Assignment 11.2

Burden Joshua

2022-06-04

Introduction to Machine Learning

These assignments are here to provide you with an introduction to the "Data Science" use for these tools. This is your future. It may seem confusing and weird right now but it hopefully seems far less so than earlier in the semester. Attempt these homework assignments. You will not be graded on your answer but on your approach. This should be a, "Where am I on learning this stuff" check. If you can't get it done, please explain why.

Include all of your answers in a R Markdown report.

```
setwd('/Users/joshua/Documents/PERSONAL_GITHUB_REPOS/dsc520/')
binary_df <- read.csv('data/binary-classifier-data.csv')
trinary_df <- read.csv('data/trinary-classifier-data.csv')
summary(binary_df)
```

```
##      label              x                y
##  Min.   :0.000   Min.   : -5.20   Min.   : -4.019
##  1st Qu.:0.000   1st Qu.: 19.77   1st Qu.: 21.207
##  Median :0.000   Median : 41.76   Median : 44.632
##  Mean   :0.488   Mean   : 45.07   Mean   : 45.011
##  3rd Qu.:1.000   3rd Qu.: 66.39   3rd Qu.: 68.698
##  Max.   :1.000   Max.   :104.58   Max.   :106.896
```

```
summary(trinary_df)
```

```
##      label              x                y
##  Min.   :0.000   Min.   :-10.26   Min.   : -1.541
##  1st Qu.:0.000   1st Qu.: 31.15   1st Qu.: 35.906
##  Median :1.000   Median : 45.59   Median : 55.073
##  Mean   :1.037   Mean   : 48.86   Mean   : 55.282
##  3rd Qu.:2.000   3rd Qu.: 66.27   3rd Qu.: 77.403
##  Max.   :2.000   Max.   :108.56   Max.   :104.293
```

Regression algorithms are used to predict numeric quantity while classification algorithms predict categorical outcomes. A spam filter is an example use case for a classification algorithm. The input dataset is emails labeled as either spam (i.e. junk emails) or ham (i.e. good emails). The classification algorithm uses features extracted from the emails to learn which emails fall into which category.

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.

Note that in real-world datasets, your labels are usually not numbers, but text-based descriptions of the categories (e.g. spam or ham). In practice, you will encode categorical variables into numeric values. - Plot

the data from each dataset using a scatter plot.

```
ggplot(binary_df,aes(x=x,y=y,color=label)) + geom_point() + ggtitle("Binary Classifier")
```
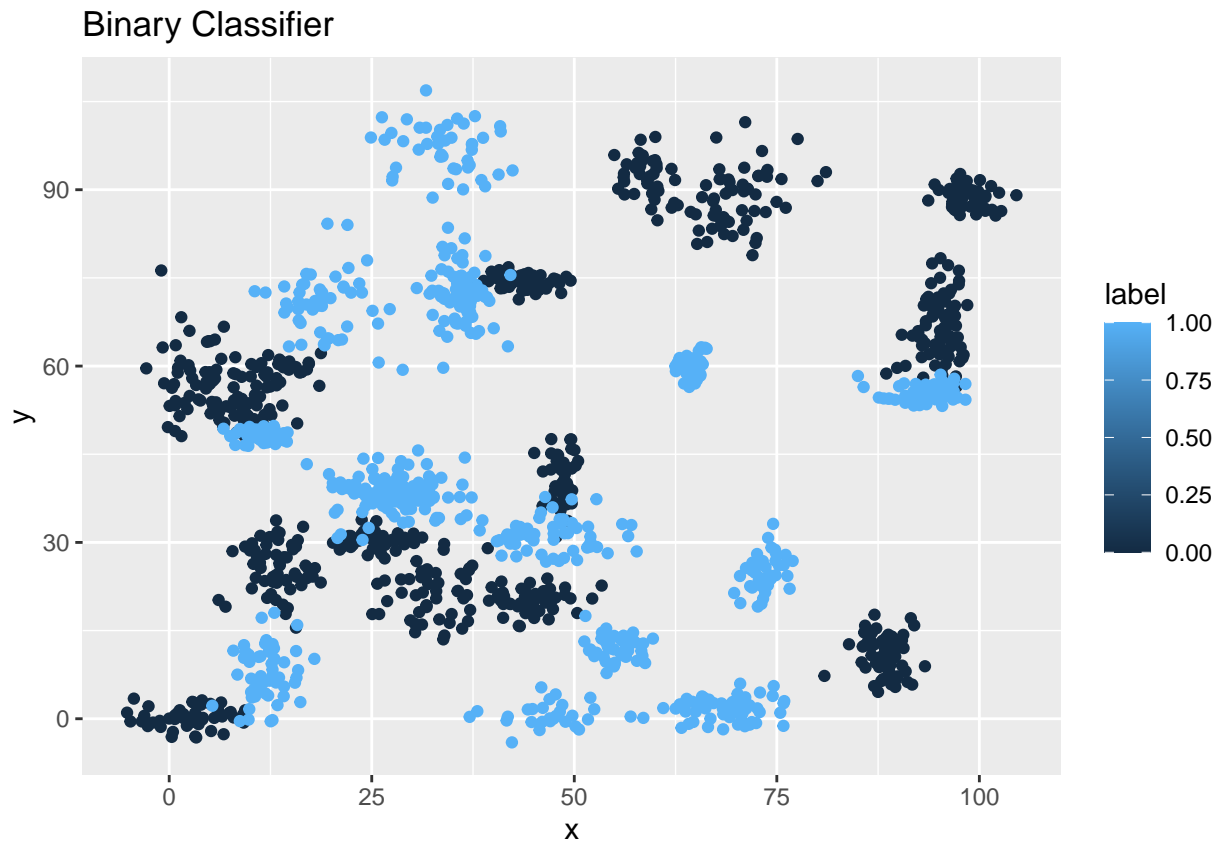


```
ggplot(trinary_df,aes(x=x,y=y,color=label)) + geom_point() + ggtitle("Trinary Classifier")
```

Trinary Classifier
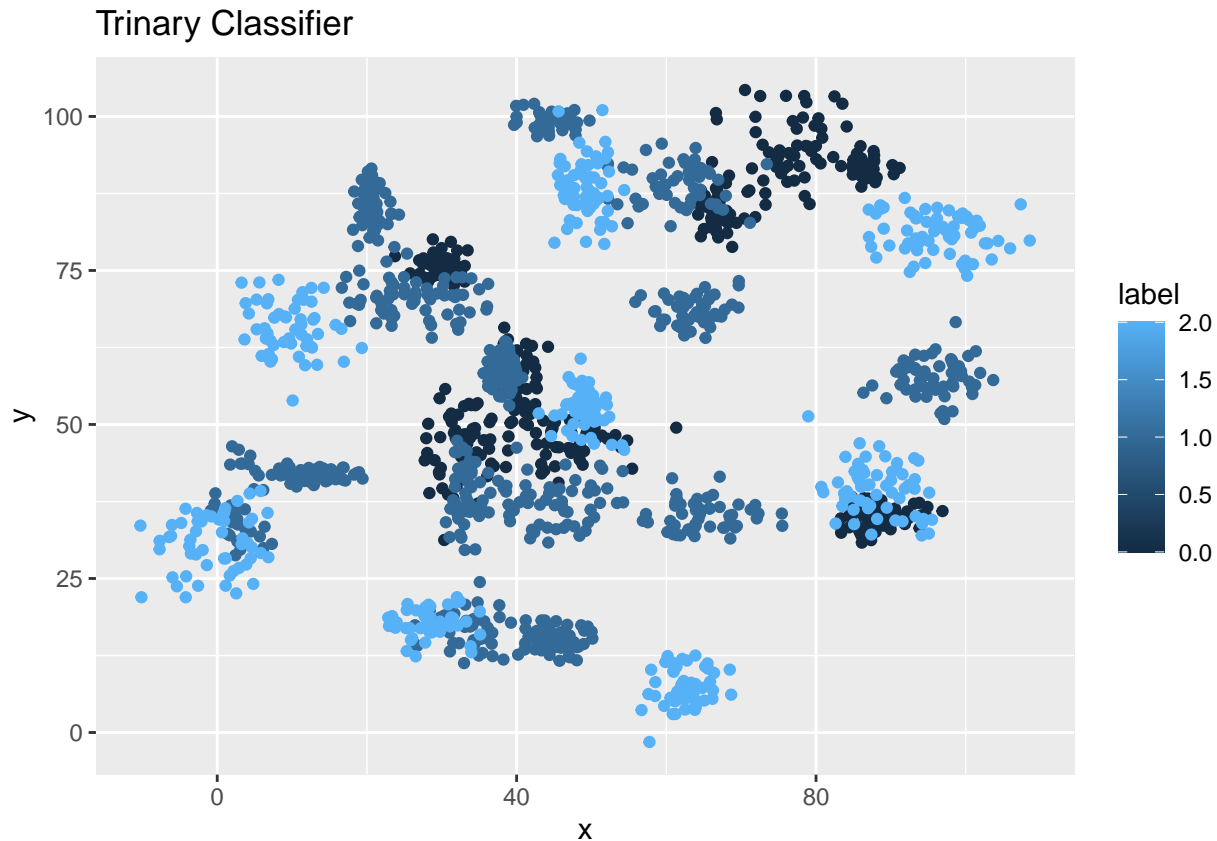
- The k nearest neighbors algorithm categorizes an input value by looking at the labels for the k nearest points and assigning a category based on the most common label. In this problem, you will determine which points are nearest by calculating the Euclidean distance between two points. As a refresher, the Euclidean distance between two points:

$$p_1 = (x_1, y_1)$$
$$p_2 = (x_2, y_2)$$
$$d = \sqrt{(x_1 - x_2)^2 + (y_1 - y_2)^2}$$

