Name(s): Michael Belmear

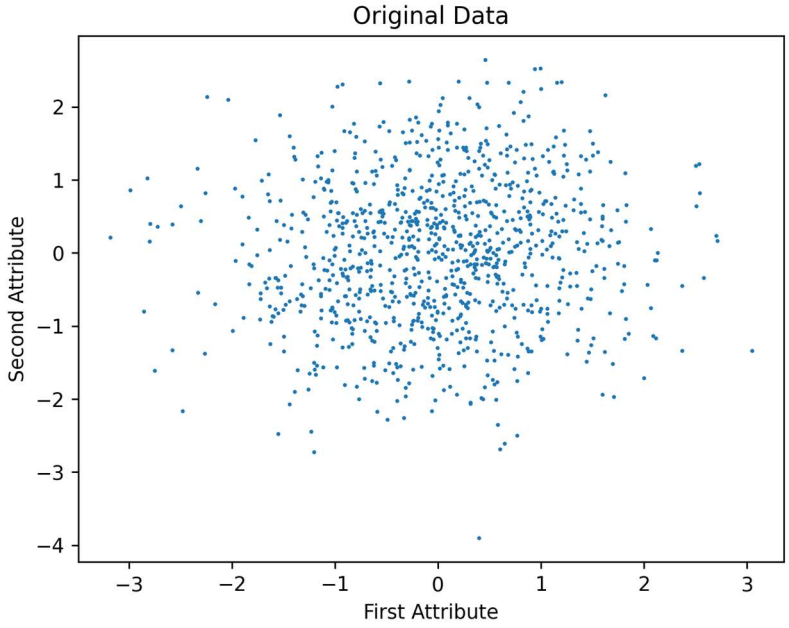
Homework 5: CSCI 347: Data Mining

Show your work. Include any code snippets you used to generate an answer, using comments in the code to clearly indicate which problem corresponds to which code.

1. [2 points] In Python, generate a (2-dimensional multivariate Gaussian) data matrix D using the following code:

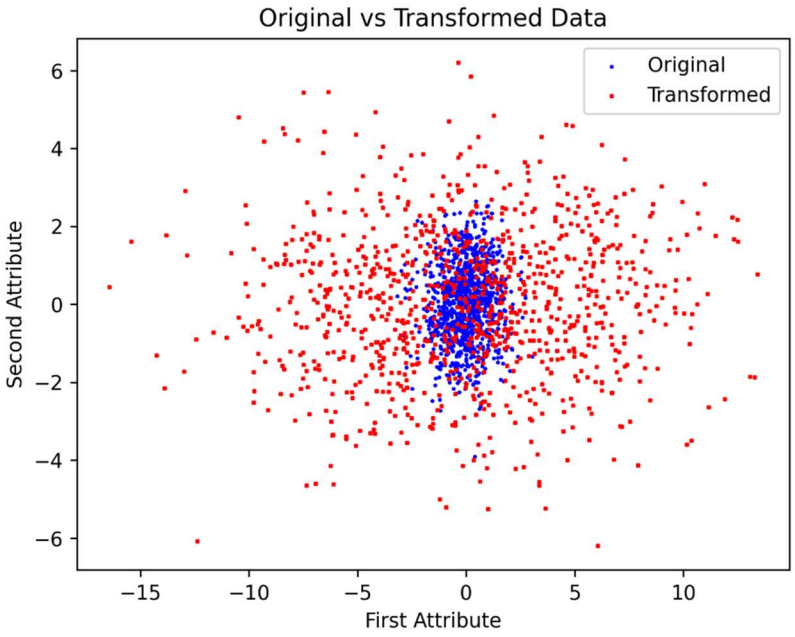
```
mu = np.array([0,0])
Sigma = np.array([[1,0], [0, 1]])
X1, X2 = np.random.multivariate_normal(mu, Sigma, 1000).T
D = np.array([X1, X2]).T
```

Create a scatter plot of the data, with the x-axis corresponding to the first attribute (column) in D, and the y-axis corresponding to the second attribute (column) in D.



