Assignment 11.2

Aarti Ramani

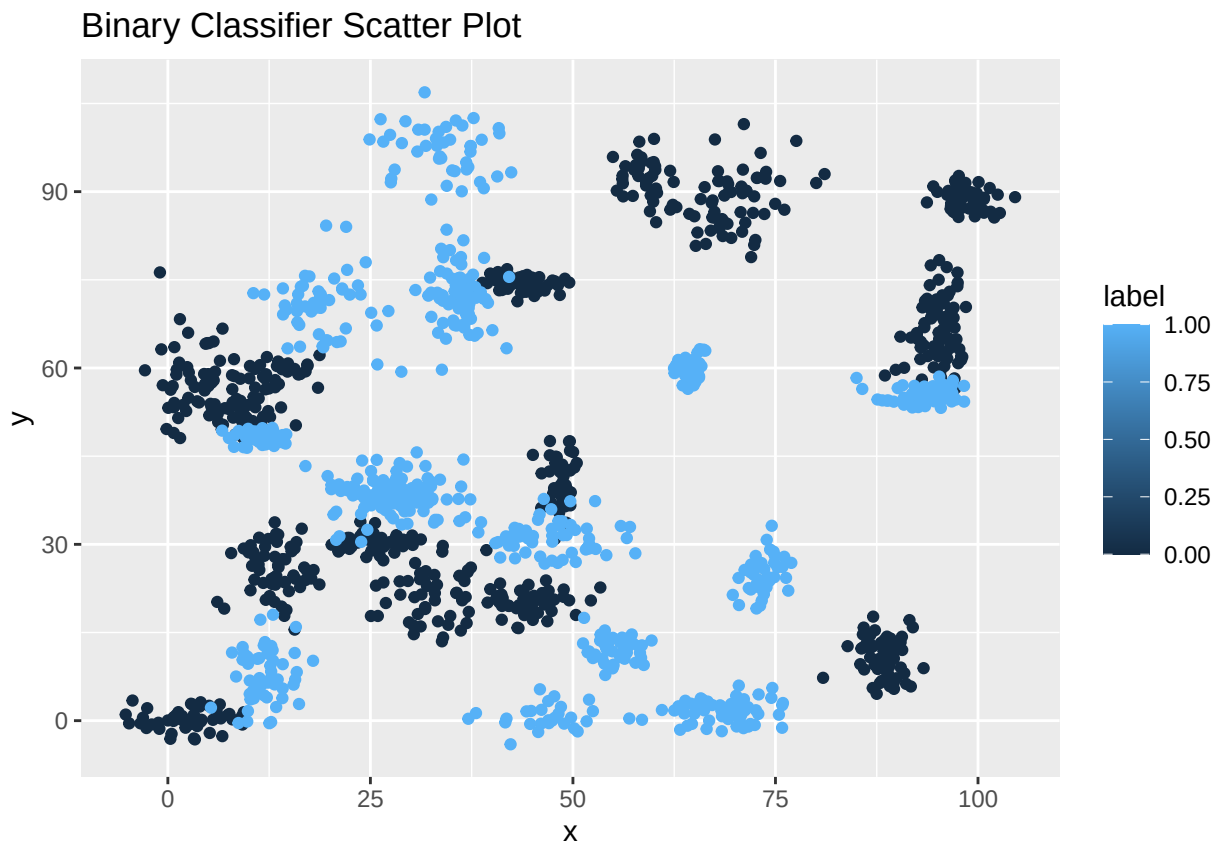
2023-02-28

REGRESSION

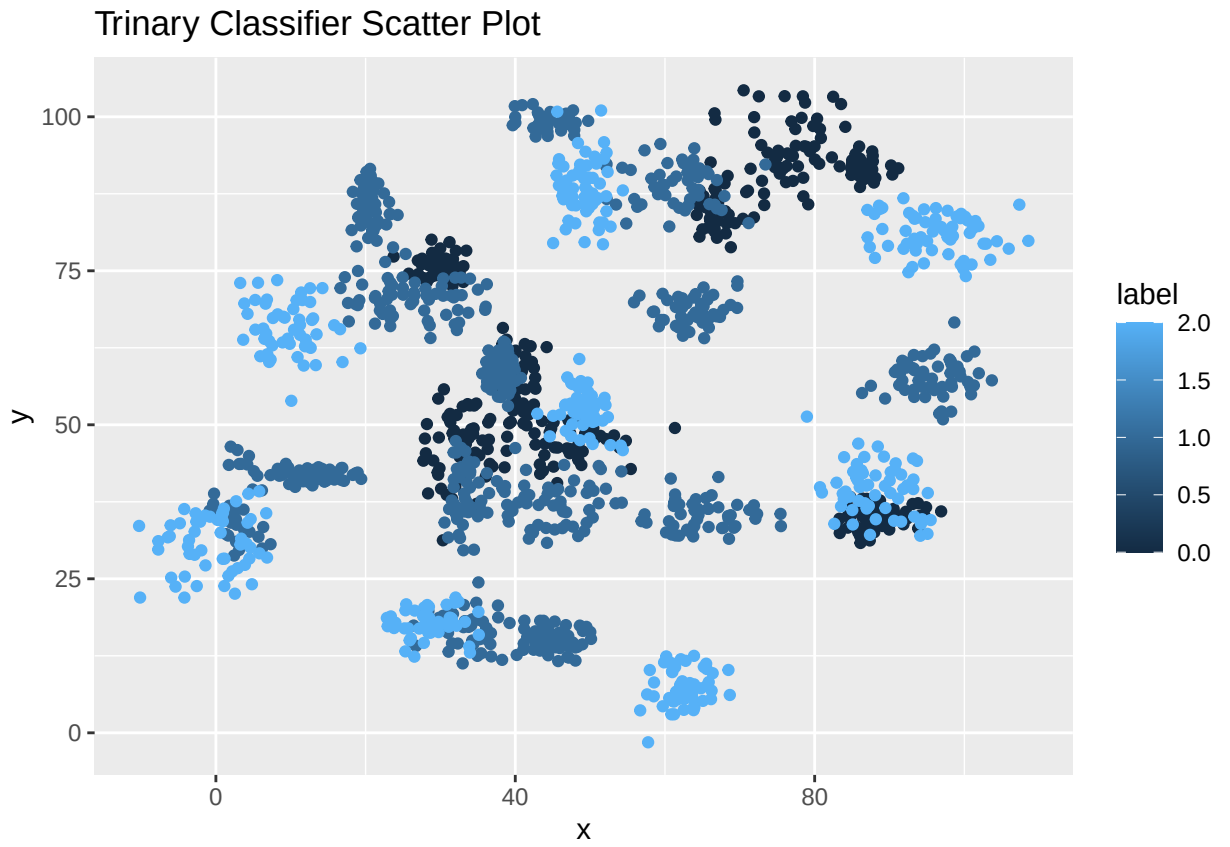
```
setwd("C:/Masters/GitHub/Winter2022/Ramani-DSC520/data/")

library(class)
library(gmodels)
library(ggplot2)
library(purrr)

binary_df <- read.csv("binary-classifier-data.csv")
trinary_df <- read.csv("trinary-classifier-data.csv")
```

Plot the data from each dataset using a scatter plot.

```
ggplot(binary_df,aes(x=x,y=y,color=label))+geom_point()+labs(title="Binary Classifier Scatter Plot")
```




```
ggplot(trinary_df,aes(x=x,y=y,color=label))+geom_point()+labs(title="Trinary Classifier Scatter Plot")
```



The k nearest neighbors algorithm categorizes an input value by looking at the labels for the k nearest points and assigning a category based on the most common label. In this problem, you will determine which points are nearest by calculating the Euclidean distance between two points. As a refresher, the Euclidean distance between two points: Fitting a model is when you use the input data to create a predictive model. There are various metrics you can use to determine how well your model fits the data. For this problem, you will focus on a single metric, accuracy. Accuracy is simply the percentage of how often the model predicts the correct result. If the model always predicts the correct result, it is 100% accurate. If the model always predicts the incorrect result, it is 0% accurate. 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.

