The exploratory data analysis I ran was on data extracted from the 1994 Census bureau database by Ronny Kohavi and Barry Becker in an effort to predict if what factors could be in play to see if a person makes over 50 thousand dollars a year. As stated on the Kaggle website,

“A set of reasonably clean records was extracted using the following conditions: ((AAGE>16) && (AGI>100) && (AFNLWGT>1) && (HRSWK>0)). The prediction task is to determine whether a person makes over $50K a year.”

The key dependent variable is “income\_over\_50k” with the independent variables being: age, education, education\_num, marital\_status, race and sex.

From my exploratory data analysis there seemed to be notable pattern on higher levels of education grouping to show there was a tendency for those to have aa higher likelihood of making over 50 thousand dollars a year. It also appeared in 1994 that being male and being either white or other (non-African American) had some role on making a successful prediction.

I feel one element that was missing that may have had some interest on the analysis would have been if the respondents had children or not as the responsibility of children could have impacted the number of hours a person had to work to increase their yearly income.

I also feel that perhaps the date of the data could have had a notable impact as compared to today in that the technology industry was not as developed yet as it is today which created a need for higher skilled workers and more opportunity to make over 50 thousand dollars a year.

Reference

Becker, K. & Kohavi, R. (1994). Adult Census Income. Retrieved July 20, 2020 from, <https://www.kaggle.com/uciml/adult-census-income>.