Homework 3B
Divya Mereddy

Consider the data files used for Homework 3 for the following tasks.

1. Consider the BCP dataset and its class variable with values “R” (Recurrence Occurred) and “N” (No Recurrence Occurred so far). Ignore the attribute that gives the number of years after which recurrence occurred or the number of years for which the patient has been free of recurrence. There are thirty other attribute values given as features measured for every patient. Use only these thirty attributes and perform the following:
    a. Run k-means algorithm with this dataset for k=4. Run it three different times and for each run show the cluster centers and the SSE values for each cluster and also the total SSE value for the clustering.

```
Centroid for K-means clustering with k-value of  4 -1st run :

 [[2.06943590e+01 2.27648718e+01 1.37466667e+02 1.33373333e+03
  1.02547179e-01 1.63605641e-01 2.03023333e-01 1.13464359e-01
  1.97112821e-01 6.11325641e-02 8.00620513e-01 1.25430769e+00
  5.66735897e+00 1.05393077e+02 6.46964103e-03 3.36861538e-02
  4.45279487e-02 1.60099487e-02 2.03479487e-02 3.91723077e-03
  2.55200000e+01 2.98020513e+01 1.71187179e+02 1.99917949e+03
  1.39948974e-01 3.88051282e-01 4.85079487e-01 2.06971282e-01
  3.24574359e-01 8.71279487e-02]
 [1.43300000e+01 2.17933766e+01 9.42207792e+01 6.40233766e+02
  1.04967013e-01 1.36427013e-01 1.28980909e-01 6.55770130e-02
  1.92972727e-01 6.59350649e-02 3.99290909e-01 1.21671558e+00
  2.88248052e+00 3.57528571e+01 7.12903896e-03 3.16628701e-02
  3.93915584e-02 1.39925974e-02 2.06202468e-02 4.17268831e-03
  1.69179221e+01 3.03093506e+01 1.13196883e+02 8.81270130e+02
  1.51371039e-01 3.89264416e-01 4.33968571e-01 1.61582727e-01
  3.37996104e-01 9.95046753e-02]
 [1.80736111e+01 2.21644444e+01 1.18891667e+02 1.01715972e+03
  1.00192778e-01 1.35127222e-01 1.49740694e-01 8.78423611e-02
  1.91263889e-01 6.05190278e-02 6.35109722e-01 1.31407222e+00
  4.42997222e+00 7.22666667e+01 6.69909722e-03 3.00197083e-02
  4.05502778e-02 1.58253333e-02 2.09968472e-02 3.91911111e-03
  2.15563889e+01 2.96759722e+01 1.43134722e+02 1.42073611e+03
  1.38678194e-01 3.29336111e-01 4.08220972e-01 1.75341667e-01
  3.11904167e-01 8.47906944e-02]
 [2.35850000e+01 2.48890000e+01 1.56520000e+02 1.75190000e+03
  1.03524000e-01 1.62960000e-01 2.30530000e-01 1.38238000e-01
  1.84800000e-01 5.97150000e-02 1.17651000e+00 1.31428000e+00
  8.06320000e+00 1.83879000e+02 5.52620000e-03 2.64240000e-02
  3.79100000e-02 1.48410000e-02 1.76760000e-02 3.31620000e-03
  3.12300000e+01 3.34770000e+01 2.09070000e+02 3.00630000e+03
  1.39790000e-01 3.47060000e-01 4.73810000e-01 2.25960000e-01
  2.89280000e-01 8.19190000e-02]]
SSE value with k-value of  4 -1st run:  11962452.98781946
```