MODELS

```
k <- c(3,5,10,15,20,25)
binary_accuracy <- NULL
for(i in 1:6)
{
    cat("KNN Binary Classisfier:",k[i])
    binary_knn<-knn(train=binary_df,test=binary_df,cl=as.factor(binary_df$label),k=k[i])
    binary_table <- CrossTable(x=binary_df$label,y=binary_knn,prop.chisq = FALSE)
    binary_accuracy[i] <-binary_table$prop.tbl[1,1]+binary_table$prop.tbl[2,2]
}
```

```
## KNN Binary Classisfier: 3
```

```
##
##    Cell Contents
## |-------------------------|
## |                       N |
## |           N / Row Total |
## |           N / Col Total |
## |         N / Table Total |
## |-------------------------|
##
##
## Total Observations in Table:  1498
##
##
##                  | binary_knn
## binary_df$label |         0 |         1 | Row Total |
## ----------------|-----------|-----------|-----------|
##               0 |       757 |        10 |       767 |
##                 |     0.987 |     0.013 |     0.512 |
##                 |     0.986 |     0.014 |           |
##                 |     0.505 |     0.007 |           |
## ----------------|-----------|-----------|-----------|
##               1 |        11 |       720 |       731 |
##                 |     0.015 |     0.985 |     0.488 |
##                 |     0.014 |     0.986 |           |
##                 |     0.007 |     0.481 |           |
## ----------------|-----------|-----------|-----------|
##    Column Total |       768 |       730 |      1498 |
##                 |     0.513 |     0.487 |           |
## ----------------|-----------|-----------|-----------|
##
##
## KNN Binary Classisfier: 5
##
##    Cell Contents
## |-------------------------|
## |                       N |
## |           N / Row Total |
## |           N / Col Total |
## |         N / Table Total |
## |-------------------------|
##
##
## Total Observations in Table:  1498
##
##
##                  | binary_knn
## binary_df$label |         0 |         1 | Row Total |
## ----------------|-----------|-----------|-----------|
##               0 |       756 |        11 |       767 |
##                 |     0.986 |     0.014 |     0.512 |
##                 |     0.981 |     0.015 |           |
##                 |     0.505 |     0.007 |           |
## ----------------|-----------|-----------|-----------|
##               1 |        15 |       716 |       731 |
```

```
##                 |     0.021 |     0.979 |     0.488 |
##                 |     0.019 |     0.985 |           |
##                 |     0.010 |     0.478 |           |
## ----------------|-----------|-----------|-----------|
##    Column Total |       771 |       727 |      1498 |
##                 |     0.515 |     0.485 |           |
## ----------------|-----------|-----------|-----------|
##
##
## KNN Binary Classisfier: 10
##
##    Cell Contents
## |-------------------------|
## |                       N |
## |           N / Row Total |
## |           N / Col Total |
## |         N / Table Total |
## |-------------------------|
##
##
## Total Observations in Table:  1498
##
##
##                 | binary_knn
## binary_df$label |         0 |         1 | Row Total |
## ----------------|-----------|-----------|-----------|
##               0 |       753 |        14 |       767 |
##                 |     0.982 |     0.018 |     0.512 |
##                 |     0.975 |     0.019 |           |
##                 |     0.503 |     0.009 |           |
## ----------------|-----------|-----------|-----------|
##               1 |        19 |       712 |       731 |
##                 |     0.026 |     0.974 |     0.488 |
##                 |     0.025 |     0.981 |           |
##                 |     0.013 |     0.475 |           |
## ----------------|-----------|-----------|-----------|
##    Column Total |       772 |       726 |      1498 |
##                 |     0.515 |     0.485 |           |
## ----------------|-----------|-----------|-----------|
##
##
## KNN Binary Classisfier: 15
##
##    Cell Contents
## |-------------------------|
## |                       N |
## |           N / Row Total |
## |           N / Col Total |
## |         N / Table Total |
## |-------------------------|
##
##
## Total Observations in Table:  1498
##
```

```
##
##                  | binary_knn
## binary_df$label |         0 |         1 | Row Total |
## ----------------|-----------|-----------|-----------|
##               0 |       752 |        15 |       767 |
##                 |     0.980 |     0.020 |     0.512 |
##                 |     0.977 |     0.021 |           |
##                 |     0.502 |     0.010 |           |
## ----------------|-----------|-----------|-----------|
##               1 |        18 |       713 |       731 |
##                 |     0.025 |     0.975 |     0.488 |
##                 |     0.023 |     0.979 |           |
##                 |     0.012 |     0.476 |           |
## ----------------|-----------|-----------|-----------|
##    Column Total |       770 |       728 |      1498 |
##                 |     0.514 |     0.486 |           |
## ----------------|-----------|-----------|-----------|
##
##
## KNN Binary Classisfier: 20
##
##    Cell Contents
## |-------------------------|
## |                       N |
## |           N / Row Total |
## |           N / Col Total |
## |         N / Table Total |
## |-------------------------|
##
##
## Total Observations in Table:  1498
##
##
##                  | binary_knn
## binary_df$label |         0 |         1 | Row Total |
## ----------------|-----------|-----------|-----------|
##               0 |       750 |        17 |       767 |
##                 |     0.978 |     0.022 |     0.512 |
##                 |     0.977 |     0.023 |           |
##                 |     0.501 |     0.011 |           |
## ----------------|-----------|-----------|-----------|
##               1 |        18 |       713 |       731 |
##                 |     0.025 |     0.975 |     0.488 |
##                 |     0.023 |     0.977 |           |
##                 |     0.012 |     0.476 |           |
## ----------------|-----------|-----------|-----------|
##    Column Total |       768 |       730 |      1498 |
##                 |     0.513 |     0.487 |           |
## ----------------|-----------|-----------|-----------|
##
##
## KNN Binary Classisfier: 25
##
##    Cell Contents
```

```
## |-------------------------|
## |                       N |
## |           N / Row Total |
## |           N / Col Total |
## |         N / Table Total |
## |-------------------------|
##
##
## Total Observations in Table:  1498
##
##
##                  | binary_knn
## binary_df$label |         0 |         1 | Row Total |
## ----------------|-----------|-----------|-----------|
##               0 |       749 |        18 |       767 |
##                 |     0.977 |     0.023 |     0.512 |
##                 |     0.979 |     0.025 |           |
##                 |     0.500 |     0.012 |           |
## ----------------|-----------|-----------|-----------|
##               1 |        16 |       715 |       731 |
##                 |     0.022 |     0.978 |     0.488 |
##                 |     0.021 |     0.975 |           |
##                 |     0.011 |     0.477 |           |
## ----------------|-----------|-----------|-----------|
##    Column Total |       765 |       733 |      1498 |
##                 |     0.511 |     0.489 |           |
## ----------------|-----------|-----------|-----------|
##
##
```

```
binary_accuracy_df <- as.data.frame(binary_accuracy)


trinary_accuracy <- NULL
for(i in 1:6)
{
    cat("KNN Trinary Classisfier:",k[i])

    trinary_knn<-knn(train=trinary_df,test=trinary_df,cl=as.factor(trinary_df$label),k=k[i])
    trinary_table <- CrossTable(x=trinary_df$label,y=trinary_knn,prop.chisq = FALSE)
    trinary_accuracy[i] <-trinary_table$prop.tbl[1,1]+trinary_table$prop.tbl[2,2]
}
```

```
## KNN Trinary Classisfier: 3
##
##    Cell Contents
## |-------------------------|
## |                       N |
## |           N / Row Total |
## |           N / Col Total |
## |         N / Table Total |
## |-------------------------|
##
##
```

```
## Total Observations in Table:  1568
##
##
##                   | trinary_knn
## trinary_df$label |         0 |         1 |         2 | Row Total |
## -----------------|-----------|-----------|-----------|-----------|
##                0 |       386 |         8 |         0 |       394 |
##                  |     0.980 |     0.020 |     0.000 |     0.251 |
##                  |     0.965 |     0.011 |     0.000 |           |
##                  |     0.246 |     0.005 |     0.000 |           |
## -----------------|-----------|-----------|-----------|-----------|
##                1 |         9 |       707 |         6 |       722 |
##                  |     0.012 |     0.979 |     0.008 |     0.460 |
##                  |     0.022 |     0.975 |     0.014 |           |
##                  |     0.006 |     0.451 |     0.004 |           |
## -----------------|-----------|-----------|-----------|-----------|
##                2 |         5 |        10 |       437 |       452 |
##                  |     0.011 |     0.022 |     0.967 |     0.288 |
##                  |     0.012 |     0.014 |     0.986 |           |
##                  |     0.003 |     0.006 |     0.279 |           |
## -----------------|-----------|-----------|-----------|-----------|
##     Column Total |       400 |       725 |       443 |      1568 |
##                  |     0.255 |     0.462 |     0.283 |           |
## -----------------|-----------|-----------|-----------|-----------|
##
##
## KNN Trinary Classisfier: 5
##
##    Cell Contents
## |-------------------------|
## |                       N |
## |           N / Row Total |
## |           N / Col Total |
## |         N / Table Total |
## |-------------------------|
##
##
## Total Observations in Table:  1568
##
##
##                   | trinary_knn
## trinary_df$label |         0 |         1 |         2 | Row Total |
## -----------------|-----------|-----------|-----------|-----------|
##                0 |       376 |        18 |         0 |       394 |
##                  |     0.954 |     0.046 |     0.000 |     0.251 |
##                  |     0.952 |     0.025 |     0.000 |           |
##                  |     0.240 |     0.011 |     0.000 |           |
## -----------------|-----------|-----------|-----------|-----------|
##                1 |        13 |       695 |        14 |       722 |
##                  |     0.018 |     0.963 |     0.019 |     0.460 |
##                  |     0.033 |     0.957 |     0.031 |           |
##                  |     0.008 |     0.443 |     0.009 |           |
## -----------------|-----------|-----------|-----------|-----------|
##                2 |         6 |        13 |       433 |       452 |
```

