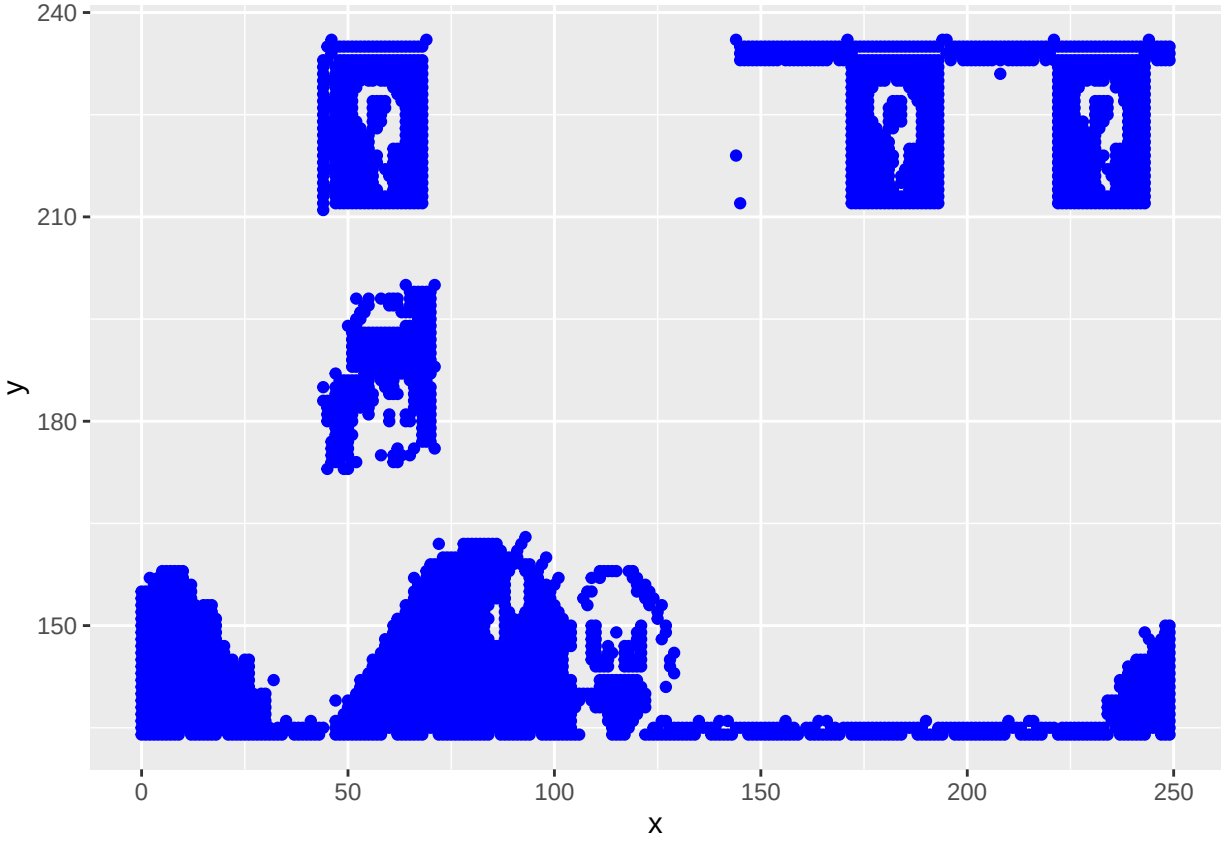
Assignment 8.3:

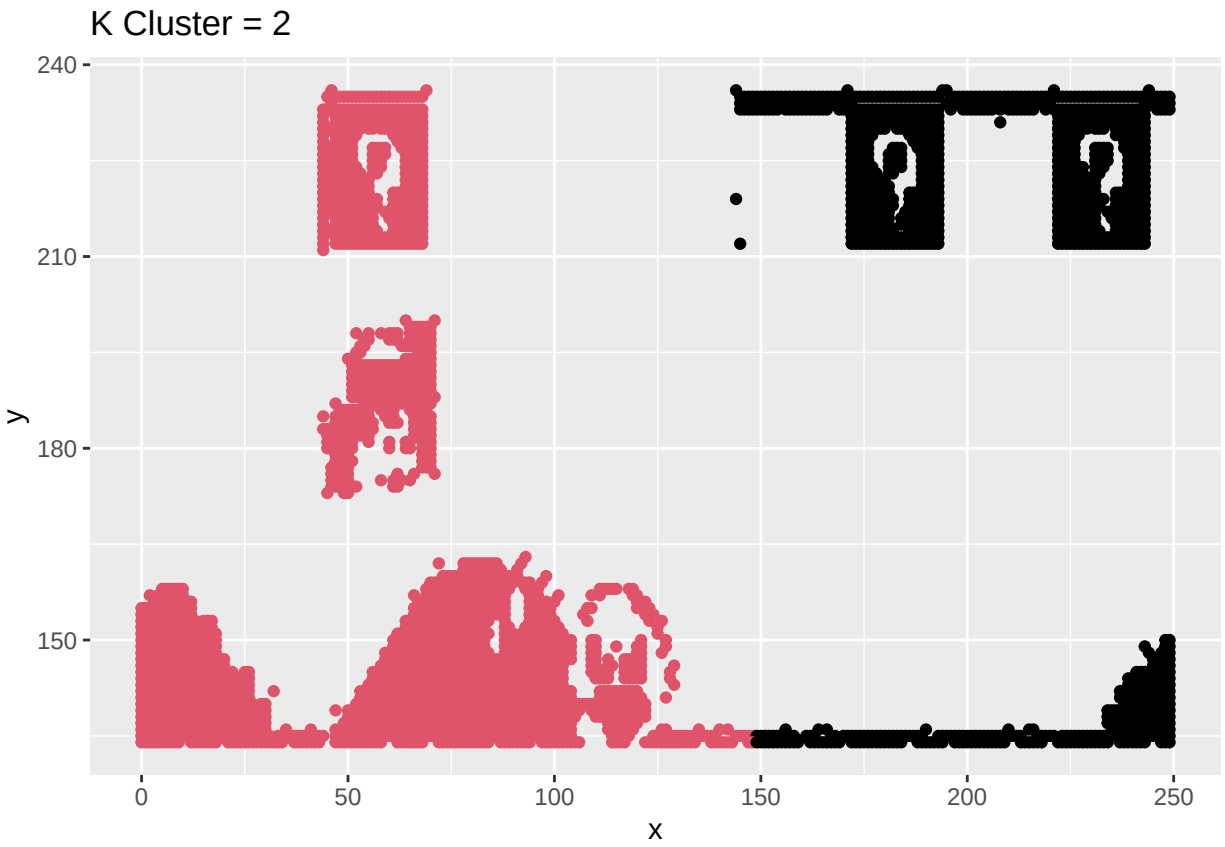
Kurt Stoneburner

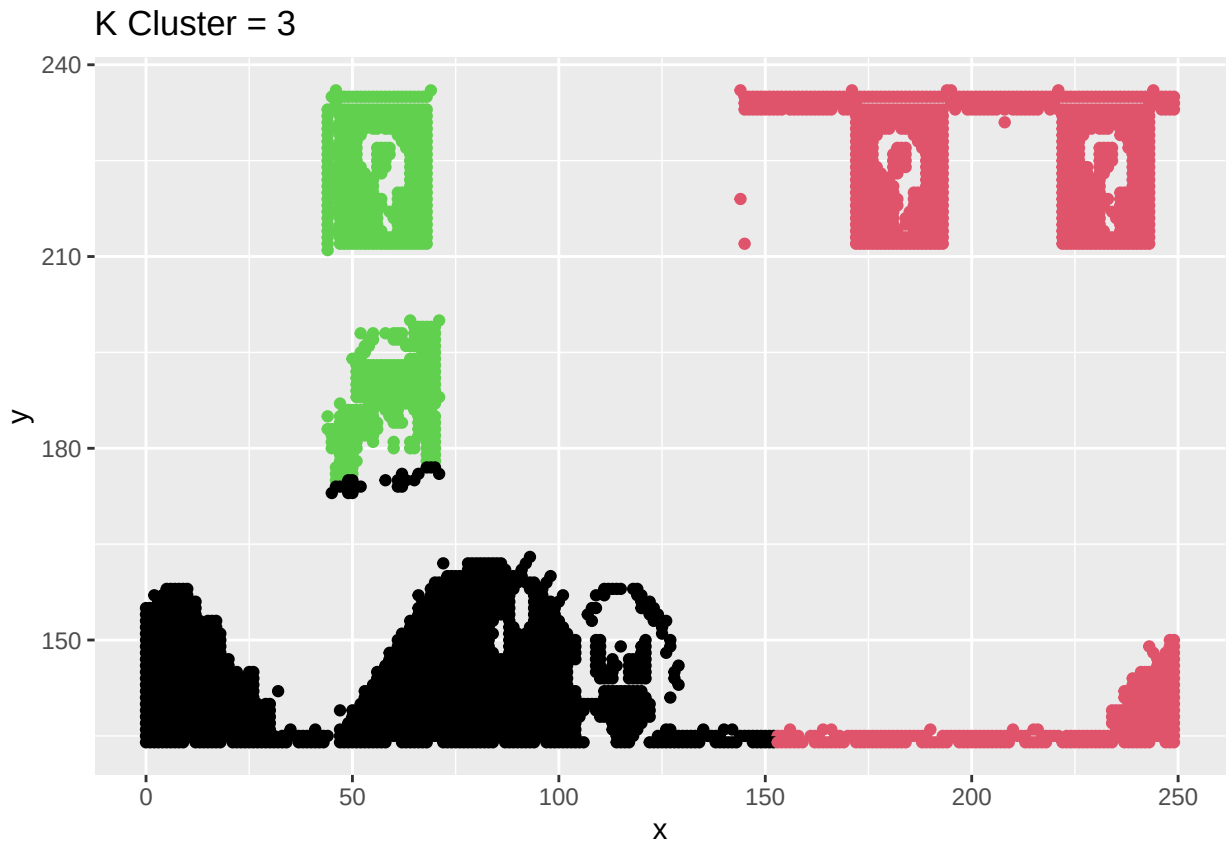