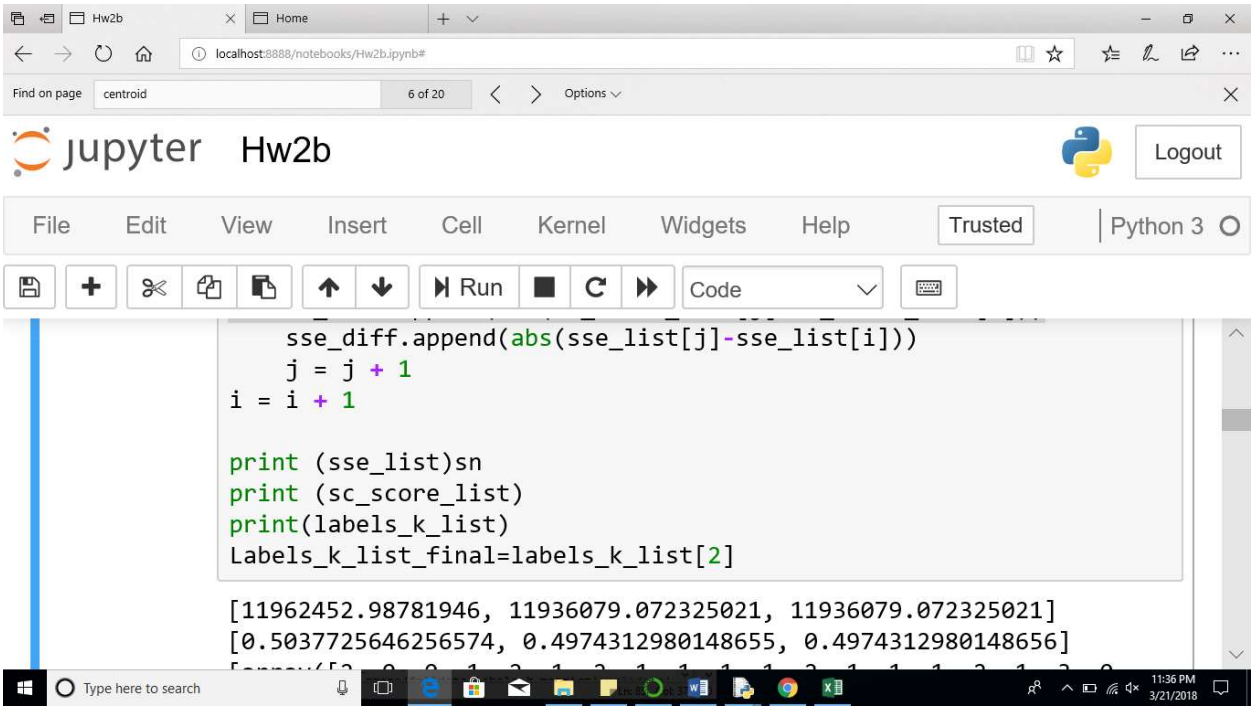
```
Labels with k-value of  4 - 1st run:  [2 0 0 1 2 1 2 1 1 1 1 2 1 1 1 2 1 3
0 2 1 2 2 2 1 2 0 2 2 1 1 1 1 1 2 1 1
 2 1 1 0 1 1 1 1 2 1 2 2 1 0 1 1 1 1 2 2 2 0 2 2 2 0 1 1 2 1 2 2 0 0 2 2 1
 1 3 0 2 1 2 1 1 1 1 2 1 0 1 1 0 1 1 1 0 0 1 1 2 0 3 2 2 2 0 1 0 1 2 1 0 2
 1 2 3 0 2 2 2 1 1 0 0 2 2 0 1 2 0 3 2 1 3 1 0 2 0 0 1 1 2 0 2 1 2 2 2 2 1
 2 2 2 0 2 2 1 0 1 2 2 0 1 3 1 1 2 2 3 2 1 0 0 0 1 1 2 0 2 2 0 1 1 2 1 0 2
 1 3 3 1 2 2 2 1 0 1 2 0 2]

Silhouette coefficient score with k-value of  4-1st run :  0.5037725646256
574

SSE0: 10066419856.979
SSE1: 23997255327.87937
SSE2: 24506604353.181778
SSE2: 9543256098.542744

Centroid for K-means clustering with k-value of  4 -2nd Run:
 [[1.41723611e+01 2.16708333e+01 9.32708333e+01 6.25316667e+02
  1.05880278e-01 1.39142639e-01 1.31299167e-01 6.65175000e-02
  1.94956944e-01 6.65120833e-02 3.98543056e-01 1.22164167e+00
  2.87844444e+00 3.52502778e+01 6.92375000e-03 3.15364167e-02
  3.86712500e-02 1.38303056e-02 2.08785972e-02 4.12936111e-03
  1.67775000e+01 3.02890278e+01 1.12335556e+02 8.66080556e+02
  1.53324583e-01 3.99206389e-01 4.45023333e-01 1.64510417e-01
  3.41261111e-01 1.01010139e-01]
 [2.04522917e+01 2.24664583e+01 1.35660417e+02 1.30115833e+03
  1.01716667e-01 1.58434583e-01 1.95998125e-01 1.10202917e-01
  1.95597917e-01 6.07514583e-02 7.80731250e-01 1.22268958e+00
  5.53164583e+00 1.01524375e+02 6.68639583e-03 3.27866667e-02
  4.51310417e-02 1.61766250e-02 2.08443750e-02 3.89443750e-03
  2.50060417e+01 2.93854167e+01 1.67866667e+02 1.91758333e+03
  1.39979375e-01 3.74533333e-01 4.74681250e-01 2.00772500e-01
  3.21000000e-01 8.60306250e-02]
 [1.77074627e+01 2.24034328e+01 1.16374627e+02 9.75891045e+02
  1.00082239e-01 1.32492985e-01 1.43834627e-01 8.39279104e-02
  1.89725373e-01 6.05682090e-02 6.09897015e-01 1.33902836e+00
  4.24725373e+00 6.80217910e+01 6.85402985e-03 3.04984776e-02
  4.04555224e-02 1.57127164e-02 2.04440746e-02 4.00385075e-03
