Assigment #3

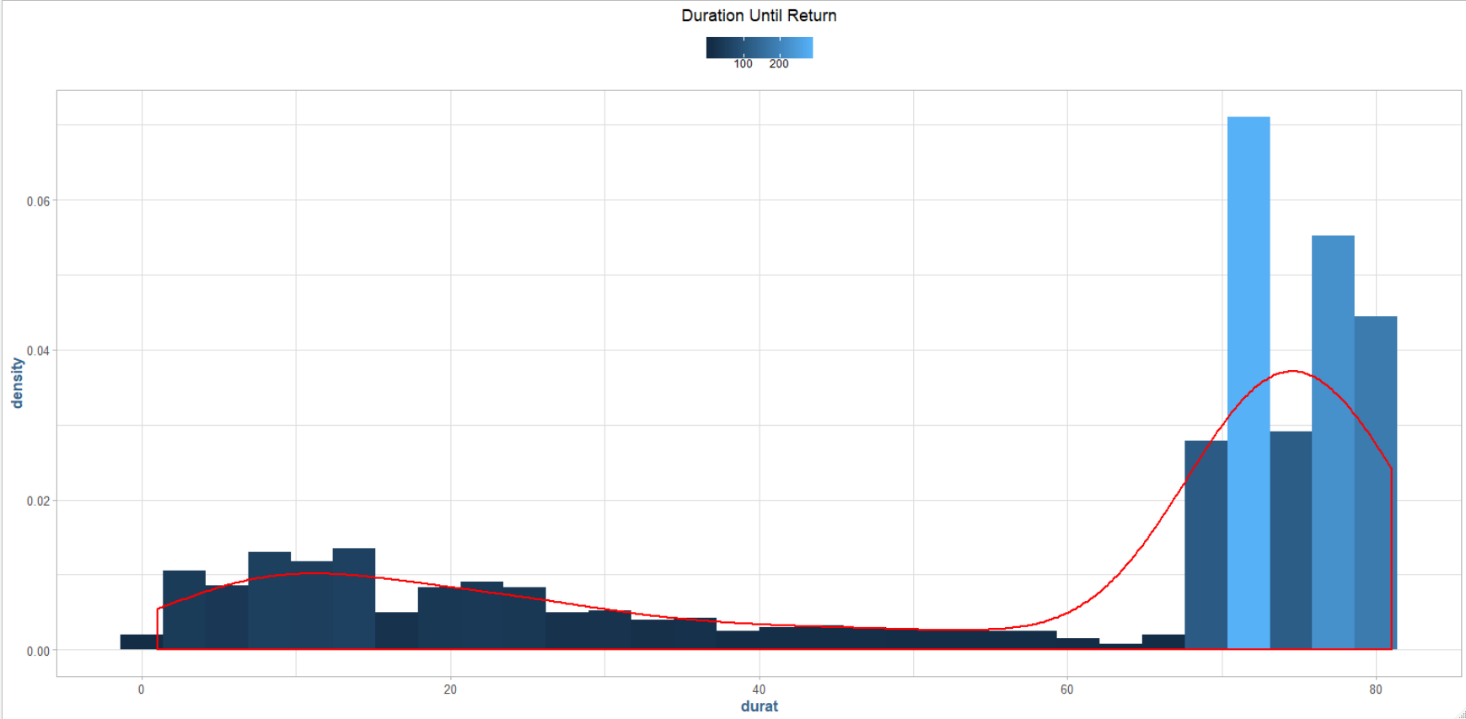
Brandon Moretz

### Multidimentional scaling

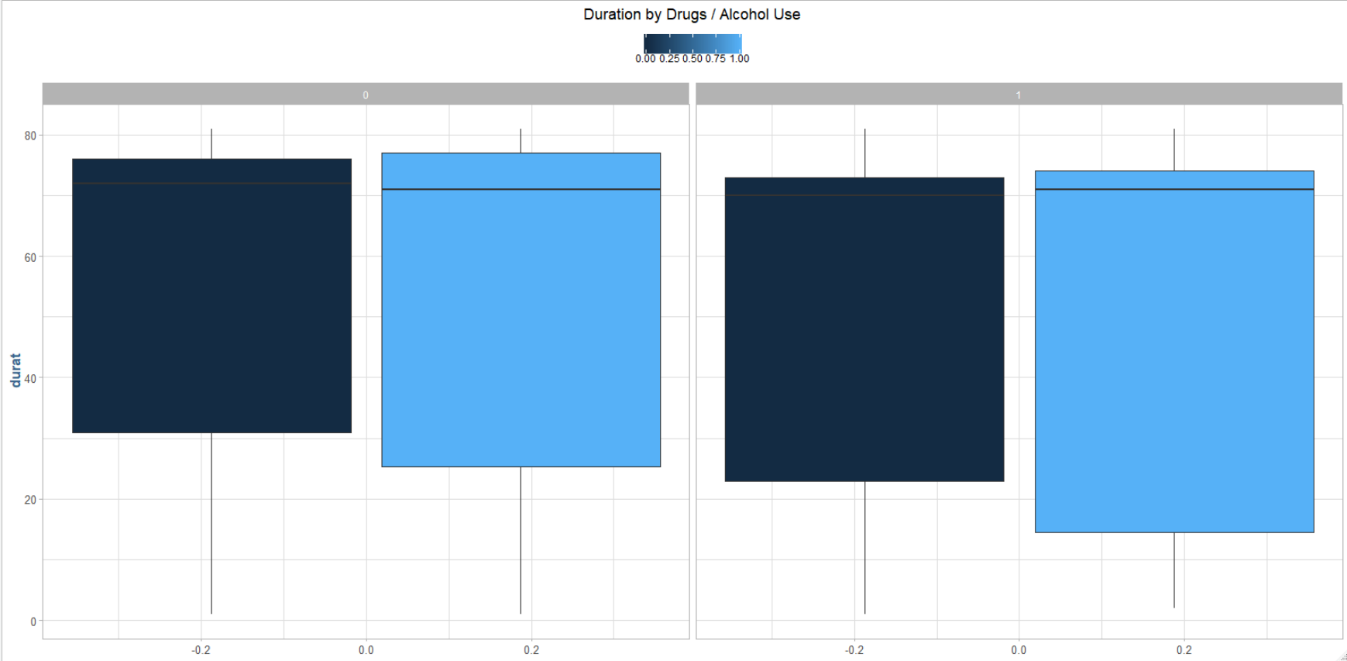
#### Assignment Tasks

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.*

