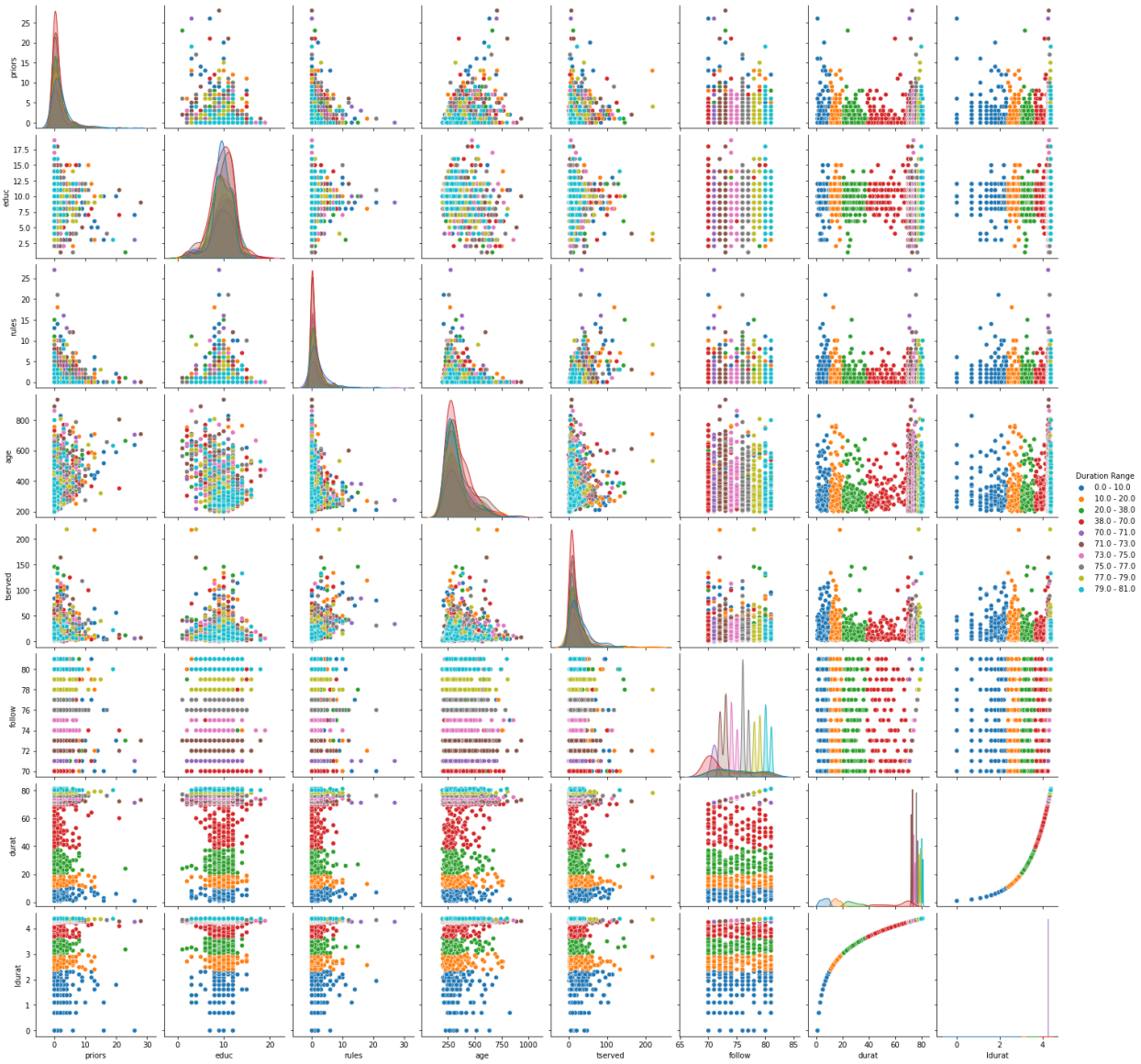
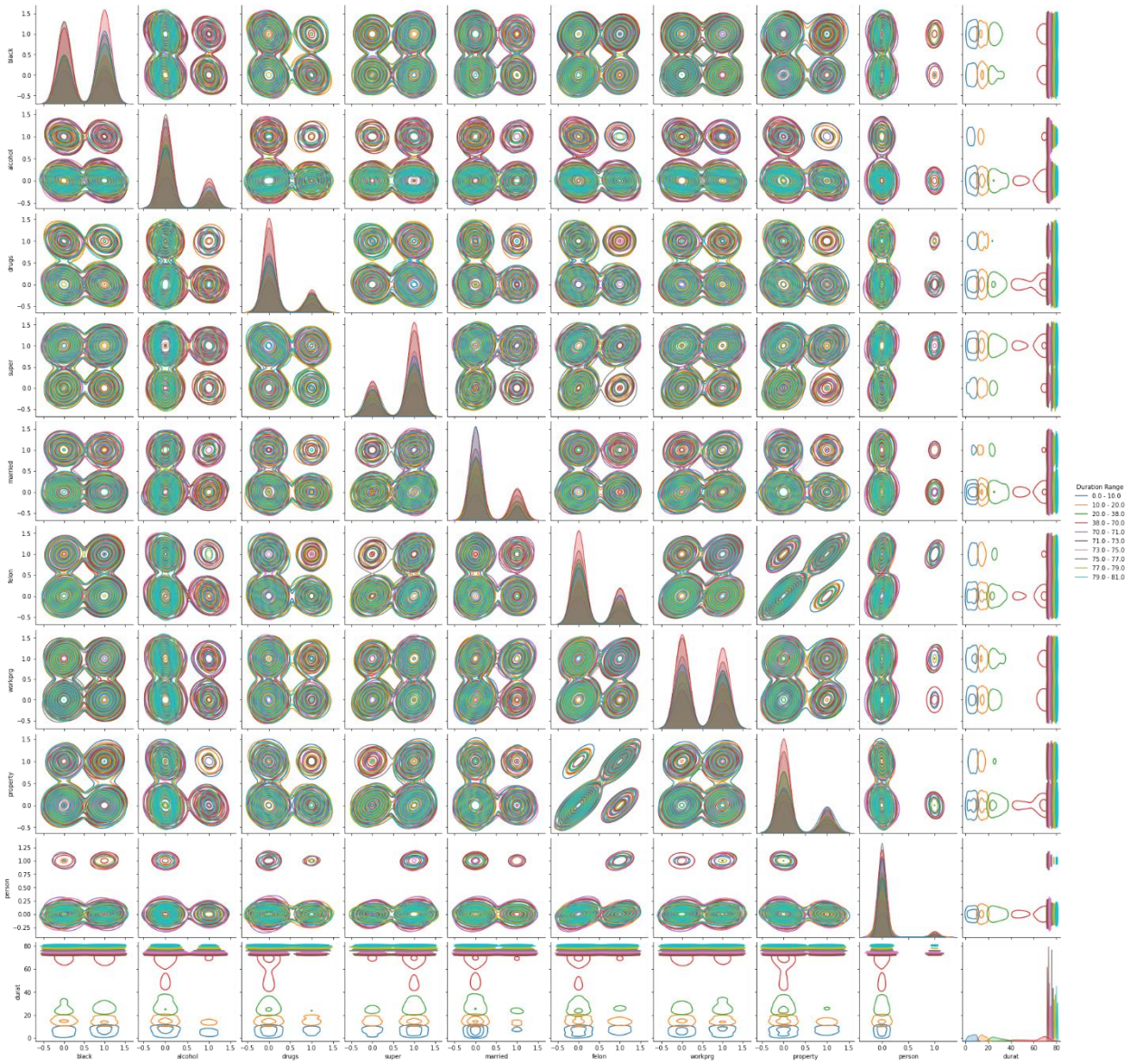
Michael Rocchio

