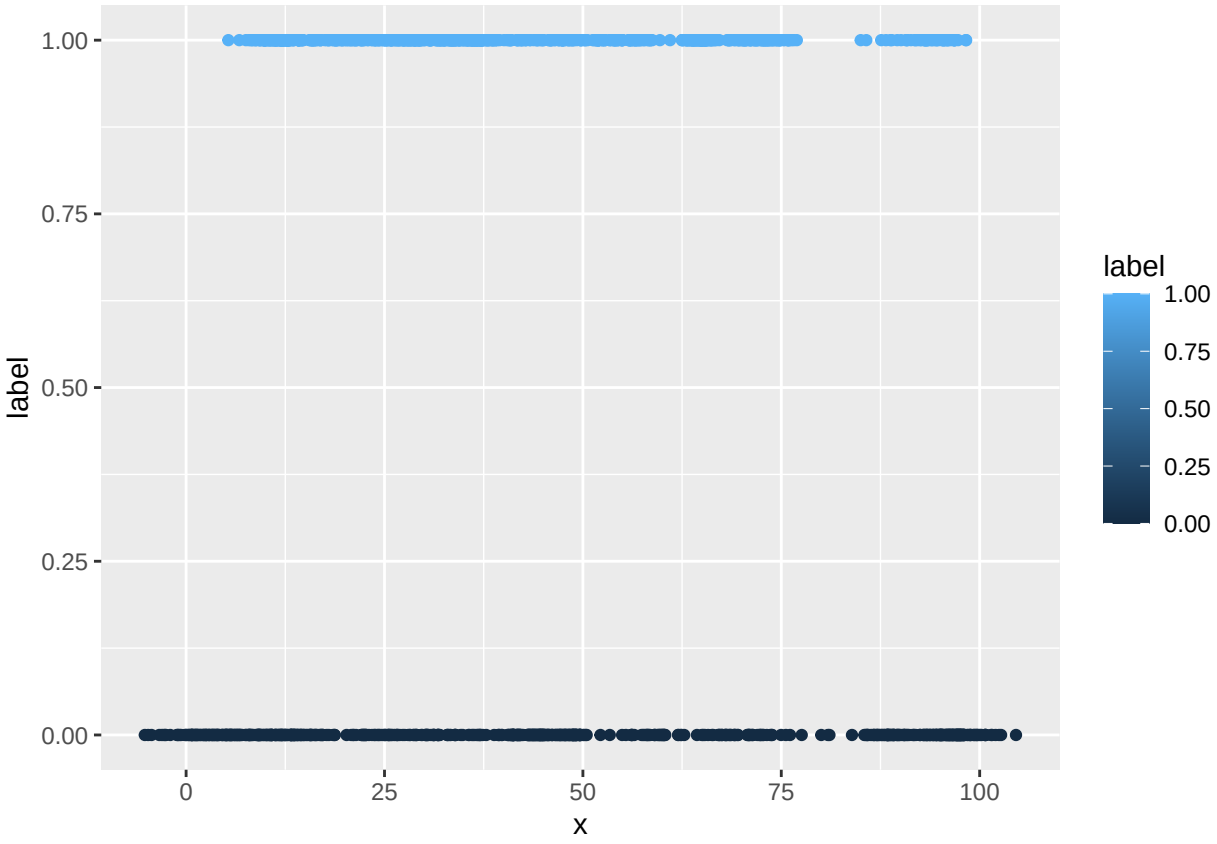
Week 11 Assignment

Mithil Patel

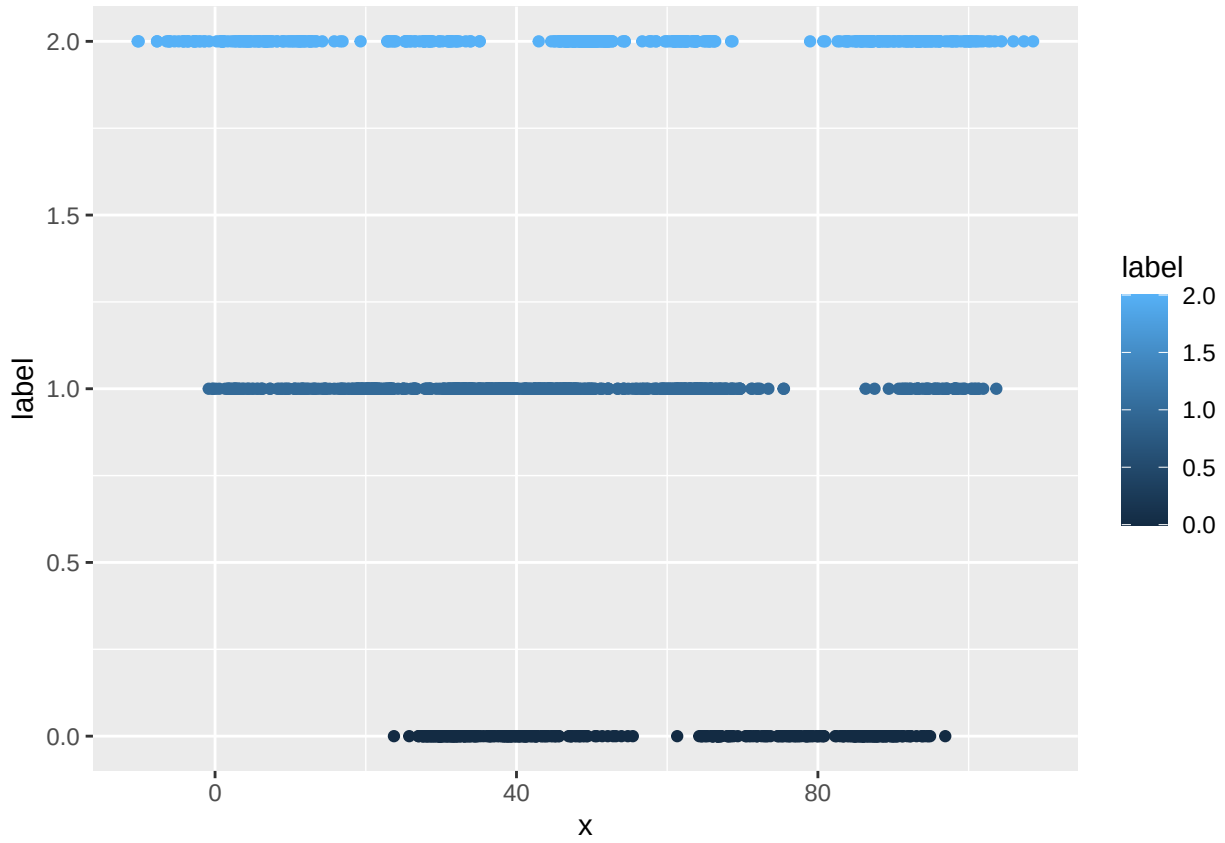
2022-06-04

K-nearest neighbor (KNN)

