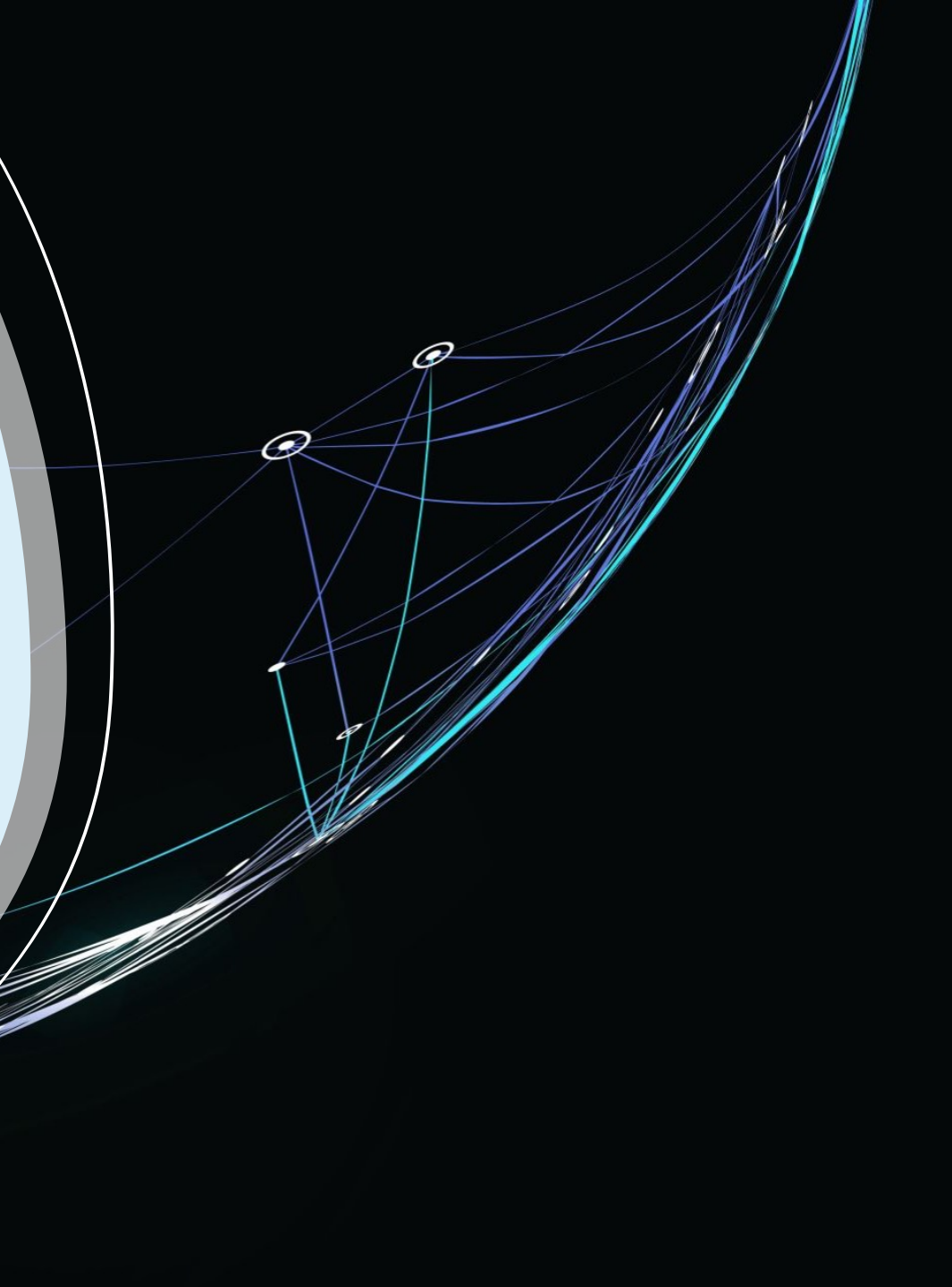
Prediction Challenge #3

Nick Whelan – ncw32

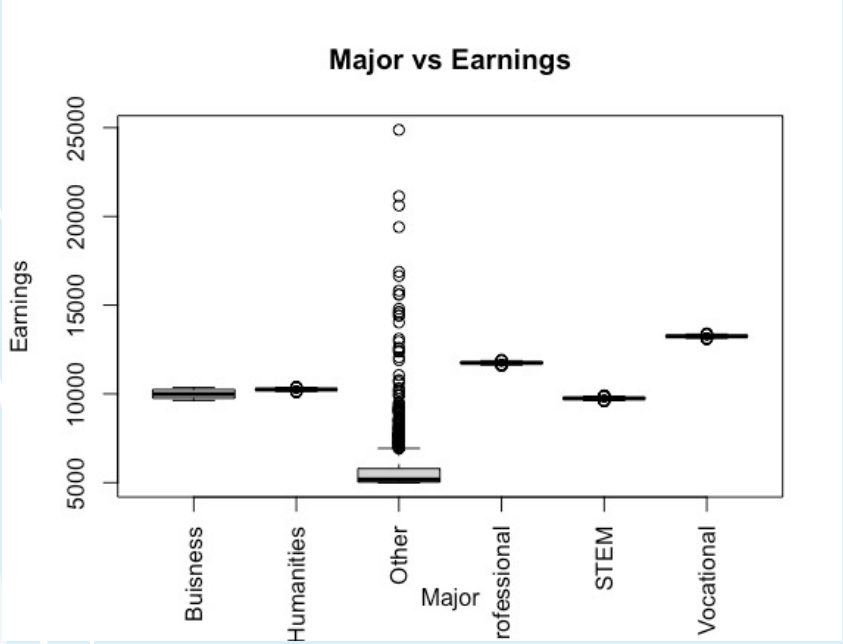


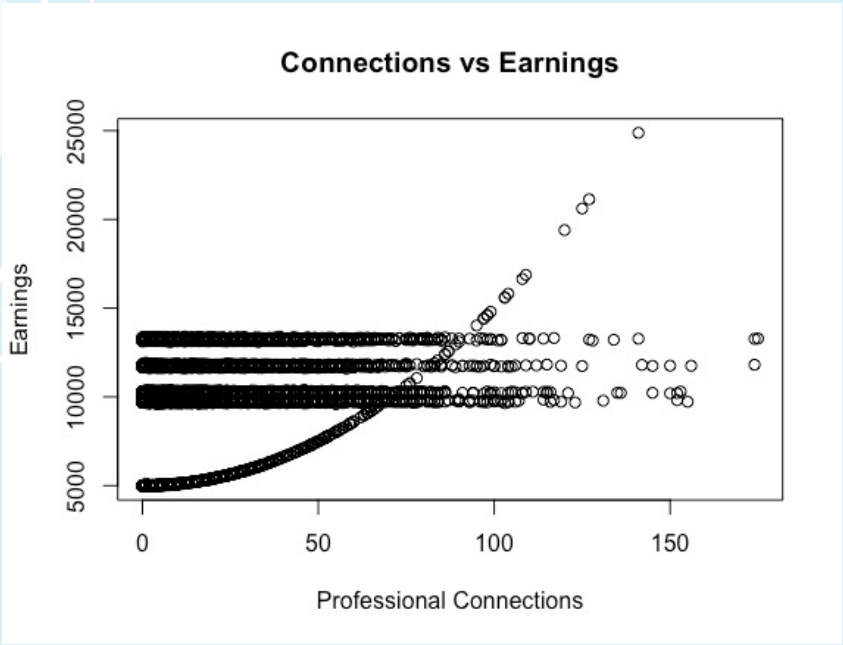
First Model Attempt

- Before really looking into the data very much I attempted to use a random forest model on the entire dataset. This yielded an MSE of about 25k, but I wanted to see if I could get a lower MSE with some other methods.

