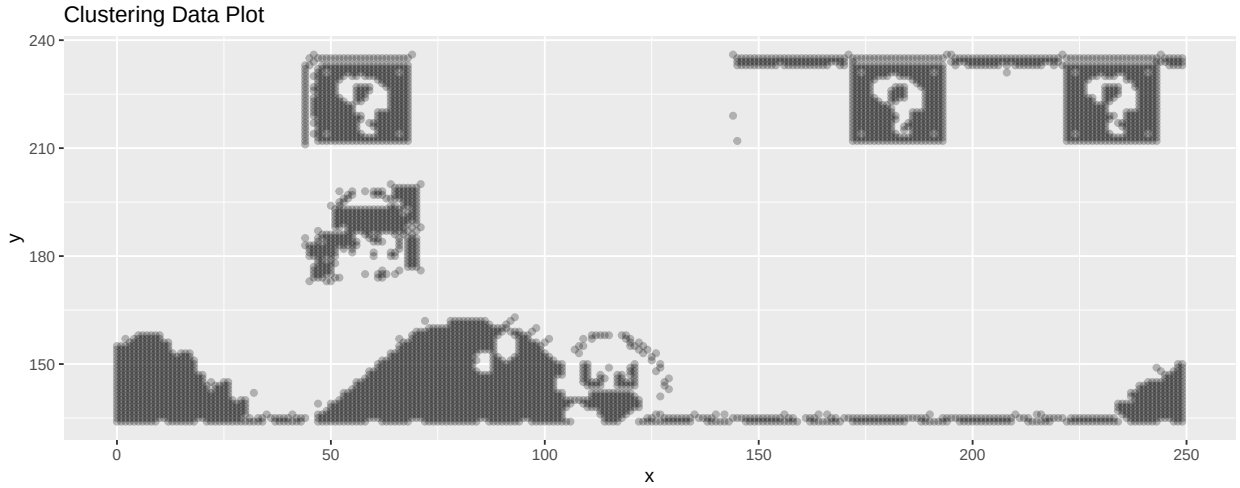
Assignment_9.2_HillZach

Zach Hill

