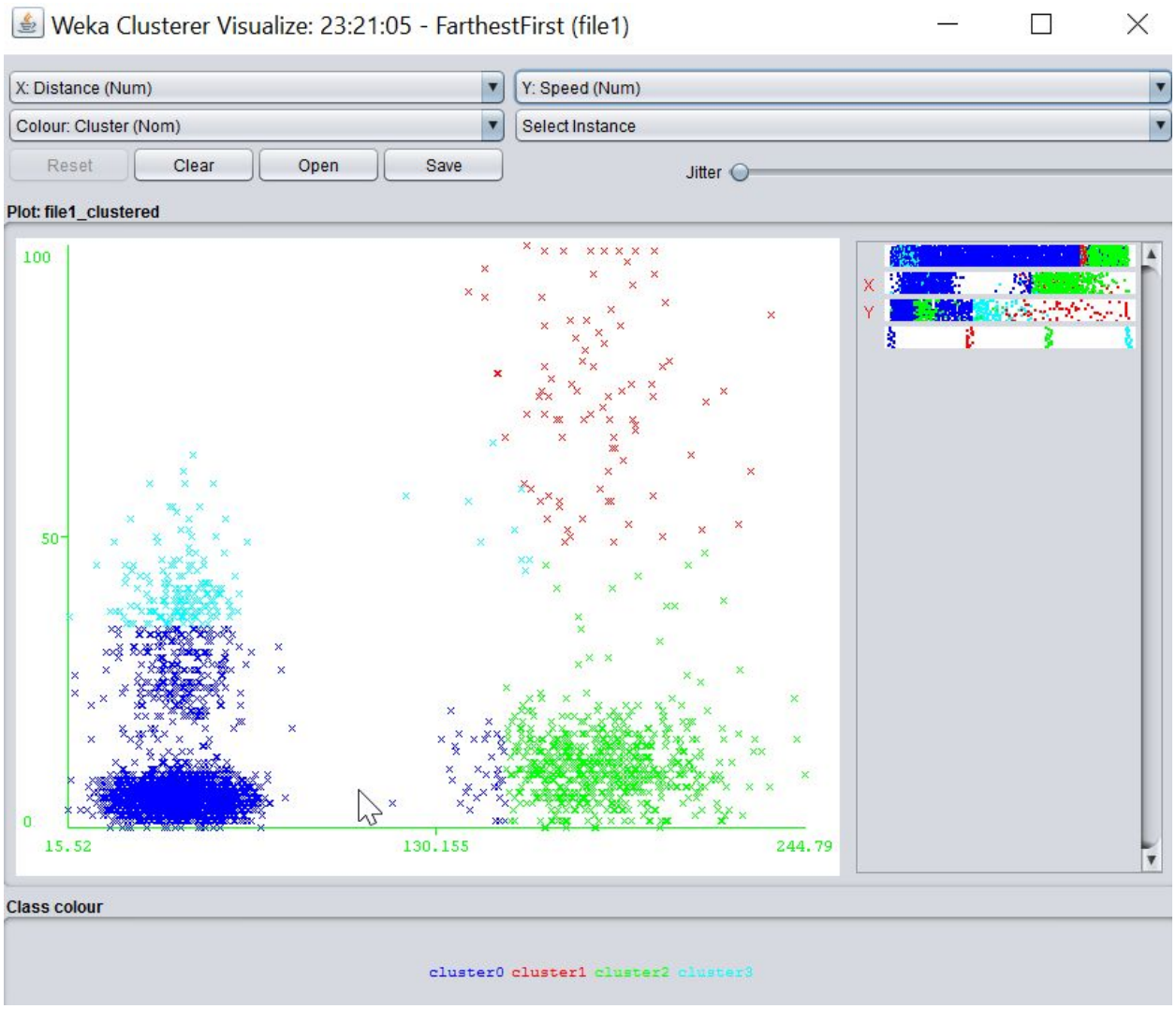
William Seery and Nahom Ogbazghi
Assignment 4 Report
3-29-17