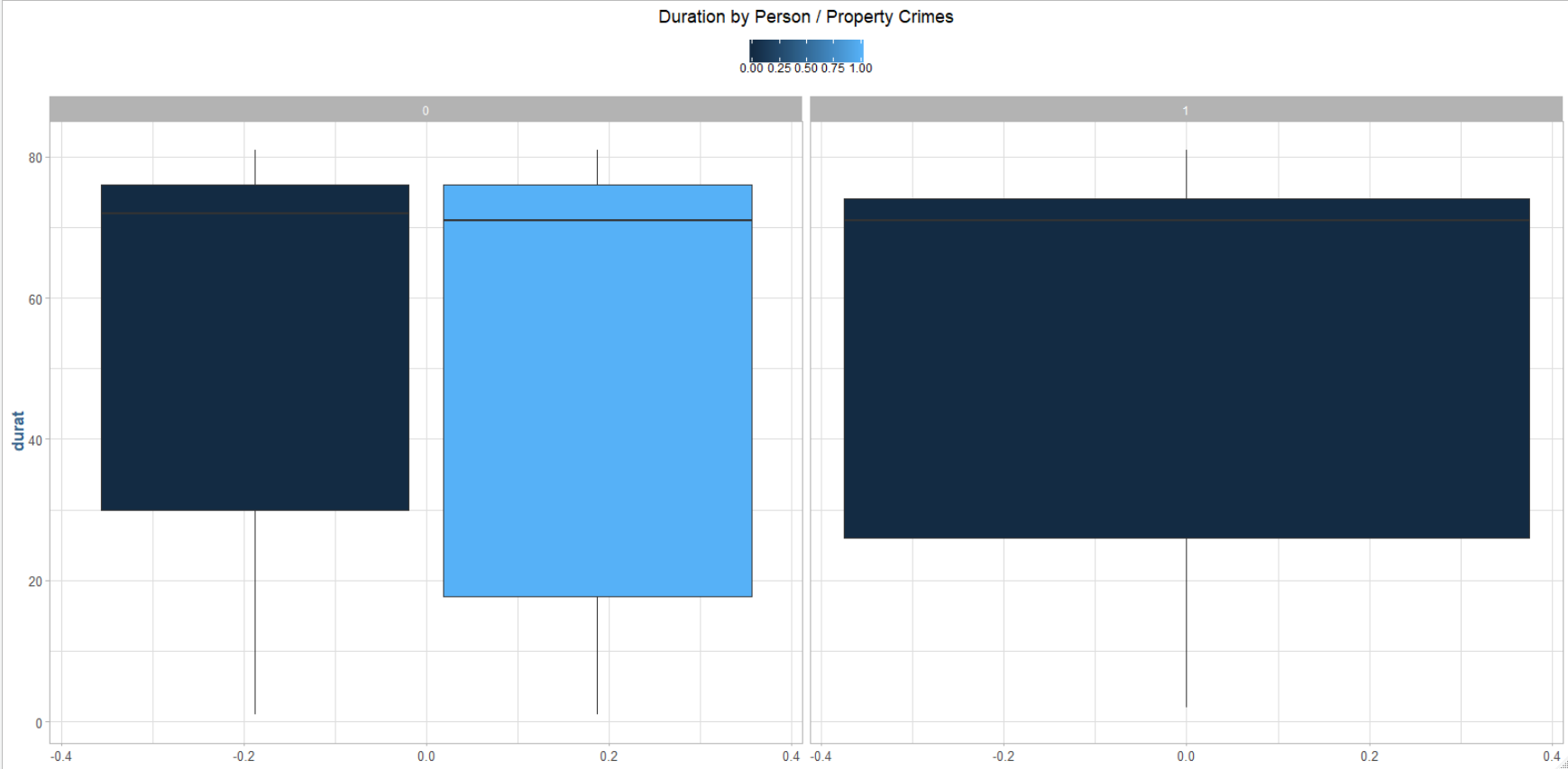
The first variable of interest in this data set is the duration variables, which represent the amount of time elapses until a person ultimately returns to prison once they have been released. We see a large clustering of values between 70 and 81 months, and we should note the maximum recorded value in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



