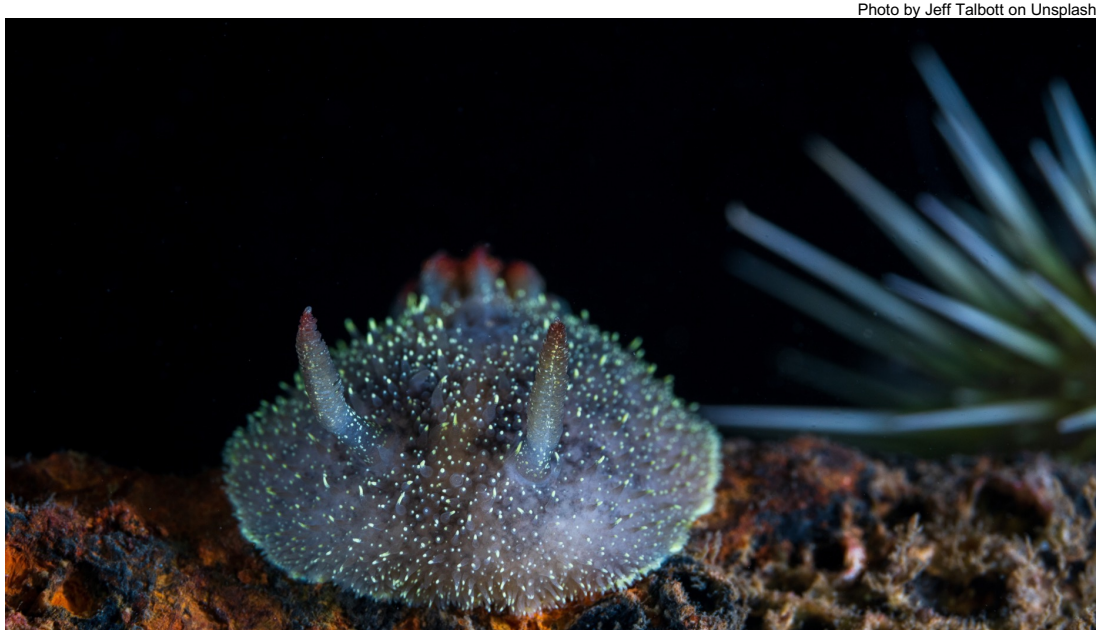
Data Science

Prof. Dr. Markus Breunig

Homework in preparation of

Lecture 5 – Outliers

Photo by Jeff Talbott on Unsplash



Pandas (Part 1)

This upcoming lecture will introduces you to the concept of ***Outliers***, which are data points that are different from the other data points. Outliers can on one hand be annoying and reduce the quality of the ML model, but on the other hand can lead to novel discoveries and strongly support assumptions.

Linear Digression is/was a Data Science podcast, covering a wide variety of subjects. It is available on all relevant podcast apps and also on the web at http://lineardigressions.com/.

We also continue the introduction to the ”pandas” library in the "Python Data Science Handbook" by Jake VanderPlas [https://jakevdp.github.io/PythonDataScienceHandbook/], as we will be discovering and handling outliers with the help of pandas.

This homework assignment consists of the following parts and expected time:

- Listen to the episode "How Outliers Helped Defeat Cholera” (in your preferred podcast app or on the web at http://lineardigressions.com/episodes/2016/2/23/how-outliers-helped-defeat-cholera). It contains a well-known example of outlier analysis, will give you an idea on why outliers can be extremely interesting. (expected time: **11min**)
- Read the rest of Chapter 3 (Data Manipulation with Pandas) (i.e. from *Combining Datasets* to the end) - **60min**
- Complete the notebook ”h05a...” - (expected time: **90min**)
- Complete the notebook ”h05b...” - (expected time: **90min**)