EXERCISE 11.2

Ramsey King

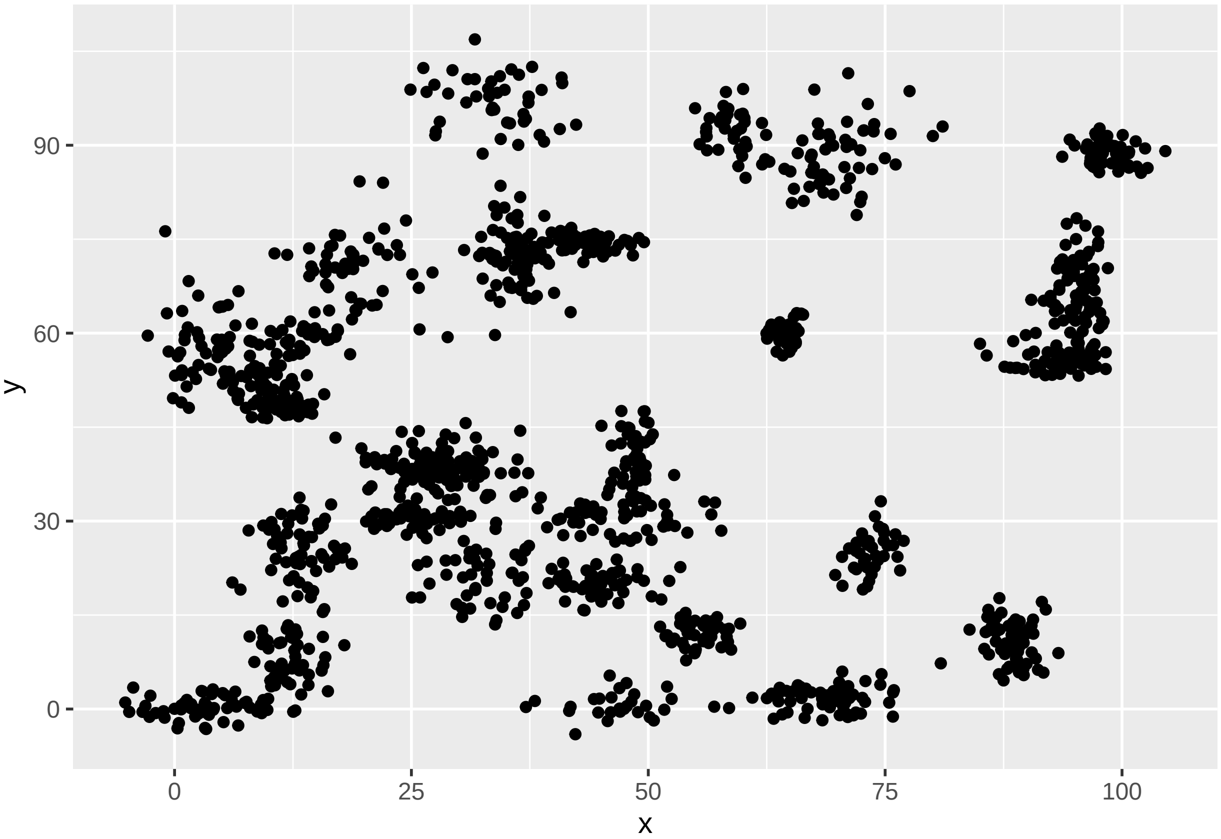
2021-05-29

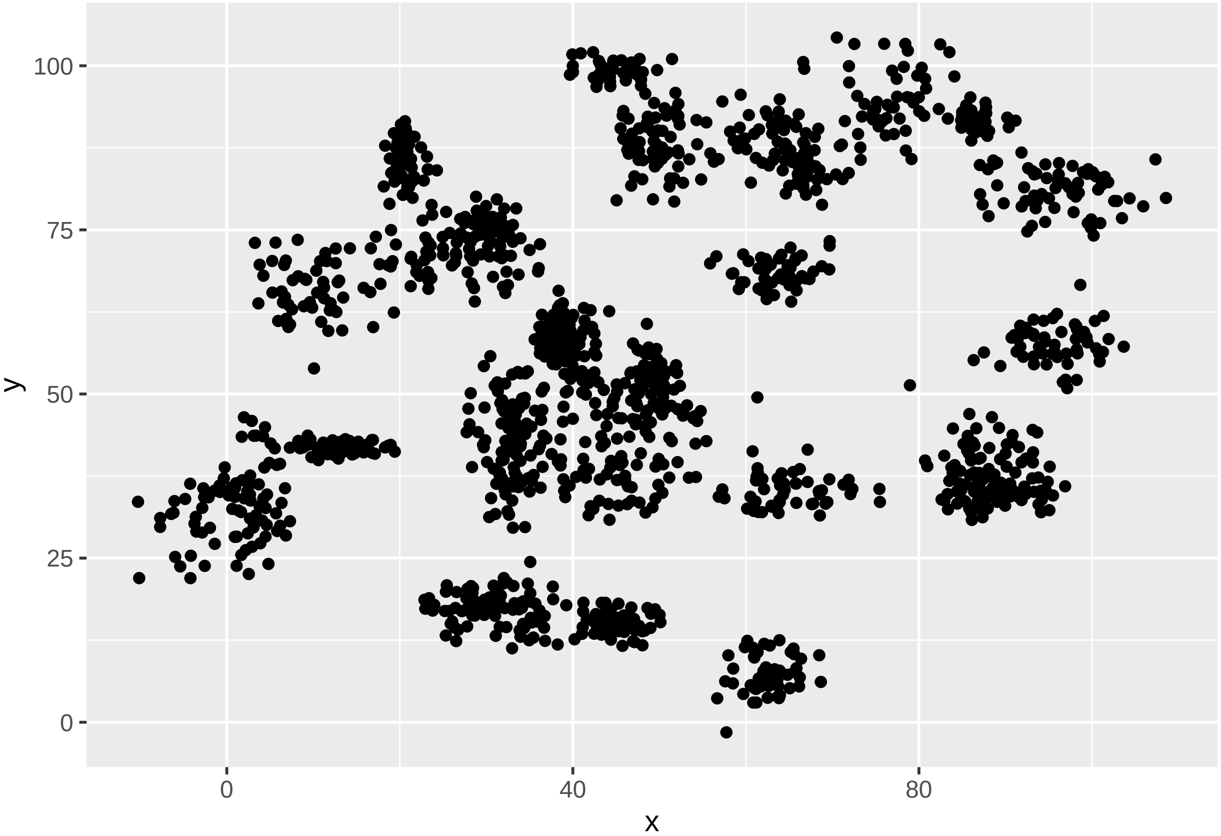
In this problem, you will use the nearest neighbors algorithm to fit a model on two simplified datasets.

The first dataset (found in binary-classifier-data.csv) contains three variables; label, x, and y. The label variable is either 0 or 1 and is the output we want to predict using the x and y variables (You worked with this dataset last week!).

The second dataset (found in trinary-classifier-data.csv) is similar to the first dataset except that the label variable can be 0, 1, or 2.