  2.11005970e+01 3.00483582e+01 1.39823881e+02 1.36289552e+03
  1.37570597e-01 3.24489552e-01 3.95588209e-01 1.70476418e-01
  3.10500000e-01 8.46374627e-02]
 [2.35563636e+01 2.46300000e+01 1.56118182e+02 1.74590909e+03
  1.01784545e-01 1.58554545e-01 2.21609091e-01 1.34490909e-01
  1.84372727e-01 5.93345455e-02 1.12993636e+00 1.27262727e+00
  7.74863636e+00 1.76058182e+02 5.47018182e-03 2.63345455e-02
  3.70290909e-02 1.49672727e-02 1.78472727e-02 3.35472727e-03
  3.09372727e+01 3.29990909e+01 2.06809091e+02 2.95145455e+03
  1.38245455e-01 3.48081818e-01 4.66627273e-01 2.26745455e-01
  2.95609091e-01 8.28236364e-02]]
SSE value with k-value of  4-2nd run :  11936079.072325021
```

```
Labels with k-value of  4 -2nd run :  [2 1 1 0 1 0 2 0 0 0 2 2 0 0 0 2 0 3
1 2 0 2 2 2 0 2 1 2 2 0 0 0 0 0 2 0 0
 2 0 0 1 0 0 0 0 2 0 1 2 0 1 0 0 0 0 2 2 2 1 2 2 2 1 0 0 2 0 2 2 1 3 2 2 0
 0 3 1 2 0 2 0 0 0 0 1 0 1 0 0 1 2 0 0 1 1 0 0 2 1 3 2 2 2 1 0 1 0 2 0 1 2
 0 2 3 1 2 1 1 0 0 1 1 2 1 1 2 2 1 3 2 0 3 0 1 1 1 1 0 0 2 1 2 0 2 2 2 2 0
 2 2 2 1 2 2 0 1 2 2 2 1 0 3 0 0 2 1 3 2 0 1 1 1 2 0 2 1 2 2 1 0 0 2 0 1 2
 0 3 3 0 1 1 2 0 1 0 2 1 2]
Silhouette coefficient score with k-value of  4 -2nd run:  0.4974312980148
655

SSE0: 10066419856.979
SSE1: 23997255327.87937
SSE2: 24506604353.181778
SSE3: 9543256098.542744

Centroid for K-means clustering with k-value of  4 -3rd Run:
 [[2.35563636e+01 2.46300000e+01 1.56118182e+02 1.74590909e+03
  1.01784545e-01 1.58554545e-01 2.21609091e-01 1.34490909e-01
  1.84372727e-01 5.93345455e-02 1.12993636e+00 1.27262727e+00
  7.74863636e+00 1.76058182e+02 5.47018182e-03 2.63345455e-02
  3.70290909e-02 1.49672727e-02 1.78472727e-02 3.35472727e-03
  3.09372727e+01 3.29990909e+01 2.06809091e+02 2.95145455e+03
  1.38245455e-01 3.48081818e-01 4.66627273e-01 2.26745455e-01
  2.95609091e-01 8.28236364e-02]
 [1.77074627e+01 2.24034328e+01 1.16374627e+02 9.75891045e+02
  1.00082239e-01 1.32492985e-01 1.43834627e-01 8.39279104e-02
