ASSIGMENT #3

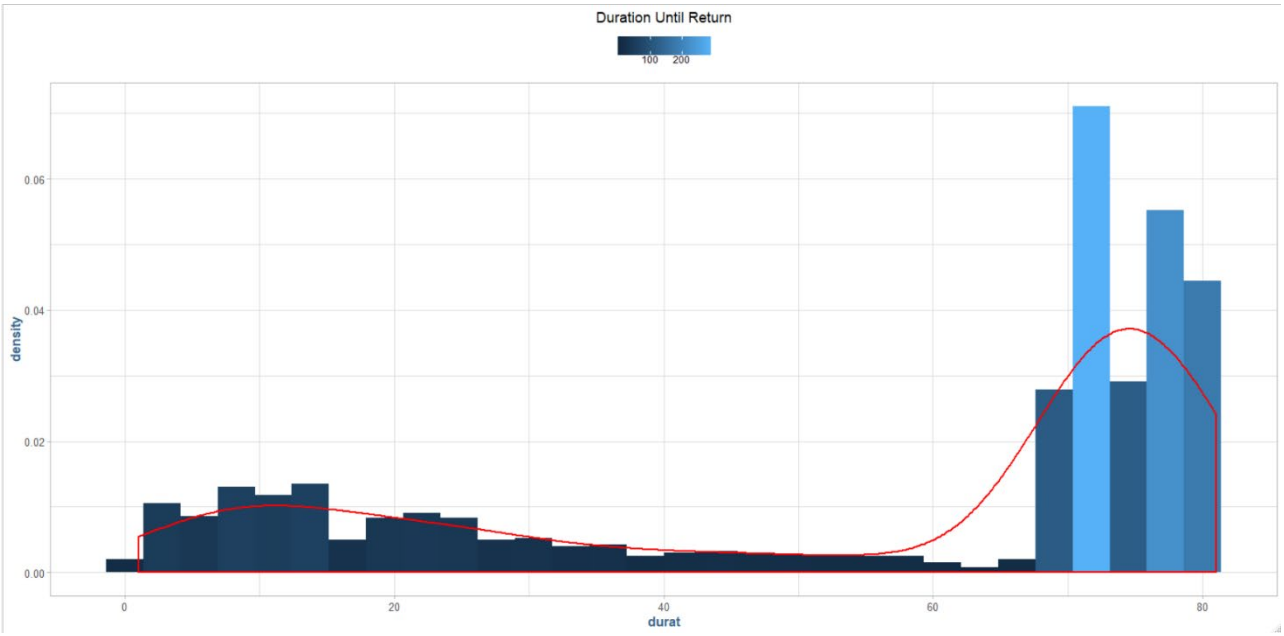
BRANDON MORETZ

MULTIDIMENTIONAL SCALING

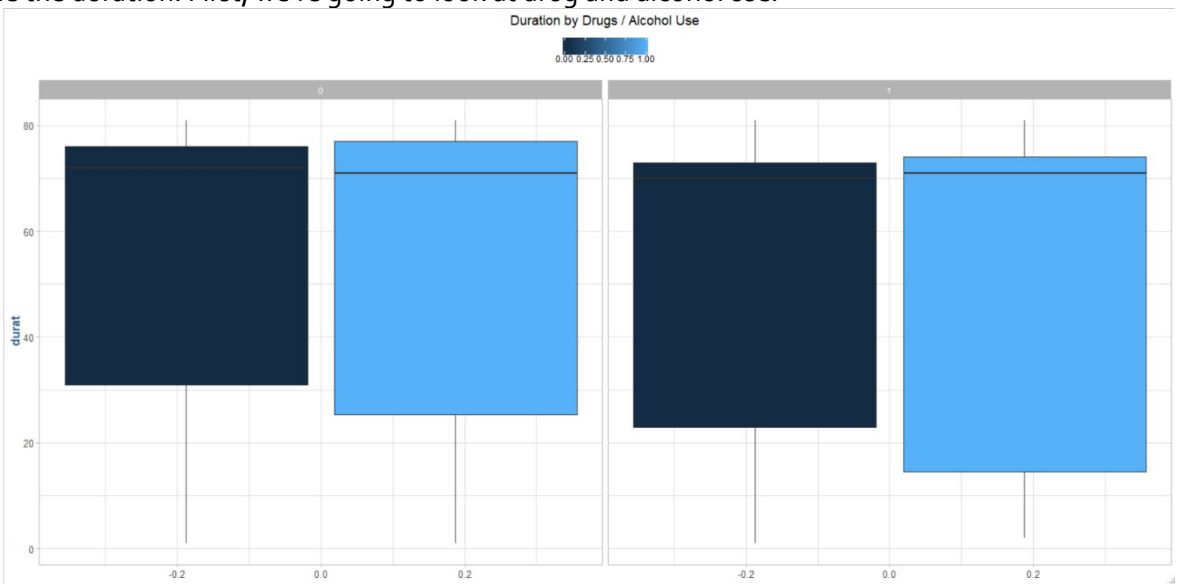
ASSIGNMENT TASKS

1) *Perform a basic Exploratory Data Analysis on the Recidivism data. Report what you have learned through this activity. Prepare the data as best you can for an upcoming MDS analysis.*

