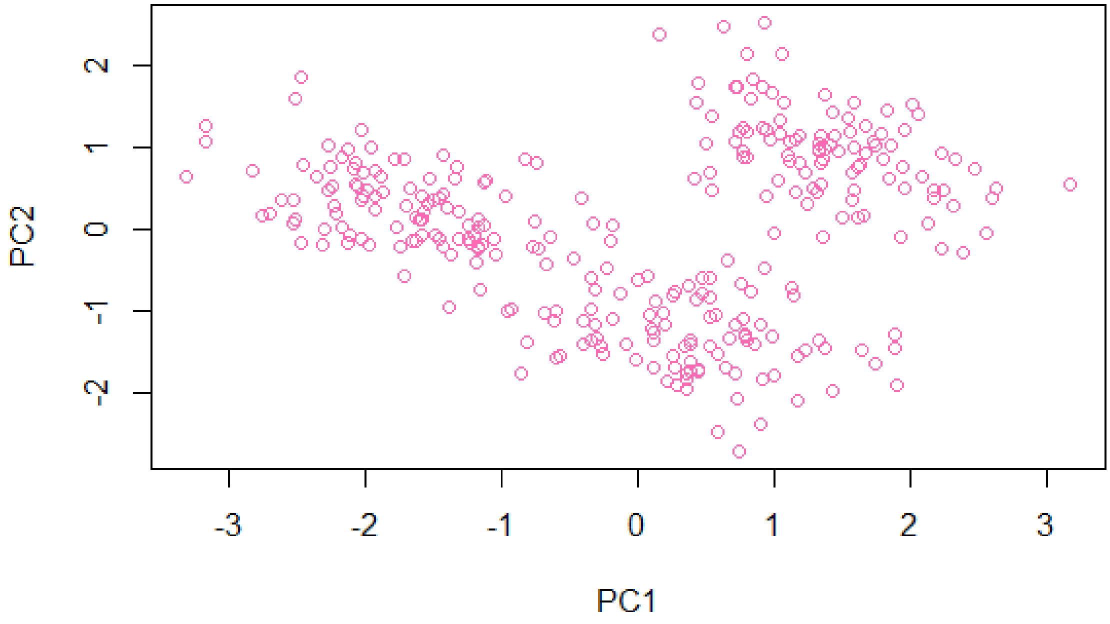
Simulate a sample of 300 observations which include 4 variables X_1, X_2, X_3, X_4 from 3 different clusters (each cluster includes 100 observations) with centroids defined as

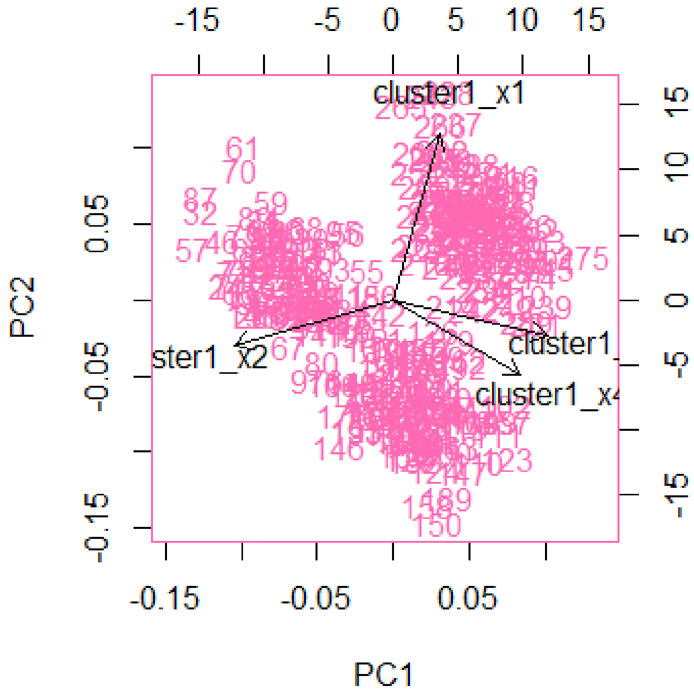
$$\begin{gathered}
Cluster\ 1: n = 100;\ X_1 \sim N(0,1); X_2 \sim N(2.5,1);\ X_3 \sim N(-2.5,1); X_4 \sim N(1,1); \\
Cluster\ 2: n = 100;\ X_1 \sim N(-2.5,1); X_2 \sim N(0,1);\ X_3 \sim N(1,1); X_4 \sim N(2.5,1); \\
Cluster\ 3: n = 100;\ X_1 \sim N(2.5,1); X_2 \sim N(-2.5,1);\ X_3 \sim N(1,1); X_4 \sim N(2.5,1);
\end{gathered}$$

Use this simulated dataset (with dimension of 300 × 4) for the following analysis. Standardize the data if appropriate.