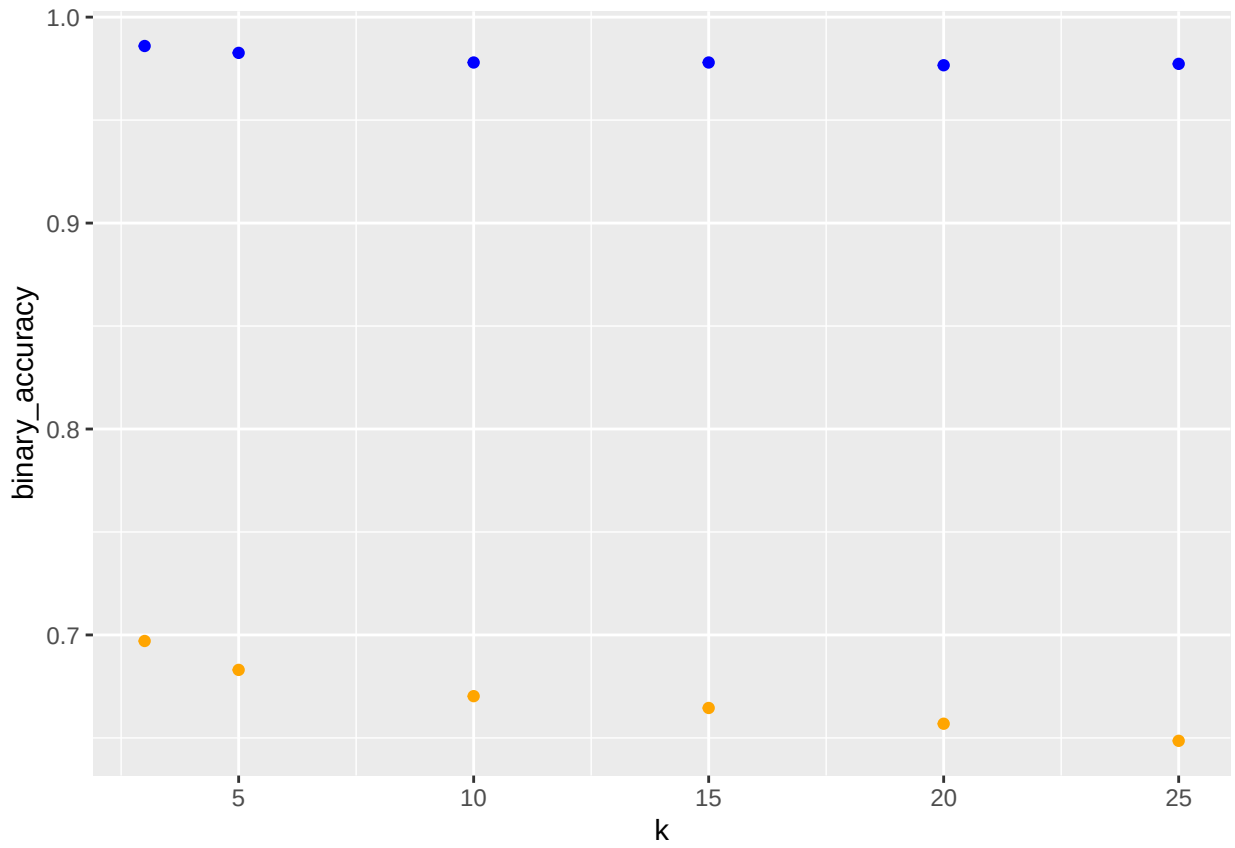
```
##                  |     0.013 |     0.029 |     0.958 |     0.288 |
##                  |     0.015 |     0.018 |     0.969 |           |
##                  |     0.004 |     0.008 |     0.276 |           |
## -----------------|-----------|-----------|-----------|-----------|
##     Column Total |       395 |       726 |       447 |      1568 |
##                  |     0.252 |     0.463 |     0.285 |           |
## -----------------|-----------|-----------|-----------|-----------|
##
##
## KNN Trinary Classisfier: 10
##
##    Cell Contents
## |-------------------------|
## |                       N |
## |           N / Row Total |
## |           N / Col Total |
## |         N / Table Total |
## |-------------------------|
##
##
## Total Observations in Table:  1568
##
##
##                  | trinary_knn
## trinary_df$label |         0 |         1 |         2 | Row Total |
## -----------------|-----------|-----------|-----------|-----------|
##                0 |       367 |        25 |         2 |       394 |
##                  |     0.931 |     0.063 |     0.005 |     0.251 |
##                  |     0.917 |     0.034 |     0.005 |           |
##                  |     0.234 |     0.016 |     0.001 |           |
## -----------------|-----------|-----------|-----------|-----------|
##                1 |        19 |       684 |        19 |       722 |
##                  |     0.026 |     0.947 |     0.026 |     0.460 |
##                  |     0.048 |     0.940 |     0.043 |           |
##                  |     0.012 |     0.436 |     0.012 |           |
## -----------------|-----------|-----------|-----------|-----------|
##                2 |        14 |        19 |       419 |       452 |
##                  |     0.031 |     0.042 |     0.927 |     0.288 |
##                  |     0.035 |     0.026 |     0.952 |           |
##                  |     0.009 |     0.012 |     0.267 |           |
## -----------------|-----------|-----------|-----------|-----------|
##     Column Total |       400 |       728 |       440 |      1568 |
##                  |     0.255 |     0.464 |     0.281 |           |
## -----------------|-----------|-----------|-----------|-----------|
##
##
## KNN Trinary Classisfier: 15
##
##    Cell Contents
## |-------------------------|
## |                       N |
## |           N / Row Total |
## |           N / Col Total |
## |         N / Table Total |
```

```
## |-------------------------|
##
##
## Total Observations in Table:  1568
##
##
##                   | trinary_knn
## trinary_df$label |         0 |         1 |         2 | Row Total |
## -----------------|-----------|-----------|-----------|-----------|
##                0 |       363 |        27 |         4 |       394 |
##                  |     0.921 |     0.069 |     0.010 |     0.251 |
##                  |     0.899 |     0.037 |     0.009 |           |
##                  |     0.232 |     0.017 |     0.003 |           |
## -----------------|-----------|-----------|-----------|-----------|
##                1 |        21 |       679 |        22 |       722 |
##                  |     0.029 |     0.940 |     0.030 |     0.460 |
##                  |     0.052 |     0.935 |     0.050 |           |
##                  |     0.013 |     0.433 |     0.014 |           |
## -----------------|-----------|-----------|-----------|-----------|
##                2 |        20 |        20 |       412 |       452 |
##                  |     0.044 |     0.044 |     0.912 |     0.288 |
##                  |     0.050 |     0.028 |     0.941 |           |
##                  |     0.013 |     0.013 |     0.263 |           |
## -----------------|-----------|-----------|-----------|-----------|
##     Column Total |       404 |       726 |       438 |      1568 |
##                  |     0.258 |     0.463 |     0.279 |           |
## -----------------|-----------|-----------|-----------|-----------|
##
##
## KNN Trinary Classisfier: 20
##
##    Cell Contents
## |-------------------------|
## |                       N |
## |           N / Row Total |
## |           N / Col Total |
## |         N / Table Total |
## |-------------------------|
##
##
## Total Observations in Table:  1568
##
##
##                   | trinary_knn
## trinary_df$label |         0 |         1 |         2 | Row Total |
## -----------------|-----------|-----------|-----------|-----------|
##                0 |       355 |        36 |         3 |       394 |
##                  |     0.901 |     0.091 |     0.008 |     0.251 |
##                  |     0.885 |     0.049 |     0.007 |           |
##                  |     0.226 |     0.023 |     0.002 |           |
## -----------------|-----------|-----------|-----------|-----------|
##                1 |        23 |       675 |        24 |       722 |
##                  |     0.032 |     0.935 |     0.033 |     0.460 |
##                  |     0.057 |     0.918 |     0.056 |           |
```

```
##                  |     0.015 |     0.430 |     0.015 |           |
## -----------------|-----------|-----------|-----------|-----------|
##                2 |        23 |        24 |       405 |       452 |
##                  |     0.051 |     0.053 |     0.896 |     0.288 |
##                  |     0.057 |     0.033 |     0.938 |           |
##                  |     0.015 |     0.015 |     0.258 |           |
## -----------------|-----------|-----------|-----------|-----------|
##     Column Total |       401 |       735 |       432 |      1568 |
##                  |     0.256 |     0.469 |     0.276 |           |
## -----------------|-----------|-----------|-----------|-----------|
##
##
## KNN Trinary Classisfier: 25
##
##    Cell Contents
## |-------------------------|
## |                       N |
## |           N / Row Total |
## |           N / Col Total |
## |         N / Table Total |
## |-------------------------|
##
##
## Total Observations in Table:  1568
##
##
##                  | trinary_knn
## trinary_df$label |         0 |         1 |         2 | Row Total |
## -----------------|-----------|-----------|-----------|-----------|
##                0 |       343 |        43 |         8 |       394 |
##                  |     0.871 |     0.109 |     0.020 |     0.251 |
##                  |     0.875 |     0.058 |     0.018 |           |
##                  |     0.219 |     0.027 |     0.005 |           |
## -----------------|-----------|-----------|-----------|-----------|
##                1 |        24 |       674 |        24 |       722 |
##                  |     0.033 |     0.934 |     0.033 |     0.460 |
##                  |     0.061 |     0.913 |     0.055 |           |
##                  |     0.015 |     0.430 |     0.015 |           |
## -----------------|-----------|-----------|-----------|-----------|
##                2 |        25 |        21 |       406 |       452 |
##                  |     0.055 |     0.046 |     0.898 |     0.288 |
##                  |     0.064 |     0.028 |     0.927 |           |
##                  |     0.016 |     0.013 |     0.259 |           |
## -----------------|-----------|-----------|-----------|-----------|
##     Column Total |       392 |       738 |       438 |      1568 |
##                  |     0.250 |     0.471 |     0.279 |           |
## -----------------|-----------|-----------|-----------|-----------|
##
##
```

```
trinary_accuracy_df <- as.data.frame(trinary_accuracy)
```

```
ggplot() +
geom_point(data=binary_accuracy_df,aes(x=k, y=binary_accuracy), color='blue') + geom_point(data=trinary_
```



I don't think a linear classifier would work well on these datasets because the plots are not in a straight line. Neither variables have a visual correlation with being in one group or the other.

