# Week 1 – Assignment 1

By: Zach Adair

MSDS 680 – Machine Learning

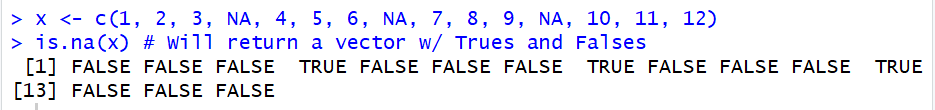
Regis University

## Introduction

For the assignment for week 1, I am tasked with working on the pre-processing techniques in RStudio and preparing my data for analysis. The types of commands I’ll be using here will help check and remove those missing records and ideally normalize the data so that it is ready for any type of analysis I would wish to run upon it.

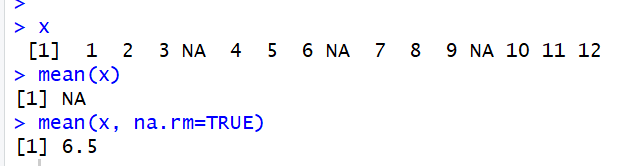
## Forming the Data with Missing Values

To start with this practice with missing data I need to first either construct or find some data with some null values within it. I will start by creating a vector with numbers and NA mixed together. Then I will use the command is.na(x) to return a vector of T/F.



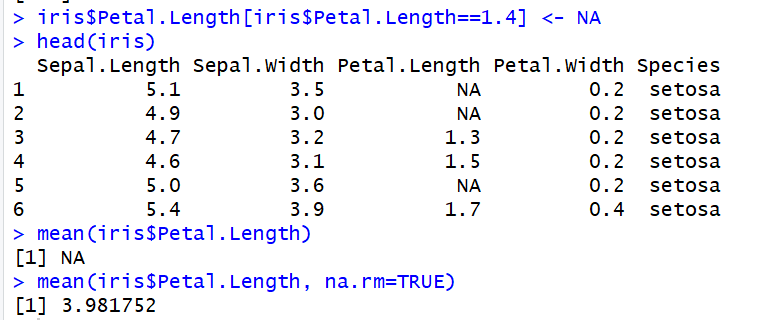
I got the result I expected, I created a vector with the numbers 1-12 and NA inserted in every 3 numbers, and as a result when I ran is.na(x) I got a sequence of FALSE FALSE FALSE TRUE several times.

Now I will exclude those missing values from the data and find the mean of x.



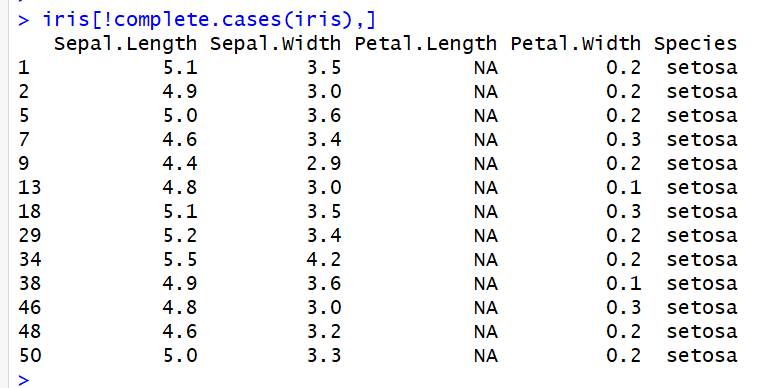
As you can see, using x again, when I first am trying to find what the mean of x is the result comes back as NA, but after removing NA from our vector we come back with an answer of 6.5.

Now I’m going to do a similar sequence but instead of creating my own vector I’m going to use the Iris dataset preloaded into RStudio.