7/21/2020

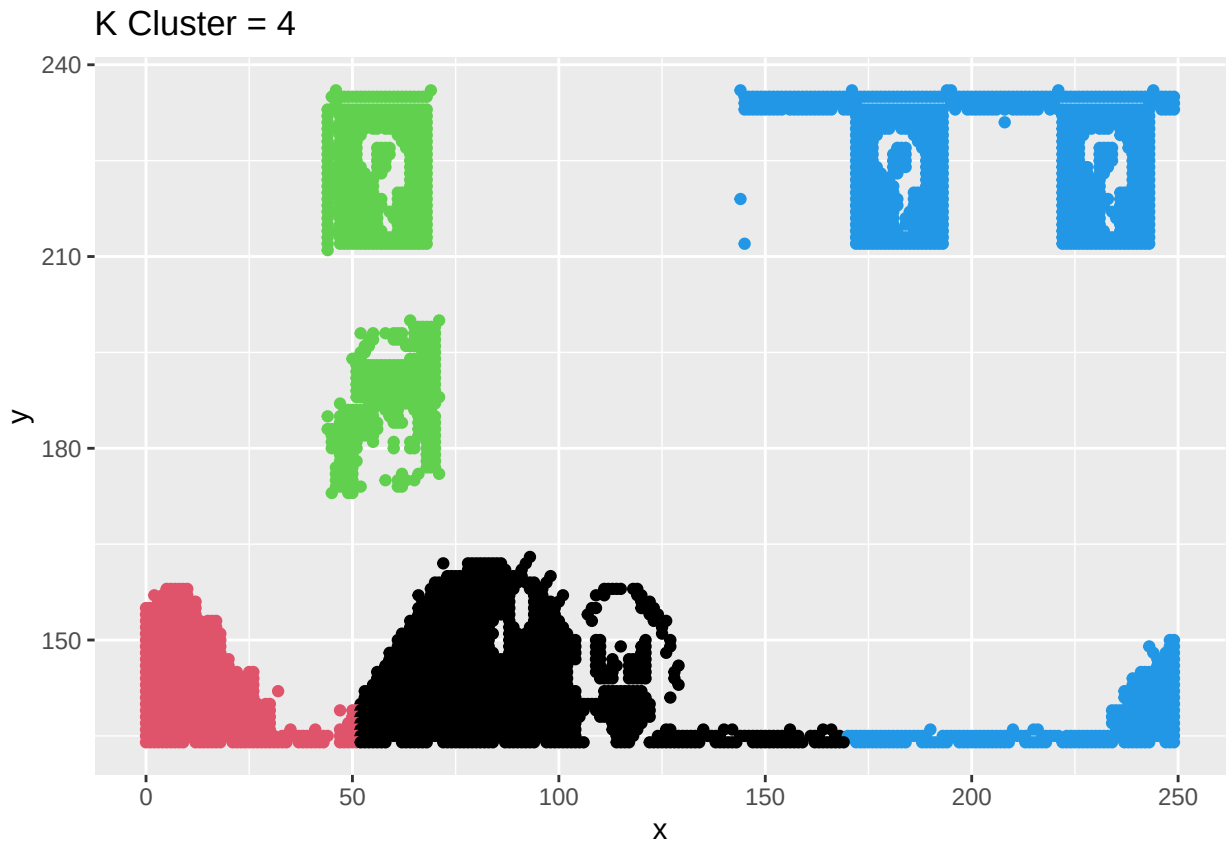
a. Plot the dataset using a scatter plot.