The standard deviations of the variables were quite close to each other, so there was no need for standardization of the data.

Generate principal components of the variables, and plot the first two PCs. Does the plot of the first two PCs suggest that there are 3 clusters in the data?



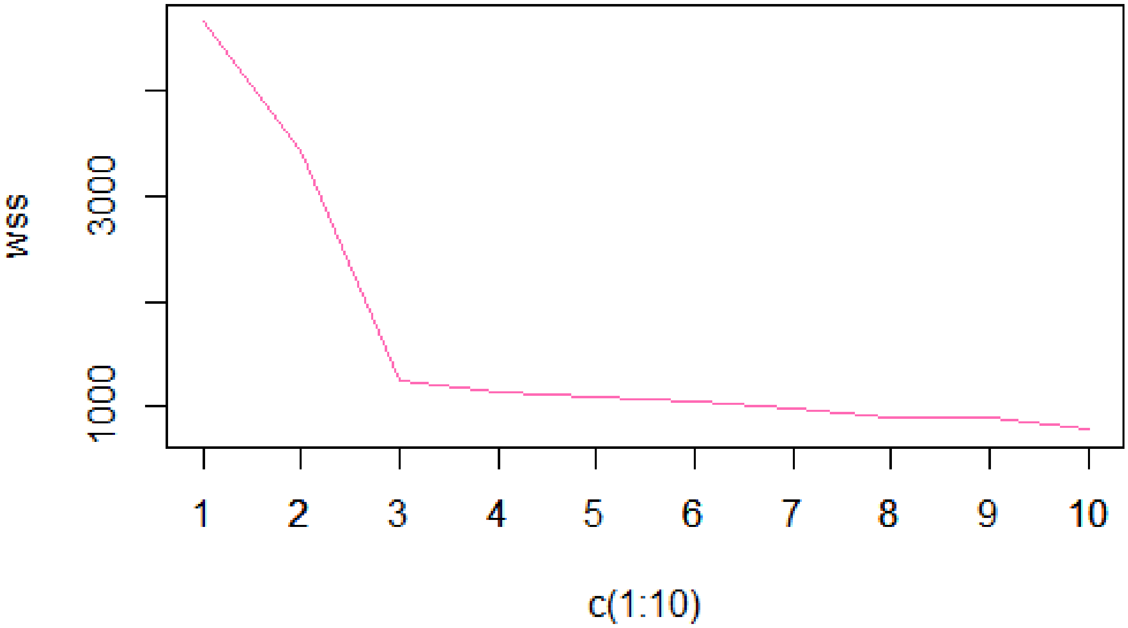


The plot of the first two principal components seem to suggest that there are 3 clusters in the data. X1 forms one cluster, X2 forms the second cluster, and X3 and X4 form the third cluster.

Use k-means clustering to discover groups in the data.

a. Analyze the within cluster sum of squares (WSS) for different numbers of clusters (i.e. vary the number of k) to decide what is the most likely number of clusters. How many potential clusters in the data?
   We will assume there are two clusters for b and c.
b. Generate the cluster centroids assuming there are two clusters.
c. Visualize the clusters using the first two PCs and different color representing different clusters assuming there are two clusters.

