300
400
500
600
0
1
2
3
4
5
6
7
8
9
10

Phase 2

```
Randomly selected row 310 for centroid mu_2.

Initial centroid mu_2:
A2     2.0
A3     1.0
A4     1.0
A5     1.0
A6     3.0
A7     1.0
A8     2.0
A9     1.0
A10    1.0
Name: 310, dtype: float64

Randomly selected row 592 for centroid mu_4.

Initial centroid mu_4:
A2     10.0
A3      3.0
A4      4.0
A5      5.0
A6      3.0
A7     10.0
A8      4.0
A9      1.0
A10     1.0
Name: 592, dtype: float64

Program ended after 4 iterations.

Final centroid mu_2:
A2   3.0472103004291844
A3   1.3025751072961373
A4   1.446351931330472
A5   1.3433476394849786
A6   2.087982832618026
A7   1.3800011310866602
A8   2.1051502145922747
A9   1.261802575107296
A10   1.109442060085837

Final centroid mu_4:
A2   7.1587982832618025
A3   6.798283261802575
A4   6.7296137339055795
A5   5.733905579399142
```

```
A6   5.472103004291846
A7   7.873965526992126
A8   6.103004291845494
A9   6.07725321888412
A10   2.5493562231759657

Final Cluster Assignment:
       Scn  Class  Predicted Class
0   1000025      2              2
1   1002945      2              4
2   1015425      2              2
3   1016277      2              4
4   1017023      2              2
5   1017122      4              4
6   1018099      2              2
7   1018561      2              2
8   1033078      2              2
9   1033078      2              2
10  1035283      2              2
11  1036172      2              2
12  1041801      4              2
13  1043999      2              2
14  1044572      4              4
15  1047630      4              2
16  1048672      2              2
17  1049815      2              2
18  1050670      4              4
19  1050718      2              2
```

Phase 3

```
Total errors: 4.3 %

Data points in Predicted Class 2: 466

Data points in Predicted Class 4: 233

Error data points, Predicted Class 2:
         Scn  Class  Predicted Class
12   1041801      4              2
15   1047630      4              2
23   1057013      4              2
25   1065726      4              2
50   1108370      4              2
51   1108449      4              2
57   1113038      4              2
59   1113906      4              2
63   1116132      4              2
```

```
65   1116998     4             2
101  1167439     4             2
103  1168359     4             2
105  1169049     4             2
222  1226012     4             2
273   428903     4             2
348   832226     4             2
356   859164     4             2
455  1246562     4             2
489  1084139     4             2

Error data points, Predicted Class 4:
        Scn  Class  Predicted Class
1    1002945     2             4
3    1016277     2             4
40   1096800     2             4
196  1213375     2             4
252  1017023     2             4
259   242970     2             4
296   616240     2             4
315   704168     2             4
319   721482     2             4
352   846832     2             4
434  1293439     2             4

Number of all data points: 699

Number of error points: 30

Error rate for class 2: 4.1 %
Error rate for class 4: 4.7 %
Total error rate: 4.3 %
```

As shown by the total error rate, the k-means algorithm was a close prediction at a rate of 4.3% total error.