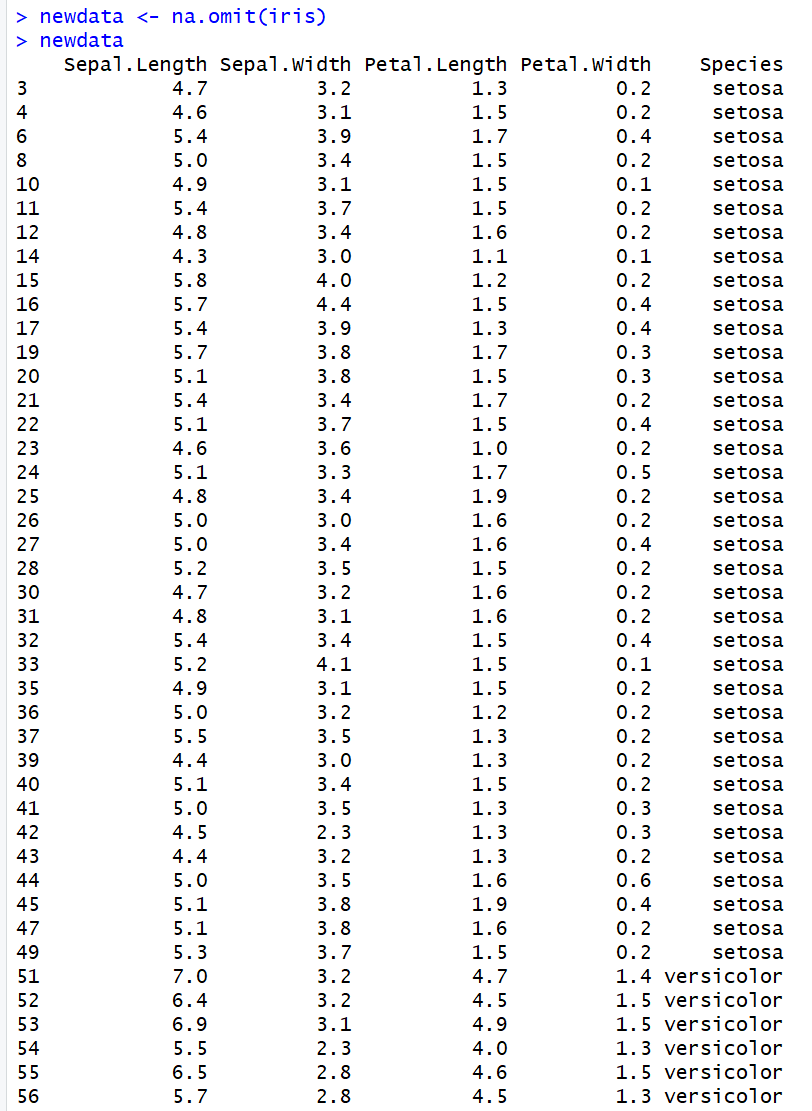


In the commands above, I seek out the iris dataset, then look at the top 5 records with the head command, and then I set any Petal Length of 1.4 to NA. From there, I went to find the mean of Petal Length, but that returned now NA so I needed to again take out the NA values that were formerly 1.4, and it returned 3.981752.

I can also list out the values with the missing data, either with the data set this can be applied with the complete.cases() command.



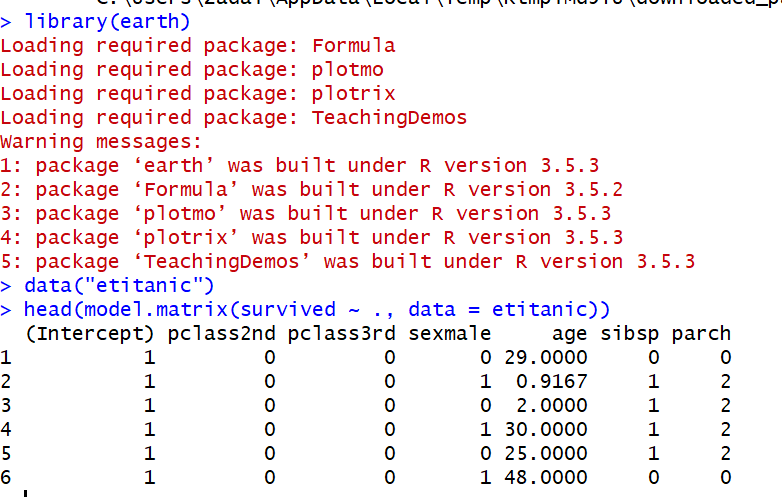
The last command I will work with be na.omit(). This will return the object with listwise deletion of missing values.



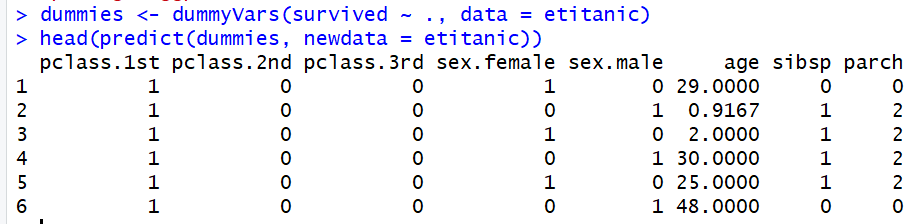
In the newdata set which was created from the Iris dataset I have omitted any record where Petal Length was 1.4, and then displayed it after doing so.

