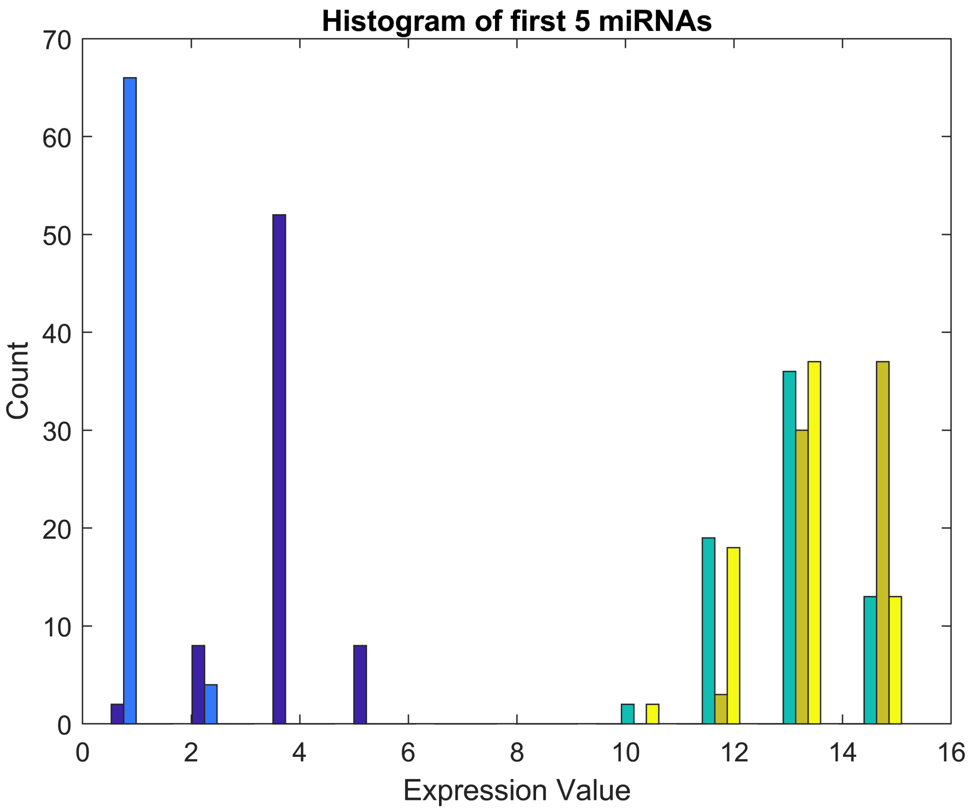
Unsupervised Learning In-Class Practice: Answer Key

Loading and examining the data

```
% Load the dataset
miRNA = readtable('miRNA_data.xlsx');
% Create a variable for patient IDs
patient_ID = miRNA.Patient_ID;
% Create a variable for patient health status
health_stat = miRNA.Health_Status;
% Create a variable for miRNA names
miRNA_names = miRNA.Properties.VariableNames(3:end);
% Create a variable for miRNA expression data
miRNA_data = table2array(miRNA(:,3:end));

% Histogram of miRNA expression data
figure
hist(miRNA_data(:,1:5))
xlabel('Expression Value'); ylabel('Count')
title('Histogram of first 5 miRNAs')
```



Answer: The expression level for two miRNAs is much lower that the other three. In other words, the distribution of expression level for different miRNAs within this dataset may be drastically different.

