Costing Portal Data Preliminary Analysis

CIL Lab

2024-04-24

# Objective:

As a lab group, we developed some code to clean the costing portal dataset

# Background:

The initial dataset was a mess!

# Methods

First we standardized data fields for analysis. We did things like make assumptions about XXX, and define Y and Z.

We then performed some summary statistics to better understand the dataset.

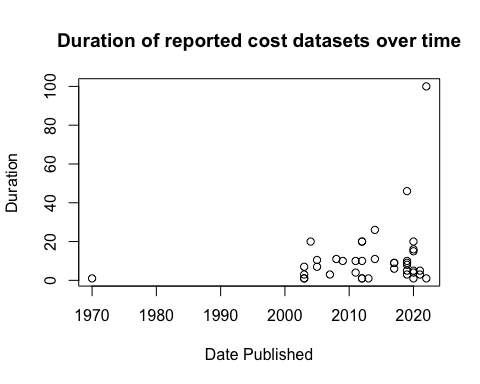
Our dataset has 78 rows and 34 columns. There are 23, 1 years with studies, the oldest is from 1970, the newest is from 2022.

Specifically, we explored the number and proportion of studies that included cost data that involved protected area (PA) establishment or management costs, and also mention of the different categories of costs across the studies.

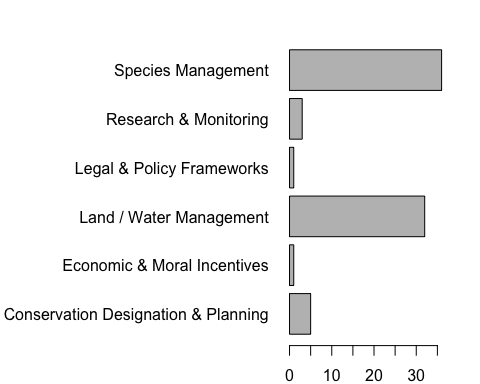
|  | count | proportion |
| --- | --- | --- |
| Yes | 40 | 0.52 |
| N/a | 5 | 0.06 |
| No | 32 | 0.42 |

|  | count | percent |
| --- | --- | --- |
| PA costs | 15 | 0.19 |
| Capital costs | 40 | 0.52 |
| Consumable costs | 23 | 0.30 |
| Labor costs | 34 | 0.44 |
| Overhead costs | 6 | 0.08 |

We explored the relationship between how many years of cost data were collected for a study (its duration) and the time it was published.



We also explored which conservation actions were described by the studies and how they were distributed.



We also explored how some of the numeric variables related to each other. This is useful for exploring predictor variables. For instance, if we wanted to formulate a regression exploring how the variation in how many years of cost data were reported in a study (duration) was related to the date it was published, whether capital and labor costs were included and whether it described protected area establishment costs (these are the numeric variables we have at the moment, no hypothesis as to why these particular predictors).

