Michael Uftring
I590 - Python
Summer 2017

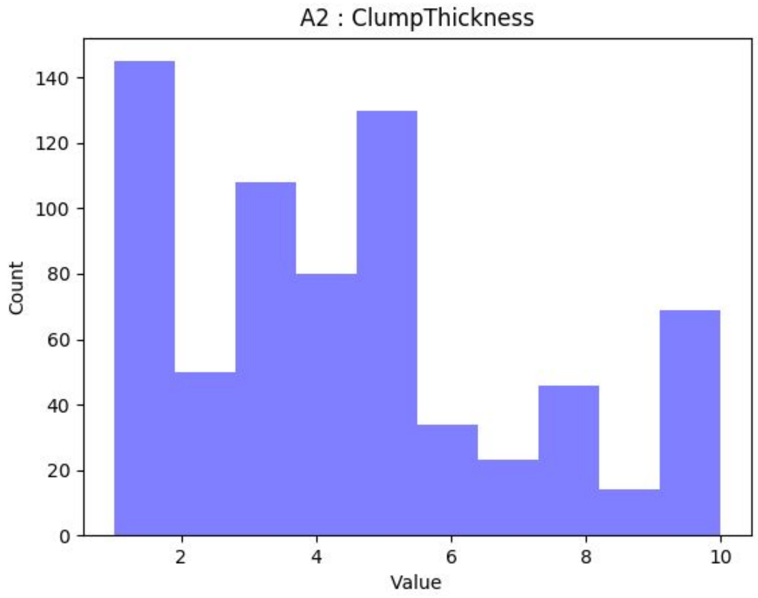
Final Project

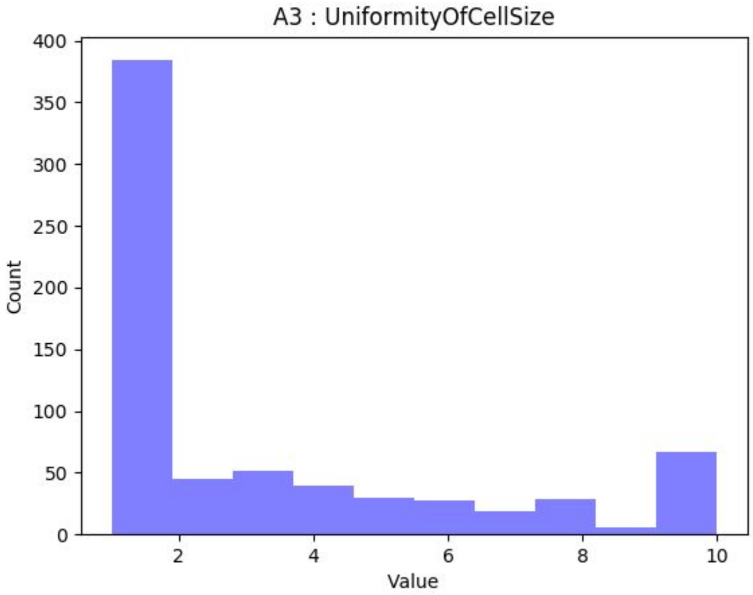
Phase 1 - Statistical Analysis

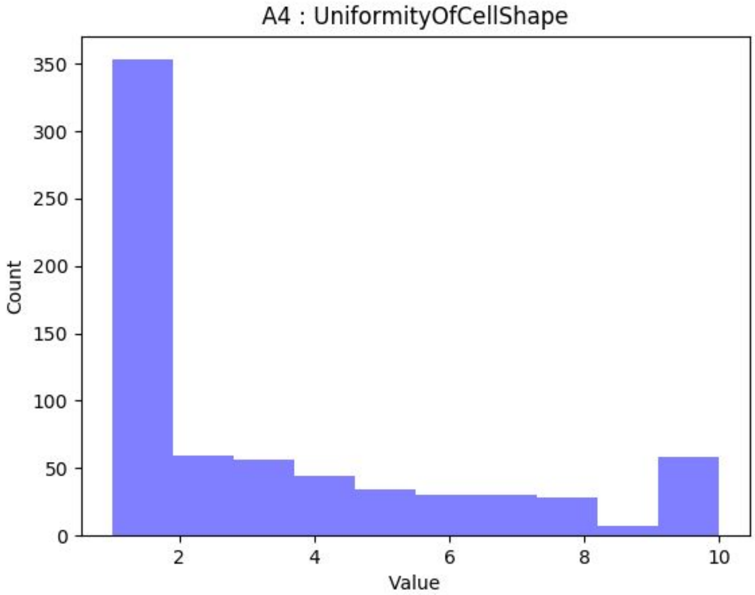
Summary Table

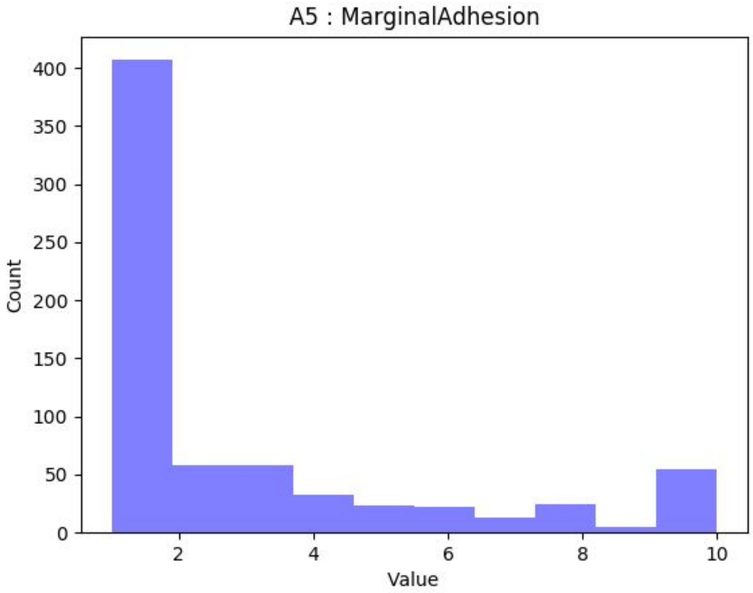
```
Attribute                                Min       Max      Mean    Median    StdDev Variance
A2  ClumpThickness                     1.000    10.000     4.418     4.000     2.816    7.928
A3  UniformityOfCellSize               1.000    10.000     3.134     1.000     3.051    9.311
A4  UniformityOfCellShape              1.000    10.000     3.207     1.000     2.972    8.832
A5  MarginalAdhesion                   1.000    10.000     2.807     1.000     2.855    8.153
A6  SingleEpithelialCellSize           1.000    10.000     3.216     2.000     2.214    4.903
A7  BareNuclei                         1.000    10.000     3.486     1.000     3.622   13.118
A8  BlandChromatin                     1.000    10.000     3.438     3.000     2.438    5.946
A9  NormalNucleoli                     1.000    10.000     2.867     1.000     3.054    9.325
A10 Mitoses                            1.000    10.000     1.589     1.000     1.715    2.941
```