So here I was able to run through setting values to NA values and also omitting them to analyze data and now I’m going to run through another example from github that will help create dummy variables and analyze data as well.

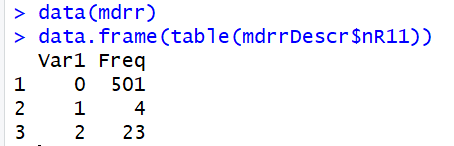
I will start by getting my data and then set up dummy variables.



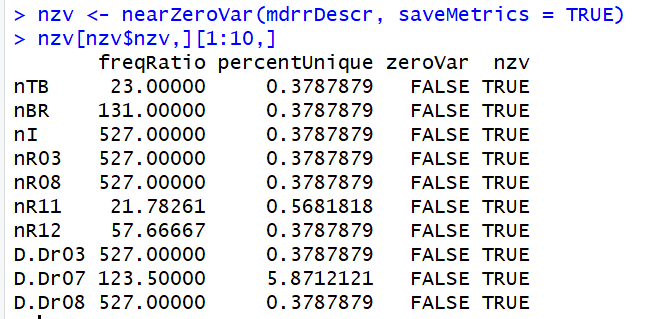
Now set the dummy variables.

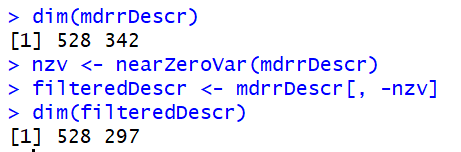


Next, I will look into zero and near zero- variance predictors.



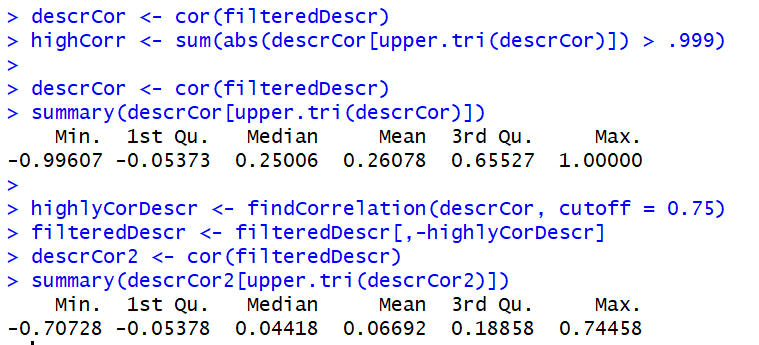
These are the zero-variance predictors when the data are split into cross-validation/bootstrap sub-samples or that a few samples may have an undue influence on the model.





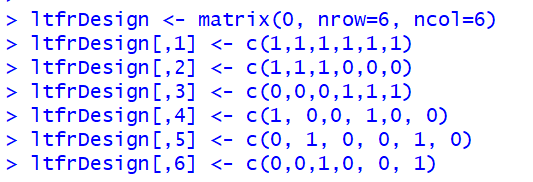
The result of this will be the nearZeroVar returns the position of the variables that are noted as problematic.

Next we want to identify correlated predictors, so starting with a correlation matrix.

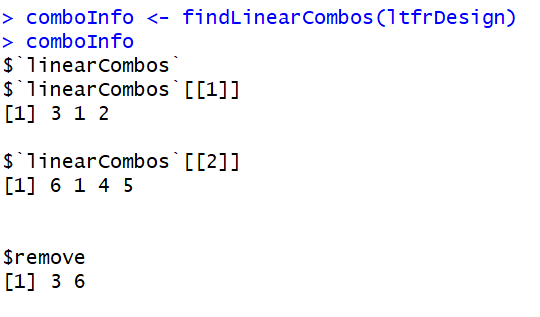


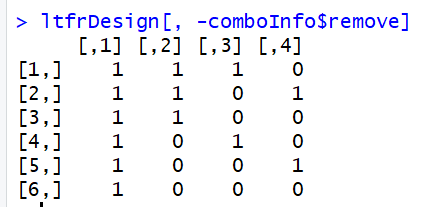
The data above shows a near perfect correlation of the data as well as when removing descriptors how that changes data set to correlations just above 0.75.

Next, I will work with linear dependencies within the data.