  1.89725373e-01 6.05682090e-02 6.09897015e-01 1.33902836e+00
  4.24725373e+00 6.80217910e+01 6.85402985e-03 3.04984776e-02
  4.04555224e-02 1.57127164e-02 2.04440746e-02 4.00385075e-03
  2.11005970e+01 3.00483582e+01 1.39823881e+02 1.36289552e+03
  1.37570597e-01 3.24489552e-01 3.95588209e-01 1.70476418e-01
  3.10500000e-01 8.46374627e-02]
 [2.04522917e+01 2.24664583e+01 1.35660417e+02 1.30115833e+03
  1.01716667e-01 1.58434583e-01 1.95998125e-01 1.10202917e-01
  1.95597917e-01 6.07514583e-02 7.80731250e-01 1.22268958e+00
  5.53164583e+00 1.01524375e+02 6.68639583e-03 3.27866667e-02
  4.51310417e-02 1.61766250e-02 2.08443750e-02 3.89443750e-03
  2.50060417e+01 2.93854167e+01 1.67866667e+02 1.91758333e+03
  1.39979375e-01 3.74533333e-01 4.74681250e-01 2.00772500e-01
  3.21000000e-01 8.60306250e-02]
 [1.41723611e+01 2.16708333e+01 9.32708333e+01 6.25316667e+02
  1.05880278e-01 1.39142639e-01 1.31299167e-01 6.65175000e-02
  1.94956944e-01 6.65120833e-02 3.98543056e-01 1.22164167e+00
  2.87844444e+00 3.52502778e+01 6.92375000e-03 3.15364167e-02
  3.86712500e-02 1.38303056e-02 2.08785972e-02 4.12936111e-03
  1.67775000e+01 3.02890278e+01 1.12335556e+02 8.66080556e+02
  1.53324583e-01 3.99206389e-01 4.45023333e-01 1.64510417e-01
  3.41261111e-01 1.01010139e-01]]
SSE value with k-value of  4-3rd Run :  11936079.072325021
```

```
Labels with k-value of  4 -3rd Run :  [1 2 2 3 2 3 1 3 3 3 1 1 3 3 3 1 3 0
2 1 3 1 1 1 3 1 2 1 1 3 3 3 3 3 1 3 3
 1 3 3 2 3 3 3 3 1 3 2 1 3 2 3 3 3 3 1 1 1 2 1 1 1 2 3 3 1 3 1 1 2 0 1 1 3
 3 0 2 1 3 1 3 3 3 3 2 3 2 3 3 2 1 3 3 2 2 3 3 1 2 0 1 1 1 2 3 2 3 1 3 2 1
 3 1 0 2 1 2 2 3 3 2 2 1 2 2 1 1 2 0 1 3 0 3 2 2 2 2 3 3 1 2 1 3 1 1 1 1 3
 1 1 1 2 1 1 3 2 1 1 1 2 3 0 3 3 1 2 0 1 3 2 2 2 1 3 1 2 1 1 2 3 3 1 3 2 1
 3 0 0 3 2 2 1 3 2 3 1 2 1]
Silhouette coefficient score with k-value of  4-3rd Run:  0.49743129801486
56


SSE0: 7661920195.430863
SSE1: 18765336587.890865
SSE2: 26129751798.96137
SSE2: 8919042209.631622
```