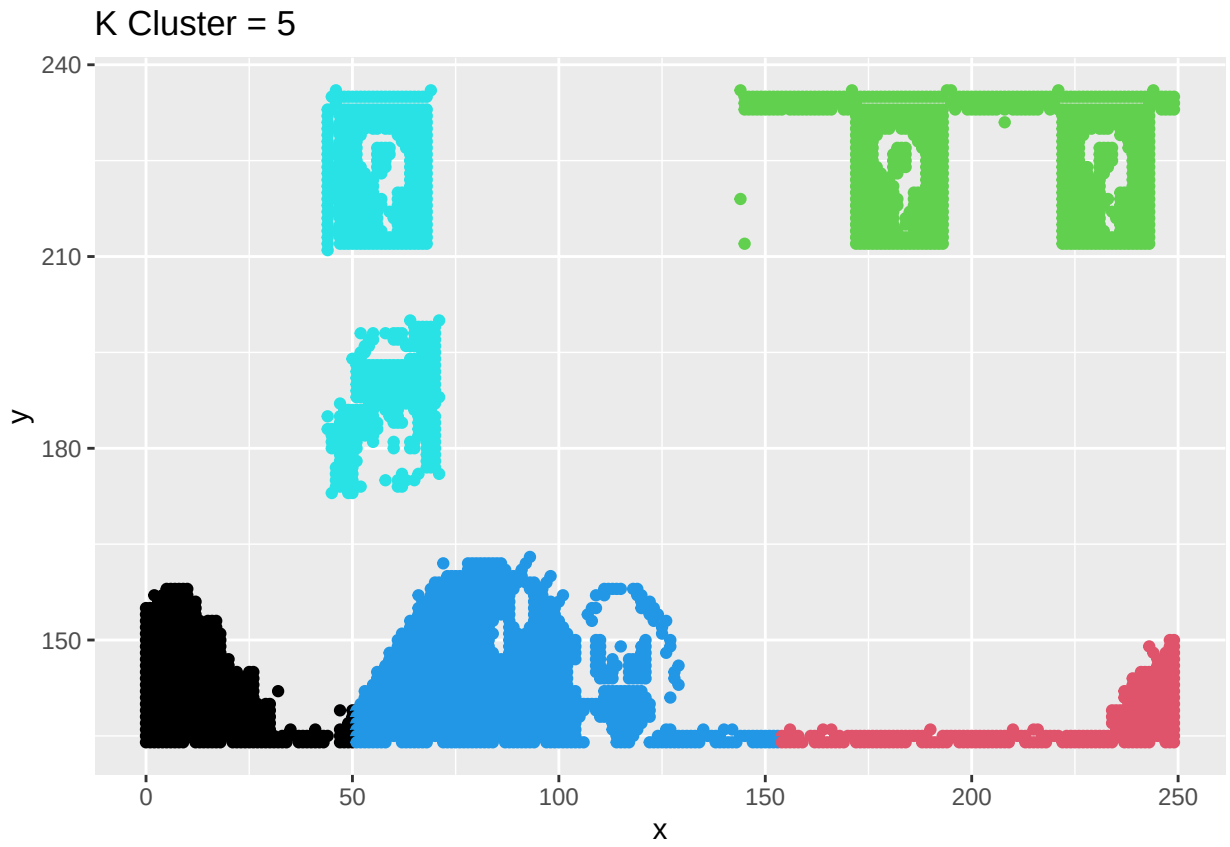


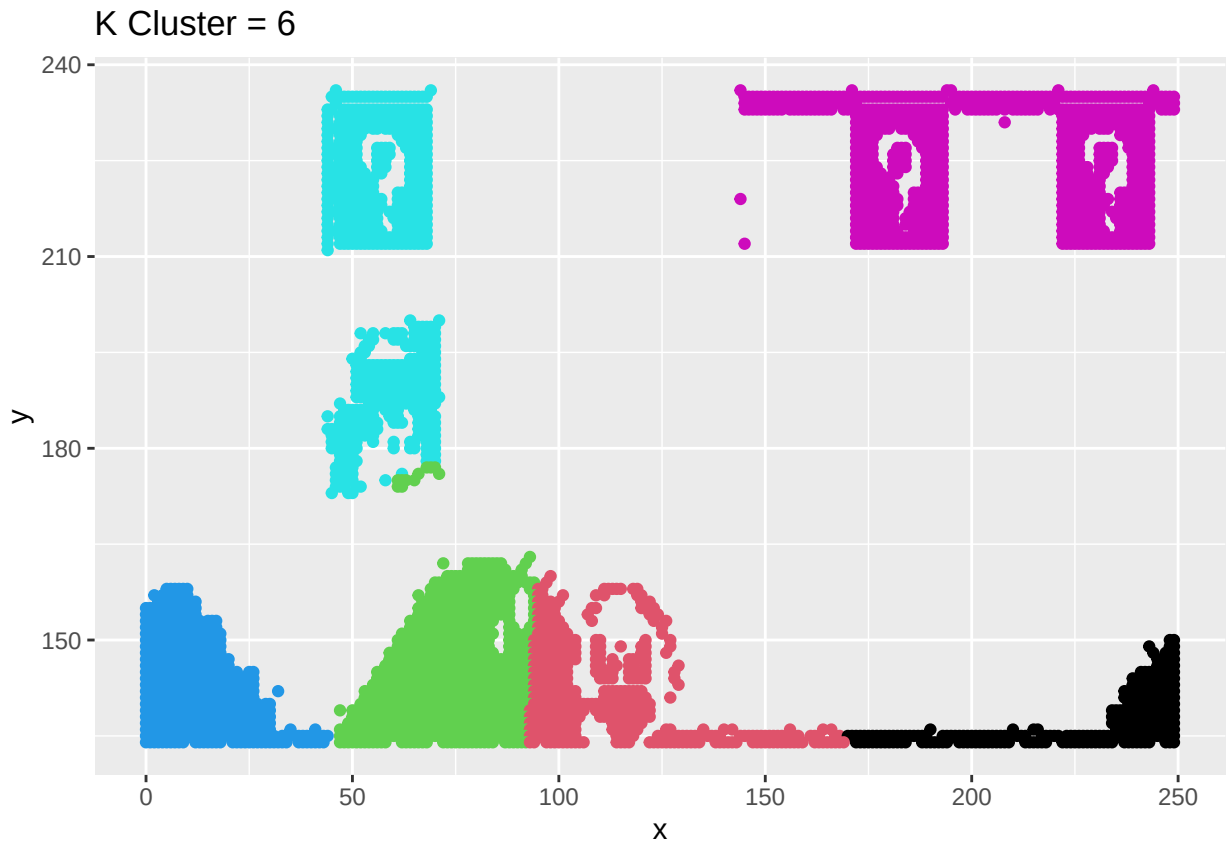
b. Fit the dataset using the k-means algorithm from k=2 to k=12. Create a scatter plot of the resultant clusters for each value of k.