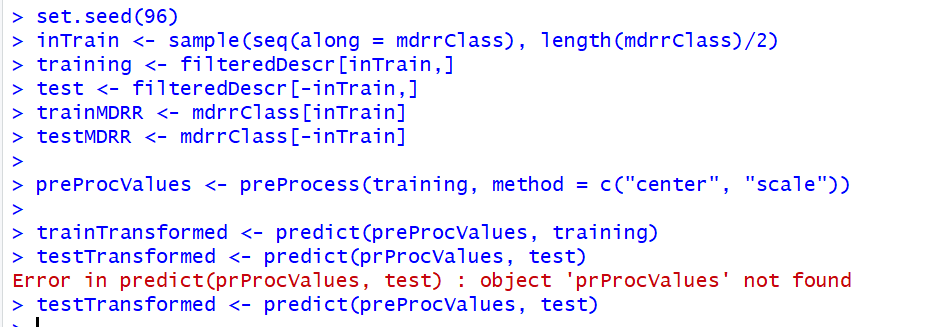
Above is the 6x6 matrix being put together. The command findLinearCombos will also return a vector of column positions can be removed eliminate the linear dependencies.





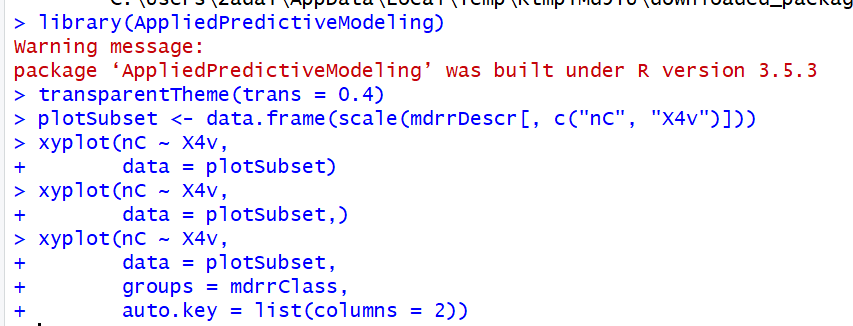
These dependencies come when large numbers of binary chemical fingerprints are used to describe the structure of a molecule.

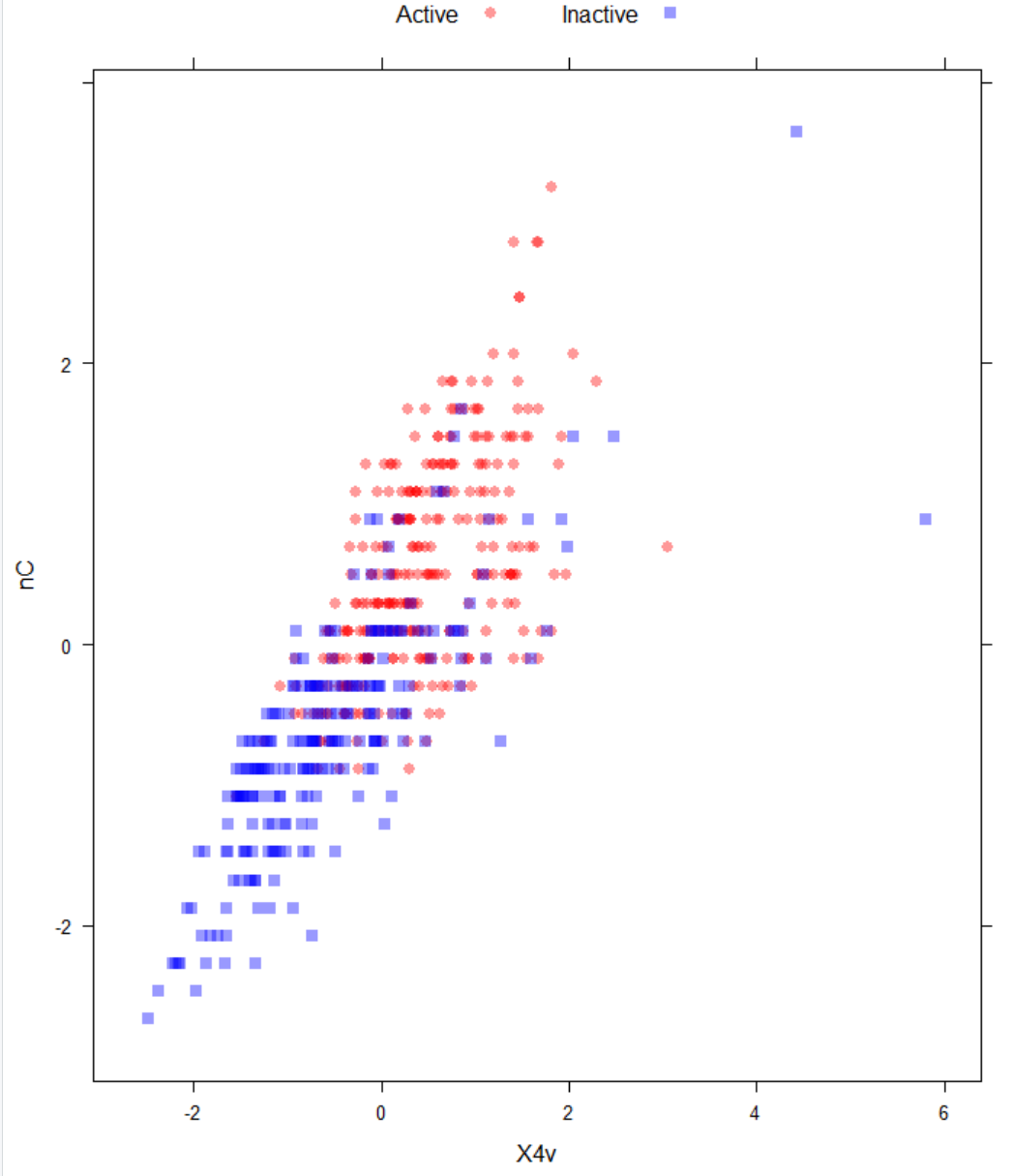
Now we will use the preProcess function to pre-process other data sets.