Initial Plots

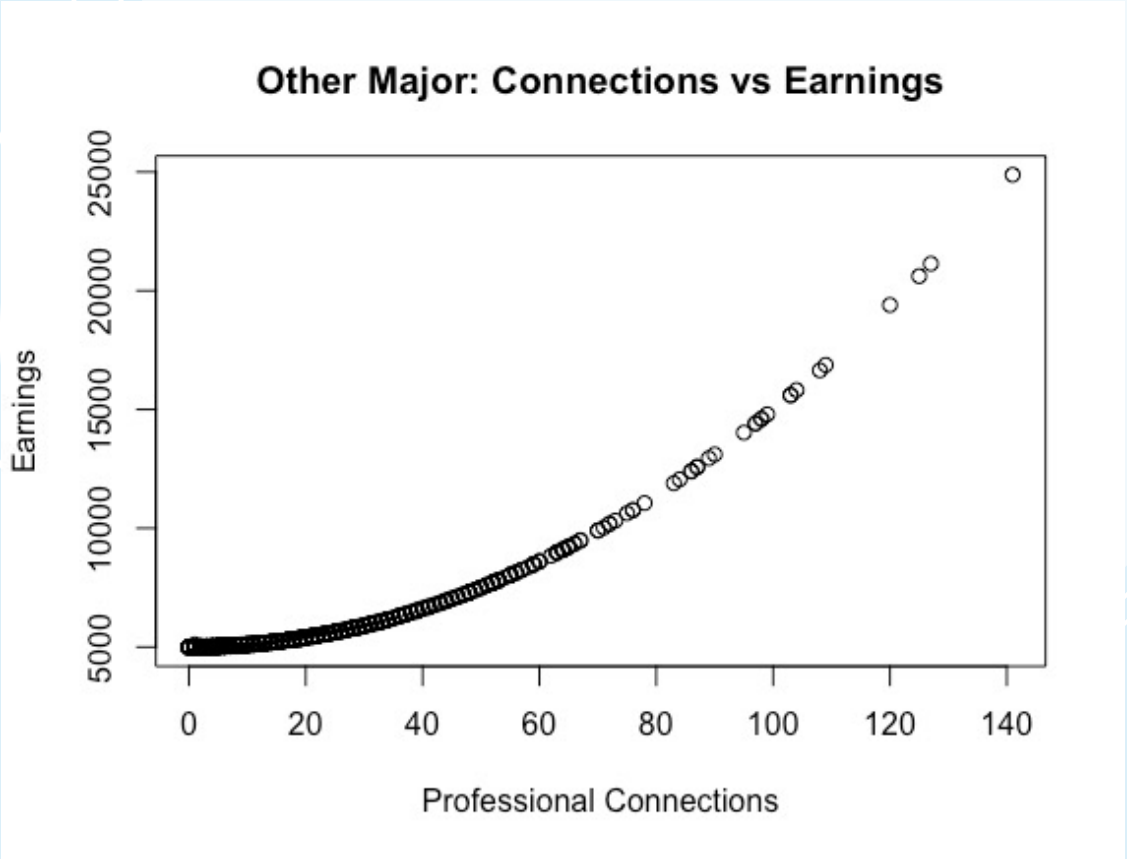
- I first plotted the each major and their earnings, and saw that earnings tightly correlated to a specific major. One interesting thig was that the “Other” major had a low average, but many outliers.
- After plotting against various other attributes, I noticed an interesting relationship in connections and earnings. It seems that some subset of the data’s earnings was an exponential function of that subset’s professional connections.





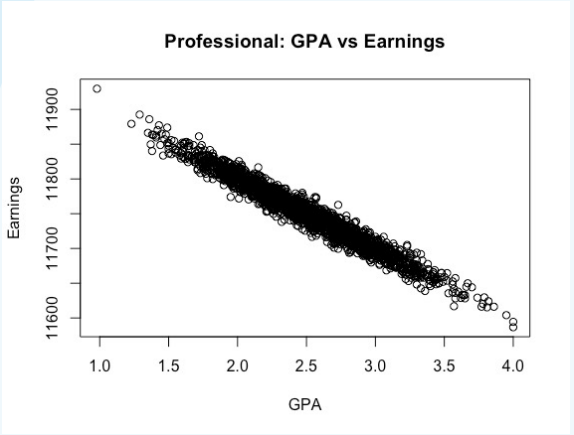
Locating The Subset

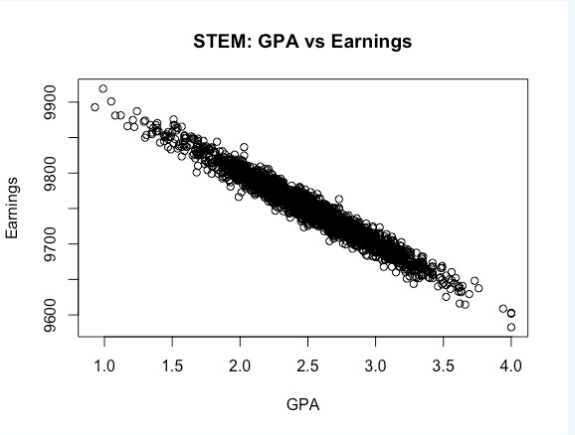
- It appeared that there were about 5 or 6 distinct subsets in the connections plot, and we have 6 majors so it made sense to start subsetting by major. After several plots I found that the ”Other” major was this subset. This made sense as in my initial box plot it had many high outliers.
- I attempted to fit this curve with some tools (nls). But ultimately it was easier to to estimate the function. I found that $f(x)=x^2+5000$ where x=Connections fit perfectly, and tested it on individual rows to be sure.