Determining optimal k value using KMC

```
% Demo for k-means clustering
```

```
idx = kmeans(miRNA_data,2);                 % identify clusters with k-means clustering
s = silhouette(miRNA_data,idx);             % determine silhouette values
s_score = mean(s)                           % calculate silhouette score
```

```
s_score = 0.5539
```

```
% Create a vector of k values ranging from 2 to 10
k_values = 2:10;

% For-loop to calculate the silhouette statistic for different k values
n = length(k_values);        % number of k values
s_score = zeros(n,1);        % variable to store silhouette statistic values
for i = 1:n
    % Use the kmeans function to cluster patients into k clusters
    idx = kmeans(miRNA_data,k_values(i));
    % Use the silhouette function to determine silhouette values
    s = silhouette(miRNA_data,idx);
    % Calculate the silhouette score by taking the mean
    s_score(i) = mean(s);
end
table(k_values',s_score)
```

ans = 9×2 table

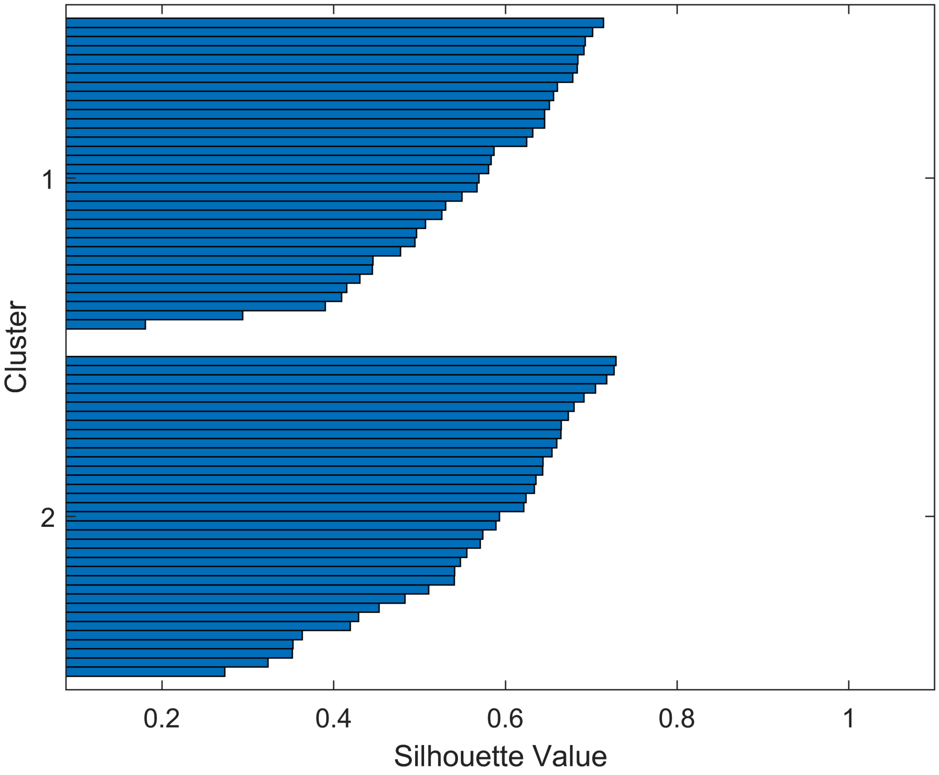
	Var1	s_score
1	2	0.5539
2	3	0.3850
3	4	0.3757
4	5	0.2745
5	6	0.2532
6	9	0.2426
7	7	0.2411
8	10	0.2087
9	8	0.1910

```
k = find(s_score==max(s_score))+1
```

```
k = 2
```

Answer: The max silhouette statistic is equal to 0.5539, and this corresponds with k = 2.

```
% Silhouette plot based on best k value
idx = kmeans(miRNA_data,k);
silhouette(miRNA_data,idx)
```



Answer: No negative values are present, but a negative silhouette value would imply that its associated observation was not well-placed into its assigned cluster.

```
cluster_1 = health_stat(idx == 1)
```

```
cluster_1 = 35×1 cell array
'healthy'
'healthy'
'healthy'
'healthy'
'healthy'
'healthy'
'healthy'
'healthy'
'healthy'
'healthy'
      ⋮
```

```
cluster_2 = health_stat(idx == 2)
```

```
cluster_2 = 35×1 cell array
'diseased'
'diseased'
'diseased'
'diseased'
'diseased'
'diseased'
'diseased'
'diseased'
'diseased'
'diseased'
```

```
       ⋮
```

Answer: The patients cluster based on health status (for the most part).