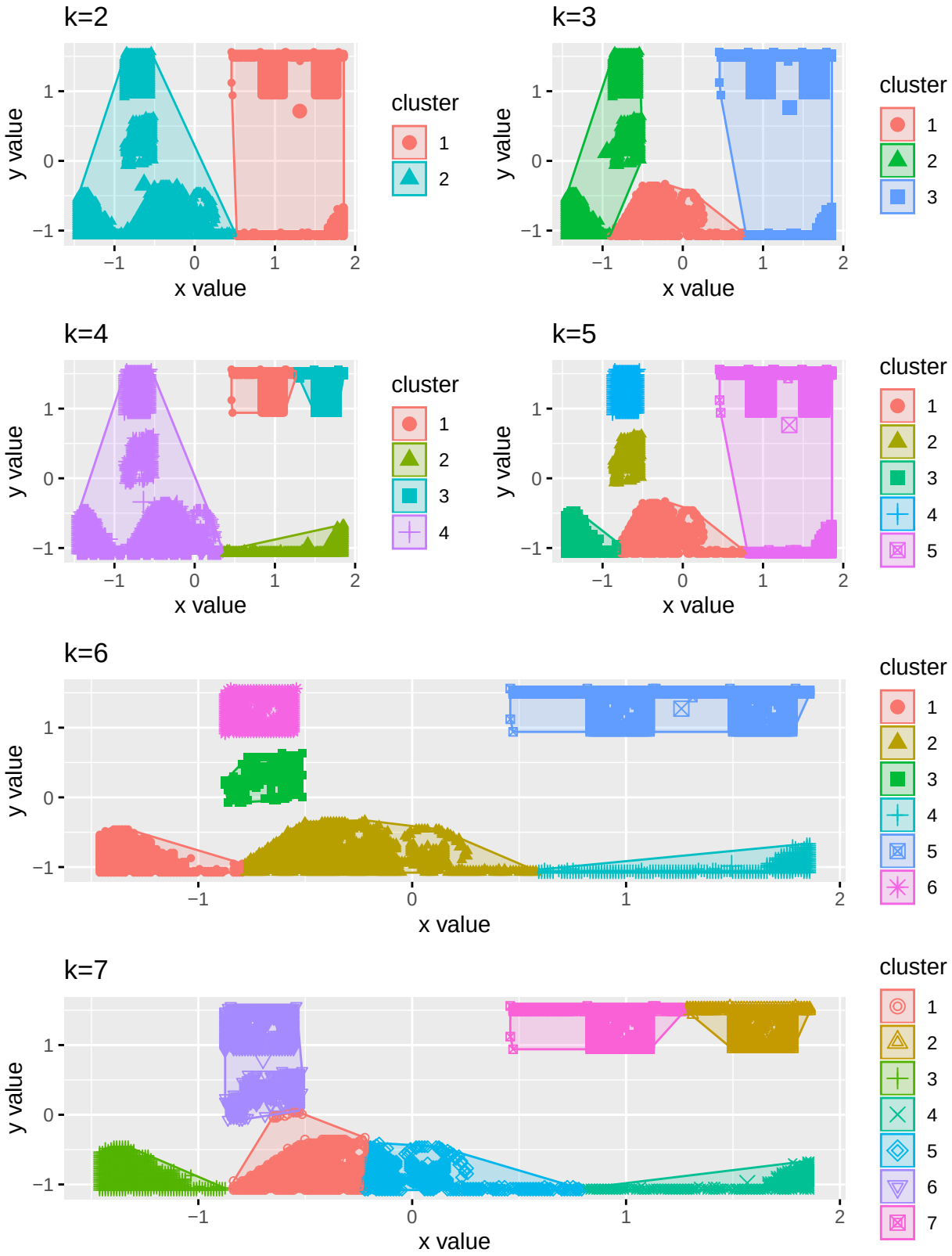
May 12, 2019

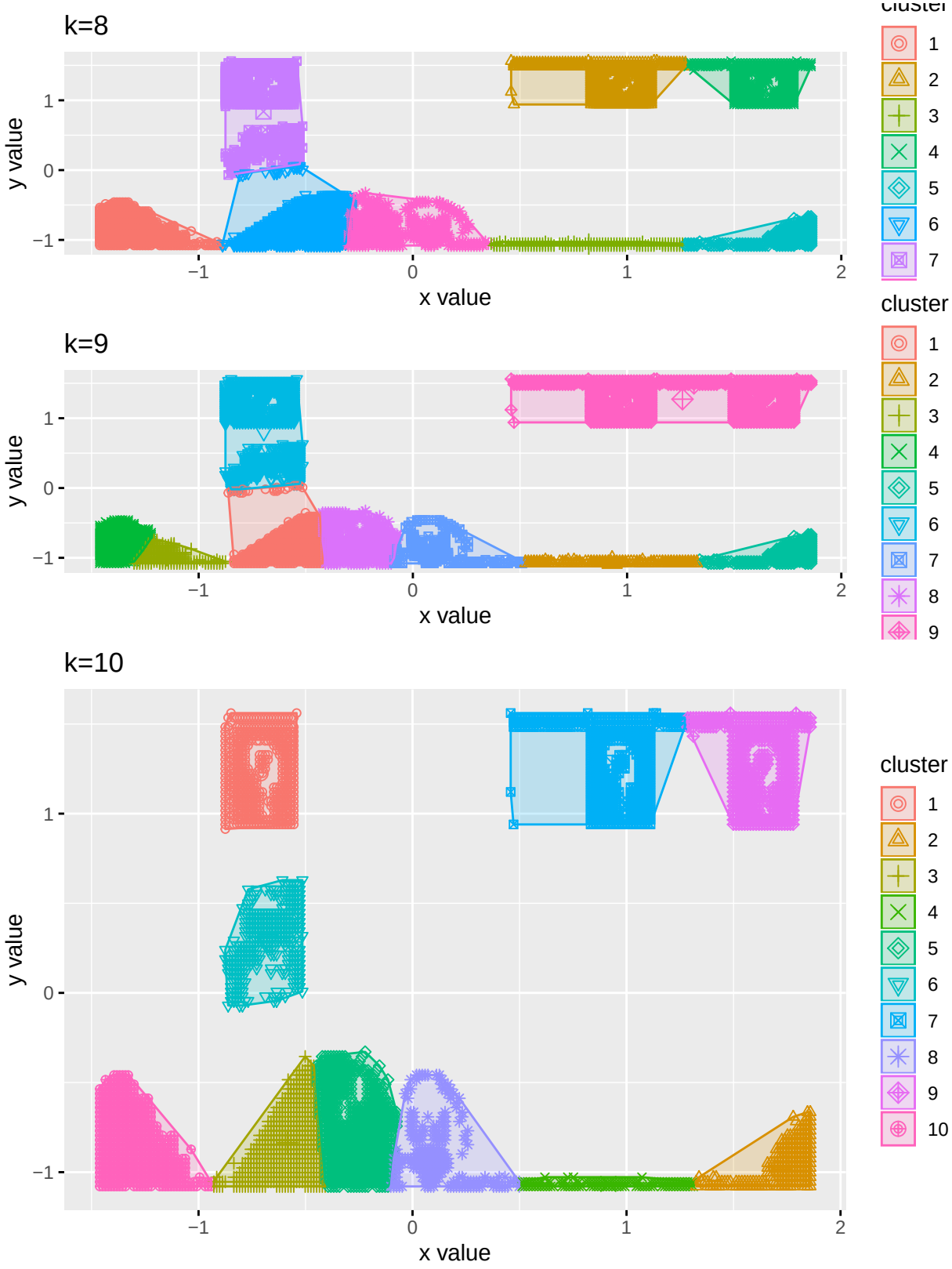