MSDS 41 - Assignment 3

Component 1

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.





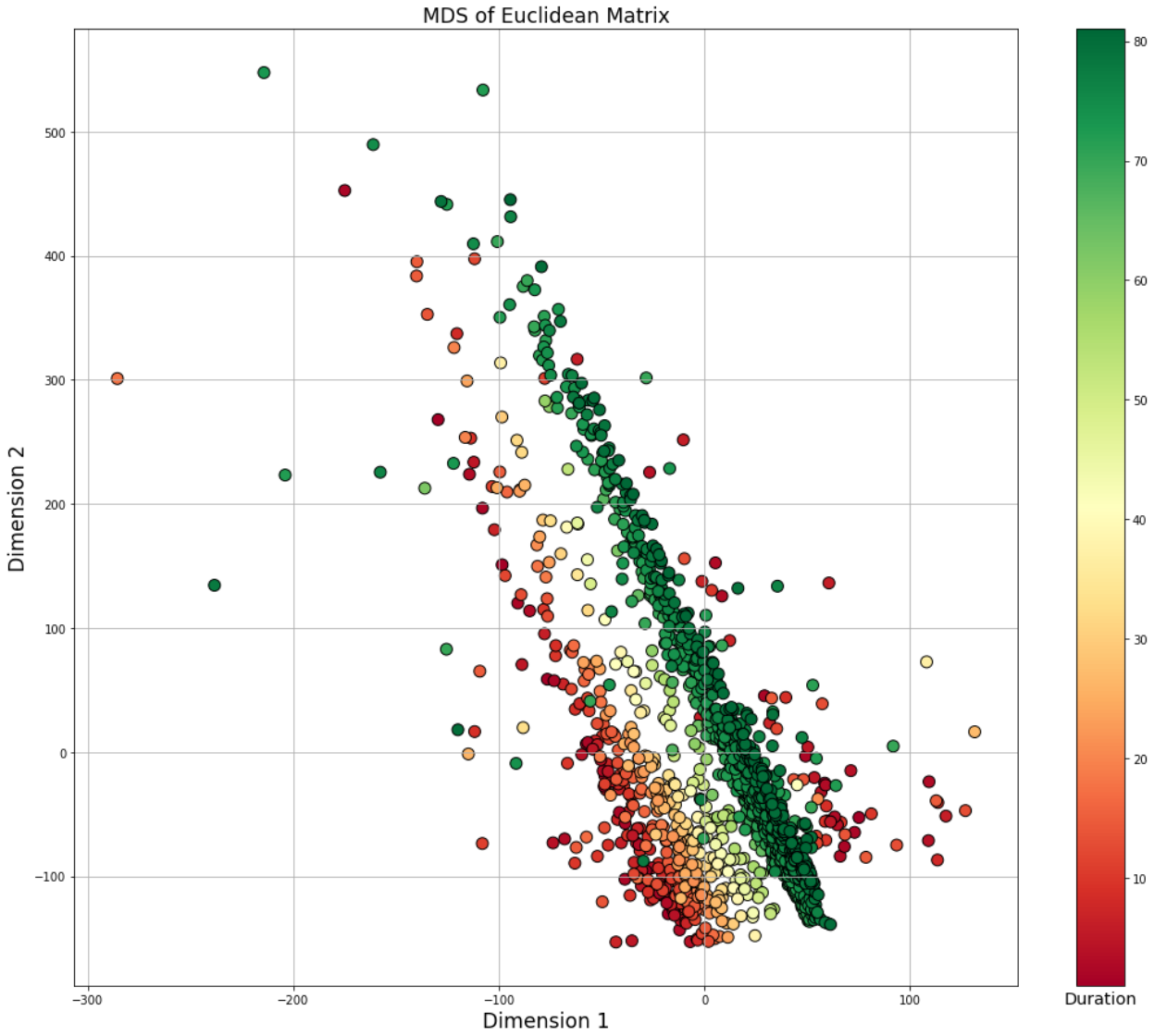
During this activity I was primarily concerned with the distribution of the data, specifically by the explanatory variable's groupings. To complete this I did a pair plot of the data for each of the types, continuous and Boolean.