Visualizing data in lower dimension using PCA

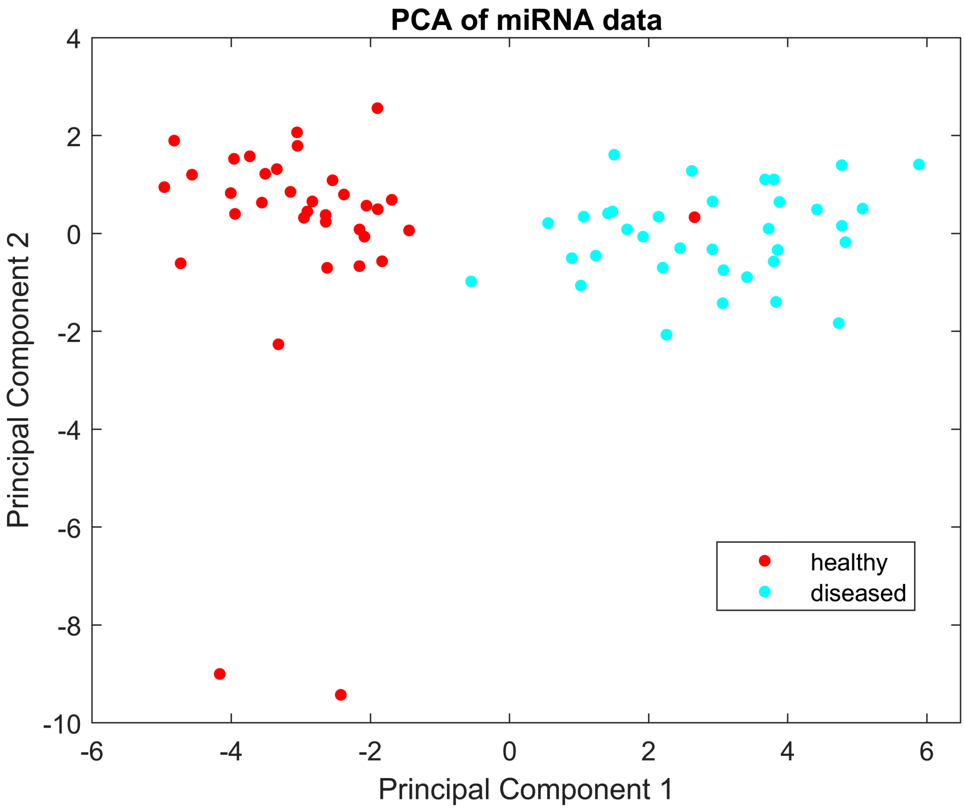
```
% Do PCA (determine PC matrix, transformed data matrix, and variance explained)
[P, Y, ~, ~, explained] = pca(miRNA_data)
```

```
P = 18×18
    0.0300    0.1534    0.3029    0.1504    0.0663    0.1182   -0.4002   -0.0045 ···
    0.0240   -0.0108    0.0908    0.0733    0.0956   -0.0254    0.2192    0.1598
   -0.2578   -0.0717    0.1162    0.1293   -0.0636    0.1584   -0.0709   -0.0355
   -0.2587   -0.0726    0.1167    0.1277   -0.0606    0.1605   -0.0716   -0.0386
   -0.2564   -0.0736    0.1184    0.1276   -0.0624    0.1609   -0.0747   -0.0376
   -0.1416   -0.3361    0.0930    0.0046    0.0537    0.1778   -0.1483    0.2345
   -0.3000   -0.2695    0.0780    0.2354    0.1714   -0.0044    0.5704    0.4512
    0.0070   -0.4171    0.1880    0.0910    0.0903   -0.2794   -0.1286   -0.1692
   -0.2214   -0.0452    0.0729    0.1145   -0.2147    0.1620   -0.0115   -0.0622
   -0.2845    0.3039    0.2076    0.3493   -0.1719    0.1774   -0.1289   -0.1124
       ⋮

Y = 70×18
   -4.8206    1.8952    0.2486    0.3666   -0.3778   -0.9746   -0.2439   -0.0500 ···