b. Select the best of the above three clusterings and explain how you chose the best candidate.

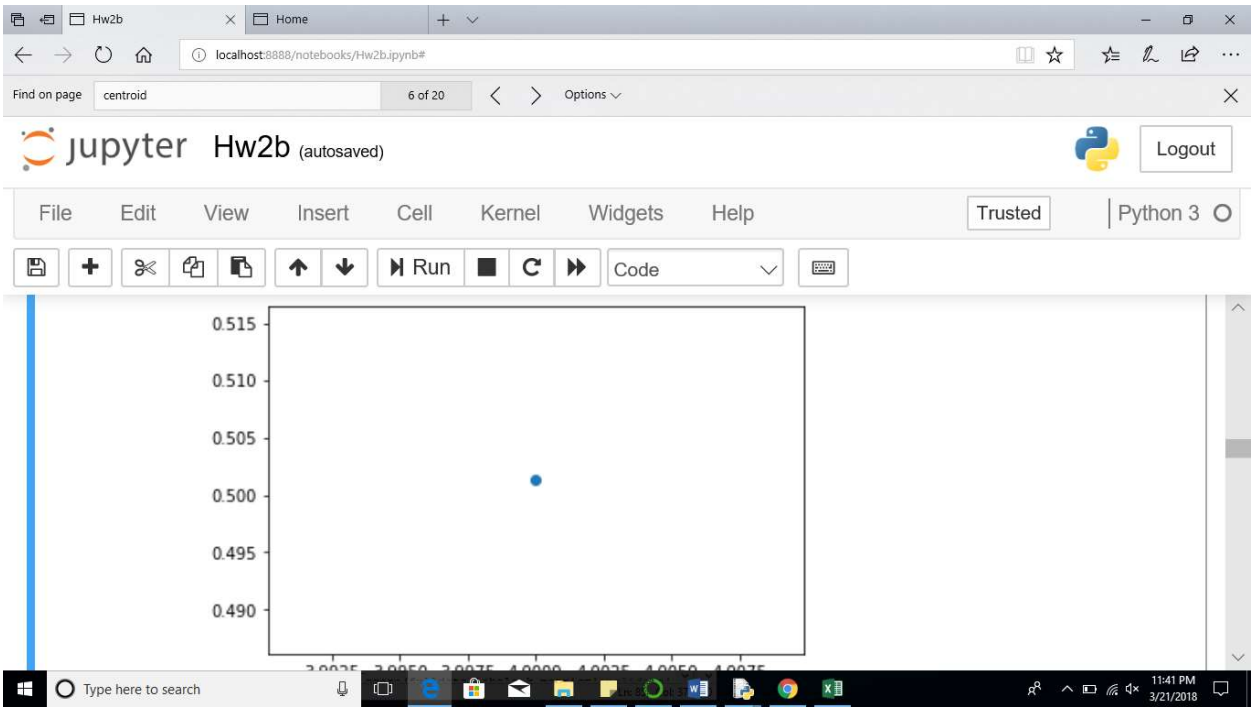
If we want to find a balance between two variables: the number of clusters (k) and the average variance of the clusters. We have to minimize the former while also minimizing the latter. Of course, as the number of clusters increases, the average variance decreases (up to the trivial case of k=n and variance=0).
In in observation, I found the SC score should be higher and SSE should be low.



