

## **About the dataset:**

Please refer to the official [Github page](#) for details and license information. The details below have also been taken from there.

Artwork: [@allison\\_horst](#)

**Palmer Archipelago (Antarctica) penguin data:** Data were collected and made available by [Dr. Kristen Gorman](#) and the [Palmer Station, Antarctica LTER](#), a member of the [Long Term Ecological Research Network](#).

Thank you to Dr. Gorman, Palmer Station LTER and the LTER Network! Special thanks to Marty Downs (Director, LTER Network Office) for help regarding the data license & use.

### **License & citation**

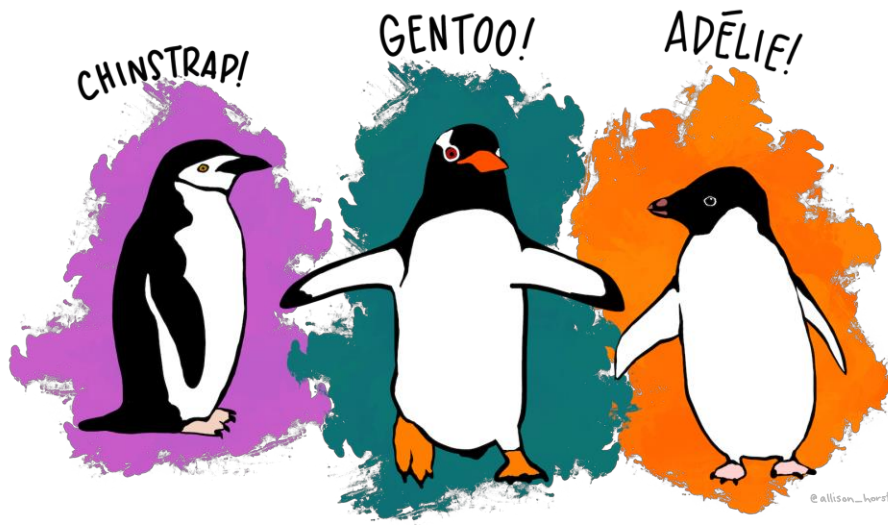
- Data are available by [CC-0](#) license in accordance with the [Palmer Station LTER Data Policy](#) and the [LTER Data Access Policy for Type I data](#).
- Please cite this data using: Gorman KB, Williams TD, Fraser WR (2014) Ecological Sexual Dimorphism and Environmental Variability within a Community of Antarctic Penguins (Genus *Pygoscelis*). PLoS ONE 9(3): e90081. doi:10.1371/journal.pone.0090081

### **Summary:**

The data folder contains a csv file: penguins\_size.csv

- penguins\_size.csv: Simplified data from original penguin data sets. Contains variables:
  - species: penguin species (Chinstrap, Adélie, or Gentoo)
  - culmen\_length\_mm: culmen length (mm)
  - culmen\_depth\_mm: culmen depth (mm)
  - flipper\_length\_mm: flipper length (mm)
  - body\_mass\_g: body mass (g)
  - island: island name (Dream, Torgersen, or Biscoe) in the Palmer Archipelago (Antarctica)
  - sex: penguin sex

### **Meet the penguins:**



What are culmen length & depth?

The culmen is "the upper ridge of a bird's beak" (definition from Oxford Languages).

**Task:** Predict the class of penguin species

**Questions to Answer:**

- Perform a detailed exploratory data analysis on the dataset
- Experiment using two different ratios of training, validation and test data ie 60-20-20 & 80-10-10. On the two different split ratios do the following
  - Implement Grid Search CV to find optimal hyperparameters for any 3 algorithms (out of LR, SVM, MLP, RF, Boosting)
  - Plot the learning curve using the learning curve function from scikit-learn to analyze the model performance. The plot should show the training score and cross validation score against the number of training examples.
  - Analyze the results on Validation set and Test set and mention which model performed the best and why?
  - Compare the performance of models(using precision, recall, accuracy, latency).
- What was the best proportion or split ratio of data from the set of experiments you conducted and why?