Data Overview & Preprocessing Steps

This section provides an overview of the data used in this research and the preprocessing steps applied prior to its utilization, including filtering, regression imputation, labeling, and feature derivation. Normalization has not been performed at this stage and will be applied according to each subsequent task's requirements. Although these procedures are time-consuming and technically demanding, they are crucial for ensuring the accuracy and reliability of all subsequent analyses and thus should be performed with the utmost care.

1. Data Description

The initial dataset was obtained from the World Bank, an international financial institution and open-data provider, and includes the following features:

Name	Description
Water productivity	GDP in constant prices divided by annual total water withdrawal
Total freshwater withdrawals	Annual total freshwater withdrawals in billion cubic meters
Sectoral withdrawals (agricultural/domestic/industrial)	Percentage of freshwater withdrawals used for each sector (agricultural/domestic/industrial)
Water stress	The ratio between total freshwater withdrawals and total renewable freshwater resources
Precipitation	Average precipitation in depth (mm per year)
Private investment	Investment in water and sanitation with private participation(current US$)
Natural disaster	Percentage of the population affected by droughts, floods, and extreme temperatures (Average from 1990 to 2009)
GDP per capita(PPP)	Gross Domestic Product per capita, in purchasing power parity (current international $)
Population	Total population by each country
Income level	Classification of each country according to the World Bank's income groups (e.g., Low income, Lower middle income, Upper

	middle income, High income).
Region	Geographical grouping of each country as defined by the World Bank (e.g., East Asia & Pacific, Latin America & Caribbean, Sub-Saharan Africa, etc.).

Note: “Natural Disaster” contains only the average value of the annual data from 1990 to 2009.

2. Data Preprocessing

2.1. R Libraries used

Library	Purpose
tidyverse	Data manipulation and visualization
rnaturalearth & rnaturalearthdata	Retrieves and manages natural Earth geographic data
sf	Spatial data handling for map visualization

2.2. Missing Data Assessment and Handling

Even though the dataset originally spanned 1960–2023 for 266 countries, a substantial share of values turned out to be missing. In order to assess and address these gaps, missing data proportions by feature were first evaluated and then visualized using a horizontal bar chart.

