Given how hard it was for me to analyze the first Capstone Project and the poor overall experiences I had, for my second project I was determined to find the right dataset, something that makes sense biologically and I can have a relatively straightforward time analyzing it. II found three candidates below:

1. UKOA cruise D366 bioassay dataset
   1. This is a dataset collected on the effect ocean acidification impacts on sea surface biology and biogeochemistry in Northwest European Shelf Seas
   2. Pros:
      1. A lot of data, each separated by a tab
      2. Detailed annotation of the variable details
      3. Overall a fairly robust dataset
   3. Cons:
      1. A lot of data to sort through and make enough sense to write a story about it - this is especially hard given this field of biology is foregin to me and would require a lot of extra work to understand the implications of each factor
      2. Following up on the previous point, the data, given how big it is, is also very “messy”. Again, for someone not familiar with ocean biology, this dataset may ultimately be too big to chew
2. Seawater carbonate chemistry and biological processes during experiments with haploid and diploid life stages of Emiliania huxleyi, Calcidiscus leptoporus and Syracosphaera pulchra dataset
   1. This is a dataset collected on the effect of elevated partial pressure of carbon dioxide (pCO2) on Emiliania huxleyi (Lohmann), Calcidiscus leptoporus (Murray and Blackman), and Syracosphaera pulchra (Lohmann)
   2. Pros:
      1. The information in this dataset makes sense - no additional digging around needed
      2. Comparisons can be made between the different species of water plants
      3. Biologically relevant
   3. Cons:
      1. Limited size
      2. Some pre-data processing I have to do as the data wasn’t downloaded cleanly for some reason
3. Dataset on birds apparently and their anatomy
   1. <https://www.kaggle.com/zhangjuefei/birds-bones-and-living-habits>
   2. This is a dataset on birds’ anatomy and their ecological niche (habitats)
   3. Pros:
      1. A very robust dataset - out of all 3 this is my favorite one
      2. Biologically relevant - I believe comparisons between the different groups of birds can be made
      3. Overall a strong candidate for a dataset for me to analyze.
   4. Cons:
      1. The dataset is present on Kaggle

Out of the three, I like option 3 much better than either 2 or 1. Between 2 or 1 I prefer option 2 just due to it being less complex - option 1 simply has way too much data, with way too many variables for me to comfortably perform any meaningful analysis on a field for which I am foregin.