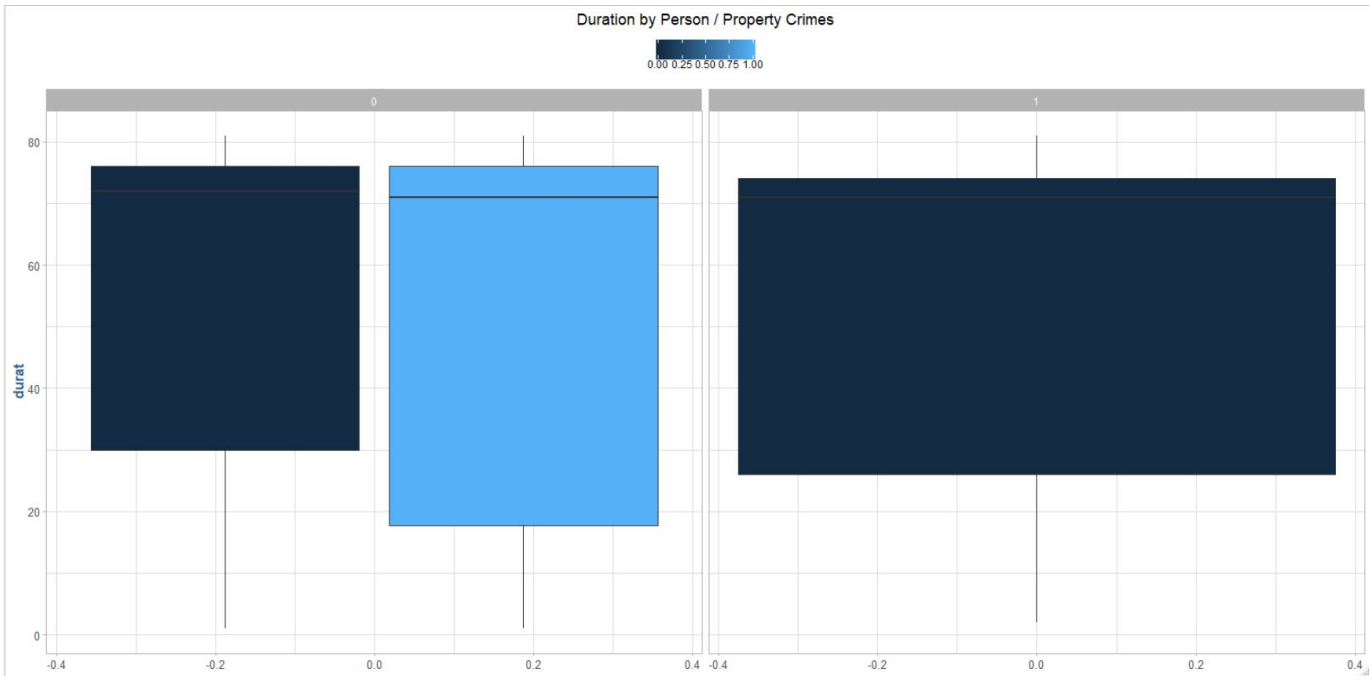
The first variable of interest in this data set is the duration variables, which represent the amount of time elapses until a person ultimately returns to prison once they have been released. We see a large clustering of values between 70 and 81 months, and we should note the maximum recorded value in this study is 81 months.



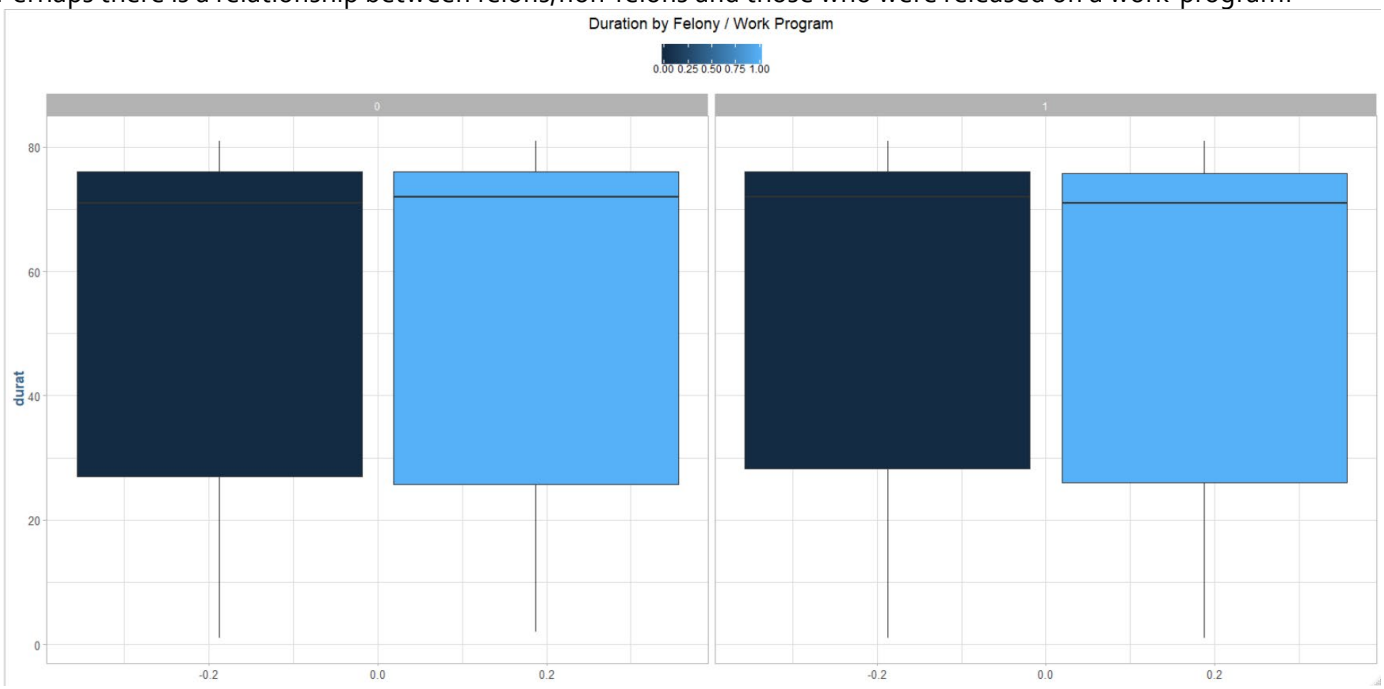
Let's look at some of the other explanatory variables to see if we can see any relationships that could help determine the duration. First, we're going to look at drug and alcohol use:



There is a large amount of variance in the inner quartile range, and the mean is right in the large tail area we saw previously. Next, we'll look at the type of crime committed to see if there is any relationship:

