Capstone 1: Intermediate Milestone Report

Problem Statement

This project looks at data gathered via an app designed to provide screening for an autism disorder 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.

The main two questions will be: Question 1) Are there any gender-specific signs to look for when considering an autism diagnosis? For each of the 10 questions, are the responses significantly different by gender when compared to the diagnosis result?

Question 2: Are there any variables that seem to be indicators of an autism diagnosis, all other factors being equal?

In this scenario, the client would be anyone designing or administering such an assessment. If certain questions are significantly correlated to other factors (gender, age, history of autism in the family, etc.), then I would recommend that the answers to those questions shouldn't be weighted into the diagnosis of people who fit into those categories.

Description of Dataset

The University of California - Irvine provides peer-reviewed, publicly available datasets in UCI Machine Learning Repository through the Center for Machine Learning and Intelligence Systems. Note that the assessment questions are not given in this file, but reading through the cited articles, they can be found. The ages are divided into four categories, with different questions given to each age group. Data analysed in this project contains the results of the assessment for children (ages 4-11), adolescents (ages 12-16), and adults (18+).

Data sources:

Children
Adolescents
Adults

Data Cleaning and Wrangling

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

There were a handful of age entries that were missing. For those values, I took the median of the other ages from that age category and replaced the missing values.

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 the 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 and typos

One of the ages was entered as 383 years old. This was replaced with the median age within that age category. There were a few clear typos: a handful of entries within the age category '12-16' years were listed as '12-15' years. I included them in the '12-16', since they would fall within that category. I did not see any other outliers, so retained all the rows and columns from the source data. There were a few suspicious points, and they are noted on each figure.

Inferential Statistical Analysis

Significant variables

Some of the variables that were first assumed to be significant either were found to be not, or were too sparsely represented to be adequately tested. The most significant variables are age group (child, adolescent, or adult), known jaundice at birth (this is a little problematic, because it is not always known, but the default assumption is 'no'), the relationship of the assessor to the assessee, and whether or not there is a family history of autism.

Significant differences between subgroups in your data

Although the assessment data was originally divided into three age categories of people (child, adolescent, and adult), I further divided the data into much smaller age groups and saw that there was a lot of variation of results within the 'adult' category.

Jaundice displays a gender-dependent relationship with assessment results, so these two subsets were teased apart for individual analysis.

Correlations between pairs of variables

Strong correlations exist between the assessment result (dependent) and many independent variables. The largest correlation exists between age and result - adults having a significantly lower result than under 18s. Other

independent variables that show a correlation to result are: whether the assessor was the parent or the person themselves, family history of autism, known jaundice at birth, and between adults and adolescents, as well as adults and children (although not a large difference between children and adolescents). No correlation was found between result and gender, or any meaningful correlation for ethnicity or country of residence.

Although there was quite a positive correlation between self-assessment and assessment by a healthcare professional, the number of data available from healthcare professionals was too small to be statistically confident of the correlation.

Statistical tests

To analyze and look for potential statistical relationships between the dependent variable and multiple independent variables, I used an independent (aka two sample) t-test of the population means, without the assumption of equal variance. This inferential test is appropriate because it requires one continuous dependent variable (in this case the resultant score from the assessment) and one two-level, categorical, independent variable (e.g. gender, age group, jaundice at birth). The null assumption of an independent t-test is that the means of both populations are equal. Usually, this test is used with an aim to reject the null hypothesis and be able to say that the two populations do not share the same mean.

Initial Findings

Average assessment results appear to vary significantly with age (**Fig. 1**), so data was broken up into age categories (4-11 years, 12-16 years, and 18+ years). Further separating the results by gender, it was clear that there was no effect of gender within each age category (**Fig. 2**). There was no statistical difference in the result means between 4-7 years and 12-16 years, but there was a significant difference between the results of under 18 years (n=414, mean=6.31, SD=2.29) and over 18 years (n=686, mean=4.87, SD=2.50), with under 18 year-olds scoring significantly higher on the assessment (p=2.69e-21).

Another statistical difference in assessment results was found between people with a prior family history of autism (significantly higher: p<.001) in the family and those without (**Fig. 3**).