2. [7 points] Using the scaling matrix and rotation matrix below to transform the data D from Question 1, by multiplying each data instance (row) x_i by RS. Let RSD be the matrix of the transformed data. That is, each 2-dimensional row vector x_i in D should be transformed into a 2-dimensional vector RSx_i in RSD.

   A. [4 points] Plot the transformed data RSD in the same figure as the original data D, using different colors to differentiate between the original and transformed data.



$$R = \begin{pmatrix} \cos(\vartheta) & -\sin(\vartheta) \\ \sin(\vartheta) & \cos(\vartheta) \end{pmatrix}, \text{ where } \vartheta = \frac{\pi}{4}, S = \begin{pmatrix} 5 & 0 \\ 0 & 2 \end{pmatrix}$$

B. [2 points] Write down the covariance matrix of the transformed data RSD.

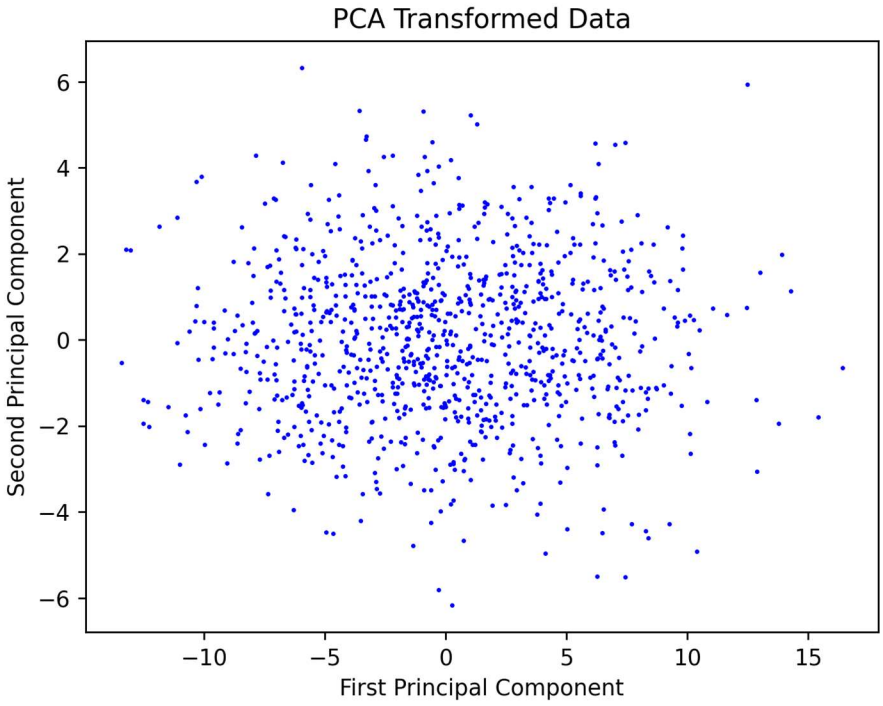
```
[25.104599   -0.46053743]
 [-0.46053743  3.82940008]]
```

C. [1 point] What is the total variance of the transformed data RSD?

```
      25.17817897000688
```

3. [8 points] Use sklearn's PCA function to transform the data matrix RSD from Question 2 to a 2-dimensional space where the coordinate axes are the principal components.

   A. [4 points] Plot the PCA-transformed data, with the x-axis corresponding to the first principal component and the y-axis corresponding to the second principal component.