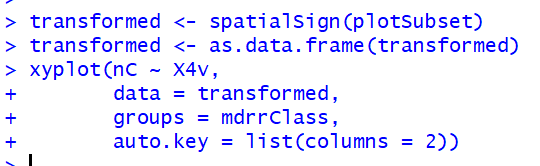
The process above just set up the training and testing data sets for pre-processing. The preprocess option “range” scales the data to the interval between zero and one.

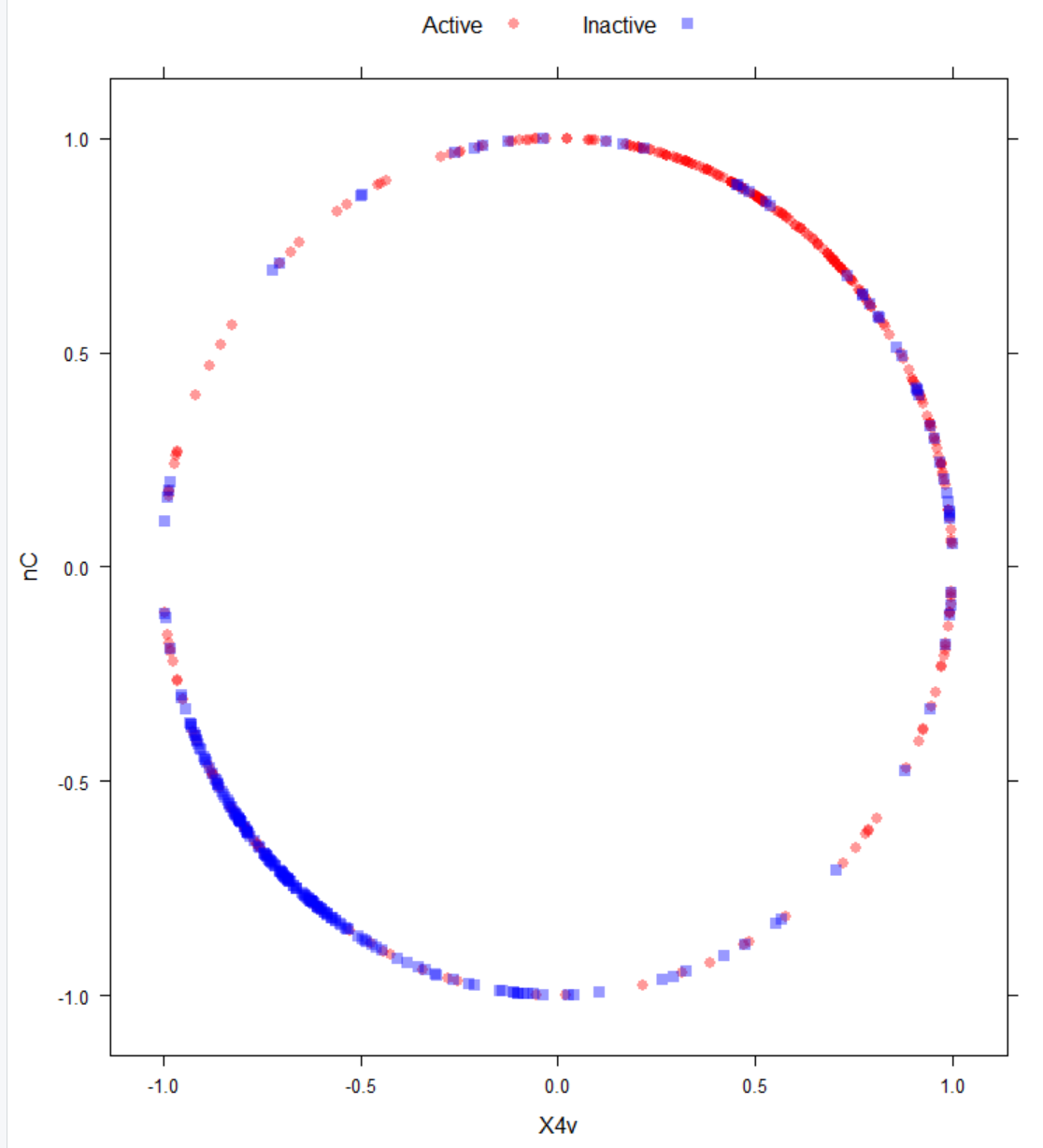
Now I will bring in the package, AppliedPredictiveModeling and with it we will also do some plots of our predictors that were just constructed.