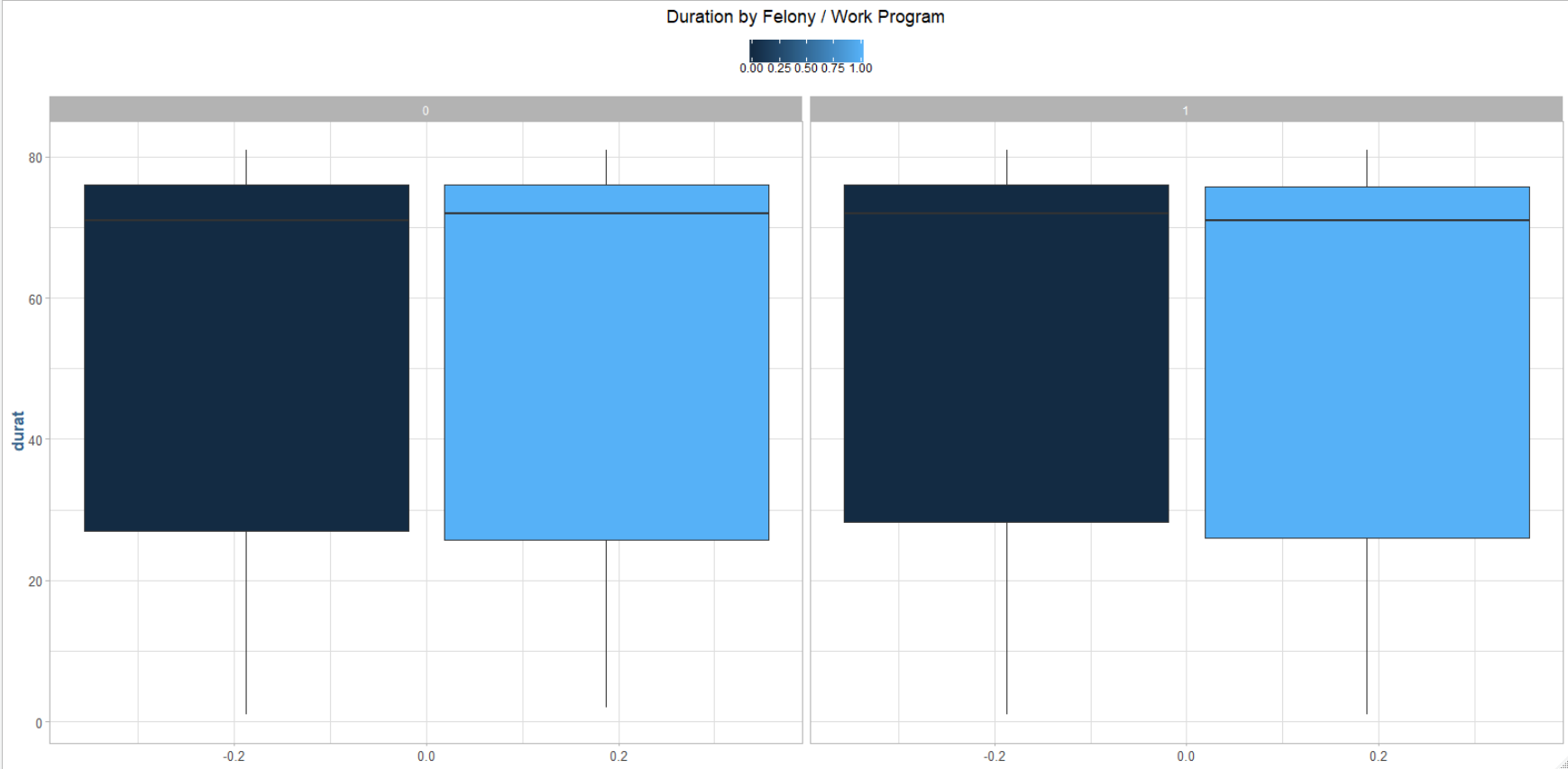
Let’s look at some of the other explanatory variables to see if we can see any relationships that could help determine the duration. First, we’re going to look at drug and alcohol use:



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:

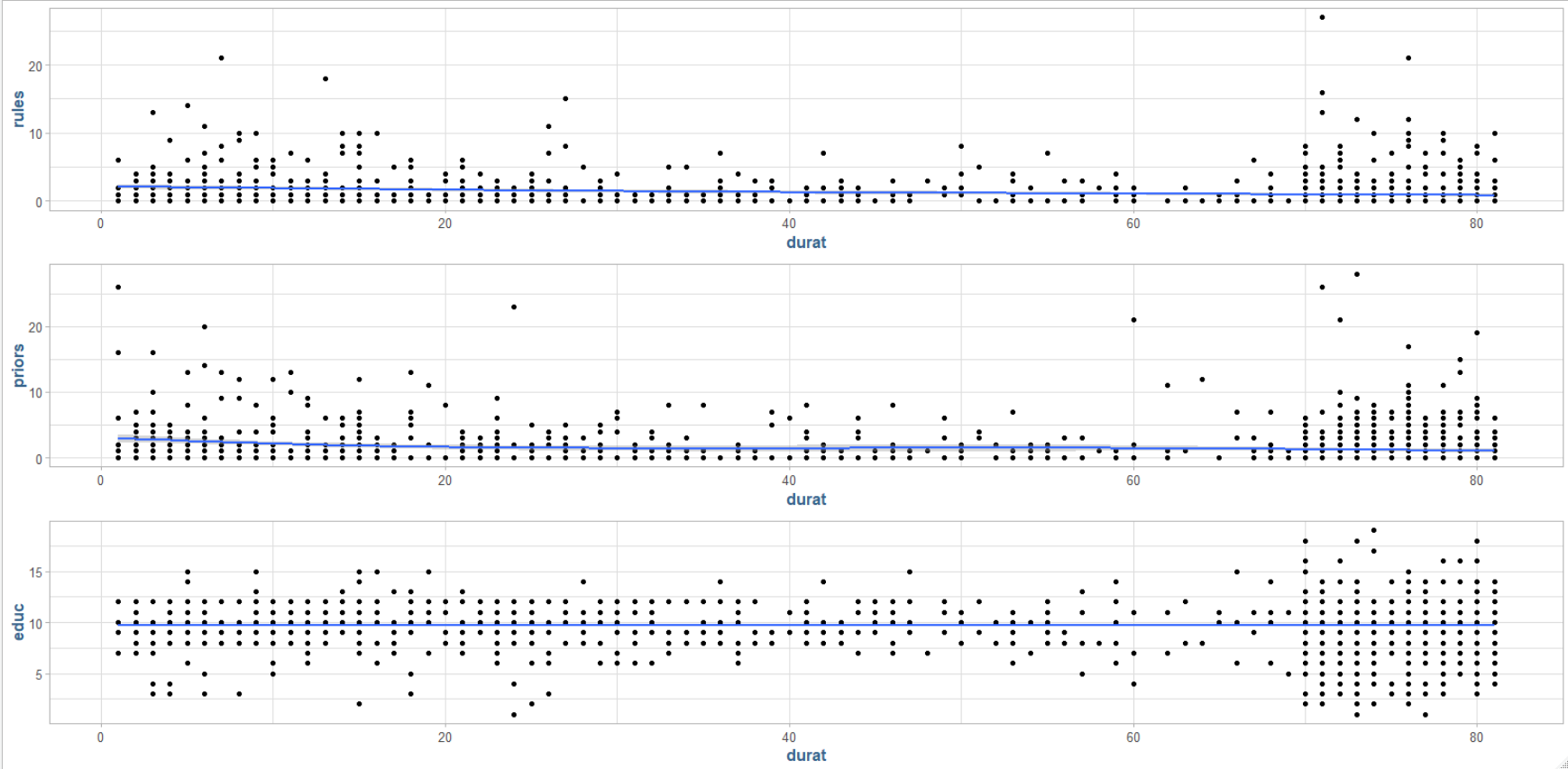


Again, we see similar behavior here as we did with drugs and alcohol, not much to discern the duration from. Perhaps there is a relationship between felons/non-felons and those who were released on a work-program?