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?

I was unable to find any meaningful patterns while graphing the data after a few attempts.

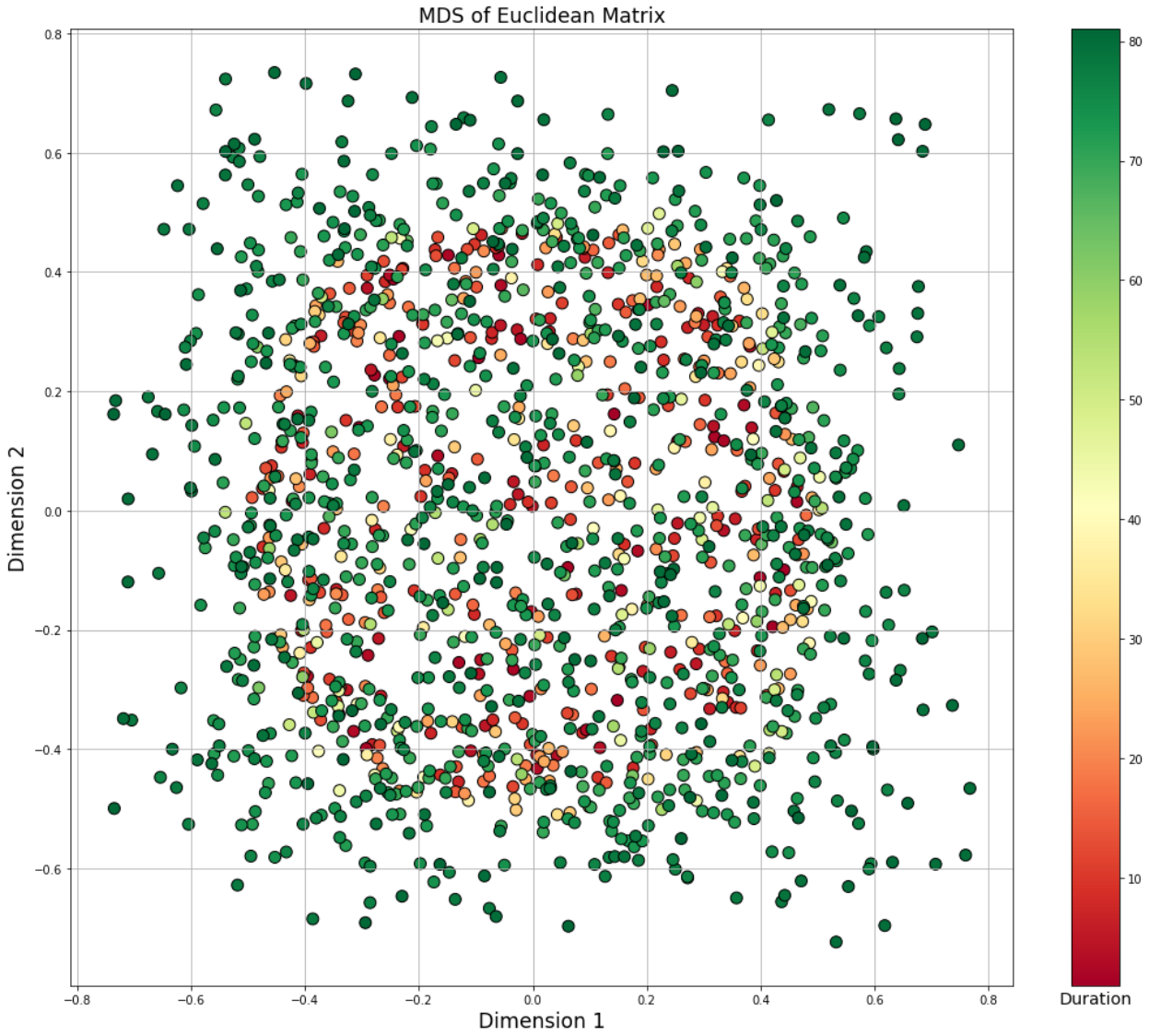
3) Conduct a classical multidimensional scaling using the Euclidean Distances dissimilarity matrix.