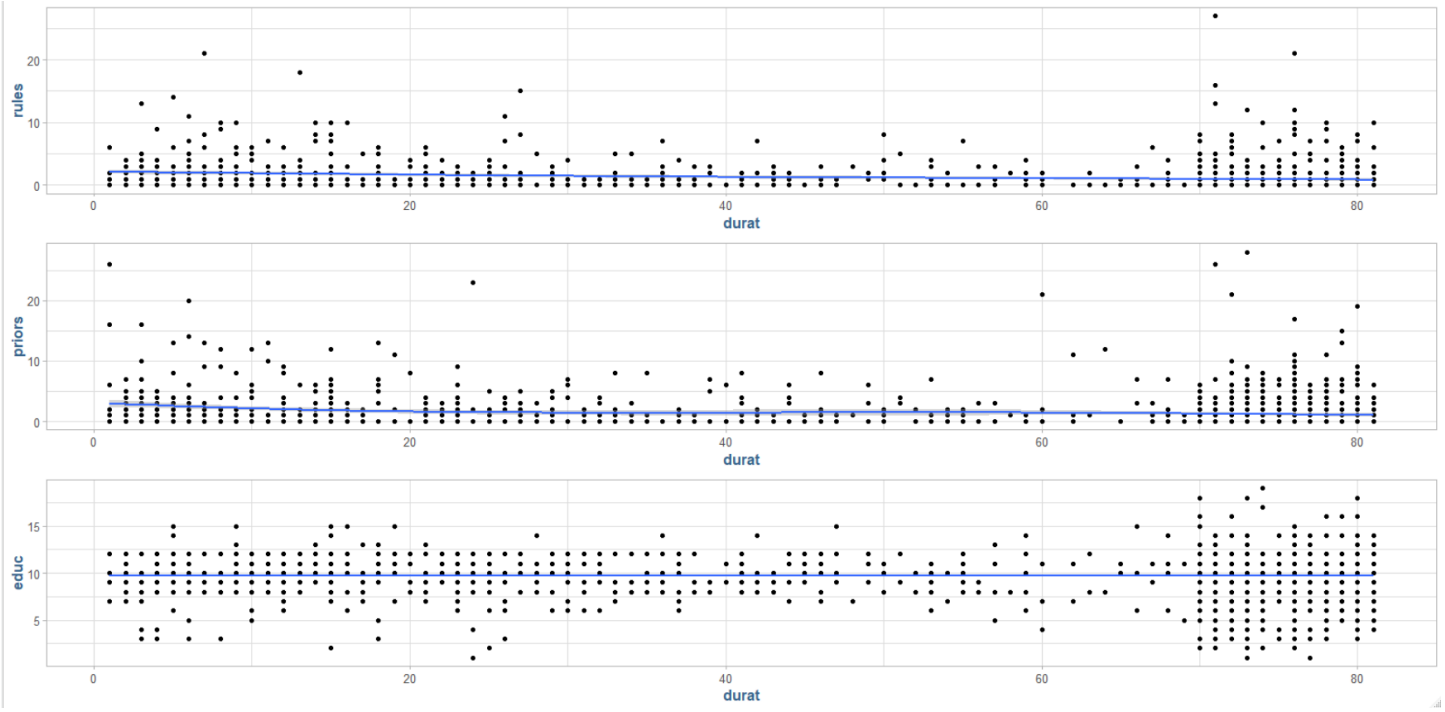


Again, we see similar behavior here as we did with drugs and alcohol, not much to discern the duration from. Perhaps there is a relationship between felons/non-felons and those who were released on a work-program?

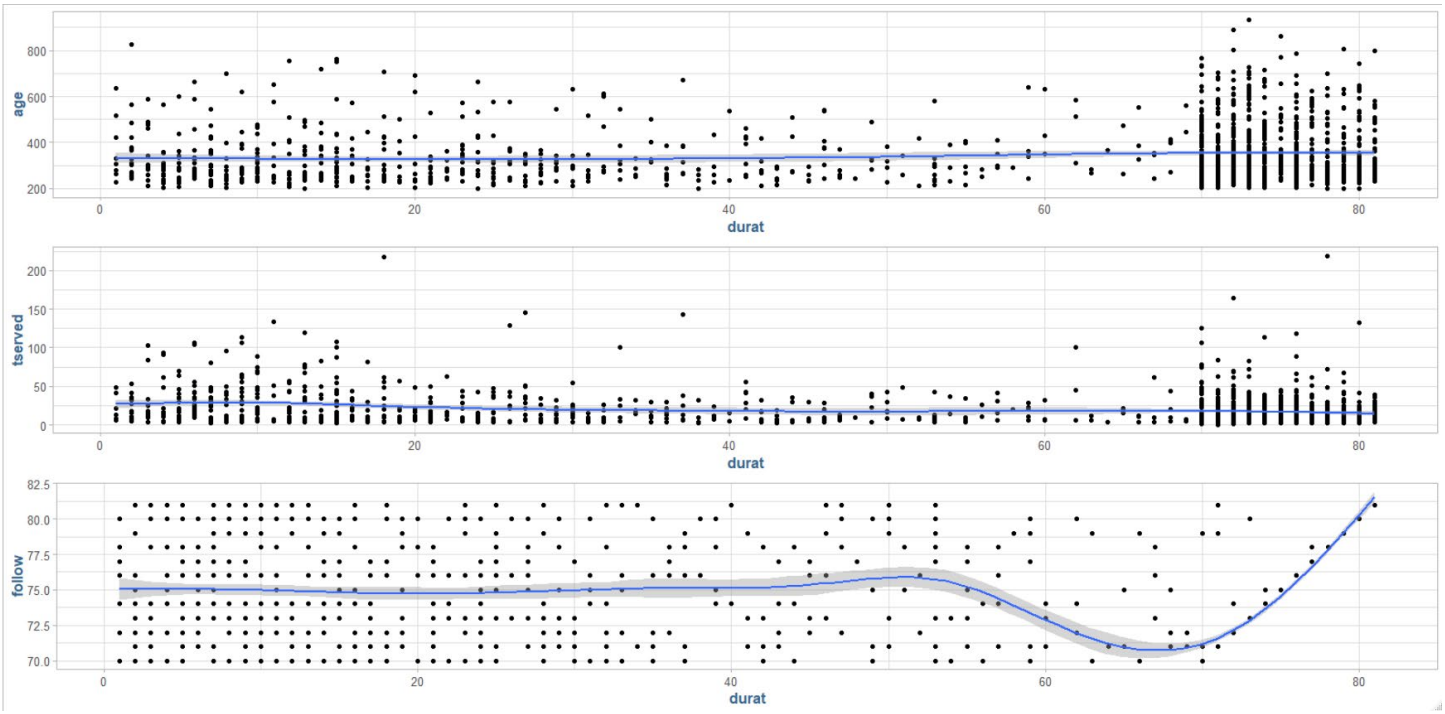


Again, we see similar results that do not explain the duration to useful degree.