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

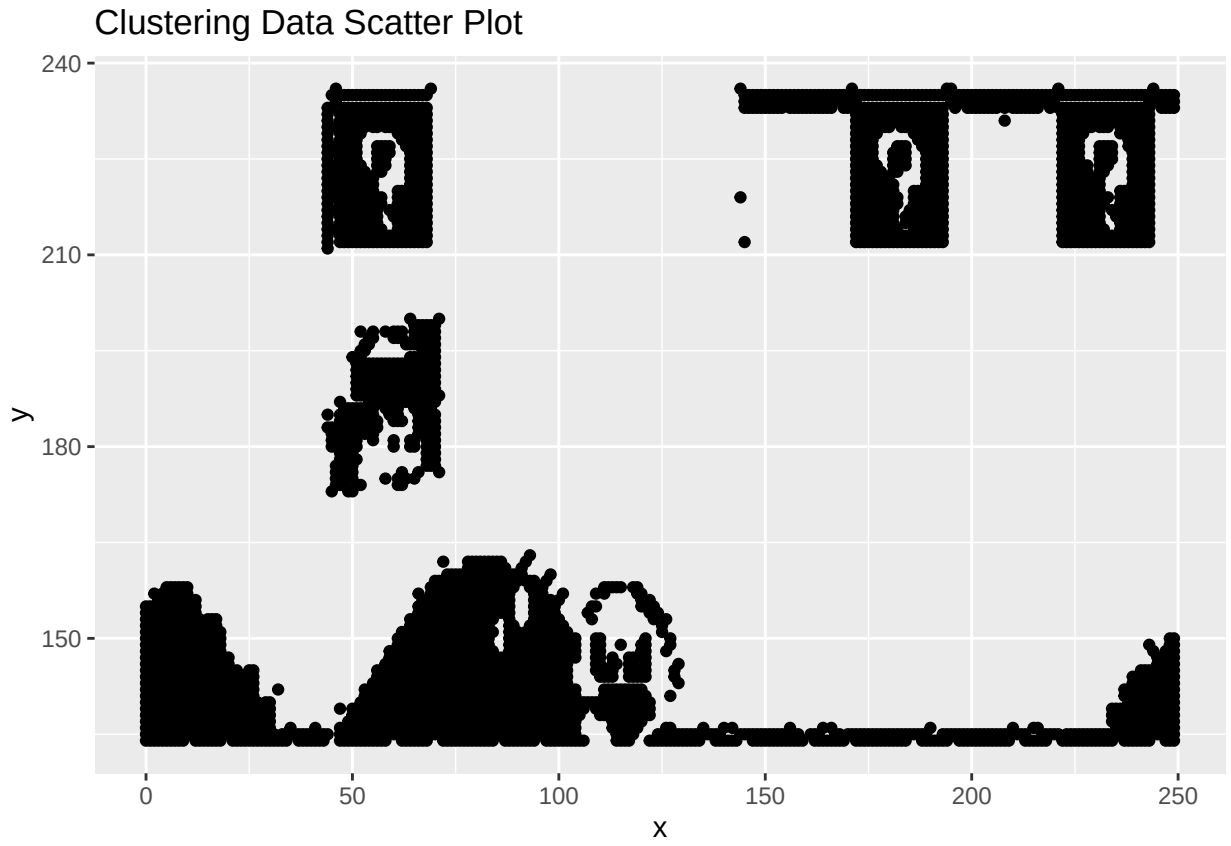
For the binary data set, week 10's accuracy was about 58%. This week it is about 97%. This is because, KNN is a non-parametric model whereas LR is a parametric model (finite number of parameters). KNN supports non-linear solutions where LR supports only linear solutions.

CLUSTERING

```
setwd("C:/Masters/GitHub/Winter2022/Ramani-DSC520/data/")
clustering_df <- read.csv("clustering-data.csv")
```

Plot the dataset using a scatter plot.

```
ggplot(clustering_df,aes(x=x,y=y))+geom_point()+labs(title="Clustering Data Scatter Plot")
```

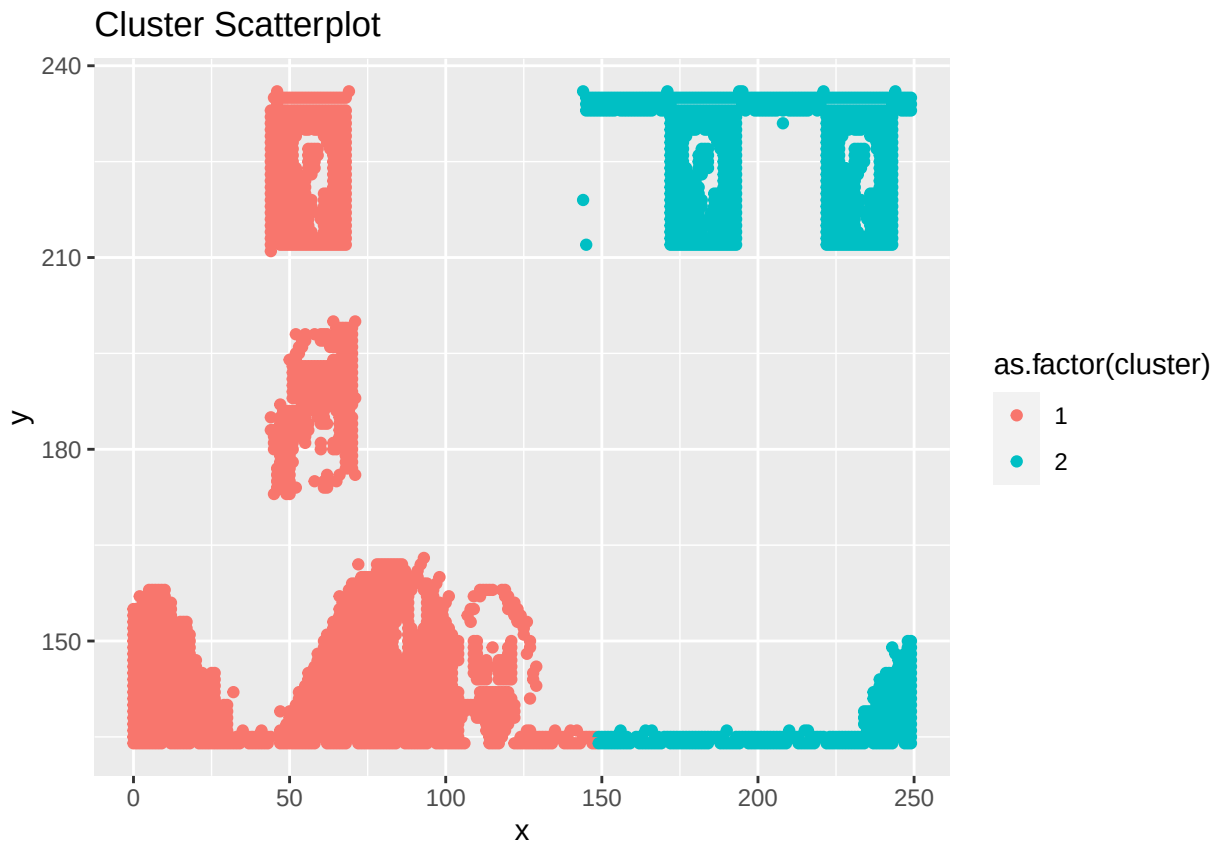


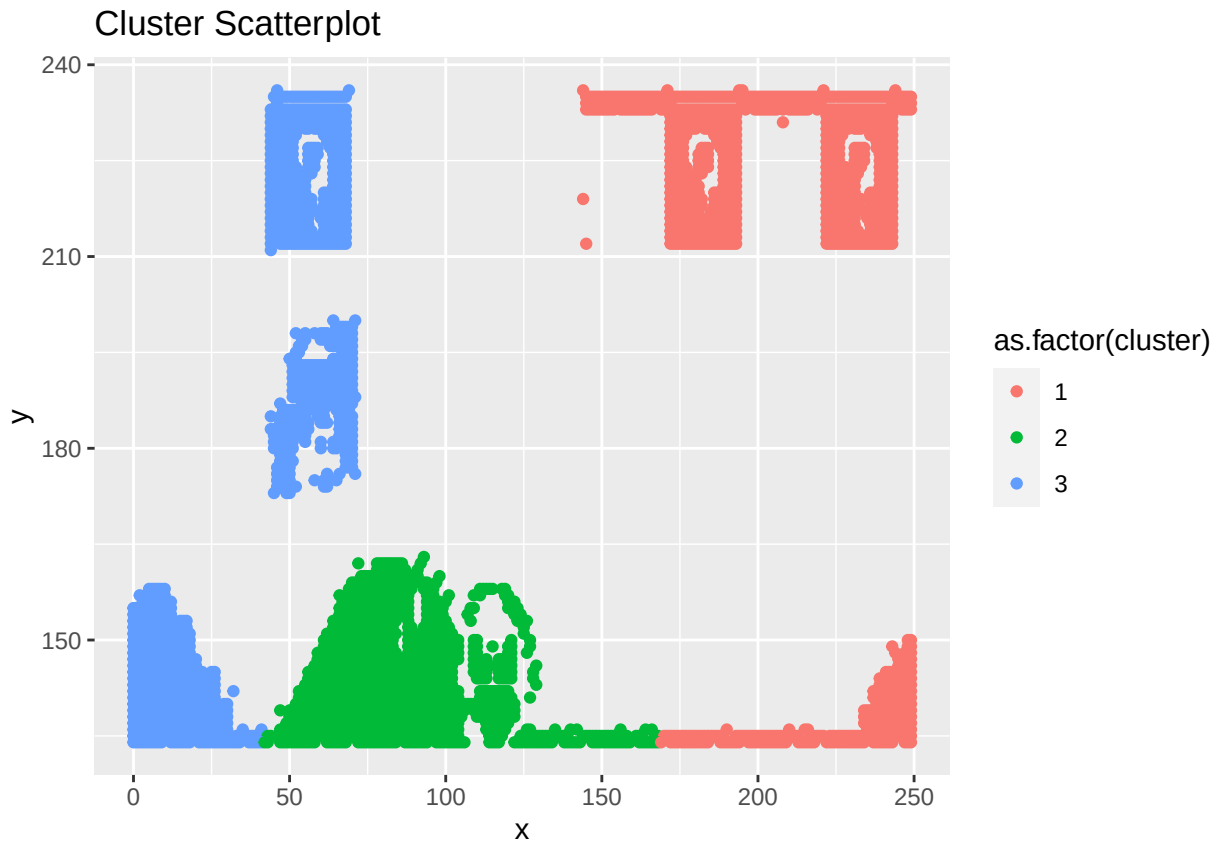
Fit the dataset using the k-means algorithm from k=2 to k=12. Create a scatter plot of the resultant clusters for each value of k.

