11.2 Exercise

Amelia Farrell

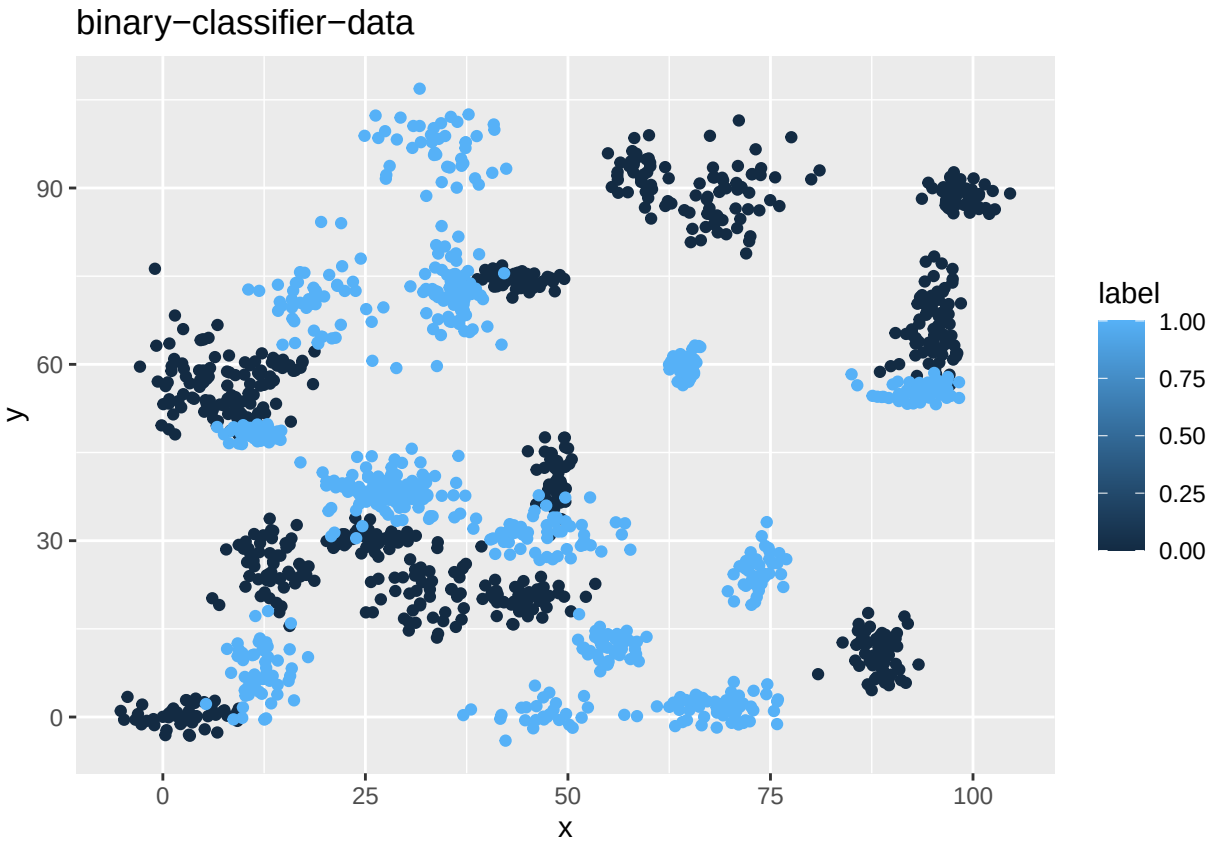
November 8st 2021

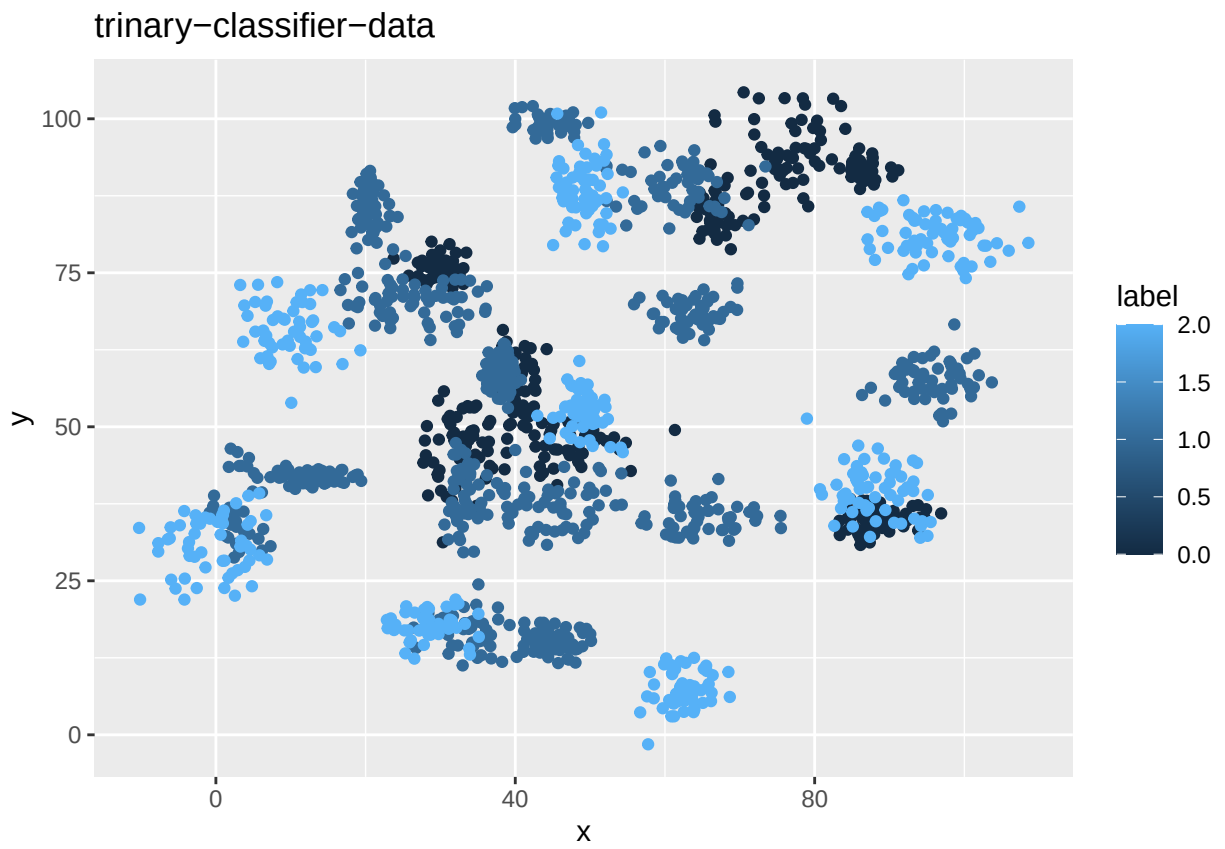
Introduction to Machine Learning

K nearest neighbors (Binary and Trinary classifier data)

i. In this section we will be fitting a k nearest neighbors algorithm to two different data sets.

Before build our modelds, lets plot the data from each dataset using a scatter plot.





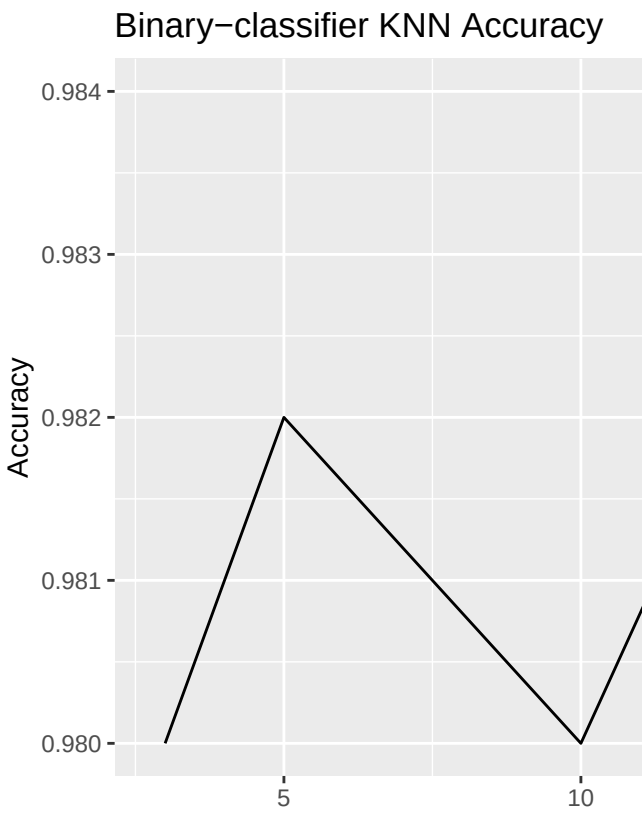
ii. Fitting a k nearest neighbors' model Next lets fit a k nearest neighbors' model to each dataset for a range of k values (k=3, k=5, k=10, k=15, k=20, and k=25).As well as print the accuracy in order to compare the performance of each.

```
## Warning in `!=.default`(t3knn, test_bc$label): longer object length is not a
## multiple of shorter object length

## Warning in is.na(e1) | is.na(e2): longer object length is not a multiple of
## shorter object length

## Warning in `!=.default`(t5knn, test_bc$label): longer object length is not a
## multiple of shorter object length

## Warning in is.na(e1) | is.na(e2): longer object length is not a multiple of
## shorter object length

## Warning in `!=.default`(t10knn, test_bc$label): longer object length is not a
## multiple of shorter object length

## Warning in is.na(e1) | is.na(e2): longer object length is not a multiple of
## shorter object length

## Warning in `!=.default`(t15knn, test_bc$label): longer object length is not a
## multiple of shorter object length
```

