Statistical Data Analysis

Chris Murdter

There are many questions to ask when doing data analysis. In most projects one would ask are there variables that are particularly significant? Are there significant differences in the data? Or Are there any strong correlations? My project is a machine learning project using image processing so it is a little difficult to apply these questions to the data I have with my current knowledge. I was looking at color pixel distribution and making comparisons based on the type of image those pixels related to. For example, a pixel distribution of an object in the water with label 0 (no ship) vs a pixel distribution of a ship in the water. Both had similar shapes when plotted as a histogram. I performed a t-test on the arrays of pixel numbers and found that the p-value was 0 or extremely close to zero. I wasn't sure what to do next since having a p-value rejects the null hypothesis. Meaning that there is a significant difference between populations. I then made boxplots of the images. But their means and shapes of boxplots were far from each other so It was hard to process what I could do with that information. Note, I did not do these tests for all images. Only the sample images that I used in the data story notebook of this capstone project.

My main concern with this project will be how accurately I can make my algorithm. Pixel distributions are just one aspect of making this algorithm work. The best way to make it more accurate would be to train the algorithm with certain features. If a machine learning algorithm can tell the difference between a ship in the ocean and a rock or some other shape then it will reduce the amount of false positives I have.