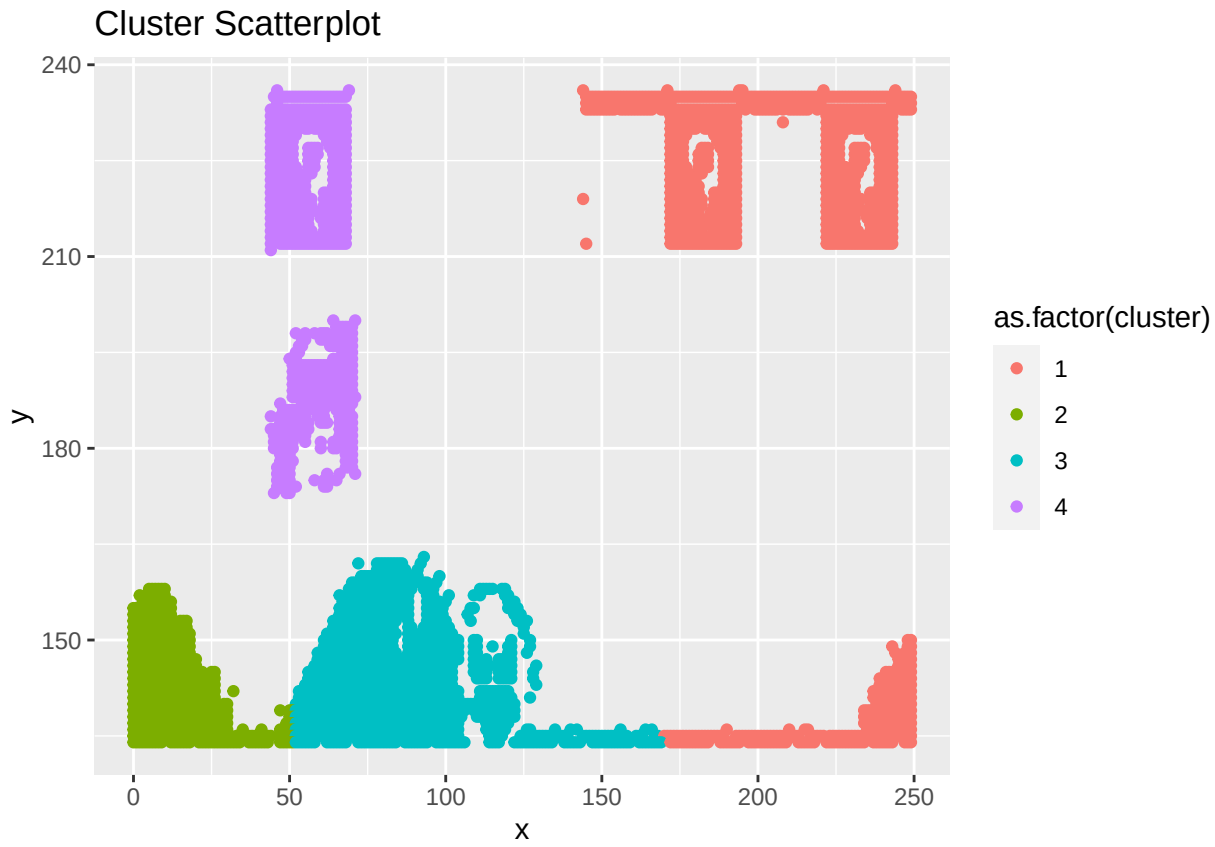
Use the average distance from the center of each cluster as a measure of how well the model fits the data. To calculate this metric, simply compute the distance of each data point to the center of the cluster it is assigned to and take the average value of all of those distances.

