Finalized Research Question:

What characteristics of an environmental protection project are the most important in determining how much money it receives from multilateral climate funds, and how well can the total amount of money raised for these projects be predicted?

Data overview:

Data Source:

We have sourced a dataset from Climate Funds Update (https://climatefundsupdate.org), an independent website supported financially by the Heinrich-Böll-Stiftung foundation in Washington, DC, and ODI, a think tank. The Climate Funds Update tracks the amount of money pledged and deployed by multilateral climate funds to assist developing nations in adapting "climate-resilient development trajectories.

Brief Data Summary

We are principally concerned with the data published by the Climate Funds Update related to funded projects. This data set is presented in an Excel file that has mostly already been curated. The dataset contains 3,428 observations of funded projects, with 26 columns/variables.

The data set includes both numerical and categorical columns. The variables that we are particularly interested in to answer our research questions are:

Predictors: variables related to the nature and geographic location of each project. Important predictors are Funding Type, Country, World Bank Region, Income Classification, Theme, Sector, Approved Year. All predictors are categorical except Approved Year (numerical).

Response: the total amount of funding for each project, which is numerical, in millions of USD.

In this EDA report, we will mainly focus on the above variables of interest. Several other variables in the data set are not directly related to our research questions. Their data types and characteristics are summarized in the attached ipynb.

Descriptive Statistics

Descriptive statistics of the numerical variables (approved year and total funding) are provided below.

	Approved year	Total Funding
count	3383.0	3428.000000
mean	2016.241502	9.764399
std	4.499021	25.935438
min	2003.0	0.000000
25%	2013.0	0.700000
50%	2017.0	2.920000
75%	2020.0	8.185000
max	2023.0	378.000000

Data preprocessing and exploration methods

We have performed the following steps to proprocess and clean the data:

- **Data missingness** in World Bank Region and Income Classification: Since the missingness happens when entries of Country column are Global, Regional, or Multiple-country, we perform both data dropping and imputation to address this issue (detailed in attached notebook).
- **Duplicate values** that differ only by lowercase/uppercase in the categorical columns: we perform case conversion to eliminate those duplications.
- **Data transformation**: We perform one-hot encoding on all the categorical variables of interest.

Data preprocessing and exploration methods of other variables are included in the attached ipynb.

Observed Patterns/Trends in the Data:

The analysis of climate funding data reveals several key patterns:

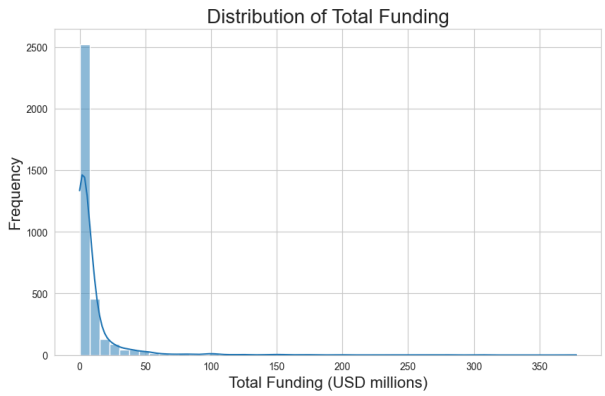
1. Funding is predominantly directed through multilateral channels, focusing on developing regions and lower-income countries that require financial assistance for climate resilience.
2. Funding distribution across different themes emphasizes adaptation and general mitigation, though projects focused on REDD are comparatively underfunded.
3. The trends over time show that while there was significant growth in climate initiatives and funding until 2019, the COVID-19 pandemic caused a notable decline, reflecting the need for resilient systems to sustain climate action during global crises.

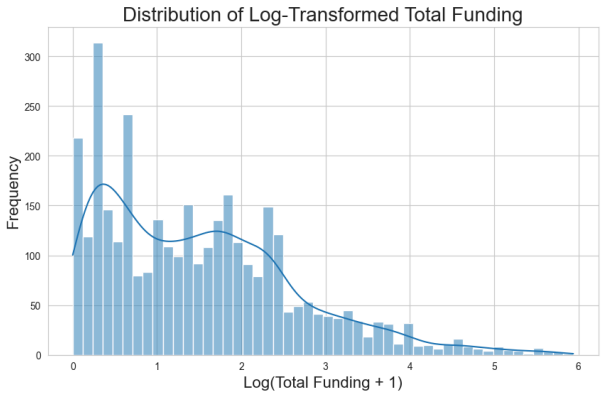
The combined analysis highlights both the progress made in supporting climate resilience and the vulnerabilities in funding mechanisms that must be addressed to ensure sustained efforts against climate change, regardless of broader global disruptions.

