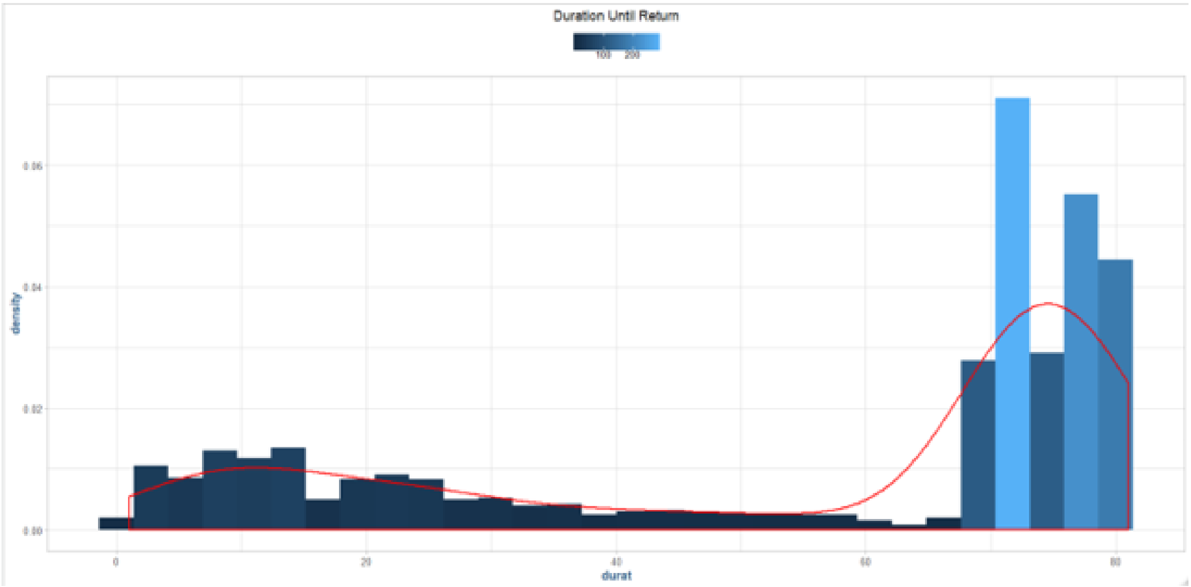
Assignment #3

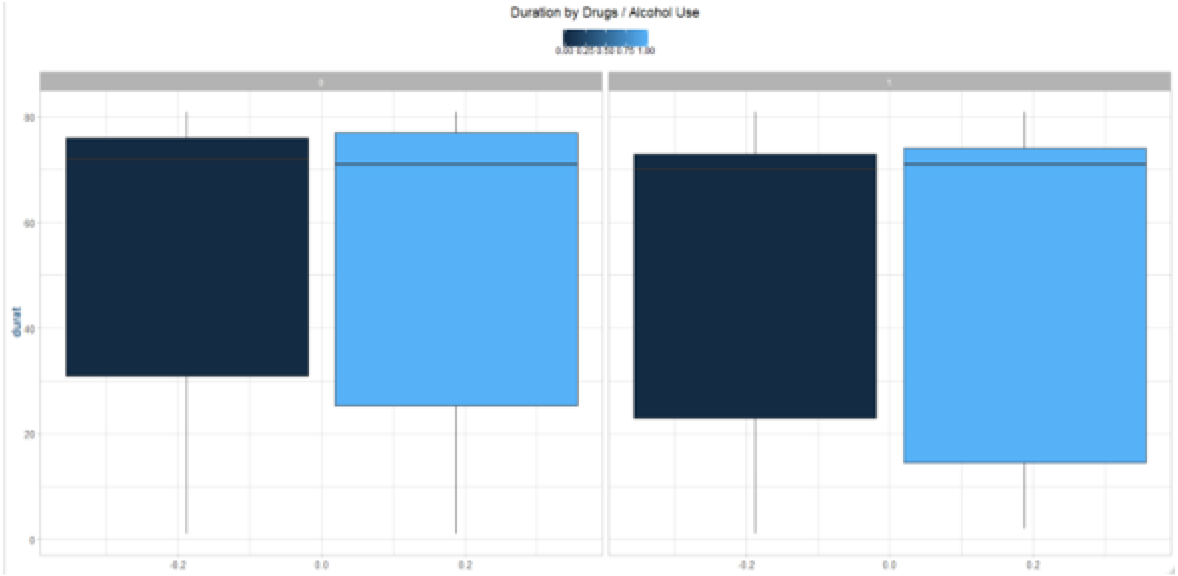
June 2021

1)

The first interesting variable in the data set is the duration variables which represent the amount of time that elapses until a person returns back to prison once they have been released. There is a large clustering of values between 70 and 81 months, although the study only goes to 81 months.