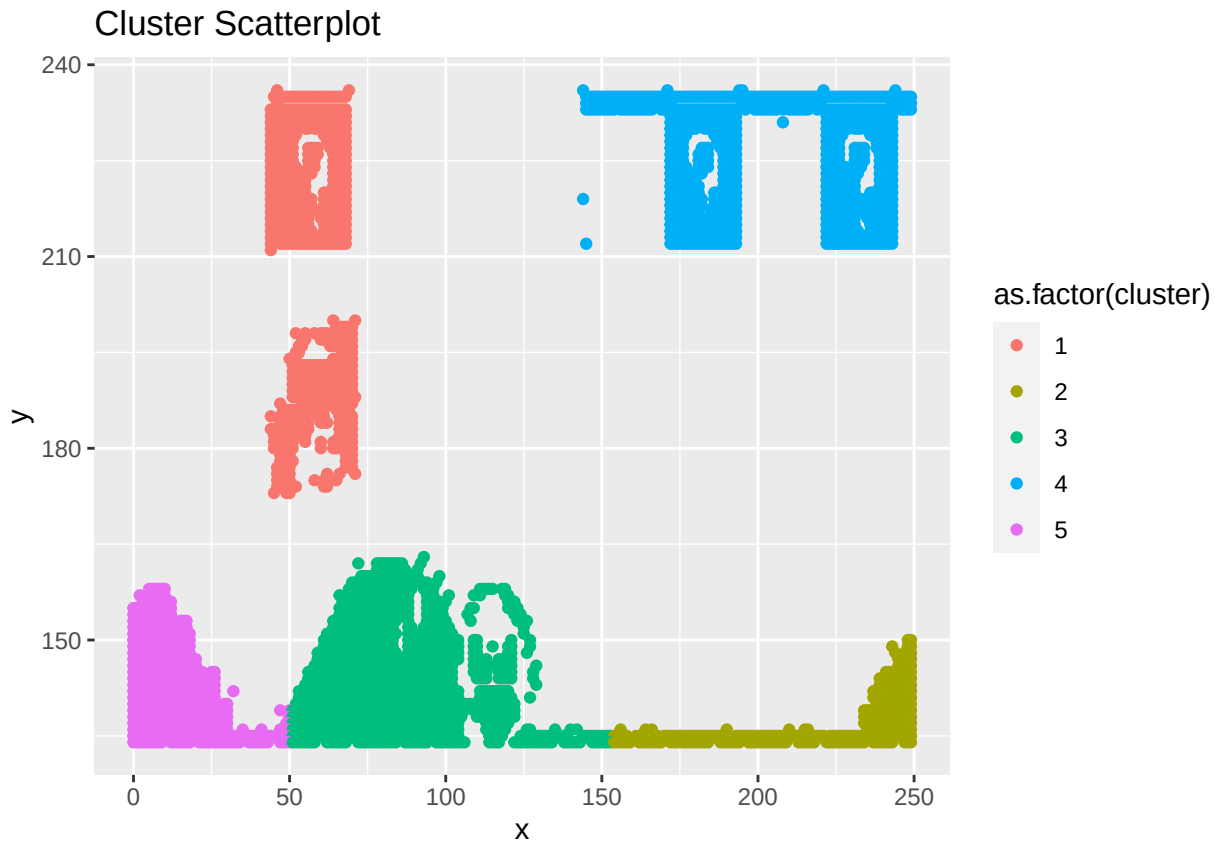
```
set.seed(2345)
clusters <-NULL
```

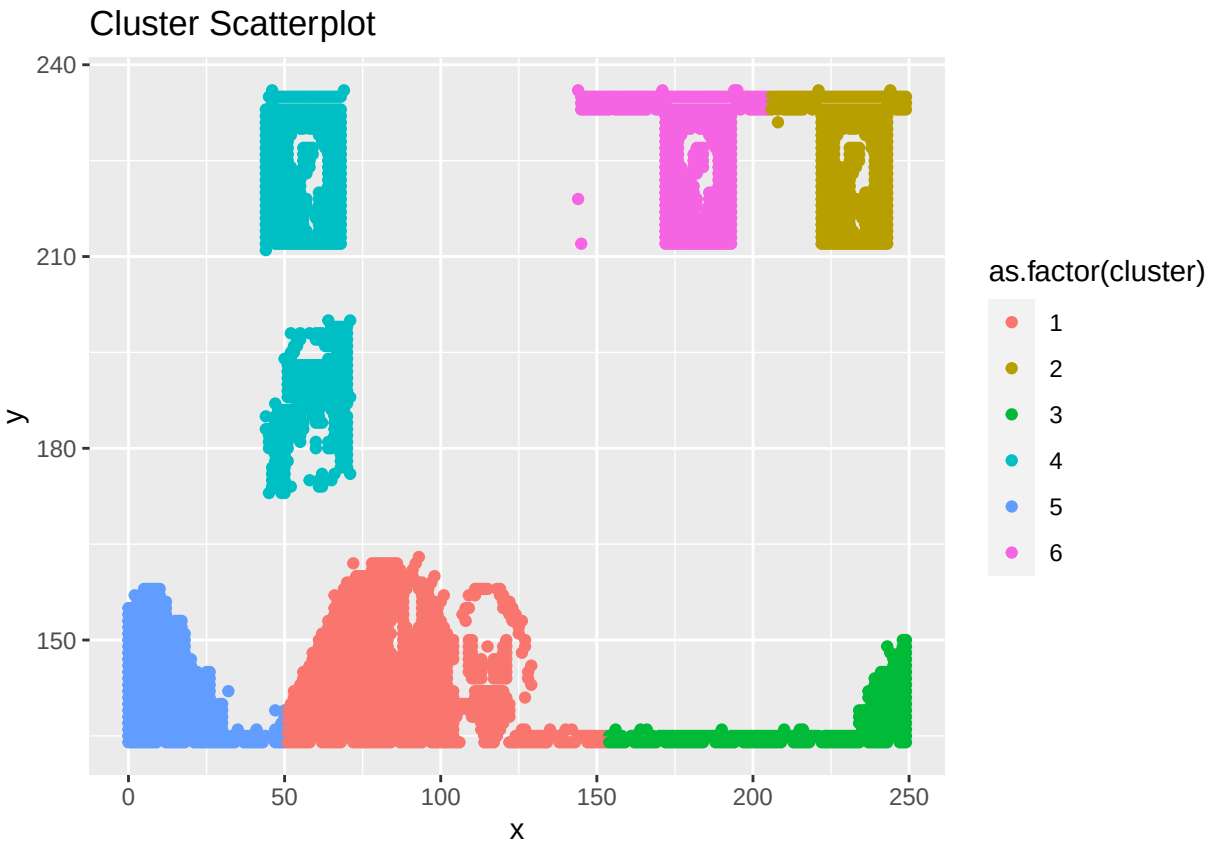
```
avg_dist <- NULL

for(i in 2:12)
{
  clusters_kMeans <- kmeans(clustering_df,i)
  clusters[i] <- as.data.frame(clusters_kMeans[["cluster"]])
  clustering_df["cluster"] <- clusters[i]

  avg_dist[i] <- sum(clusters_kMeans[["withinss"]]/clusters_kMeans[["size"]])

  print(ggplot(clustering_df,aes(x=x,y=y,color=as.factor(cluster))) +
  geom_point() + labs(title = "Cluster Scatterplot"))

}
```

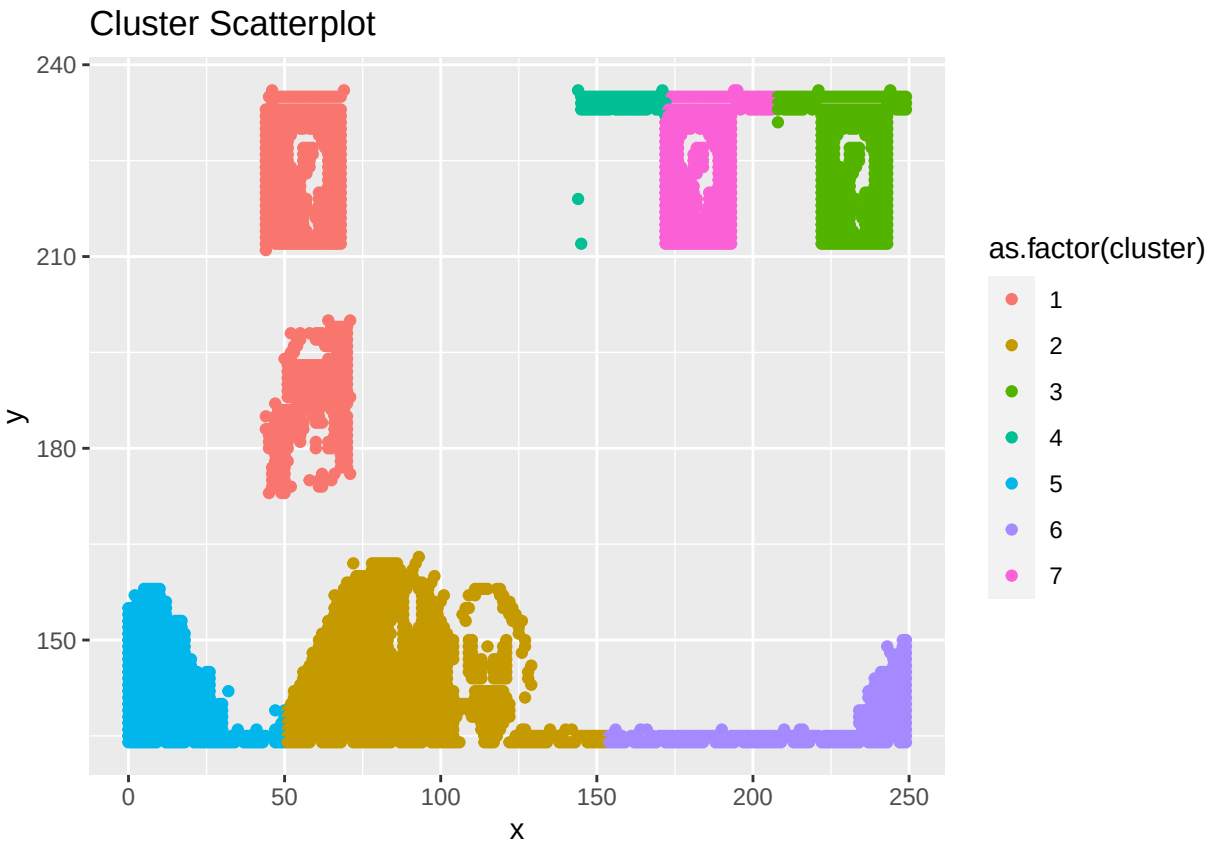


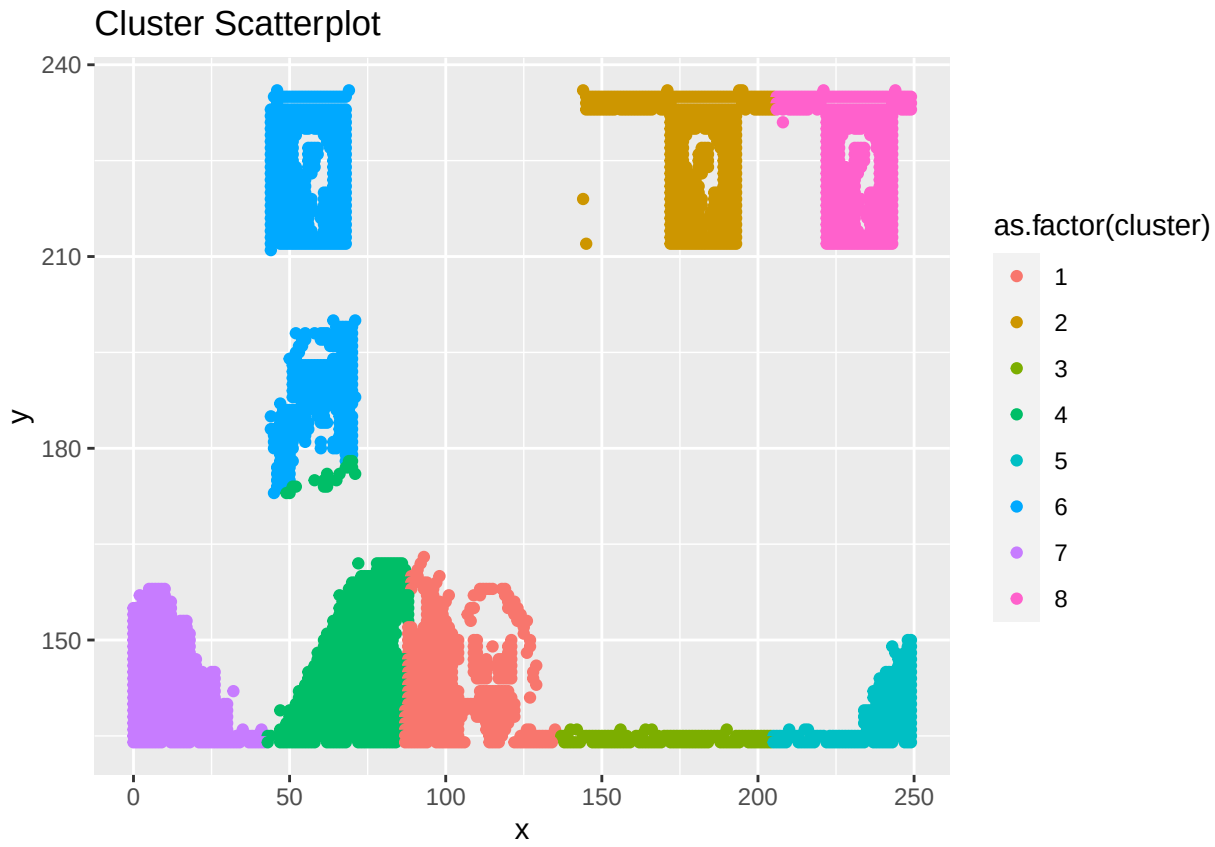

Cluster Scatterplot
240
210
180
150
y
0
50
100
150
200
250
x
as.factor(cluster)
1
2
3