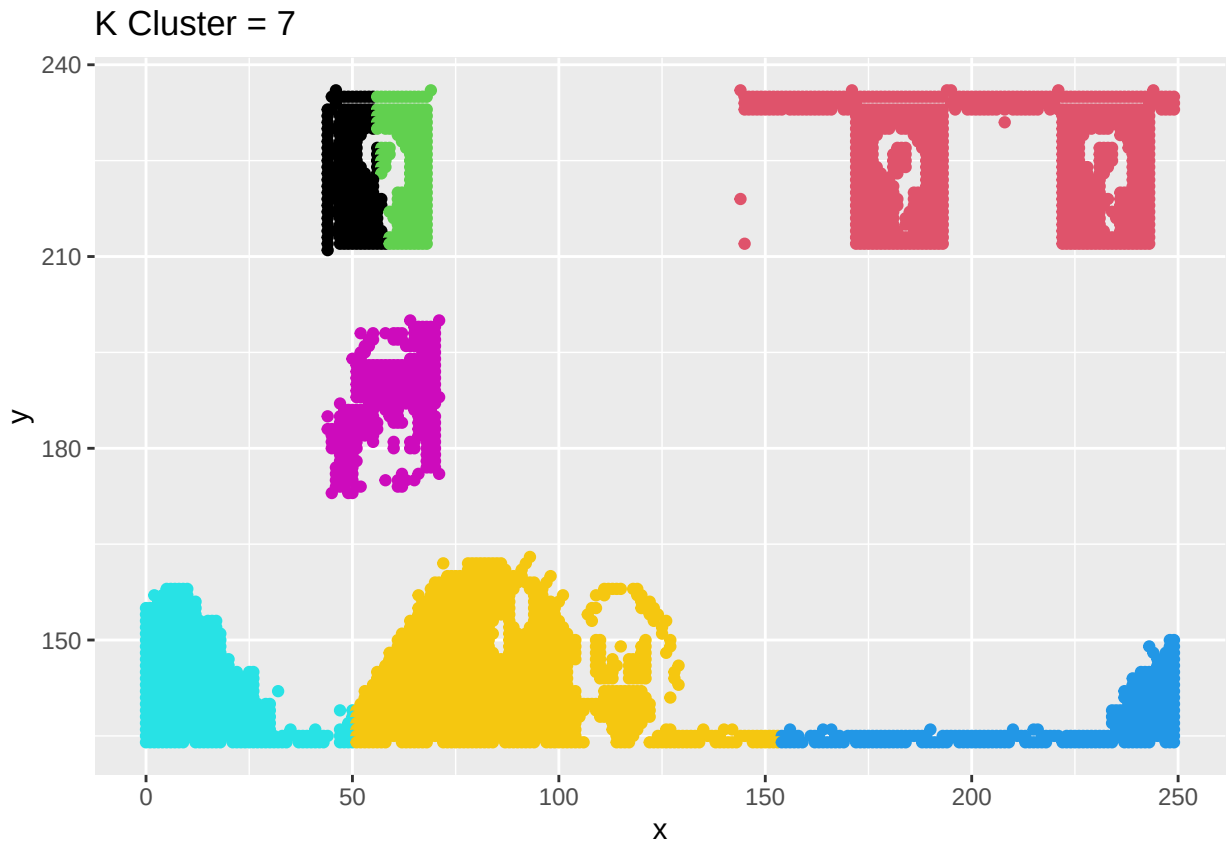

K Cluster = 2
240
210
180
150
y
0
50
100
150
200
250
x

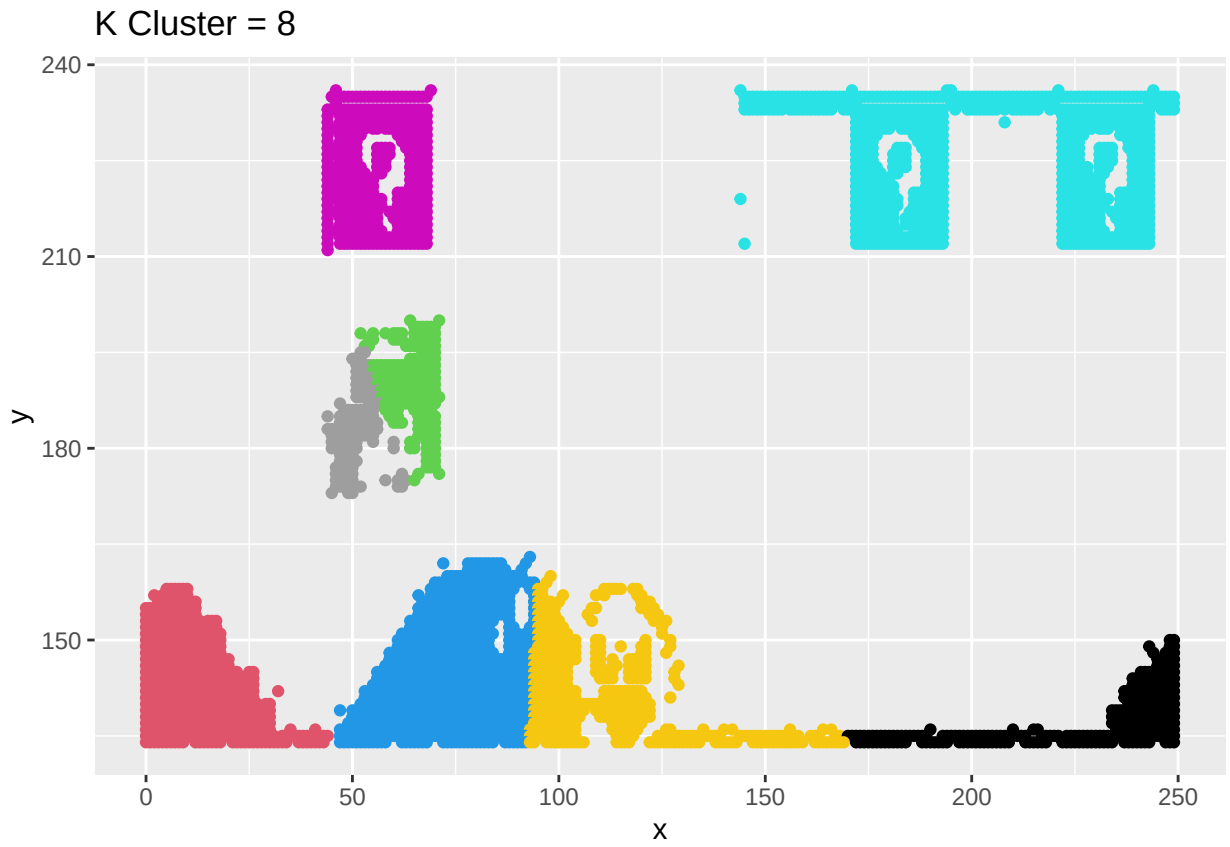

K Cluster = 3
240
210
180
150
y
0
50
100
150
200
250
x

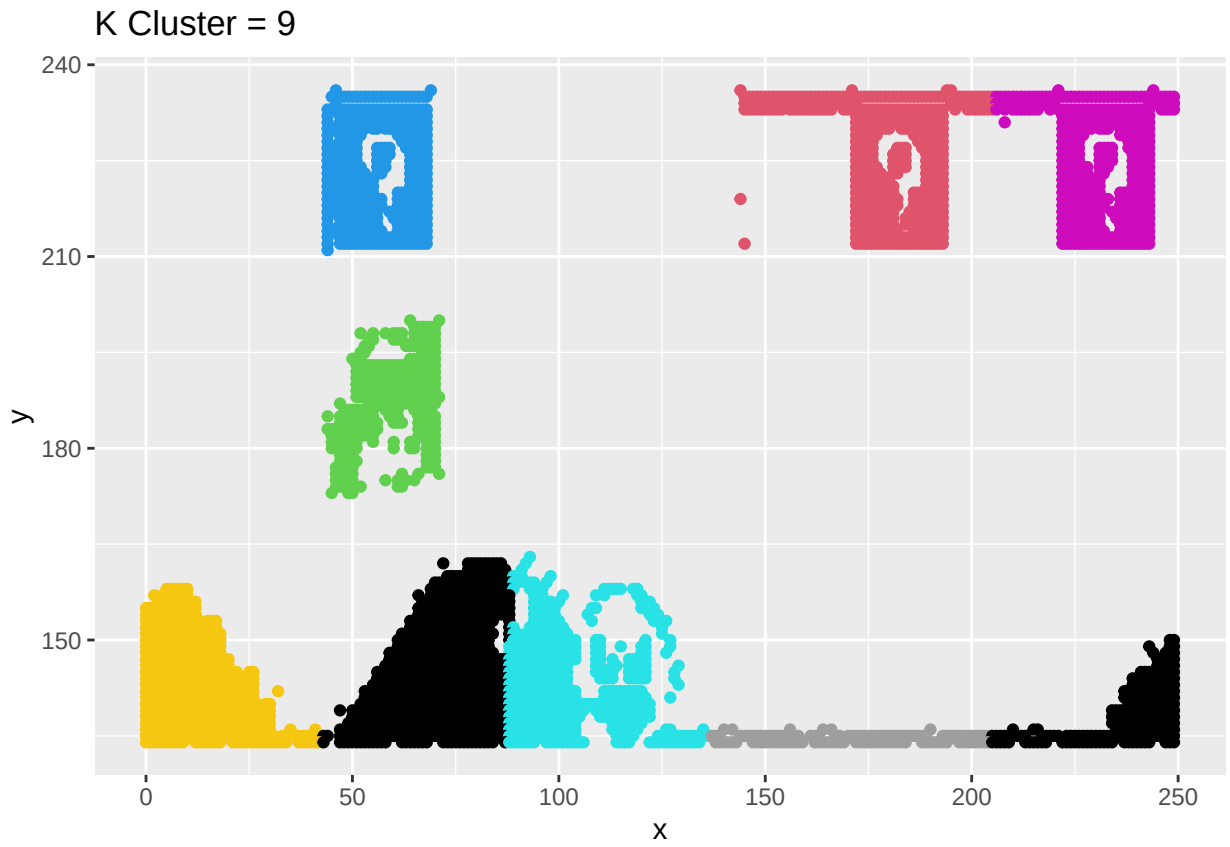

K Cluster = 4
240
210
180
150
y
0
50
100