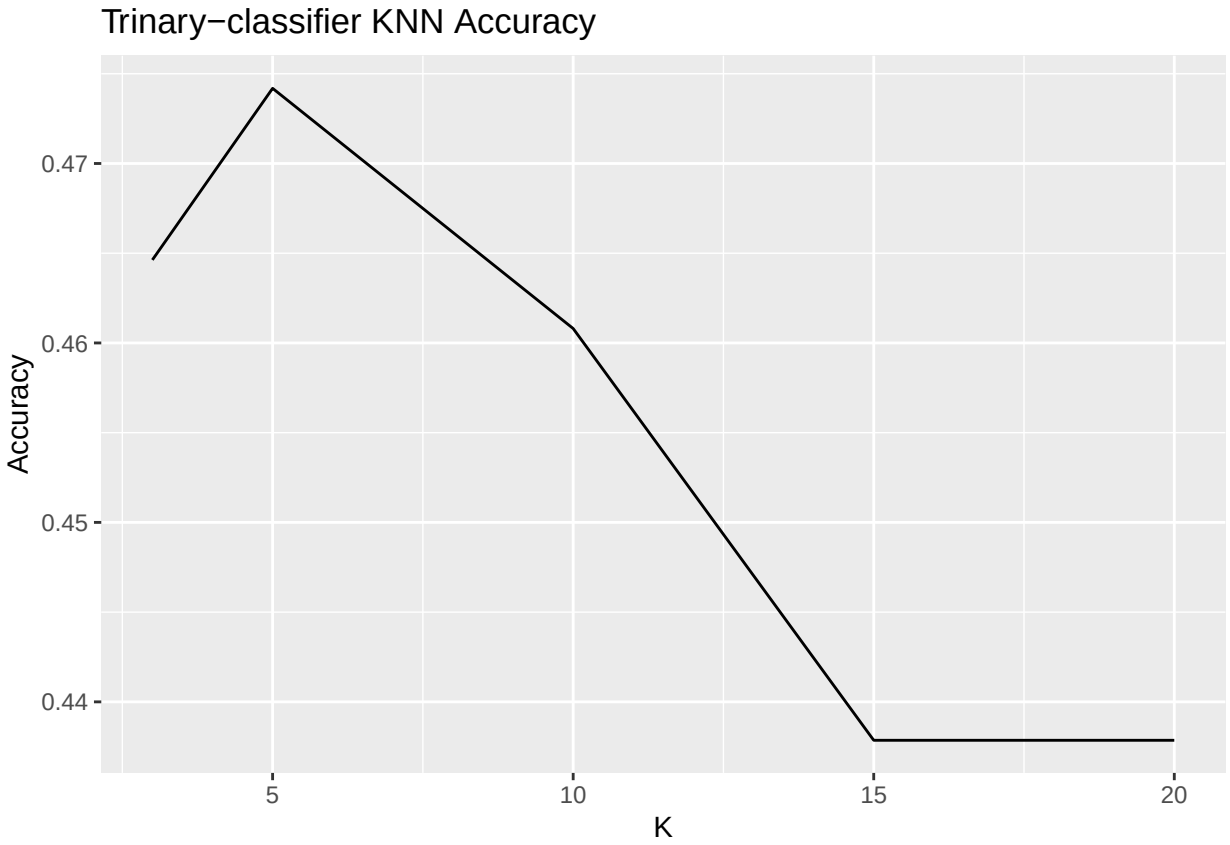
```
## Warning in is.na(e1) | is.na(e2): longer object length is not a multiple of
## shorter object length
```

```
## Warning in `!=.default`(t20knn, test_bc$label): longer object length is not a
## multiple of shorter object length
```

```
## Warning in is.na(e1) | is.na(e2): longer object length is not a multiple of
## shorter object length
```



Next we can plot our accuracy to compare the performance of each K.



iii. Does a linear classifier would work well on these datasets?

Looking at the above charts, we can conclude that the KNN model did a great job at predicting for the Binary-classifier data set. The line graph may look drastic, but it's accuracy stays between 98% and 98.4% accurate for our differing K values. The KNN model did not do so well for the Trinary-classifier data set. It's accuracy only reached 48% at K=5 and dropped with each increase in K there after. Why would this be? Looking back at our scatter plot we can clearly see what confused our KNN model. May of our Lables over lap one another. This would lead the KNN model to incorrectly classify observations since the groupings are so close to one another. This also further explains why the accuracy decreases as out number of Ks increases. Based off this, KNN may not be the best model for this data set.

Figure 1: Trinary-classifier data

iv. How does the accuracy of your logistic regression classifier from last week compare? The KNN model preformed much better for the Binary-classifier data set than the logistic regression model we ran

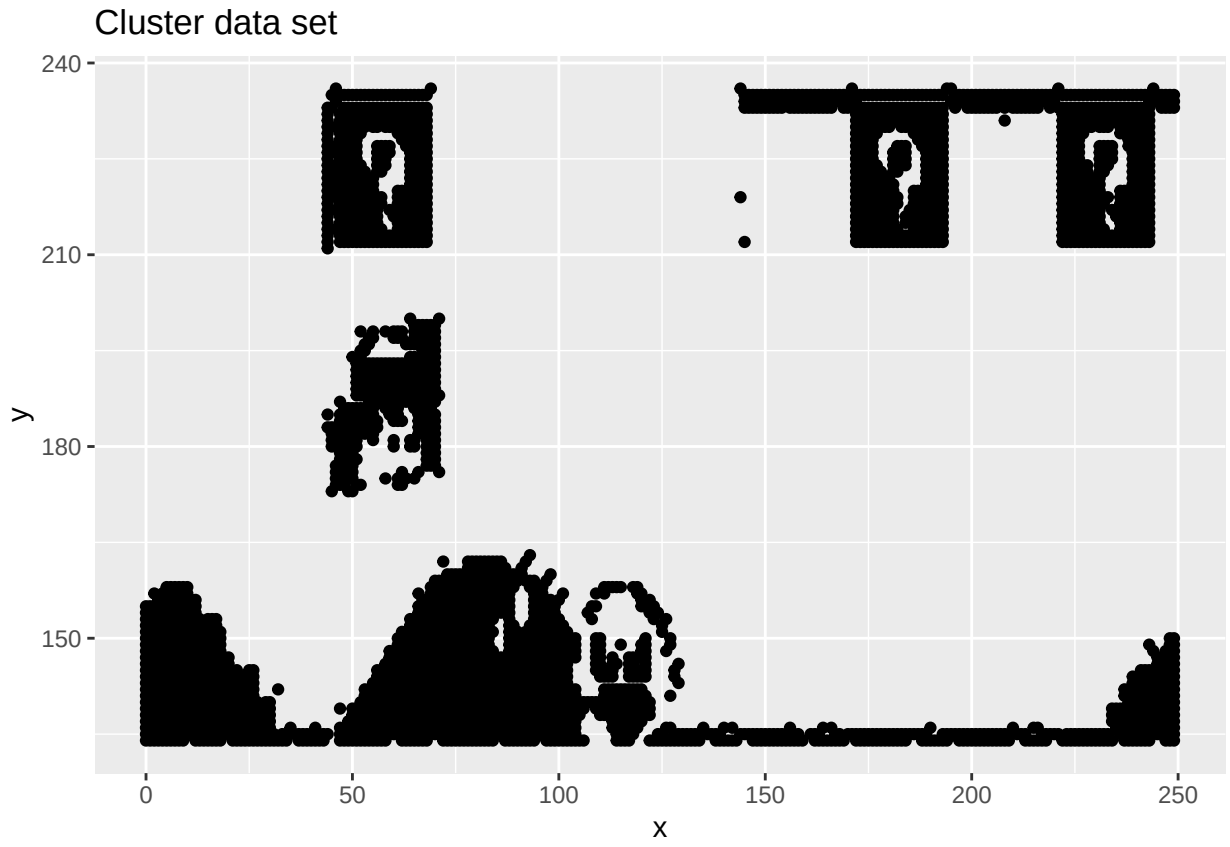
last week (58% accurate). So why would the logistic regression model preform far worse than the KNN model? Well look at our data set in the first plot. The data is certainly not linear. Logistic regression should not be used in problems where the data is not linear and their is alot of "noise". KNN developed to handle non-parametric data and will almost always preform better on this type of data when compared to regression.

