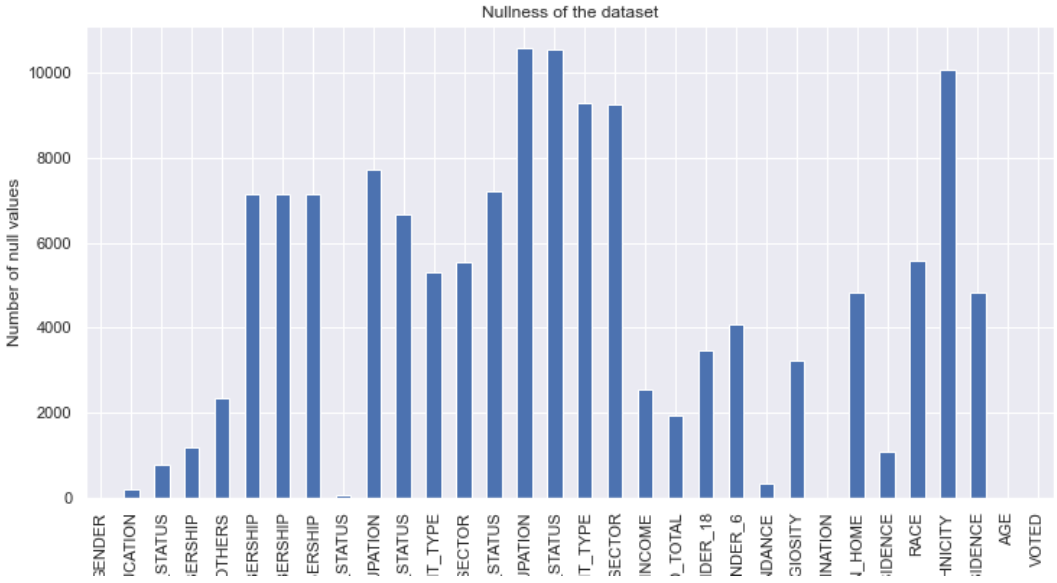
I first changed the name of the variables so that it can be understood better during the analysis.

I also deleted the first row as it was similar to the index.

I, then, checked the high or low feature variability by looking for the number of unique values in each variable.

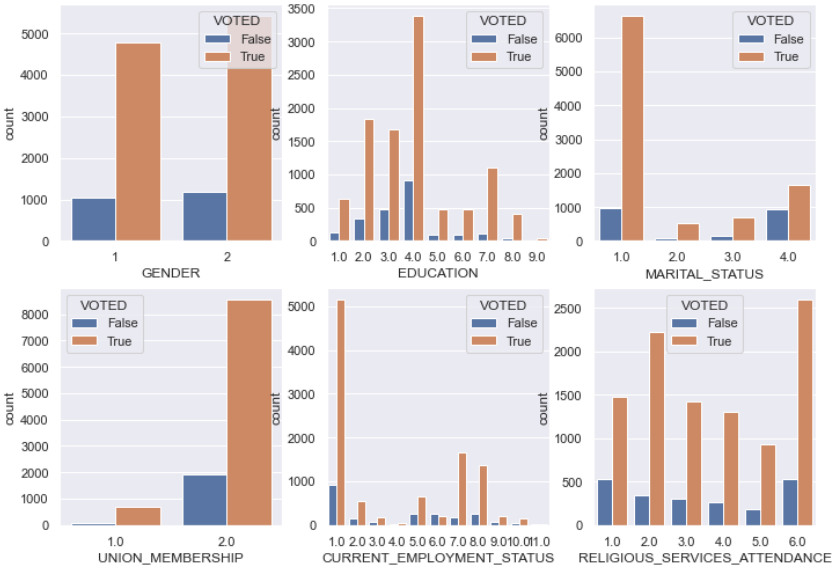
In order to check as to whether there are odd values and data collection mistakes, and it appeared that there are several variables that took values outside the scope of the values given in the codebook i.e. RELIGIOUS_DENOMINATION.

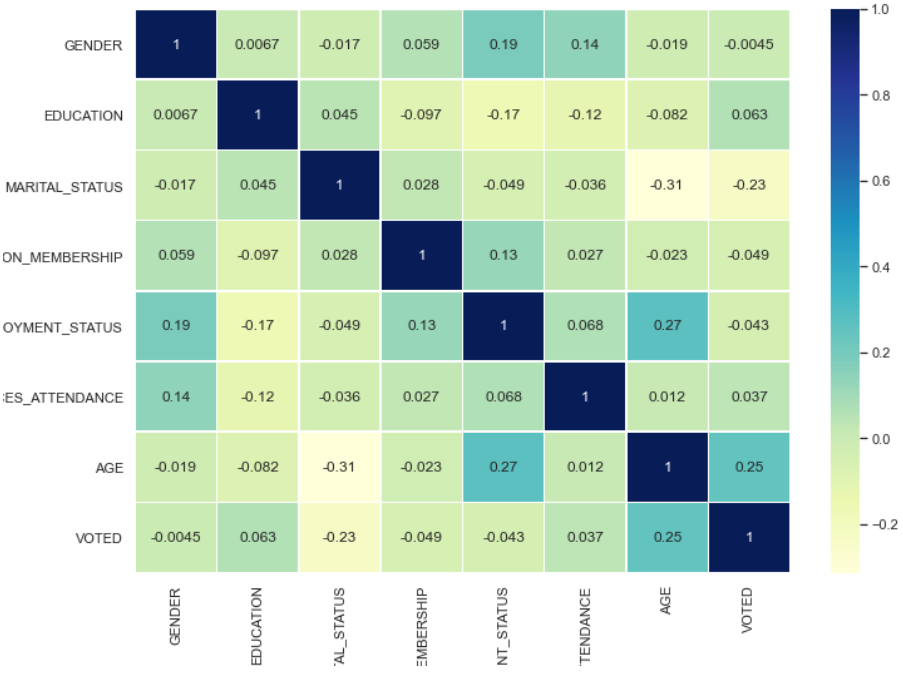
Then, I looked at the codebook and coded refused, don't know and missing as missing values. For education, since there was also a value for early childhood education, I dropped also observations for no education. One factor pushing me to do that was related to not to complicate the process by recoding the existing values. I then, checked the number of missing values in the entire dataset and plot it in a bar graph in order to better see the distribution of missing values across the variables.