Graph a 2-dimensional solution and interpret the result.



There is quite a lot of vertical clustering towards the various duration times, specifically towards Dimension 1. Dimension 2 seems to have a much amount of impact.

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?



I am unsure if I coded this incorrectly, but this seemed to scatter the data in an uninterpretable manner.

There seems to be no correlation between the two dimensions.

<u>Component 2</u>

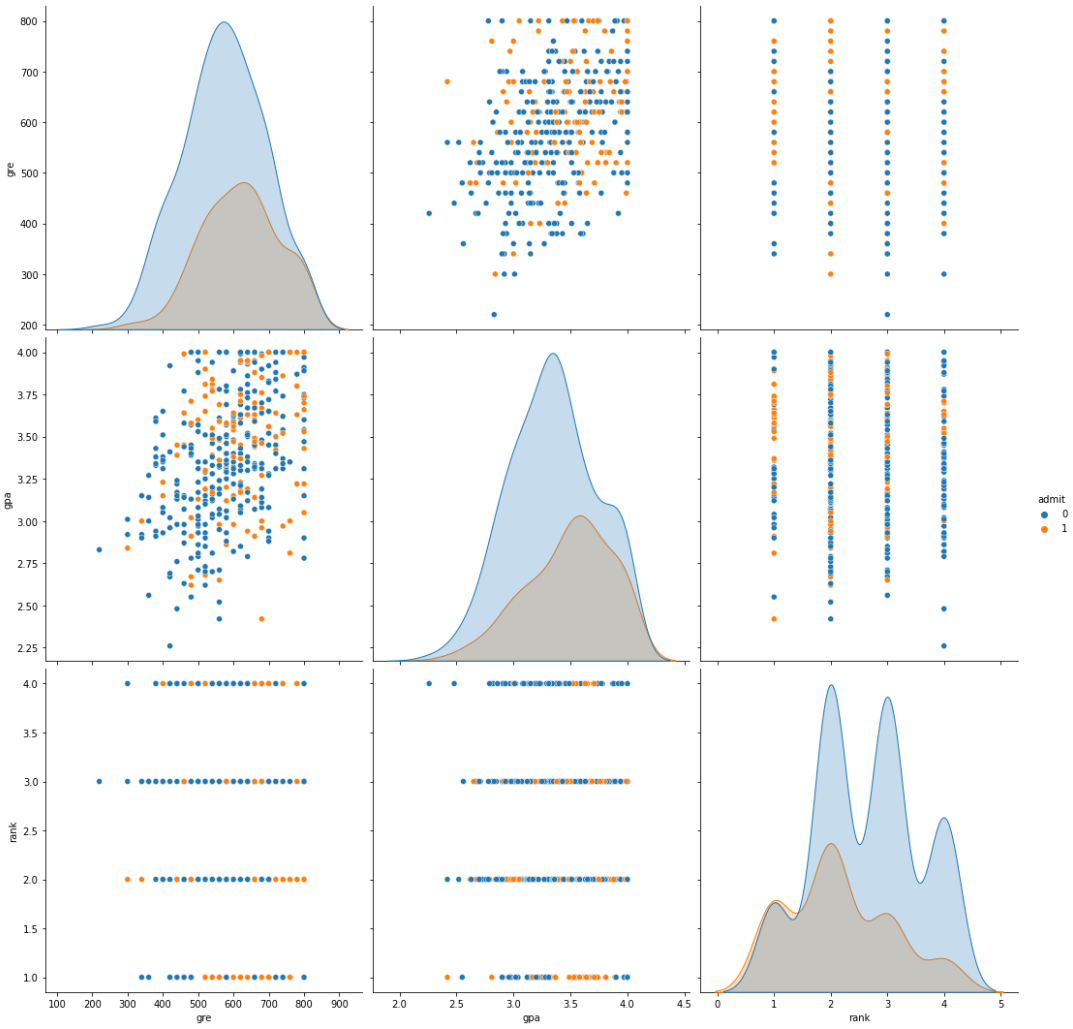
5) Exploratory Data Analysis [EDA] and Data Preparation for the College Acceptance Data.

