Data Science Reflection: Individual Research, Summer 2020

March 13, 2020 was the day all Washington and Lee students were ordered to leave campus due to the spread of coronavirus across the country. Thankfully, I was able to still make something of the following summer as I was allowed to participate in research on campus. Dr. David Marsh, a professor at Washington and Lee, wanted to determine if park characteristics were good predictors of herp species richness. Once this was answered, we sought to find if park-specific or even city-specific factors contradicted the predictor findings. These results eventually came to culminate into a scientific paper published in “Urban Ecosystems”.

The first thing that had to be done was obtain data from a national database known as iNaturalist. iNaturalist is simply a site for users to log species sightings and give descriptions of what they saw. We obtained herp sighting data from large parks across the United States and cleaned the dataset of any observations that were missing data (logs without description or proof of the species in question). After the data was cleaned in RStudio, several scripts were written which helped determine presence or absence of species, species richness, and predictive analysis of species richness by park. Although these scripts were written mostly by Dr. Marsh, our research team was made aware of how they worked and the reasoning behind them. qGIS, an application used for geospatial data analysis, was also used to create .shp files so that readers could see clear boundaries of the parks in question.

I ended up learning several things from this experience which culminated in data analysis for a published paper. The functions I learned in R were used to provide raw count data which Dr. Marsh later was able to use in connectivity and diversity analysis. Other functions used were for cleaning of datasets in order to be able to carryout data counts. qGIS was used as mentioned before, to create .shp files so readers could see the boundary of areas under investigation. By creating these boundaries, we could either include of exclude data points and observations used for analysis. Finally, Excel skills came into play in that we could get a preliminary view of how our data would look when modeled and examined closer (such as finding a SUM, AVERAGE, MEDIAN and MODE measurement).