**Autism Case Study**

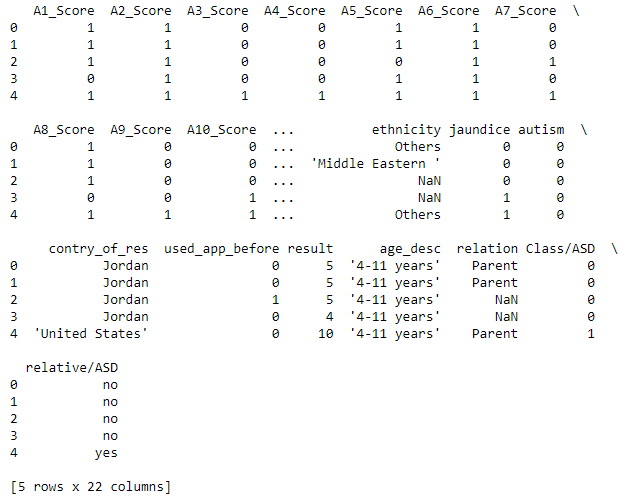
The dataset I am working with is regarding autism. There is a series of ten Boolean questions and a final score adding up the ten questions responses. A number of factors such as age, gender, ethnicity, having a family member with autism, being born with jaundice, and country of residence are also recorded. Based on these factors, is there some type of correlation regarding who will be diagnosed as autistic and who will not. One thing to consider is that this dataset is based on children ages 4 to 11. The Class/ASD variable is whether a relative has autism.

Steps for Graphics Analysis

Step 1: I loaded a dataset I found at <https://archive.ics.uci.edu/ml/datasets.php?format=&task=&att=&area=&numAtt=&numIns=&type=&sort=nameUp&view=table>

Step 2: The dimensions of the table are 292 rows and 22 columns.

Step 3: Display the first 5 rows of data



Items of interest:

All of the scores, jaundice, autism, used\_app\_before, result and Class/ASD fields are Boolean values.

Missing data is represented as “Nan”

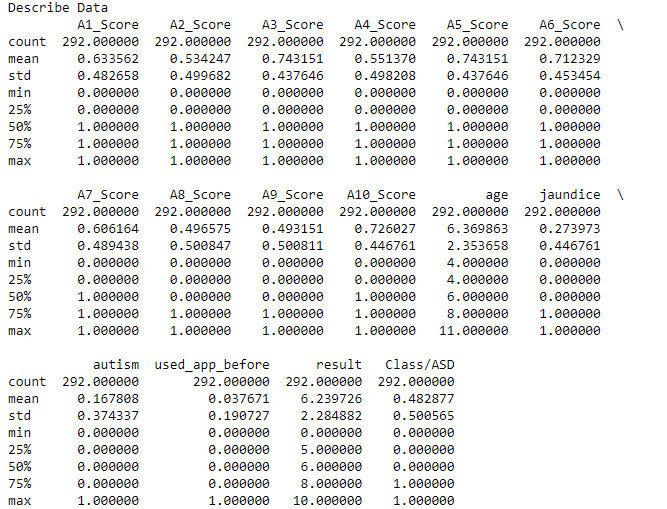
The autism variable is the “target”, the other variables are the “features”

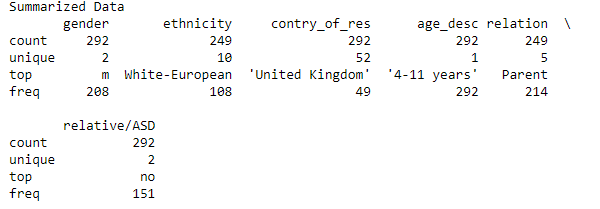
Step 4: Questions to consider regarding who will have autism?

What are the types of variables, are they numerical, Boolean, or categorical? How many different categories are in the categorical variables? What distributions do the numerical variables have?

Are there difference autistic rates in the categories, do more boys than girls have autism? Is there a correlation between autism and any of the variables such as having a family member with autism, or jaundice at birth? Is there a particular answer score that correlates to autism?