Binary Dataset Scatter plot



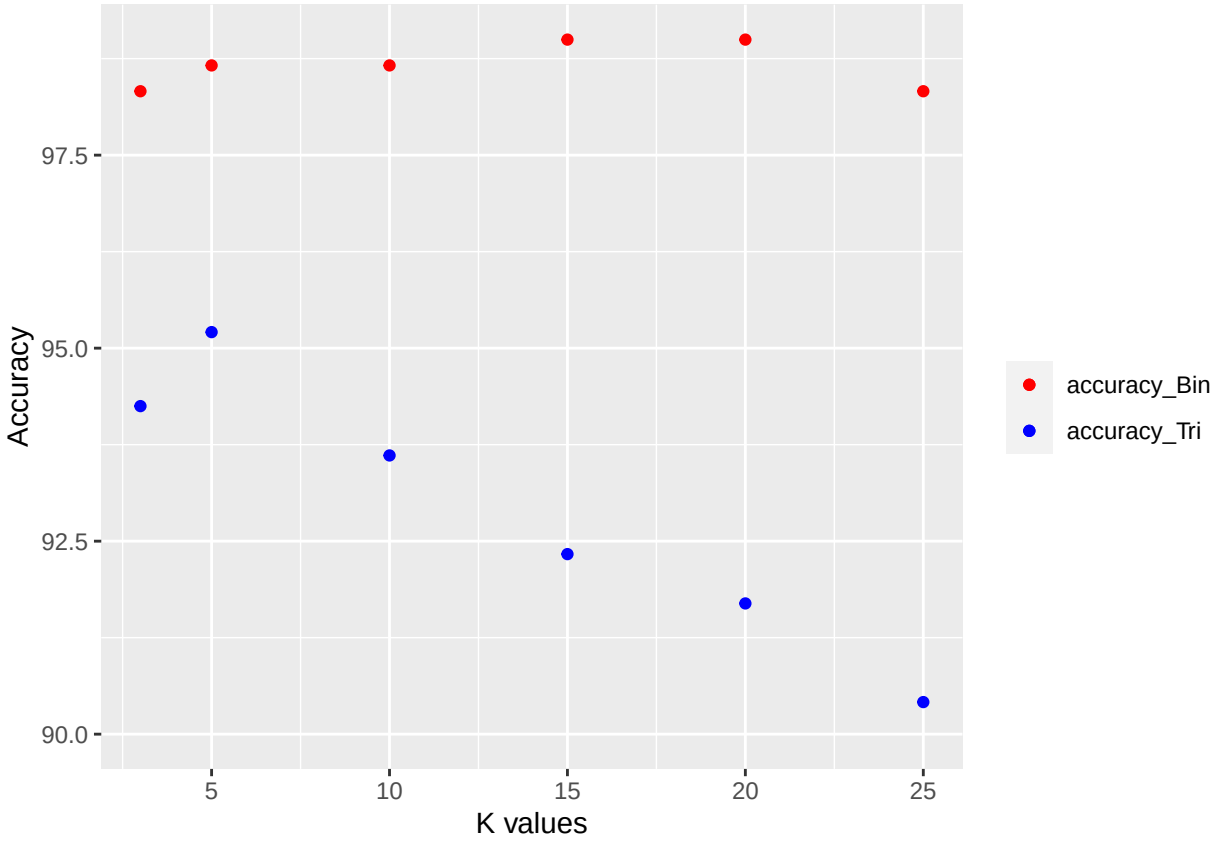
Trinary Dataset Scatter plot



KNN model accuracy with different k values

	k_number	accuracy_Bin	accuracy_Tri
1	3	98.32776	94.24920
2	5	98.66221	95.20767
3	10	98.66221	93.61022
4	15	98.99666	92.33227
5	20	98.99666	91.69329
6	25	98.32776	90.41534

Accuracy plot

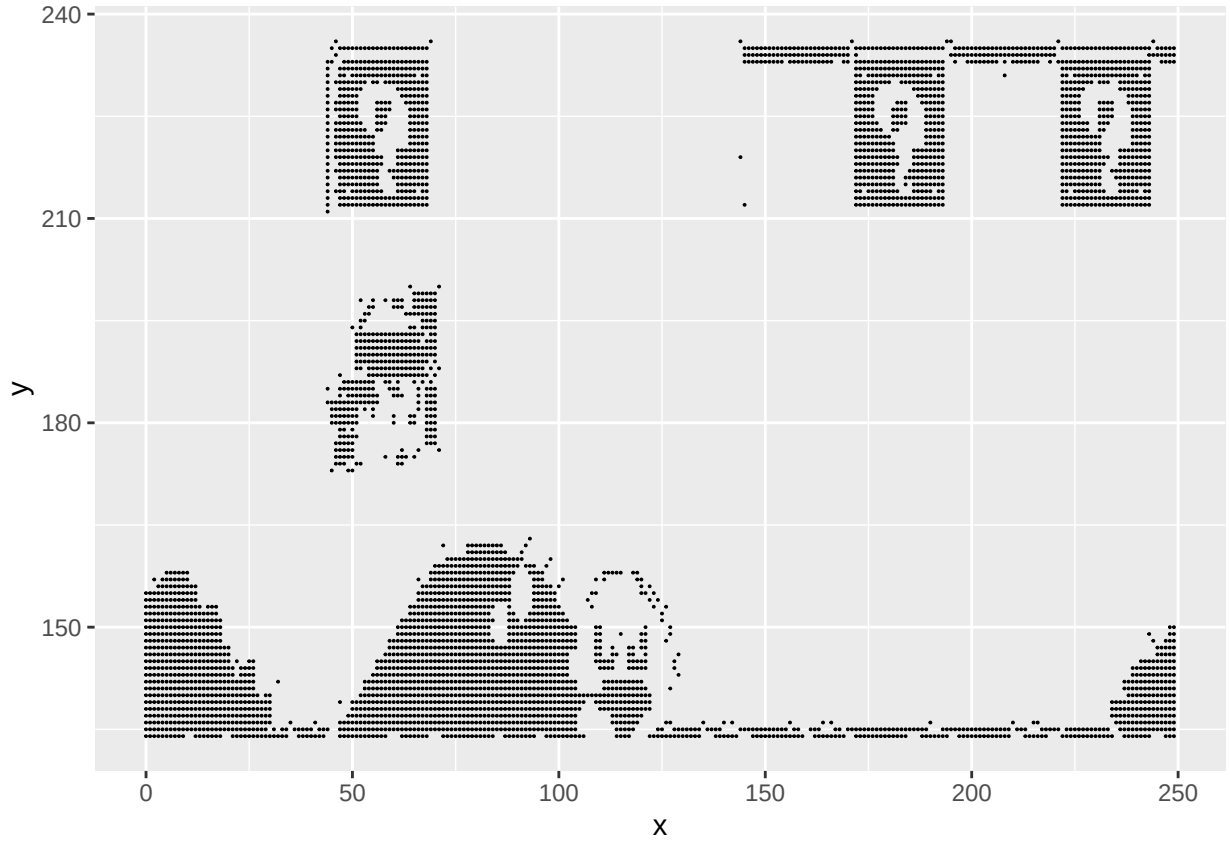


1.e.v) Based on the accuracy of the models, a linear classifier such as K-nearest neighbor (KNN) is the most effective model when dealing with datasets containing ordinal outputs (i.e. 0's and 1's). With accuracy values above 90% for the KNN model with different k values, the model is a good representation of the given datasets and it can be used to predict the output with high confidence. One key difference between both datasets is that the accuracy on the trinary classifier dataset appears to decrease as the k value increases, thus suggesting that the k=5 is the ideal parameter for our dataset and increasing the k value will further decrease the accuracy of the model.