1. Plot the data from each dataset using a scatter plot.

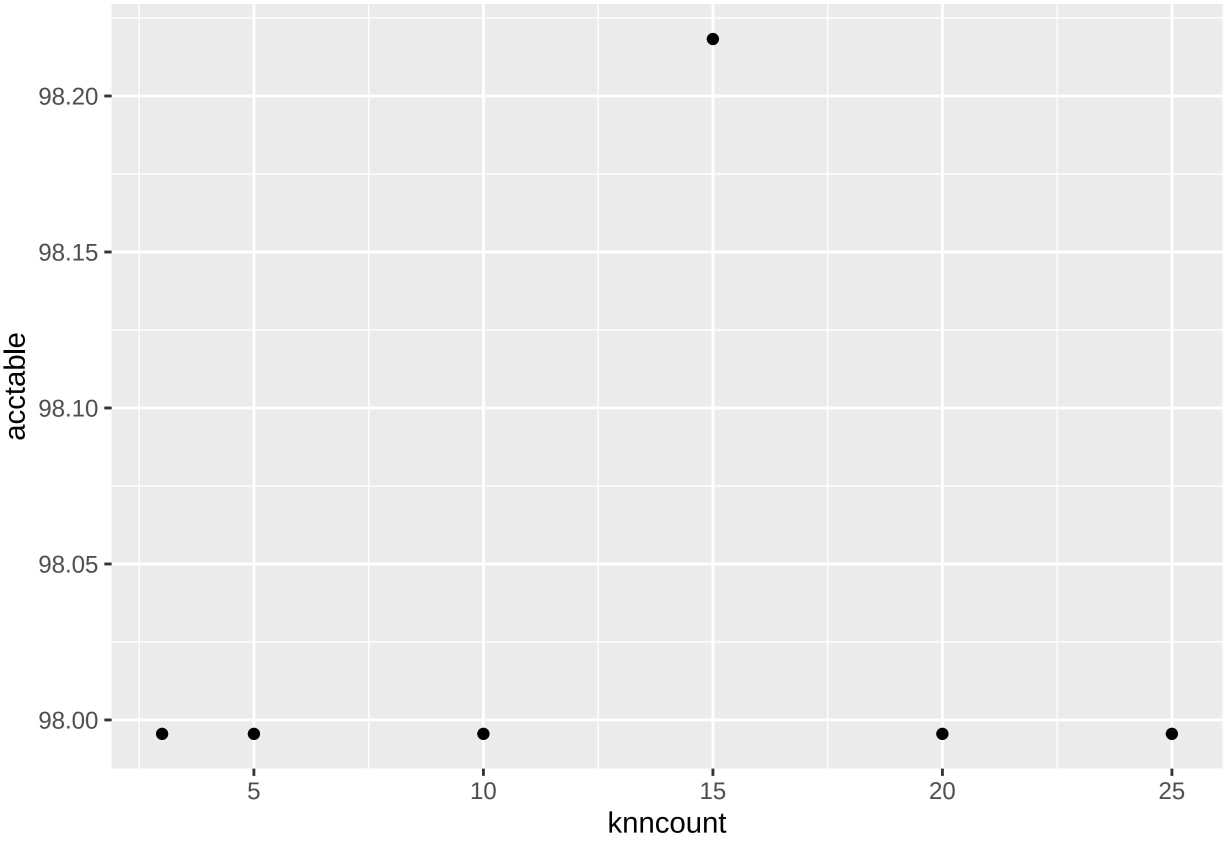




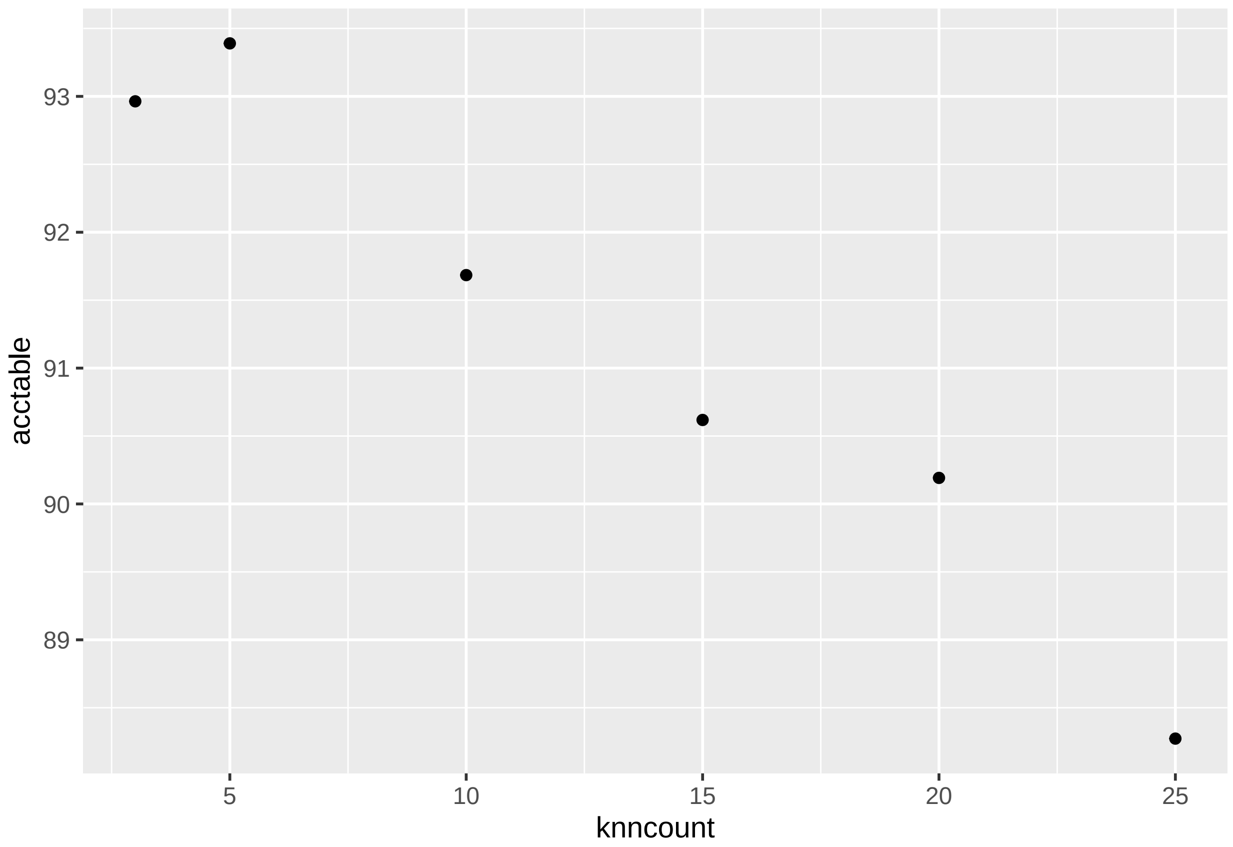
2. Fit a k nearest neighbors' model for each dataset for k=3, k=5, k=10, k=15, k=20, and k=25. Compute the accuracy of the resulting models for each value of k. Plot the results in a graph where the x-axis is the different values of k and the y-axis is the accuracy of the model.

```
## Warning: package 'caret' was built under R version 4.0.5
```

```
## Loading required package: lattice
```



To me, I think something is incorrect with my code because the accuracy values are really high, and very similar to each other. I did the best I could to write the code in the manner that would help me to obtain the correct values for accuracy.



This graph looks more realisitic in terms of accuracy due to the values being more varied.