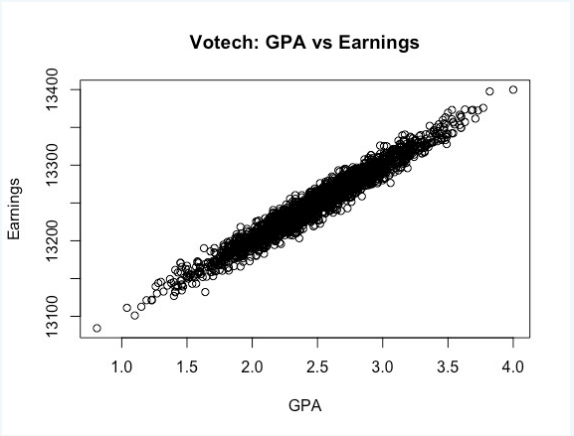


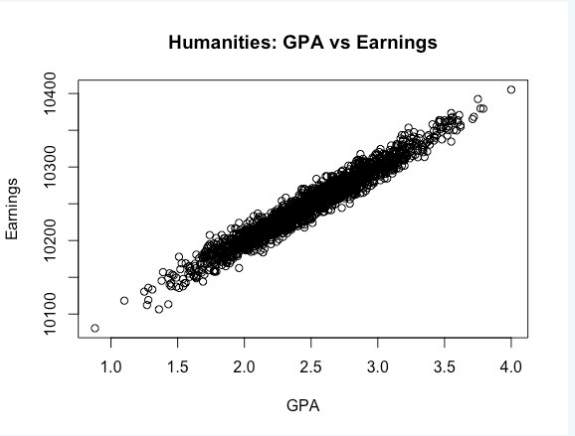
Continued Analysis

- Once I realized that Other's earnings was a function of their connections, I was curious if other majors were functions of some other attribute.
- After some plotting, I found that Professional, STEM, Humanities, and Vocational major's earnings all seemed to be linear functions of their GPA. (Some inverse)
- For all 4 of these subsets I used linear models [lm()] and got a good MSE of ~100.