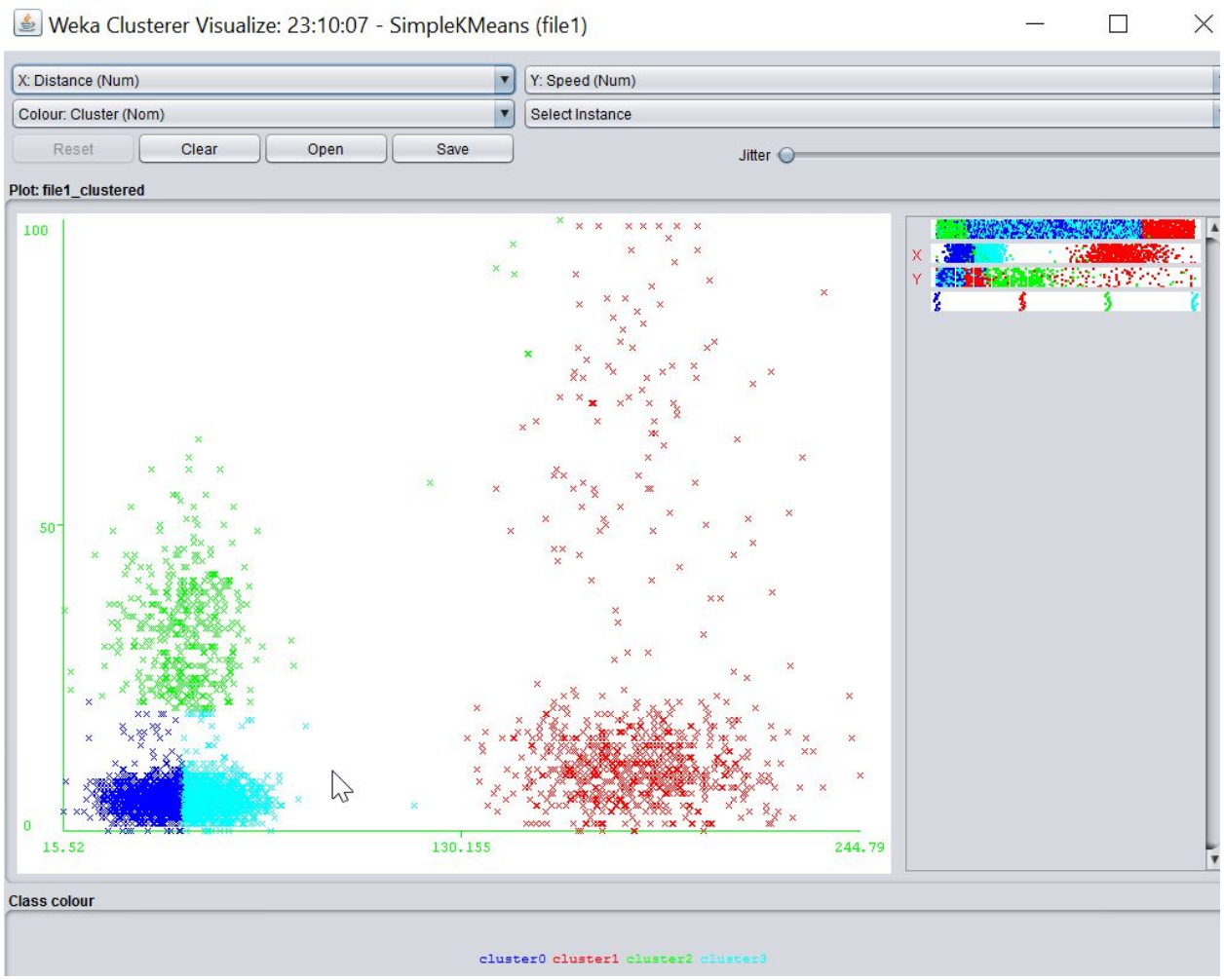
The datasets we used are a log of the speed and distance a set of truck drivers completed and the first two sets of the iris-setosa, classifying different types of iris flowers. We processed the data because we wanted to reduce the data to something that was more manageable, so we removed all other attributes but two attributes per dataset. We also removed the labels for the attributes from the first line of the data.

After testing my data against Weka's Simple K Means, I received similar coordinates for k = 4, but there were some deviations in certain coordinates. We did about 10 iterations of WSS and took the average due to deviations after each iteration. Our average came out to be approximately 926.4 for file1 and 6.1 for file2.

We have pictures of the graphs from weka of k means clustering with 4 clusters and farthest first clustering with 4 clusters. These charts are for the driver dataset, file1. The results for the other dataset were similar.

There was a difference between k means and farthest first. K means had centroids located differently, making the quantity of items more uniform between centroids. For the farthest first clustering, the item allocation was more uniform based on distance in the overall graph. For pre processing, it would help if you take out labels for the attributes to make the data easier to work with and to take out unrelated attributes. If there are missing values in the dataset, that could be an issue for some clustering algorithms, so it would also be useful to remove items that are missing missing values using pre processing.


Weka Clusterer Visualize: 23:21:05 - FarthestFirst (file1)
X: Distance (Num)
Y: Speed (Num)
Colour: Cluster (Nom)
Select Instance
Reset
Clear
Open
Save
Jitter
Plot: file1_clustered
100
50
0
15.52
130.155
244.79
X
Y
Class colour
cluster0 cluster1 cluster2 cluster3


Weka Clusterer Visualize: 23:10:07 - SimpleKMeans (file1)
X: Distance (Num)
Y: Speed (Num)
Colour: Cluster (Nom)
Select Instance
Reset
Clear
Open
Save
Jitter
Plot: file1_clustered
100
50
0
15.52
130.155
244.79
X
Y
Class colour
cluster0 cluster1 cluster2 cluster3