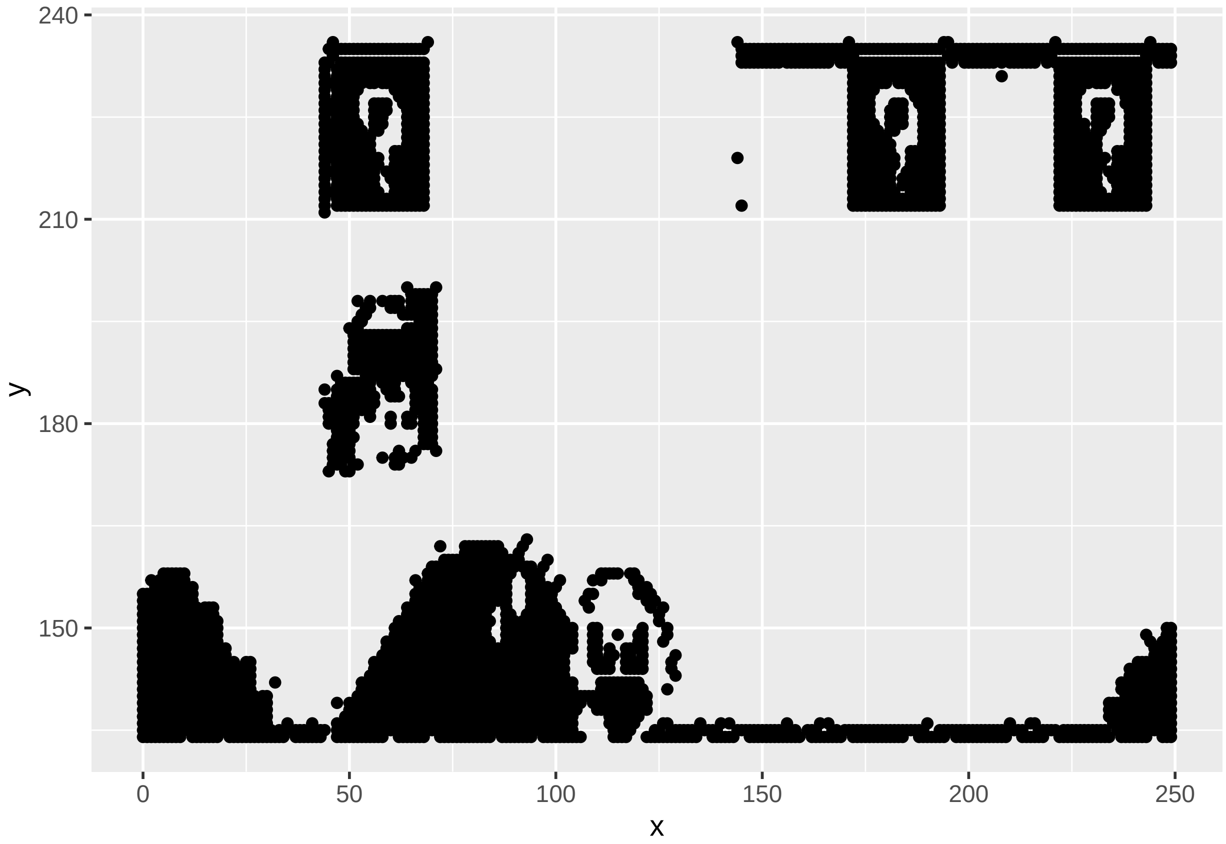
3. Looking back at the plots of the data, do you think a linear classifier would work well on these datasets?

I do not think that a linear classifer would work well on these data sets due to the sporadic nature and clustering of the data points.

4. How does the accuracy of your logistic regression classifier from last week compare? Why is the accuracy different between these two methods?

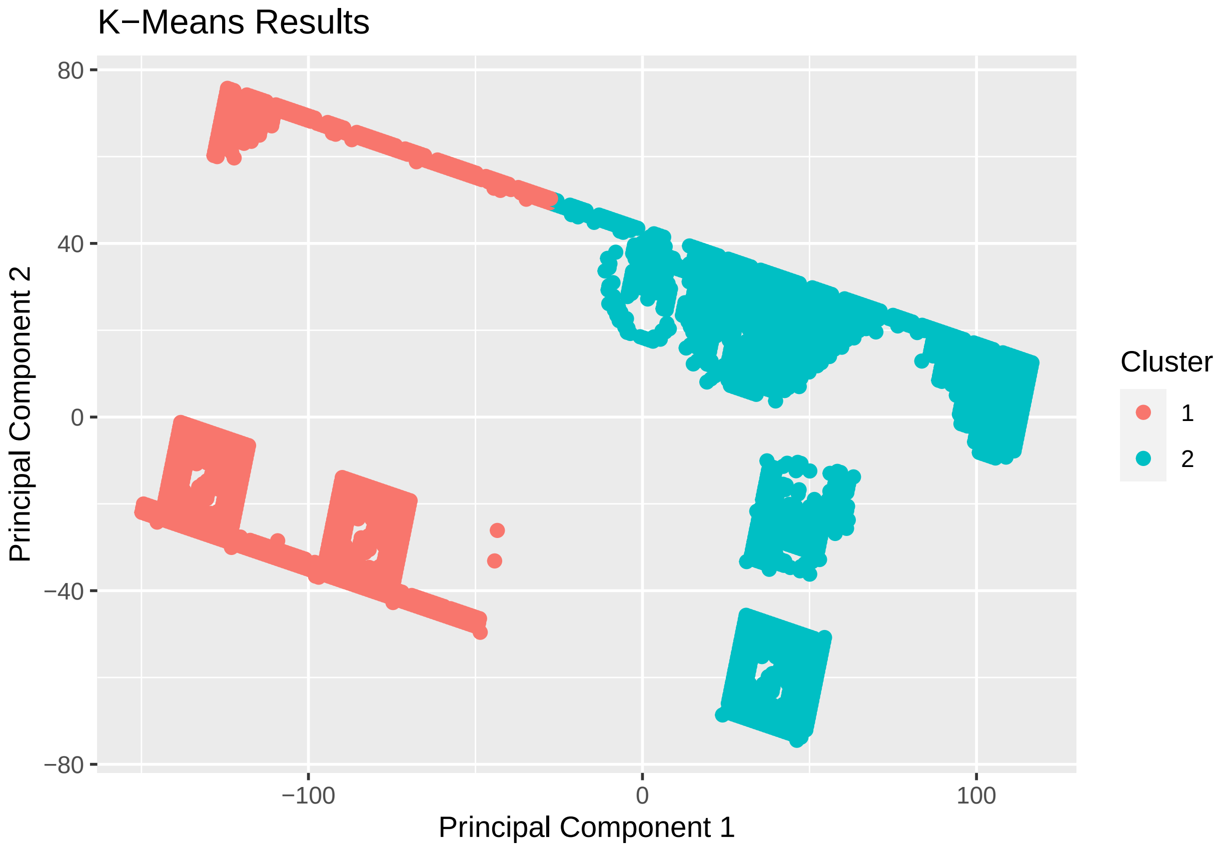
For the binary data set, the logistic regression classifer had an accuracy of 0.601 or 60.1%. This week, the accuracy was 99%. I believe that the accuracy is different because of the ability to use a non-linear method of prediction versus a linear method.