The dichotomous variables do not seem to yield much explanatory information, let's look to some of the discrete variables and how they relate to the return duration.



Above we see the number of rules broken, number of priors, years of schooling and the only thing valuable we can see is that the average inmate has about 10 years of schooling. We'll finish off the discrete variable with age, time served and length of follow-up time, all measured in months:

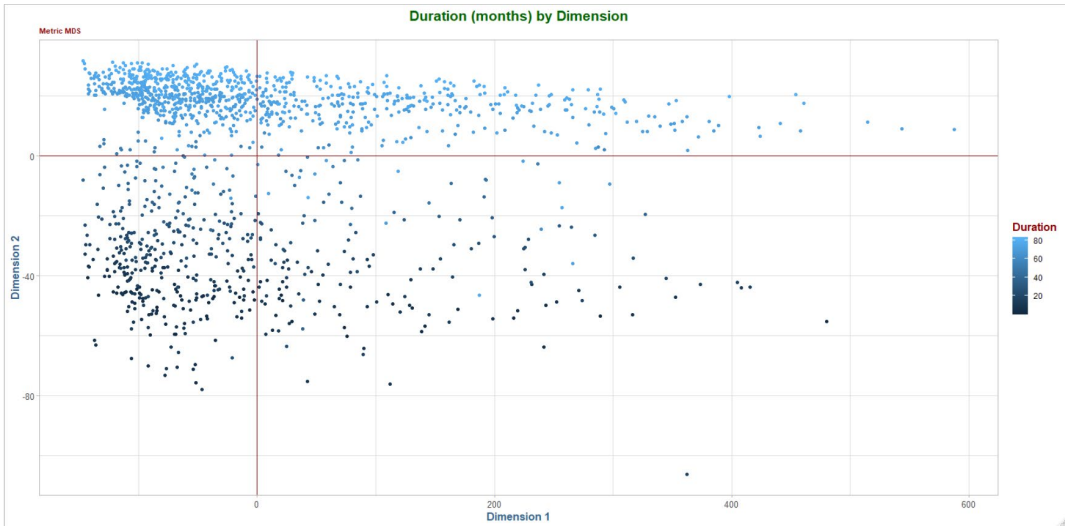


The most interesting thing we have found so far is that the length of the follow-up time seems to be uniformly distributed until past 70 months from release. After this period, it seems that their follow-up period is directly proportional to their return time. The exploratory data analysis did not yield a great deal of explanatory information that could be deemed actionable.

2) *Obtain a dissimilarity matrix using Euclidean Distances. There are a lot of cells in this matrix, but can you see any patterns at this point?*

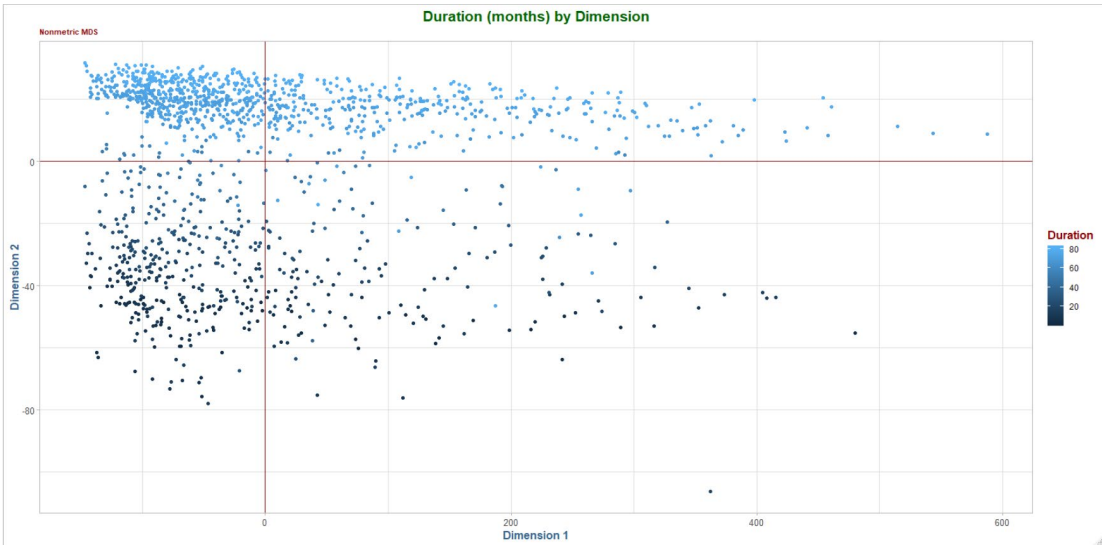
There are no obvious patterns discernable in the raw Euclidean distances. It seems to just be a regular column-wise increment.

3) *Conduct a classical multidimensional scaling using the Euclidean Distances dissimilarity matrix. Graph a 2-dimensional solution and interpret the result.*



There is prominent clustering behavior of the inmates with a longer duration of returning to prison. We explored almost all the variables in the data set in relation to this variable and saw no discernable patterns. However, with this multidimensional scaling dimensionality reduction approach, we are clearly able to see there are hidden relationships in the data as the inmates with longer return times seem to be heavily clustered in the upper left quadrant.