   -2.6433    0.3761    0.3641    1.1508   -1.7826    0.6302    0.3440   -0.1008
   -2.9064    0.4478   -0.5345    0.8140   -0.8675   -0.9135   -0.0047    0.1337
    4.7762    1.3941    0.2017   -3.8618    1.7310    0.2733   -0.0400   -1.1672
    1.6909    0.0815   -0.2951    0.5706    1.0015    0.7312   -0.2389   -0.3995
    3.8011    1.1033    0.3089    1.3176    0.5251   -1.1256   -0.4901   -0.4099
   -4.0075    0.8242    0.7472    0.8301   -0.9192    0.3098    0.6441   -0.0136
   -3.9458    0.3981   -0.1171   -1.4169    0.7272    0.5166   -0.0070   -0.1275
    5.0765    0.5061   -2.6832   -1.0818    0.0429   -0.7550    0.9973    0.2080
    1.0668    0.3413   -1.6572   -0.4395   -0.6223    1.0007   -0.8566    0.8757
       ⋮

explained = 18×1
   46.9119
   15.7222
   10.7405
    7.2704
    5.5997
    3.6613
    2.0605
    1.7339
    1.5851
    1.3276
       ⋮
```

```
% Visualize transformed data on PC1 vs. PC2 plot (label = patient health status)
figure
gscatter(Y(:,1), Y(:,2), health_stat)
xlabel('Principal Component 1')
ylabel('Principal Component 2')
title('PCA of miRNA data')
```



