My Growth as a Data Scientist

Over the course of my time in STAT 331, I have grown as a data scientist significantly. Of my time in college, this is the first course I have taken that has exposed me to the type of work I would be doing as a future data scientist. In particular, my skills in filtering data sets and making creative visualizations of my data have become stronger during my time in STAT 331.

In the beginning of the class when I had learned basic functions for cleaning and editing a data set, my skills for filtering data were very limited. I only knew how to filter rows with specific characteristics using the filter function, however I did not understand the complexity of filtering on a vector of values before completing and editing Lab 4. However, upon completing my edits, I learned how to effectively use filtering joins on data sets to create new data sets based on certain criteria, and then join or exclude those new data sets with my original data set. These two functions provide a new way of filtering information in a data set when the information to include or exclude is found in a vector of values. Initially, I was also creating many intermediate data sets that I did not need to eventually obtain a new data set that filtered out only the information I wanted. By eliminating these data sets, I made my code a lot more concise and easy to follow for someone else who was reading it. Lab 4 Question 2 exhibits the changes I have made.

In Lab 5, I learned to go above and beyond in making creative plots using my data using the forcats package. During the beginning of the course, I only knew the bare minimum of how to create visualizations for categorical and numerical data, and I did not take the time to explore the various arguments I could include to make my plots more creative and readable. On question 1 of Lab 5, I discovered how to reorder boxplots by a specific variable using the forcats reorder function. Rather than outputting all of my boxplots in a meaningless order, the reorder function allowed me to display the distribution of species weights in descending order, which displayed a better narrative to the viewer. I also utilized the fct\_collapse function in question 3 of Lab 5 to create a segmented bar graph of the number of rodents captured each day, however the categories were collapsed into the categories of “weekday” or “weekend”. Using these new functions I learned to elevate my visualizations of data demonstrates how my skills as a data scientist have improved.