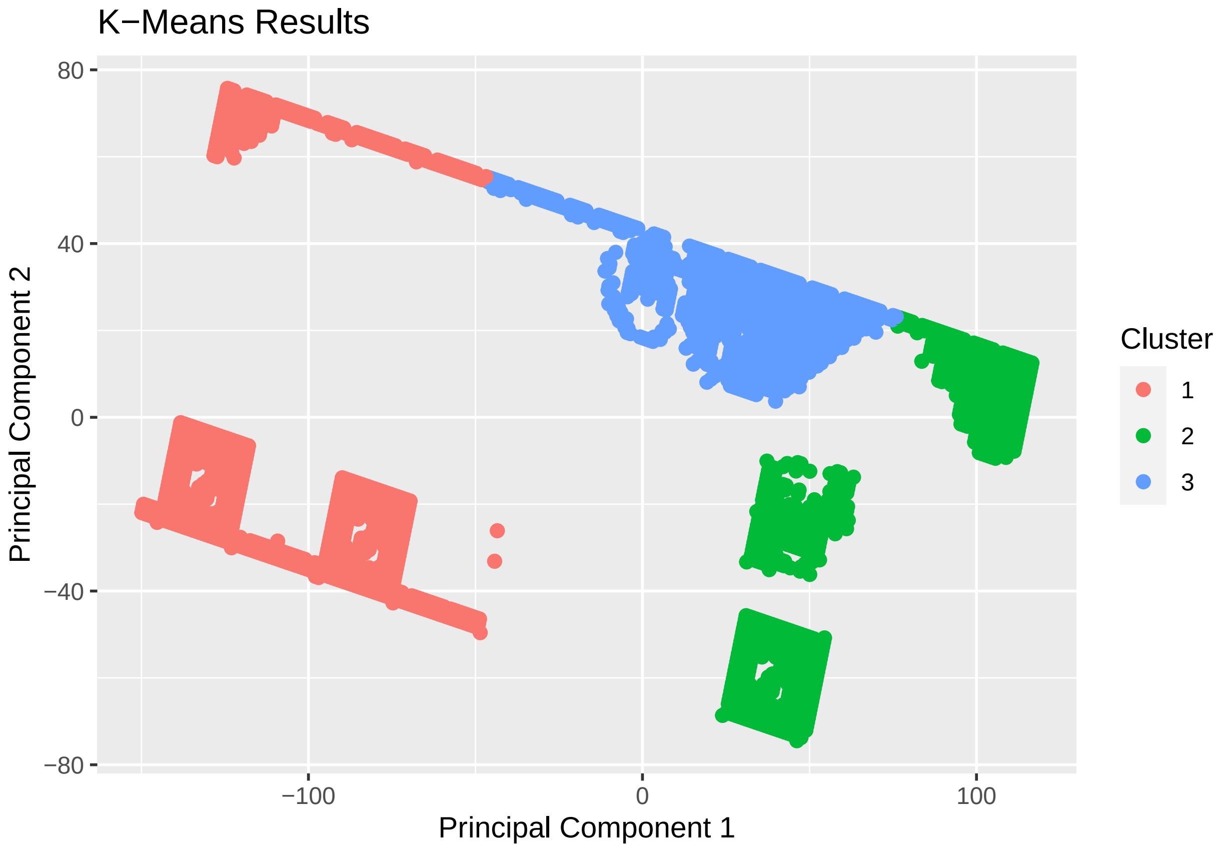
5. In this problem, you will use the k-means clustering algorithm to look for patterns in an unlabeled dataset. The dataset for this problem is found at data/clustering-data.csv. Plot the dataset using a scatter plot.

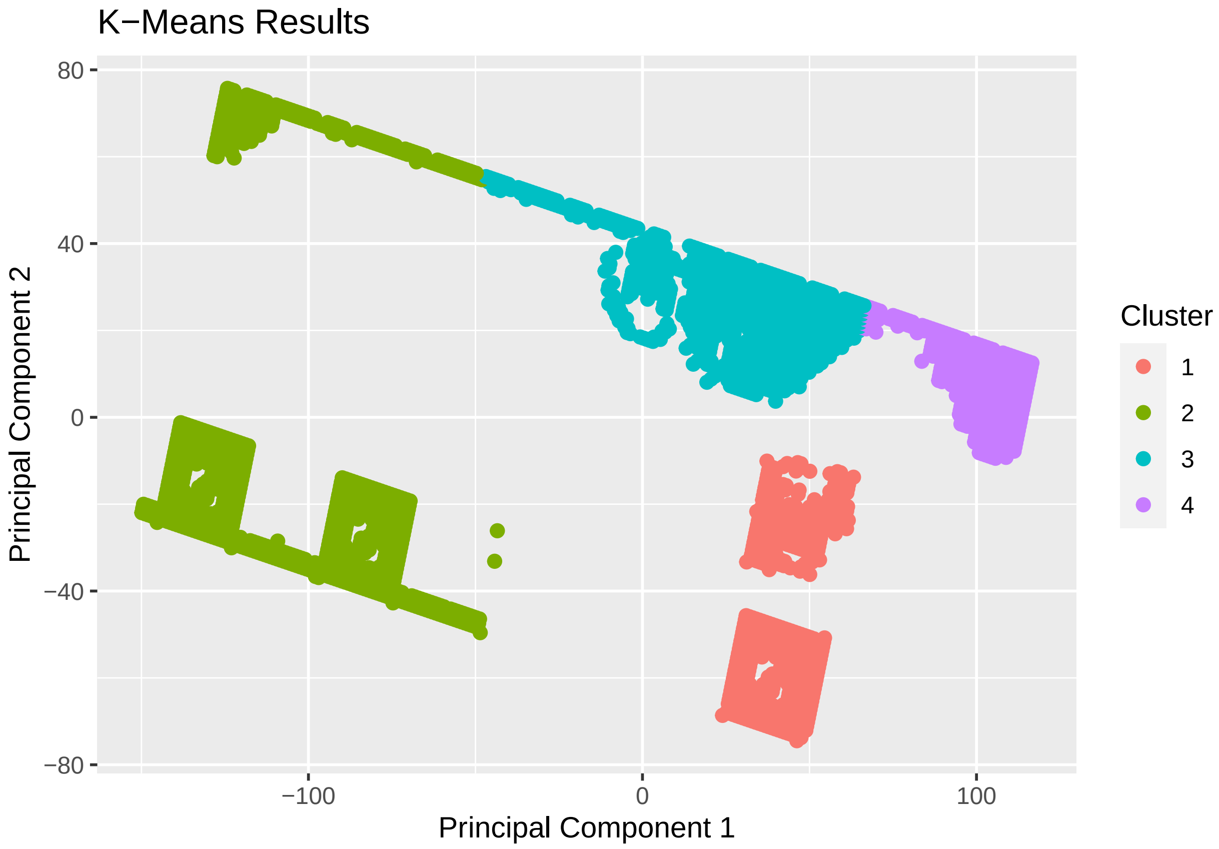


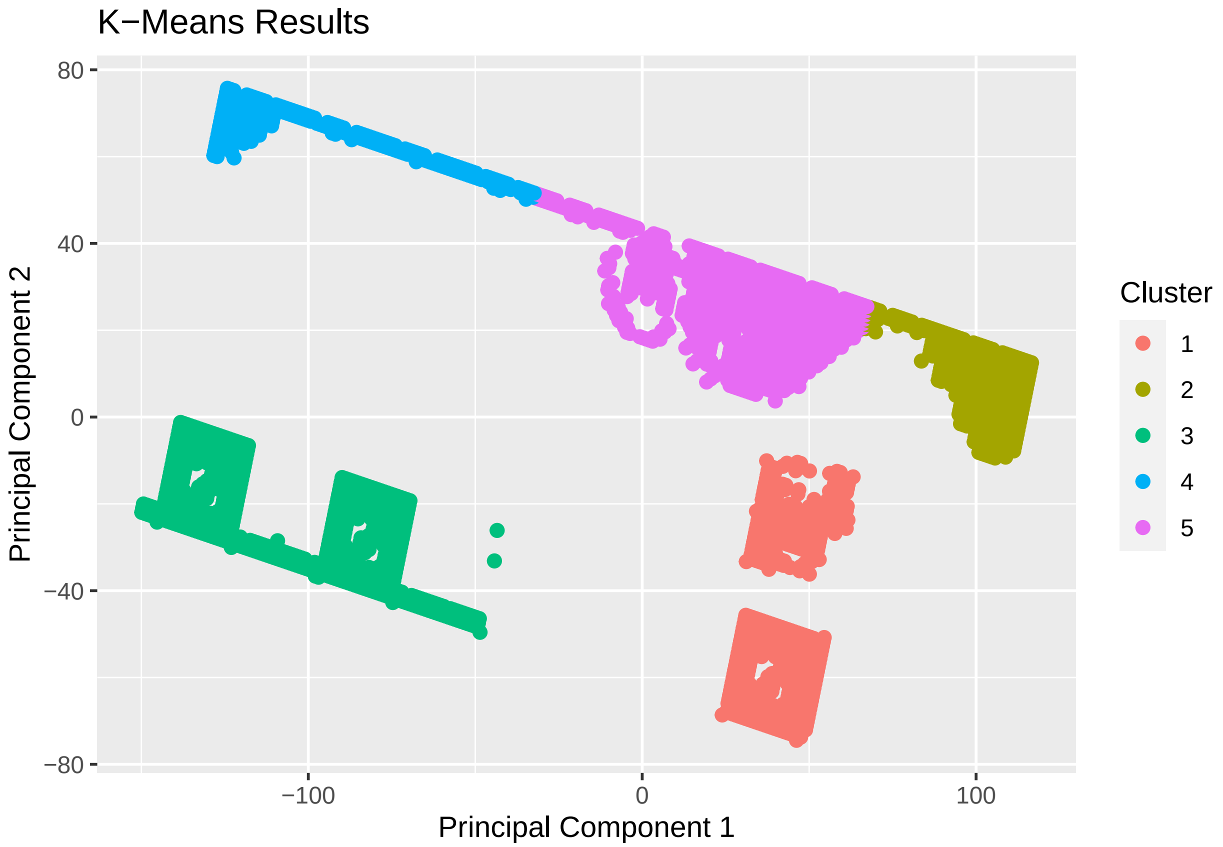
6. Fit the dataset using the k-means algorithm from k=2 to k=12. Create a scatter plot of the resultant clusters for each value of k.

