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 had to check that they ran smoothly before being published. 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. For one, I relearned how to clean datasets in RStudio, something I had forgotten how to do. I also learned how to implement complex models in RStudio and conduct statistical tests. Several new packages and functions were also learned and implemented in the research done. qGIS was not an application I had heard of before, much less used. By the end of the research experience though, I knew how to create various geometric shapes that told me important factors about an area of land, such as area, perimeter, and tree cover, for example. I have not used qGIS since the research experience but have used the R packages and functions in various ways since then. Examples include modeling inequality in the United States or penguin population growth in South America.