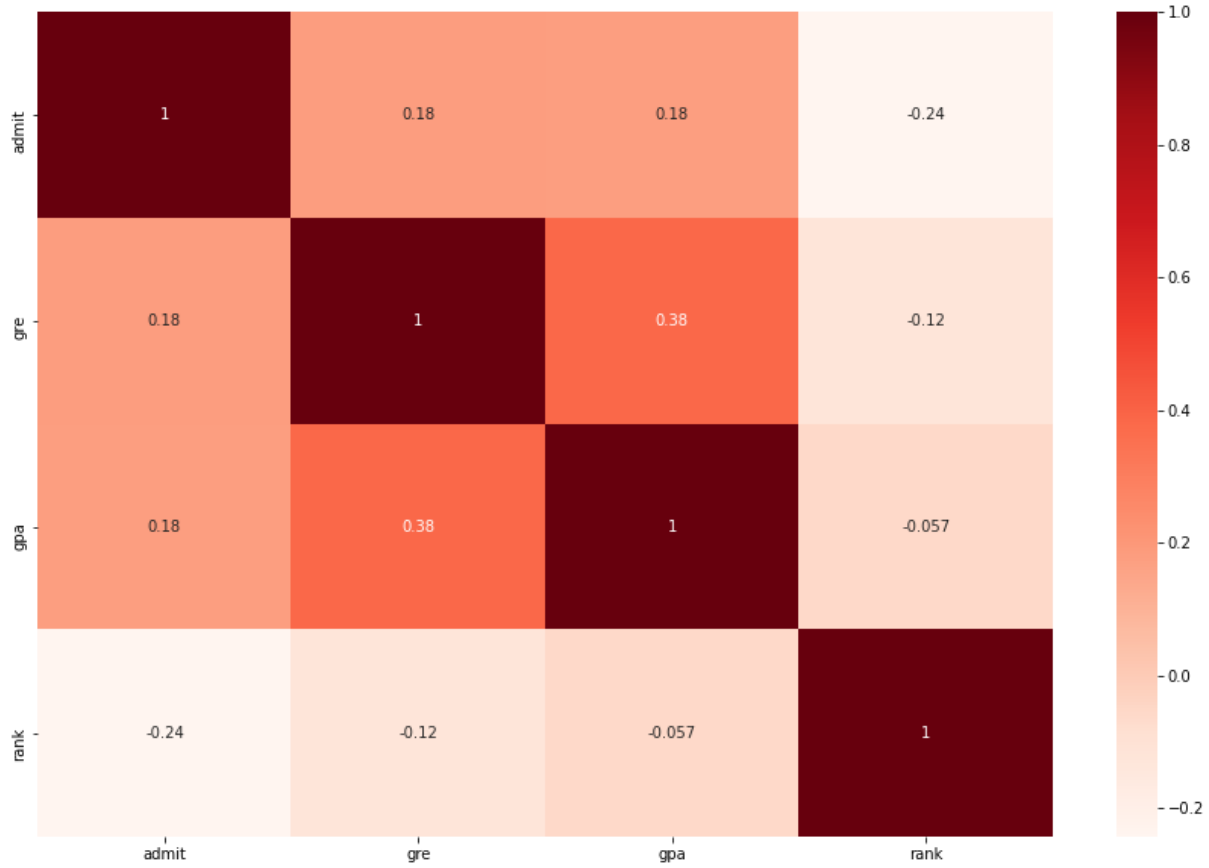
Perform EDA on the data set and report your findings.

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.





As the first pair plot shows there are some variables with a much that have a much stronger correlation than others. In order to prevent the model from being weighted incorrectly we will need to scale the explanatory variables.

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.

I used 10,000 epochs to fit the model but only 8,000 were required to achieve the desired result.

Determine the appropriate grid size for the SOM. Report the method that you used.