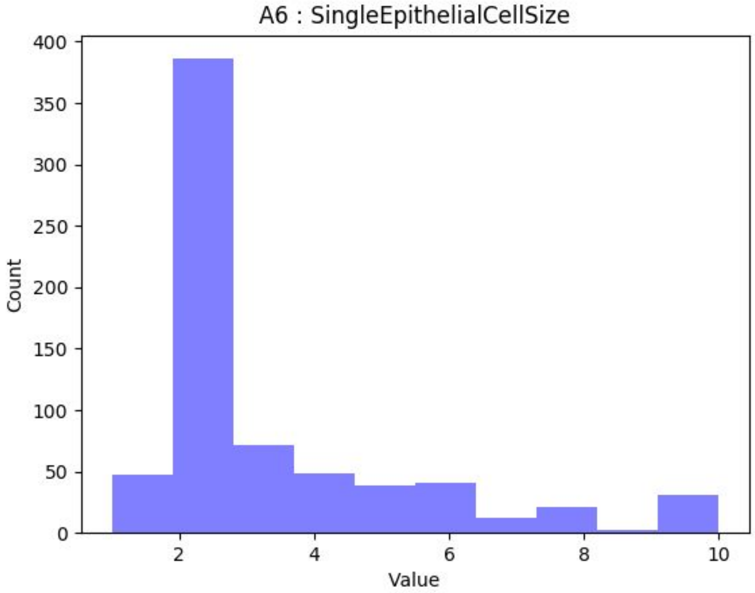
Histograms

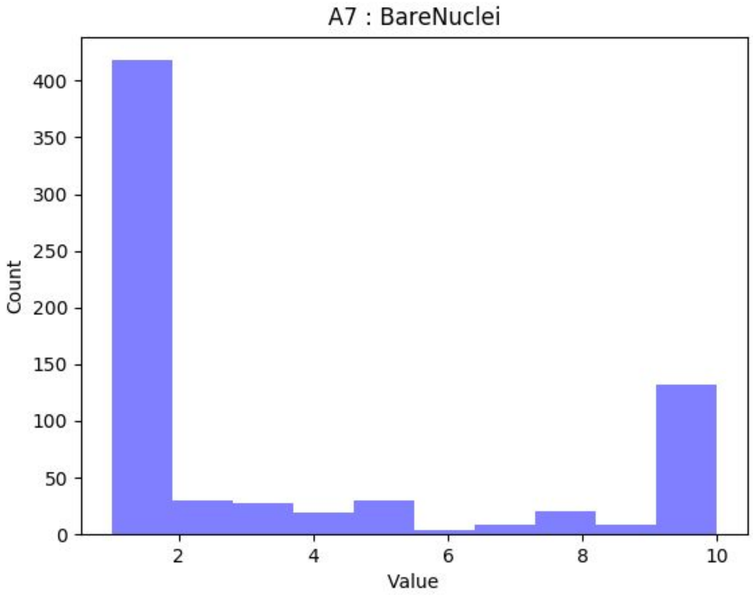


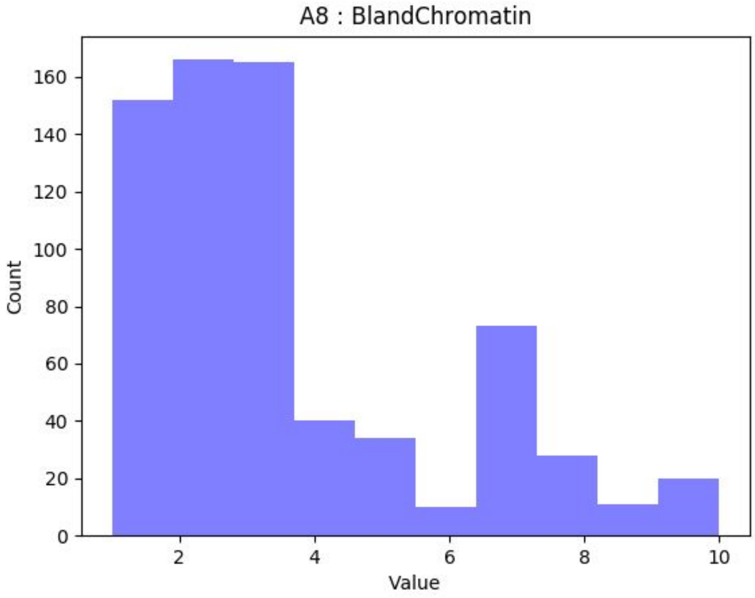

A3 : UniformityOfCellSize
Count
0
50
100
150
200
250
300
350
400
Value
2
4
6
8
10