So I went with 3rd run which has comparatively good balance between higher SC score and lower SSE.

c. For the best candidate clustering chosen by you plot the Silhouette coefficient for the clustering. Compute and report the average Silhouette Coefficient for each cluster of the chosen clustering.

Average Silhouette Coefficient Plotting:




```
Cluster-1:0.67536589856
Cluster-2:0.358476874655
Cluster-3:0.556746574734
Cluster-4:0.482764746756

Silhouette coefficient score with k-value of  4-3rd Run:  0.49743129801486
56
```

d. Now consider the class label for each data point in each cluster (“R” or “N”). To each cluster assign the label that belongs to most of the data points in that cluster.

cluster center

```
Centroid for K-means clustering with k-value of  4:
```

```
 [[2.35563636e+01 2.46300000e+01 1.56118182e+02 1.74590909e+03
  1.01784545e-01 1.58554545e-01 2.21609091e-01 1.34490909e-01
  1.84372727e-01 5.93345455e-02 1.12993636e+00 1.27262727e+00
  7.74863636e+00 1.76058182e+02 5.47018182e-03 2.63345455e-02
  3.70290909e-02 1.49672727e-02 1.78472727e-02 3.35472727e-03
  3.09372727e+01 3.29990909e+01 2.06809091e+02 2.95145455e+03
  1.38245455e-01 3.48081818e-01 4.66627273e-01 2.26745455e-01
  2.95609091e-01 8.28236364e-02]
 [1.77074627e+01 2.24034328e+01 1.16374627e+02 9.75891045e+02
  1.00082239e-01 1.32492985e-01 1.43834627e-01 8.39279104e-02
  1.89725373e-01 6.05682090e-02 6.09897015e-01 1.33902836e+00
  4.24725373e+00 6.80217910e+01 6.85402985e-03 3.04984776e-02
  4.04555224e-02 1.57127164e-02 2.04440746e-02 4.00385075e-03
  2.11005970e+01 3.00483582e+01 1.39823881e+02 1.36289552e+03
  1.37570597e-01 3.24489552e-01 3.95588209e-01 1.70476418e-01
  3.10500000e-01 8.46374627e-02]
 [2.04522917e+01 2.24664583e+01 1.35660417e+02 1.30115833e+03
  1.01716667e-01 1.58434583e-01 1.95998125e-01 1.10202917e-01
  1.95597917e-01 6.07514583e-02 7.80731250e-01 1.22268958e+00
  5.53164583e+00 1.01524375e+02 6.68639583e-03 3.27866667e-02
  4.51310417e-02 1.61766250e-02 2.08443750e-02 3.89443750e-03
  2.50060417e+01 2.93854167e+01 1.67866667e+02 1.91758333e+03
  1.39979375e-01 3.74533333e-01 4.74681250e-01 2.00772500e-01
  3.21000000e-01 8.60306250e-02]
 [1.41723611e+01 2.16708333e+01 9.32708333e+01 6.25316667e+02
  1.05880278e-01 1.39142639e-01 1.31299167e-01 6.65175000e-02
  1.94956944e-01 6.65120833e-02 3.98543056e-01 1.22164167e+00
  2.87844444e+00 3.52502778e+01 6.92375000e-03 3.15364167e-02
  3.86712500e-02 1.38303056e-02 2.08785972e-02 4.12936111e-03
  1.67775000e+01 3.02890278e+01 1.12335556e+02 8.66080556e+02
  1.53324583e-01 3.99206389e-01 4.45023333e-01 1.64510417e-01
  3.41261111e-01 1.01010139e-01]]

Labels with k-value of  4 :  [1 2 2 3 2 3 1 3 3 3 1 1 3 3 3 1 3 0 2 1 3 1
1 1 3 1 2 1 1 3 3 3 3 3 1 3 3
 1 3 3 2 3 3 3 3 1 3 2 1 3 2 3 3 3 3 1 1 1 2 1 1 1 2 3 3 1 3 1 1 2 0 1 1 3
 3 0 2 1 3 1 3 3 3 3 2 3 2 3 3 2 1 3 3 2 2 3 3 1 2 0 1 1 1 2 3 2 3 1 3 2 1
 3 1 0 2 1 2 2 3 3 2 2 1 2 2 1 1 2 0 1 3 0 3 2 2 2 2 3 3 1 2 1 3 1 1 1 1 3
 1 1 1 2 1 1 3 2 1 1 1 2 3 0 3 3 1 2 0 1 3 2 2 2 1 3 1 2 1 1 2 3 3 1 3 2 1
 3 0 0 3 2 2 1 3 2 3 1 2 1]


its SSE : 11936079.072325021

fraction of points that have the class label.

Population of R and N classes in clusters:
Cluster-1- for N :  6
Cluster-1- for R :  5
Cluster-2- for N :  49
```

```
Cluster-2- for R :  18
Cluster-3- for N :  35
Cluster-3- for R :  13
Cluster-4- for N :  61
Cluster-4- for R :  11
```

class label:

```
[1, 0, 0, 0]
```

e. Now, use the cluster centers and the class labels as a new classifier. Consider each data point again as belonging to your test set.

For each data point predict its class label to be the one that belongs to the cluster center that is closest to the data point.

```
Predicted labels:

array([1, 0, 0, 2, 1, 2, 1, 2, 2, 2, 2, 1, 2, 2, 2, 1, 2, 3, 0, 1, 2, 1,
1, 1, 2, 1, 0, 1, 1, 2, 2, 2, 2, 2, 1, 2, 2, 1, 2, 2, 0, 2, 2, 2, 2, 1, 2,
1, 1, 2, 0, 2, 2, 2, 2, 1, 1, 1, 0, 1, 1, 1, 0, 2, 2, 1, 2, 1, 1, 0, 0, 1,
1, 2, 2, 3, 0, 1, 2, 1, 2, 2, 2, 2, 1, 2, 0, 2, 2, 0, 2, 2, 2, 0, 0, 2, 2,
1, 0, 3, 1, 1, 1, 0, 2, 0, 2, 1, 2, 0, 1, 2, 1, 3, 0, 1, 0, 1, 2, 2, 0, 0,
1, 0, 0, 2, 1, 0, 3, 1, 2, 3, 2, 0, 0, 0, 0, 2, 2, 1, 0, 1, 2, 1, 1, 1, 1,
2, 1, 1, 1, 0, 1, 1, 2, 0, 2, 1, 1, 0, 2, 3, 2, 2, 1, 1, 3, 1, 2, 0, 0, 0,
1, 2, 1, 0, 1, 1, 0, 2, 2, 1, 2, 0, 1, 2, 3, 3, 2, 1, 1, 1, 2, 0, 2, 1, 0,
1])
```