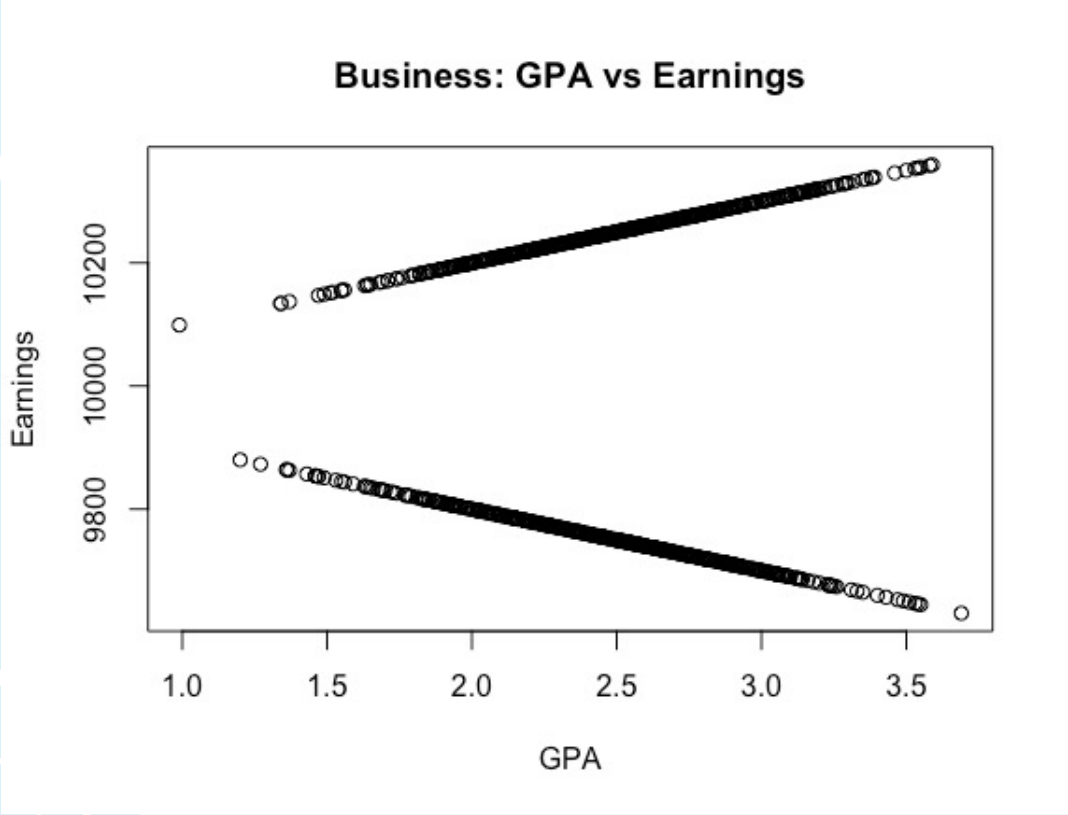






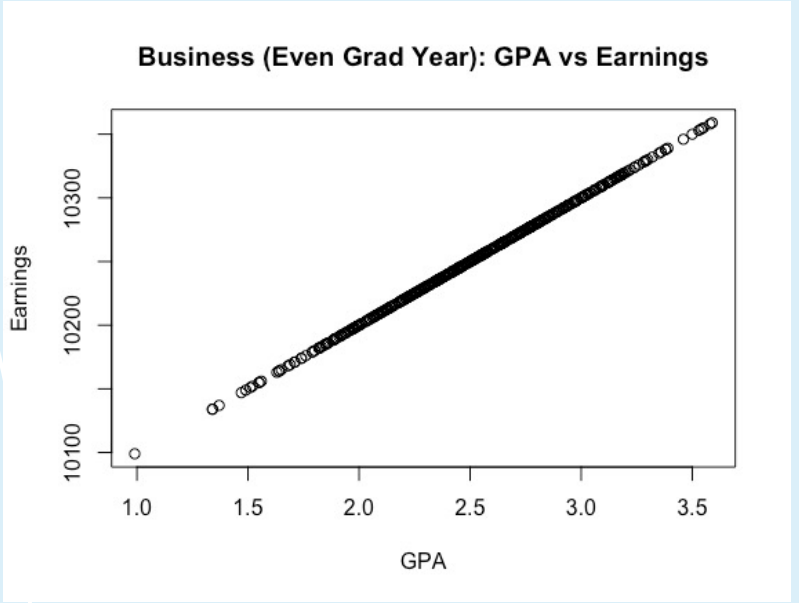
Business Weirdness

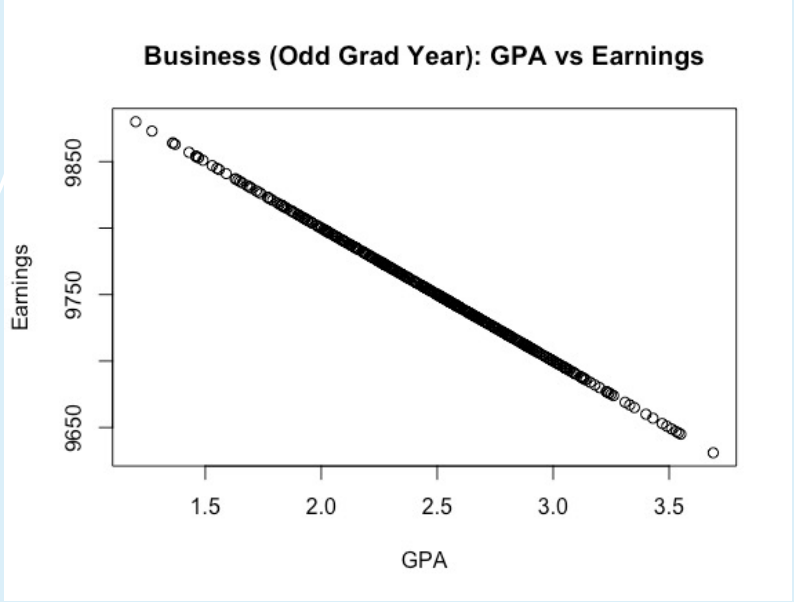
- The last major to make a model for was business. In this case I found a more interesting plot.
- It seems like there was another subset in business that determined which function it would be under.
- Luckily, one line was right above 10k earnings and one below, so I simply split Business once more and looked at the data for each subset. I quickly noticed that the rows that had over 10k in earnings only belonged to even numbered graduation years and vice versa.



Business by Grad Year

- After subsetting by even and odd graduation years I found that business grads were an even tighter function of GPA.
- Since there was such little variance, using lm() yielded an extremely low MSE. I calculated it to be 1.505e-23. Essentially zero.





Putting it all back together

- After putting my results from each model back into the training table I calculated my total MSE to be 90.96. Much better than using any one model on the entire dataset.
- All I did at this point was use each model on the corresponding subset of the testing data. After building my submission table and submitting to Kaggle I received an initial MSE of 82.98723.