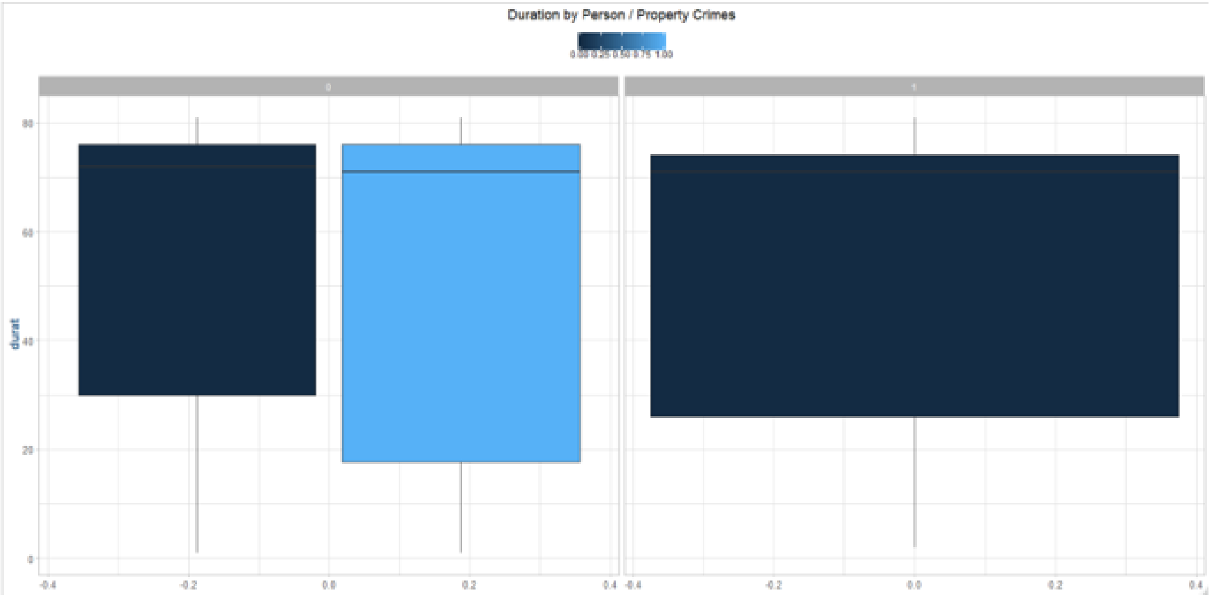


Drug and alcohol usage:

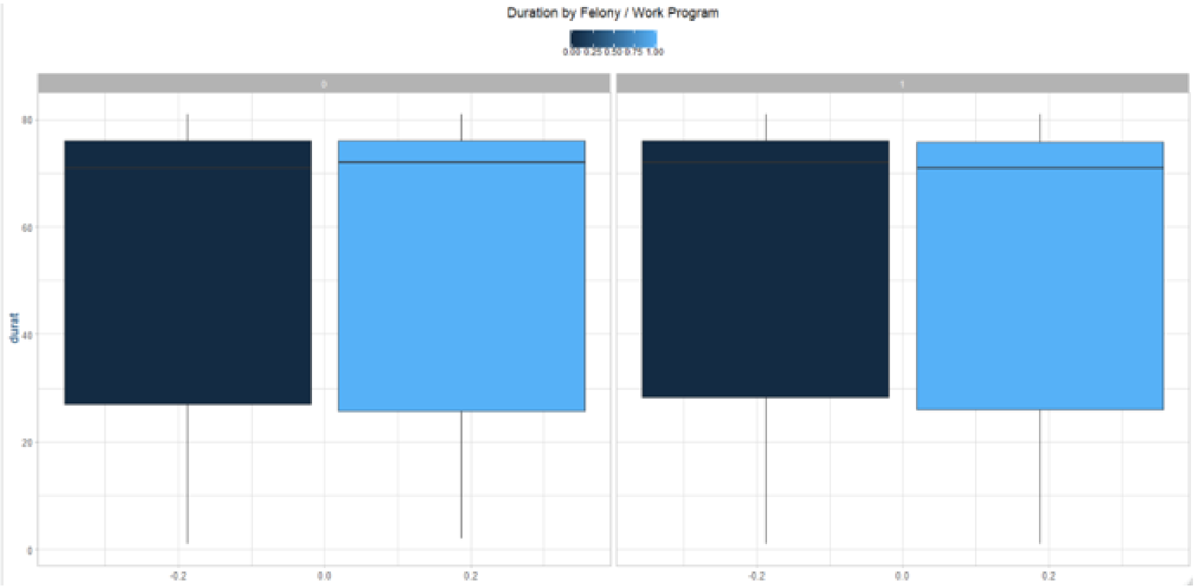


There is a large amount of variance and the mean is right in the large tail.