Predicted class label:

```
array([0, 0, 0, 1, 0, 1, 0, 1, 1, 1, 0, 0, 1, 1, 1, 0, 1, 0, 0, 0, 1, 0,
0, 0, 1, 0, 0, 0, 0, 1, 1, 1, 1, 1, 0, 1, 1, 0, 1, 1, 0, 1, 1, 1, 1, 0, 1,
0, 0, 1, 0, 1, 1, 1, 1, 0, 0, 0, 0, 0, 0, 0, 0, 1, 1, 0, 1, 0, 0, 0, 0, 0,
0, 1, 1, 0, 0, 0, 1, 0, 1, 1, 1, 1, 0, 1, 0, 1, 1, 0, 0, 1, 1, 0, 0, 1, 1,
0, 0, 0, 0, 0, 0, 0, 1, 0, 1, 0, 1, 0, 0, 1, 0, 0, 0, 0, 0, 0, 1, 1, 0, 0,
0, 0, 0, 0, 0, 0, 0, 0, 1, 0, 1, 0, 0, 0, 0, 1, 1, 0, 0, 0, 1, 0, 0, 0, 0,
1, 0, 0, 0, 0, 0, 0, 1, 0, 0, 0, 0, 0, 1, 0, 1, 1, 0, 0, 0, 0, 1, 0, 0, 0,
0, 1, 0, 0, 0, 0, 0, 1, 1, 0, 1, 0, 0, 1, 0, 0, 1, 0, 0, 0, 1, 0, 1, 0, 0,
0])
```

compute its accuracy, precision and recall values.

```
Accuracy:  58.080808080808076
Precision:  1.0101010101010102
recall:  1.0101010101010102
```

f. Compare these performance results with those obtained by you in HW3 Q1. Comment on the possible causes for the differences between the two sets of performance values.

- I guess as decision tree is able to classify data further into smaller decision sub tress it gave us good performance than the current clustering
- May be if we increase number of clusters, the k-means can cluster the data more effectively and give good accuracy.

2. Mix the datasets for the red and white wines in one dataset. Perform k-means clustering on this large dataset for the values of k to be: 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, and 14.

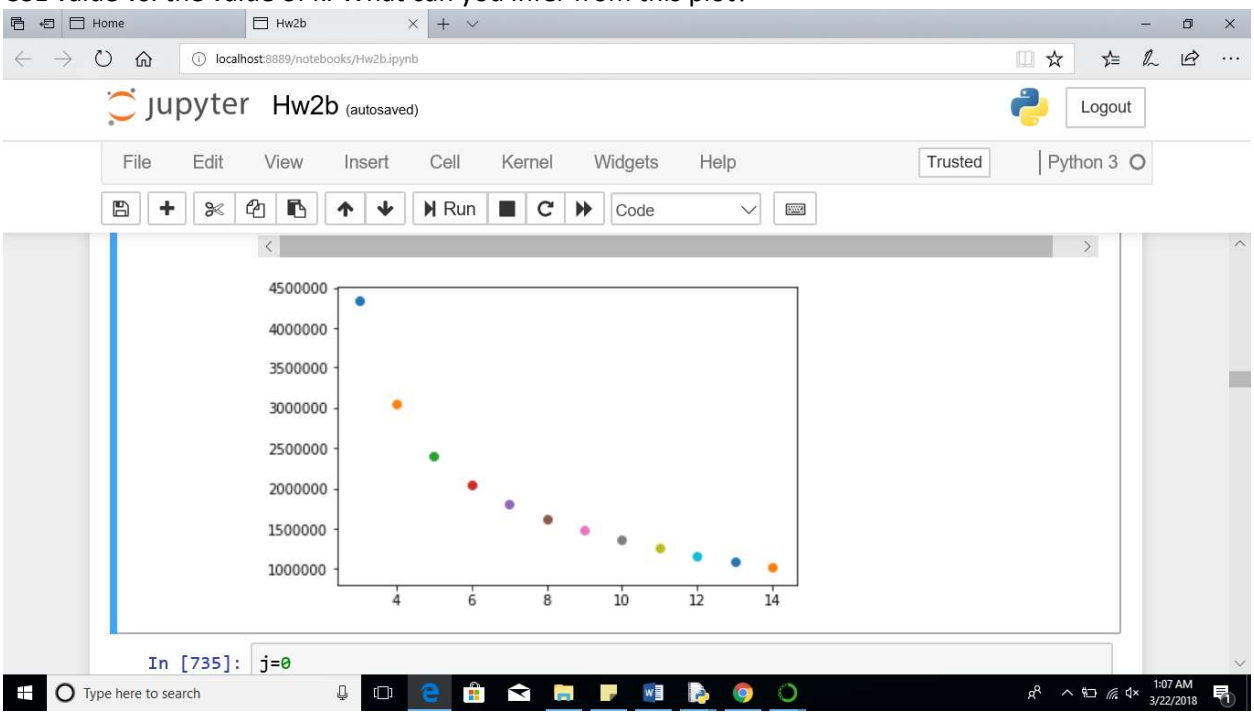
For each value of k report the lowest total SSE value after selecting the best of the 3-runs for each value of k.