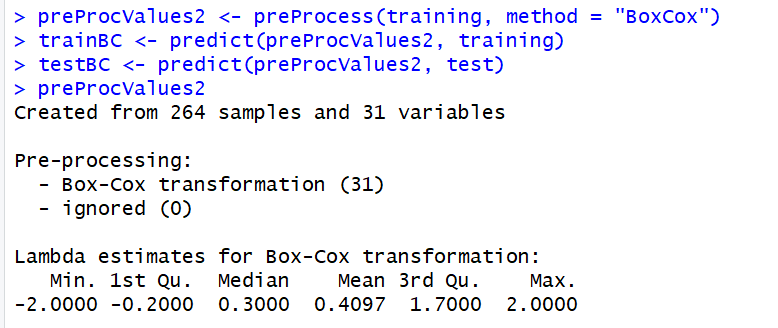


In the plot we are able to see which records are active and which ones are inactive, but now we will see what happens after the spatial sign:



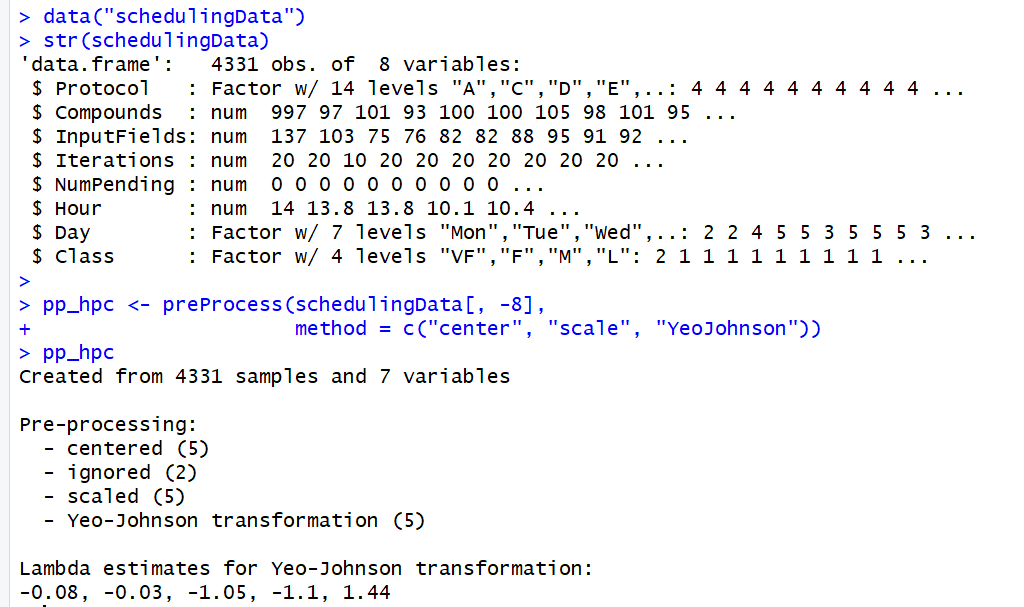


Setting the data and transforming it to a spatial plane and then setting it as a data frame this is what our plot looks like now, One big circle with both Active and Inactive records making it up. Now another option is a Box-Cox transformation which estimates on the predictors if the data are greater than zero.