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.

```
k <- c(3,5,10,15,20,25)
binary_accuracy <- NULL
for(i in 1:6)
{
  cat("k=NN 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]
}
```

```
## k=NN Binary Classisfier: 3
##
##    Cell Contents
## |-------------------------|
## |                       N |
## |           N / Row Total |
## |           N / Col Total |
## |         N / Table Total |
## |-------------------------|
##
##
## Total Observations in Table:  1498
##
##
##               | binary_knn
## binary_df$label |         0 |         1 | Row Total |
## ---------------|-----------|-----------|-----------|
##             0 |       757 |        10 |       767 |
##               |     0.987 |     0.013 |     0.512 |
##               |     0.986 |     0.014 |           |
##               |     0.505 |     0.007 |           |
## ---------------|-----------|-----------|-----------|
##             1 |        11 |       720 |       731 |
##               |     0.015 |     0.985 |     0.488 |
##               |     0.014 |     0.986 |           |
##               |     0.007 |     0.481 |           |
## ---------------|-----------|-----------|-----------|
##    Column Total |       768 |       730 |      1498 |
##               |     0.513 |     0.487 |           |
## ---------------|-----------|-----------|-----------|
##
##
## k=NN Binary Classisfier: 5
##
##    Cell Contents
## |-------------------------|
## |                       N |
## |           N / Row Total |
## |           N / Col Total |
## |         N / Table Total |
## |-------------------------|
##
##
## Total Observations in Table:  1498
##
##
##               | binary_knn
## binary_df$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 |            |
## ----------------|-----------|-----------|-----------|
##
##
## k=NN 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 |        752 |        15 |        767 |
##                |      0.980 |      0.020 |      0.512 |
##                |      0.979 |      0.021 |            |
##                |      0.502 |      0.010 |            |
## ----------------|-----------|-----------|-----------|
##              1 |         16 |        715 |        731 |
##                |      0.022 |      0.978 |      0.488 |
##                |      0.021 |      0.979 |            |
##                |      0.011 |      0.477 |            |
## ----------------|-----------|-----------|-----------|
##    Column Total |        768 |        730 |       1498 |
##                |      0.513 |      0.487 |            |
## ----------------|-----------|-----------|-----------|
##
##
## k=NN Binary Classisfier: 15
##
##     Cell Contents
## |-------------------------|
## |                       N |
## |           N / Row Total |
## |           N / Col Total |
```

```
## |            N / Table Total |
## |-------------------------|
##
##
## Total Observations in Table:  1498
##
##
##               | binary_knn
## binary_df$label |          0 |          1 | Row Total |
## ----------------|-----------|-----------|-----------|
##              0 |       752 |        15 |       767 |
##                |     0.980 |     0.020 |     0.512 |
##                |     0.977 |     0.021 |           |
##                |     0.502 |     0.010 |           |
## ----------------|-----------|-----------|-----------|
##              1 |        18 |       713 |       731 |
##                |     0.025 |     0.975 |     0.488 |
##                |     0.023 |     0.979 |           |
##                |     0.012 |     0.476 |           |
## ----------------|-----------|-----------|-----------|
##    Column Total |       770 |       728 |      1498 |
##                |     0.514 |     0.486 |           |
## ----------------|-----------|-----------|-----------|
##
##
## k=NN Binary Classisfier: 20
##
##    Cell Contents
## |-------------------------|
## |                       N |
## |          N / Row Total |
## |          N / Col Total |
## |        N / Table Total |
## |-------------------------|
##
##
## Total Observations in Table:  1498
##
##
##               | binary_knn
## binary_df$label |          0 |          1 | Row Total |
## ----------------|-----------|-----------|-----------|
##              0 |       750 |        17 |       767 |
##                |     0.978 |     0.022 |     0.512 |
##                |     0.975 |     0.023 |           |
##                |     0.501 |     0.011 |           |
## ----------------|-----------|-----------|-----------|
##              1 |        19 |       712 |       731 |
##                |     0.026 |     0.974 |     0.488 |
##                |     0.025 |     0.977 |           |
##                |     0.013 |     0.475 |           |
## ----------------|-----------|-----------|-----------|
##    Column Total |       769 |       729 |      1498 |
##                |     0.513 |     0.487 |           |
```

```
## ----------------|-----------|-----------|-----------|
##
##
## k=NN 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 |           |
## ----------------|-----------|-----------|-----------|
##
##
```

```r
trinary_accuracy <- NULL
for(i in 1:6)
{
  cat("k=NN 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]
}
```

