Ryan Frank

DSC530 Term Project

The overall outcome of my EDA was the creation of a series of models that attempt to predict income based on education, gender, marital status, age, number of children, location and type of occupation. I feel like I ended up in a good place with my final model having an adjusted R-Square of 0.386, which seems like a decent result for using a limited set of variables.

I think if I did a deeper dive and included more variables there would be an opportunity to improve my work. One of the variables that I did not end up gathering was college degree major. While general education attainment did have a statistically significant relationship with income, I think delving into more detail could get a better result. There was additional categorial data that had an ever more detailed breakdown of location and job title that could be useful that were not among the variables I extracted. I am also interested in looking at variables that indicate primary language spoken as well.

None of my assumptions seem to have been factually incorrect, however some of the variables chosen had less of an impact than I expected. Gender for instance worked well as a proxy when using just a couple of variables, however once other variables that account for lifestyle and career choices were included the variation that it covered was much lower.

The major challenge for me was selecting the variables to use in the original extract. While none of my assumptions seem to be flat out wrong, I suspect having a larger pool of variables to choose from and additional time to analysis all of them could produce a better result.