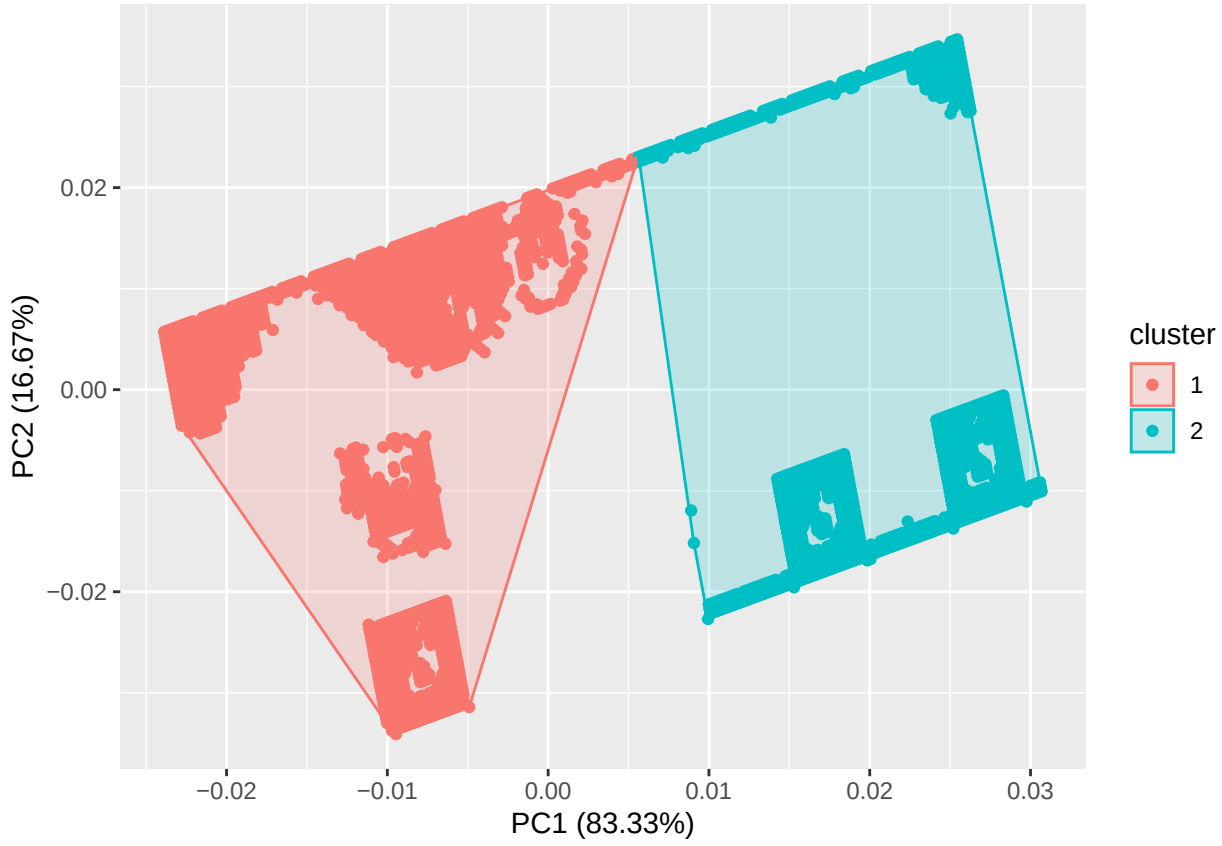
1.e.vi) Based on the accuracy of both models, K-nearest neighbor (KNN) performed significantly better than logistic regression at predicting categorical data with 98% accuracy compared to the logistic model's 58% accuracy. Since the dependent value ('label') of our binary dataset contains categorical data, logistic regression, which is often used to find the probability of an event occurring using a sigmoid function, will either bump up or drop the expected probability to a discrete value, thereby compromising the accuracy of the logistic model. Additionally, the KNN model generally gives you highly accurate predictions with no real-life relatability compared to the logistic regression.