Insights Helpful for Modeling

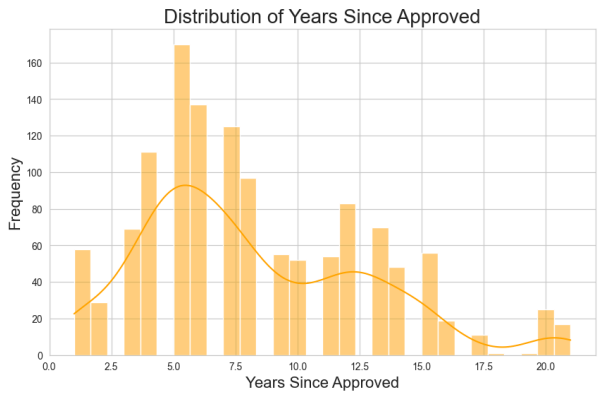
In this section, we summarize the insights from our attached ipynb that we believe will be most useful for modeling purposes.

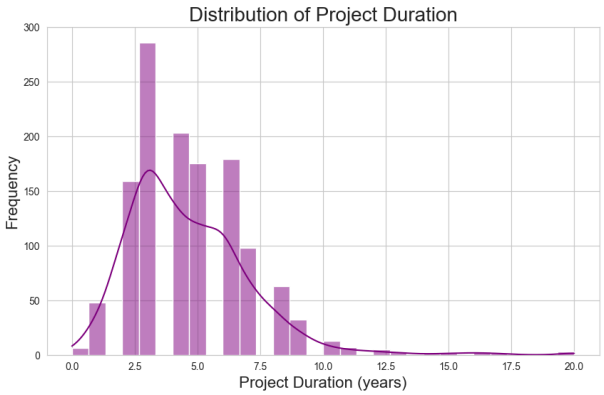
The ‘Total Funding’ variable we aim to predict is right-skewed with a few projects receiving disproportionately large funding. This skewness can violate the assumptions of linear regression, which we aim to use for our baseline model. Applying a log transformation will reduce the impact of outliers and make the data closer to approximate normality.





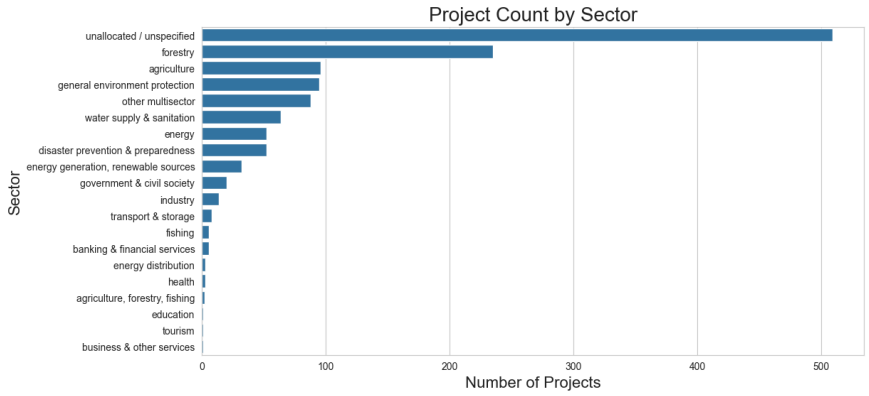
There is a positive trend in the number of projects over time, indicating a growing commitment to addressing climate change. To help our model better learn this pattern, we could transform ‘Approved Year’ into ‘Years Since Approved,’ which provides a relative measure that better captures the temporal dynamics. Instead of having ‘Approved Year’ as an effective categorical variable, ‘Years Since Approved’ would provide a single continuous feature and simplify the feature space. Additionally, we can create a new feature representing the project duration from the ‘Approved year’ and ‘End year’ to catch more temporal trends.





Another insight that we should be aware of for modeling is the presence of outliers. Using the IQR method, there are 140 instances of ‘Total Funding’ deemed as outliers. We could consider removing or capping these outliers. These outliers do, however, make up 11% of the data, indicating that they are relevant and should be taken into account. Either way, it is important to be aware of the outliers during modeling to make informed decisions.

We deal with many categorical predictors, some containing significant class imbalances. The following plot displays the class imbalance for the Sector predictor. We could apply resampling techniques or adjust class weights to stop our model from becoming biased toward majority classes.



Baseline model or implementation plan:

Because we are predicting a numerical amount of funding, and because all of the independent variables in our dataset except for Year are categorical, we will implement a Lasso regularized linear regression model. Lasso regularization will help with reducing our feature space to only those features that are the most meaningful in making predictions.

All of our categorical features must be implemented with one-hot encoding; however, the “Theme/Objective” field includes a value, “Multiple Focii,” that seems to group many themes into that one value. We’d want the model to utilize a more diversified “Theme” input, so to satisfy our added requirement for APCOM209a, we’ll use Bag of Words and a clustering technique to derive a new categorical variable for the project’s theme / objective based on project names, their project summaries (text field that includes a description of the project), and a “keywords” field.