   B. [2 points] What is the (sample) covariance matrix of the PCA-transformed data?

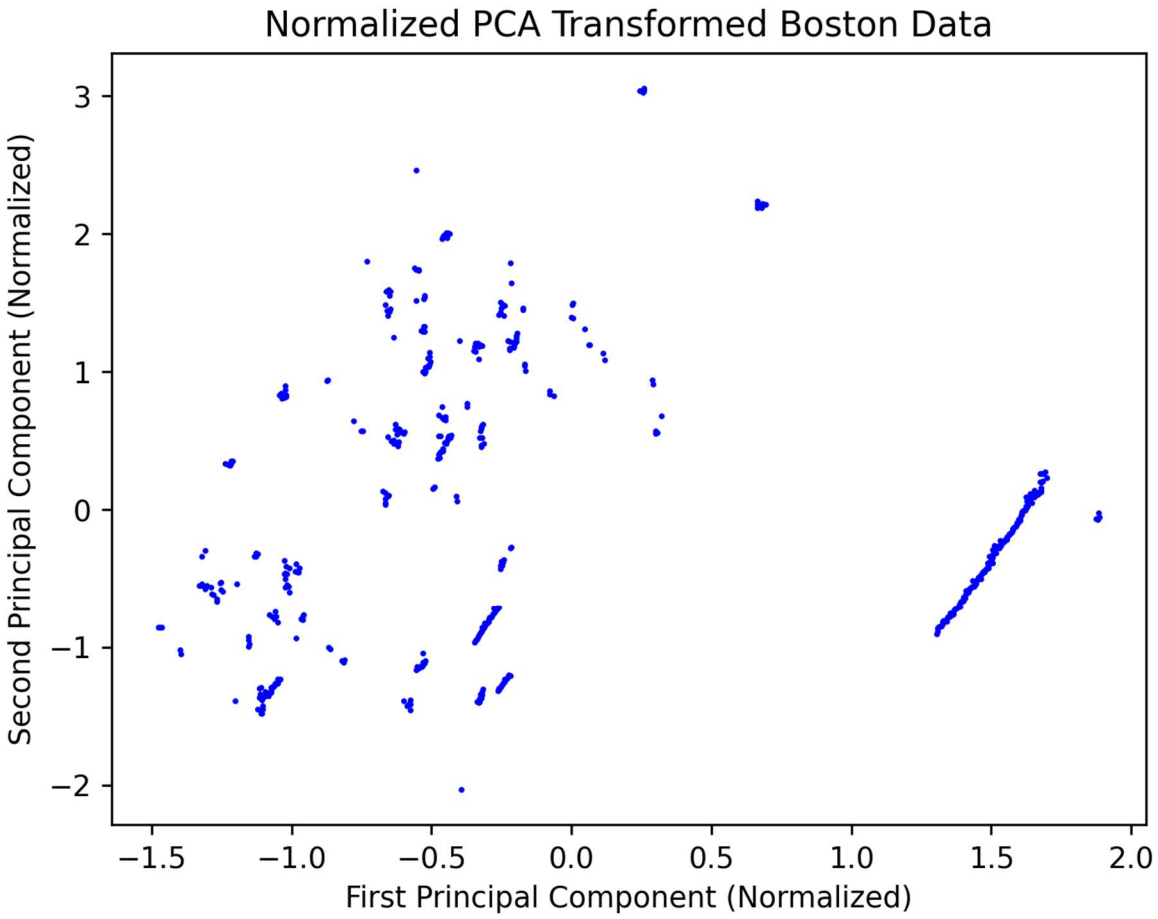
```
[[ 2.49690893e+01 -5.69003192e-17]
 [-5.69003192e-17  4.19462429e+00]]
```

   C. [2 points] What is the fraction of the total variance captured in the direction of the first principal component? What is the fraction of the total variance captured in the direction of the second principal component?

