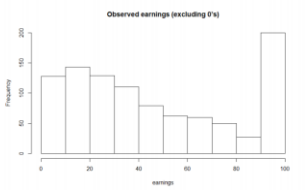
Data Analytics Assignment 1: Data imputation using leaner and logistic regression.
Student Name: Roman Shaikh
Student Number: 183000989

Dataset:
Social Indicators Survey, a telephone survey of New York City families conducted every two years by the Columbia University School of Social Work. This dataset consists of 1501 records with 944 variables (Columns), which makes a total of 1416944 data points.

Goal:
There are many missing values in the dataset for earning column (rearn+tearn). the missing-data process must be modeled to perform imputations correctly. We do this in using logistic and leaner regression models. Now implementing this in R using Rstudio as the editor.

Process:

1. Dataset Visualization: First the dataset is loaded in R. The earnings are then simplified by summing up the rearn(the candidates earning) and tearn(the spouses earning).
2. We then take only and revalent variables sex, race, educ_r, r_age, earnings, police found by using correlation. Topcode values to reduce the sensitivity of the results to the highest values which in this survey go up to the millions (of earnings).
3. We now perform random imputation of single variable. And plot histogram of earnings vs Observed



4. Deterministic imputation of single variable:
   We now filter and refine our dependent variables - earnings, earnings.top, male, over65, white, immig, educ_r, workmos, workhrs.top, any.ssi, any.welfare, any.charity and perform the deterministic imputation of single variable. A new column is generated which replaces positive values in earning with 1, 0 and negative values as 0 while keeping NA values constant. With the help of this new variable, Logistic regression is applied, and predictions are taken on the NA values.
   Newly predicted values from logistic regression algorithm; negative values are updated as 0, positive values as well as NA values are retained as it is. and Linear regression model is finally applied. The predicted values are updated on the original dataset.

5. Standard deviation and Mean of the data is calculated from this imputed dependent variable
   Results: Standard Deviation = 45267 and Mean = 29246