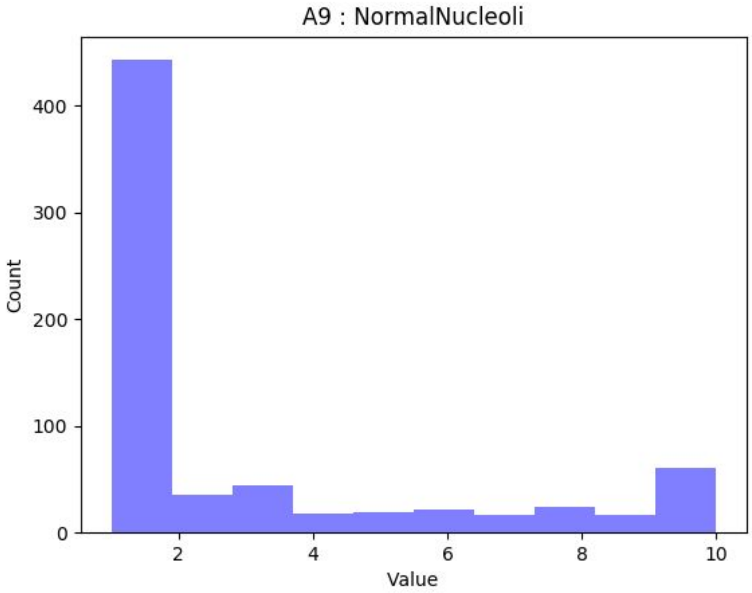

A4 : UniformityOfCellShape
Count
0
50
100
150
200
250
300
350
Value
2
4
6
8
10

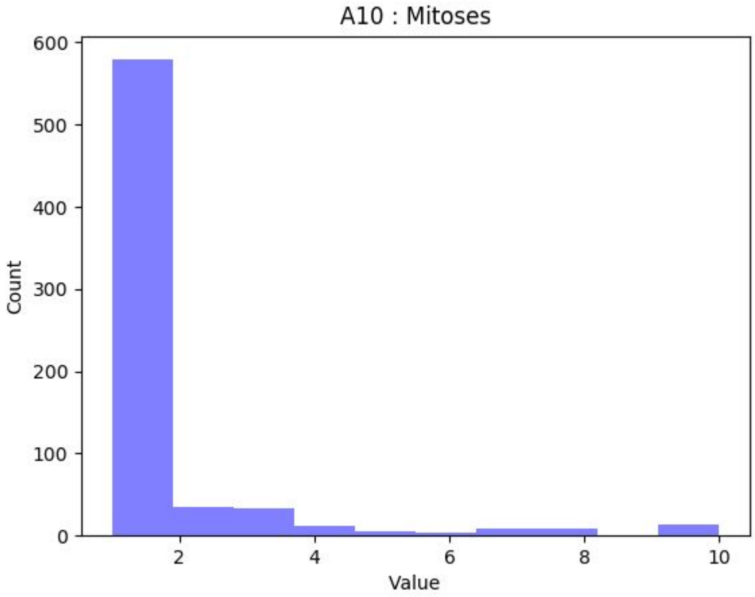

A5 : MarginalAdhesion
Count
0
50
100
150
200
250
300
350
400
2
4
6
8
10
Value


A6 : SingleEpithelialCellSize
Count
0
50
100
150
200
250
300
350
400
2
4
6
8
10
Value


A7 : BareNuclei
Count
0
50
100
150
200
250
300
350
400
Value
2
4
6
8
10


A8 : BlandChromatin
Count
0
20
40
60
80
100
120
140
160
Value
2
4
6
8
10


A9 : NormalNucleoli
Count
0
100
200
300
400
Value
2
4
6
8
10


A10 : Mitoses
Count
0
100
200
300
400
500
600
Value
2
4
6
8
10

Phase 2 - k-means Clustering

Program Output

```
$ time ./main.py
Please enter file name [default: Breast-Cancer-Wisconsin.csv]:

Generating Histograms...

Statistic Summary Table
Attribute                           Min      Max     Mean   Median   StdDev Variance
A2  ClumpThickness                1.000   10.000    4.418    4.000    2.816    7.928
A3  UniformityOfCellSize          1.000   10.000    3.134    1.000    3.051    9.311
A4  UniformityOfCellShape         1.000   10.000    3.207    1.000    2.972    8.832
A5  MarginalAdhesion              1.000   10.000    2.807    1.000    2.855    8.153
A6  SingleEpithelialCellSize      1.000   10.000    3.216    2.000    2.214    4.903
A7  BareNuclei                    1.000   10.000    3.486    1.000    3.622   13.118
A8  BlandChromatin                1.000   10.000    3.438    3.000    2.438    5.946
A9  NormalNucleoli                1.000   10.000    2.867    1.000    3.054    9.325
A10 Mitoses                       1.000   10.000    1.589    1.000    1.715    2.941

Running k-means Classification with 1500 iterations (please be patient)

Initial Means (all columns)
u2: Scn=673637.00000 A2=3.00000 A3=1.00000 A4=1.00000 A5=1.00000 A6=2.00000 A7=5.00000 A8=5.00000
A9=1.00000 A10=1.00000 CLASS=2.00000
u4: Scn=1115293.00000 A2=1.00000 A3=1.00000 A4=1.00000 A5=1.00000 A6=2.00000 A7=1.00000 A8=1.00000
A9=1.00000 A10=1.00000 CLASS=2.00000
....................................................................................................
....................................................................................................
....................................................................................................
....................................................................................................
....................................................................................................
....................................................................................................
....................................................................................................
....................................................................................................