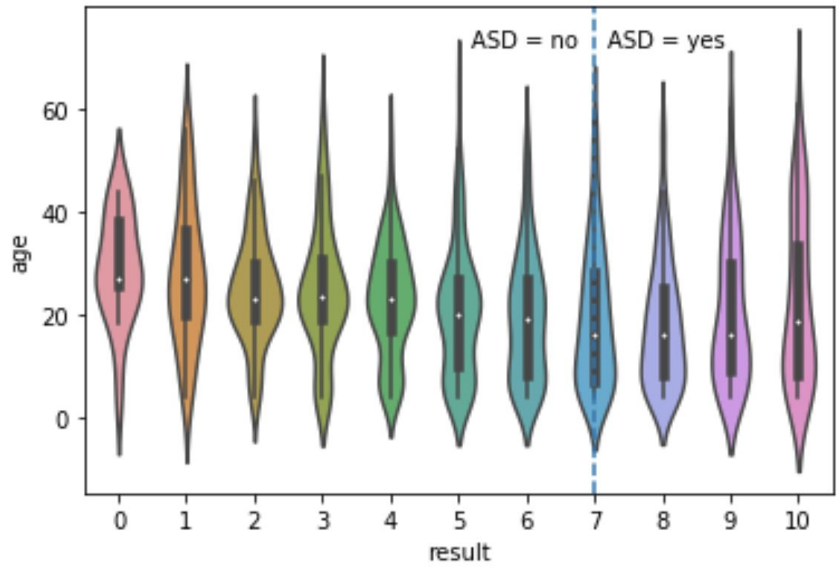

Figure 1. Violin plot of autism assessment result vs age

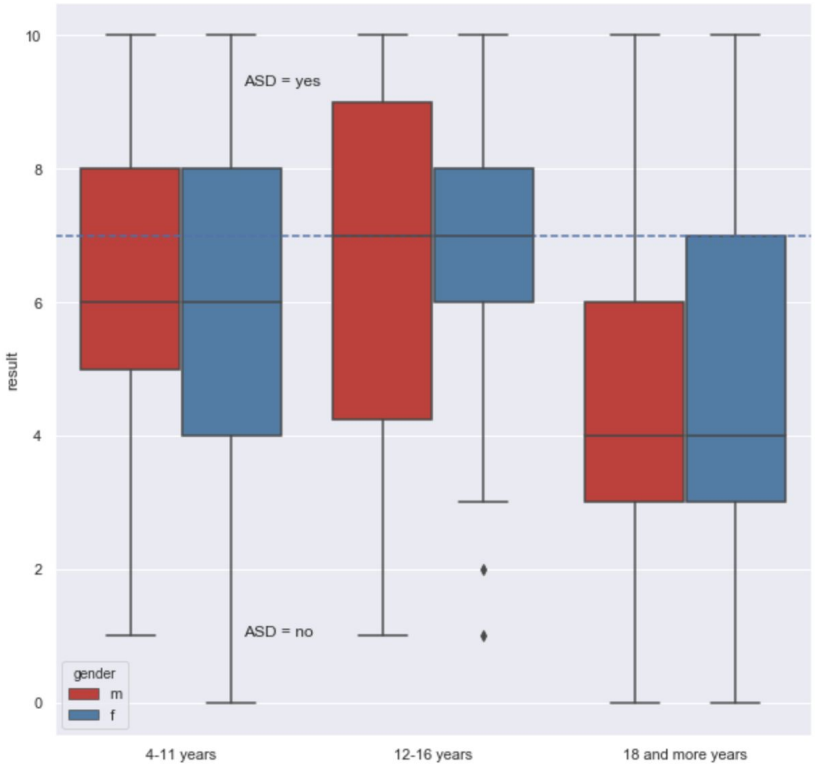

Figure 2. Box plot of autism assessment result vs age group by gender

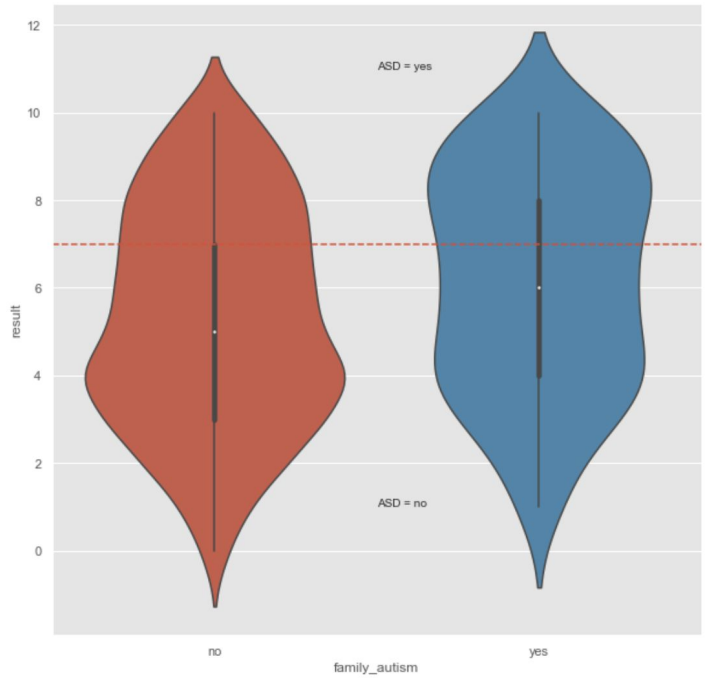

Figure 3. Violin plot of people with a family history of autism (n=154, mean=6.15, SD=2.52) and those without (n=946, mean=5.29, SD=5.50).

There was also a significantly higher result in males known to be born with jaundice (**Fig. 4**, *p*=1.47e-05), while there was no difference for females.