4) *Conduct 2 similar analyses using nonmetric scaling and Ramsey's method. Graph and interpret the two-dimensional solutions. How do these solutions compare with the classical approach?*



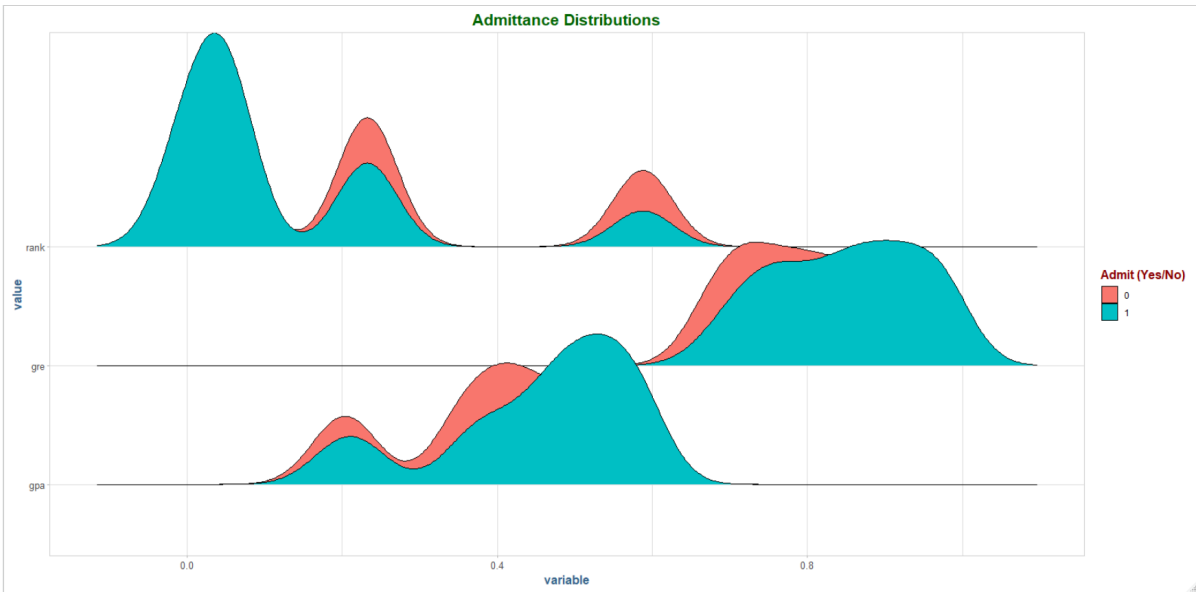
The non-metric multidimensional scaling is basically identical to the classical / metric version above. There are a few data points that change position, however, there are relatively few of these points that change coordinates.

ASSIGNMENT TASKS

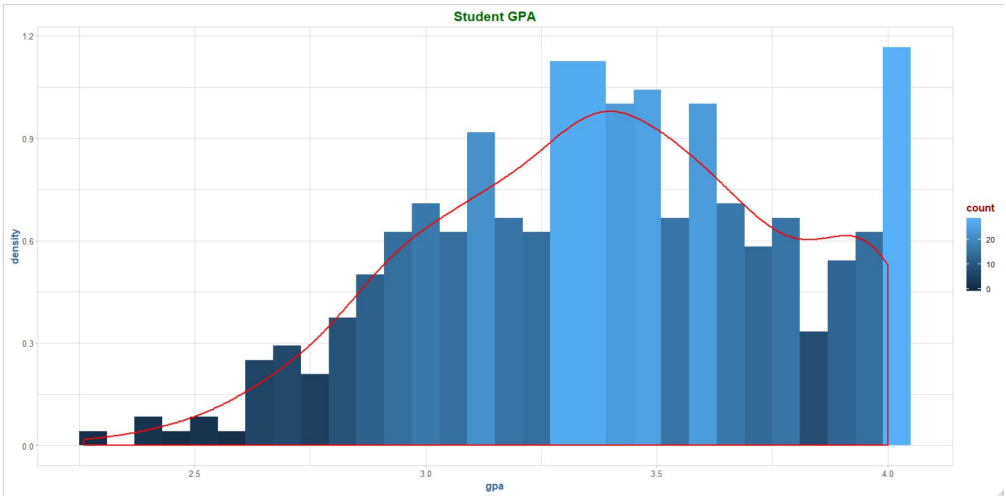
5) Exploratory Data Analysis [EDA] and Data Preparation for the College Acceptance Data.

- *Perform EDA on the data set and report your findings.*

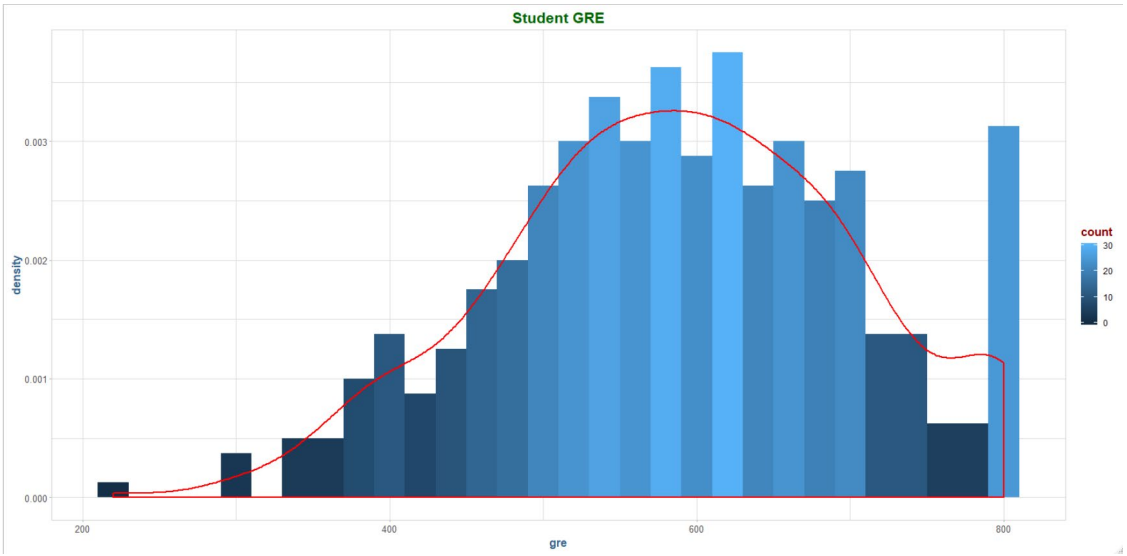
First, let's look at the distributions for the variables we would typically associate with college admittance that are available in the data set:



Let's look closer at the student's GPAs to see how they're distributed:

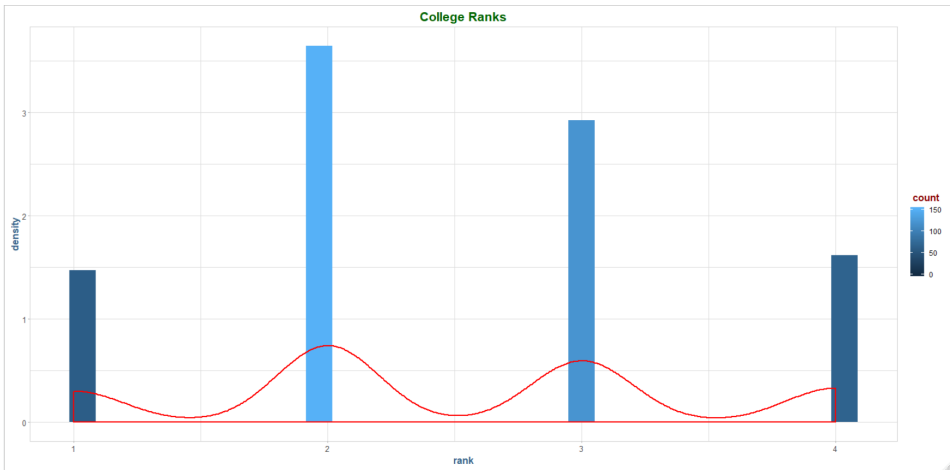


Now distribution of their GRE's:

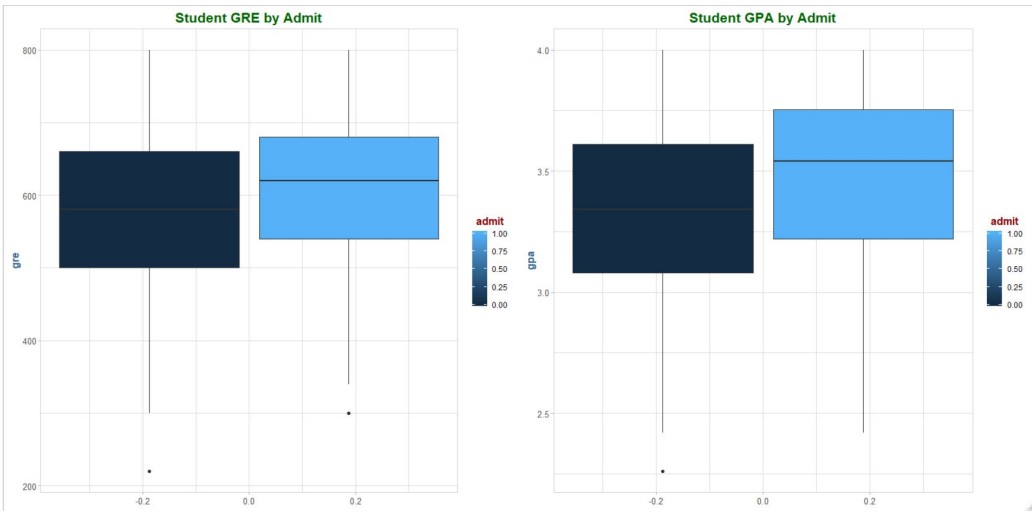


Immediately we notice that there are lots of maximum values here, where the GPA and GRE are capped at 4.0 and 800 respectively.

And finally, the college ranks:

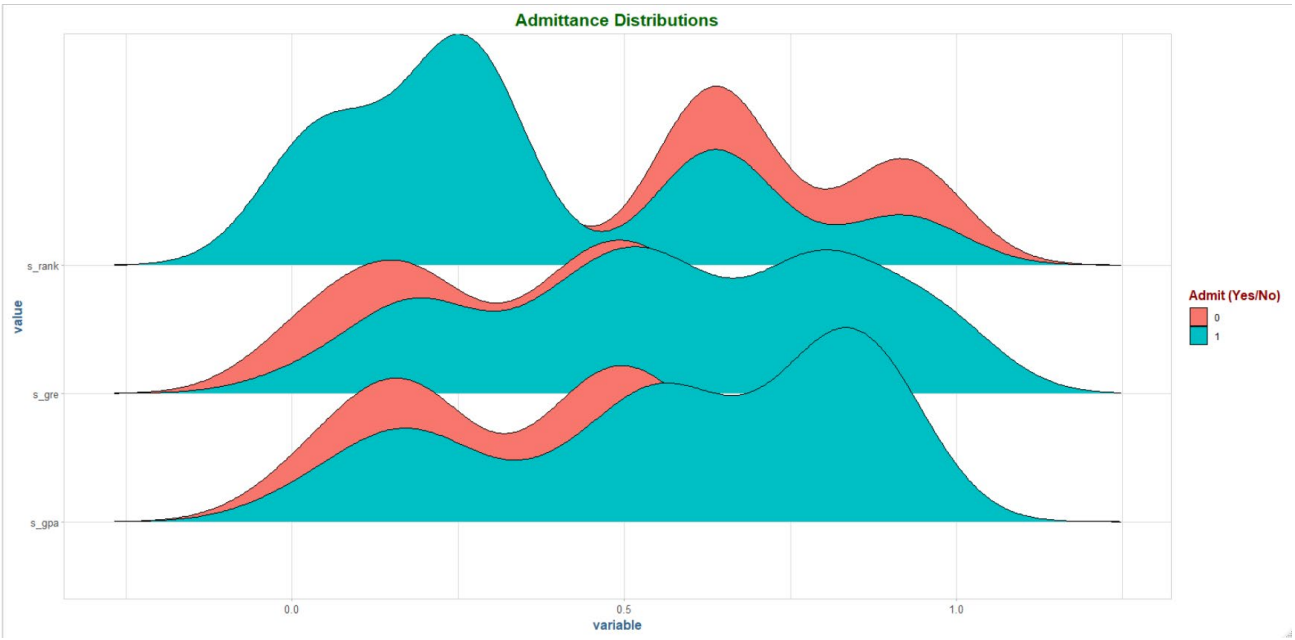


Let's separate the students into admit yes/no categories and compare their GRE/GPA:



- *Prepare the dataset for modeling as appropriate. Should scaling or normalization be applied? Why or why not?*

Yes, we need to scale the data in order to help ensure that no individual variable has too much influence in the mapping. We can see the post-scaled distributions maintain their respective shapes from above, and all the variables are now on the same scale. Admit will not be scaled, as it's a dichotomous variable.