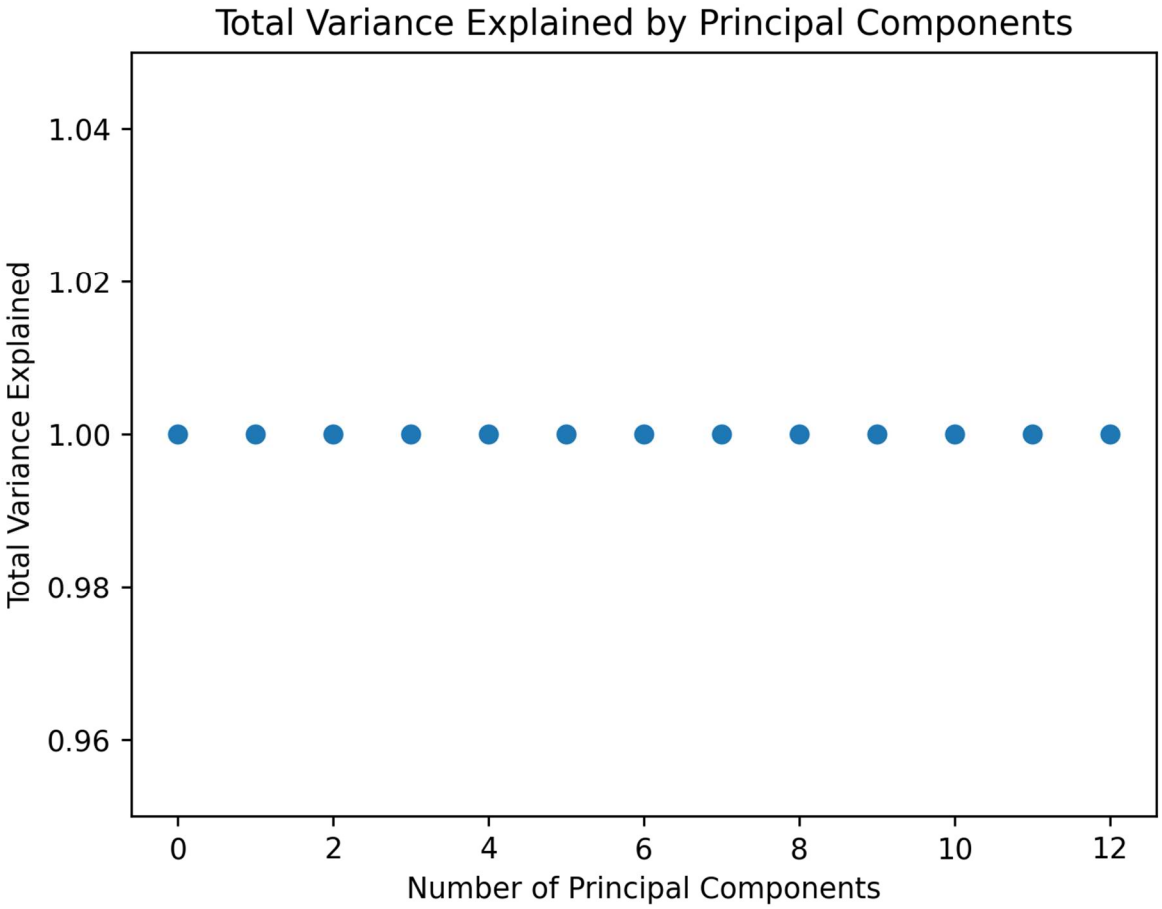
```
0.8673090216061319
0.13269097839386812
```

4. [18 points] Load the boston data set into Python (find the dataset in assignment section on D2L (under assignment 5). Use sklearn's PCA function to reduce the dimensionality of the data to 2 dimensions.

   1. [5 points] First, standard-normalize the data. Then, create a scatter plot of the 2-dimensional, PCA-transformed normalized Boston data, with the x-axis corresponding to the first principal component and the y-axis corresponding to the second principal component.



2. [3 points] Create a plot of the fraction of the total variance explained by the first components for $r = 1,\ldots,13$.