150
200
250
x

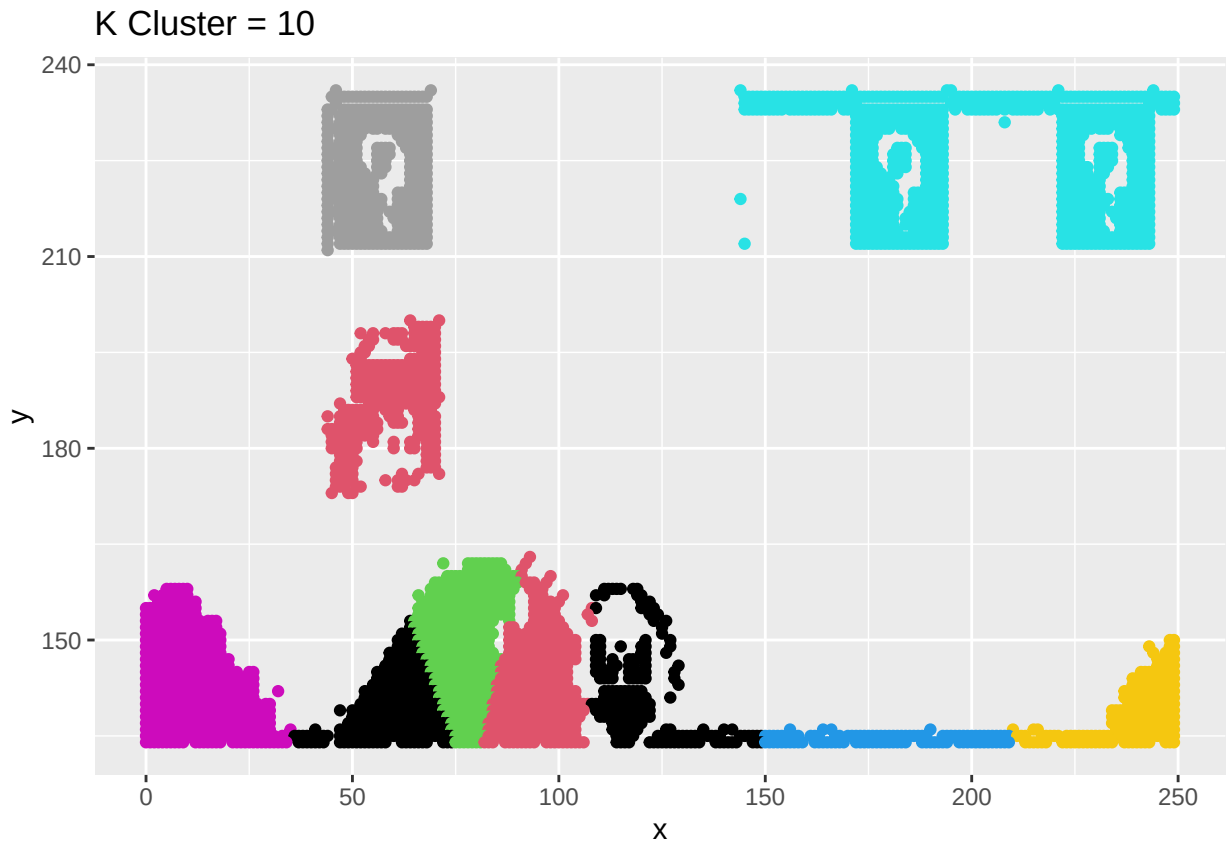

K Cluster = 5
240
210
180
150
y
0
50
100
150
200
250
x

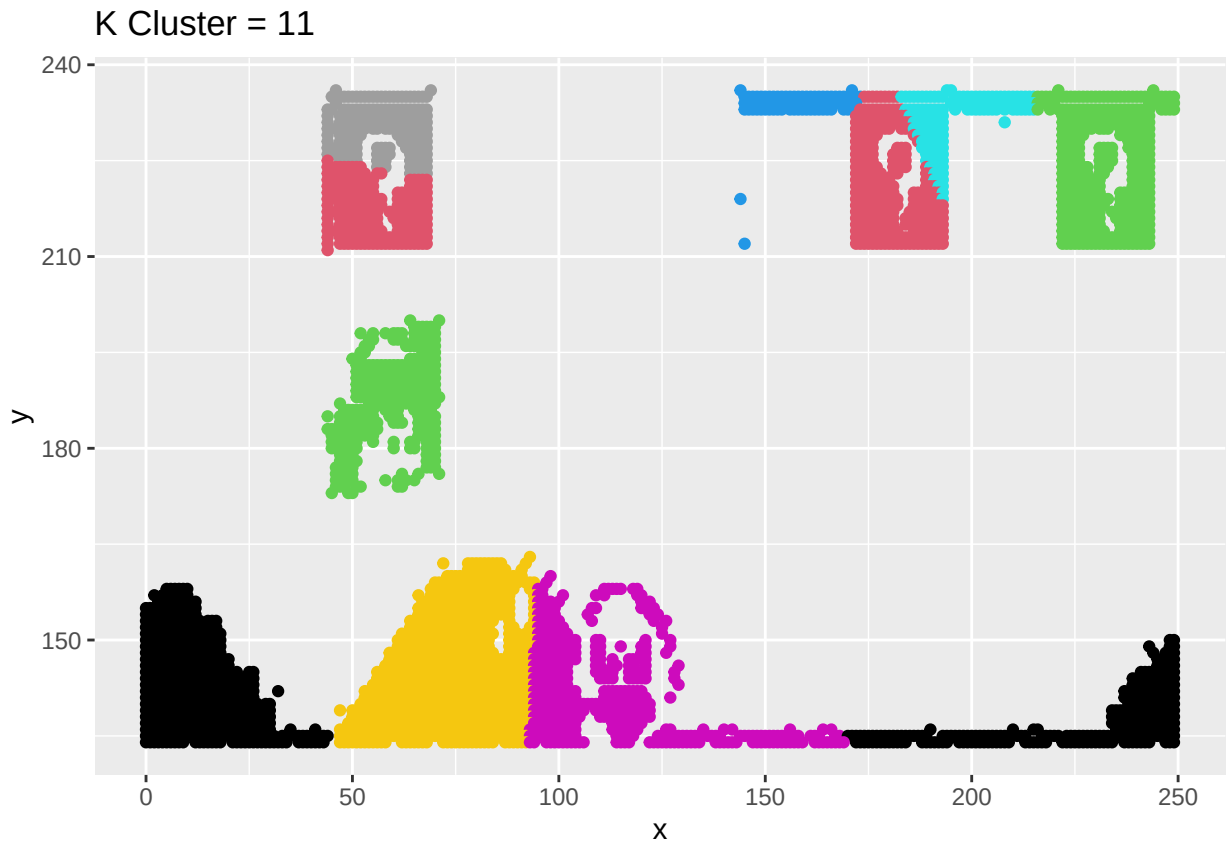

K Cluster = 6
240
210
180
150
y
0
50
100
150
200
250
x

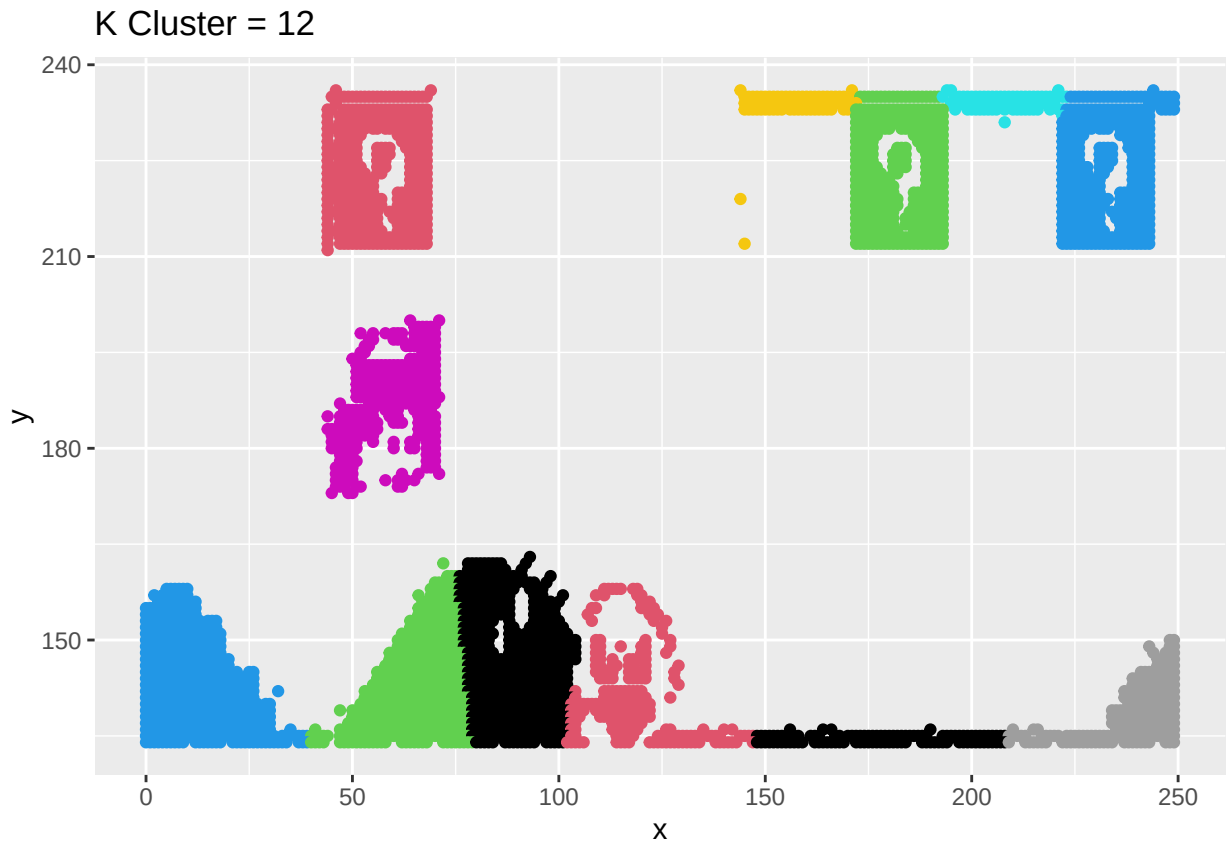

K Cluster = 7
240
210
180