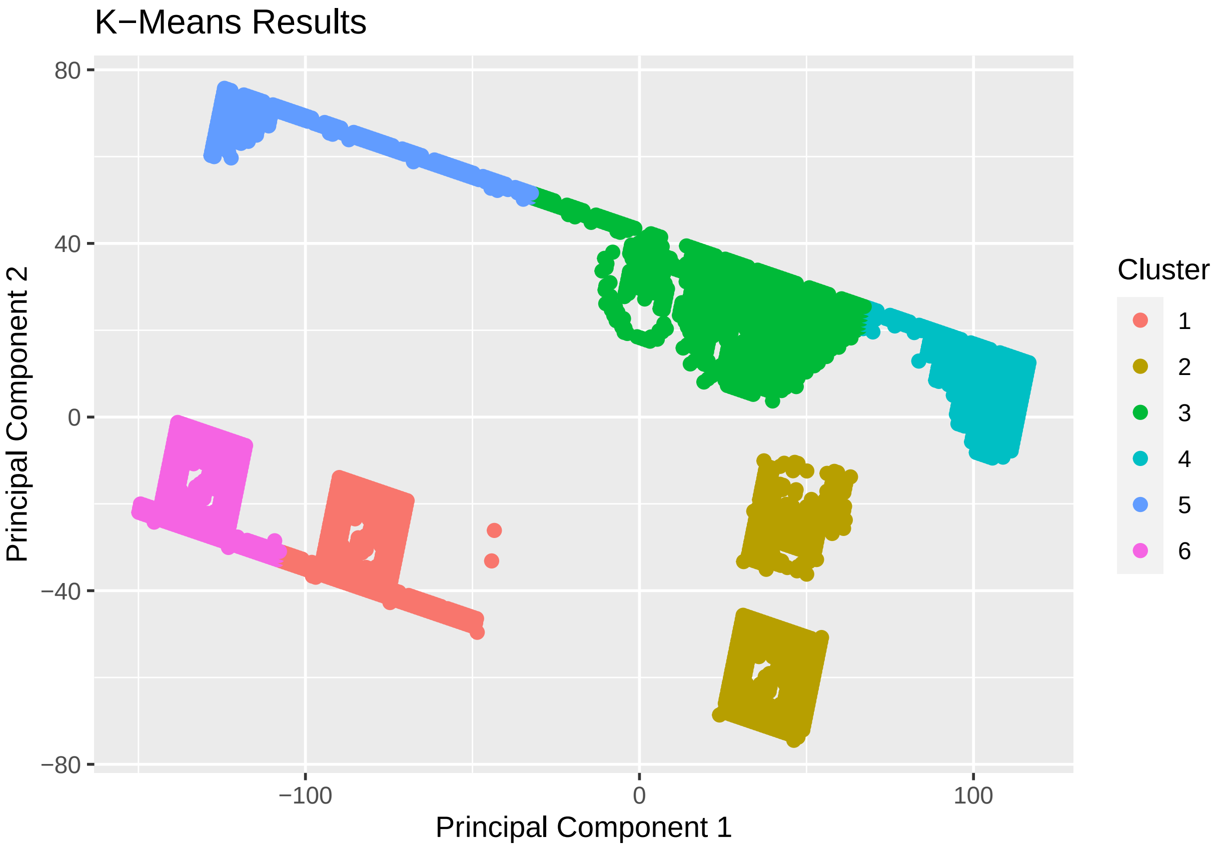
```
## Warning: package 'useful' was built under R version 4.0.5
```

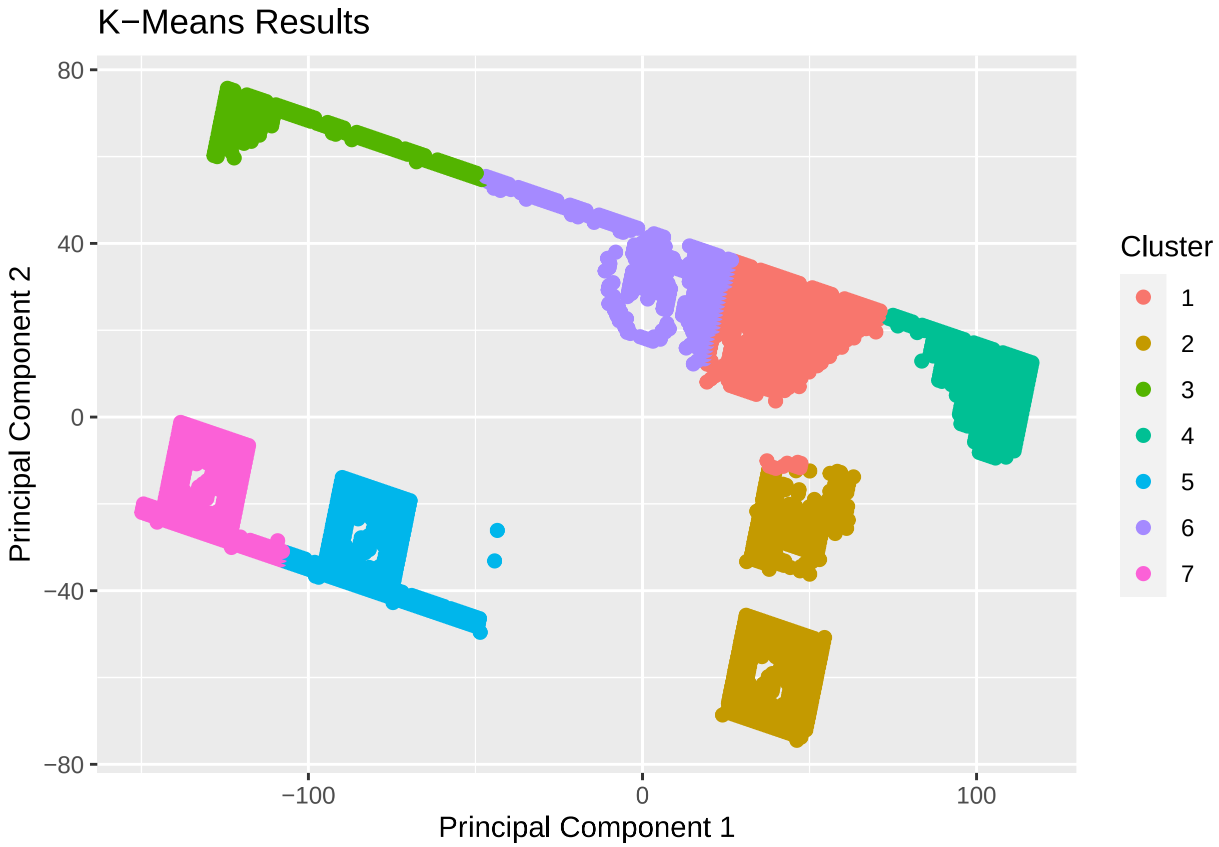

K−Means Results
Principal Component 2
80
40
0
−40
−80
−100
0
100
Principal Component 1
Cluster
1
2