Clustering

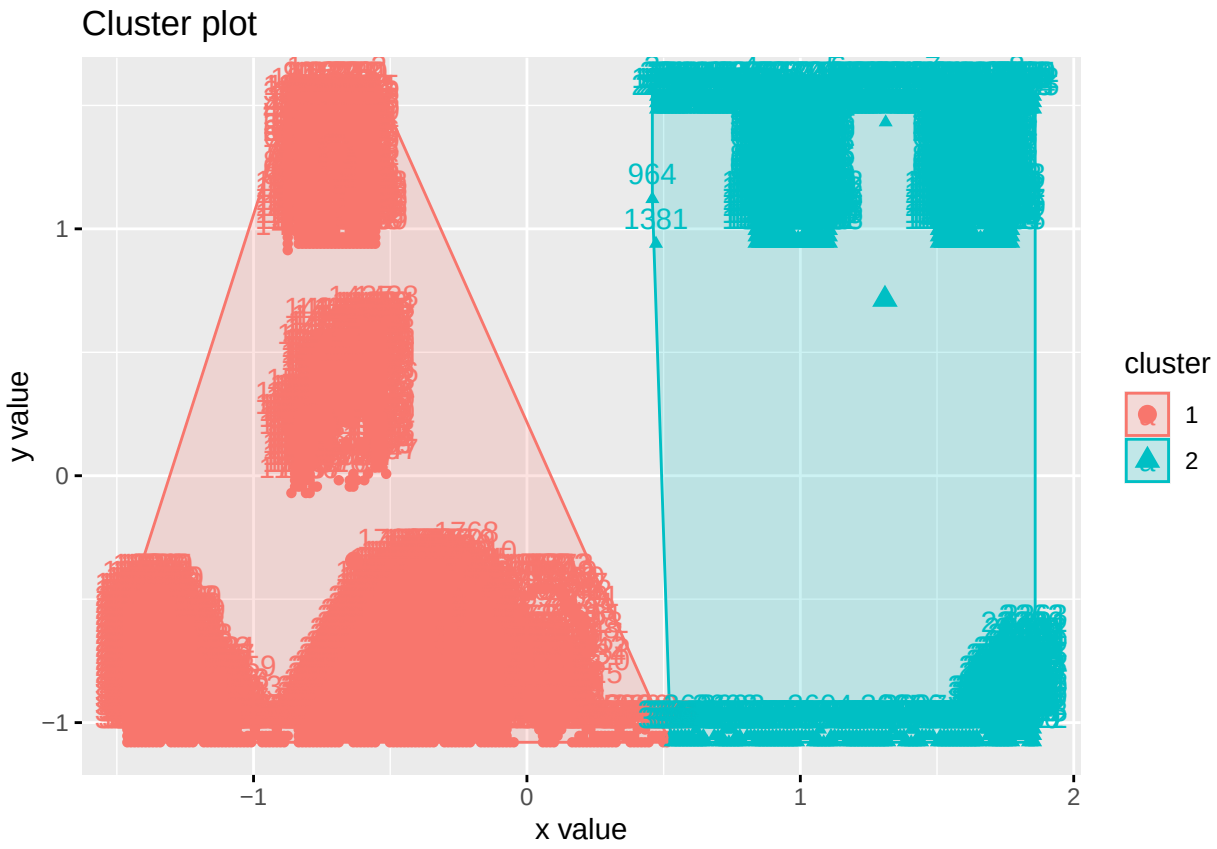
Now lets take a look at K-means clustering using the clustering data set.

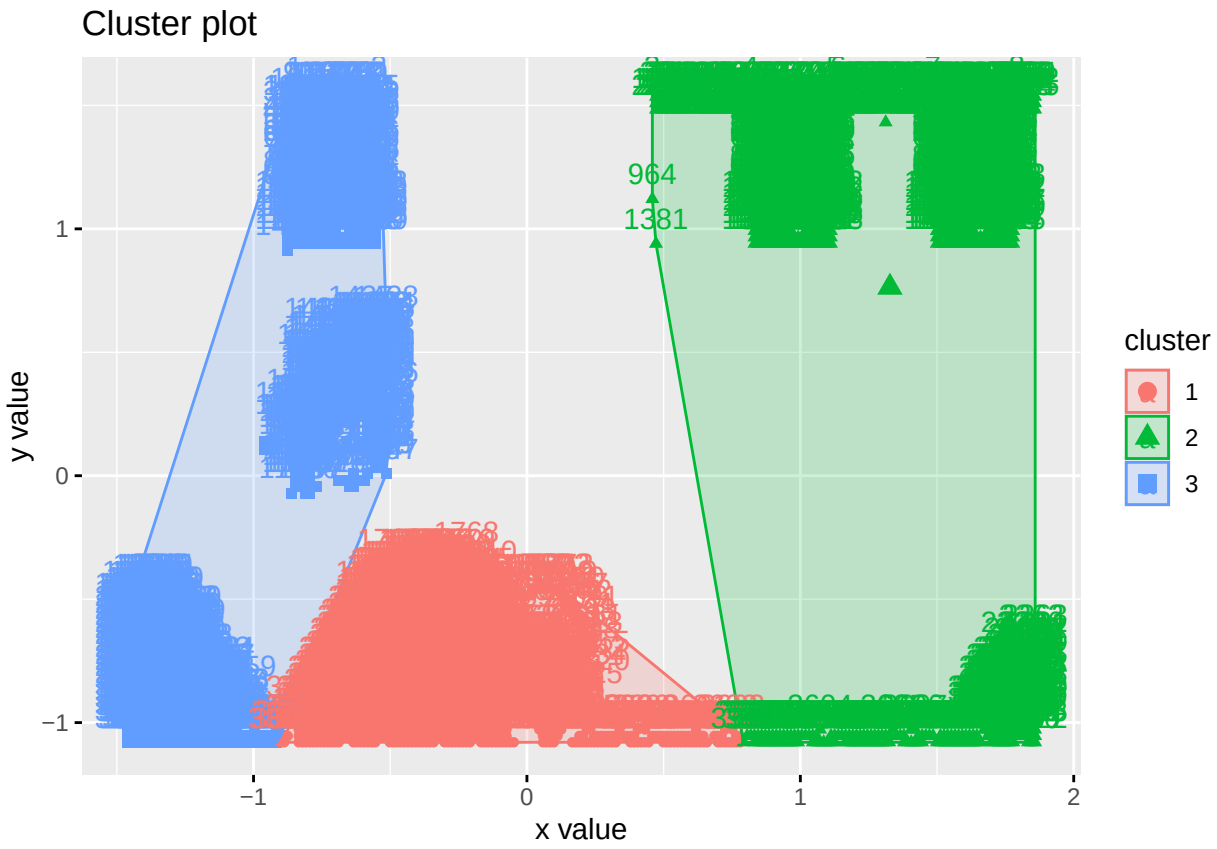
i. We will fist look at the columns we are working with then plot our data.

