Coding Exercise

Please complete the following exercise using Python or R and submit the full code along with your results.

1. Download the csv file with the data from https://github.com/adebayoj/fairml/blob/master/doc/example_notebooks/propublica_data_for_fairml.csv directly into your code. Any lines with missing data should be deleted.

The data contains information concerning attributes of a group of convicts in the United States, including racial identification, age group, sex, number of priors, and whether or not their offense was a misdemeanor. An ML model was used to predict whether each convict would reoffend in the short run. The binary variable score_report represents the result of the model's prediction and another binary variable, Two_yr_Recidivism, indicates whether the convict actually reoffended in the subsequent two years.

2. Assume that the convicts that do not identify with any of the mentioned racial categories (African American, Asian, Hispanic, Native American, Other) identify as white. Using this criterion only, add a column to the data set indicating which convicts identify as white.
3. The convicts are divided into 4 categories according to their number of priors.

Category	Number of Priors
A	0
B	1-3
C	4-10
D	More than 10

   Add a column to the data set indicating the category of each convict.

4. Run regression models to determine the influence of the variables Number_of_Priors, Age_Below_TwentyFive, African_American, Hispanic, Asian, Native_American, Female, and Misdeamenor on the variables Two_yr_Recidivism and score_factor. Conclude on your results.
5. Regressions allow one to measure a variables' dependence on particular attributes. However, indirect effects may not be accounted for when the attributes used in the model are co-linear. Comment on whether this is the case in your regression models and why.
6. What experiments or analysis would you conduct in order to determine whether there is racial, gender, and/or age bias in the way that the ML model is predicting recidivism?