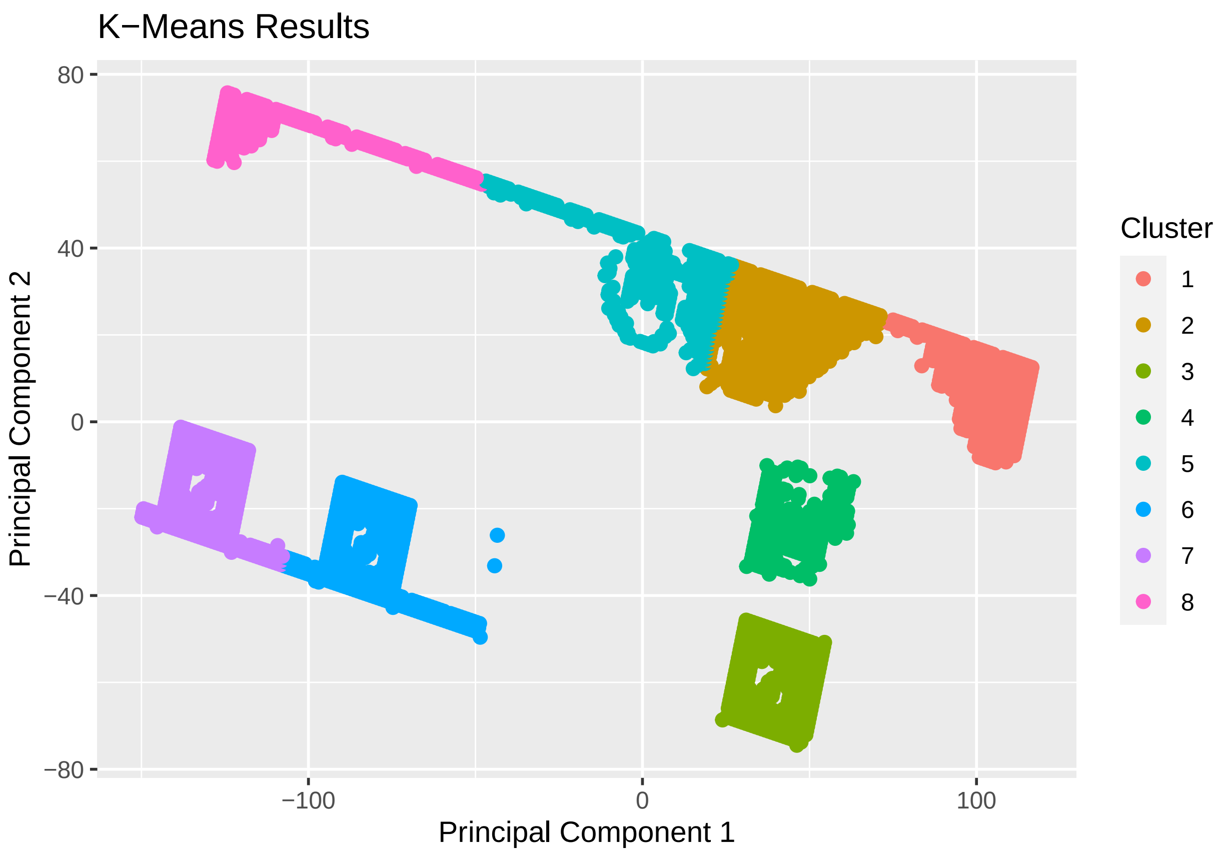

K−Means Results
80
40
0
−40
−80
Principal Component 2
−100
0
100
Principal Component 1
Cluster
1
2
3

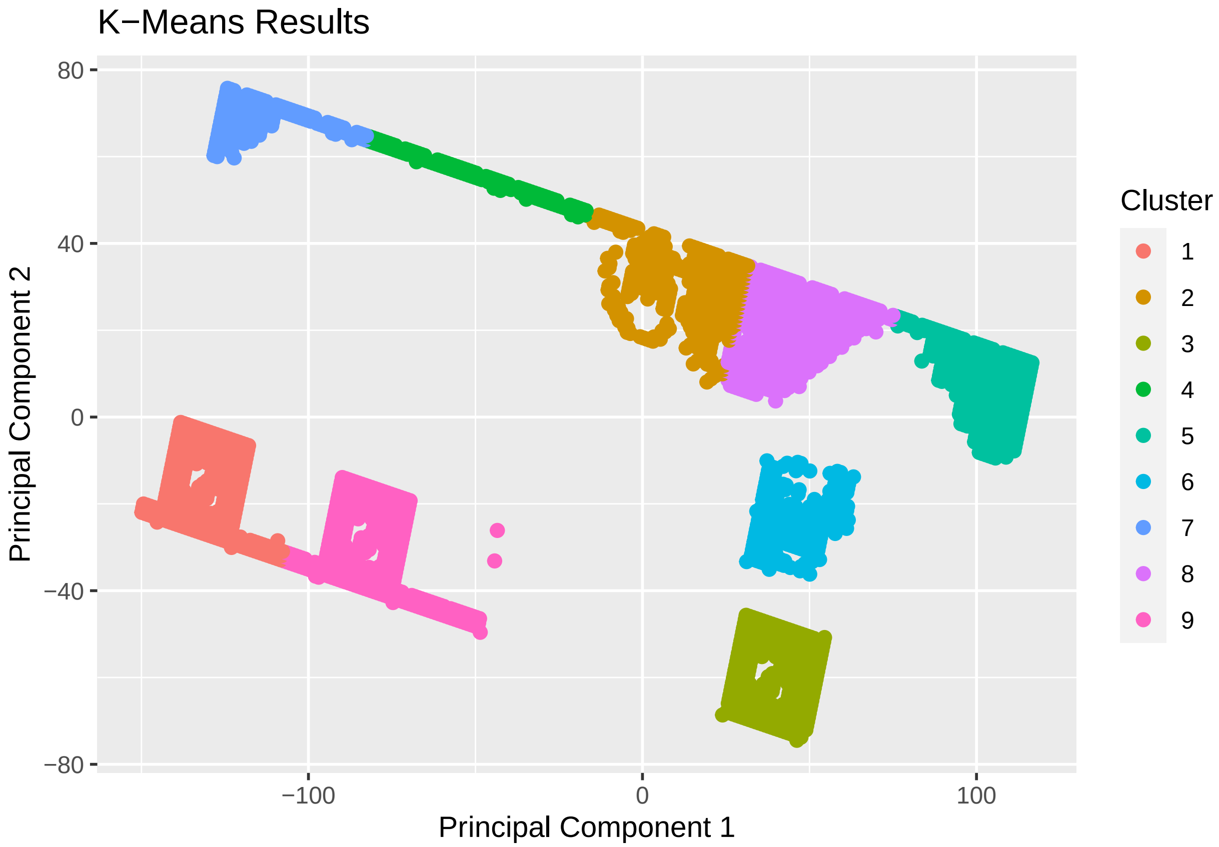

K−Means Results
Principal Component 2
80
40
0
−40
−80
−100
0
100
Principal Component 1
Cluster
1
2
3
4