K-Mean Clustering

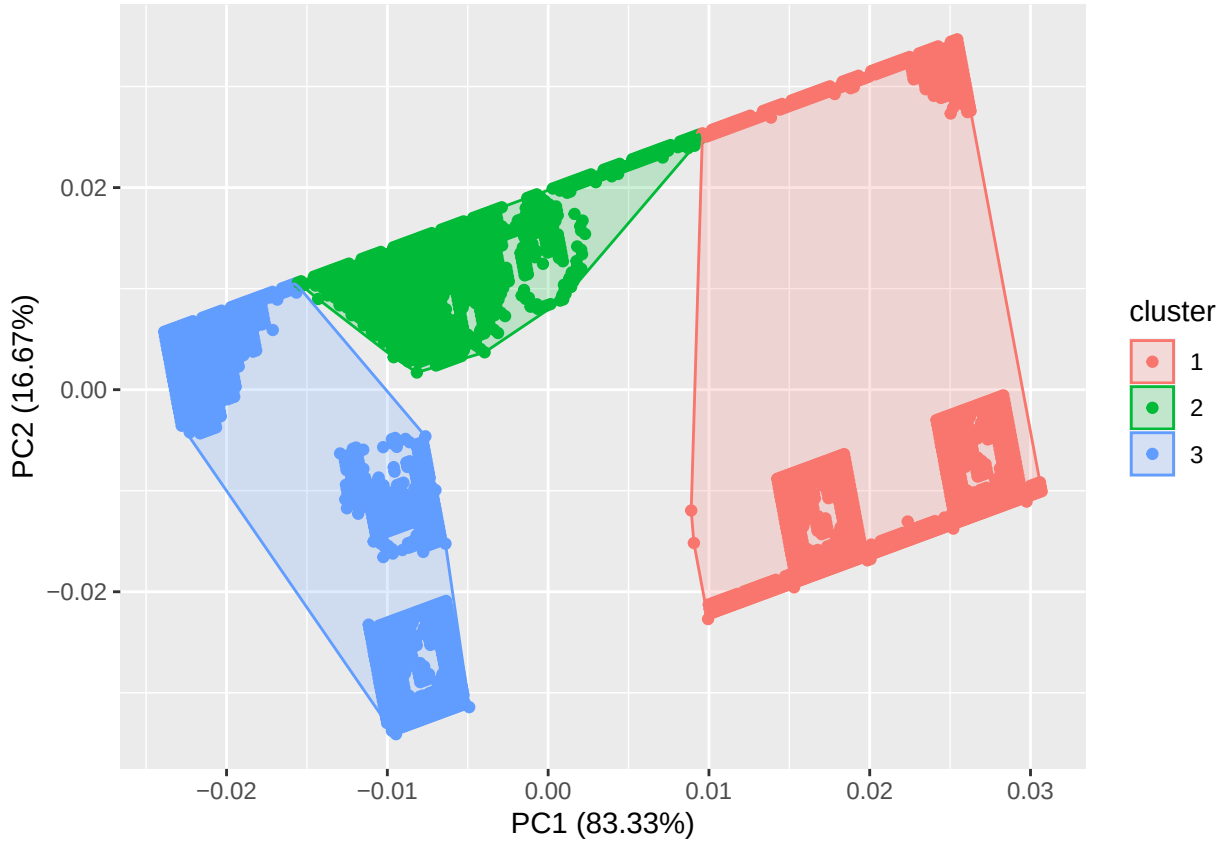
Cluster Dataset Scatter Plot



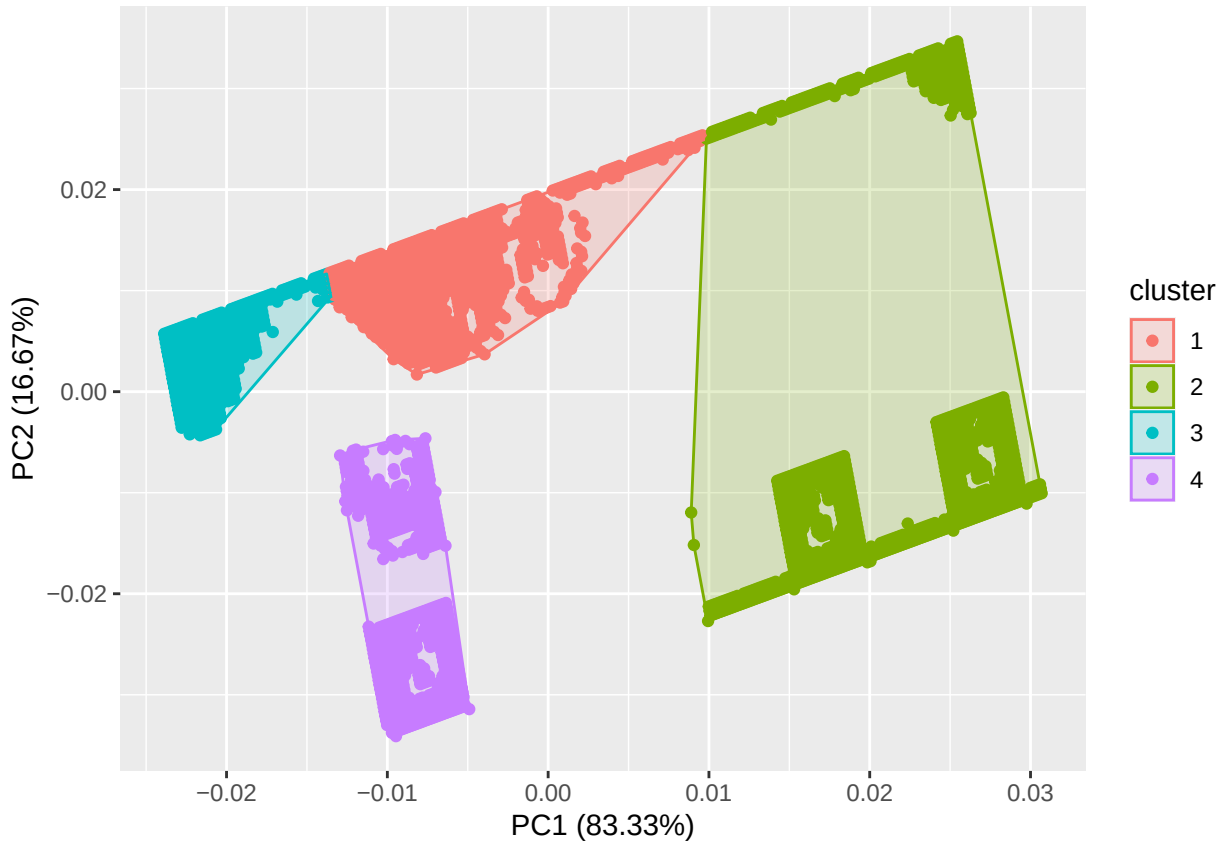
Visualizing Clusters



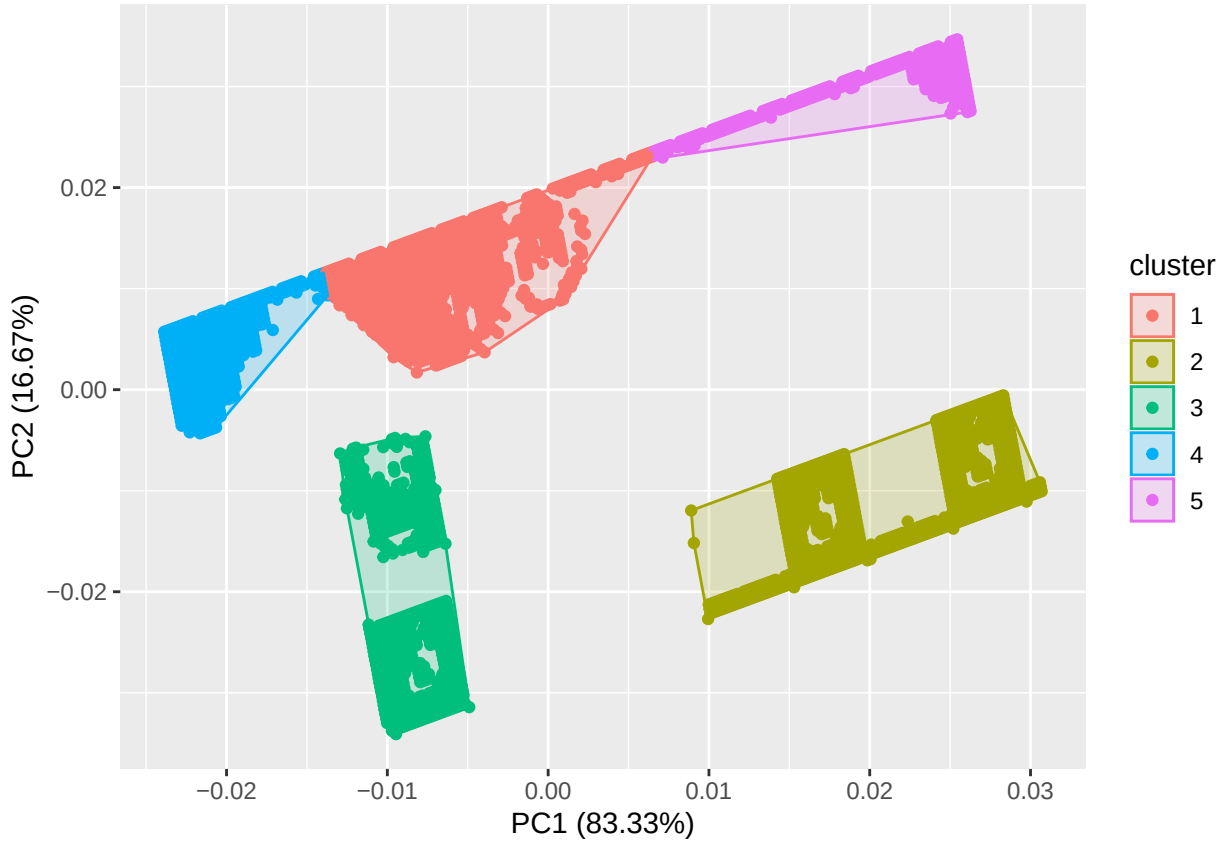

```
[1] "Cluster plot with k =  2"
