18 November 2022

For our first meeting, we looked over our dataset. We looked at the important context of the dataset. Some notable features we found with the dataset were that the age of the abalone could be found using the number of rings + 1.5. We also noted that it could be a problem for us, or at least our conclusions would be less concrete given that the context of weather, species, or location is not provided. Before we had any more discussion about it, we decided to do a quick visualization of the attributes that we would later analyze. To do this, we displayed the distributions of all 8 attributes. From seeing the distribution, we found that there was something irregular in our dataset. This was in the age distribution. It shows that the distribution steadily increases until the age of about 12 before sharply decreasing at the very next point. Without enough context, we can only guess the reason for this was that was the age the abalone would be harvested. Next, we wanted to produce some important questions about the data. After some discussion about what could be insightful to visualize, we landed on attributes of age. We found that this would be useful to find because age is important for abalone. Because as we saw in the distribution of age, there was a certain age that the population would sharply decline. It is obvious that the people who harvested the abalone did not harvest them based on their age but on their size attributes (most likely length). We concluded that if we were to find correlations to this age and other attributes of size, the data might have practical uses such as using age to accurately predict sizes in certain populations, seeing if the populations are growing at a regular or abnormal rate. For the rest of this meeting, we discussed the tools needed for this analysis and visualization. Some tools listed were boxplots, scatter pilots, and heat maps. We will use Matplotlib, seaborn, and pandas.

25 November 2022

For our second meeting, we did a lot of the implementation. As we had already discussed what we would need to do in the prior meeting, there was not as much to talk about. This meeting focused on reviewing materials we learned in class to make the implementations. Alex had created a heat map of the attributes while Corentin created the scatter plots of the attributes. After these plots had been made, we again began our discussions, trying to draw conclusions from the data. We saw, to our surprise, that the correlations on the heat map looked very weak when looking at age correlations. Most of the correlations landed between .5 and .6. Looking at the heap map, it seems our hypothesis was incorrect, and that age cannot be determined by size. But for this reason, we also made scatter plots of the correlations to see them on a graph. From here, there was an obvious trend in attributes of age and size. As age increased, so did all the other attributes. From this, as the age continues to increase, the correlation becomes much weaker, with the scatter plot becoming very spread out, but the trend still exists. We created more scatter plots focusing on attributes that have high correlation in addition to our lower correlation age plots. These were to show in our presentation the difference against stronger correlated attributes such as length and diameter, and how we expected our age correlations to look similar. We also focused on weight correlations which also were strong. At the end of this session, we discussed other potential visualizations we could add. This brought us to the topic of marine life. From Alex’s experience with fish and other aquatic life such as clams and snails, he mentioned the idea that in lots of marine life, it is hard to determine the sex of the animal based on attributes of size. Instead, many look to appearance or other determining factors. From this, we suggested that abalone would follow this pattern, and the size attributes of abalone would not be determined by the sex of the individual. If true, then it would help us establish age as a stronger factor in finding correlations, while also answering more questions about abalone. If this is not true, how much of a role does sex play and how can we determine the sex or the other attributes from it? To visualize and analyze this, we plan to use box plots, scatter plots, and heat maps.

2 December 2022

For our last meeting, we did our visualizations of sex. In our last meeting, we did not consider the percentage of the population that was either male or female, or an infant (the sex cannot be determined as the individual is too young). If one group is dominating the population or one group is underrepresented, then this might cause the data to be skewed or affected by some other factor. To see this, we represented the population in either M, F, or I. What we found is that the population was roughly split into thirds based on the three categories we made. This will be good for our data and analysis. To visualize this as we did our visualizations and analysis of age, we decided to show the correlations of each group within their own heat maps, as well as add them to the same scatter plots. Deviating from our previous implementations, we also created a box plot to see the difference in age between the sexes. We also looked at the differences in the mean of all the attributes (including age) between M, F, and I, and found that, while there wasn’t a significant difference for males vs females, infants were smaller in all attributes, which makes sense as they are still infants. After the implementation was complete, we began our discussions to draw conclusions from the data. In our heat maps, we noticed that the heap maps for our males and females had weak correlations between age and the other attributes, with the females having the weakest correlation. But what we noticed with our infant heat map was that there was a strong correlation between age and the other attributes, higher than both males and females. When looking at our scatter plot, we can see that as the individual is younger or smaller, and they fall into the category of infant, their correlations between attributes are very strong, with the points being very close. But as the size increases, and the age increases, we can see that the correlation between attributes starts to weaken. In this scatterplot, we can see both males and females losing correlation at about the same rate, keeping the same level of weak correlation. This shows that not only is age directly correlated to the other size attributes, especially at a younger age, but sex plays a weak role in determining these attributes of size.

As we have finished our visualizations and analysis, we still need to work on creating a presentation and GitHub. For our PowerPoint, it was designated that Alex would complete the first half (age visualizations), and Corentin would complete the second half (sex visualizations). For creating the GitHub, we have been keeping up with everything as we have been doing our meetings, so all that is left is to compile it all. This is our last meeting, so to complete all the preparations for presenting and handing our project in, we will work on our designated PowerPoint topics for the presentation and compilation. If anything more needs to be discussed, it will be on our own time.