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

```
## 
## Total Observations in Table:  1568
## 
## 
##                  | trinary_knn 
## trinary_df$label |         0 |         1 |         2 | Row Total | 
## -----------------|-----------|-----------|-----------|-----------|
##                0 |       386 |         8 |         0 |       394 | 
##                  |     0.980 |     0.020 |     0.000 |     0.251 | 
##                  |     0.965 |     0.011 |     0.000 |           | 
##                  |     0.246 |     0.005 |     0.000 |           | 
## -----------------|-----------|-----------|-----------|-----------|
##                1 |         9 |       707 |         6 |       722 | 
##                  |     0.012 |     0.979 |     0.008 |     0.460 | 
##                  |     0.022 |     0.975 |     0.014 |           | 
##                  |     0.006 |     0.451 |     0.004 |           | 
## -----------------|-----------|-----------|-----------|-----------|
##                2 |         5 |        10 |       437 |       452 | 
##                  |     0.011 |     0.022 |     0.967 |     0.288 | 
##                  |     0.013 |     0.014 |     0.986 |           | 
##                  |     0.003 |     0.006 |     0.279 |           | 
## -----------------|-----------|-----------|-----------|-----------|
##     Column Total |       400 |       725 |       443 |      1568 | 
##                  |     0.255 |     0.462 |     0.283 |           | 
## -----------------|-----------|-----------|-----------|-----------|
## 
## 
## k=NN Trinary Classisfier: 5
## 
##    Cell Contents
## |-------------------------|
## |                       N |
## |           N / Row Total |
## |           N / Col Total |
## |         N / Table Total |
## |-------------------------|
## 
## 
## Total Observations in Table:  1568
## 
## 
##                  | trinary_knn 
## trinary_df$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 |            |
## ----------------|-----------|-----------|-----------|-----------|
##
##
## k=NN 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 |        368 |         24 |          2 |        394 |
##                   |      0.934 |      0.061 |      0.005 |      0.251 |
##                   |      0.927 |      0.033 |      0.005 |            |
##                   |      0.235 |      0.015 |      0.001 |            |
## ----------------|-----------|-----------|-----------|-----------|
##                 1 |         16 |        686 |         20 |        722 |
##                   |      0.022 |      0.950 |      0.028 |      0.460 |
##                   |      0.040 |      0.944 |      0.045 |            |
##                   |      0.010 |      0.438 |      0.013 |            |
## ----------------|-----------|-----------|-----------|-----------|
##                 2 |         13 |         17 |        422 |        452 |
##                   |      0.029 |      0.038 |      0.934 |      0.288 |
##                   |      0.033 |      0.023 |      0.950 |            |
##                   |      0.008 |      0.011 |      0.269 |            |
## ----------------|-----------|-----------|-----------|-----------|
##     Column Total |        397 |        727 |        444 |       1568 |
##                   |      0.253 |      0.464 |      0.283 |            |
## ----------------|-----------|-----------|-----------|-----------|
##
##
## k=NN Trinary Classisfier: 15
##
##      Cell Contents
## |-----------------------|
## |                     N |
## |         N / Row Total |
## |         N / Col Total |
```

```
## |         N / Table Total |
## |-------------------------|
##
##
## Total Observations in Table:  1568
##
##
##               | trinary_knn
## trinary_df$label |         0 |         1 |         2 | Row Total |
## ----------------|-----------|-----------|-----------|-----------|
##              0 |       364 |        26 |         4 |       394 |
##                |     0.924 |     0.066 |     0.010 |     0.251 |
##                |     0.899 |     0.036 |     0.009 |           |
##                |     0.232 |     0.017 |     0.003 |           |
## ----------------|-----------|-----------|-----------|-----------|
##              1 |        21 |       679 |        22 |       722 |
##                |     0.029 |     0.940 |     0.030 |     0.460 |
##                |     0.052 |     0.937 |     0.050 |           |
##                |     0.013 |     0.433 |     0.014 |           |
## ----------------|-----------|-----------|-----------|-----------|
##              2 |        20 |        20 |       412 |       452 |
##                |     0.044 |     0.044 |     0.912 |     0.288 |
##                |     0.049 |     0.028 |     0.941 |           |
##                |     0.013 |     0.013 |     0.263 |           |
## ----------------|-----------|-----------|-----------|-----------|
##    Column Total |       405 |       725 |       438 |      1568 |
##                |     0.258 |     0.462 |     0.279 |           |
## ----------------|-----------|-----------|-----------|-----------|
##
##
## k=NN Trinary Classisfier: 20
##
##    Cell Contents
## |-------------------------|
## |                       N |
## |         N / Row Total |
## |         N / Col Total |
## |       N / Table Total |
## |-------------------------|
##
##
## Total Observations in Table:  1568
##
##
##               | trinary_knn
## trinary_df$label |         0 |         1 |         2 | Row Total |
## ----------------|-----------|-----------|-----------|-----------|
##              0 |       354 |        36 |         4 |       394 |
##                |     0.898 |     0.091 |     0.010 |     0.251 |
##                |     0.881 |     0.049 |     0.009 |           |
##                |     0.226 |     0.023 |     0.003 |           |
## ----------------|-----------|-----------|-----------|-----------|
##              1 |        23 |       674 |        25 |       722 |
##                |     0.032 |     0.934 |     0.035 |     0.460 |
```

```
##                 |     0.057 |     0.918 |     0.058 |           |
##                 |     0.015 |     0.430 |     0.016 |           |
## ----------------|-----------|-----------|-----------|-----------|
##               2 |        25 |        24 |       403 |       452 |
##                 |     0.055 |     0.053 |     0.892 |     0.288 |
##                 |     0.062 |     0.033 |     0.933 |           |
##                 |     0.016 |     0.015 |     0.257 |           |
## ----------------|-----------|-----------|-----------|-----------|
##    Column Total |       402 |       734 |       432 |      1568 |
##                 |     0.256 |     0.468 |     0.276 |           |
## ----------------|-----------|-----------|-----------|-----------|
##
##
## k=NN 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 |
##                 |     0.873 |     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 |        26 |        21 |       405 |       452 |
##                 |     0.058 |     0.046 |     0.896 |     0.288 |
##                 |     0.066 |     0.028 |     0.927 |           |
##                 |     0.017 |     0.013 |     0.258 |           |
## ----------------|-----------|-----------|-----------|-----------|
##    Column Total |       393 |       738 |       437 |      1568 |
##                 |     0.251 |     0.471 |     0.279 |           |
## ----------------|-----------|-----------|-----------|-----------|
##
##
```
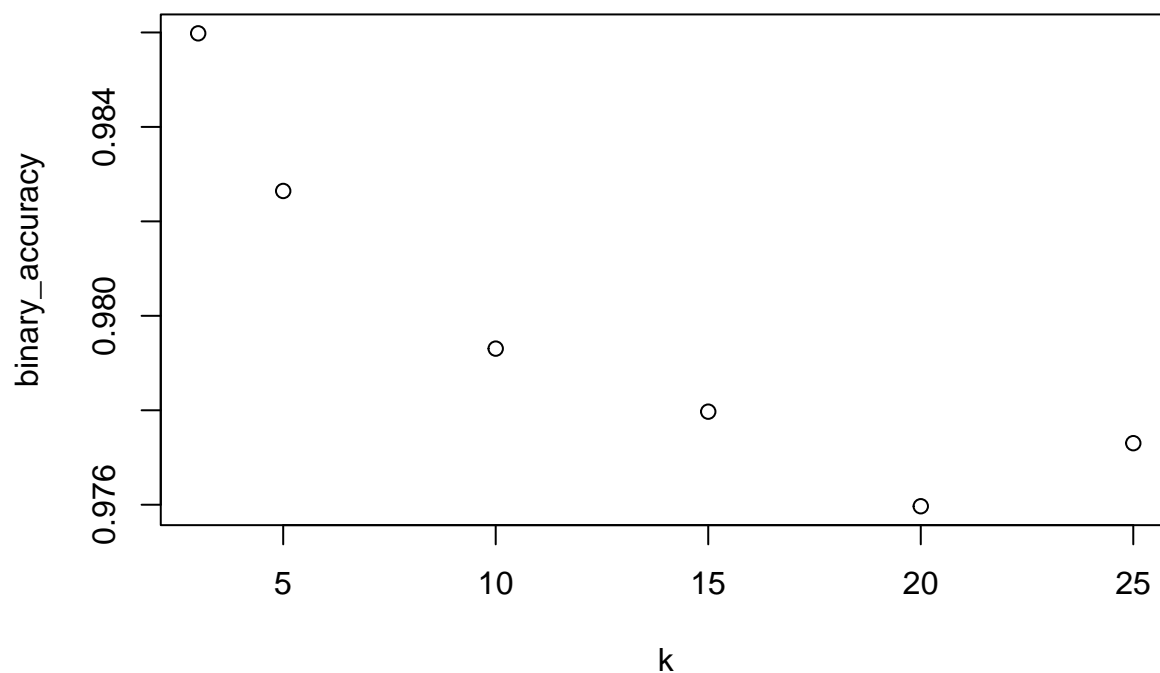
```
plot(k,binary_accuracy,main="Binary k-NN Classsification Accuracy")
```
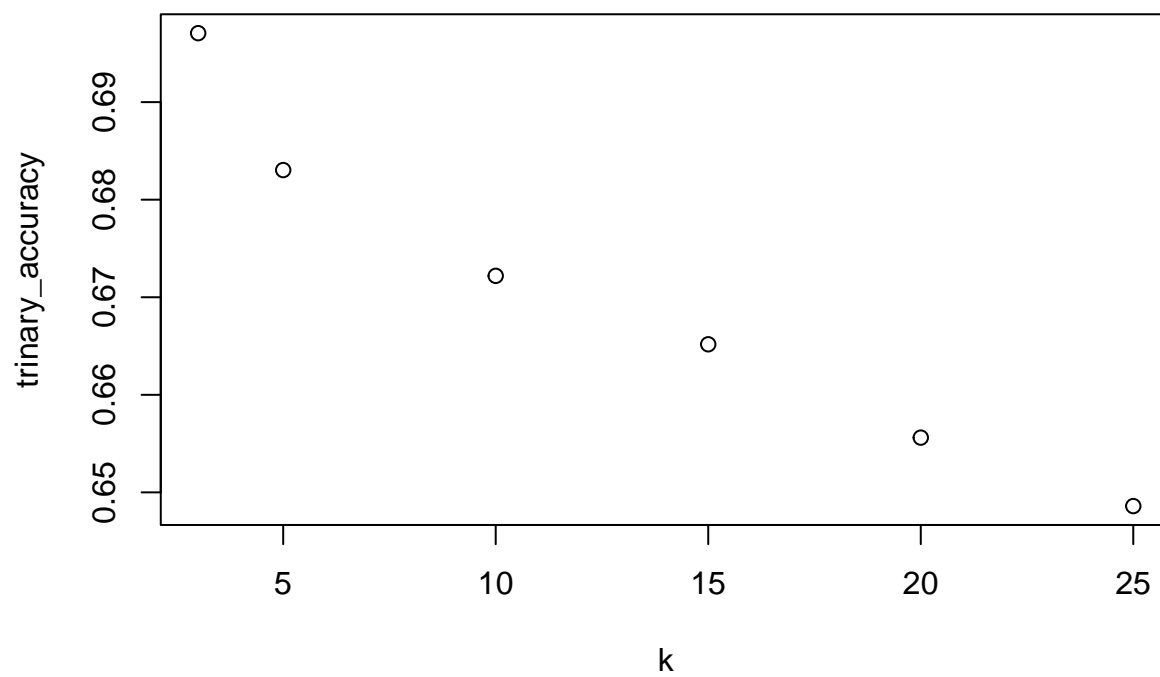
Binary k−NN Classsification Accuracy


```
plot(k,trinary_accuracy,main="Trinary k-NN Classsification Accuracy")
```

Trinary k−NN Classsification Accuracy

Clustering

These assignments are here to provide you with an introduction to the "Data Science" use for these tools. This is your future. It may seem confusing and weird right now but it hopefully seems far less so than earlier in the semester. Attempt these homework assignments. You will not be graded on your answer but on your approach. This should be a, "Where am I on learning this stuff" check. If you can't get it done, please explain why.