```



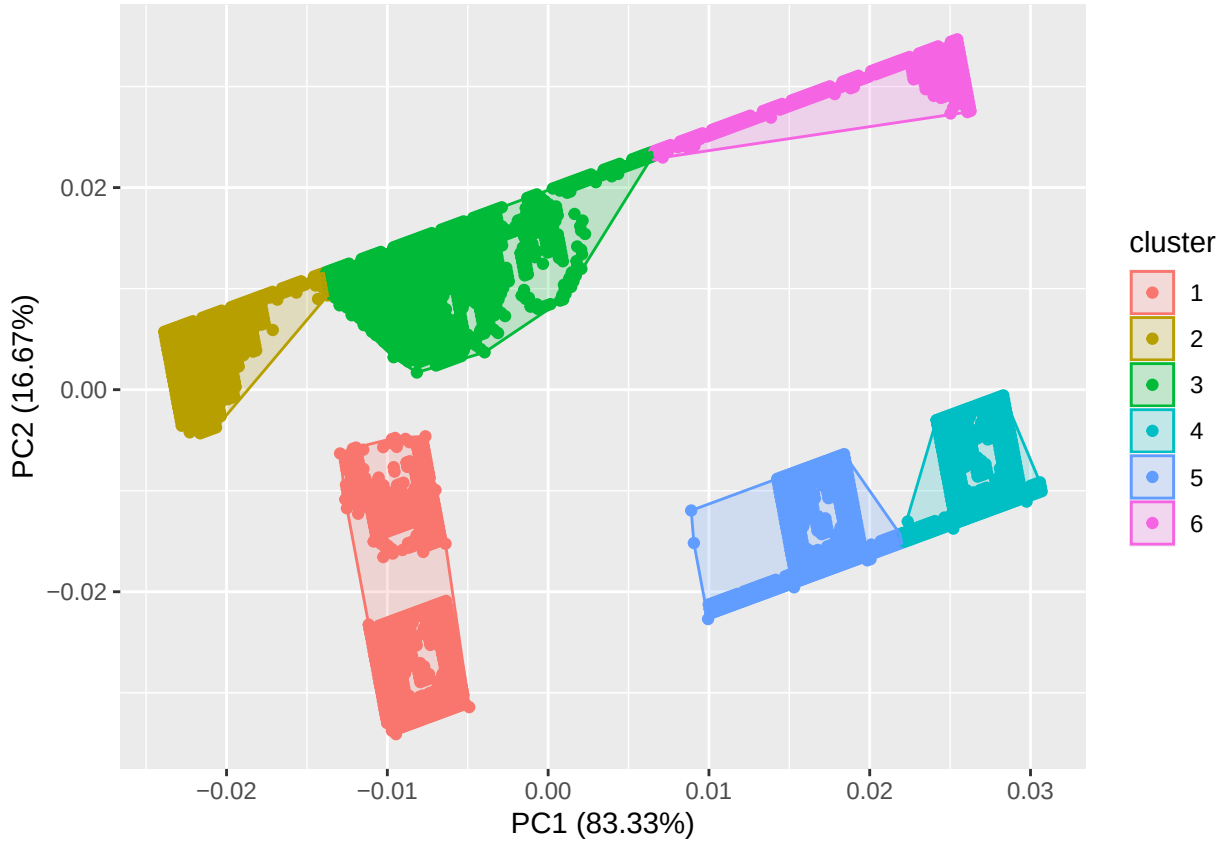

```
[1] "Cluster plot with k =  3"
```



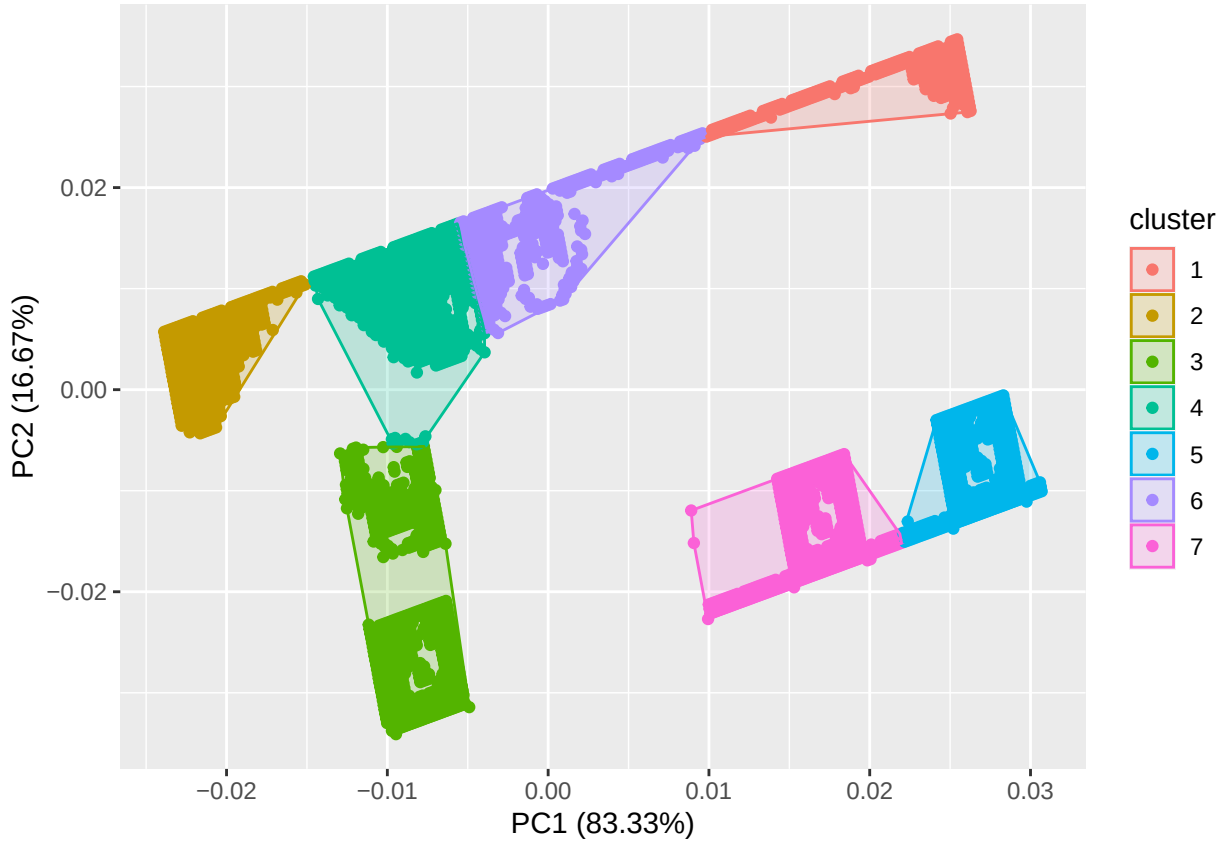

```
[1] "Cluster plot with k =  4"
```



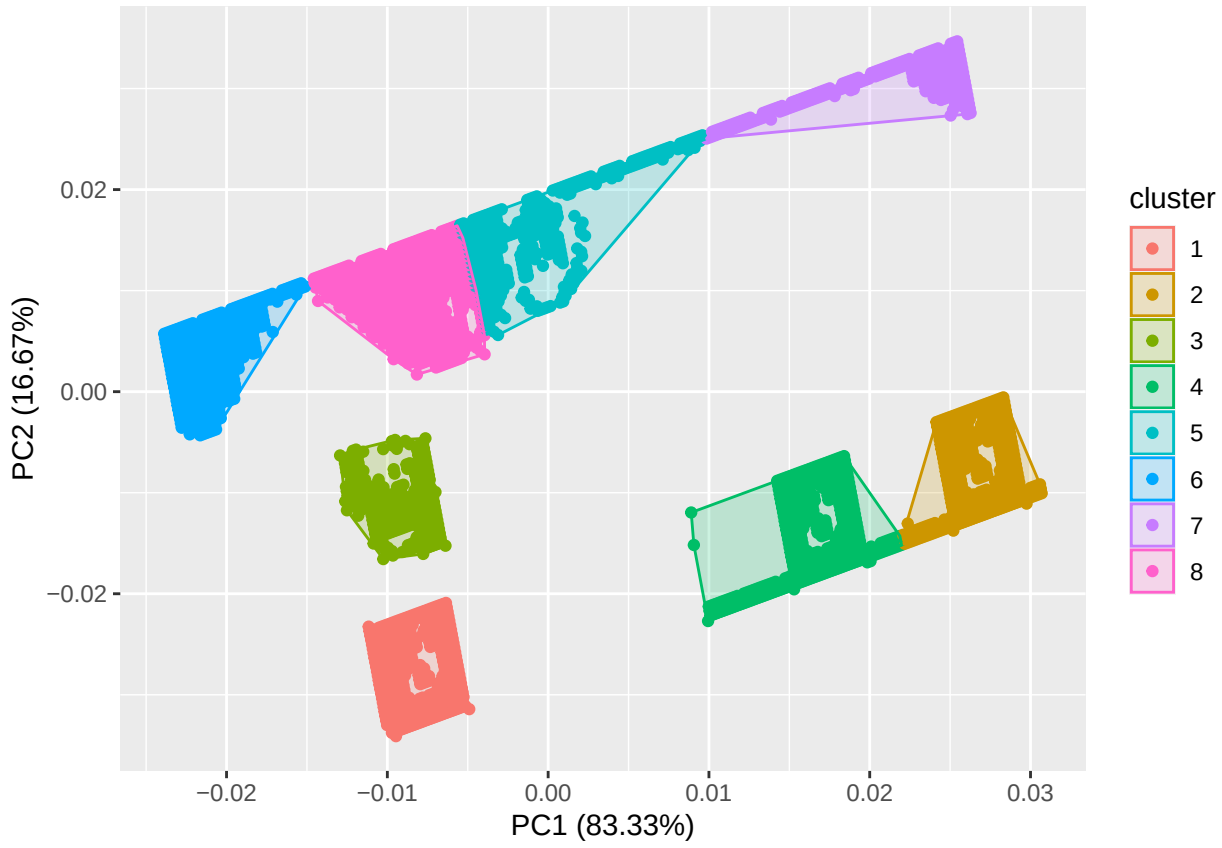

```
