# Pregistration of An Already-Existing Data Set

Fiel Dimayacyac

2022-09-17

---

## ## Study Information

**Title: Clustering Size Measurements of Extinct and Extant Bird Species**

**Research Questions**

- RQ1: Do bird species cluster by environment?
- RQ2: Do Extinct and Extant species cluster separately?

**Hypotheses**

- RQ1: Because environments (in this case living on islands) have been shown to have size-specific effects on organisms; i.e., the island effect, I expect organisms to cluster by environment.
- RQ2: I expect that since the paper found there were statistically significant differences in size and lifestyle between extinct and extant birds, I will find the same trend.

## ## Data description

**Datasets Used**

To answer these questions I will use the data set provided by Fromm & Meiri 2021 of the bone sizes and body size estimates of extinct and extant bird species. They found evidence of 469 species of bird driven to extinction by humans. They found that extinct species tended to be larger by median estimated body size and that island birds were larger than mainland birds. However, within taxa, these size differences were only slight.

**Data availability**

The dataset is publicly available.

**Data Identifiers**

https://doi.org/10.5061/dryad.1rn8pk0tb

#### Access Date

Initially downloaded September 8, 2022.

#### Data Collection Procedures

Data was collected via literature review. They recorded geographic range, ability to fly, and body size of each of the species. When necessary they estimated mass using ML-based linear regressions.

## ## Variables

#### Measured Variables

K-means: The amount of clusters found during the clustering process.

#### Unit of Analysis

Will be including all of the species in this analysis. I will only exclude species if the data entry is missing a significant amount of data.

[Question title]

#### Missing Data

Missing data will not be included when performing PCA of body measurements.

#### Statistical Outliers

Statistical outliers in this sense will be species that do not cluster with any of the groups. In this case I will not include them in any conclusions.

#### Sampling Weights

Sampling weights will not be included.

## *Knowledge of data*

#### Prior Knowledge

I have no personal prior knowledge of the data except for reading the conclusions that the authors came to.

## *Analyses*

#### Statistical Models

For the hypothesis testing I will be using a k-means clustering algorithm through the tidymodels R package. I will assess the performance of the clustering by analyzing the total of within sum squares by k size. I may perform dimensionality reduction before analysis to reduce the data to two dimensions via PCA.

### Inference Criteria

For all research questions we will use the best amount of k-means clusters determined by total of within sum squares to determine if birds cluster by any particular category.

### Reliability and Robustness Testing

To assess our analysis we will measure the total within sum squares.

### Exploratory Analysis

I will also explore if any other factors such as order, flightlessness, or wing/body size ratios impact the analysis.