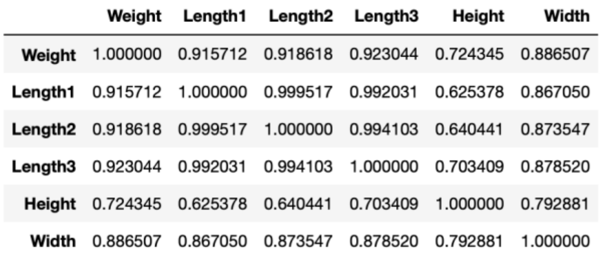
Madison Christiansen
Final Project
DSC530 Week12

Outcome of your EDA

The data was sourced from Kaggle, which contains many open-source quality data sets. My data is a record of 7 common fish species in fish market sales. The variables include species name, weight; the weight of the fish in grams, lenght1; vertical length in cm, lenght2; diagonal length in cm, length3; cross length in cm, height; height in cm, and width; diagonal width in cm. To clean the dataset, I took the only the columns that were needed for the analysis. Here I made a correlation matrix to better understand how the variables correlates and can help show the relationships among them.



	Weight	Length1	Length2	Length3	Height	Width
Weight	1.000000	0.915712	0.918618	0.923044	0.724345	0.886507
Length1	0.915712	1.000000	0.999517	0.992031	0.625378	0.867050
Length2	0.918618	0.999517	1.000000	0.994103	0.640441	0.873547
Length3	0.923044	0.992031	0.994103	1.000000	0.703409	0.878520
Height	0.724345	0.625378	0.640441	0.703409	1.000000	0.792881
Width	0.886507	0.867050	0.873547	0.878520	0.792881	1.000000

When choosing the right statistical methods, I wanted to help visualize the data and provide lots of graphical visuals. This included the histograms, scatterplots, line graphs, and a threshold line when needed. Using the mix of multivariate and univariate graphical methods I believe shows a good analysis of the data.

For this project I used fish market data to make a complete analysis on the given data. During the analysis I feel that something I missed was understand the probability mass functions for each variable. When calculating it for all the graph it looked good although the x-axis was the variables. I also see that having the three different length variables might have not been needed. Another huge thing that I would have benefited from was a better understanding of the variable meanings. A few variables that could have helped the analysis would be age/maturity and sex. These two would allow for the better understanding throughout the whole analysis and could offer more routes to be taken with the regression analysis and the analytical distribution. These two variables I think would have also been better than using the three different length variables. When looking at the permutation test the threshold line was giving me a tad trouble, but I believe this shows a decent representation of the data. Looking at the analytical distribution I think the data is along the model line although it is not a great representation, and this could be an incorrect assumption. Throughout this project I had trouble finding a good dataset that I fully understood. I also think the PMF and CDF were two areas that needed more attention and understanding.