Crime committed:

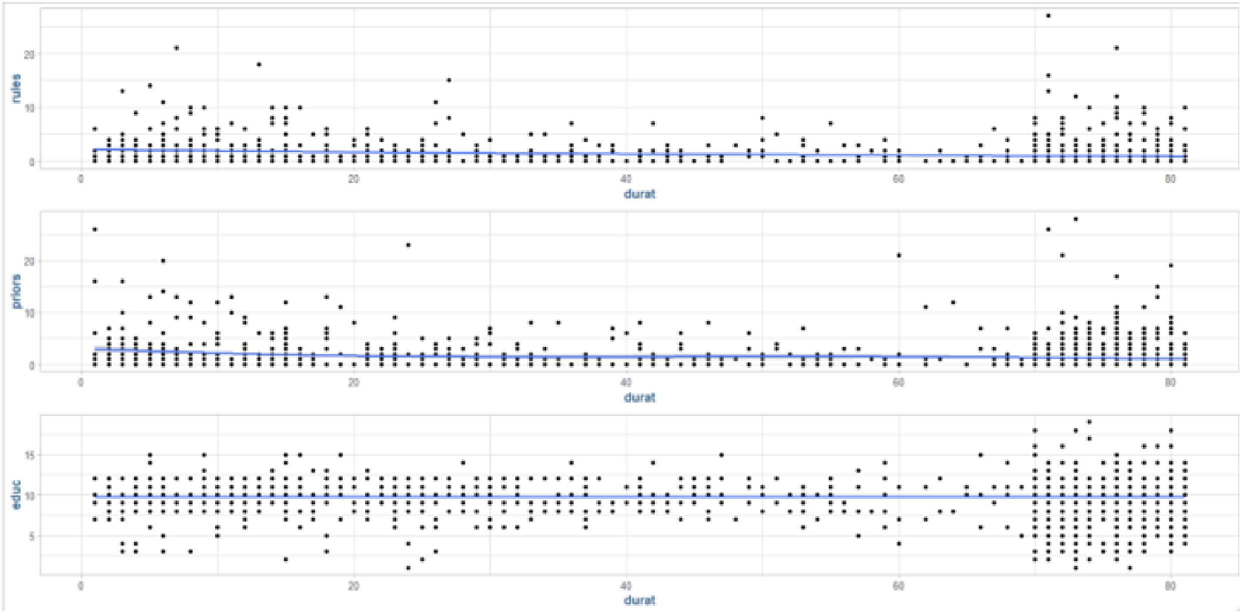


Felons, non-felons, vs work program:

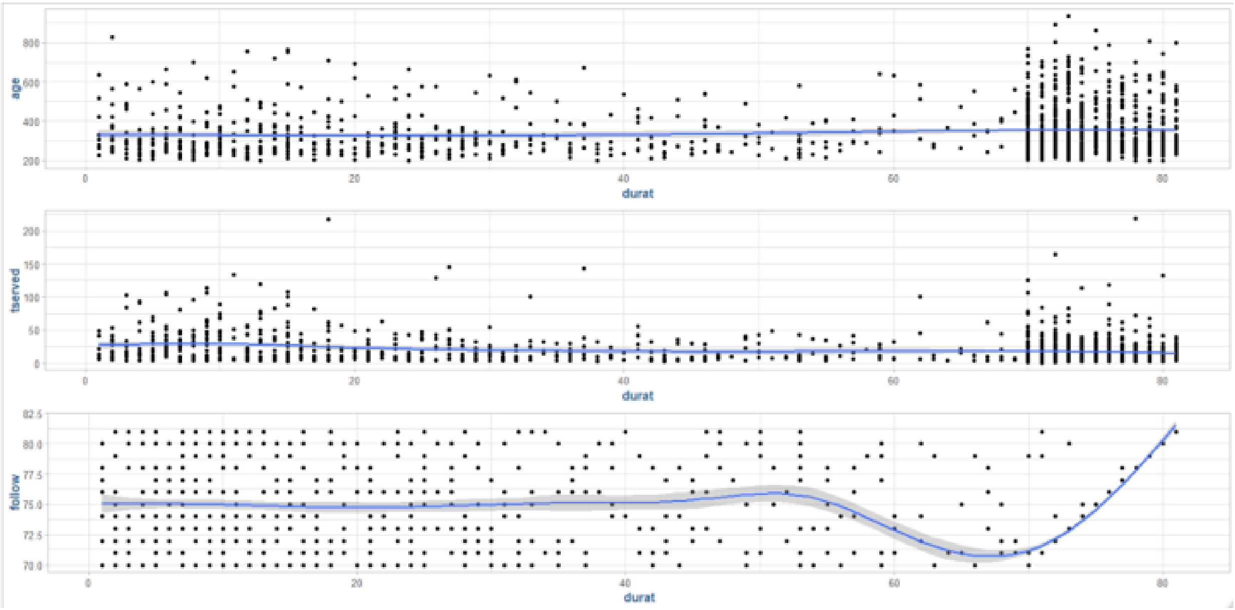


These variables show similar results and don’t really explain duration well enough.

The dichotomous variables do not yield a lot of information about the variables, so next I’ll look at the discrete variables and their relationship to return duration.



The numbers of rules broken, priors, and years of schooling show that the average inmate has close to 10 years of schooling.

