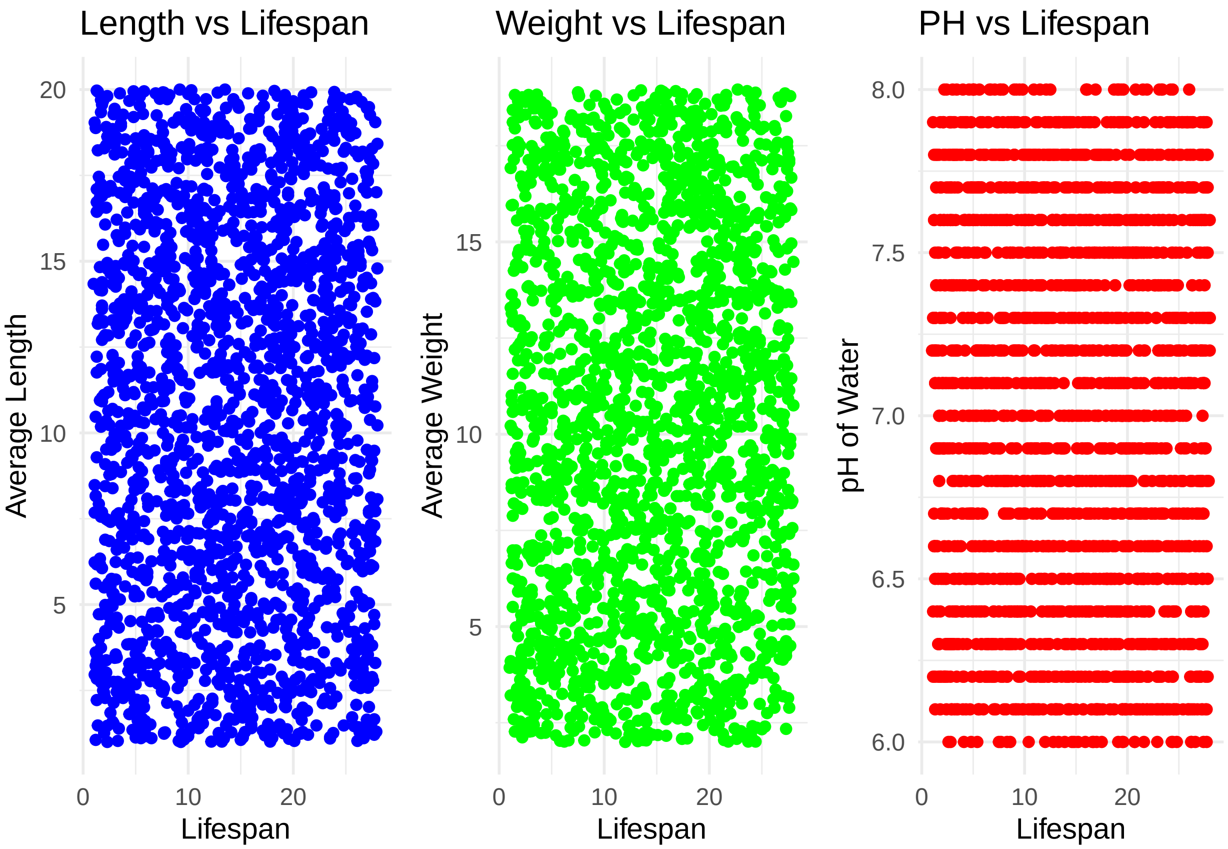
Not Project 1

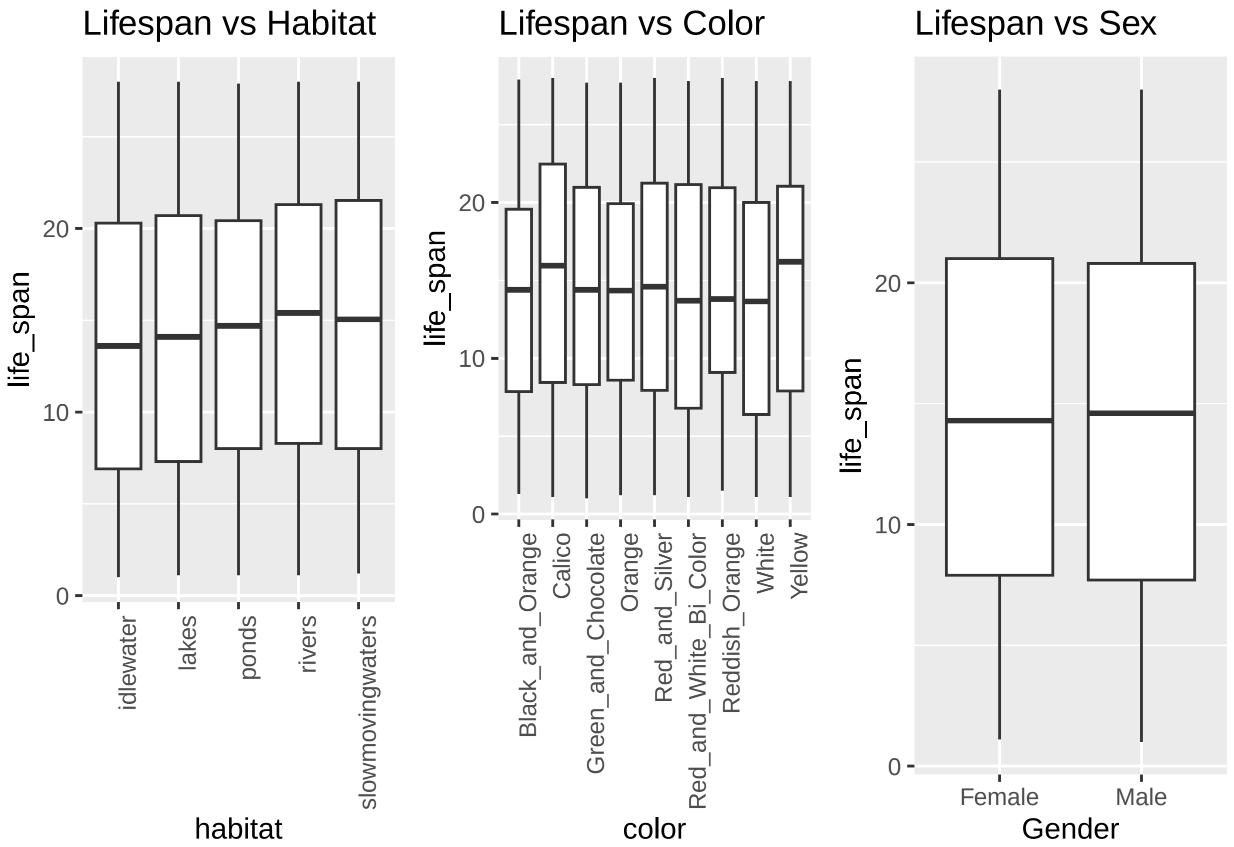
Caleb King

This project covers the life expectancy of a comet goldfish depending on various factors. Comet goldfish, a popular freshwater pet, often exhibit varying lifespans influenced by multiple components. I will be using care, environment, and genetics to learn more about comet goldfish and statistically predict the lifespan of these fish based on the data provided. I would like to learn more about what goes into taking good care of a comet gold fish. When I was young I had a comet gold fish that lived for 10 years and others that passed away within months. I want to learn more about the factors that go into a comet goldfishes life to better improve the quality of life for goldfish I have in the future. The data was also nicely organized with good variables to learn more about the factors that go into a goldfish's life.

As for cleaning the data, I had to clean much of the data at the start of the project. This included changing the variable types of Habitat and Color to factors. This change was made so I could be able to preform numerical analysis on the variables. The next change I did when cleaning my data was change the Gender variable to a factor as well with the levels of male and female. The original data set used true or false variables which made the results pretty confusing. For the original data set I had to change some of the names of the columns using the colnames() function. I changed the names of average weight and length because the column names had incorrect units of measurement listed. Lastly I had to eliminate the goldfish with NA for some of the factors. To do this I used na.omit() function on the data that had been cleaned. After these changes to the data, I was ready to start modeling and performing statistical analysis.



From the visualization above it is clear that most of the data is completely random. There is not much to decipher from these graphs except that the data is completely random. There is no discernible pattern or relationship between the variables being plotted. Since there is no correlation between these variables, it suggests that these variables are completely independent from each other.



From the visualization above you can see that the data is pretty uniform around the center of the graph. There is not much variation between these variables and lifespan. This would lead me to believe that these variables are not the greatest for predicting lifespan. There is a slight increase in lifespan between the different habitats but that may also be because of the number of goldfish in each habitat. There is not an equal amount of distribution between these variables so that may be the resulting issue.

If I was given more time I would try to test the other predicting variables against life span. In my previous project I learned that most of the predictors are not statistically significant in predicting lifespan. I used a regression model, 95%CI, Linear model, Gam model, Regression tree, Ridge regression model, and Lasso model. I didn't have too many challenges analyzing the data but the difficulty was from trying to predict life span.

https://github.com/KingCalebDS