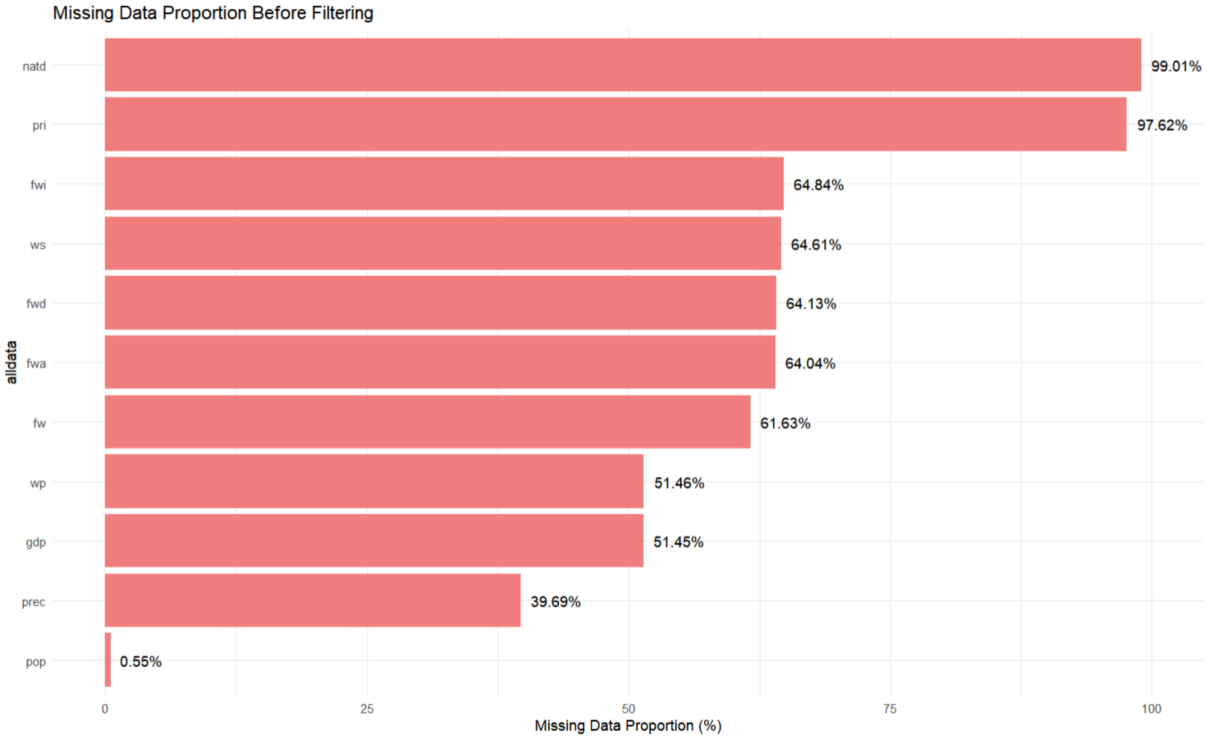


Figure - Missing Data Proportion by Feature

According to the generated plot, natural disaster and private investment exhibited extremely high missing rates (99.01% and 97.62% respectively), while population had an exceptionally low rate (0.55%). The remaining features displayed a missingness range of 39.69%–64.84%. Features were subsequently classified into two subgroups based on their missing value proportions: “normal” and “problematic.” The “problematic” subgroup consisted solely of natural disaster and private investment, while all other features fell into the “normal” category. This grouping was designed to facilitate the use of distinct missing data handling approaches by subgroup.

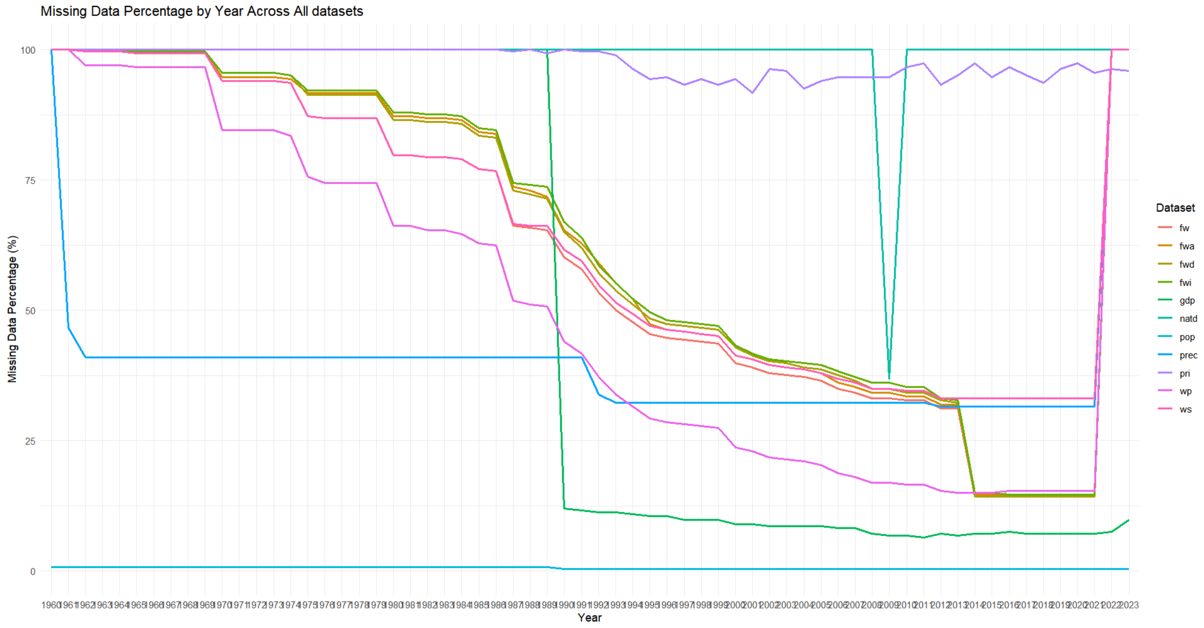


Figure - Missing Data Percentage by Year Across All Features

This line chart tracks missing value proportions for each feature from 1960 to 2023. Missingness is extremely high for all features in the very early decades but steadily declines. This downward trend likely reflects the gradual establishment of standardized reporting protocols, expansion of national statistical agencies, or improvements in data collection technologies. Two anomalies stand out in the chart: the 2009 spike in natural disaster data due to the aggregation of 20 years of data and a modest rebound in missingness for 2022 and 2023 in most features, likely due to reporting delays for the most recent time.

