DSC680 Project 3 Check-In

# Katie Briggs Winter 2020

# Domain Surprises

There were a few surprises within the domain perspective. This project looks at the witness of location by latitude and longitude, but also what time of year the sighting happened, the weather and other non-formatted information. The data set is been split into two with the locations in one csv file and the meta data of the encounter of the sighting in a JSON file. There aren’t a lot of data entries, which I also didn’t have in my last project. So, this has me concerned that the accuracy during training and testing will be off.

# Dataset As Expected

For the most part the dataset is as I would have expected. The Bigfoot Research Organization (BFRO) is an enthusiast website without a professional database to hold entries. Due to this, there are a lot of nulls and missing data within the columns. I will need to keep this in mind for the feature selection part of the project. I do need to merge the two separate data sets into one and will create new columns and titles. The next steps are to find correlations with the attributes within the data and the locations of the sightings.

# Adjust Approach

My approach does not require an adjustment at this time. The tools that I am planning on using are pandas, numpy, seaborn, and sklearn, I will plot the season and location against each other to learn insights and which are the most frequent and point toward a migration pattern. If I require an adjustment, I will research the recommender process in more depth.

# Method

My method is working so far. If I need to adjust my methods, I will conduct that as needed. I am planning on following the CRISP-DM and what I have learned over the past year for each process. I have used online resources and will cite the sources throughout my code. After I learn the insights about the time of year and the locations, I will conduct a process for predictors and fitting the predictors and the target. Training, testing, and splitting will be conducted, and I am hoping to keep it at the 30 percent size, but I will adjust as needed due to the size of the dataset. Cross validations will be completed with scoring as accuracy and precision. A confusion matrix will also be conducted with test and predict.

# Challenges

As noted, my major challenges are cleaning the data to where it is useable and being mindful of which sightings occur with time and conditions. The time that is taking to clean the data has taken up most of my time. I have not looked at relationship of variables, correlation, or regression as of yet, and have not focused on a type of data preparation and visualizations. I believe my audience will be anyone that is interested in the topic of Bigfoot, Sasquatch, Skunk Ape or Cyptids and the Travel Channel’s television show Expedition Bigfoot.