As it appeared form the graph, several variables have too many missing values. First, I dropped the columns whose half of the rows have valid values. However, this has dramatically reduced the remaining dataset as there remained less than one thousand observations in total. As such, I drop columns that have less than 11,000 valid values, rending the dataset to keep 9995 observations. While knowing that most of the variables are dropped as a result of this decision, I prioritized keeping large number of observations as possible.

Then, I plotted graphs in order to see the distribution of values on outcome variable and also a correlation matrix. The matrix revealed that there is a correlation between age and outcome variable.

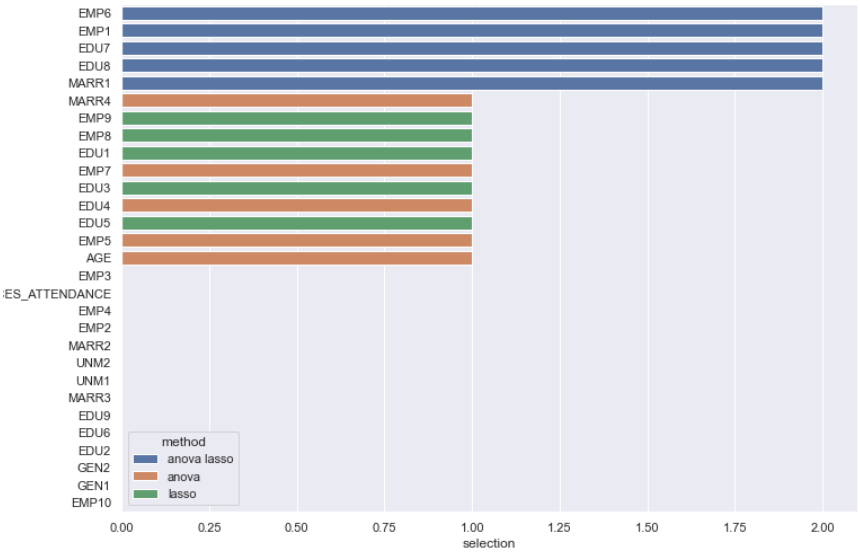


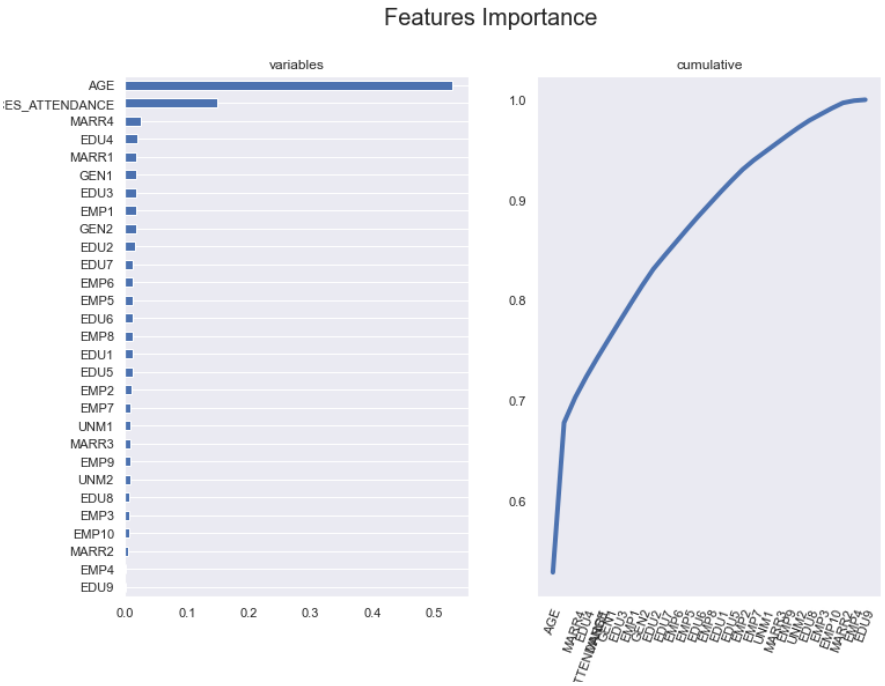


Then, I hot encoded categorical variables in order to make them appropriate for the ensuing analysis by one by.

Then, I created feature variables and a outcome variable and split data into training and test datasets.

Based on the feature set, I try to understand which variables are more important in the final version of the dataset in order to select only the features appearing to have more significance. Hence, I plot a graph that contrast the results of the ANOVA and Lasso regularization and it appears that certain variables appear to be selected by both methods. I also, applied random forest classifier in order to detect the most important features and it showed that age and religious service attendance are the most important ones. I also included other most important variables in the graph, too. As such, I readjusted my feature variables and once again split old dataset into train and test sets.





I applied, Gaussian Naïve Bayes model to the dataset. As, it has naïve assumptions, I did not make any changes for hyperparameters.

Then, I applied logistic regression and decision tree models with the default parameters. Then, I changed the hyperparameters with Grid Search. Comparison of the results showed that decision tree model with hyperparameter tuning is the best model among the three with an accuracy score of 0.85