Labeled data is not always available. For these types of datasets, you can use unsupervised algorithms to extract structure. The k-means clustering algorithm and the k nearest neighbor algorithm both use the Euclidean distance between points to group data points. The difference is the k-means clustering algorithm does not use labeled data.

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.

```
library(stats)
setwd('/Users/joshua/Documents/PERSONAL_GITHUB_REPOS/dsc520/')
cluster_df <- read.csv('data/clustering-data.csv')
summary(cluster_df)
```

```
##        x                y
##  Min.   :  0.0   Min.   :134.0
##  1st Qu.: 56.0   1st Qu.:141.0
##  Median : 82.0   Median :154.0
##  Mean   :109.6   Mean   :175.7
##  3rd Qu.:180.0   3rd Qu.:218.0
##  Max.   :249.0   Max.   :236.0
```

Plot the dataset using a scatter plot.

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

Clustering 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.

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

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

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

  print(ggplot(cluster_df,aes(x=x,y=y,color=as.factor(cluster)))+geom_point()+labs(title = "Cluster Sca
  print(avg_dist[i])
}
```
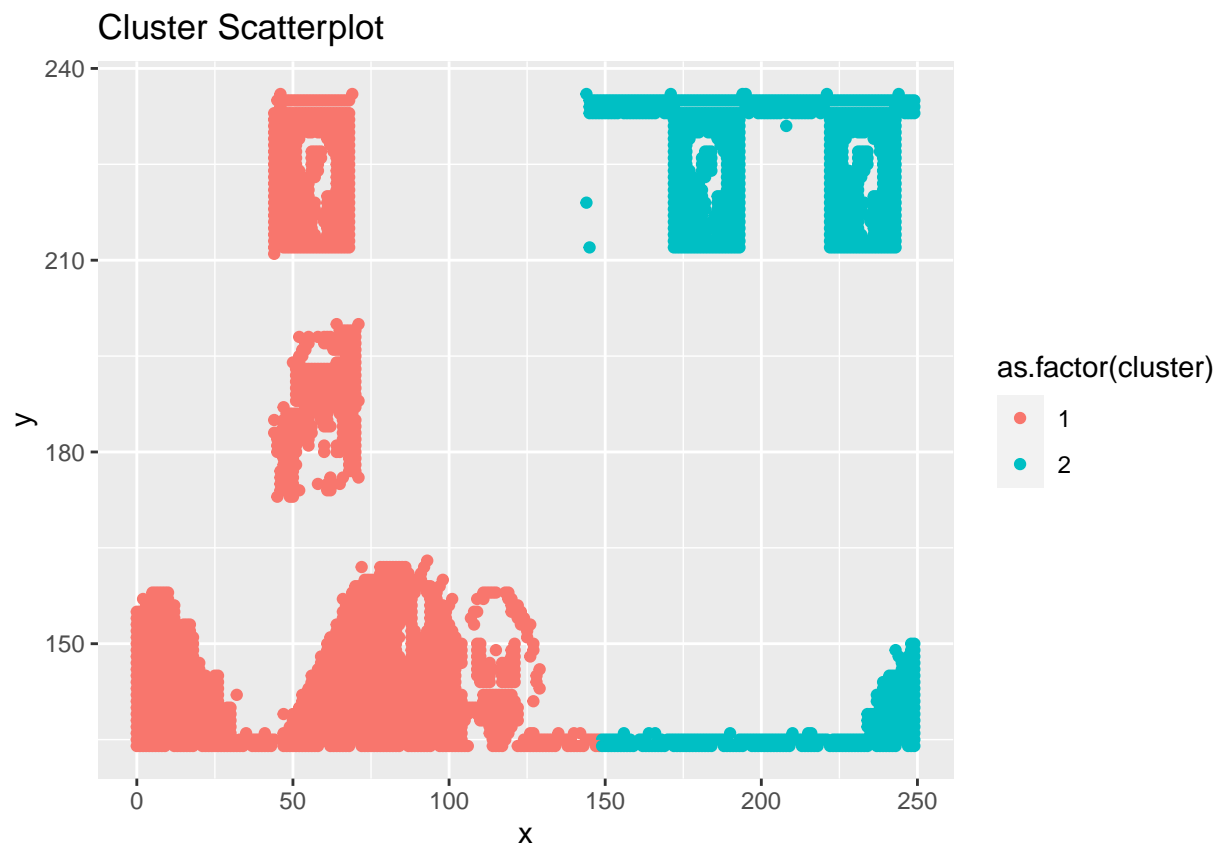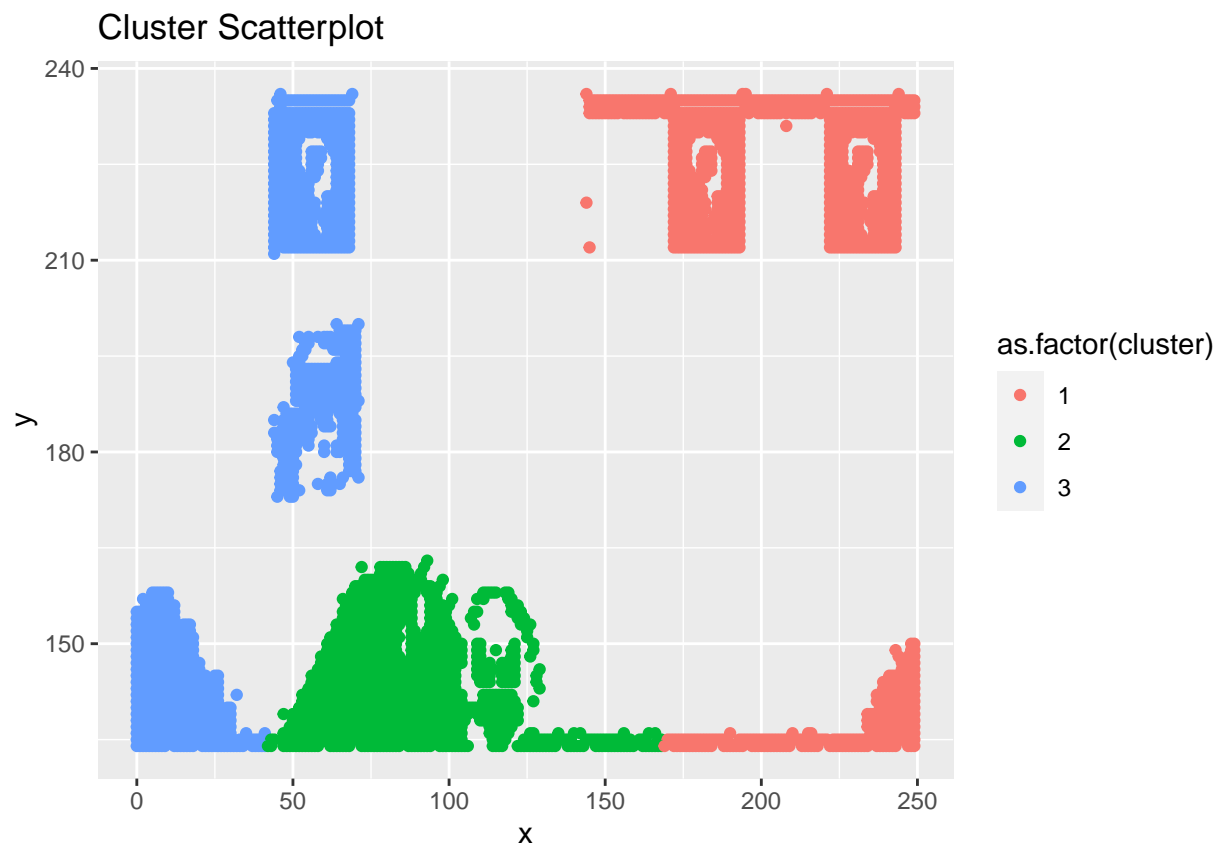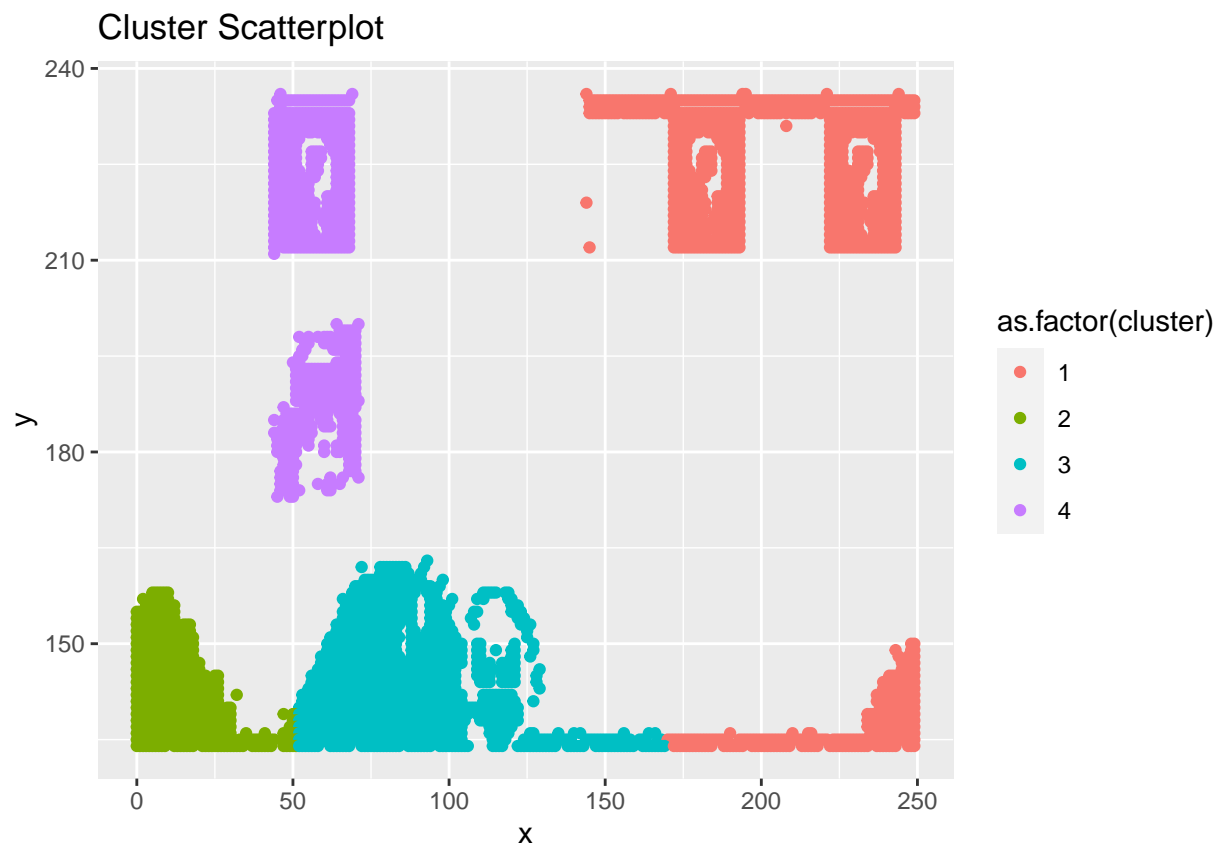
Cluster Scatterplot