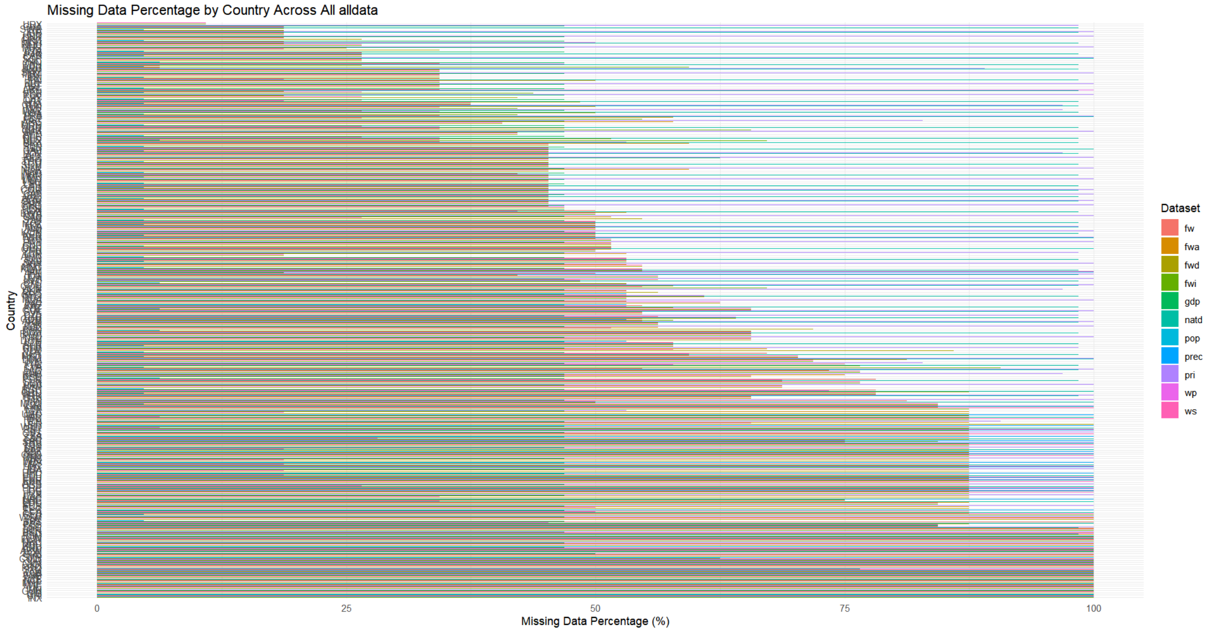


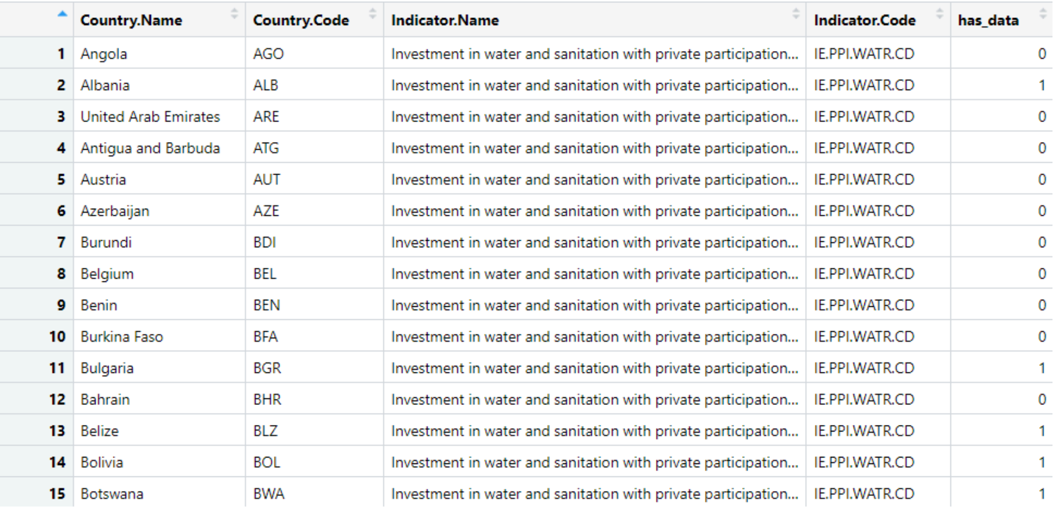
Figure - Missing Data Percentage by Country Across All Datasets