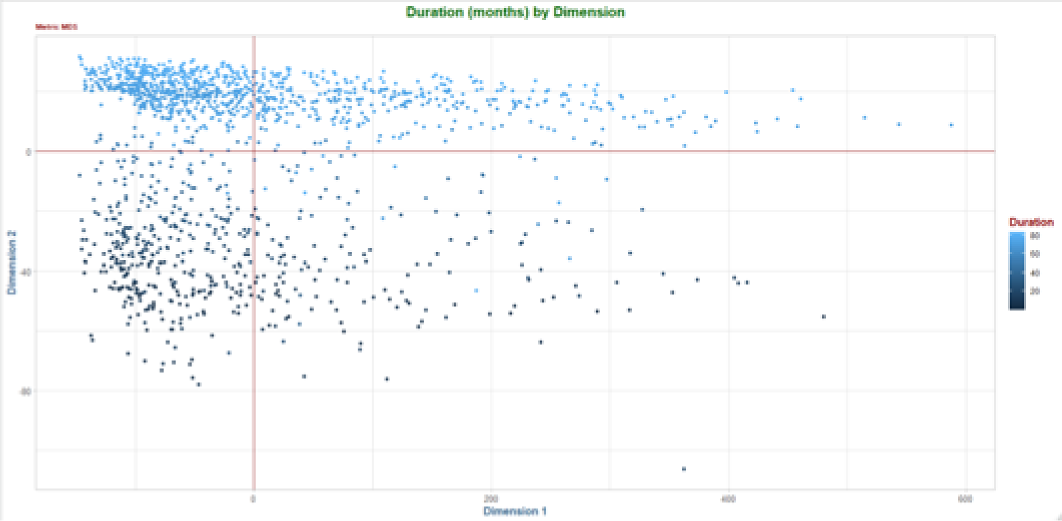


The length of follow up time seems to be uniformly distributed until after 70 months from release, after this, it seems their follow up period is directly proportional to their return time. These variables did not yield too much information though.

2)

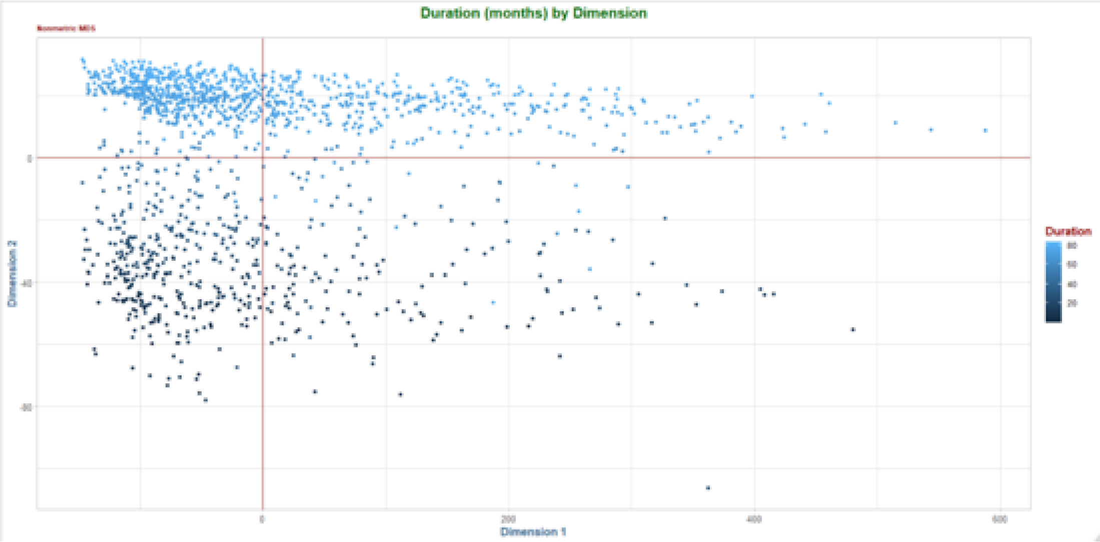
There are no discernable patterns in the raw Euclidian distances, they seem to be uniform columns with standard increments.

3)



The most prominent clusters are around inmates with a longer duration of returning to prison. After viewing nearly all the variables in the dataset as compared to this variable there is no discernable pattern. However, the multidimensional scaling dimensionality reduction approach, we can see there are hidden relationships in the data as inmates with longer return times are clustered together.

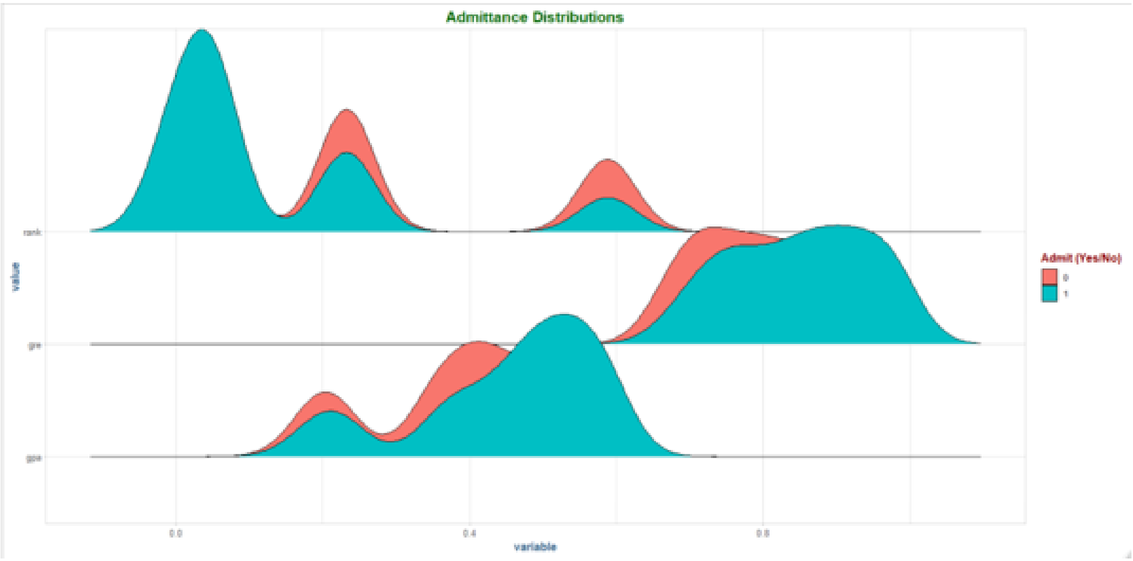
4)