....................................................................................................
....................................................................................................
....................................................................................................
....................................................................................................
....................................................................................................
....................................................................................................
....................................................................................................
Final Means
u2: A2=7.14957 A3=6.77778 A4=6.71368 A5=5.72650 A6=5.45726 A7=7.83761 A8=6.08974 A9=6.07692
A10=2.54274
```

```
u4: A2=3.04301 A3=1.30108 A4=1.44301 A5=1.33763 A6=2.08817 A7=1.29677 A8=2.10323 A9=1.25161
A10=1.10968

Cluster Assignment: u2-Benign
     N Index        Id     Class Predicted
     0     1   1002945         2         2
     1     3   1016277         2         2
     2     5   1017122         4         2
     3    14   1044572         4         2
     4    18   1050670         4         2
     5    20   1054590         4         2
     6    21   1054593         4         2
     7    25   1065726         4         2
     8    32   1072179         4         2
     9    36   1080185         4         2
    10    38   1084584         4         2
    11    39   1091262         4         2
    12    40   1096800         2         2
    13    41   1099510         4         2
    14    42   1100524         4         2
    15    43   1102573         4         2
    16    44   1103608         4         2
    17    46   1105257         4         2
    18    49   1106829         4         2
    19    52   1110102         4         2

Cluster Assignment: u4-Malignant
     N Index        Id     Class Predicted
     0     0   1000025         2         4
     1     2   1015425         2         4
     2     4   1017023         2         4
     3     6   1018099         2         4
     4     7   1018561         2         4
     5     8   1033078         2         4
     6     9   1033078         2         4
     7    10   1035283         2         4
     8    11   1036172         2         4
     9    12   1041801         4         4
    10    13   1043999         2         4
    11    15   1047630         4         4
    12    16   1048672         2         4
    13    17   1049815         2         4
    14    19   1050718         2         4
    15    22   1056784         2         4
    16    23   1057013         4         4
    17    24   1059552         2         4
    18    26   1066373         2         4
    19    27   1066979         2         4

real    43m27.623s
user    43m21.896s
sys     0m3.906s
```