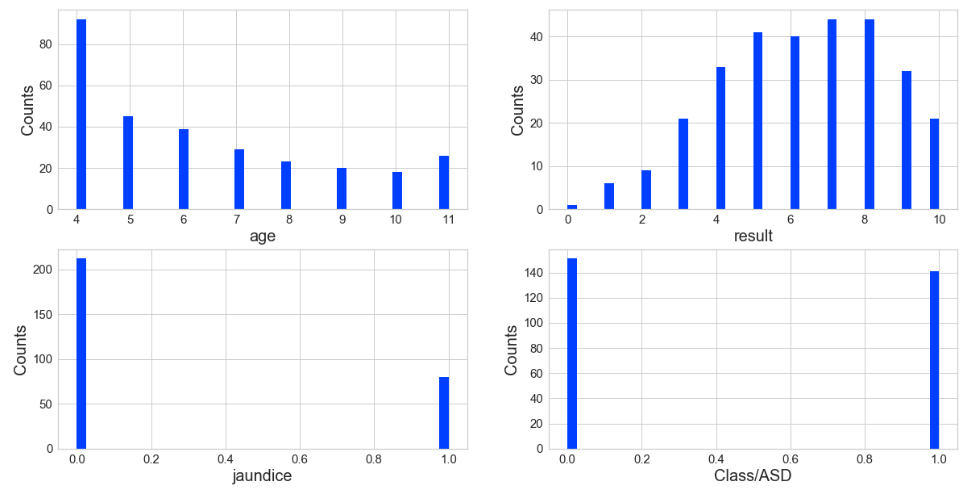
Step 5: Look at the summary information:





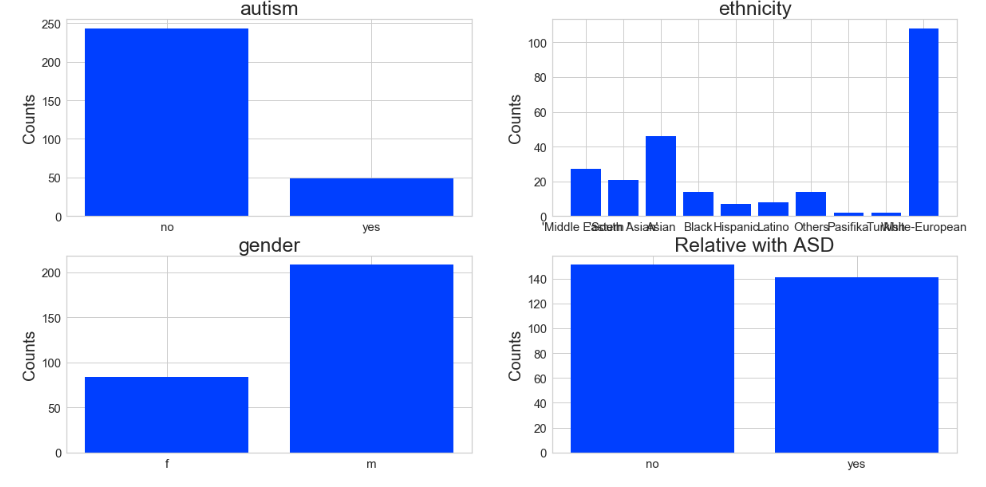
Based on the summary, there are more males than females with autism. The top ethnicity is White-European, and top family member with autism is a parent.

Step 6 Make histograms of the following variables: age, result, jaundice, Class/ASD



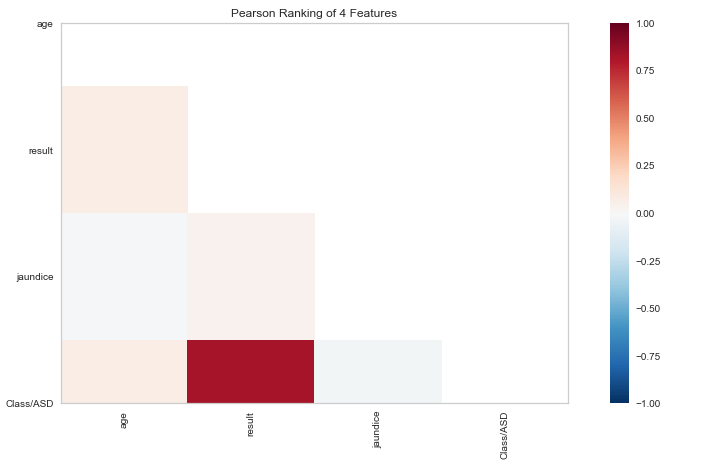
Many of the children were age 4 when diagnosed. The results distribution is normal and slightly skewed to the right. More children were not born with jaundice. The number of children with a relative with autism is fairly evenly split.