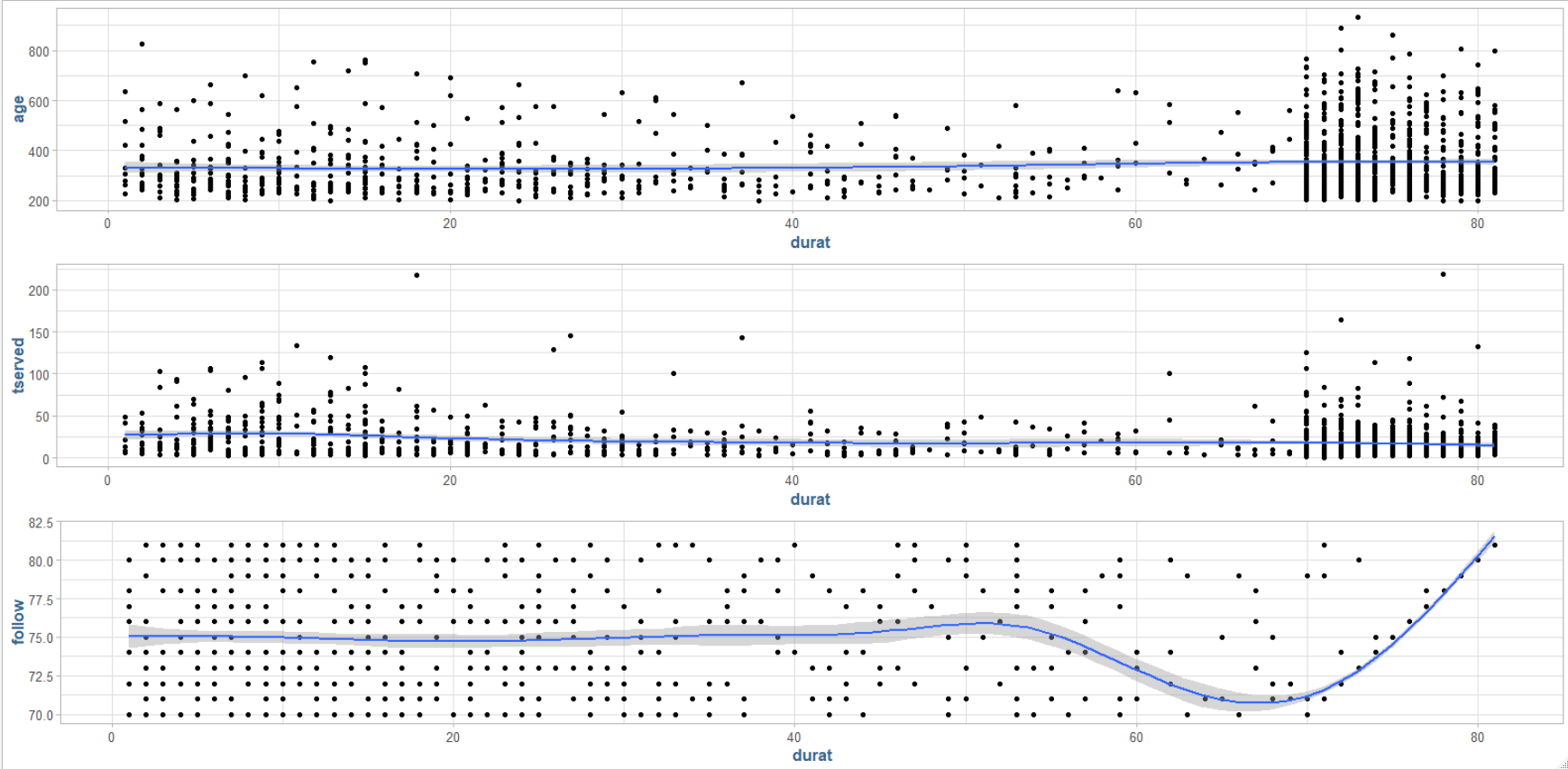


Again, we see similar results that do not explain the duration to useful degree.

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:

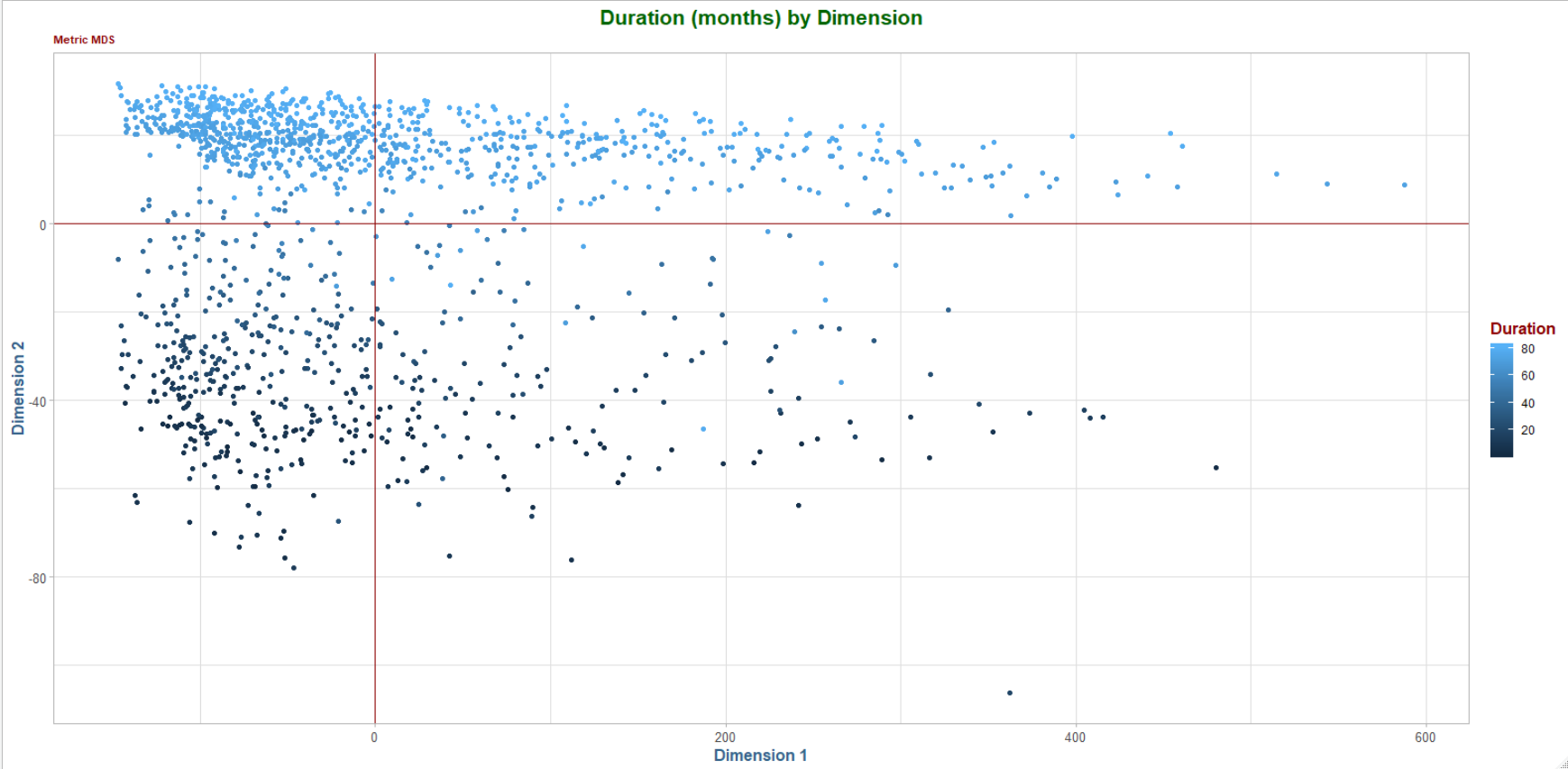


The most interesting thing we have found so far is that the length of the follow-up time seems to be uniformly distributed until past 70 months from release. After this period, it seems that their follow-up period is directly proportional to their return time. The exploratory data analysis did not yield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