[1] "Cluster plot with k =  5"
```



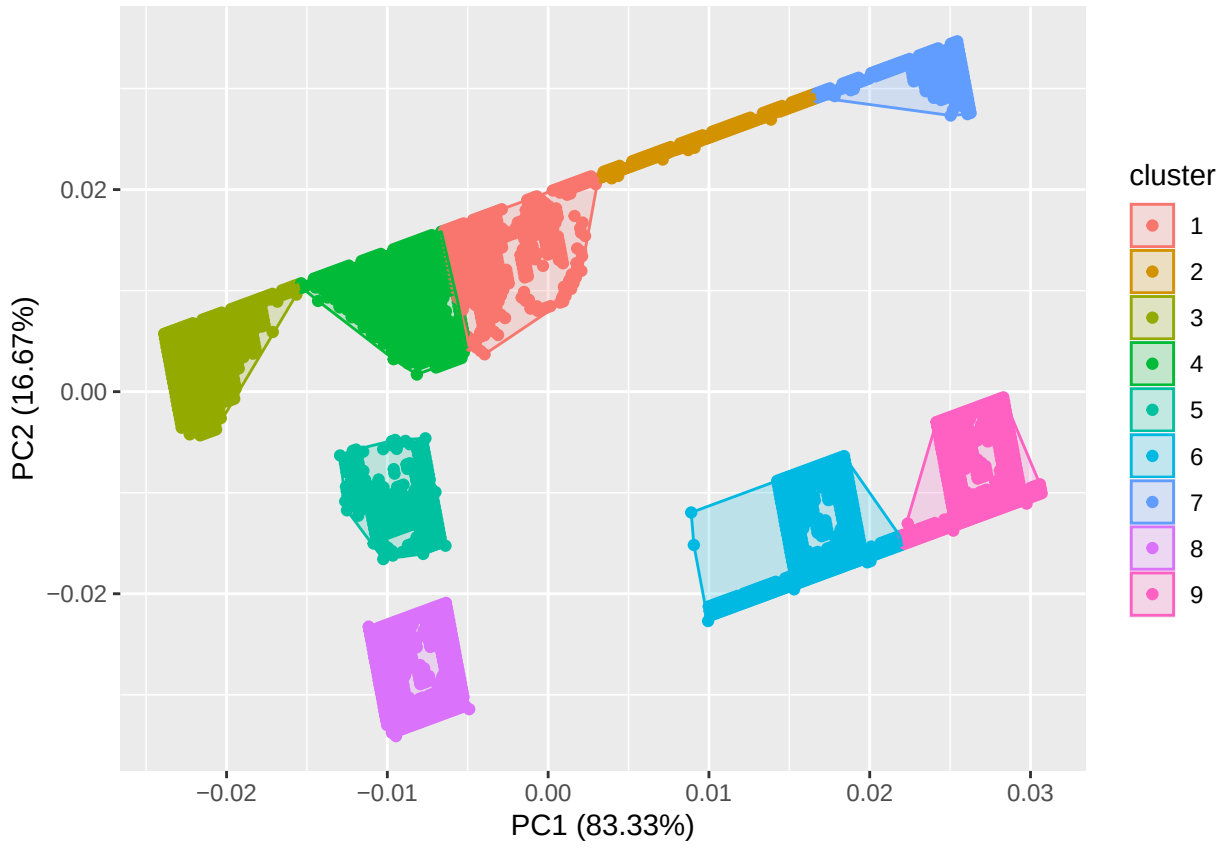

```
[1] "Cluster plot with k =  6"
```



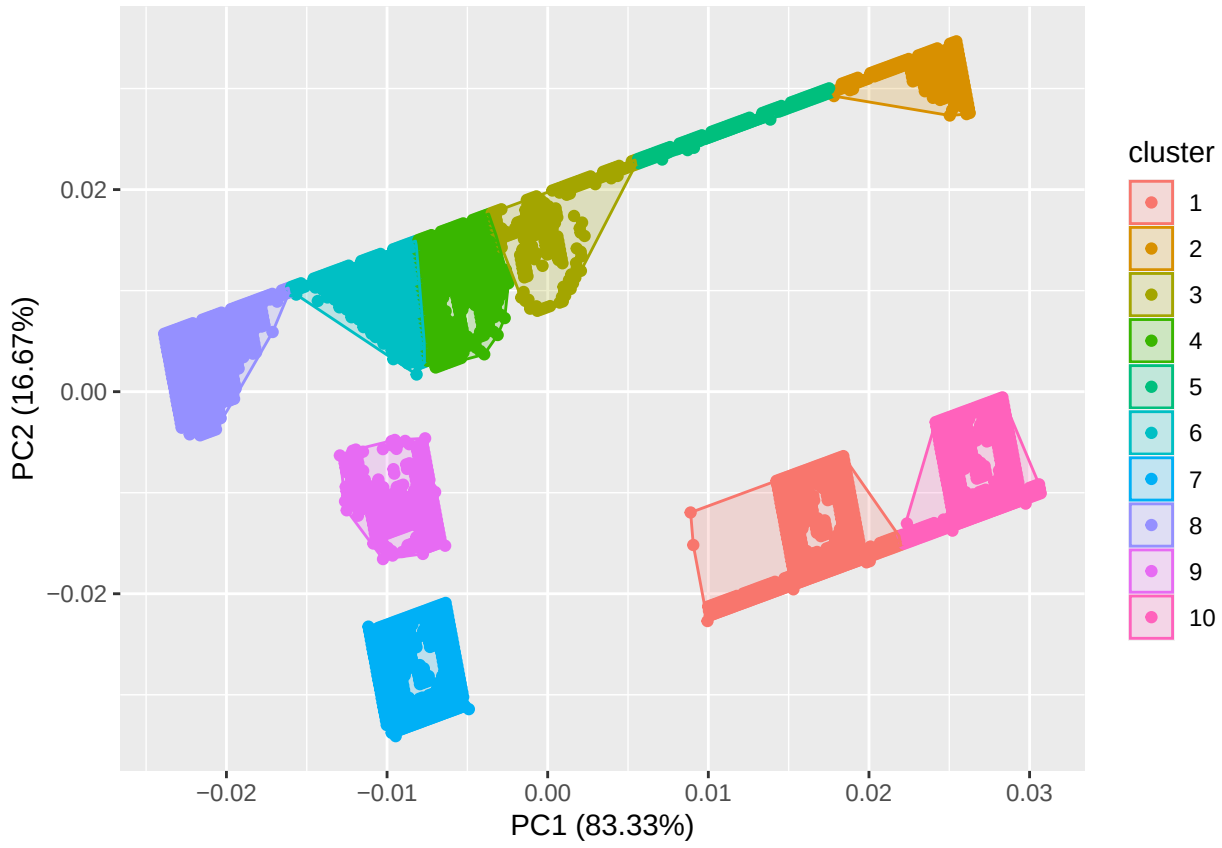

```
[1] "Cluster plot with k =  7"
```



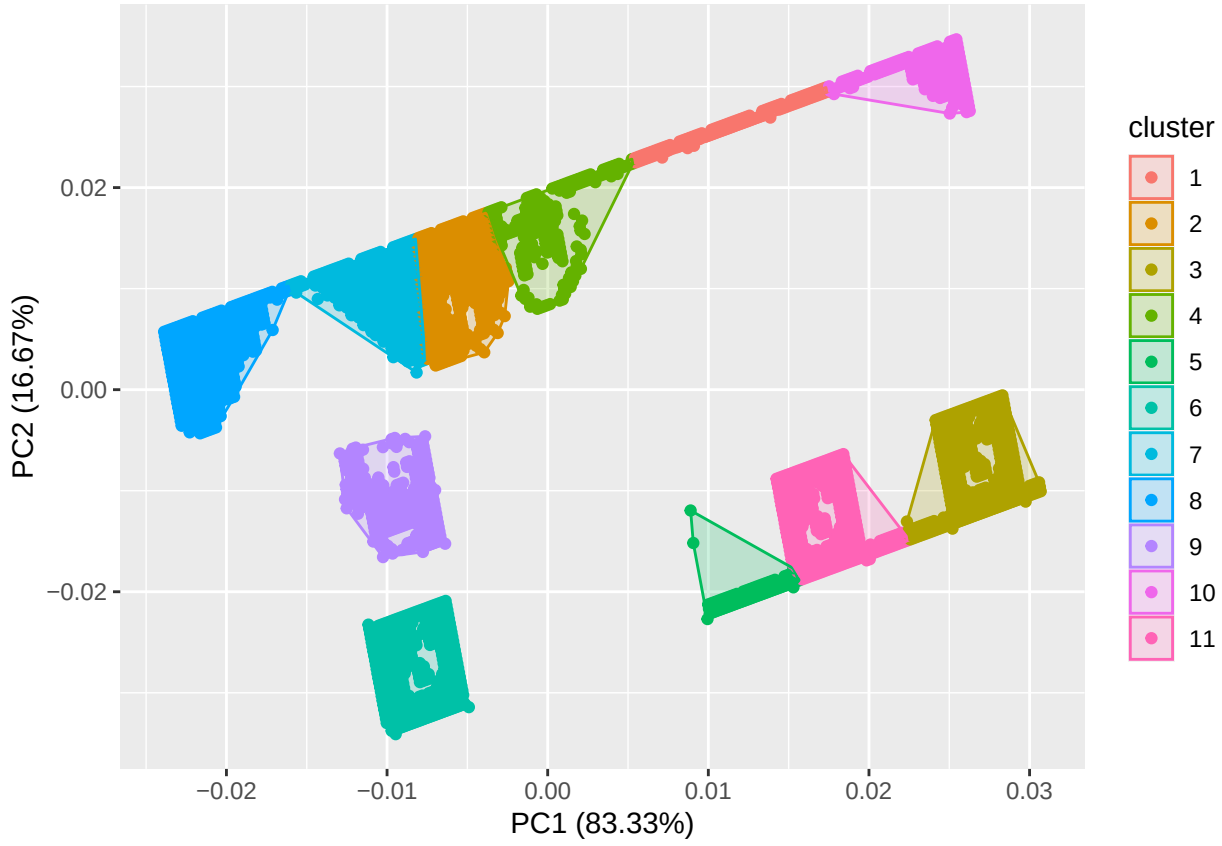