Step 7: Make bar charts of the following variables: autism, ethnicity, gender, Class/ASD.



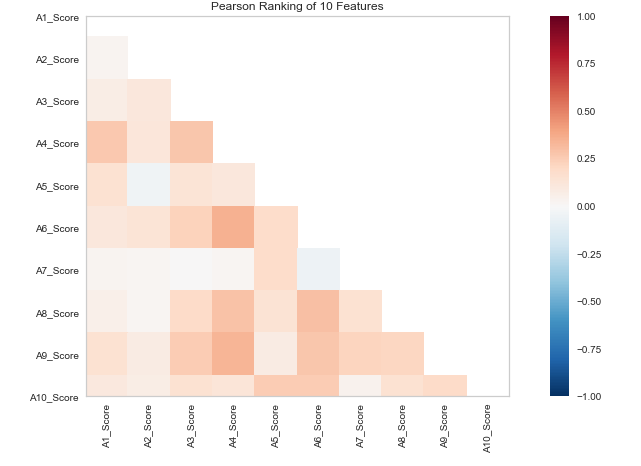
Step 8: Make some Pearson Ranking charts.

First, look for correlations between age, result, jaundice, and Class/ASD



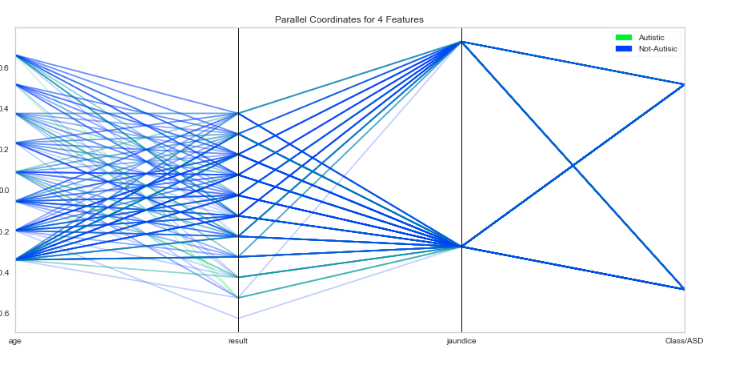
There is very high correlation between the result and Class/ASD variables. This means the results of the questions is correlated to having a relative with autism.

Next look for correlations between the answer scores

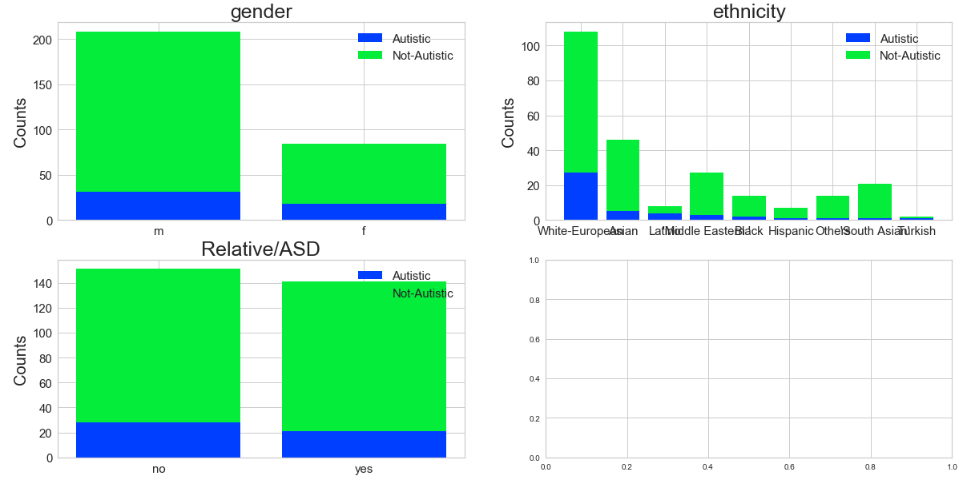


There appears to be some slight correlation between the A4 and A6 scores, as well as the A4 and A9 scores.

