Exploratory Data Analysis: Extremism

Course: DSC 530

Author: Holly Figueroa

Date: 11/20/21

Exploratory data analysis was conducted to seek possible predictors, however slight, for the unique social phenomenon of extremism. The data set was large, but variables relating to age, gender, religion, social media use, political view, and racism were the focus of these analyses. Whether any of the variables had any relationship to extremism score was the hypothesis being tested. Results showed evidence for relationships between extremism and racism scores and political leaning. Regression results supported potential influences of age and gender as well.

To better find relationships to extremism, the sample was split into higher and lower scores to compare results for certain plots. PMFs created showed that respondents with higher extremism scores had higher probabilities concentrated for right-leaning political views.

The CDF for extremism scores showed the data was poorly suited for parametric tests as over 80% of the respondents scored 2 or less for extremism. Scatter plots comparing racism scores showed that those with higher extremism scores had a wider range, including higher racism scores compared to those with lower extremism scores. The difference between mean racism scores for both groups was tested for significance. Results showed the more extreme group had a -0.96 racism mean while the less extreme group had a -0.05 racism score mean, with 0.0 p-value.

Kendalls Tau was used to correlate extremism scores and racism scores. Results supported a small but significant negative tau at -0.179. Given the coding, this means higher extremism is related to higher racism. Variables were then explored for inclusion in a multiple regression model for extremism scores. After some trial and error, the included variables for regression were age, gender, racism scores, boolean -racism scores < 0, and a boolean for strong-right leaning politics (scores < -1) . The predictive power was modest, with an adjusted R2 of .168. The f statistic was strong and the p value well under 0.00. The coefficients for each variable were supported by significant p values, all under 0.00.

The pool of variables provided by the original dataset was impressive, however, I believe scores for racism and extremism should have been derived from the sum of more questions than they were derived from. This created a limited opportunity to capture these variables compared to others.

The most concerning assumption the analysis cannot address is the lack of population representation in the data set. Demographics of the respondents are very disproportionate in race, religion, and gender in the UK.

The main challenge provided by this data set was the nature of the dependent variable. Extremism scores were mostly zero. With a severe skew to the distribution extremism scores and most other variables, parametric tests were not optional. Splitting the data by extremism score allowed for some important differences to be seen, but none could offer much for the purposes of prediction. Another challenge was interpreting test results as some variables scaling was inconsistent and not intuitive.