Phase 3 - Performance Assessment

Additional output was added to show the performance of the classification. This is highlighted below.

Program Output

```
Please enter file name [default: Breast-Cancer-Wisconsin.csv]:

Generating Histograms...

Statistic Summary Table
Attribute                            Min      Max     Mean   Median   StdDev Variance
A2  ClumpThickness                 1.000   10.000    4.418    4.000    2.816    7.928
A3  UniformityOfCellSize           1.000   10.000    3.134    1.000    3.051    9.311
A4  UniformityOfCellShape          1.000   10.000    3.207    1.000    2.972    8.832
A5  MarginalAdhesion               1.000   10.000    2.807    1.000    2.855    8.153
A6  SingleEpithelialCellSize       1.000   10.000    3.216    2.000    2.214    4.903
A7  BareNuclei                     1.000   10.000    3.486    1.000    3.622   13.118
A8  BlandChromatin                 1.000   10.000    3.438    3.000    2.438    5.946
A9  NormalNucleoli                 1.000   10.000    2.867    1.000    3.054    9.325
A10 Mitoses                        1.000   10.000    1.589    1.000    1.715    2.941

How many iterations for k-means Classification [default: 1500]: 10
Running k-means Classification with 10 iterations (please be patient)

Initial Means (all columns)
u2: Scn=1238777.00000 A2=1.00000 A3=1.00000 A4=1.00000 A5=1.00000 A6=2.00000 A7=1.00000 A8=1.00000
A9=1.00000 A10=1.00000 CLASS=2.00000
u4: Scn=274137.00000 A2=8.00000 A3=8.00000 A4=9.00000 A5=4.00000 A6=5.00000 A7=10.00000 A8=7.00000
A9=8.00000 A10=1.00000 CLASS=4.00000
..........
Final Means
u2: A2=3.04301 A3=1.30108 A4=1.44301 A5=1.33763 A6=2.08817 A7=1.29677 A8=2.10323 A9=1.25161
A10=1.10968
u4: A2=7.14957 A3=6.77778 A4=6.71368 A5=5.72650 A6=5.45726 A7=7.83761 A8=6.08974 A9=6.07692
A10=2.54274

Cluster Assignment: u2-Benign
     N Index        Id     Class Predicted
     0     0   1000025         2         2
     1     2   1015425         2         2
     2     4   1017023         2         2
     3     6   1018099         2         2
     4     7   1018561         2         2
     5     8   1033078         2         2
     6     9   1033078         2         2
     7    10   1035283         2         2
     8    11   1036172         2         2
     9    12   1041801         4         2
    10    13   1043999         2         2
    11    15   1047630         4         2
    12    16   1048672         2         2
    13    17   1049815         2         2
    14    19   1050718         2         2
    15    22   1056784         2         2
    16    23   1057013         4         2
    17    24   1059552         2         2
    18    26   1066373         2         2
    19    27   1066979         2         2

Cluster Assignment: u4-Malignant
     N Index         Id      Class Predicted
     0     1   1002945         2         4
     1     3   1016277         2         4
     2     5   1017122         4         4
     3    14   1044572         4         4
     4    18   1050670         4         4
     5    20   1054590         4         4
     6    21   1054593         4         4
     7    25   1065726         4         4
     8    32   1072179         4         4
     9    36   1080185         4         4
    10    38   1084584         4         4
    11    39   1091262         4         4
    12    40   1096800         2         4
    13    41   1099510         4         4
    14    42   1100524         4         4
    15    43   1102573         4         4
    16    44   1103608         4         4
    17    46   1105257         4         4
    18    49   1106829         4         4
    19    52   1110102         4         4

Confusion Matrix:
      Benign   Malignant   <-- classified as
         447          11 |     Benign
          18         223 |     Malignant

Error Rates:
  ErrorB       = 0.02402
  ErrorM       = 0.07469
  Total Error = 0.09871
```