Step 9: Use Parallel Coordinates visualization to compare the distributions of numerical variables between children with autism and those without autism.

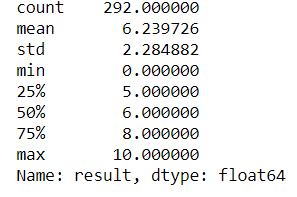


Step 10: Use Stack Bar Charts to compare children with autism to those who don’t based on the following variables: gender, ethnicity, and relative/ASD.

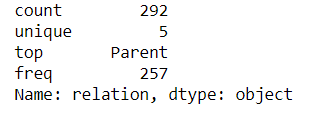
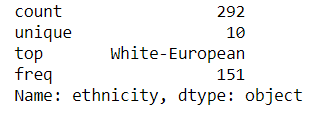


Steps for Feature Reduction and Filling in Missing Values

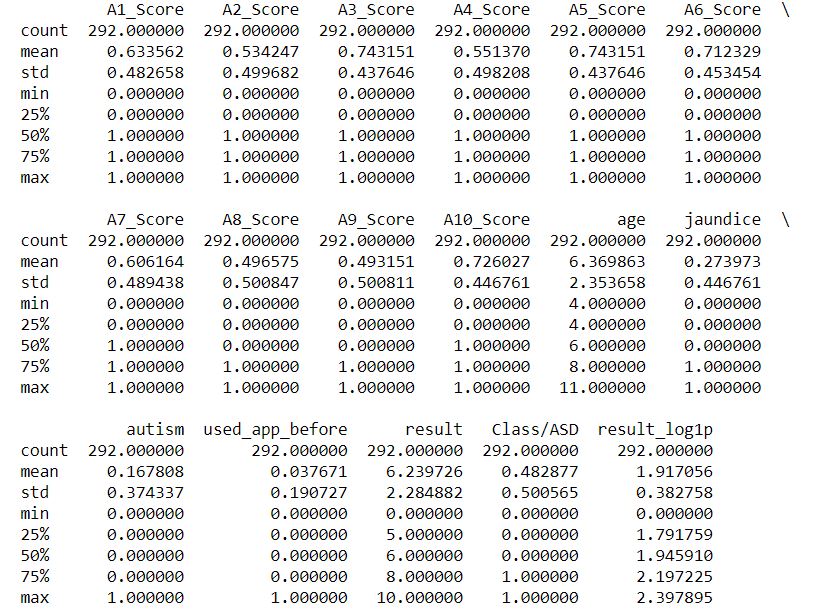
Step 11: It doesn’t appear that any of the fields are missing numeric values, but just to be safe, I ran code to fill in any missing values in the result feature with the median. I printed the results here:



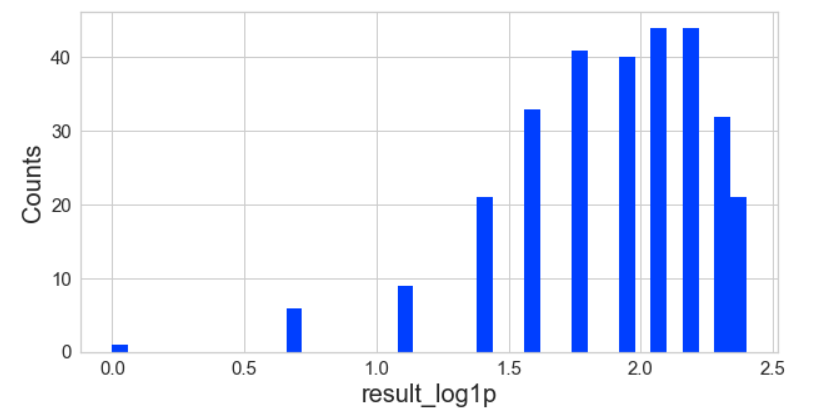
I did have missing values in the features for ethnicity and relation, so I filled those in with the most represented value.

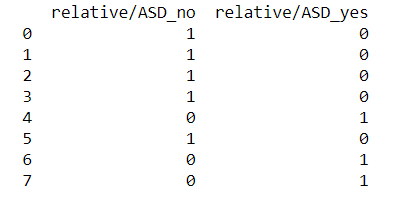
I used a log transformation on the result feature and looked at its distribution.