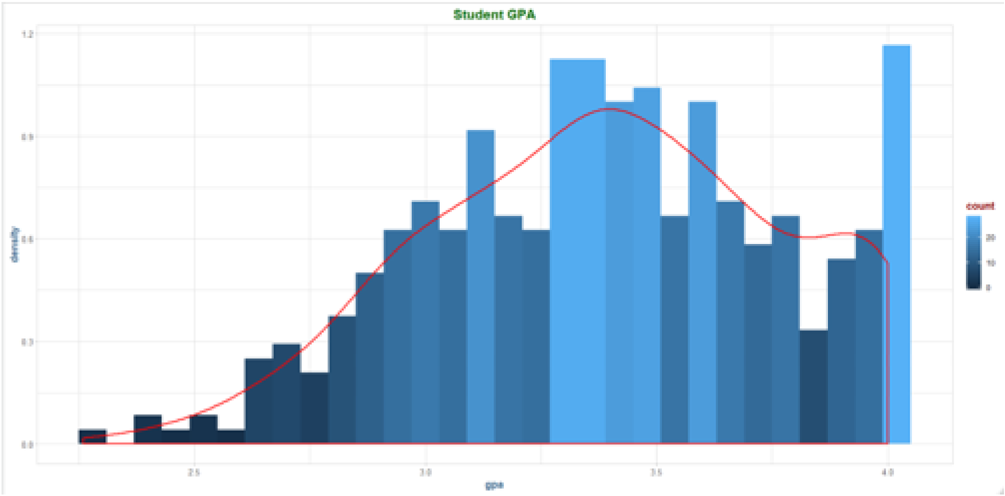
The non-metric multidimensional scaling is very similar to the version from the previous model with some changes in data points.

5)

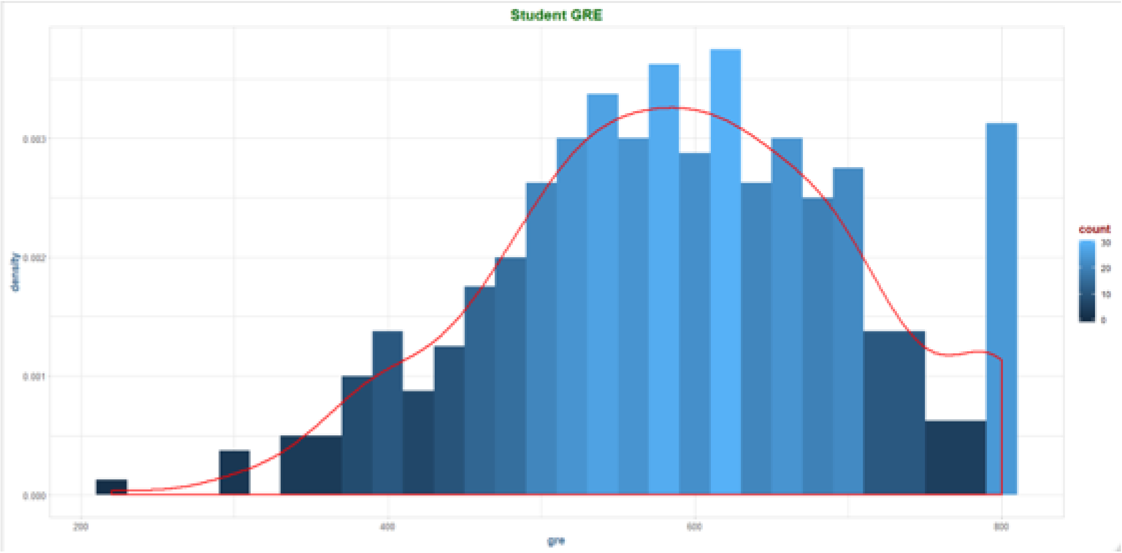
Distributions of variables associated with college admittance:


Admittance Distributions
rank
gre
gpa
value
0.0
0.4
0.8
variable
Admit (Yes/No)
0
1

GPA's:

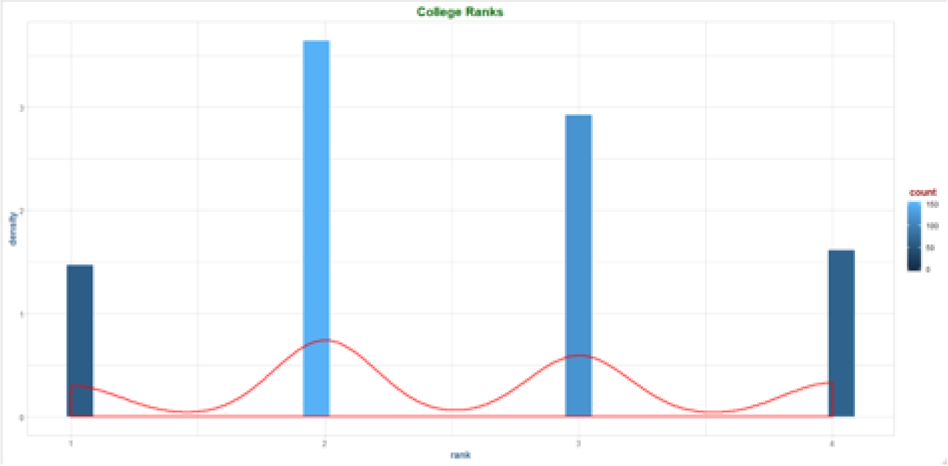

Student GPA
1.2
0.9
0.6
0.3
0.0
density
2.5
3.0
3.5
4.0
gpa
count
20
10
0

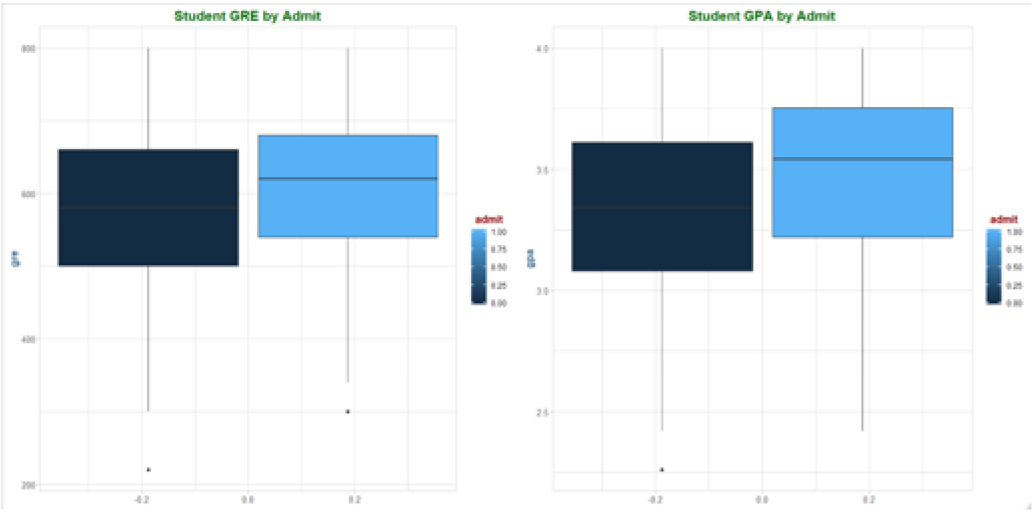
GRE's:



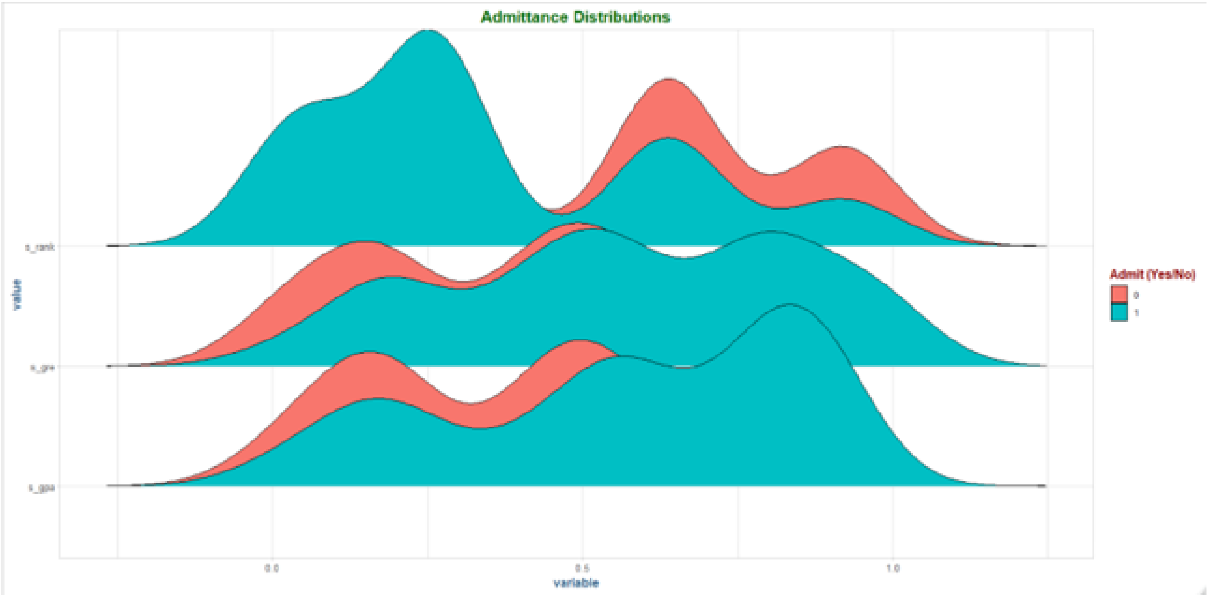
There are a lot of maximum values with both GPA’s and GRE’s hitting the maximums.