Answer: There seems to be two distinct clusters with two outliers. PC1 and PC2 account for 62.6% of the variance.

Answer: Yes, it seems like the clusters we see from PCA match with our results from k-means clustering if we separate the data along PC1.

```
% Create table of genes matched to coefficients of best PC
PC1 = P(:,1);
table(miRNA_names',PC1)
```

```
ans = 18×2 table
```

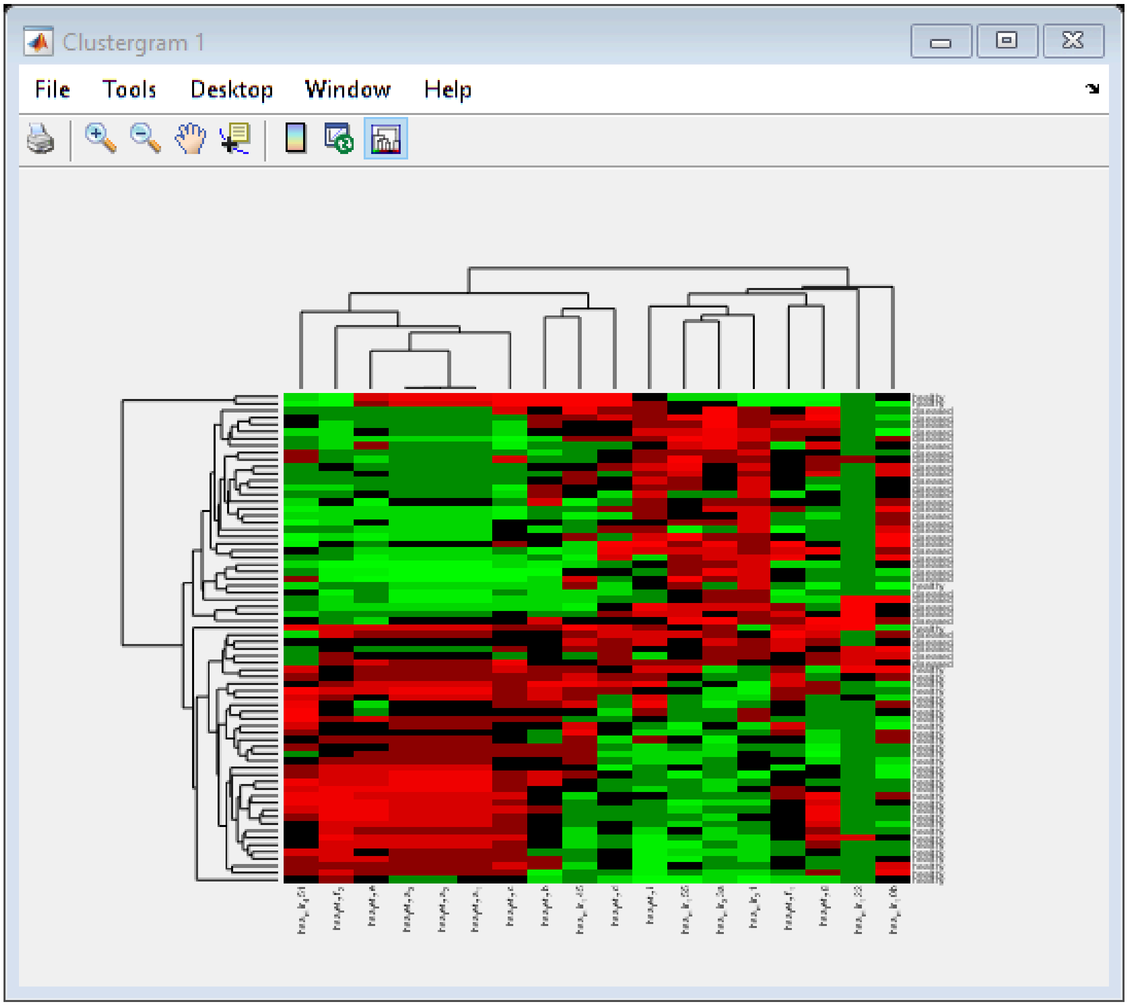
	Var1	PC1
1	'hsa_mir_33a'	0.4131
2	'hsa_mir_21'	0.3526
3	'hsa_mir_155'	0.2916
4	'hsa_mir_10b'	0.1040
5	'hsa_let_7i'	0.0786
6	'hsa_let_...	0.0300
7	'hsa_mir_122'	0.0240
8	'hsa_let_7d'	0.0070
9	'hsa_let_7g'	0.0058

	Var1	PC1
10	'hsa_mir_145'	-0.0859
11	'hsa_let_7b'	-0.1416
12	'hsa_let_7e'	-0.2214
13	'hsa_let_...	-0.2564
14	'hsa_let_...	-0.2578
15	'hsa_let_...	-0.2587
16	'hsa_let_...	-0.2845
17	'hsa_let_7c'	-0.3000
18	'hsa_mir_451'	-0.3937

Answer: miRNA 33a has the greatest weight (0.4131) for PC1.