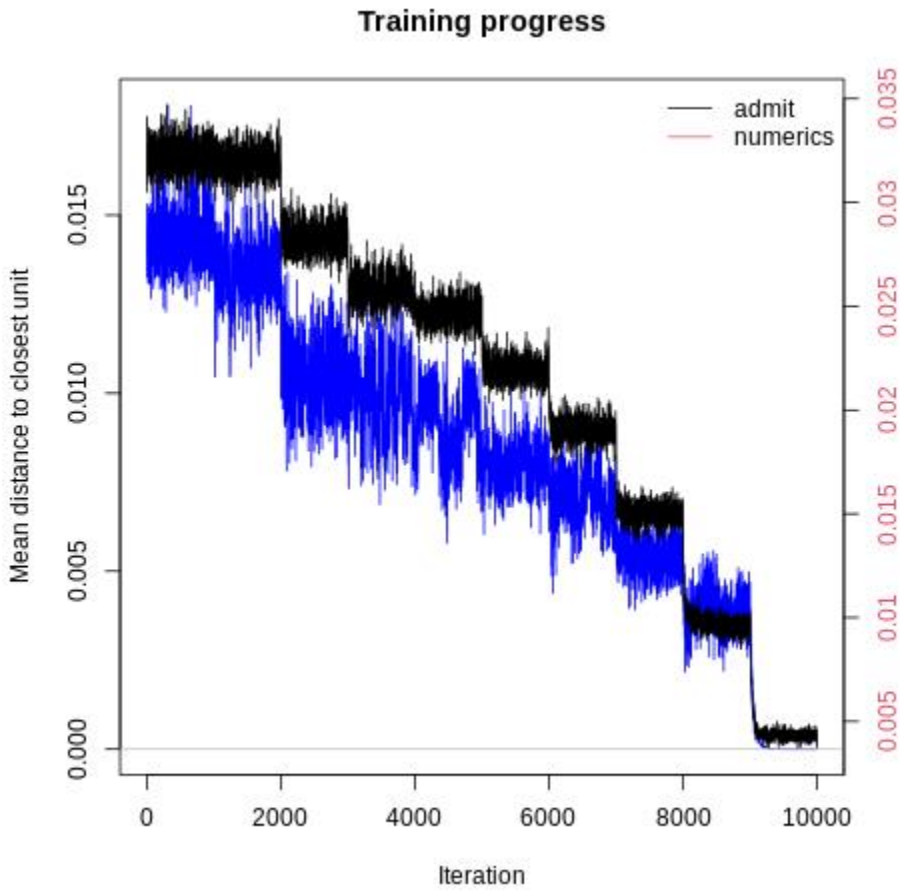
I utilized a 16 x 18 grid after attempting numerous other iterations.

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.

The seed was set to 42 and I fitted the model*

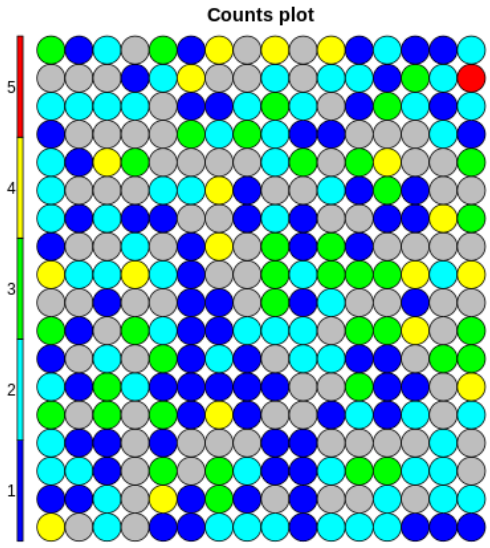
7) Evaluate the SOM model. To do this you need to address the following:

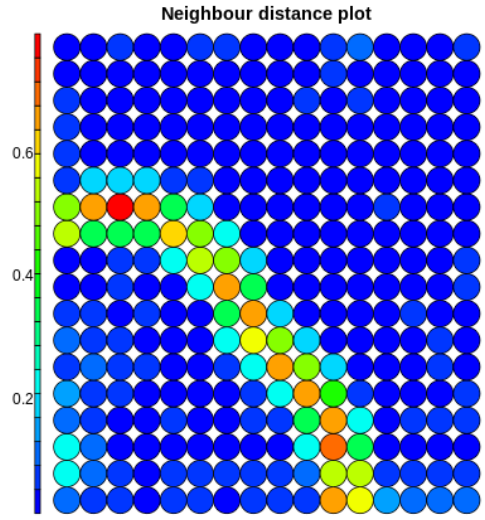
Was the epochs value selected in PART B adequate to train the model? Include a copy of the visualization that was used to make that determination. If the model needs additional training, adjust the epochs value and retrain the model before continuing.