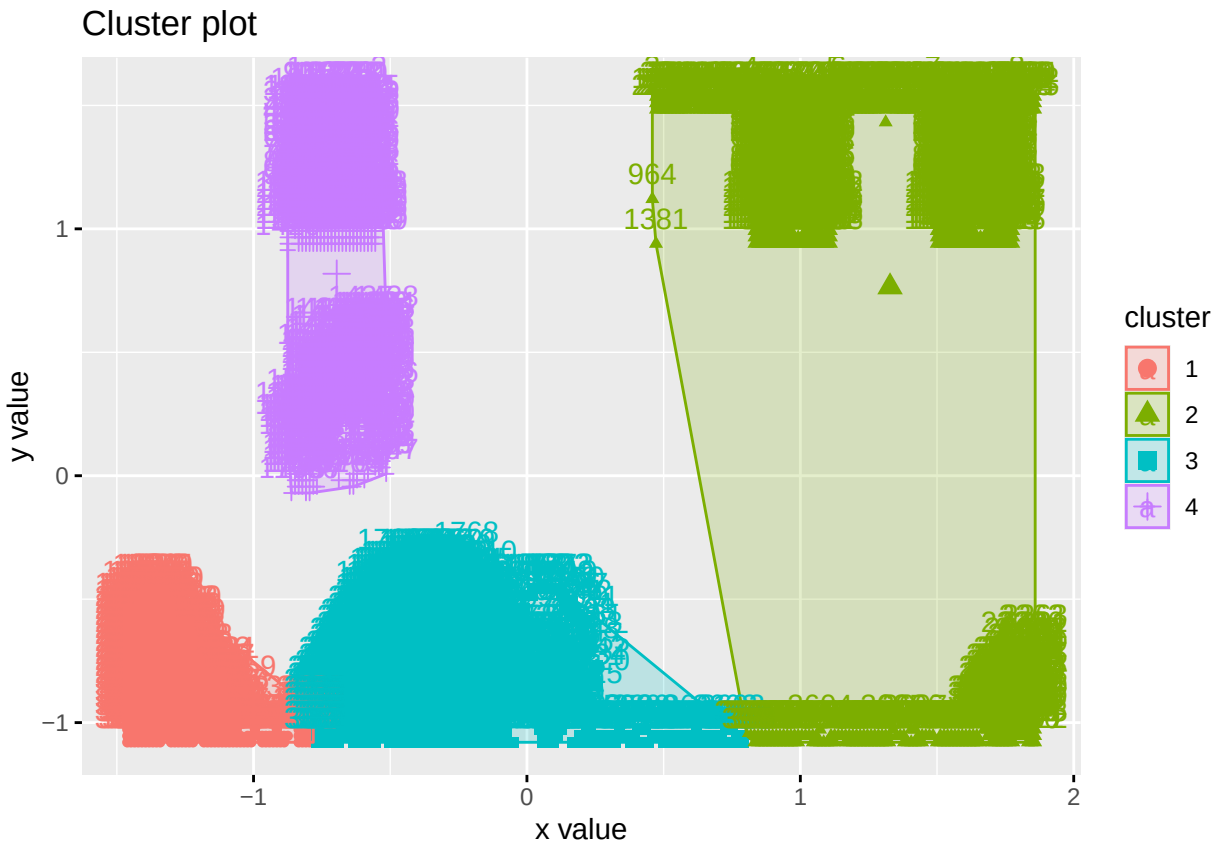
```
## [1] "x" "y"
```

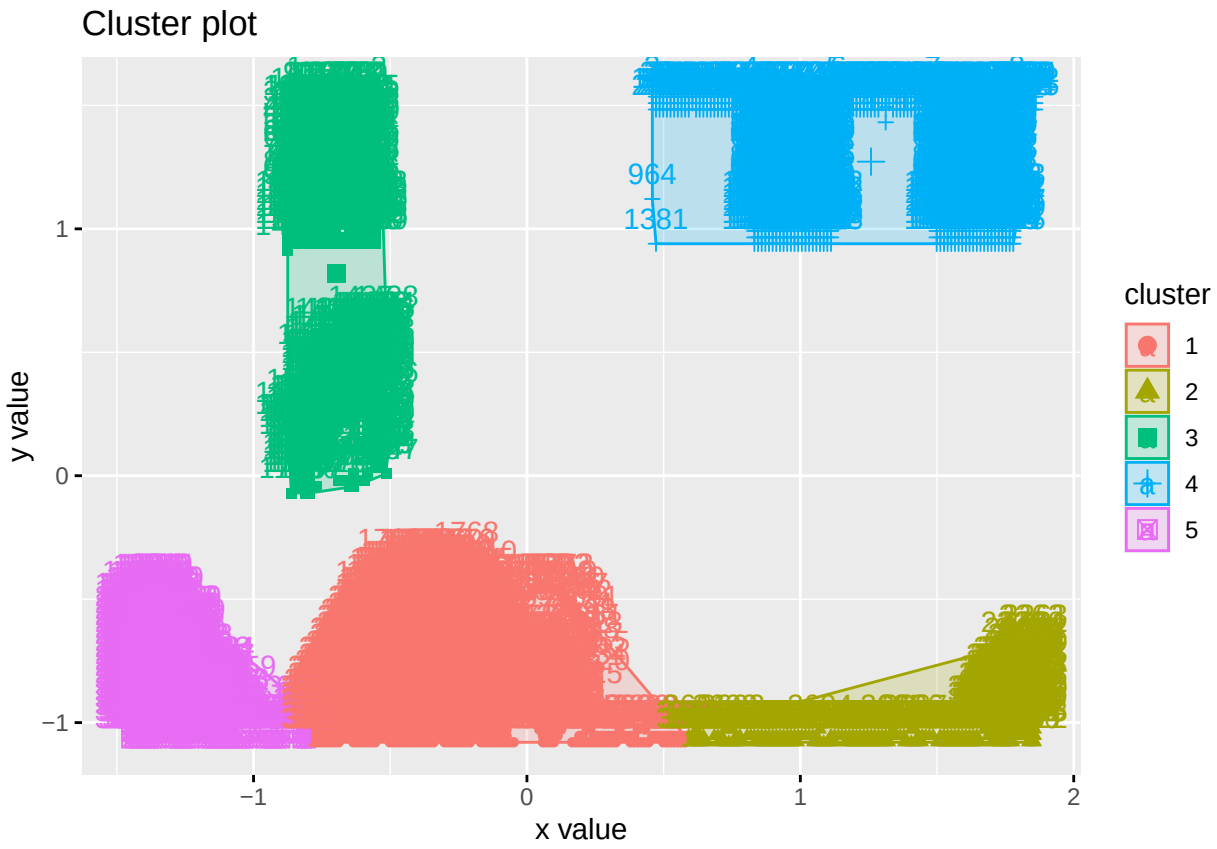


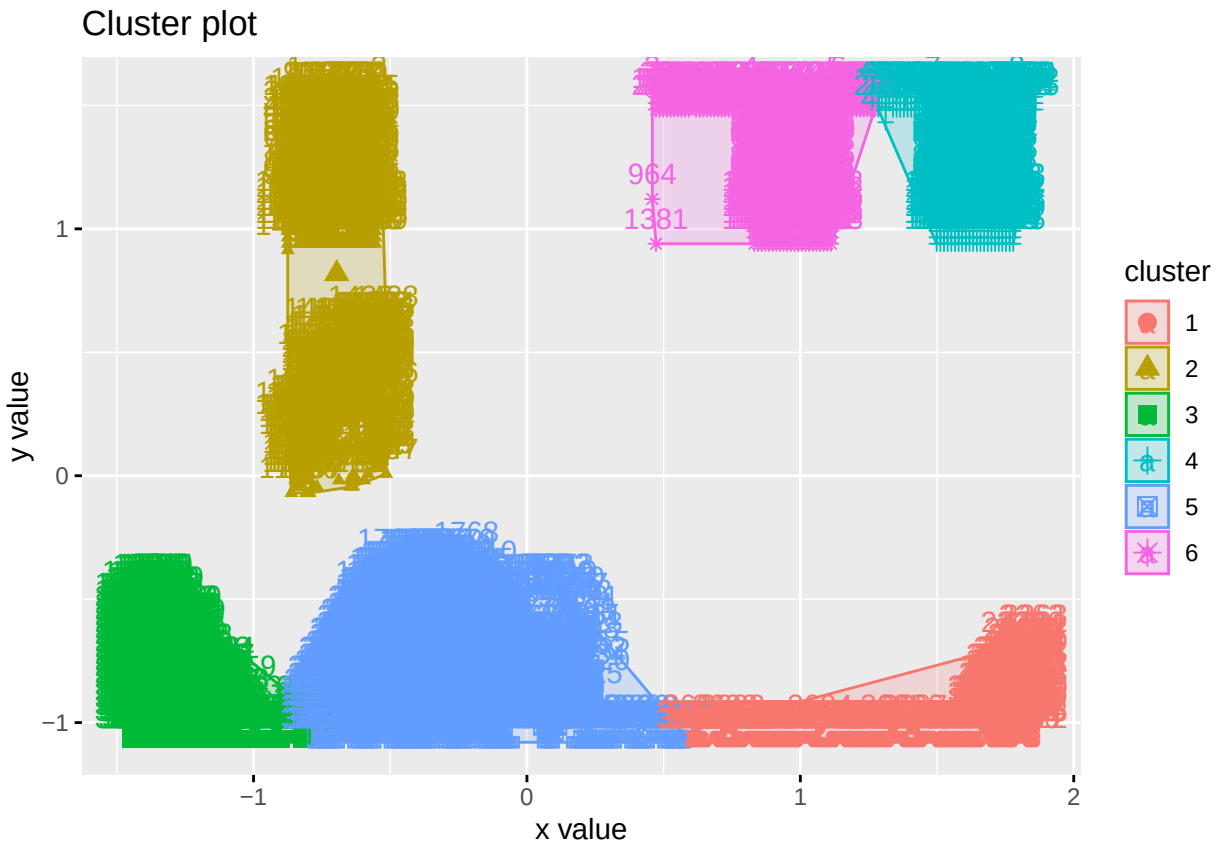
ii. Fitting the dataset using the k-means algorithm from k=2 to k=12 and plotting the resultant clusters for each using the fviz_cluster function from the factoextra library.