```
## [1] 4322.313
```

Cluster Scatterplot

```
## [1] 4562.045
```

Cluster Scatterplot


```
## [1] 3366.235
```

Cluster Scatterplot

[1] 2765.524

Cluster Scatterplot

```
## [1] 2280.247
```

Cluster Scatterplot

[1] 2261.702

Cluster Scatterplot

[1] 1806.729

Cluster Scatterplot

```
## [1] 1584.719
```

Cluster Scatterplot

[1] 1743.035

Cluster Scatterplot


```
## [1] 1828.966
```

Cluster Scatterplot
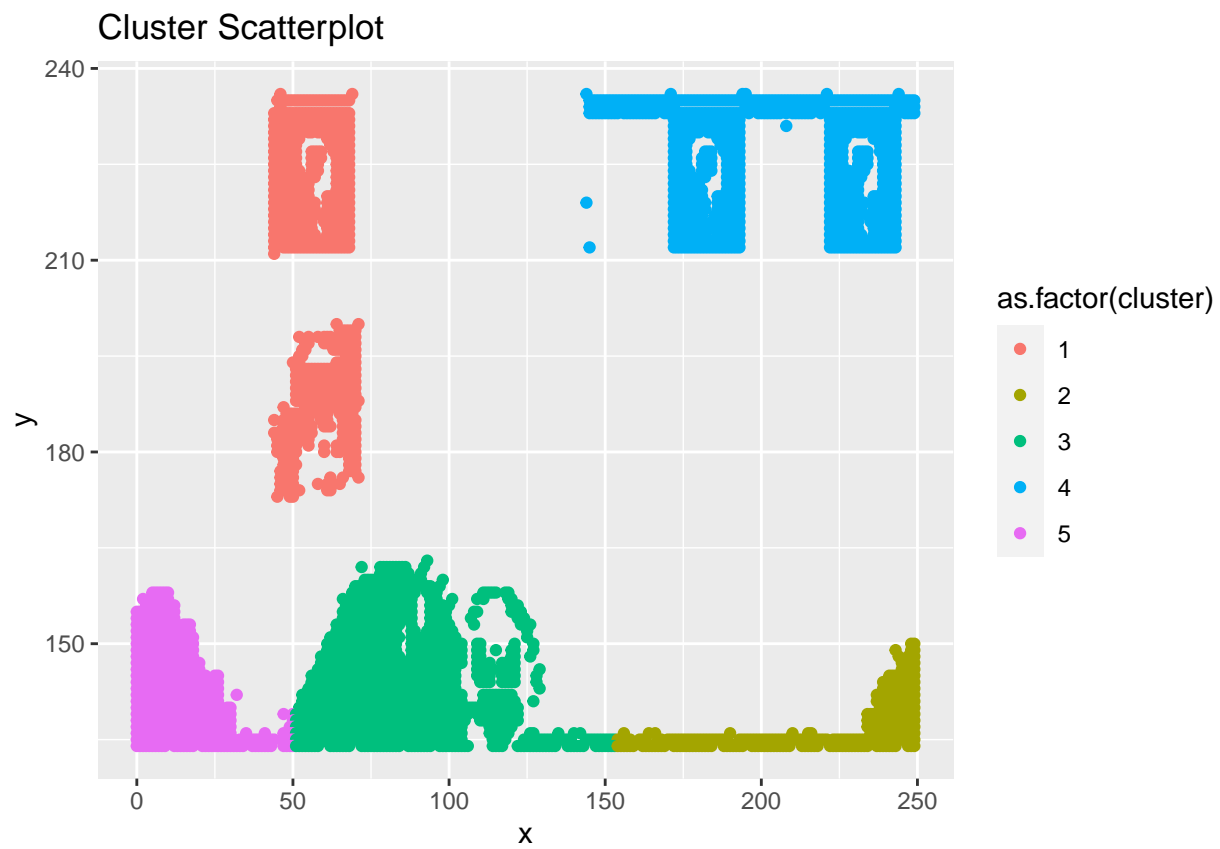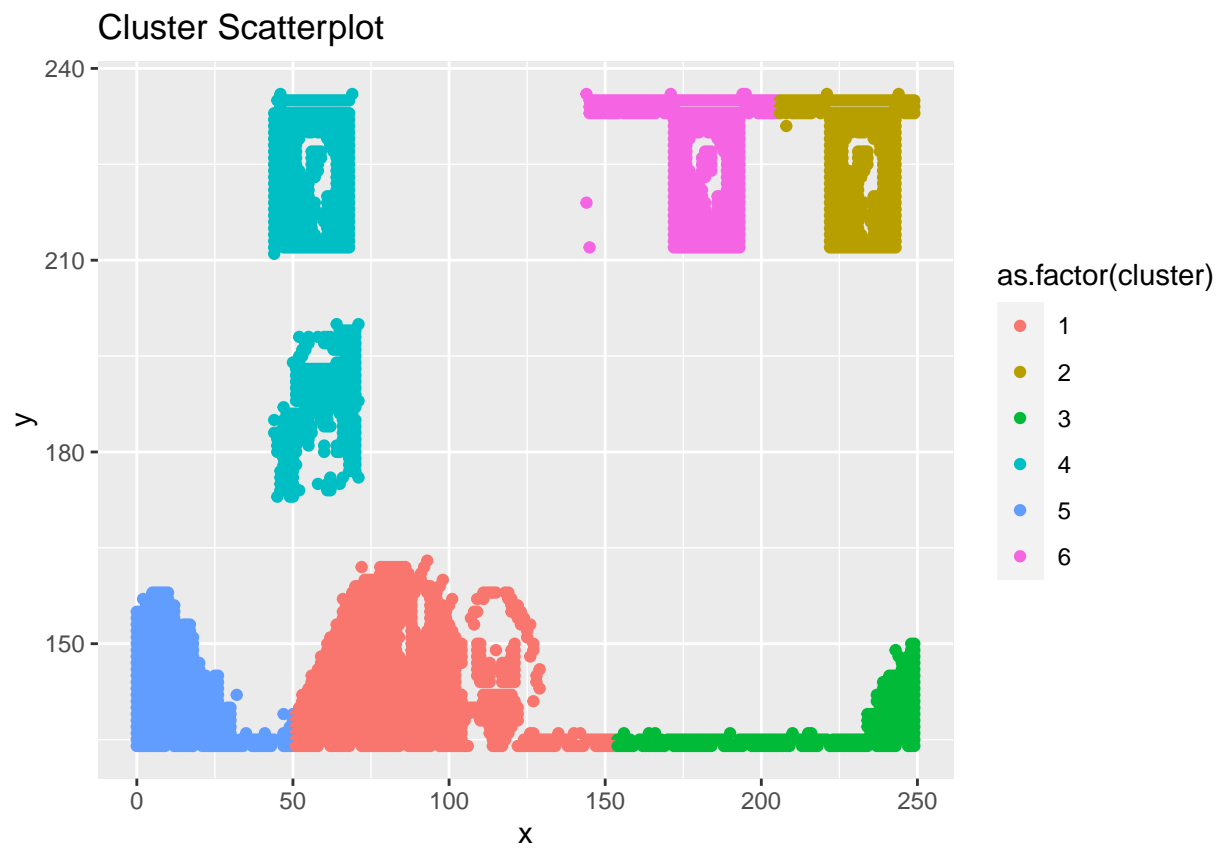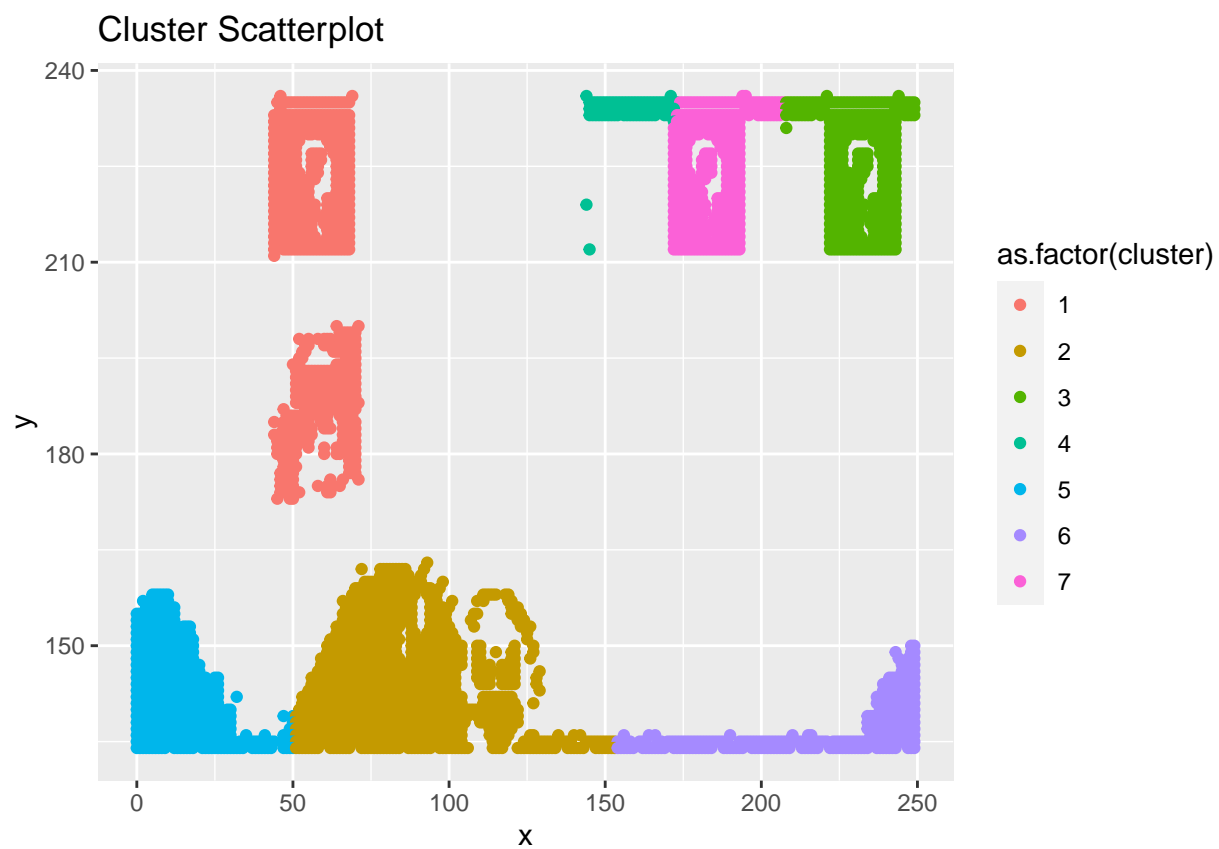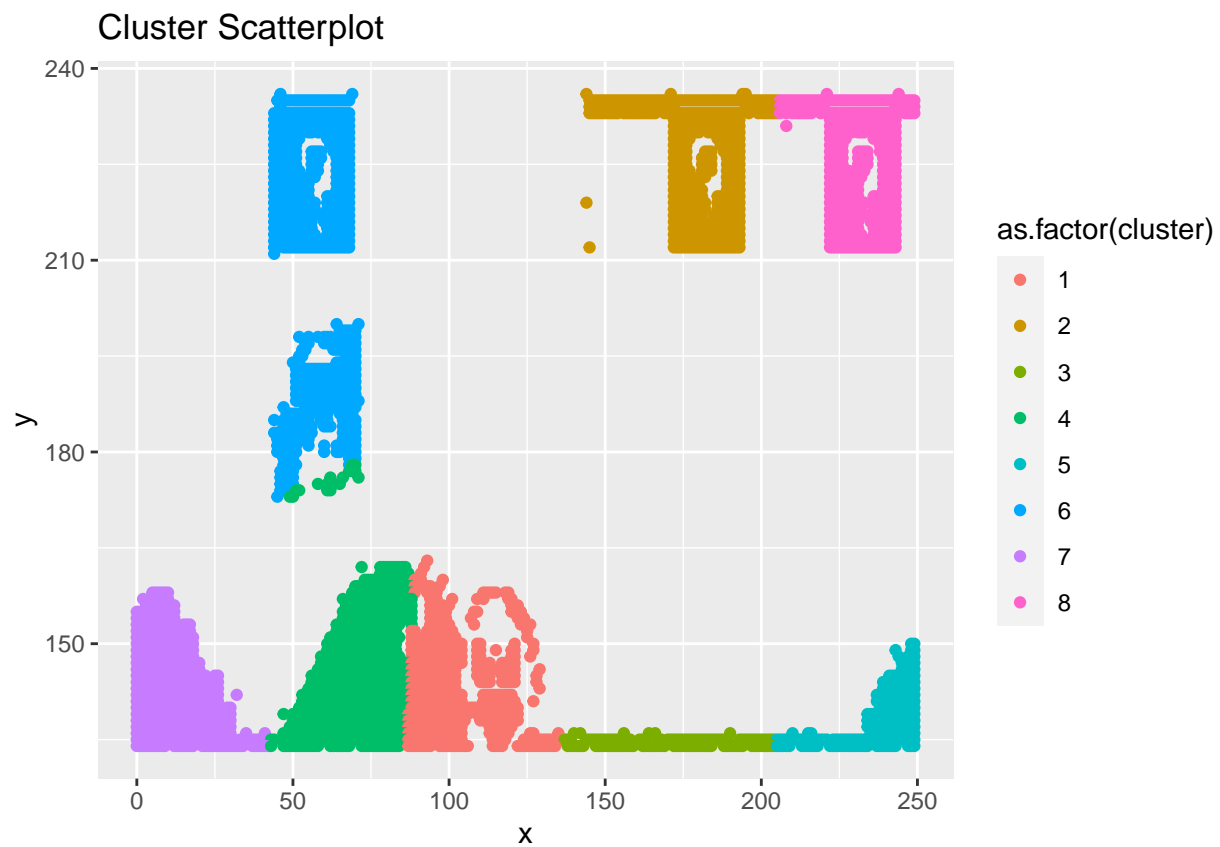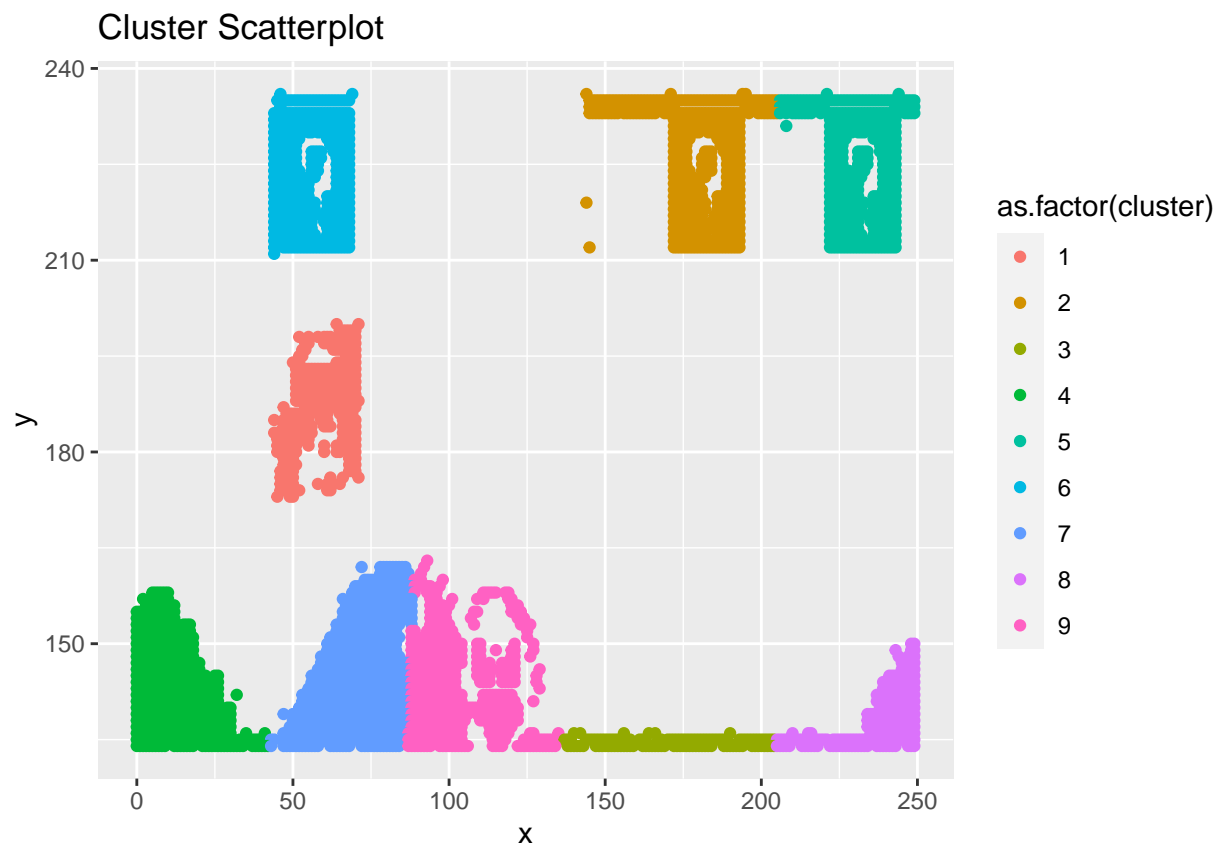