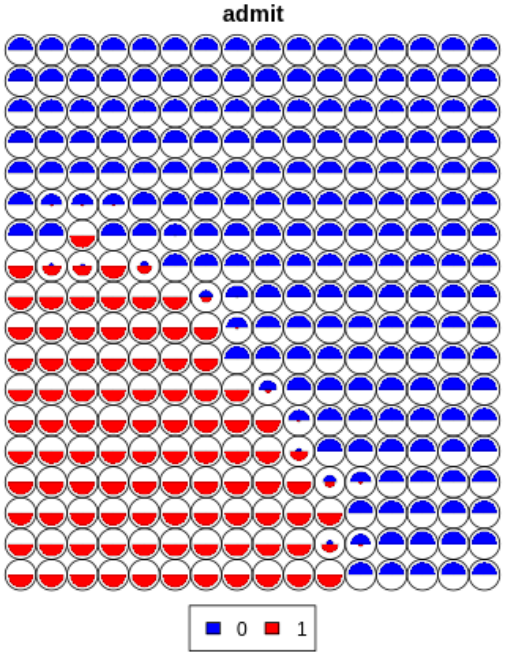


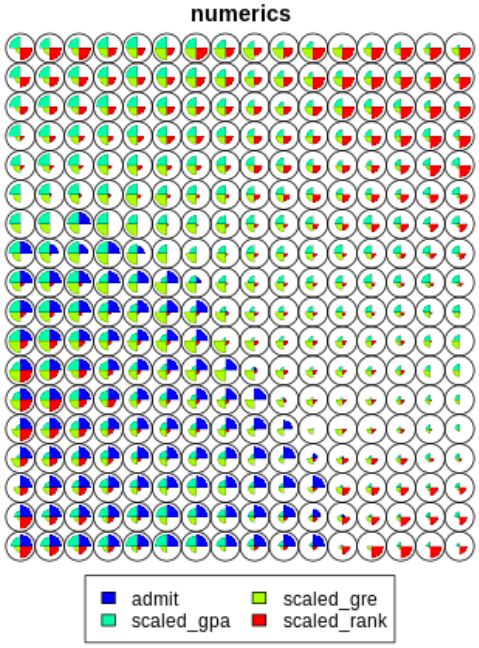
Based on the chart above the model had adequate training at around 9,000 epochs.

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.



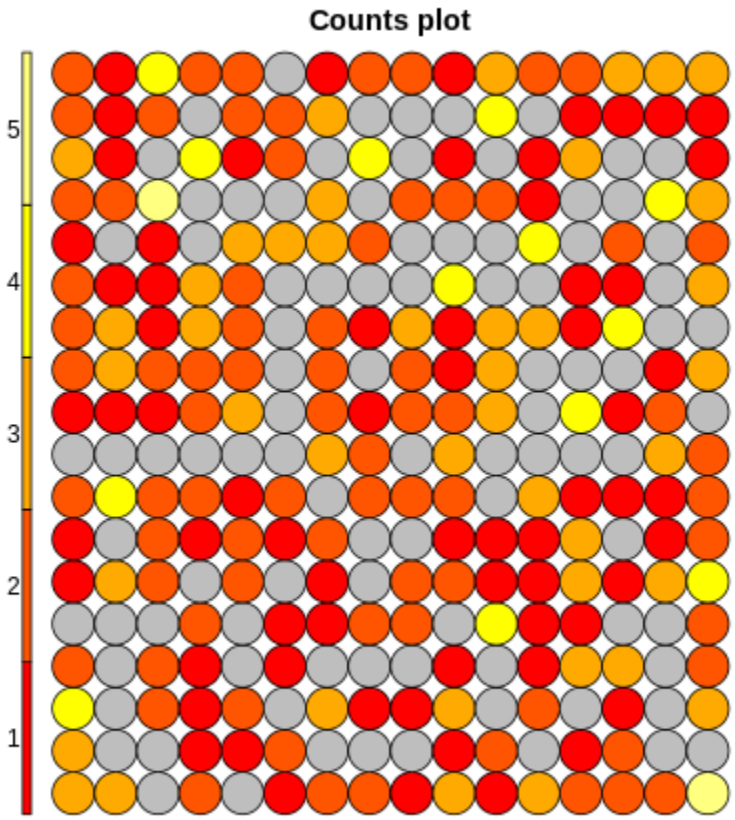






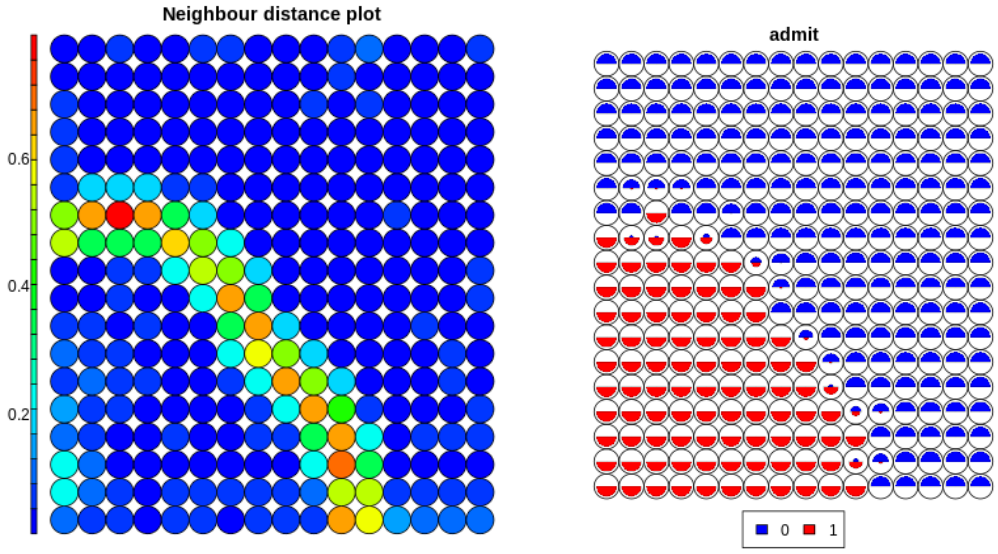
The charts above are what I produced to visualize the SOM model. Based on the grouping of the data, the chart size was adequate as there are clear boundaries between admit and non-admitted students.

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



The average number of observations per node is 2.04.

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



Yes, the nodes appear to cluster in a line starting within the plot. This is the boundary between the y variable, admittance, as shown in the right plot.