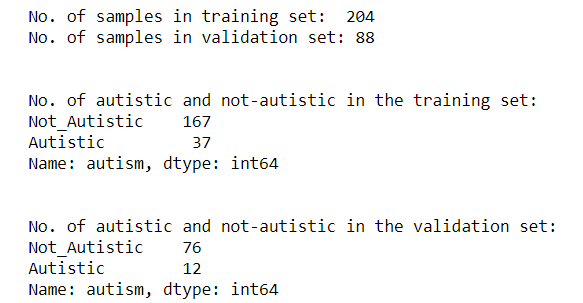
Step 12: I plotted a histogram of the log transformed feature for result, to view its results.



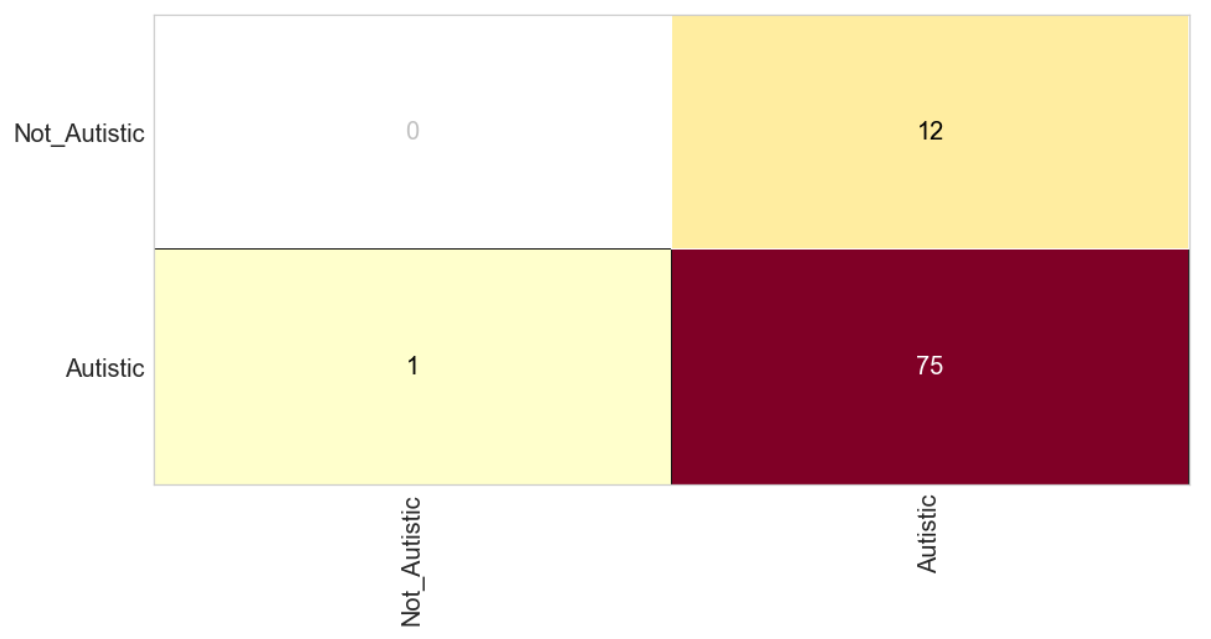
Step 13: I had one feature that had categorical values to be converted to numbers. I created dummy variables for the relative/ASD feature using One Hot Encoding.