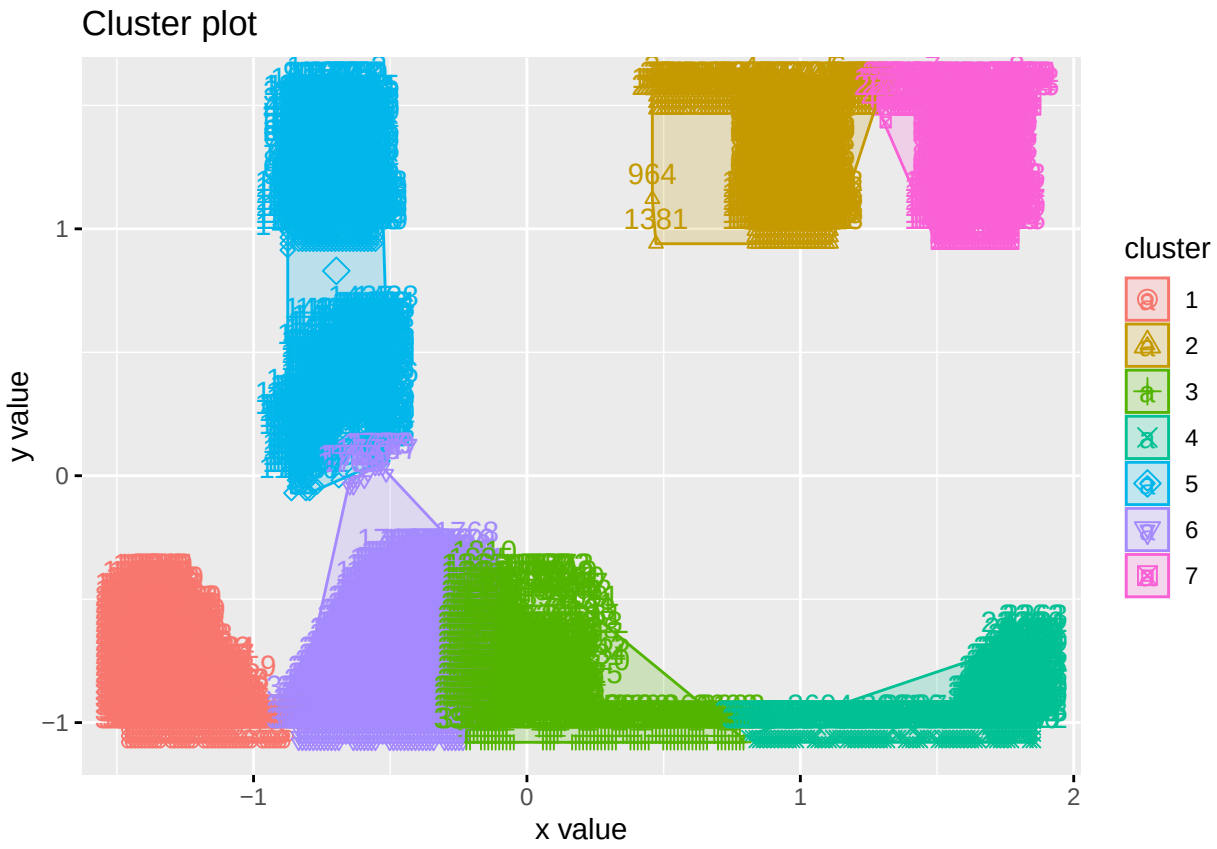

Cluster plot
964
1381
1
0
−1
−1
0
1
2
y value
x value
cluster
1
2

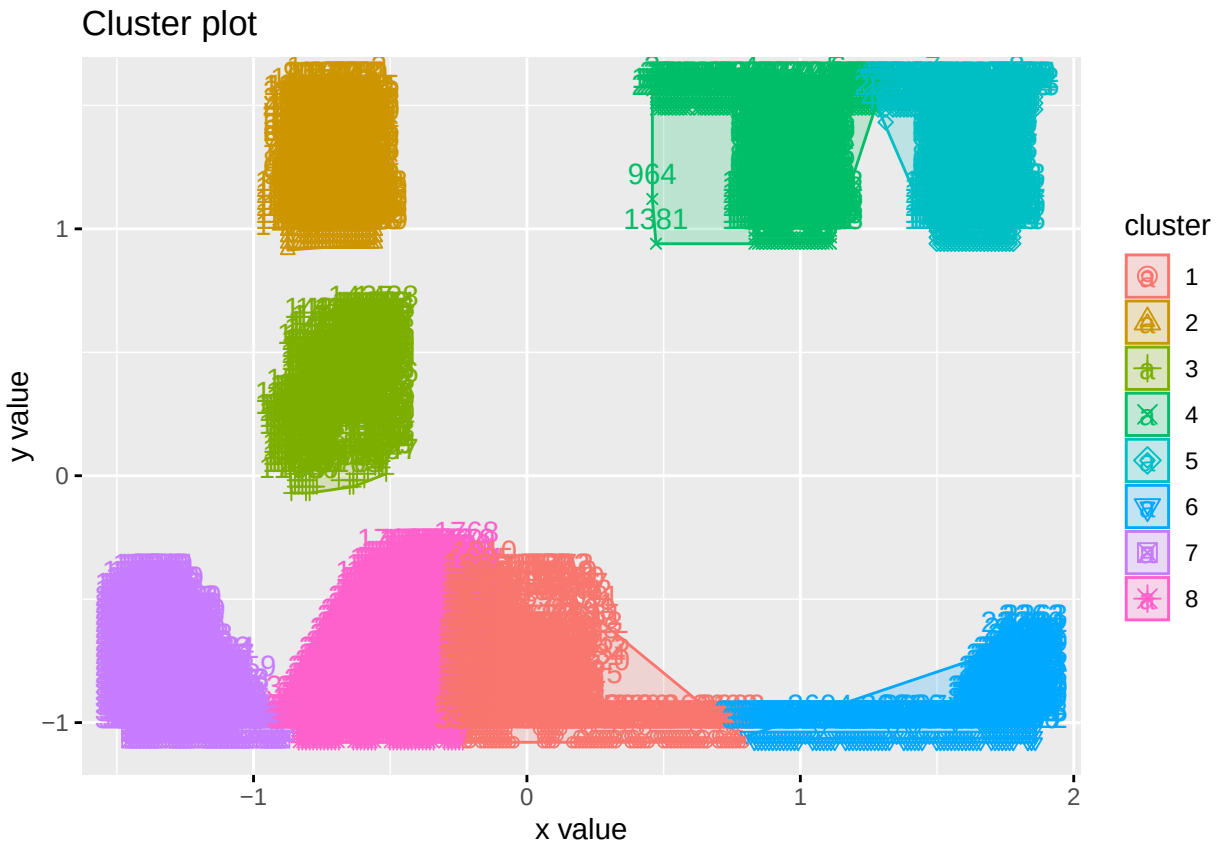

Cluster plot
y value
x value
cluster
1
2
3
964
1381
1
0
−1
−1
0
1
2

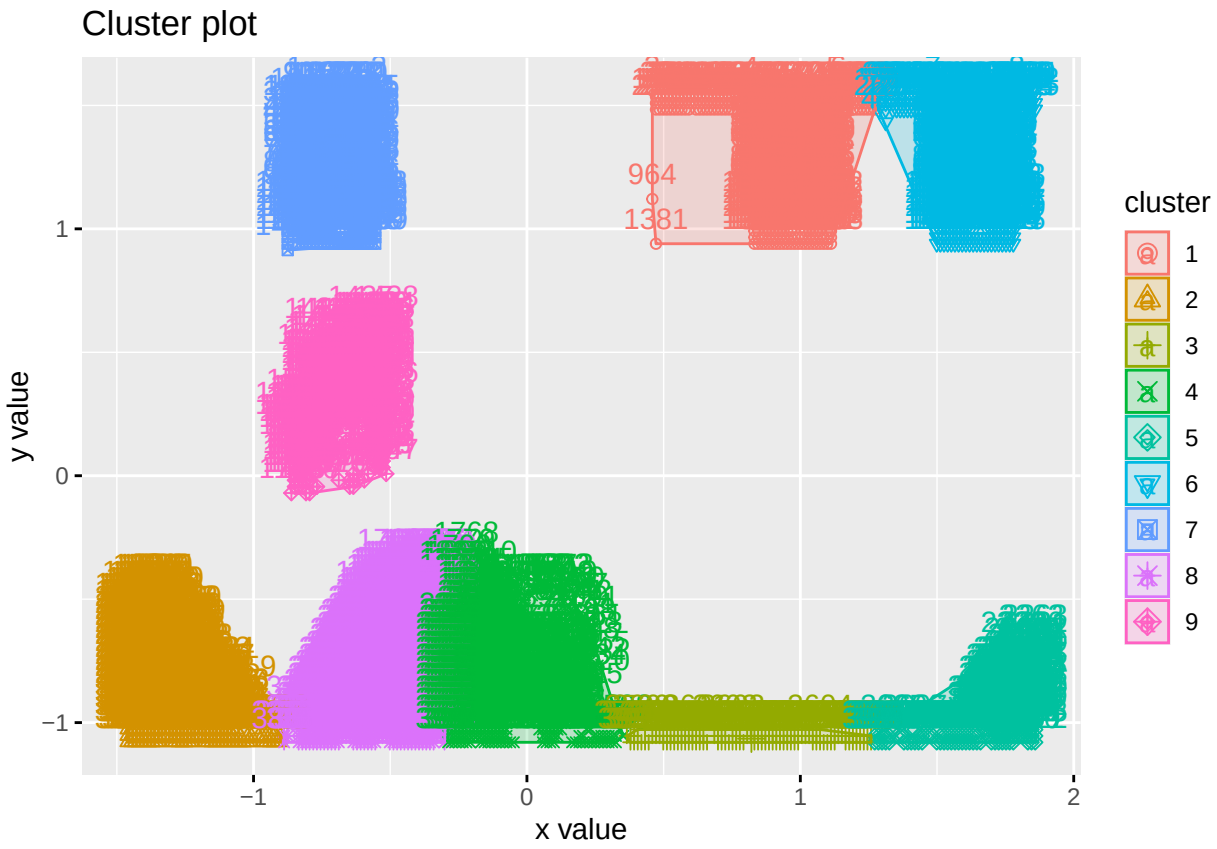

Cluster plot
y value
x value
cluster
1
2
3
4
964
1381
1
0
−1
−1
0
1
2

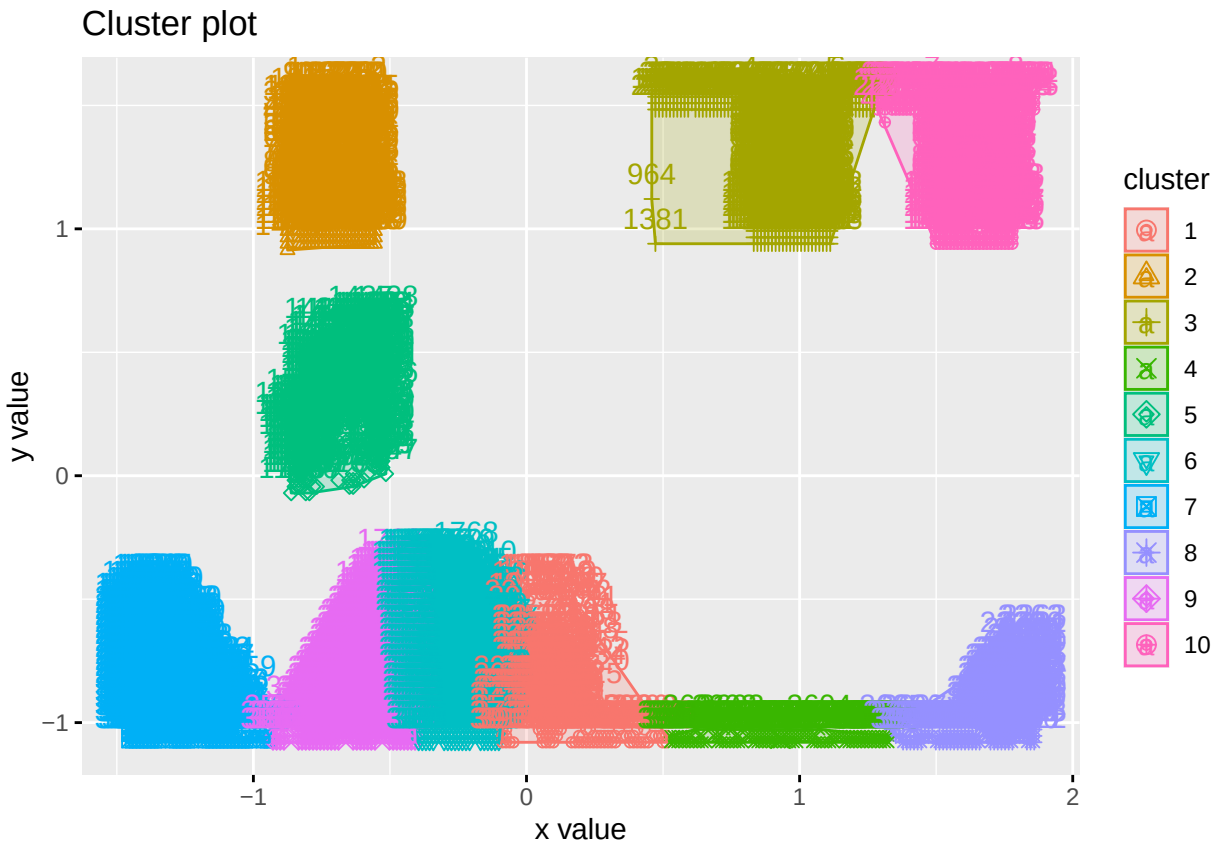

Cluster plot
y value
x value
cluster
1
2
3
4
5
964
1381
1
0
−1
−1
0
1
2

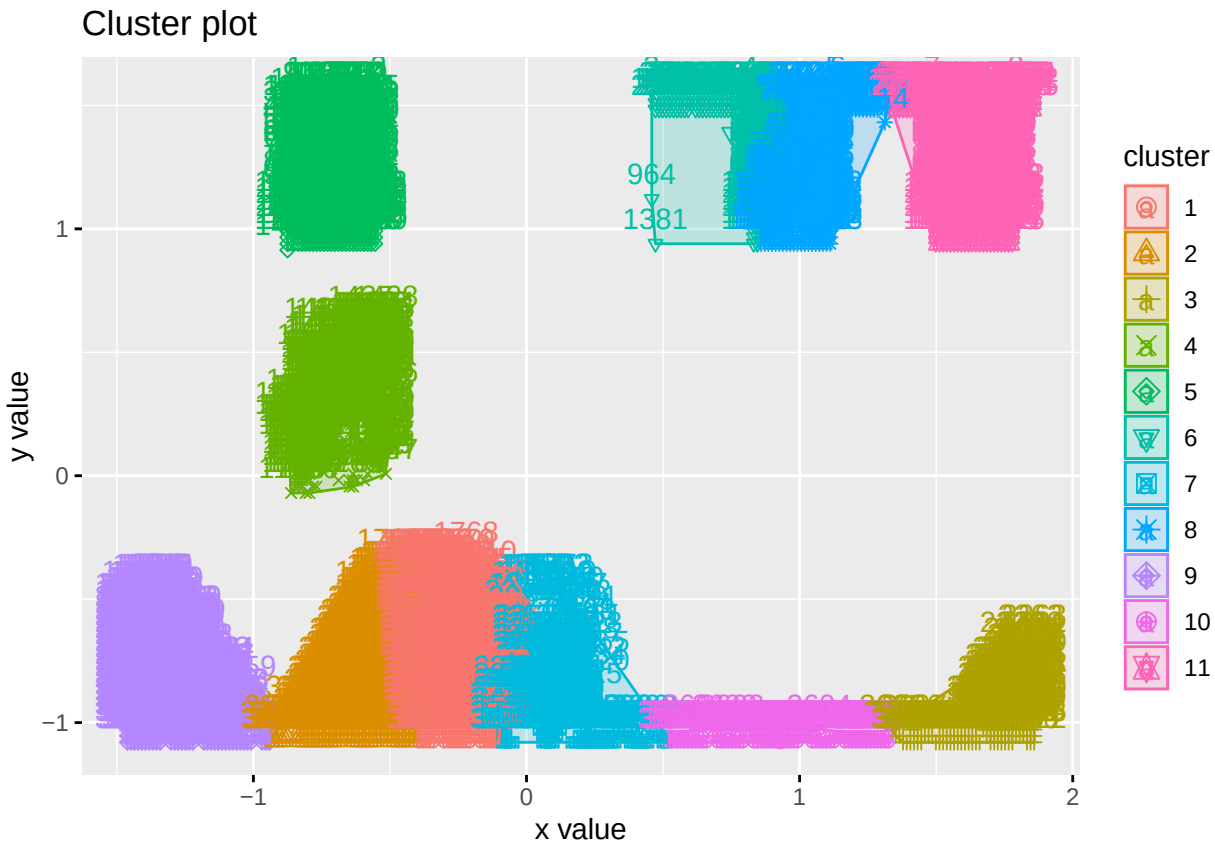

Cluster plot
y value
x value
cluster
1
2
3
4
5
6
964
1381
1
0
−1
−1
0
1
2