Part A

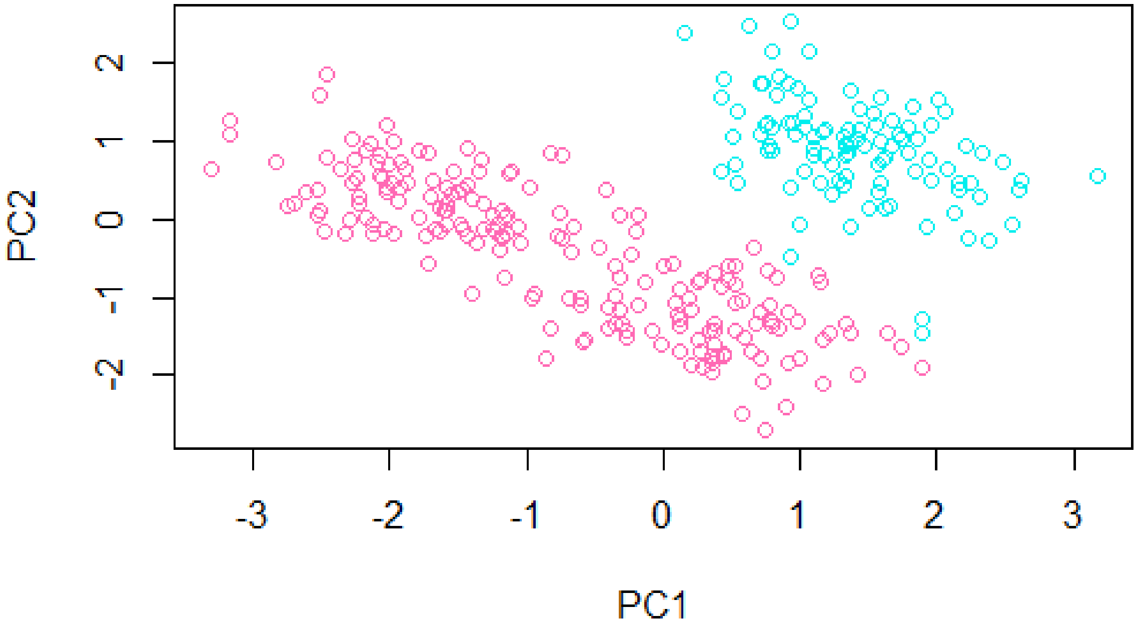


The graph of WSS for different number of clusters suggest that there is most likely 3 clusters.

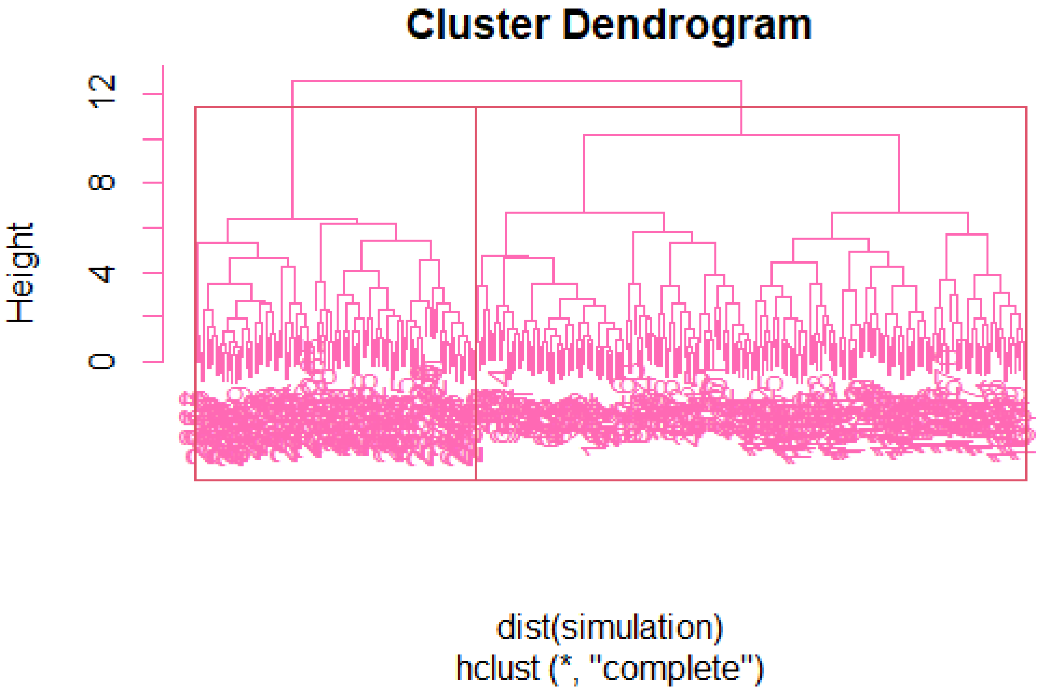
Part B

$$cluster\ means = \begin{bmatrix} 2.388625 & -2.464411 & 1.0110115 & 2.386470 \\ -1.226704 & 1.267717 & -0.8835077 & 1.743877 \end{bmatrix}$$