```
[1] "Cluster plot with k =  8"
```



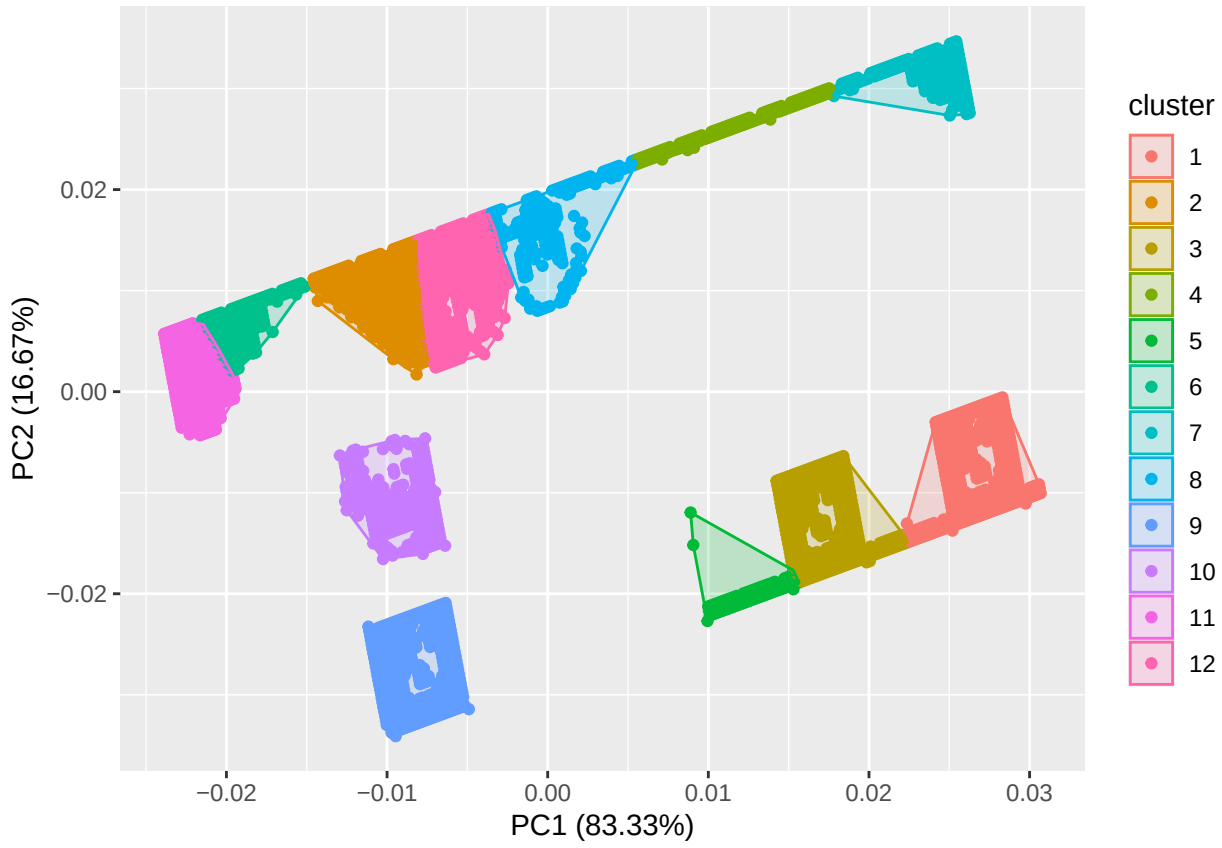

```
[1] "Cluster plot with k =  9"
```




```
[1] "Cluster plot with k =  10"
```

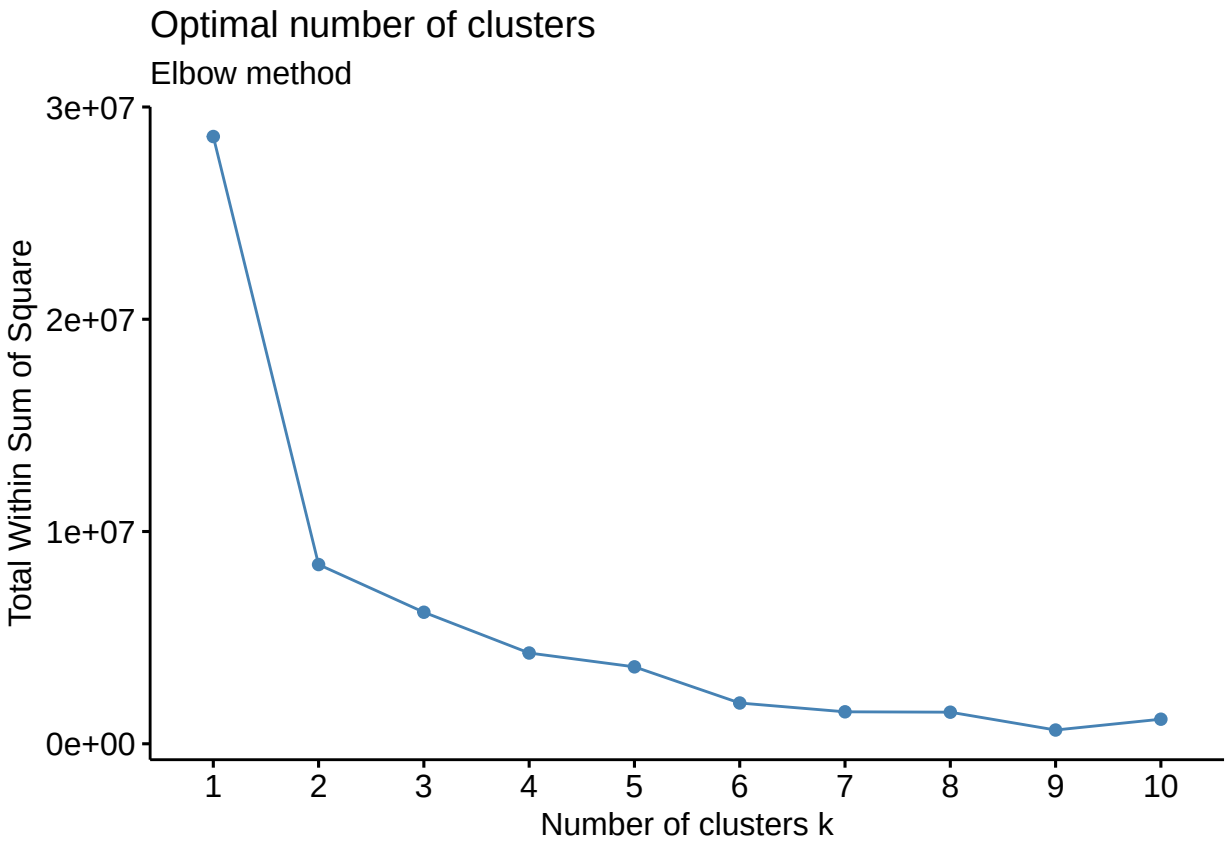



```
[1] "Cluster plot with k =  11"
```




```
[1] "Cluster plot with k =  12"
```

Elbow Method



Based on the Elbow Method chart above, we can conclude the optimal number, or the 'Elbow Point', of the dataset is k = 5.