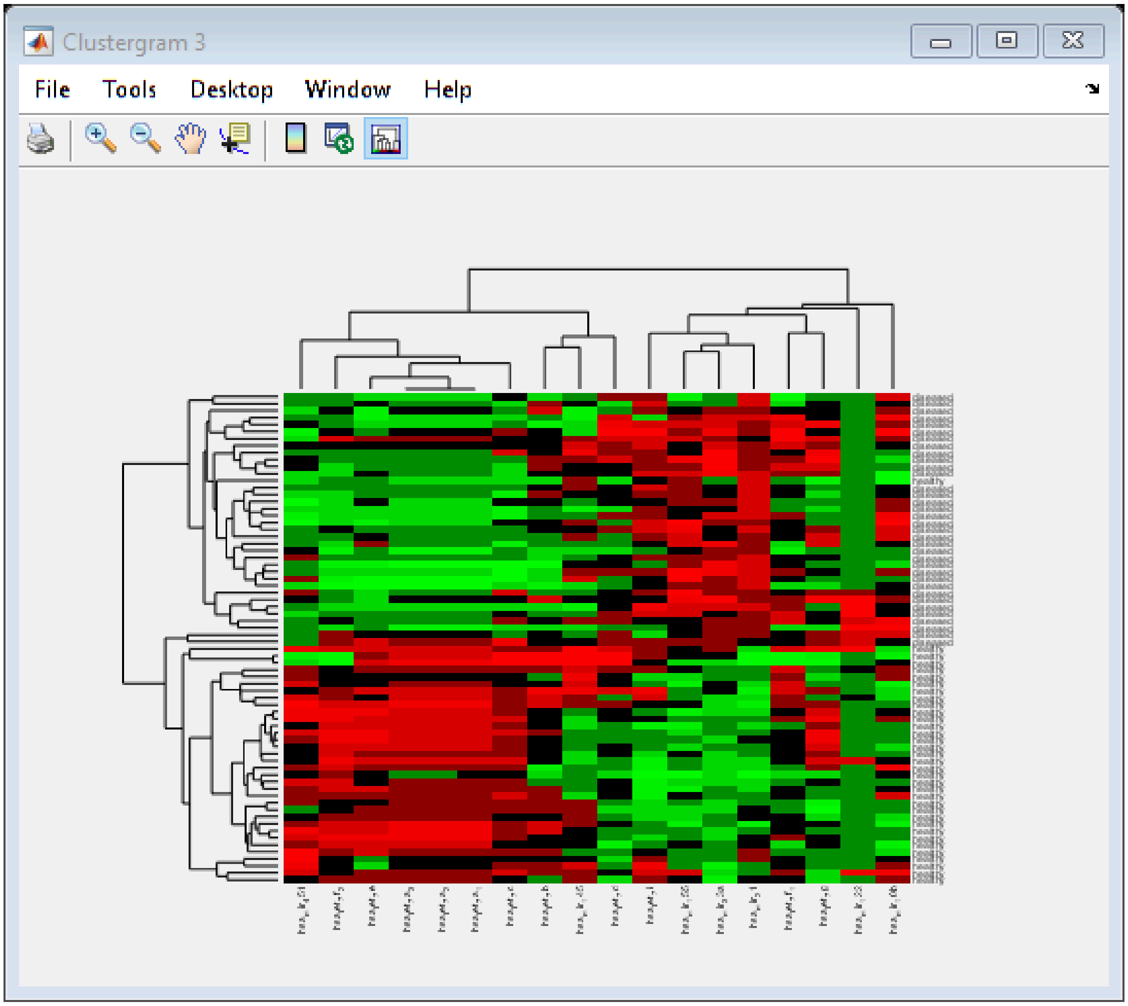
Identifying differentially expressed miRNAs using HC

```
% Create clustergram (label rows, label columns, and standardize by columns)
miRNA_cg = clustergram(miRNA_data,...
    'RowLabels',health_stat,...
    'ColumnLabels',miRNA_names,...
    'Standardize','column');
```



Answer: It seems that the outliers were picked out at the edges, but the two clusters we've seen before are not clearly separated.

```
% Create clustergram (label rows, label columns, standardize by columns
% and use correlation as distance metric for rows and columns)
miRNA_cg_corr_rbc = clustergram(miRNA_data,...
    'RowLabels',health_stat,...
    'ColumnLabels',miRNA_names,...
    'Standardize','column',...
    'RowPDist','correlation',...
    'ColumnPDist','correlation');
```



<u>Answer</u>: As opposed to before, this clustergram better separates patients according to their health status. This matches closer to our previous results from KMC and PCA.

<u>Answer</u>: Based on the clustergram, the following miRNAs appear to be differentially expressed between healthy and diseased individuals: miR-451, *Let*-7, miR-155, miR-33, and miR-21.