This horizontal bar chart displays the share of missing values for each country. It is observed that missingness varies considerably by different countries as well, possibly due to a lack of statistical infrastructure, political or social instability, or reporting lags.

```
> year_threshold
[1] 0.55
> country_threshold
[1] 0.57
```

R Console Output - Thresholds for the Normal Dataset

Since missingness was concentrated in earlier years and in certain countries, a two-stage filtering strategy was applied to the "normal" datasets (all features except natural disaster and private investment) to address missing values. First, years with more than 55% missing values were dropped by intersecting, across all nine features, the sets of years whose missing value proportions did not exceed the 0.55 threshold. Next, countries exceeding 57% missingness were removed by the same logic. Finally, each dataset was reduced to retain only those valid years and countries, yielding filtered "normal" data frames with substantially improved completeness. This approach also ensures that features share a common set of years and countries, facilitating consistent, panel-wide analyses and seamless data frame manipulation.



	Country.Name	Country.Code	Indicator.Name	Indicator.Code	has_data
1	Angola	AGO	Investment in water and sanitation with private participation...	IE.PPI.WATR.CD	0
2	Albania	ALB	Investment in water and sanitation with private participation...	IE.PPI.WATR.CD	1
3	United Arab Emirates	ARE	Investment in water and sanitation with private participation...	IE.PPI.WATR.CD	0
4	Antigua and Barbuda	ATG	Investment in water and sanitation with private participation...	IE.PPI.WATR.CD	0
5	Austria	AUT	Investment in water and sanitation with private participation...	IE.PPI.WATR.CD	0
6	Azerbaijan	AZE	Investment in water and sanitation with private participation...	IE.PPI.WATR.CD	0
7	Burundi	BDI	Investment in water and sanitation with private participation...	IE.PPI.WATR.CD	0
8	Belgium	BEL	Investment in water and sanitation with private participation...	IE.PPI.WATR.CD	0
9	Benin	BEN	Investment in water and sanitation with private participation...	IE.PPI.WATR.CD	0
10	Burkina Faso	BFA	Investment in water and sanitation with private participation...	IE.PPI.WATR.CD	0
11	Bulgaria	BGR	Investment in water and sanitation with private participation...	IE.PPI.WATR.CD	1
12	Bahrain	BHR	Investment in water and sanitation with private participation...	IE.PPI.WATR.CD	0
13	Belize	BLZ	Investment in water and sanitation with private participation...	IE.PPI.WATR.CD	1
14	Bolivia	BOL	Investment in water and sanitation with private participation...	IE.PPI.WATR.CD	1
15	Botswana	BWA	Investment in water and sanitation with private participation...	IE.PPI.WATR.CD	1