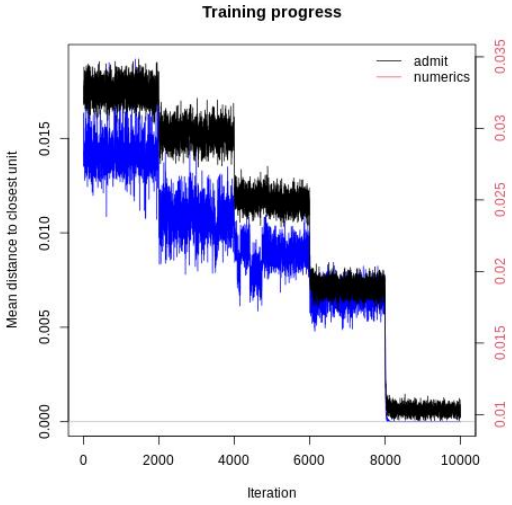
8) Experiment with the SOM model. To accomplish this task, you will need to:

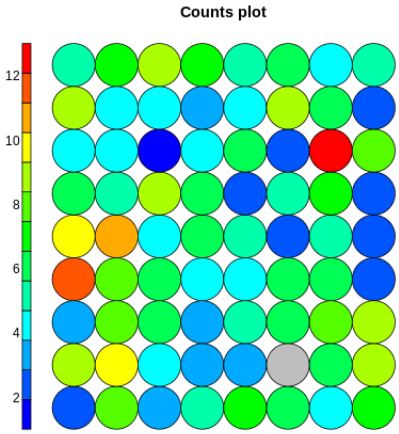
Change the grid size for the SOM and retrain the model. Discuss whether you increased or decreased the grid size and why.

I decided to test the decrease in my model as it has a larger number of rows and the average observations per node was quite low.

Compare this new SOM to the SOM created in PART B. Does the new grid size improve the SOM?

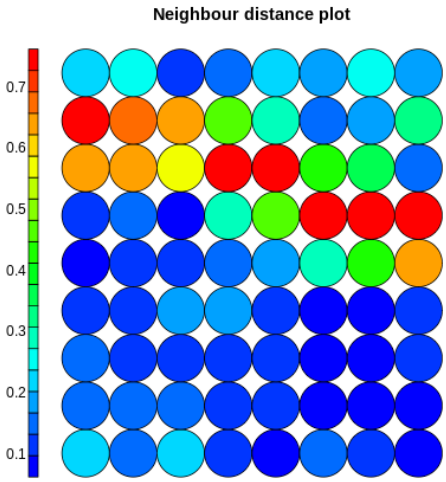
Discuss how grid size impacts the SOM.





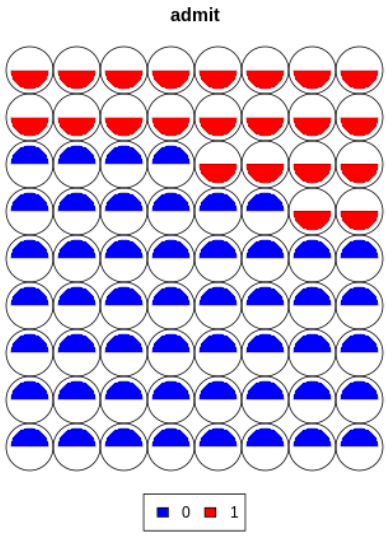
The new grid, in my opinion, decreased the precision of the plot and thus made it less effective.

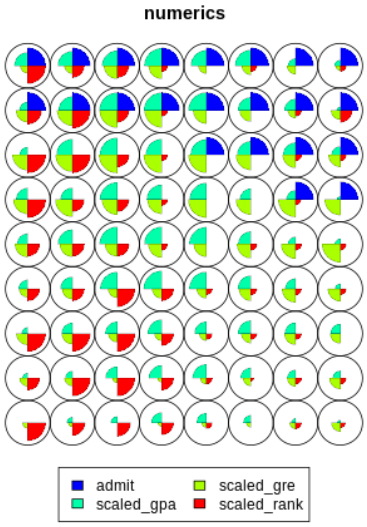
Generate a distance map and attach a copy of it here. Are any nodes quite distant from their neighbors?



As with the previous plot the nodes at the boundary between the y variable are distant.

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





This shows that the attribute weights do not have an even distribution and that scaled_gpa has the heaviest weight, it also shows that there is clear clustering between those who were admitted with the explanatory variables.

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

This assignment was very informative as well as quite difficult to me as I have not had any exposure to SOM modeling in the past. It is quite unusual and I would like to try it with a larger dataset. I realize that scaling is very important as uneven explanatory variables can really throw off your model. I did find it interesting how the model was able to separate out the Boolean variable and how it had a clear distance boundary between admitted and non-admitted students.

The MDS exercise was less interesting overall as I have worked with similar data sets in the past and I know of far better methods to utilize when analyzing similar data sets. However, the visualizations I created were pretty telling of the predictability of the output variable.

Overall, this lab was an enjoyable one but it was extremely time consuming; however I had looked forward to working on it each day during the past week when coming home from work.