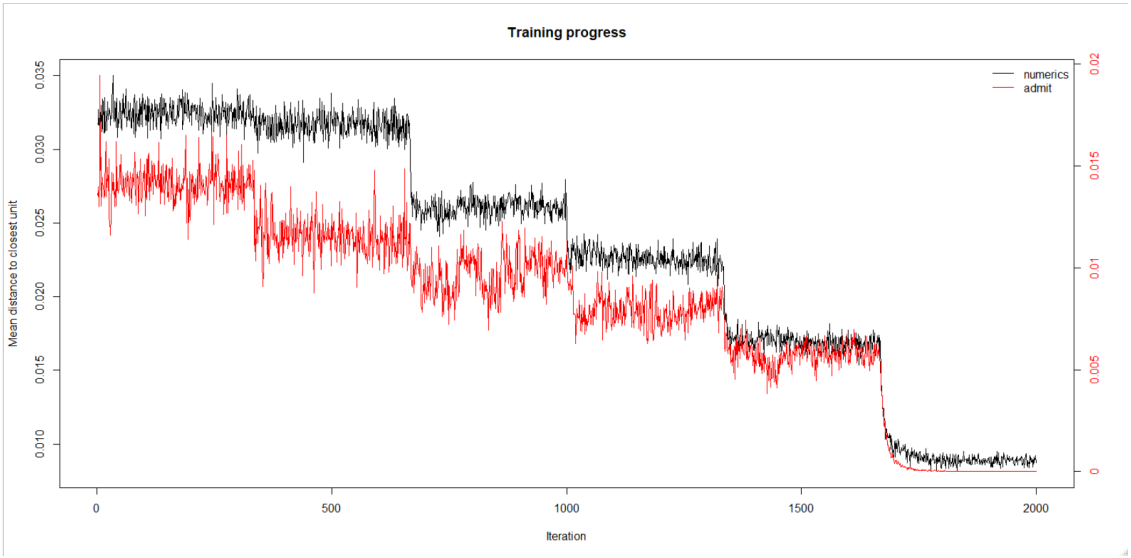
- Use only the variables provided in the dataset or variables you create by modifying or combining the variables provided.

I created 3 new variables, s_*, for the scaled versions of GPA/GRE and rank.

6) *Fit the SOM model. In the process you need to:*

- *Determine and report the number of epochs that will be used to train the model.* **We'll use 2,000 to start.**
- *Determine the appropriate grid size for the SOM. Report the method that you used.* **10x10 for a data set with 400 rows.**
- Fit the model using the R *kohonen* package or like the dataset that you prepared in *PART A*. Use the grid size and epochs that you selected in 1 and 2.

Seed = 123.

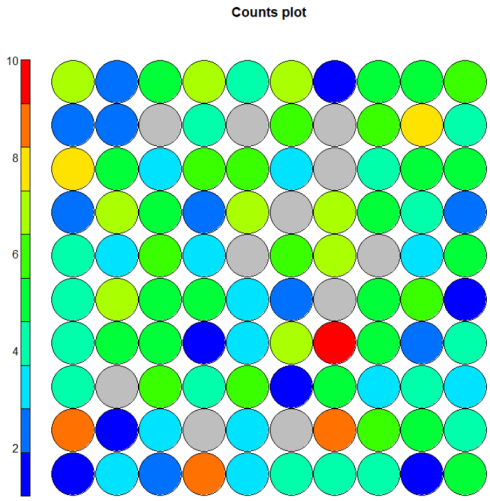


7) *Evaluate the SOM model. To do this you need to address the following:*

- Was the epochs value selected in PART B adequate to train the model? Include a copy of the visualization that was used to make that determination. If the model needs additional training, adjust the epochs value and retrain the model before continuing.

  Yes, 2,000 epochs was enough to reach a minimum plateau of zero at around ~1,800 runs.

- Was the grid size selected in PART B adequate? Explain why the grid size was or was not adequate and attach the visualizations used to make that determination.

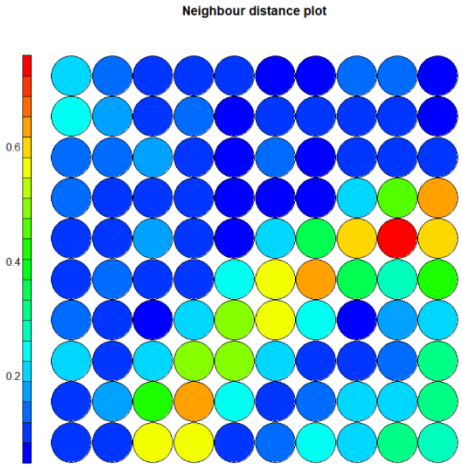


  There is a lot of grey in the counts map, suggesting that the grid is too big. There is one area with a high count, but there are not enough of these to suggest that the map is too big.

- What is the average number of observations assigned to the nodes?

  Average node count: 4.5

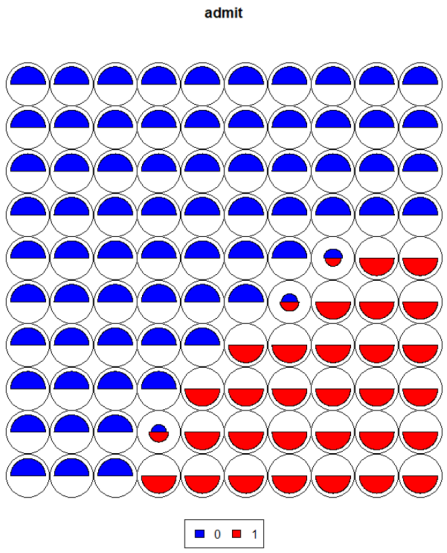
- Generate a distance map and attach a copy of it here.



  Are any nodes quite distant from their neighbors?