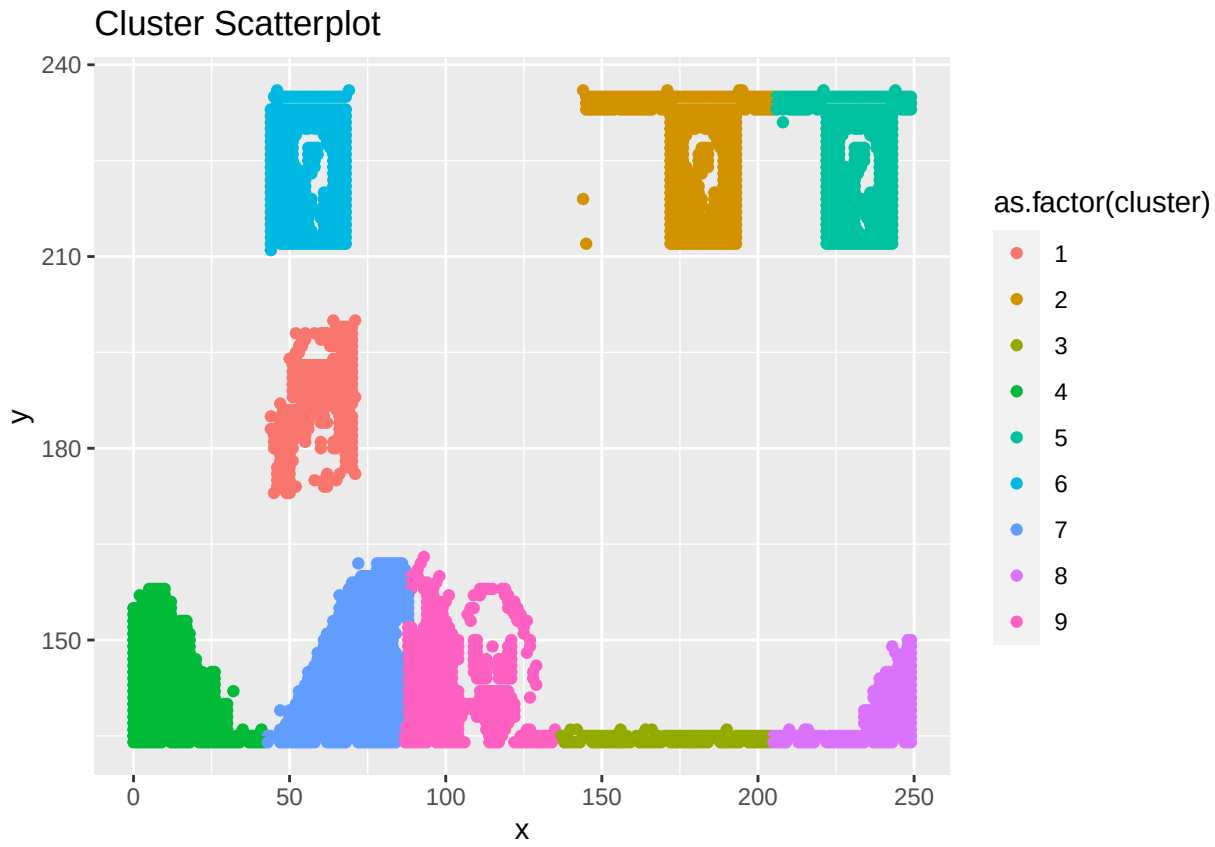

Cluster Scatterplot
240
210
180
150
y
0
50
100
150
200
250
x
as.factor(cluster)
1
2
3
4

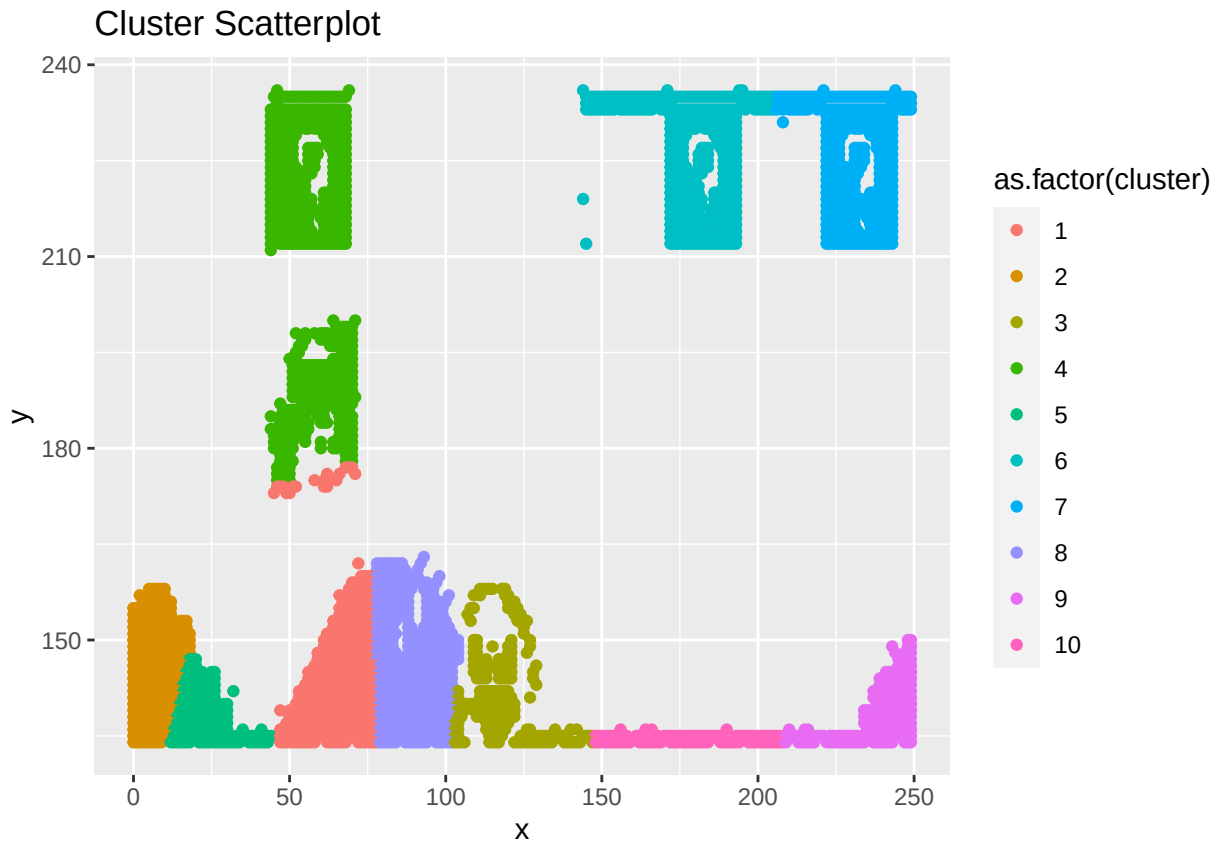

Cluster Scatterplot
240
210
180
150
y
0
50
100
150
200
250
x
as.factor(cluster)
1
2
3
4
5