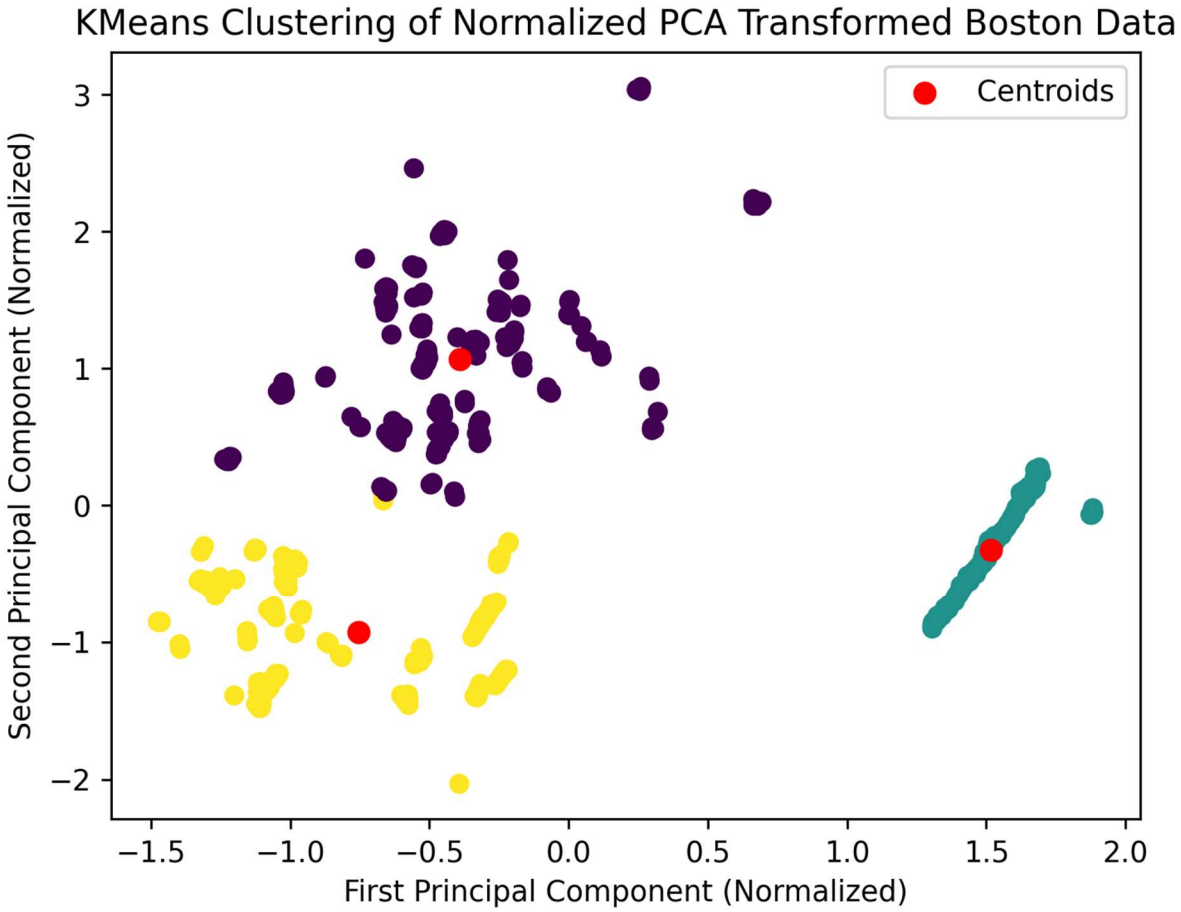
3. [2 points]
   1. [1 point] If we want to capture at least 90% of the variance of the normalized Boston data, how many principal components (i.e., what dimensionality) should we use?

```
(17.2580963142208+0j) and we would only need one principal component
```

   2. [1 point] If we use two principal components of the normalized Boston data, how much (what fraction or percentage) of the total variance do we capture?

```
(0.9999999999999994+1.9521546814784524e-17j)
```

4. [4 points] Use scikit-learn's implementation of k-means to find 2 clusters in the two-dimensional, PCA-transformed normalized Boston data set (the input to k-means should be the data that was plotted in part 4.1). Plot the 2-dimensional data with colors corresponding to predicted cluster membership for each point. On the same plot, also plot the the two means found by the k-means algorithm in a different color than the colors used for the data.



5. [4 points] Use scikit-learn's implementation of DBSCAN to find clusters in the two-dimensional, PCA-transformed normalized Boston data set (the input to DBSCAN should be the data that was plotted in part 4.1). Plot the 2-dimensional data with colors corresponding to predicted cluster membership for each point. Noise points should be colored differently than any of the clusters. How many clusters were found by DBSCAN?