Best of 3 runs:

```
SSE value with k-value of 3 : 4336387.441652566
Population of the cluster with k-value of 3 : [2925 1618 1954]
Silhouette coefficient score with k-value of 3 : 0.5036522116797673

SSE value with k-value of 4 : 3043614.856177952
Population of the cluster with k-value of 4 : [1385 2074 1091 1947]
Silhouette coefficient score with k-value of 4 : 0.45025231513275665 SSE

value with k-value of 5 : 2398816.868905013
Population of the cluster with k-value of 5 : [1797 1176 1470 1295 759]
Silhouette coefficient score with k-value of 5 : 0.4171867446934559

SSE value with k-value of 6 : 2046288.0114755668
Population of the cluster with k-value of 6 : [1095 1375 1643 1031 959
394]
Silhouette coefficient score with k-value of 6 : 0.39015411615926876

SSE value with k-value of 7 : 1801204.051114595
Population of the cluster with k-value of 7 : [ 723 856 1148 916 354 1143
1357]
Silhouette coefficient score with k-value of 7 : 0.3640047912311862

SSE value with k-value of 8 : 1628497.027658349
Population of the cluster with k-value of 8 : [ 910 1359 753 717 300 1125
604 729]
Silhouette coefficient score with k-value of 8 : 0.36378035013716425

SSE value with k-value of 9 : 1488004.1257619474
Population of the cluster with k-value of 9 : [ 610 1006 626 301 1183 605
781 807 578]
Silhouette coefficient score with k-value of 9 : 0.34209719445413533

SSE value with k-value of 10 : 1369241.4218815402
Population of the cluster with k-value of 10 : [ 644 606 1114 594 915 831
7 316 555 915]
Silhouette coefficient score with k-value of 10 : 0.3453672924416606

SSE value with k-value of 11 : 1264699.9898976563 Population of the
cluster with k-value of 11 : [730 662 528 729 7 737 338 561 505 971 729]
Silhouette coefficient score with k-value of 11 : 0.33718810661218596 SSE
```

```
value with k-value of 12 : 1163452.5411679726
Population of the cluster with k-value of 12 : [719 438 718 540 640 617
268 459 542 7 895 654]
Silhouette coefficient score with k-value of 12 : 0.3352483584670191

SSE value with k-value of 13 : 1087585.7601274415

Population of the cluster with k-value of 13 : [369 818 710 656 612 507
538 227 518 635 7 376 524]

Silhouette coefficient score with k-value of 13 : 0.3294782065055398
SSE value with k-value of 14 : 1025676.7660481313
Population of the cluster with k-value of 14 : [509 710 351 704 505 7 486
169 353 513 576 423 571 620]
Silhouette coefficient score with k-value of 14 : 0.3264670933208289
```

Plot the SSE value vs. the value of k. What can you infer from this plot?



As per graph analysis as k value increases the error decreased
But when very highest k value taken it may lead to over fitting properly and doesn't work well for real time prediction