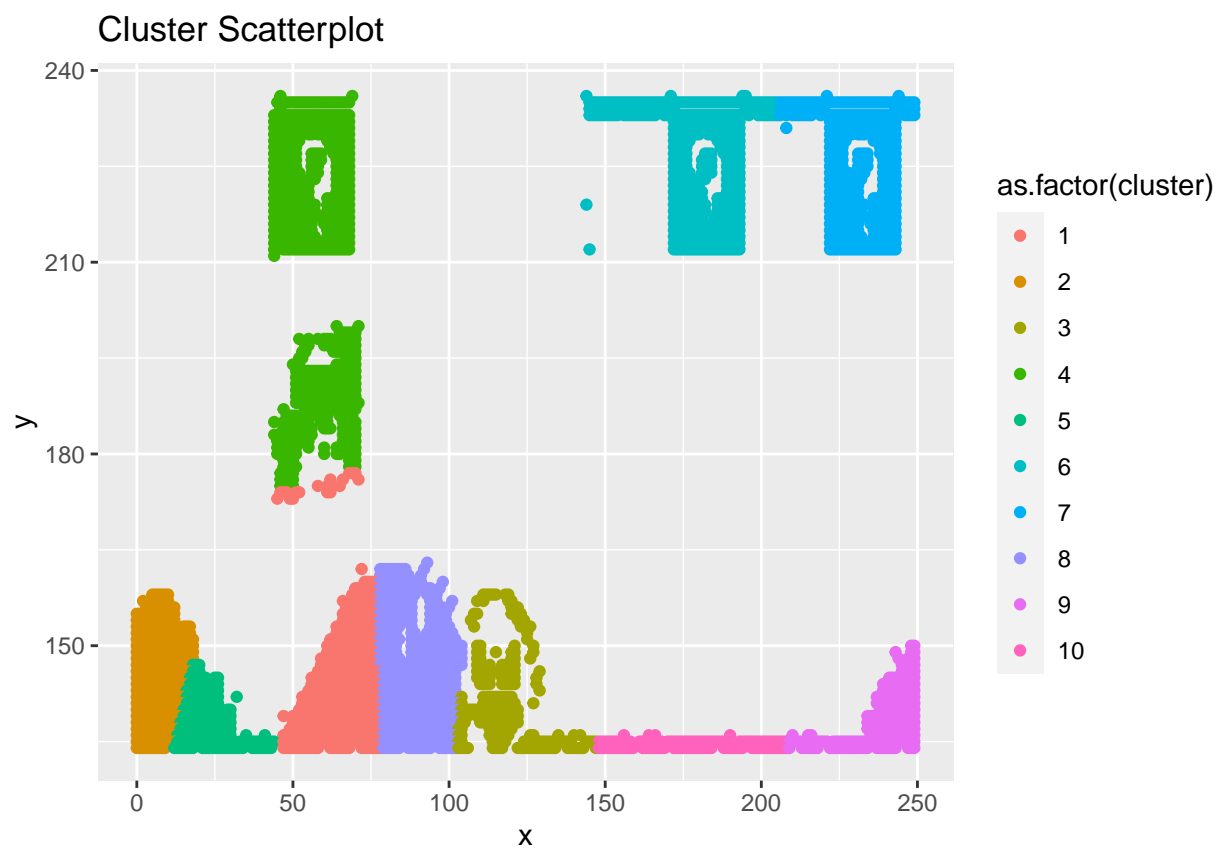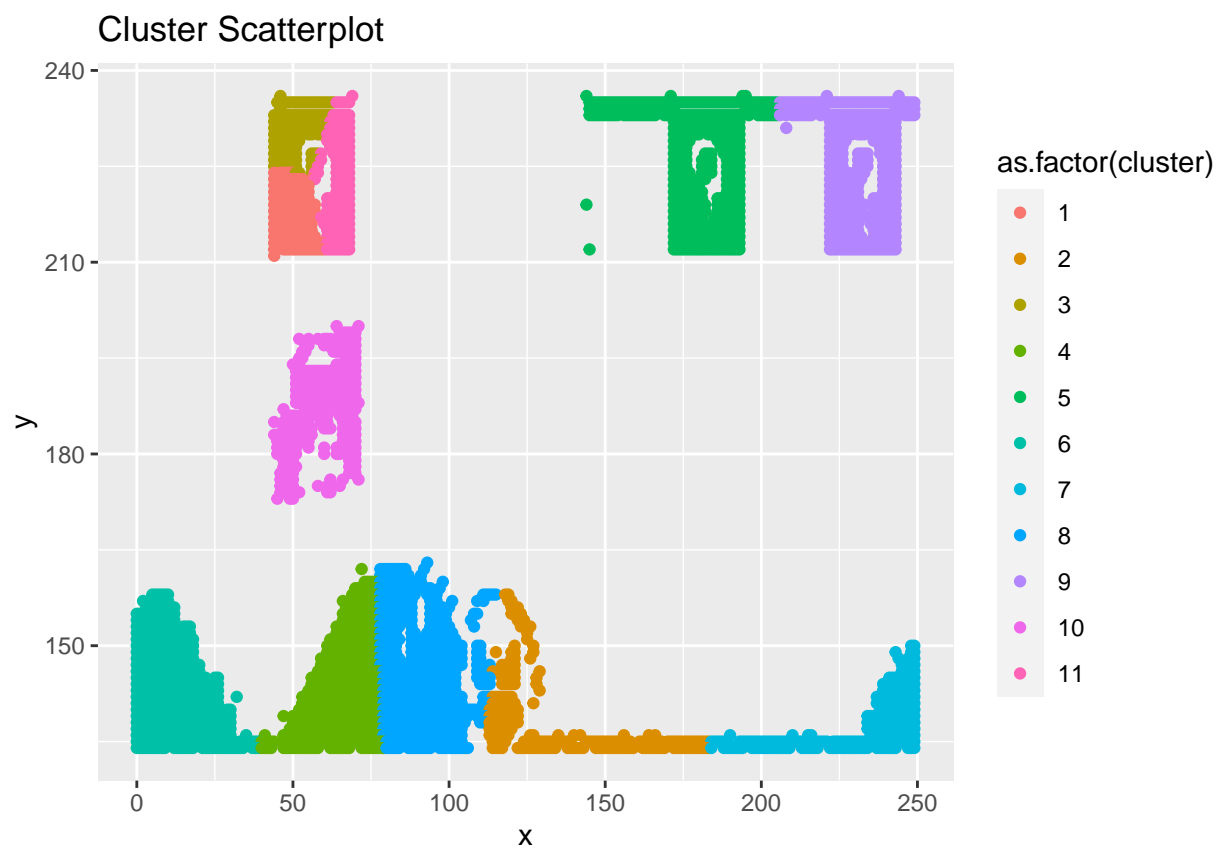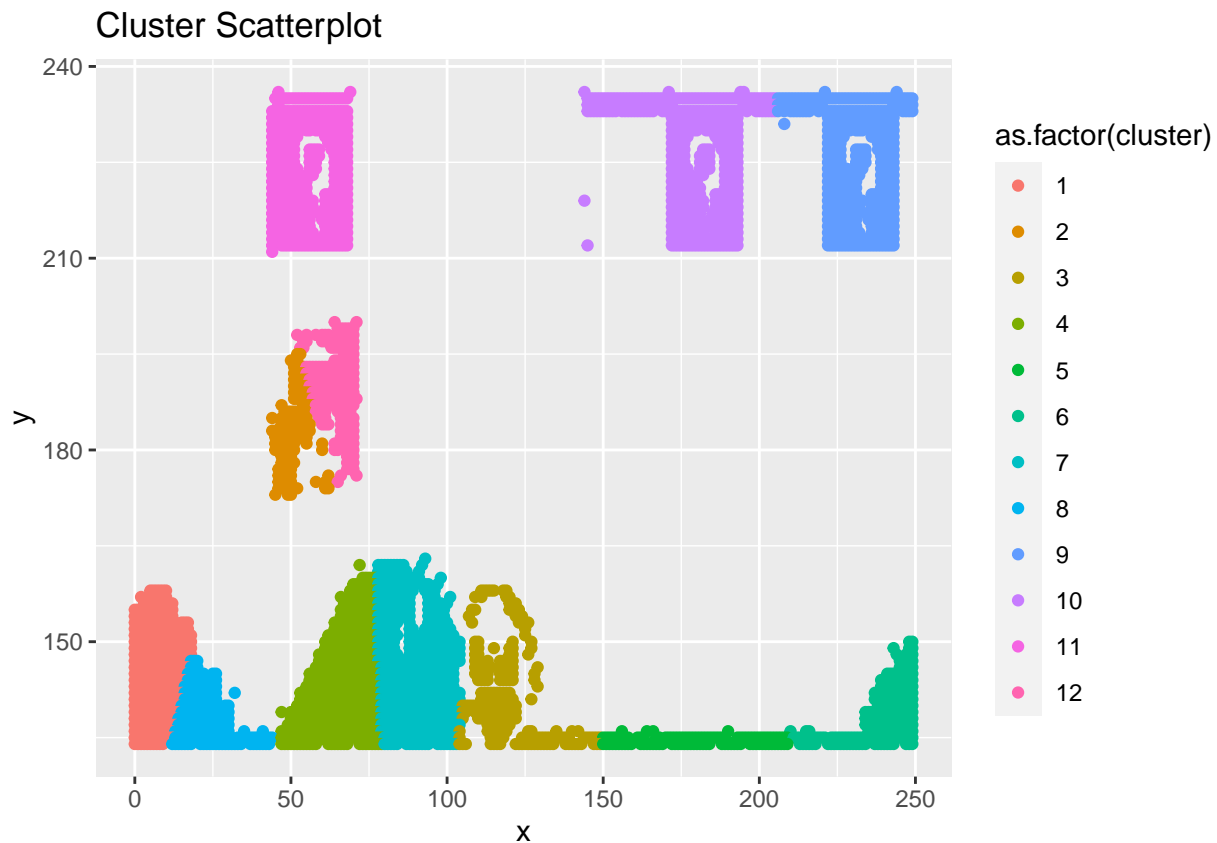
```
## [1] 1522.473
```

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. 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.

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.