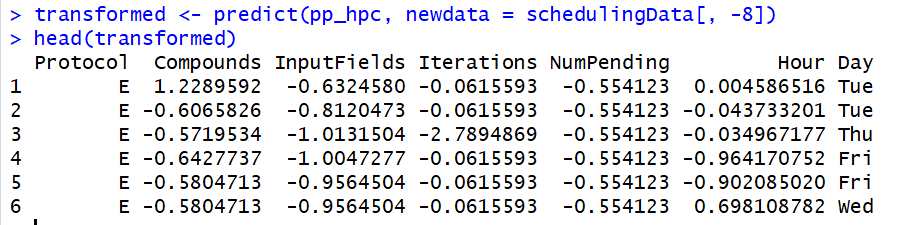


The NA values correspond to the predictors that could not be transformed, which requires the data to be greater than zero.

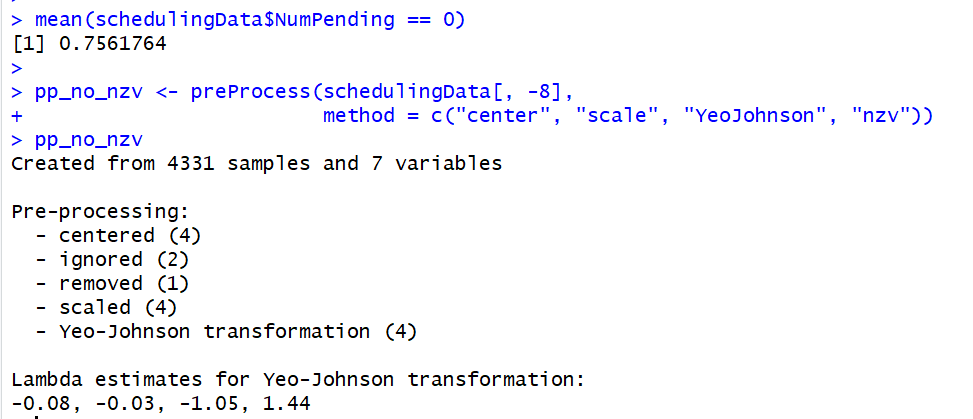
What is next to do is put all of these steps together, and we will do that with the schedulingData.

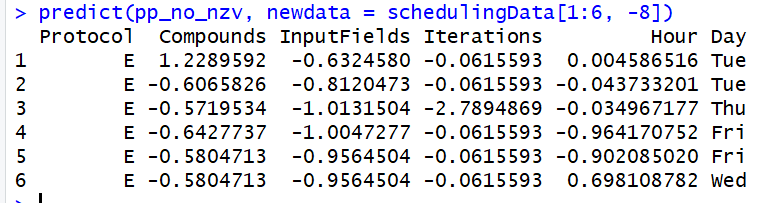


Now transform the pp\_hpc data.

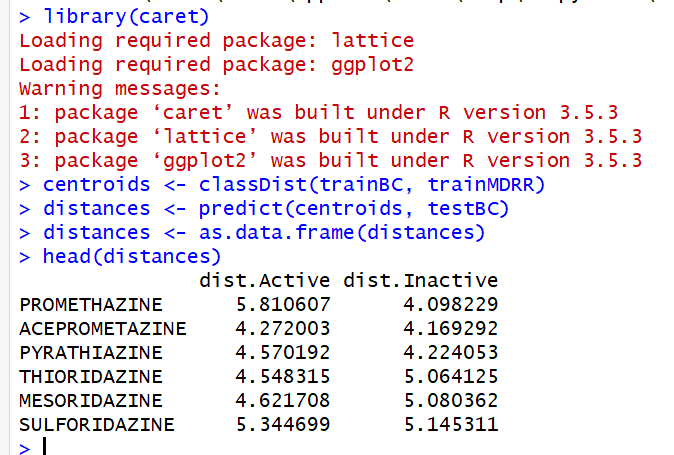


The two predictors labeled “ignored” in the output are the two factor predictors. These are not altered but the numeric predictors are transformed, but, the predictor for the number of pending jobs, has a very sparse and unbalanced distribution.