College Ranks:





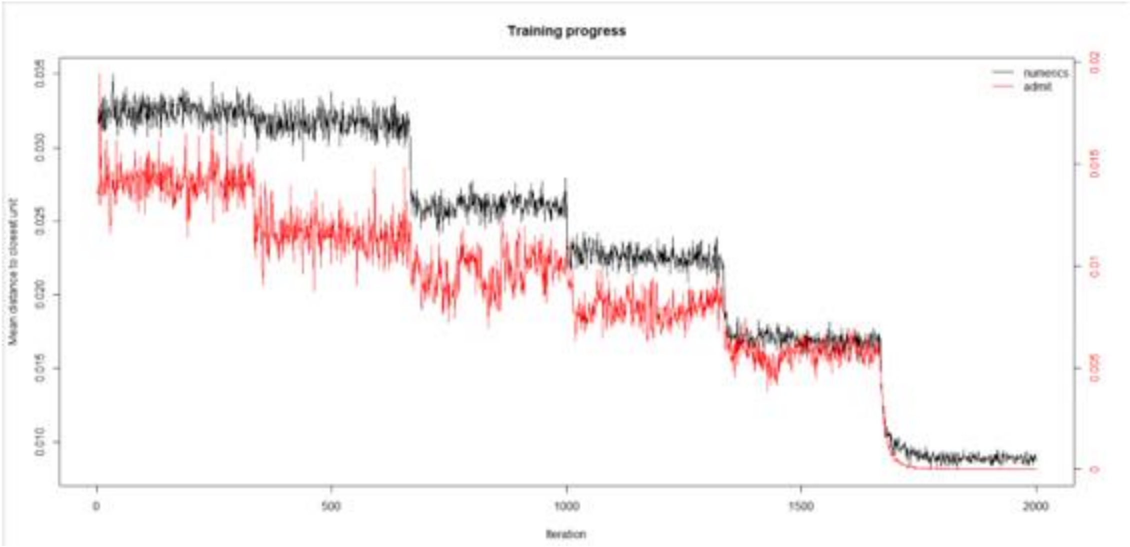
Yes I believe we need to scale the data in order to ensure that no individual variable has too much influence in our mapping. The post-scaled distributions maintain their respective shapes from the previous model, and all the variables are now on the same scale with the admit variable being omitted since its not dichotomous.



I created 3 new variables in the model above for the scaled version of gpa, gre, and rank.

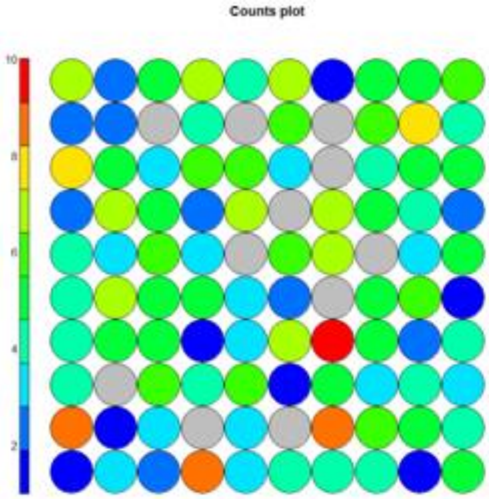
6)

I will use 2,000 epochs in the beginning, with a 10 x 10 data set with 400 rows.



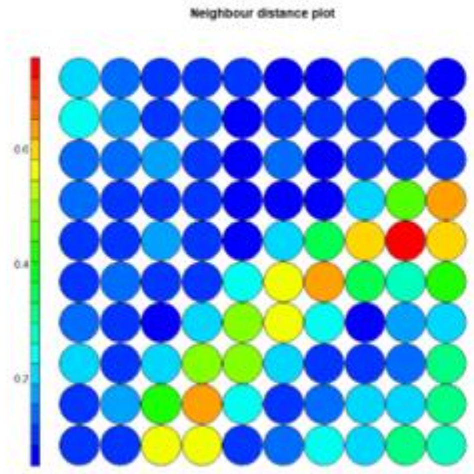
7)

Yes, the 2,000 epochs were adequate enough to train the model and reach a plateaus at zero around 1,800 runs.

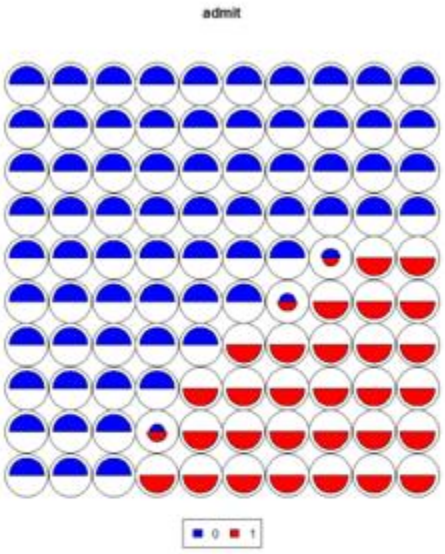


There are a lot of grey bubbles in the map, suggesting the grid is too big. There is one area with a high count but there are not enough to suggest the map is too big.