Classification Performance Results

In order to assess whether the implementation of k-mean classification was performing well, and reliably, several runs were performed and the results captured and analyzed. Each execution performed just 25 iterations of k-means clustering. Limiting the number of iterations kept the run-time of each to about 1 minute, perhaps sacrificing robustness and correctness of the model execution. The output is below, and is truncated to just show the initial means selection (including class), the final means computation, and the performance metrics (confusion matrix, and error rates).

Run #1

```
Initial Means (all columns)
u2: Scn=1114570.00000 A2=5.00000 A3=3.00000 A4=3.00000 A5=2.00000 A6=3.00000 A7=1.00000 A8=3.00000
A9=1.00000 A10=1.00000 CLASS=2.00000
u4: Scn=836433.00000 A2=5.00000 A3=1.00000 A4=1.00000 A5=3.00000 A6=2.00000 A7=1.00000 A8=1.00000
A9=1.00000 A10=1.00000 CLASS=2.00000
.........................
Final Means
u2: A2=7.14957 A3=6.77778 A4=6.71368 A5=5.72650 A6=5.45726 A7=7.83761 A8=6.08974 A9=6.07692
A10=2.54274
u4: A2=3.04301 A3=1.30108 A4=1.44301 A5=1.33763 A6=2.08817 A7=1.29677 A8=2.10323 A9=1.25161
A10=1.10968

Confusion Matrix:
      Benign   Malignant   <-- classified as
          11         447 |     Benign
         223          18 |     Malignant

Error Rates:
  ErrorB      = 0.97598
  ErrorM      = 0.92531
  Total Error = 1.90129
```

Run #2

```
Initial Means (all columns)
u2: Scn=1042252.00000 A2=3.00000 A3=1.00000 A4=1.00000 A5=1.00000 A6=2.00000 A7=1.00000 A8=2.00000
A9=1.00000 A10=1.00000 CLASS=2.00000
u4: Scn=1183240.00000 A2=4.00000 A3=1.00000 A4=2.00000 A5=1.00000 A6=2.00000 A7=1.00000 A8=2.00000
A9=1.00000 A10=1.00000 CLASS=2.00000
.........................
Final Means
u2: A2=3.04301 A3=1.30108 A4=1.44301 A5=1.33763 A6=2.08817 A7=1.29677 A8=2.10323 A9=1.25161
A10=1.10968
u4: A2=7.14957 A3=6.77778 A4=6.71368 A5=5.72650 A6=5.45726 A7=7.83761 A8=6.08974 A9=6.07692
A10=2.54274
```

```
Confusion Matrix:
      Benign   Malignant   <-- classified as
         447          11 |     Benign
          18         223 |     Malignant

Error Rates:
  ErrorB      = 0.02402
  ErrorM      = 0.07469
  Total Error = 0.09871
```