  The cluster in the middle right of the plot oriented around the one red dot seem to be fairly distant. Also, there are two yellow nodes in the bottom right that could also quality.

- Generate a *codes* plot and attach a copy of it. Discuss what this plot tells us about the applications and college acceptance.

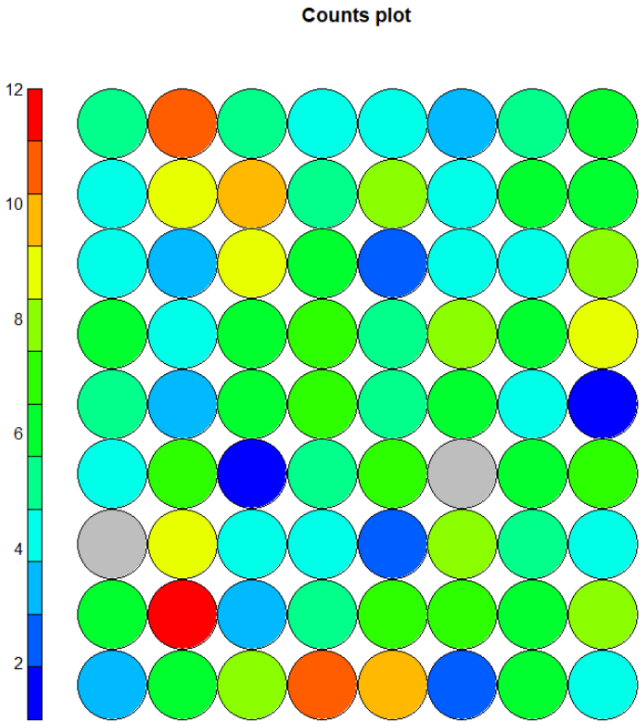


It seems that, as we would expect, the students who got accepted have similar GRE/GPA scores and college rank criterion, as they cluster together. There are some who got admitted who appear to be somewhat outliers, in the last diagonal. They could have been accepted with lower scores or had higher scores and got into a lower rank.

8) Experiment with the SOM model. To accomplish this task, you will need to:

- *Change the grid size for the SOM and retrain the model. Discuss whether you increased or decreased the grid size and why.*

  I decreased the size of the grid, there was a lot of grey/empty areas for the previous run. I changed the grid size to an 8x9 matrix, deciding to not keep symmetry because the data seems to fit an asymmetric matrix better (the low dimensionality perhaps). The resulting plot has 2 gray areas and 1 red, 2 orange. Seems like a fit.

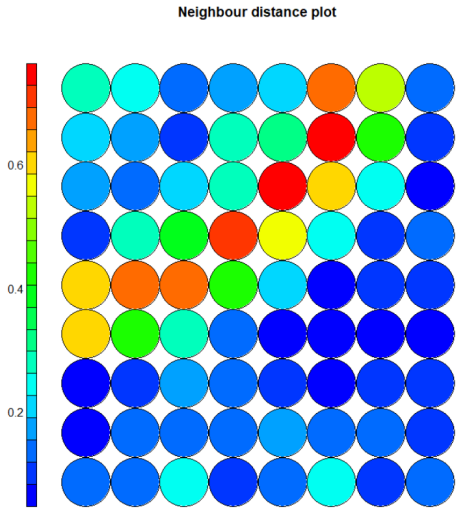


- *Compare this new SOM to the SOM created in PART B. Does the new grid size improve the SOM? Discuss how grid size impacts the SOM.*

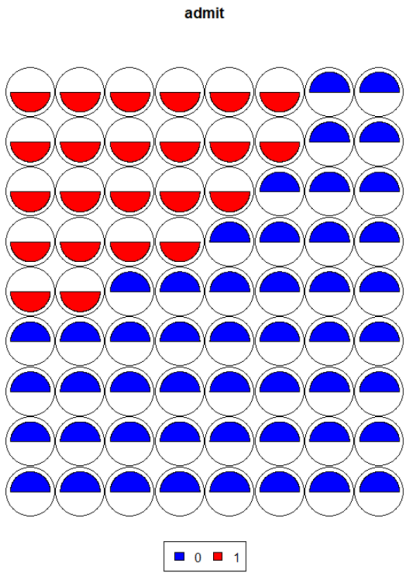
  The mean observations per node increases to 5.71 with the new grid size.

  The distances plot has more "red" dots overall, however, the colors are more concentrated in a few specific areas and overall display less dispersion.

- *Generate a distance map and attach a copy of it here. Are any nodes quite distant from their neighbors?*



- *Generate a codes plot and attach a copy of it. Discuss what this plot tells us about the applications and college acceptance.*



There is a clearer separation of those who were admitted and those who were not in this codes diagram than the previous. We don't see the nodes mixing admit/non-admit together, and the attribute weights appear to be clustered appropriately.

CONCLUSION

9) Please write a reflection on your MDS and SOM modeling experiences.

This assignment was an informative one, in that it covered two drastically different unsupervised learning techniques to uncover hidden relationships in the data. The first example was much more interesting to me in that, by traditional methods of exploratory data analysis there truly seemed to be no connection to the response variable, return duration. There are a great deal of attributes in that data set, and the level of dispersion in all of them in how they relate to the return duration was a frustrating endeavor. The unsupervised method was clearly able to pick up on some latent trait, and at least split the data into two distinct parts: those above 0 where the group that were ~65-80 months to return, while those below zero were 0-65.

The college admissions data was less interesting in general as the relationships are intuitively known before the analysis started, however, the technique of building a self-organizing map is indeed an interesting one. I think a higher dimensional data set would have been interesting here, although the basic structure of the data did make it easier to focus exclusively on the modeling technique at hand. The grid size parameter was of particular interest, as I tried many combinations after the original suggestion of a 10x10 was clearly too big, however, trying to keep the matrix symmetrical (I suppose just for the sake of mathematical purity) did not yield overall higher quality results. The smaller grids packed too much information and then relationships started to get lost in the resulting model. Definitely a goldilocks situation with the grid size toggling.

Overall, this lab was an informative and enjoyable one. I would feel comfortable at least having a high-level conversation about these topics having performed it, as I had never been introduced to either of these methods in the past. Time well spent.