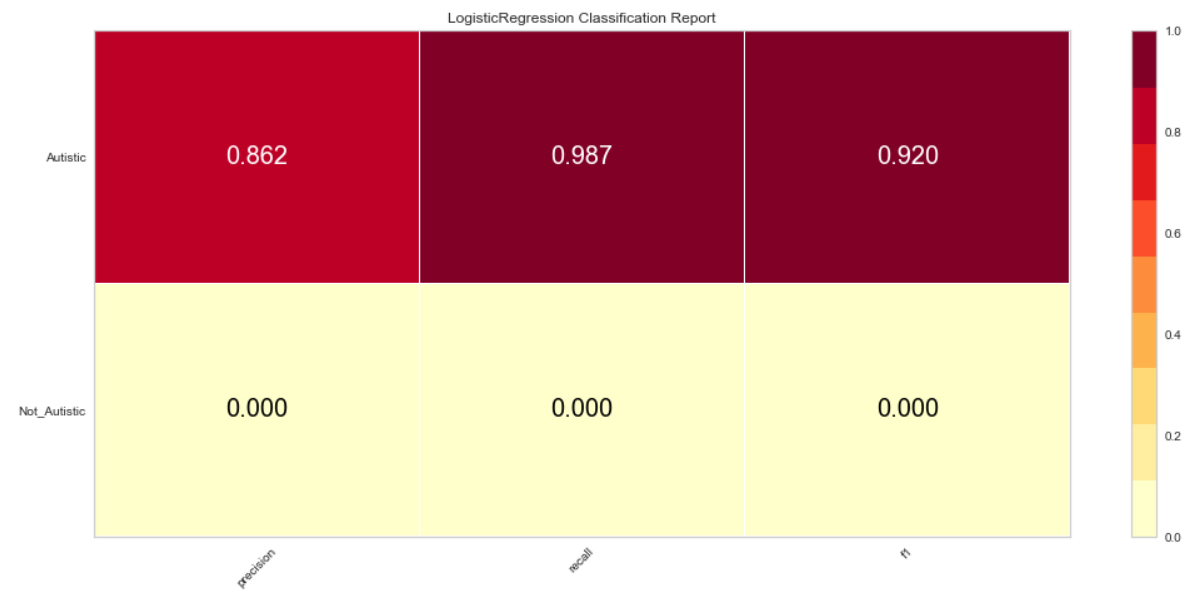
Step 14: I split the data into two training sets so that both contain children with and without autism.

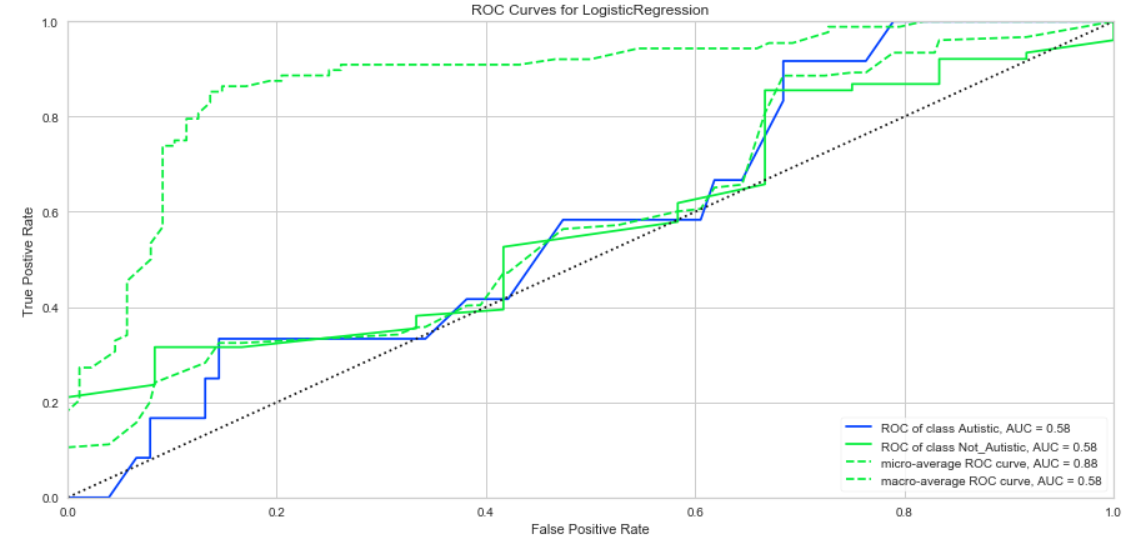


Step 15: I created a logistic regression model to predict whether or not a child would be autistic based on the features of jaundice, age, and the answer scores to the survey questions. I created a confusion matrix to show how accurate the model was with the training and testing data.



I visualized the Precision, Recall, and F1 Score:

I also plotted the ROC and AUC curves:



Evaluation: There is a correlation between a relative with autism and a child having autism. However, there do not appear to be any strong prediction between the age, gender, or ethnicity and whether or not a child will have autism. This supports what we know from the medical field, as they continue to look for patterns and links into what causes autism in children.