Run #3

```
Initial Means (all columns)
u2: Scn=1227210.00000 A2=10.00000 A3=5.00000 A4=5.00000 A5=6.00000 A6=3.00000 A7=10.00000 A8=7.00000
A9=9.00000 A10=2.00000 CLASS=4.00000
u4: Scn=1155546.00000 A2=2.00000 A3=1.00000 A4=1.00000 A5=2.00000 A6=3.00000 A7=1.00000 A8=2.00000
A9=1.00000 A10=1.00000 CLASS=2.00000
.........................
Final Means
u2: A2=7.14957 A3=6.77778 A4=6.71368 A5=5.72650 A6=5.45726 A7=7.83761 A8=6.08974 A9=6.07692
A10=2.54274
u4: A2=3.04301 A3=1.30108 A4=1.44301 A5=1.33763 A6=2.08817 A7=1.29677 A8=2.10323 A9=1.25161
A10=1.10968

Confusion Matrix:
      Benign   Malignant   <-- classified as
          11         447 |     Benign
         223          18 |     Malignant

Error Rates:
  ErrorB      = 0.97598
  ErrorM      = 0.92531
  Total Error = 1.90129
```

Run #4

```
Initial Means (all columns)
u2: Scn=1188472.00000 A2=1.00000 A3=1.00000 A4=1.00000 A5=1.00000 A6=1.00000 A7=1.00000 A8=3.00000
A9=1.00000 A10=1.00000 CLASS=2.00000
u4: Scn=1115282.00000 A2=5.00000 A3=3.00000 A4=5.00000 A5=5.00000 A6=3.00000 A7=3.00000 A8=4.00000
A9=10.00000 A10=1.00000 CLASS=4.00000
.........................
Final Means
u2: A2=3.04301 A3=1.30108 A4=1.44301 A5=1.33763 A6=2.08817 A7=1.29677 A8=2.10323 A9=1.25161
A10=1.10968
u4: A2=7.14957 A3=6.77778 A4=6.71368 A5=5.72650 A6=5.45726 A7=7.83761 A8=6.08974 A9=6.07692
A10=2.54274

Confusion Matrix:
      Benign   Malignant   <-- classified as
         447          11 |     Benign
          18         223 |     Malignant

Error Rates:
  ErrorB      = 0.02402
  ErrorM      = 0.07469
  Total Error = 0.09871
```

Run #5

```
Initial Means (all columns)
u2: Scn=1199983.00000 A2=1.00000 A3=1.00000 A4=1.00000 A5=1.00000 A6=2.00000 A7=1.00000 A8=3.00000
A9=1.00000 A10=1.00000 CLASS=2.00000
u4: Scn=877291.00000 A2=6.00000 A3=10.00000 A4=10.00000 A5=10.00000 A6=10.00000 A7=10.00000
A8=8.00000 A9=10.00000 A10=10.00000 CLASS=4.00000
.........................
Final Means
u2: A2=3.04301 A3=1.30108 A4=1.44301 A5=1.33763 A6=2.08817 A7=1.29677 A8=2.10323 A9=1.25161
A10=1.10968
u4: A2=7.14957 A3=6.77778 A4=6.71368 A5=5.72650 A6=5.45726 A7=7.83761 A8=6.08974 A9=6.07692
A10=2.54274

Confusion Matrix:
      Benign   Malignant   <-- classified as
         447          11 |     Benign
          18         223 |     Malignant

Error Rates:
  ErrorB      = 0.02402
  ErrorM      = 0.07469
  Total Error = 0.09871
```

Summary

Based on several previous executions during phase 2, I was developing the belief that the classes of the initial randomly selected means had a great impact on the final outcome (I called this the *First Random Selection Class assumption*). However, after analyzing the results from the five executions performed for this assessment, and reading more about the k-means classification algorithm this does not necessarily hold true.

While this First Random Selection Class assumption seems to hold true in Run #1 (looking only at the initial means and the final performance results). And in Run #3 where, where the initial means for each class are the opposite of their named groups, the final results are also dismal.

However, clearly Run #2 proves the First Random Selection Class assumption wrong. In Run #2 both initial means are class 2 (Benign) and the final results show excellent performance with ErrorB = 0.024 and ErrorM = 0.075.

Conclusion

Writing a k-means classifier using Python, Pandas, NumPy, and Matplotlib proved a very interesting exercise. The ease of loading data, and manipulating and processing the data makes this technology combination very appealing for a broad range of problems. Of particular note is how easy it is to work with DataFrames and Series, and how concise and elegant solutions can be with them.