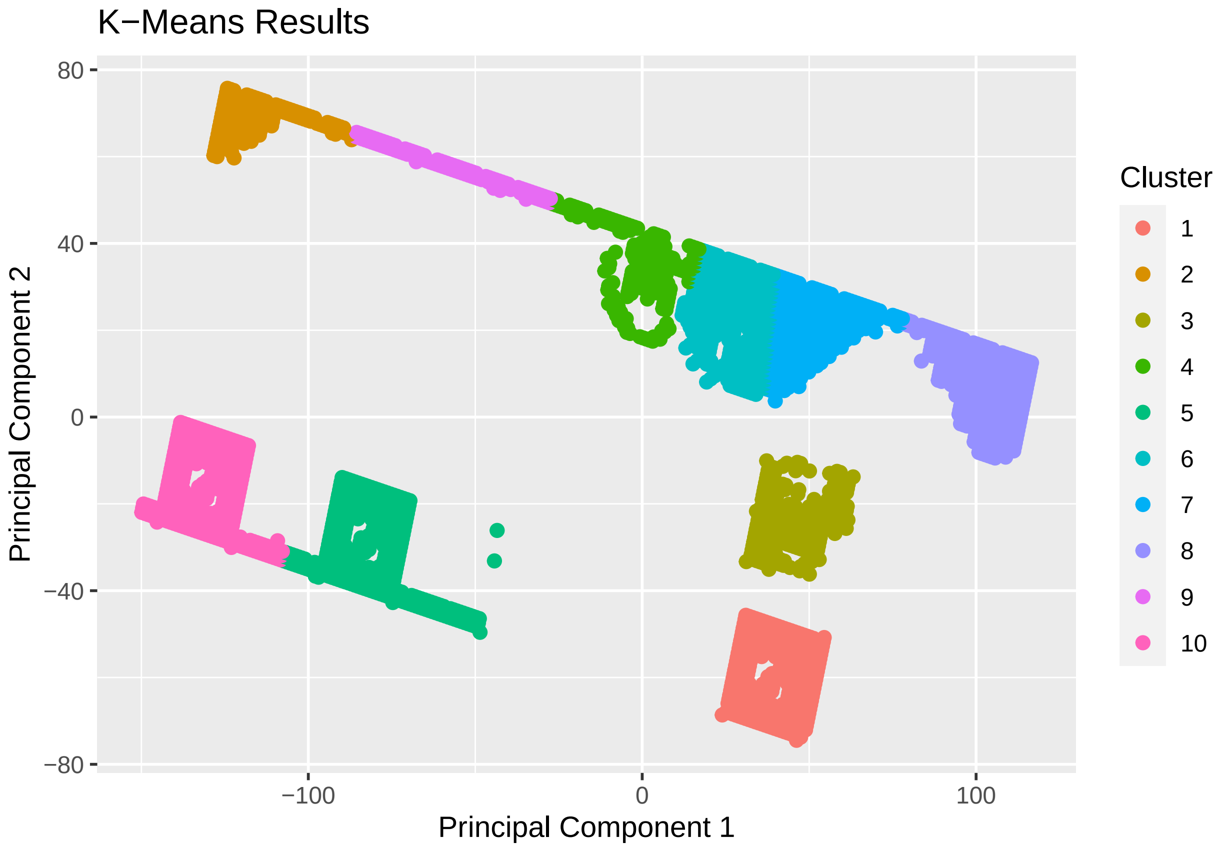

K−Means Results
80
40
0
−40
−80
Principal Component 2
−100
0
100
Principal Component 1
Cluster
1
2
3
4
5

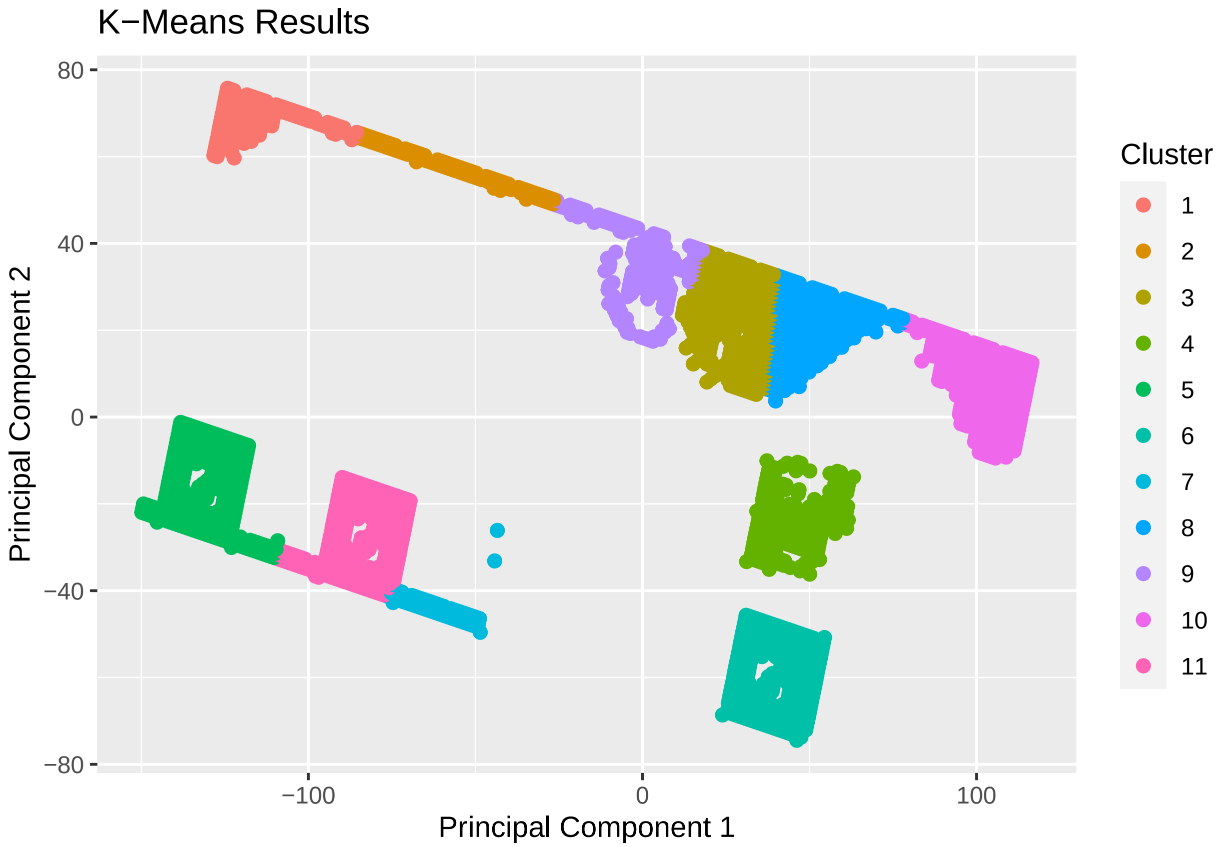

K−Means Results
Principal Component 2
80
40
0
−40
−80
Principal Component 1
−100
0
100
Cluster
1
2
3
4
5
6