Figure - Country Level Labelling in Private Investment Data

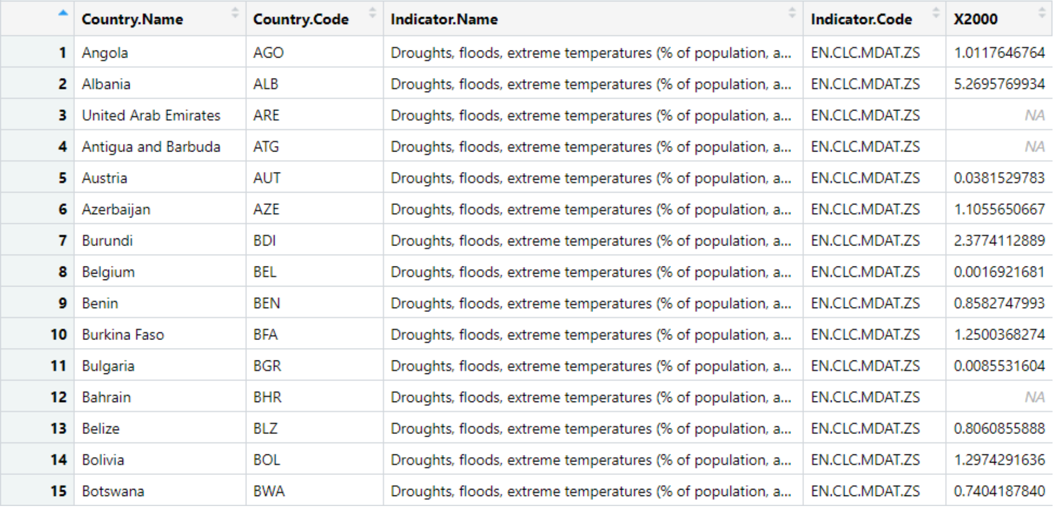
For the "problematic" dataset, a different strategy was adopted. Since the private investment data exhibited a 97.62% of missingness, it was converted into a binary label at the country level. Any country with at least one valid observation received a label of 1, while countries with no data at all were labeled 0. This approach is intended to avoid numeric imputation, which would be unreliable given the extreme sparsity and lack of domain knowledge.

```
> table(pri_labelled$has_data)
```

```
  0   1
205  61

> table(pri_labelled_final$has_data)

 0  1
83 32
```

R Console Output - Class Distribution of Private Investment Labels Before and After Filtering

Initially, the number of countries with no records at all was significantly higher than the number of countries with at least one record. Therefore, the labeled panel was filtered to include only those countries retained by the normal dataset, in order not only to ensure consistency across datasets, but also to further improve class balance. This approach is based on the assumption that countries removed earlier for excessive missingness in normal features are also likely to not have private investment records. After applying this filter, the class size ratio narrowed from 3.4:1 (205 zeros vs. 61 ones) to 2.6:1 (83 zeros vs. 32 ones), making the potential subsequent classification task more manageable. However, class imbalance is still significant—positives account for only about 28% of cases—so applying resampling techniques (e.g. SMOTE) or incorporating class weights in the model would be advisable.