150
y
0
50
100
150
200
250
x


K Cluster = 8
240
210
180
150
y
0
50
100
150
200
250
x

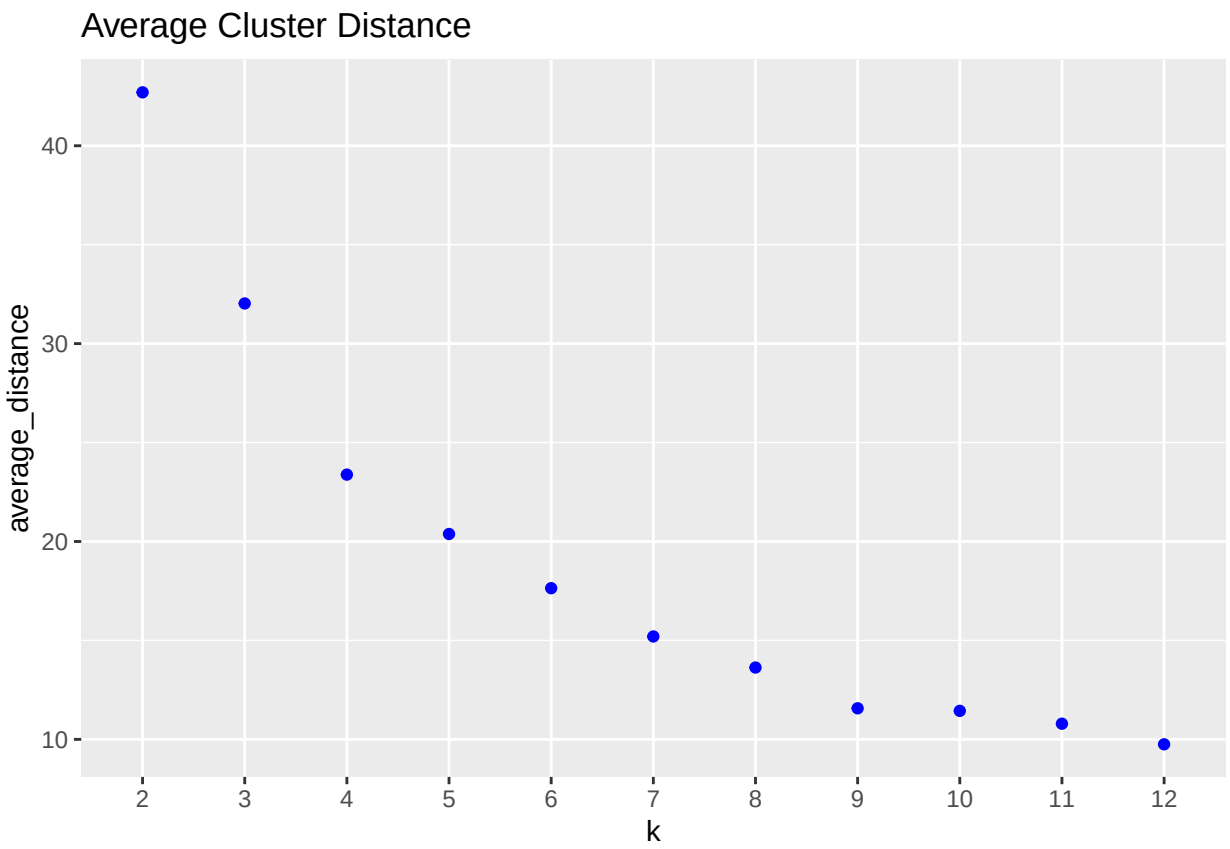

K Cluster = 9
240
210
180
150
y
0
50
100
150
200
250
x


K Cluster = 10
240
210
180
150
y
0
50
100
150
200
250
x


K Cluster = 11
240
210
180
150
y
0
50
100
150
200
250
x



c. 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. Looking at the graph you generated in the previous example, what is the elbow point for this dataset?



The ‘elbow joint’ is at a k of 6 or 7. Looking at the scatterplots k=6 is the better choice.