Capstone Project 1: Data Wrangling

This project looks at data gathered via an app designed to provide an autism screening assessment. Researchers collected responses to 8 factors (e.g. age, gender, country of residence, prior history within immediate family) and 10 questions, which were tailored to each age group. I am using response data from three age groups (4-11, 12-16, and 18+).

Cleaning Steps

I found datasets from the University of California Irvine Machine Learning Repository because I wanted to find publicly available data which has been vetted via peer-review. The data was in .arff file type, which I converted to .csv and stored as data frames in the Jupyter notebook with Pandas file reader. I used .isnull() and .info() and .value_counts() to inspect each data frame and column for missing values, non-uniform formatting and spelling errors, which were added, standardized and replaced, respectively. The three data frames were then merged into one by .concat() and now the visual exploratory data analysis has begun.

Missing Values and Special Characters

Fortunately for data analysis, responses to each of the assessment questions were required and reported in full. I found that one data frame was missing one column that the other two had. Luckily, this was only the 'id' column, which only has meaning to the researchers with personal information about each of the respondents. So, an arbitrary counter column was added so that each data frame would have the same number of columns and column labels.

However, upon inspecting each column, I found that some of the categorical questions (ethnicity, prior history, etc.) were filled with a '?' when an answer was not given. There was no way to fill in this missing data based on responses from neighboring rows, or average column values, because the data frames are full of unsorted entries. I decided to add a category of 'unknown' for these values, and keep them, since the other data in the rows were usable.

Within certain columns, responses were formatted differently from each other, and occasionally contained extraneous leading or bracketing special characters. I removed and/or replaced these characters and labels with standard formatting.

Aside from the missing data, throughout the data, there were many misspellings and mismatched capitalizations that were fixed - these were not found within responses given, but within the original assessment.

Combining Data:

I merged the three data frames into one, after verifying that I had each one containing same number of columns, that the columns were in the same order and labeled identically. Because the columns were identical, I was able to do a simple concatenation of data frames. This project combines three data sets that have the same names for columns that mean different things (for example, the questions that are asked of each participant vary depending on the age range, but are assigned the same question number). I decided to leave these in place, since this project is not delving into each individual question.

Outliers:
None of these data sets had any obvious outliers, so I retained all the rows and columns from the source data. There were a few suspicious points, and they are noted on each figure.

Source data:
Adults
Adolescents
Children