CS614 – Prof. Berardi
Homework #6 (Lecture 10)

In this assignment, you will use analyses we reviewed in class to identify and manage outliers and missing data. The data you will be using was crowdsourced from Amazon Mechanical Turk, where Turkers examined over 90 years of Time Magazine issues and extracted and annotated faces from the corpus. Detailed info about the datasets can be found in this data set publication – you will be using augmented versions of Datasets 2 and 3 (see above). For the interested student, here is a publication we created with this data that examined trends in the representation of women in Time Magazine throughout the 20th century and how they related to prevailing cultural trends.

I've deleted a few columns from Dataset 2 and added an area variable, that details the area of the cropped face image. Dataset 3 (Turker demographics) is unchanged.

For this data, perform the following tasks:

1. In the demographic table, there is a proficiency variable, which essentially captures how accurately a Turker annotated extracted faces. Identify outliers in the proficiency variable using each of the three methodologies discussed in class: 1.) +/- IQR; 2.) +/- 3SDs from the mean; and 3.) inferential statistics. Visualization may be used as well. Compare the results from each approach, including the number of outliers identified and the range of values that were removed. Based on your results, how would you propose to treat the outliers in subsequent analyses?

2. Identify all missing variables in the annotation data frame (df). Perform diagnostic tests to determine if the NAs appear to be MCAR or MAR. Then address the missing values in two ways: 1.) list-wise deletion and 2.) multiple imputation. For each case, build a linear regression model that predicts area based on all other variables. Contrast the results of these modeling efforts. Which approach do you think is most appropriate?