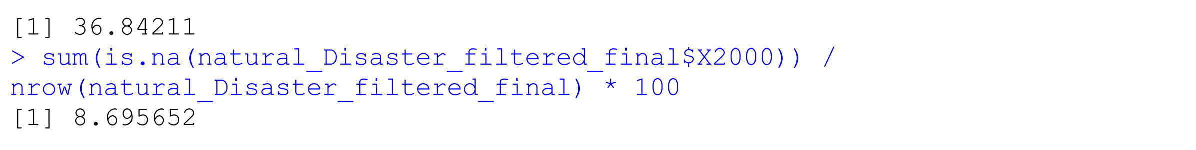


	Country.Name	Country.Code	Indicator.Name	Indicator.Code	X2000
1	Angola	AGO	Droughts, floods, extreme temperatures (% of population, a...	EN.CLC.MDAT.ZS	1.0117646764
2	Albania	ALB	Droughts, floods, extreme temperatures (% of population, a...	EN.CLC.MDAT.ZS	5.2695769934
3	United Arab Emirates	ARE	Droughts, floods, extreme temperatures (% of population, a...	EN.CLC.MDAT.ZS	NA
4	Antigua and Barbuda	ATG	Droughts, floods, extreme temperatures (% of population, a...	EN.CLC.MDAT.ZS	NA
5	Austria	AUT	Droughts, floods, extreme temperatures (% of population, a...	EN.CLC.MDAT.ZS	0.0381529783
6	Azerbaijan	AZE	Droughts, floods, extreme temperatures (% of population, a...	EN.CLC.MDAT.ZS	1.1055650667
7	Burundi	BDI	Droughts, floods, extreme temperatures (% of population, a...	EN.CLC.MDAT.ZS	2.3774112889
8	Belgium	BEL	Droughts, floods, extreme temperatures (% of population, a...	EN.CLC.MDAT.ZS	0.0016921681
9	Benin	BEN	Droughts, floods, extreme temperatures (% of population, a...	EN.CLC.MDAT.ZS	0.8582747993
10	Burkina Faso	BFA	Droughts, floods, extreme temperatures (% of population, a...	EN.CLC.MDAT.ZS	1.2500368274
11	Bulgaria	BGR	Droughts, floods, extreme temperatures (% of population, a...	EN.CLC.MDAT.ZS	0.0085531604
12	Bahrain	BHR	Droughts, floods, extreme temperatures (% of population, a...	EN.CLC.MDAT.ZS	NA
13	Belize	BLZ	Droughts, floods, extreme temperatures (% of population, a...	EN.CLC.MDAT.ZS	0.8060855888
14	Bolivia	BOL	Droughts, floods, extreme temperatures (% of population, a...	EN.CLC.MDAT.ZS	1.2974291636
15	Botswana	BWA	Droughts, floods, extreme temperatures (% of population, a...	EN.CLC.MDAT.ZS	0.7404187840

Figure - Overview of the Natural Disaster Data

Natural Disaster data—originally recorded as a 1990–2009 average in the 2009 column—was processed by extracting the X2009 column and renaming it to X2000, which is the midpoint year of 1990–2009. The resulting data frame was then filtered to the same set of valid countries retained in the normal dataset, in order to ensure consistency across all panels.

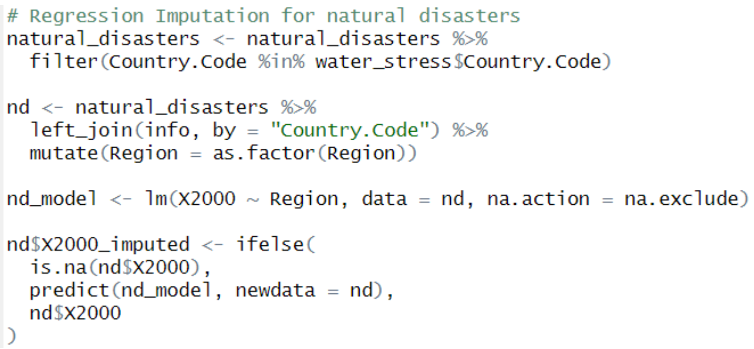
```
> sum(is.na(natural_Disaster_filtered$X2000)) /
nrow(natural_Disaster_filtered) * 100
```



```
[1] 36.84211
> sum(is.na(natural_Disaster_filtered_final$X2000)) /
nrow(natural_Disaster_filtered_final) * 100
[1] 8.695652
```

R Console Output - Missingness of Natural Disaster Data Before and After Country Filtering

Filtering also reduced missingness in the natural disaster data from 36.84% to 8.70%, indicating that missingness in this feature is not independent of missingness in other features.

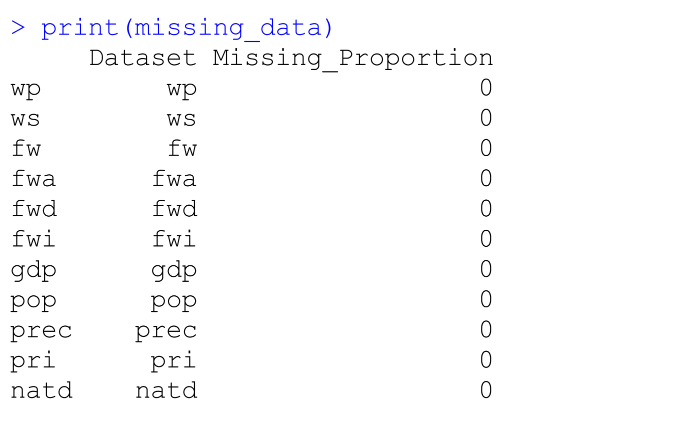


```
# Regression Imputation for natural disasters
natural_disasters <- natural_disasters %>%
  filter(Country.Code %in% water_stress$Country.Code)

nd <- natural_disasters %>%
  left_join(info, by = "Country.Code") %>%
  mutate(Region = as.factor(Region))

nd_model <- lm(X2000 ~ Region, data = nd, na.action = na.exclude)

nd$X2000_imputed <- ifelse(
  is.na(nd$X2000),
  predict(nd_model, newdata = nd),
  nd$X2000
)
```

