Project Write-up:

In this project, I aim to build a prediction model of every country's threatened species. I minor in environmental science and always want to use my data science knowledge for data building and collection in EE field for protecting the environment & sustainability, so I choose this topic for my project. I downloaded my data form https://data.un.org/_Docs/SYB/PDFs/SYB66_313_202310_Threatened%20Species.pdf, which is official data containing vertebrates, invertebrates, plants and total threatened species in each country in the years of 2004, 2010, 2015, 2019, 2020, 2021, 2022.

At first, I need to write a clean file to clean the downloaded csv data. The original csv is messed up in terms of categorical data, there's an extra T25 column with numerical data, some extra commas in between lines of data, unmatching header and data with source & footnote, extra commas in header. I made the clean file function a separate module, reading the input and process rows with various field counts, avoiding UnequalLengths error. Then, it trims the white space between extra commas while using the writer builder to generate the modified cleaned file. Then, in header, make sure "T25" and "Series" are removed while the order is changed to "Region/Country/Area","Year","Threatened species","Value","Source","Footnotes". In data body, it iterates through each line and removes the first extra column. Then, since the data has values like "1,875" that are supposed to be usize but read as String, I convert them back by removing the extra "" and commas. I trim it again, swapping col_field[4] and [5] in liens that have 6 fields, which are sources and footnote to match the header. Then, I call mod clean in main to generate the new cleaned csv file.

In main function, I use serde(rename) to make sure the struct name matches with header names, which will become the keys in hashmap. I created a hashmap of reading the values, with outer key country name, inner key species type, and value for (year, value). In read data function, the data is deseralized into CountryData struct and put into a hashmap. The keys are country names, and the values are nested hashmaps where the keys are species types ("Total", "Plants", "Vertebrates", "Invertebrates") and the values are vectors containing year-value pairs.

In the fit_model function, it performs linear regression to model the relationship between years and the number of threatened species for a given country and species type. The years and values are converted into f64 types, formatted into matrices where x is Array2 and y is Array1, used x and y to create a new dataset and fit the model into linear regression. After fitting the model, it calculates the mean absolute error of the model to measure prediction accuracy. In function predict next years, it uses the fitted linear regression model to predict the number of threatened species for the next n years, based on the last available year in

the dataset. It generates predictions by inputting future years into the regression model and returns a vector of tuples containing the year and the predicted value.

For the main function, I make the users input the country they wish to predict at and the type of species types ("Total", "Plants", "Vertebrates", “Invertebrates”) to get output predictions for next n years. The n can is customized to change the total numbers of prediction, but if the input if invalid it will be defaulted to 10 times. After main, I created a mock csv for separate sample testing. In test, I created 3 unit tests – one reads the mock CSV file and ensures that the data contains the expected country, one fits a model on a small set of data (year-value pairs) and checks that predictions are generated correctly, one tests the prediction of future years using the linear regression model and verifies the results.