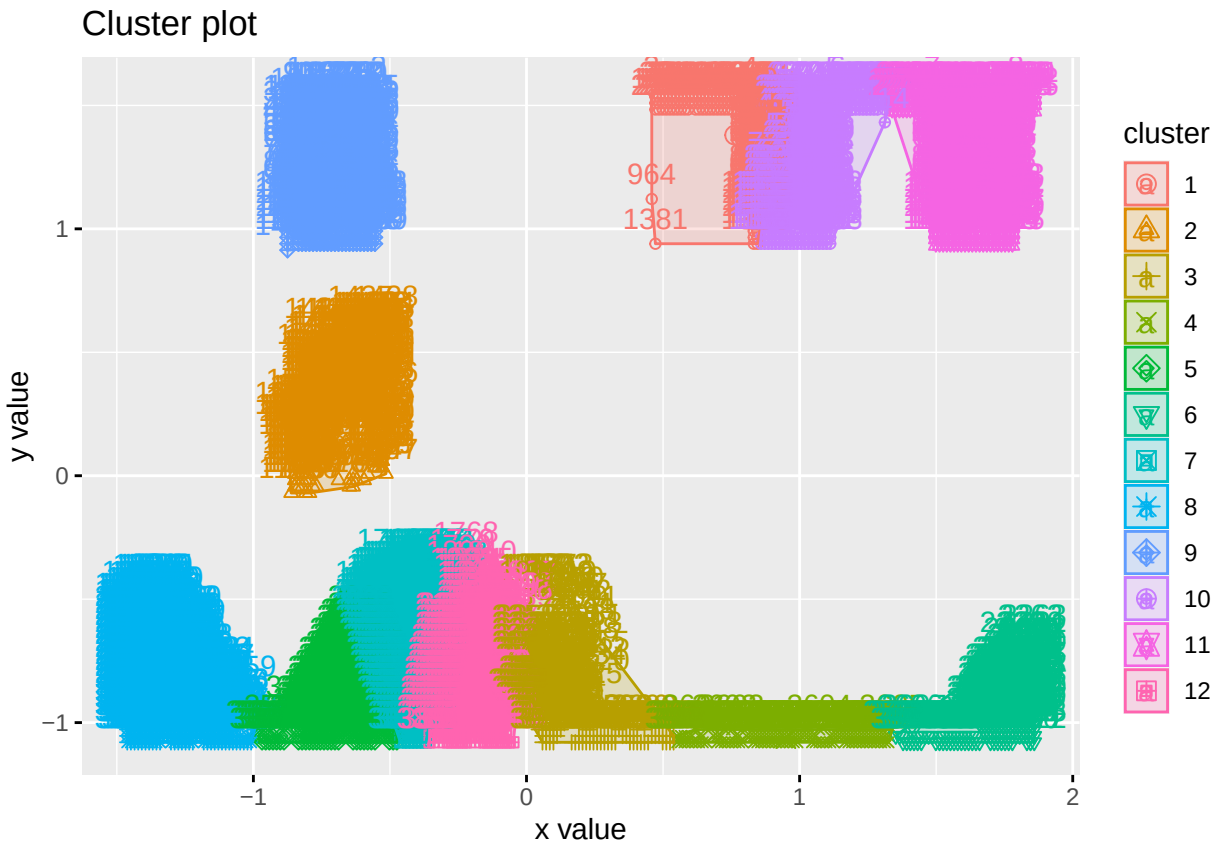

Cluster plot
y value
x value
cluster
1
2
3
4
5
6
7
964
1381
−1
0
1
2

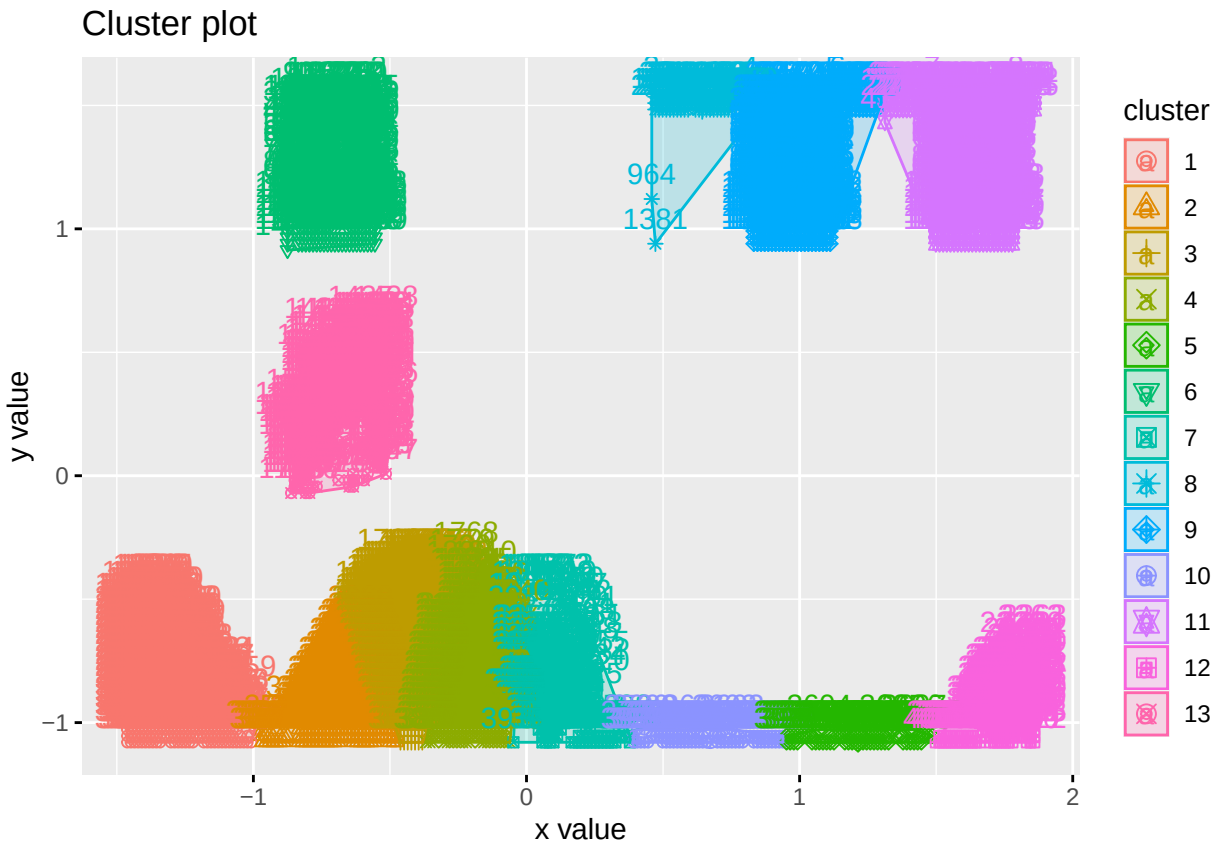

Cluster plot
y value
x value
cluster
1
2
3
4
5
6
7
8
964
1381
−1
0
1
2

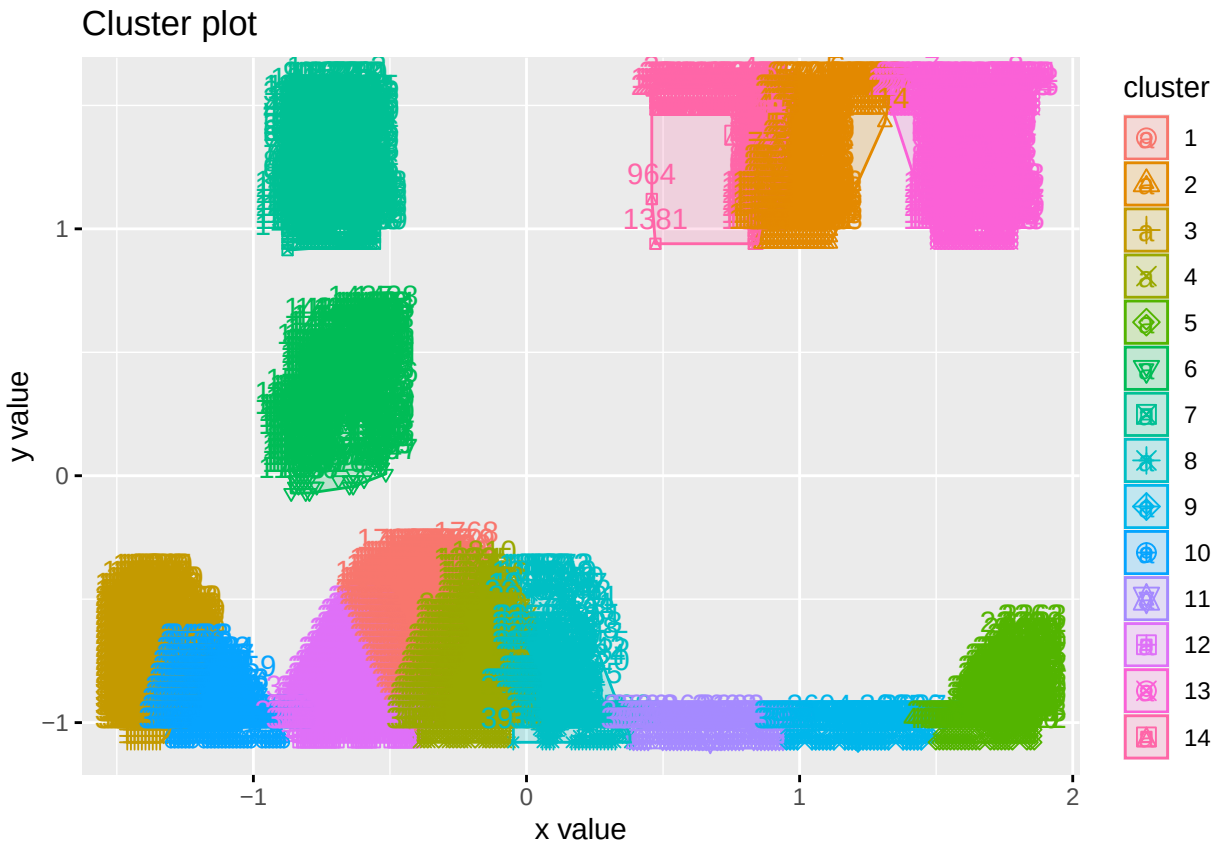

Cluster plot
y value
x value
cluster
1
2
3
4
5
6
7
8
9
964
1381
−1
0
1
2

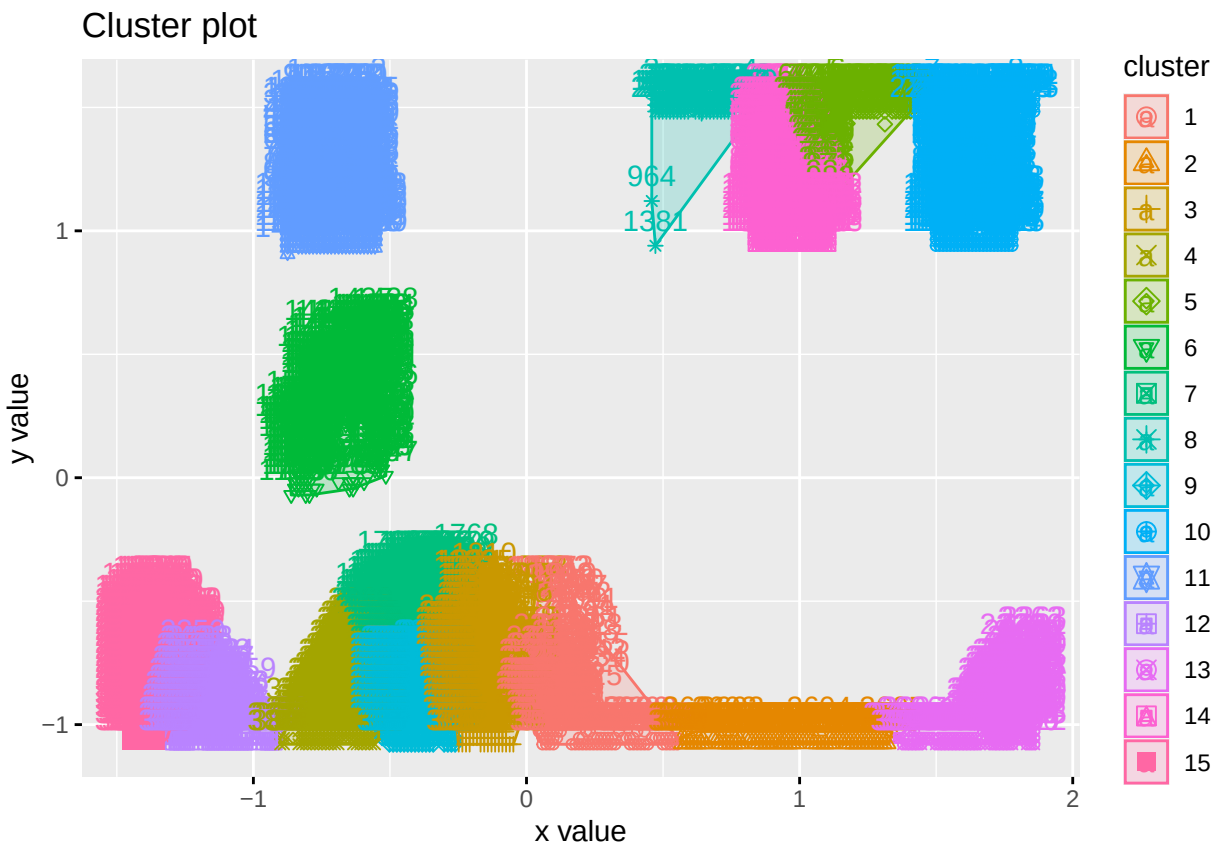

Cluster plot
y value
x value
cluster
1
2
3
4
5
6
7
8
9
10
964
1381
1768


Cluster plot
y value
x value
cluster
1
2
3
4
5
6
7
8
9
10
11
964
1381
−1
0
1
2


Cluster plot
y value
x value
cluster
1
2
3
4
5
6
7
8
9
10
11
12
964
1381


Cluster plot
y value
x value
cluster
1
2
3
4
5
6
7
8
9
10
11
12
13
964
1381
−1
0
1
2


Cluster plot
y value
x value
cluster
1
2
3
4
5
6
7
8
9
10
11
12
13
14
964
1381
−1
0
1
2



iii.

e. Plot of the average distance from the center of each cluster.

f. Based off the graph above we can conclude that the optimal number of clusters is 5. We can conclude this due to 5 being the "elbow point" in our line graph. We choose this "elbow point" as the optimal number of clusters because as we go further to the right of the line chart the number of clusters gets closer and closer to the number of data points (when the within sum of squares becomes Zero, k equals the number of data points). Therefore, we want the least sum of squares without over fitting the data, hence why we chose the "elbow point" in our sum of squares line graph.

References

Field, A., J. Miles, and Z. Field. 2012. Discovering Statistics Using R. SAGE Publications. https://books.google.com/books?id=wd2K2zC3swIC.

Lander, J. P. 2014. R for Everyone: Advanced Analytics and Graphics. Addison-Wesley Data and Analytics Series. Addison-Wesley. https://books.google.com/books?id=3eBVAgAAQBAJ.