PQHS 471 (2019) Midterm

The data are from a US census in the past. It has 15 variables. One of them, `income`, is our outcome. Most variables are self explanatory. The variable `fnlwgt` is the final sampling weight, a number often calculated in survey sampling. I do not know what the values of the `relationship` variable mean exactly in our context. The dataset was probably extracted from a larger database in which the variable would make sense. You can treat it as a potentially informative variable. (I do not know if it is informative or not.)

The data have been split into a training set, `census_train.csv`, of 25000 observations and a test set, `census_test.csv`, of 7561 observations.

Study the relationship among the predictors and between the predictors and the outcome. Build prediction models using the training set. The test set should be used only for evaluating your final model(s). Specifically,

1. Perform exploratory analyses. I do not want to see all the details. I would like to see that you have made an effort to understand the data. Turn in a summary of what you have tried and a summary of main findings (in words and/or graphs) from the exploratory analyses.
2. Build a prediction model or prediction models. I would like to see that you have made an effort to improve your models' performance. Turn in a summary of what you have tried (in words and/or graphs), your final model(s) (in words and formula and code), and the performance on the test sets.

Turn in your code (with comments) as a file, and describe your results in a report.

The midterm is due by the midnight of Saturday, March 9, 2019.