A: Plotted data from original datasets

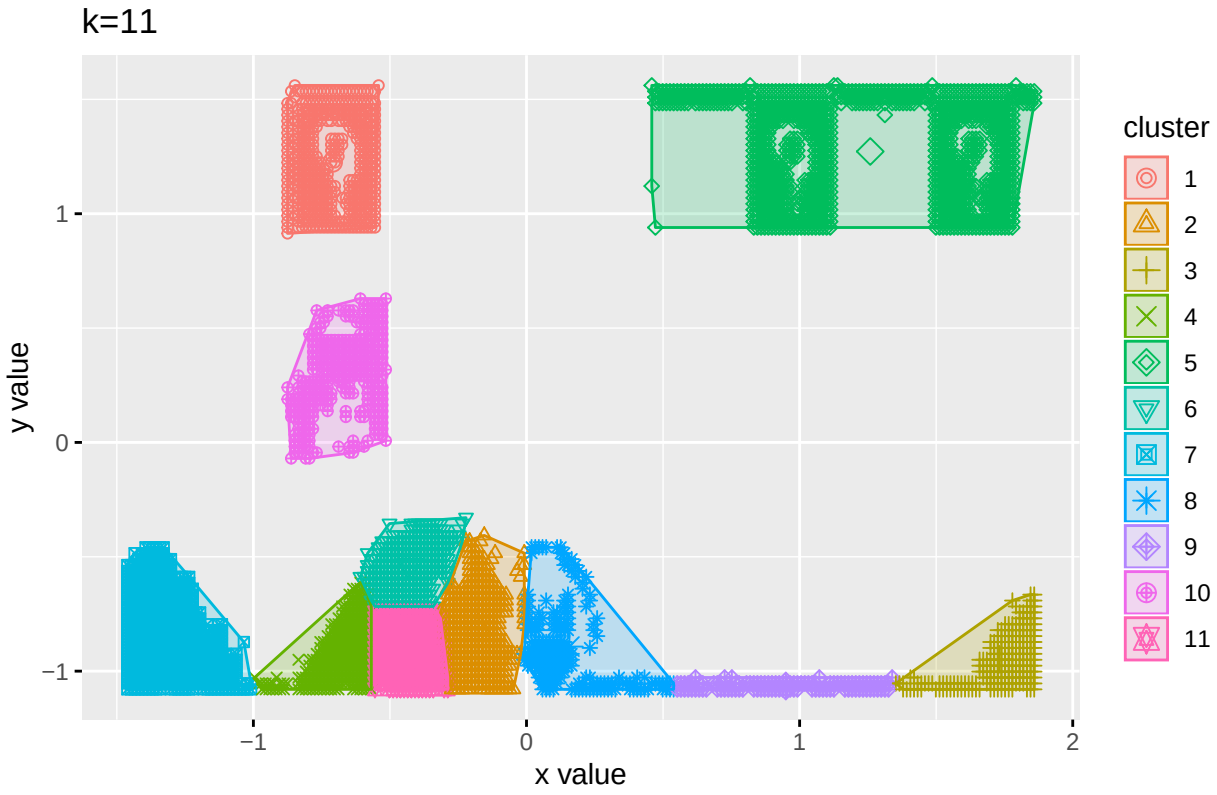