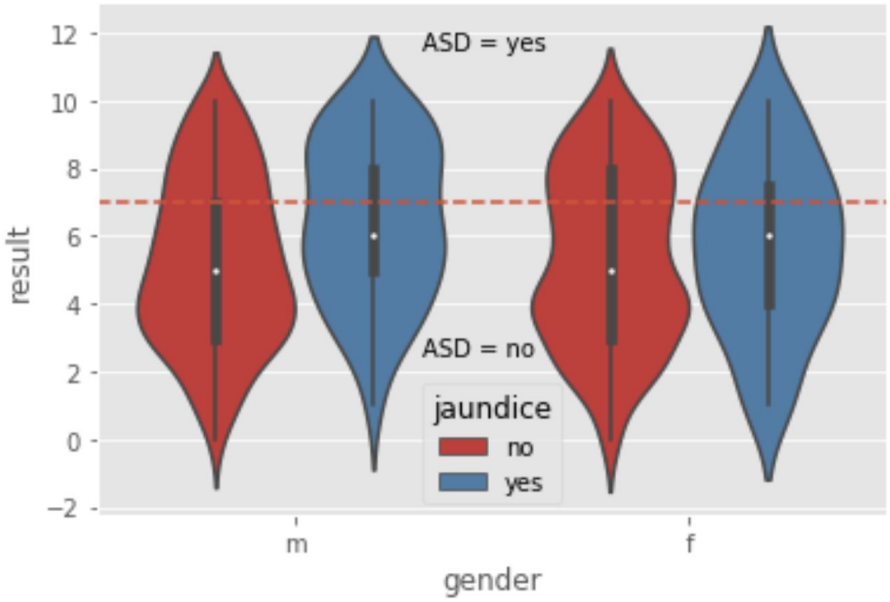

Figure 4. Violin plot of males known to be born with jaundice (n=98, mean=6.42, SD=2.35) and those without (n=527, mean=5.24, SD=2.48), and females (no statistical difference).

The last positive significant difference in results (p=8.91e-08) was for those whom a parent completed the assessment versus those who self-assessed. There looks to be an even larger, higher result for health care professionals completing the assessment, but the number of data points (n= is not large enough to say anything conclusive. (**Fig. 5**)

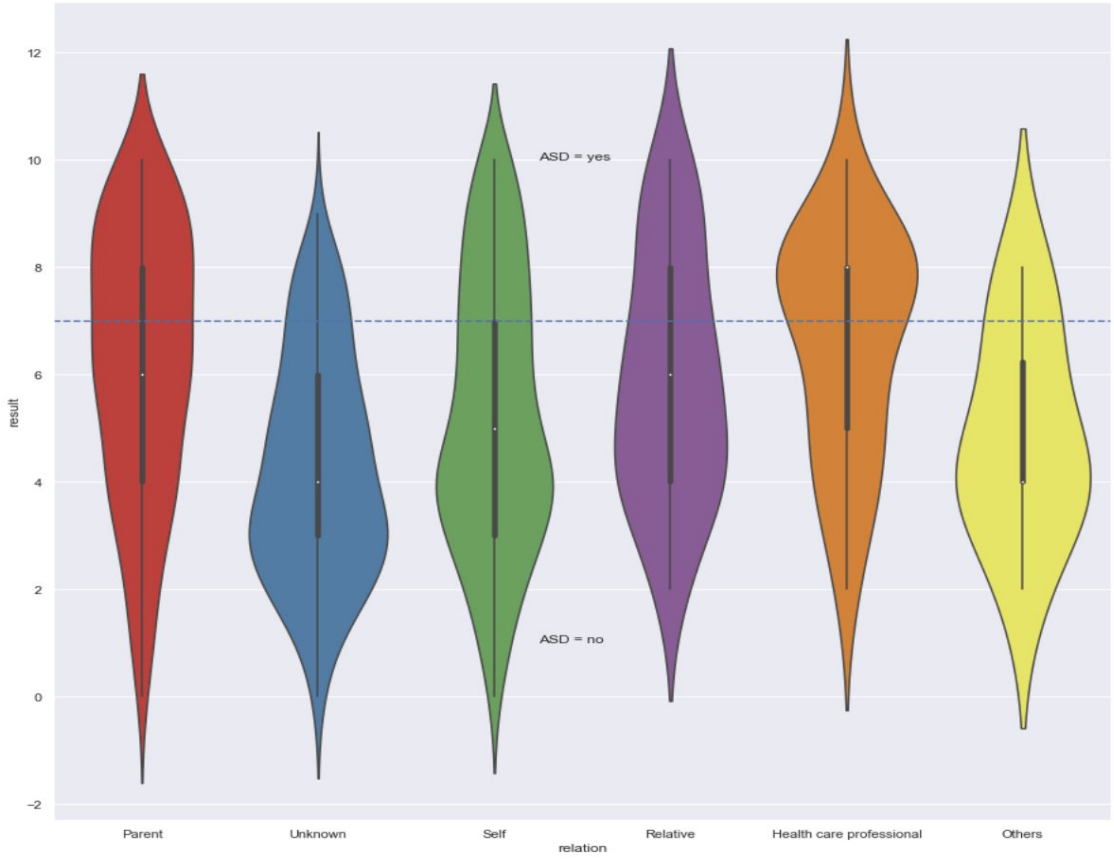

Figure 5. Violin plot of relation of assessor to assessee. A significant difference was found between parents (n=300, mean=6.19, SD=2.510) and those who self-assessed (n=572, mean=5.22, SD=2.53)