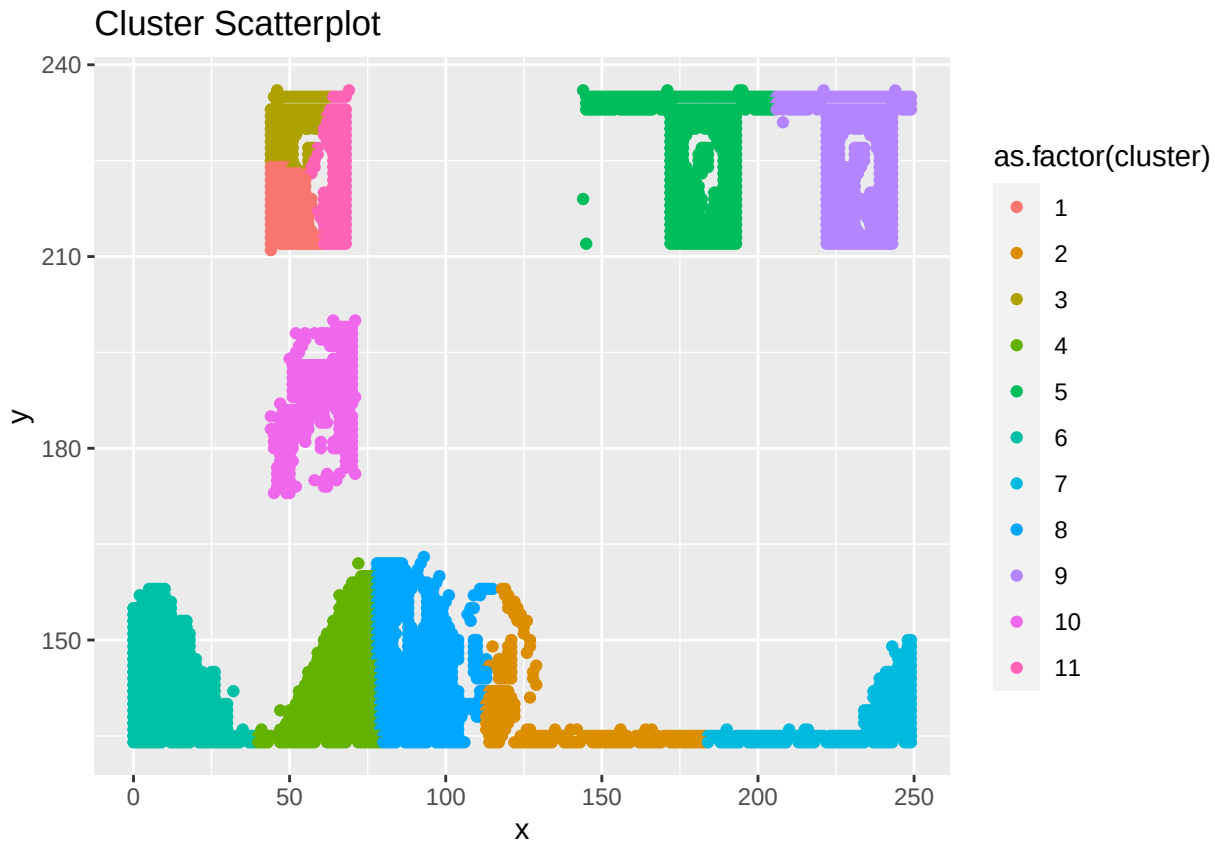

Cluster Scatterplot
240
210
180
150
y
0
50
100
150
200
250
x
as.factor(cluster)
1
2
3
4
5
6

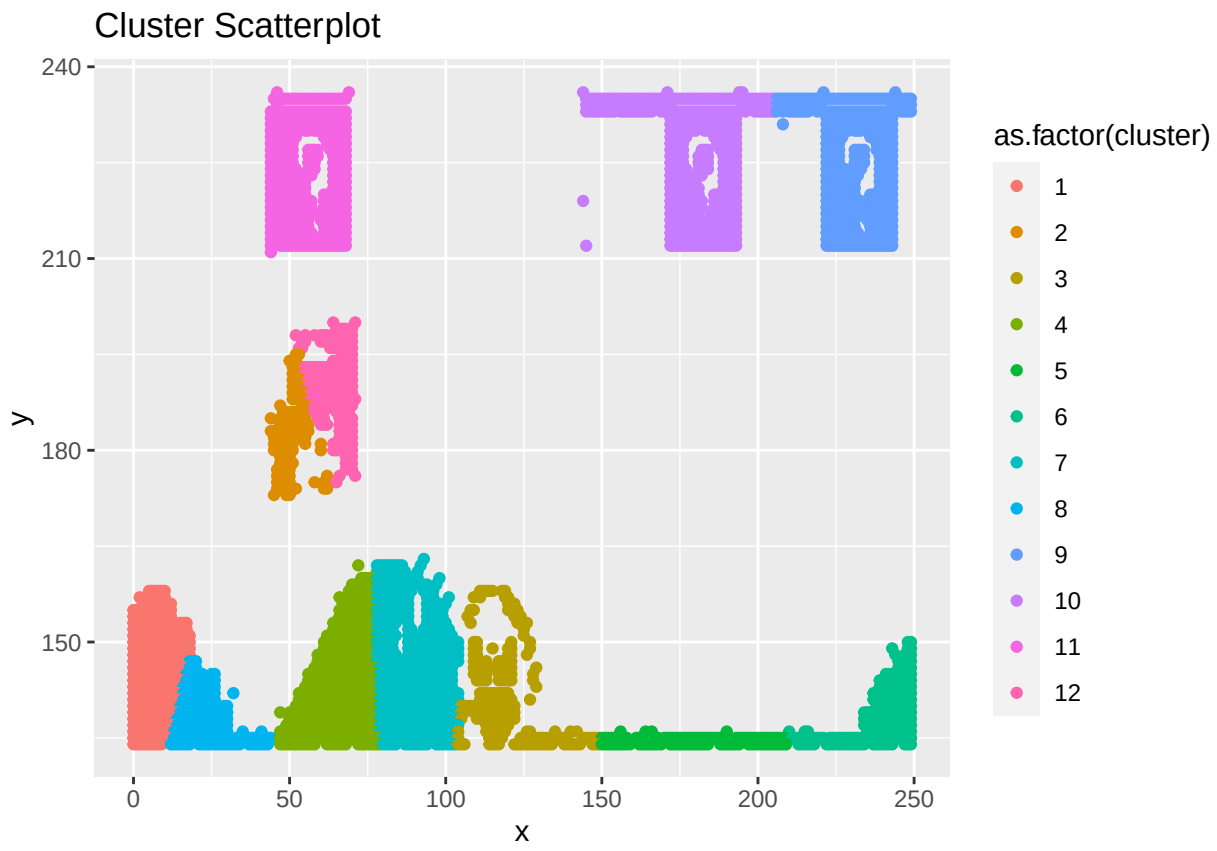

Cluster Scatterplot
240
210
180
150
y
0
50
100
150
200
250
x
as.factor(cluster)
1
2
3
4
5
6
7


Cluster Scatterplot
240
210
180
150
y
0
50
100
150
200
250
x
as.factor(cluster)
1
2
3
4
5
6
7
8


Cluster Scatterplot
240
210
180
150
y
0
50
100
150
200
250
x
as.factor(cluster)
1
2
3
4
5
6
7
8
9


Cluster Scatterplot
240
210
180
150
y
0
50
100
150
200
250
x
as.factor(cluster)
1
2
3
4
5
6
7
8
9
10

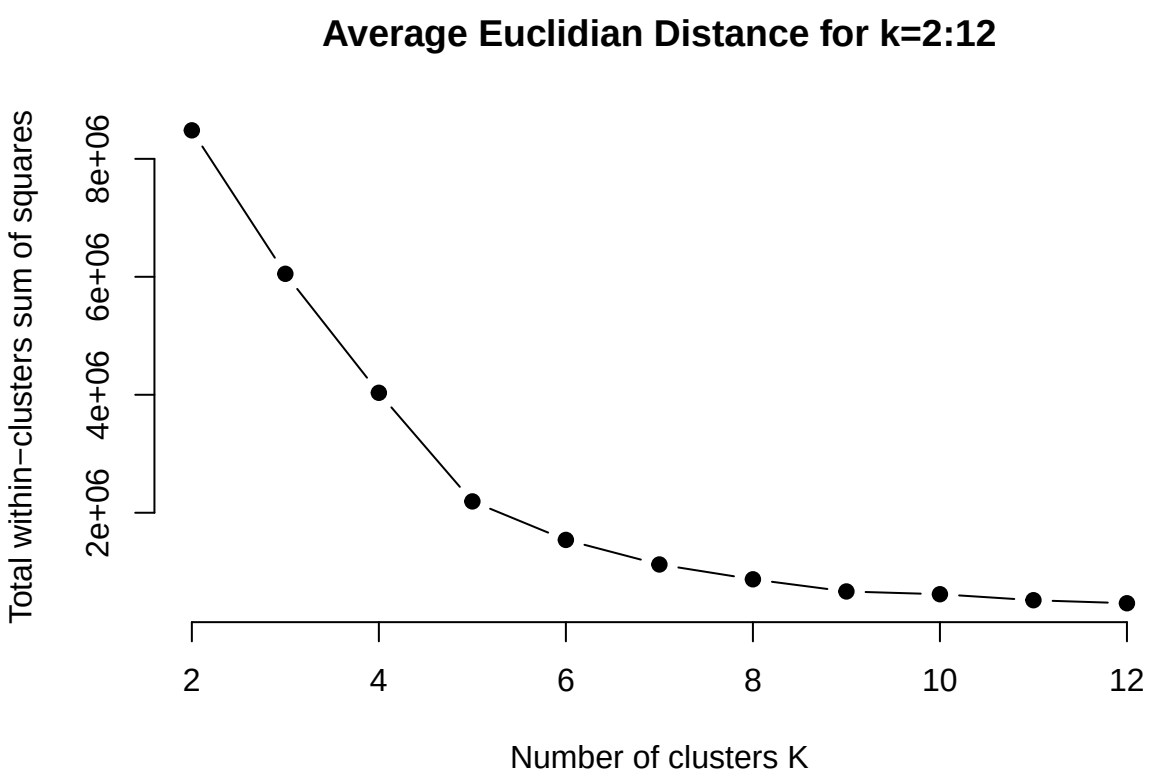

Cluster Scatterplot
240
210
180
150
y
0
50
100
150
200
250
x
as.factor(cluster)
1
2
3
4
5
6
7
8
9
10
11




```
avg_dist
```

```
##  [1]       NA 4322.313 4562.045 3366.235 2765.524 2280.247 2261.702 1806.729
##  [9] 1584.719 1743.035 1828.966 1522.473
```

```
#Calculate this average distance from the center of each cluster for each value of k and plot it as a l
set.seed(2345)
wss <- function(k) {
  kmeans(clustering_df, k, nstart = 10 )$tot.withinss
}

# Compute and plot wss for k = 1 to k = 15
k.values <- 2:12

# extract wss for 2-12 clusters
wss_values <- map_dbl(k.values, wss)

plot(k.values, wss_values,main="Average Euclidian Distance for k=2:12",
       type="b", pch = 19, frame = FALSE,
       xlab="Number of clusters K",
       ylab="Total within-clusters sum of squares")
```



What is the elbow point for this dataset?

The elbow point is k=5. At k=5, you stop getting as much accuracy per increase in k.