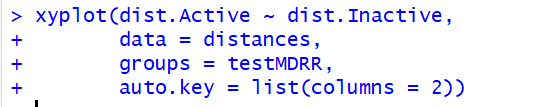


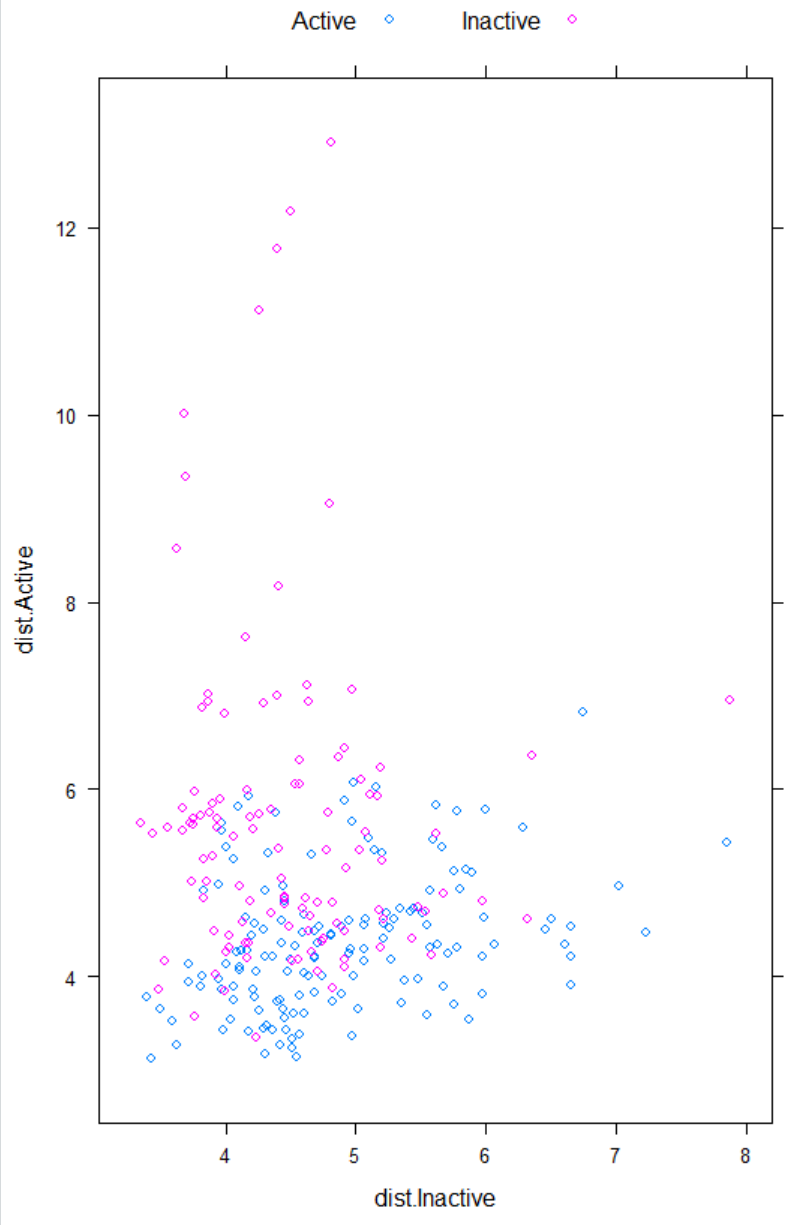


Here we started by finding the mean of the scheduling Data column for Num Pending at zero, then set up pp\_no\_nzv and took it through the pre-processing command. Then finally predicted the newly made data set.



Now I will plot the data.





Now that it is plotted, we can see the difference between Active and Inactive data points. But that is the end of the github demo of the pre-processing techniques an analyst can use in RStudio.

## Summary

After going through a couple of the demos assigned for this week I found it really fun and informative. Both seemed very extensive and each step was well explained, making them easy to follow. I liked that I got to use a wide amount of different analytic techniques, from data preparation and manipulation, to plotting and even running predictor models I felt like there was a lot to absorb and learn from the assignment overall and a great introduction to the course.