```r
plot(2:12,avg_dist[2:12], main="Average Euclidian Distance for k=2:12")
```

Average Euclidian Distance for k=2:12
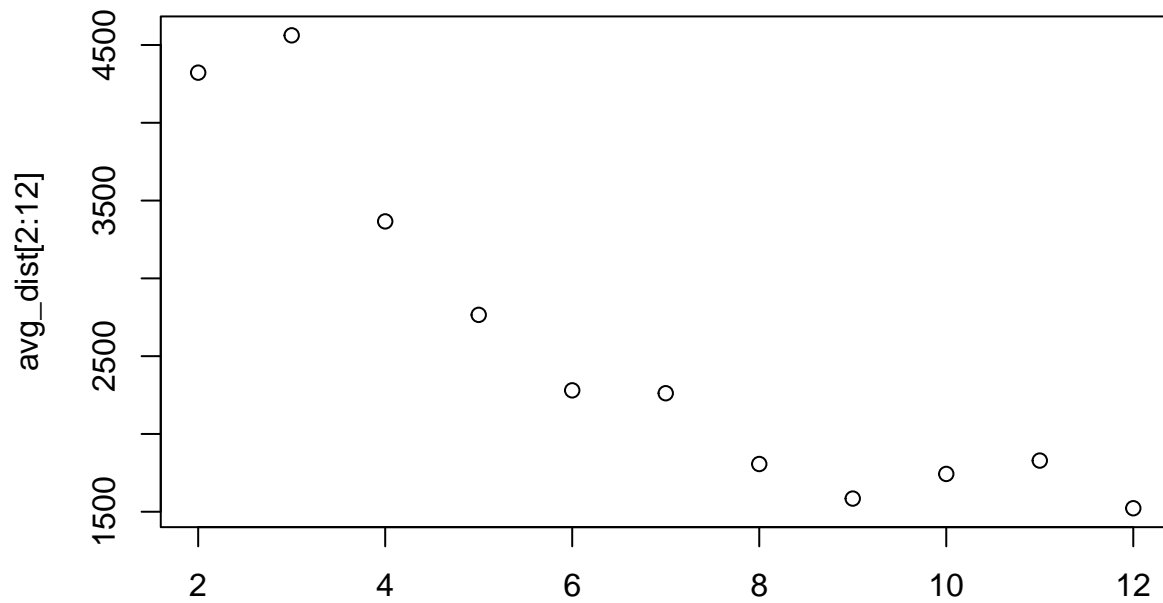
2:12   ###

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 occurs at K=9