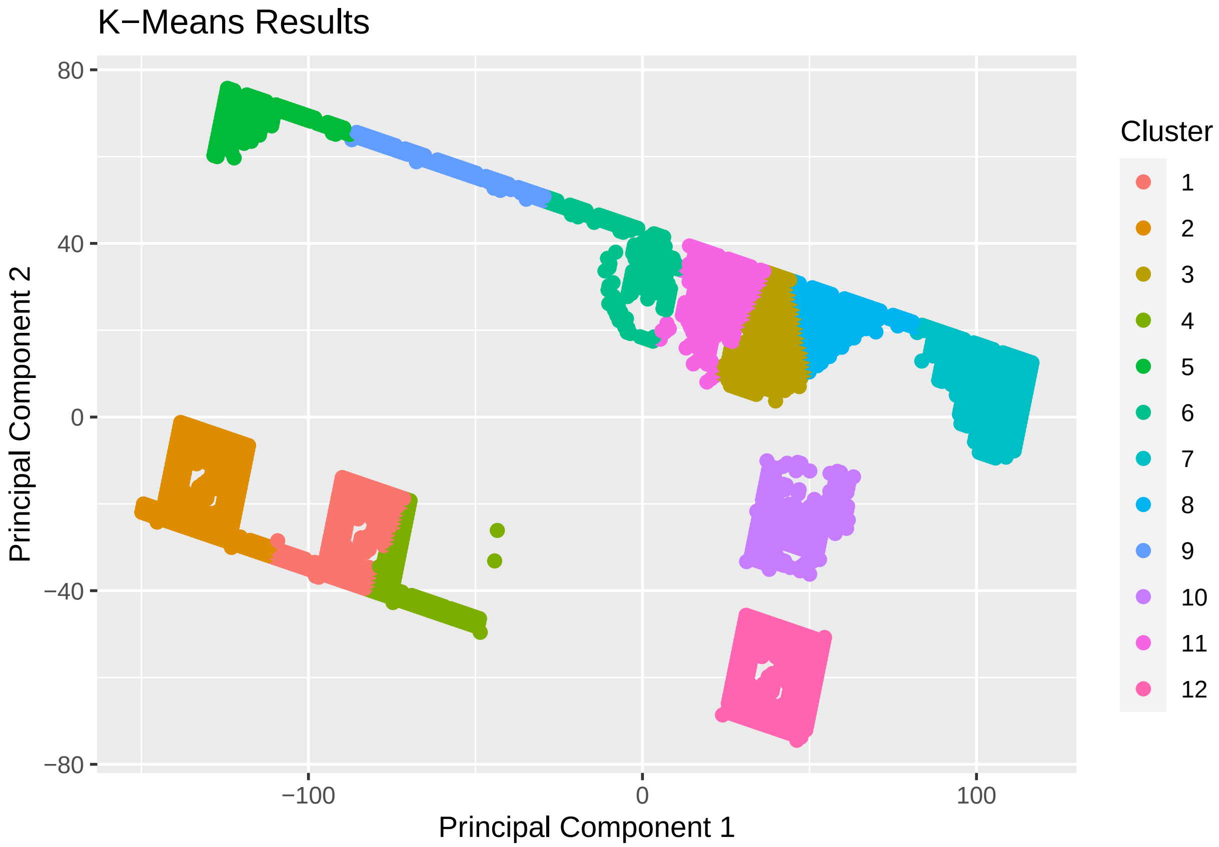

K−Means Results
Principal Component 2
80
40
0
−40
−80
−100
0
100
Principal Component 1
Cluster
1
2
3
4
5
6
7


K−Means Results
Principal Component 2
80
40
0
−40
−80
Principal Component 1
−100
0
100
Cluster
1
2
3
4
5
6
7
8


K−Means Results
Principal Component 2
80
40
0
−40
−80
−100
0
100
Principal Component 1
Cluster
1
2
3
4
5
6
7
8
9


K−Means Results
80
40
0
−40
−80
Principal Component 2
−100
0
100
Principal Component 1
Cluster
1
2
3
4
5
6
7
8
9
10

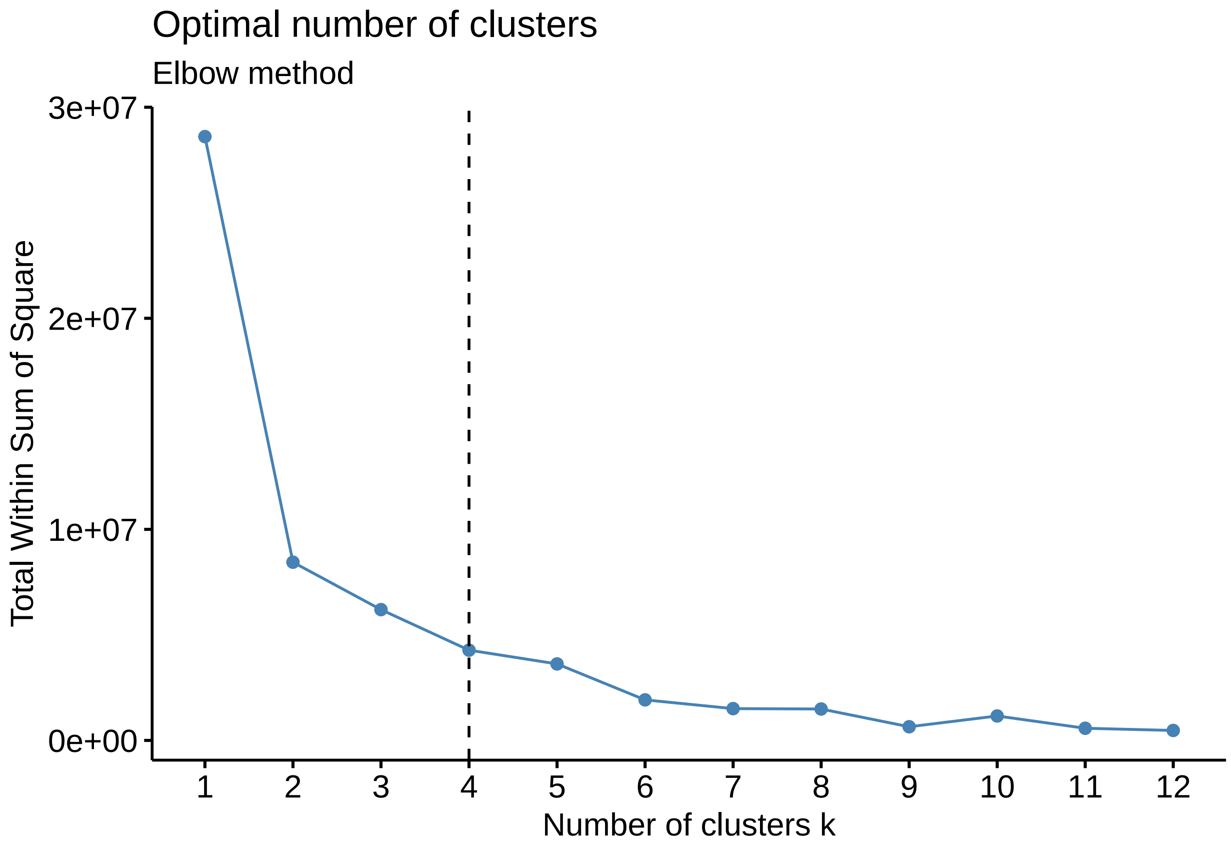

K−Means Results
80
40
0
−40
−80
Principal Component 2
−100
0
100
Principal Component 1
Cluster
1
2
3
4
5
6
7
8
9
10
11



7. As k-means is an unsupervised algorithm, you cannot compute the accuracy as there are no correct values to compare the output to. Instead, you will use the average distance from the center of each cluster as a measure of how well the model fits the data. Calculate this average distance from the center of each cluster for each value of k and plot it as a line chart where k is the x-axis and the average distance is the y-axis.

```
## Warning: package 'factoextra' was built under R version 4.0.5
```

```
## Welcome! Want to learn more? See two factoextra-related books at https://goo.gl/ve3WBa
```



8. One way of determining the "right" number of clusters is to look at the graph of k versus average distance and finding the "elbow point". Looking at the graph you generated in the previous example, what is the elbow point for this dataset?

The elbow point for this dataset is 4.