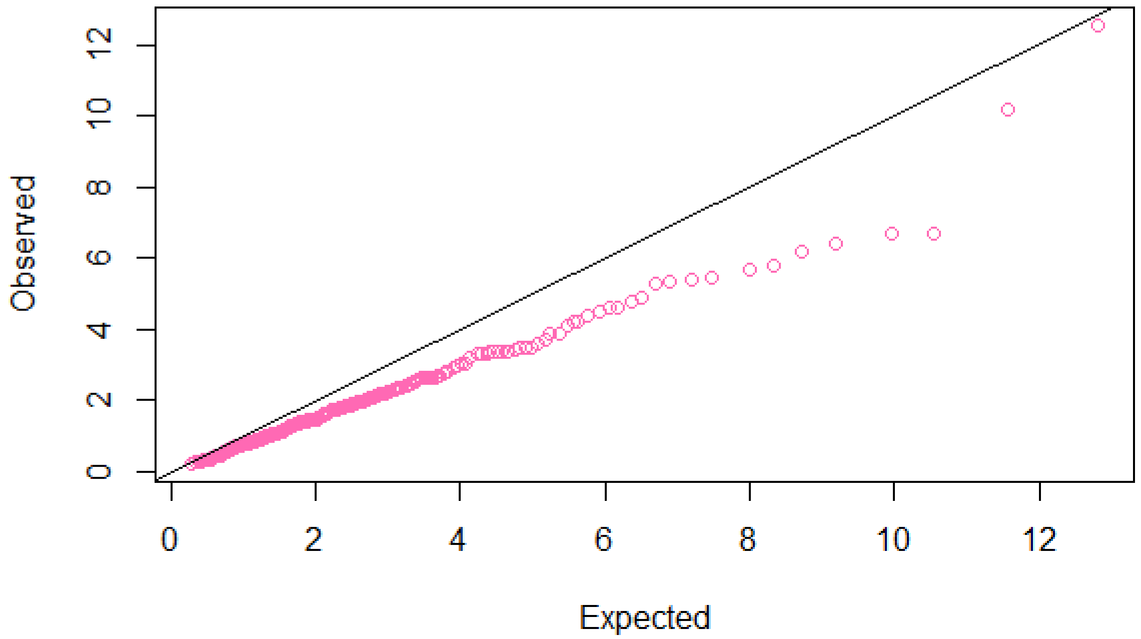
Part C

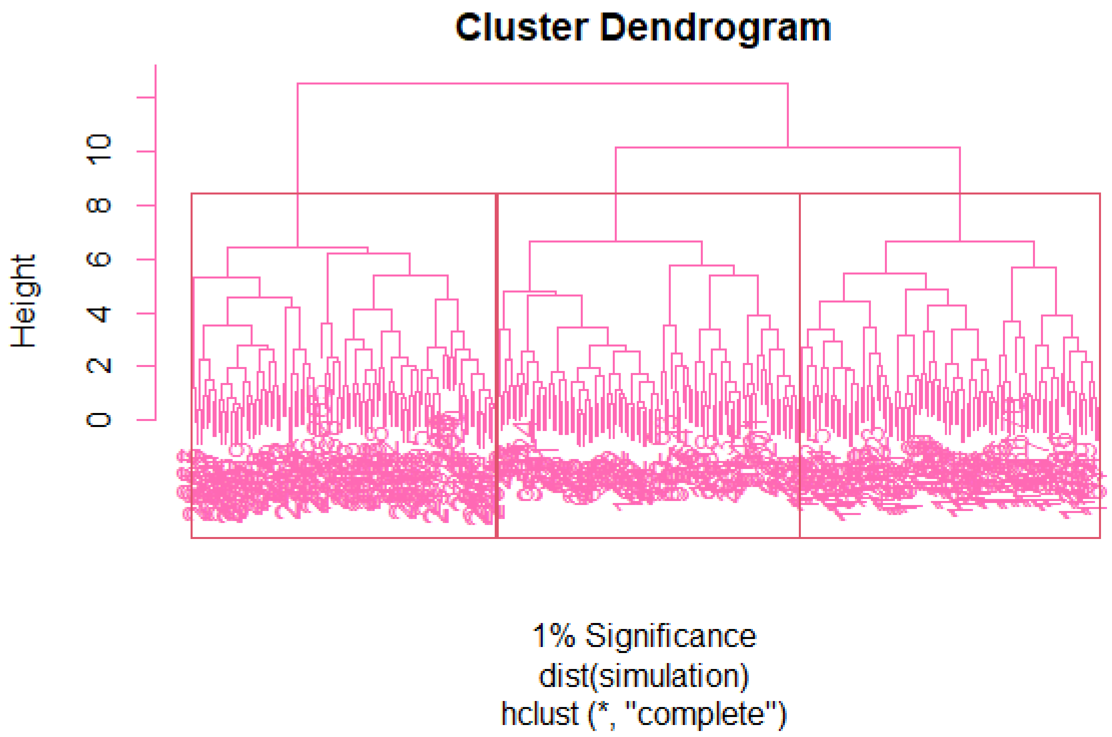


Use hierarchical clustering with complete linkage to generate a dendrogram of the data.



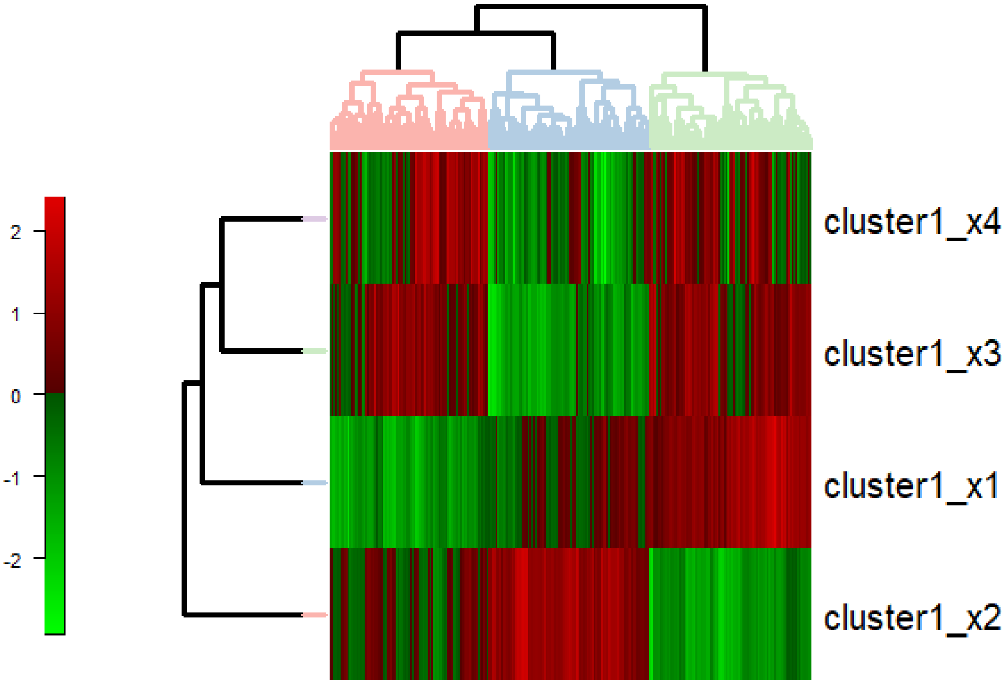
Use the resampling technique shown in class to decide whether there are evident clusters.





The Q-Q plot shows significant departure from the reference line, which suggests there are clusters. At 1% significance, cutting at a height of 9.976784 gives 3 clusters.

Generate a heatmap of the data and show the significant clusters based on the resampling approach at 1% level of significance.



There are 3 significant clusters at the 1% significance level

Compare the agreement for the clusters discovered with k-means clustering and hierarchical clustering based on resampling approach at 1% level of significance.

		Hierarchical Clustering		
		1	**2**	**3**
K-Means Clustering	**1**	100	0	0
	2	0	99	0
	3	0	0	101

Comparing k-means clustering and hierarchical clustering, all 300 observations have consistent clustering sorting.