1. *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?*

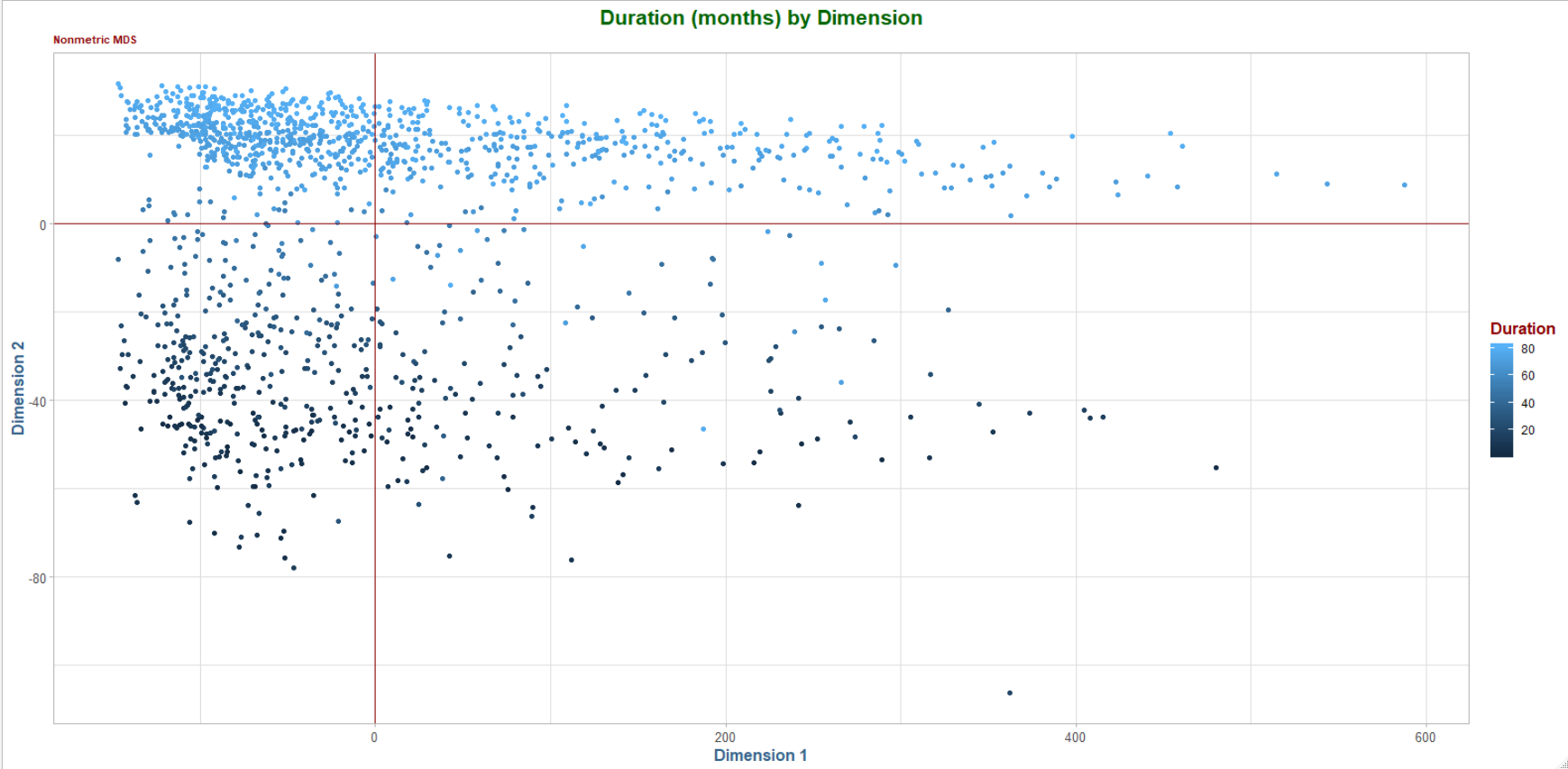
There are no obvious patterns discernable in the raw Euclidean distances. It seems to just be a regular column-wise increment.

1. *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.

1. *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 actually change coordinates.

### Self organizing maps

#### Assignment Tasks

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

* Perform EDA on the data set and report your findings.
* Prepare the dataset for modeling as appropriate. Should scaling or normalization be applied? Why or why not?
* Use only the variables provided in the dataset or variables you create by modifying or combining the variables provided.

1. 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.
* Determine the appropriate grid size for the SOM. Report the method that you used.
* Fit the model using the R *kohonen* package or similar to the dataset that you prepared in *PART A*. Use the grid size and epochs that you selected in 1 and 2. Be sure to set the seed before fitting the model so that the results may be reproduced.

1. 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.
* 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.
* What is the average number of observations assigned to the nodes?
* 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.

1. 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.
* 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.
* 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.

### Conclusion

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