The average node count is 4.5



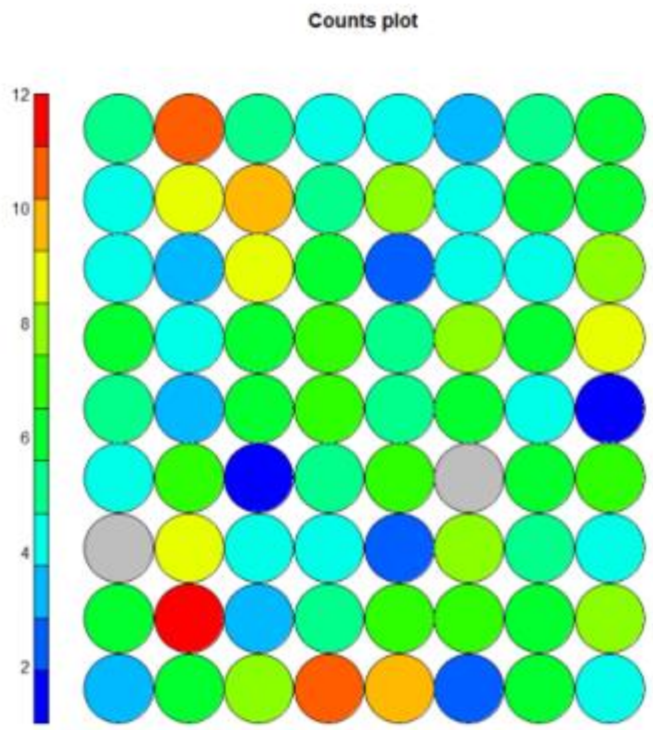
The cluster in the middle right is surrounding the one red dot seems to be fairly distant. There are also two yellow bubbles in the bottom right that could also qualify.



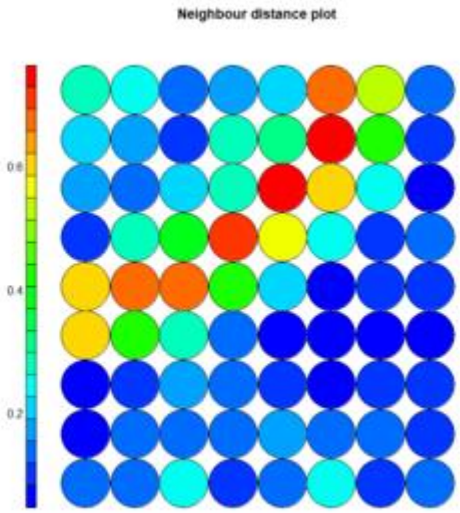
The students who got accepted have similar gre and gpa scores and college rank criterion and they are clustered together. There are some outliers which is also to be expected in the last diagonal showing students who could have been accepted with lower score or had higher scores and go into a lower ranked school.

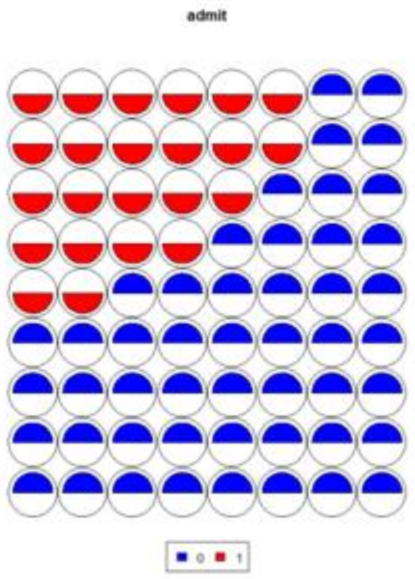
8)

I minimized the grid due to the large amount of gray bubbles in the previous model, and changed it to an 8 x 9 matrix, deciding not to keep a 9 x 9 because the data seems to fit the asymmetric matrix better. The resulting plot seems to be a better fit.



The mean observations per node increases to 5.71 with the new grid size and the distance plot has more red dots overall, but the colors seem to be more concentrated in specific areas and overall display less dispersion.





There seems to be a clearer separation between those who were admitted to those who were not in this codes diagram. The nodes mixing admit and non-admit aren't present and the attribute weights are clustered appropriately.

9)

I learned quite a bit in this assignment and was exposed to a lot of great new things. I have never created an MDS or SOM model and it was great to be able to create those and work through them. The first model seemed to be difficult to find explanatory data connections via the response variable despite there being a great number of variables in the data set. How they all relate to our response variable of return duration did not prove to be fruitful. However, the unsupervised learning model was able to split the data into clusters, splitting the data from 0-65 month return, and 65-85 months return.

The self-organizing map was interesting to create and view as I had never built one before. Higher dimensional data might have been more interesting here, although the basic structure of the data made it easier to utilize. The grid size parameter was fun to explore and see which iteration created the best results. The smaller grids seem to be too loaded with info to utilize and thus adjusting it was important.