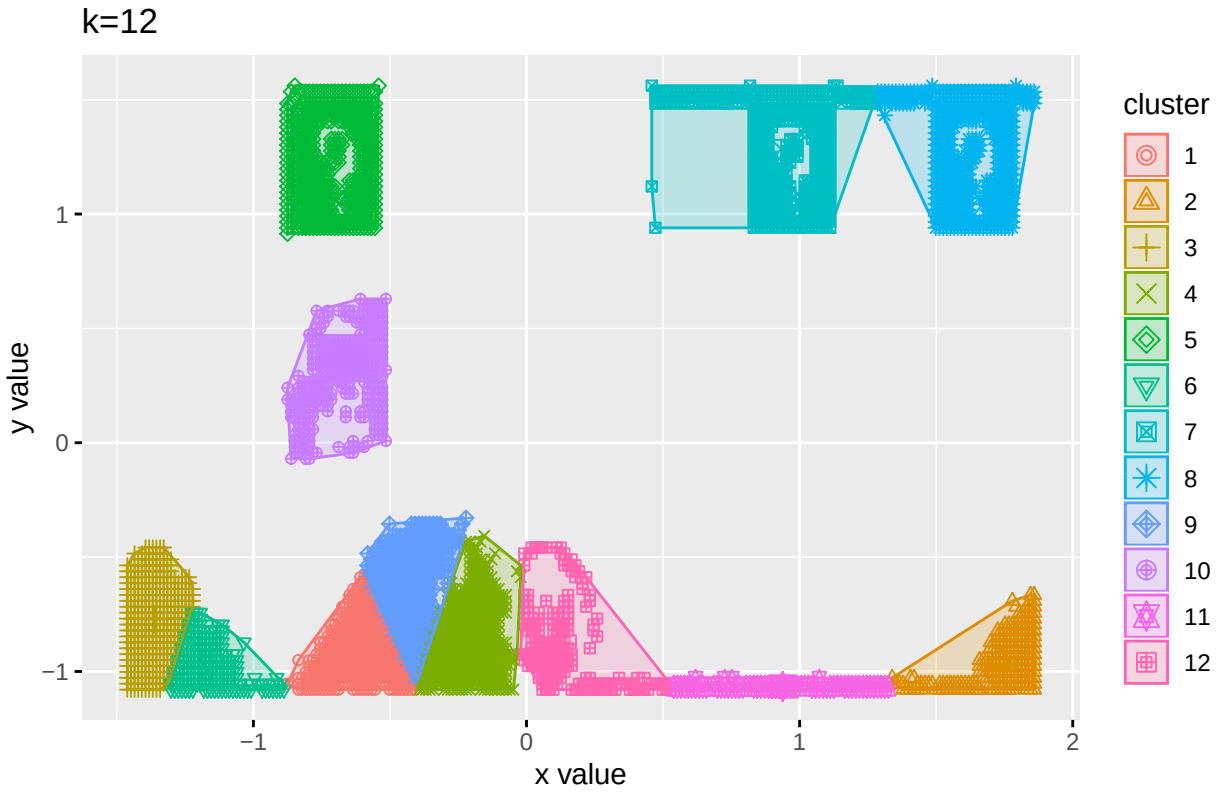
Why, hello there Mario. How is World 1-1 treating you?

B: Fitting the dataset using kmeans



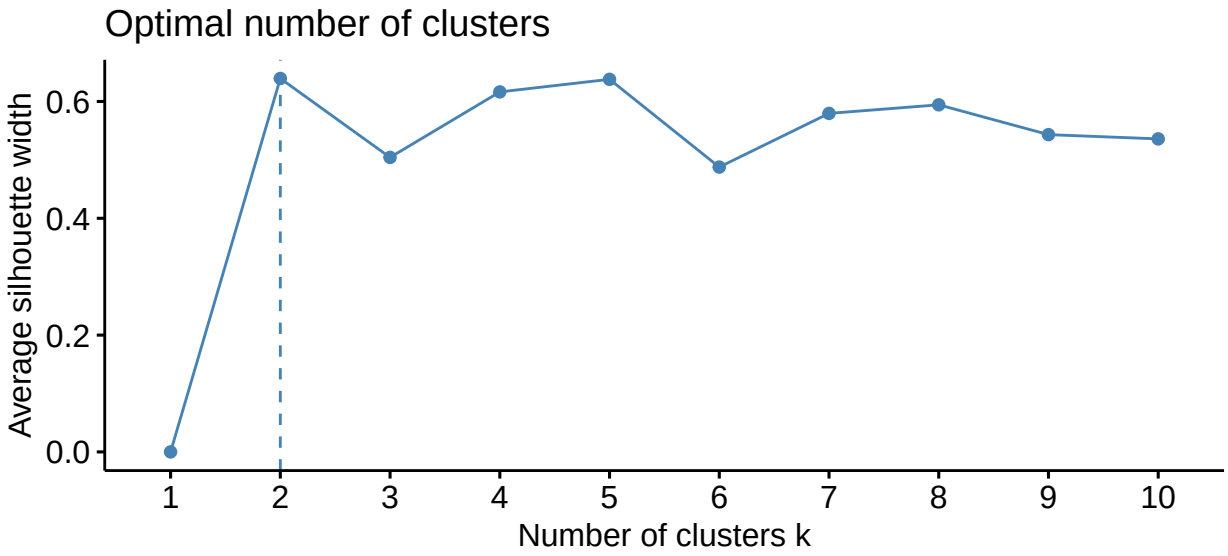

k=8
y value
x value
cluster
1
2
3
4
5
6
7
k=9
y value
x value
cluster
1
2
3
4
5
6
7
8
9
k=10
y value
x value
cluster
1
2
3
4
5
6
7
8
9
10


k=11
y value
x value
cluster
1
2
3
4
5
6
7
8
9
10
11

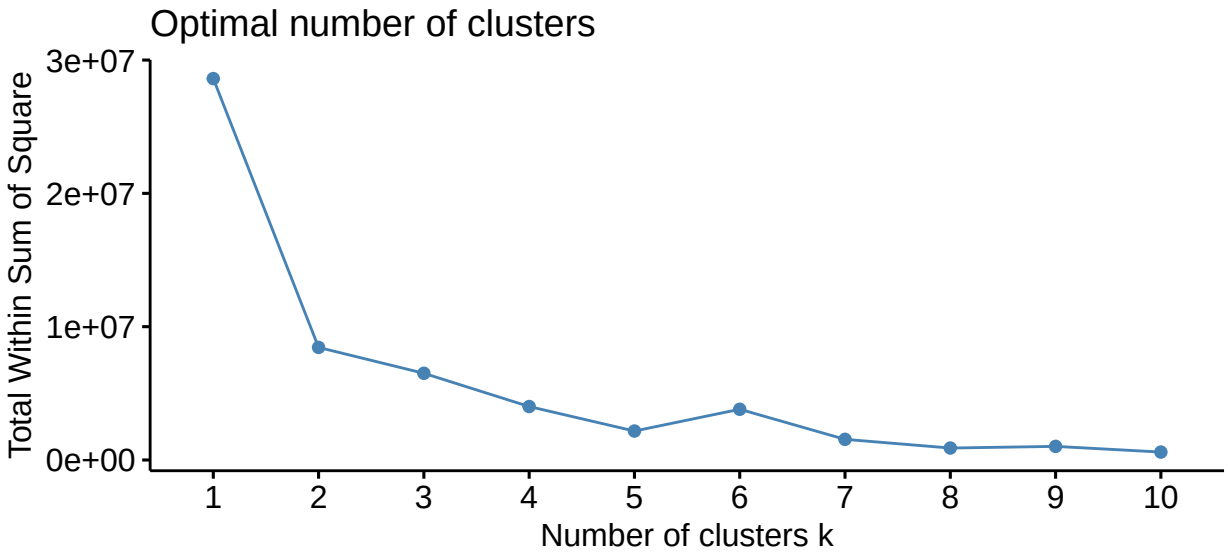

k=12
y value
x value
cluster
1
2
3
4
5
6
7
8
9
10
11
12

C: Plotting the mean distance from the centroid of each cluster given a value k

```
fviz_nbclust(data, kmeans, method = "silhouette")
```




```
fviz_nbclust(data, kmeans, method = "wss")
```



kmeans uses the sum of the squares of the euclidean distance of all points to the center of the cluster. This is used in the elbow method but is different than requested, as the mean of the distance between all points in a cluster and the cluster's centroid is not the same as the sum of squares. I have not found anything in my reading which indicates how to do the latter properly although it should be doable by adding a euclidean distance feature to each observation relative to its assigned cluster and creating a plot based on the means of those distances in each cluster. Given the specific request for the elbow method, however, I think the sum of sqaures method is what is actually being requested and I'm overthinking the assignment.

D: Analyzing the elbow

The elbow appears to be 2 but successive iterations have shown 3, 4 and 5 as well. The Silhouette method however says two is the optimal number of clusters.