Code Snippet - Regression Imputation for Natural Disaster Data

Finally, remaining gaps in the natural disaster data were filled by regressing the midpoint year value (X2000) on each country's region. By definition, the natural disaster data represents the percentage of the population affected by droughts, floods, and extreme temperatures. Therefore, an assumption has been made that countries within the same region experience broadly similar exposure to such extreme weather events, and conditional means can serve as plausible proxies for missing national values.

Given the relatively low post-filter missing rate of 8.7%, leveraging regional conditional means was not expected to introduce substantial error. However, more rigorous validation—such as comparing the imputed values against independent disaster-impact datasets or conducting cross validation within the panel—would strengthen confidence in the assumption of regional homogeneity.

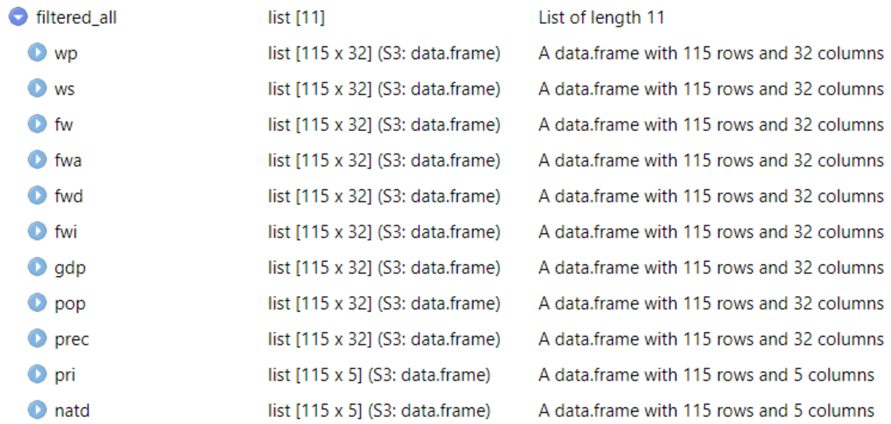


```
> print(missing_data)
     Dataset Missing_Proportion
wp        wp                  0
ws        ws                  0
fw        fw                  0
fwa      fwa                  0
fwd      fwd                  0
fwi      fwi                  0
gdp      gdp                  0
pop      pop                  0
prec    prec                  0
pri      pri                  0
natd    natd                  0
```

R output console - Final Missing Data Proportion

As a result of the applied filtering, labeling, and imputation steps, all datasets no longer contained missing values.

2.3 Data Overview After Handling Missing Data



```
filtered_all    list [11]                              List of length 11
  wp            list [115 x 32] (S3: data.frame)       A data.frame with 115 rows and 32 columns
  ws            list [115 x 32] (S3: data.frame)       A data.frame with 115 rows and 32 columns
  fw            list [115 x 32] (S3: data.frame)       A data.frame with 115 rows and 32 columns
  fwa           list [115 x 32] (S3: data.frame)       A data.frame with 115 rows and 32 columns
  fwd           list [115 x 32] (S3: data.frame)       A data.frame with 115 rows and 32 columns
  fwi           list [115 x 32] (S3: data.frame)       A data.frame with 115 rows and 32 columns
  gdp           list [115 x 32] (S3: data.frame)       A data.frame with 115 rows and 32 columns
  pop           list [115 x 32] (S3: data.frame)       A data.frame with 115 rows and 32 columns
  prec          list [115 x 32] (S3: data.frame)       A data.frame with 115 rows and 32 columns
  pri           list [115 x 5] (S3: data.frame)        A data.frame with 115 rows and 5 columns
  natd          list [115 x 5] (S3: data.frame)        A data.frame with 115 rows and 5 columns
```

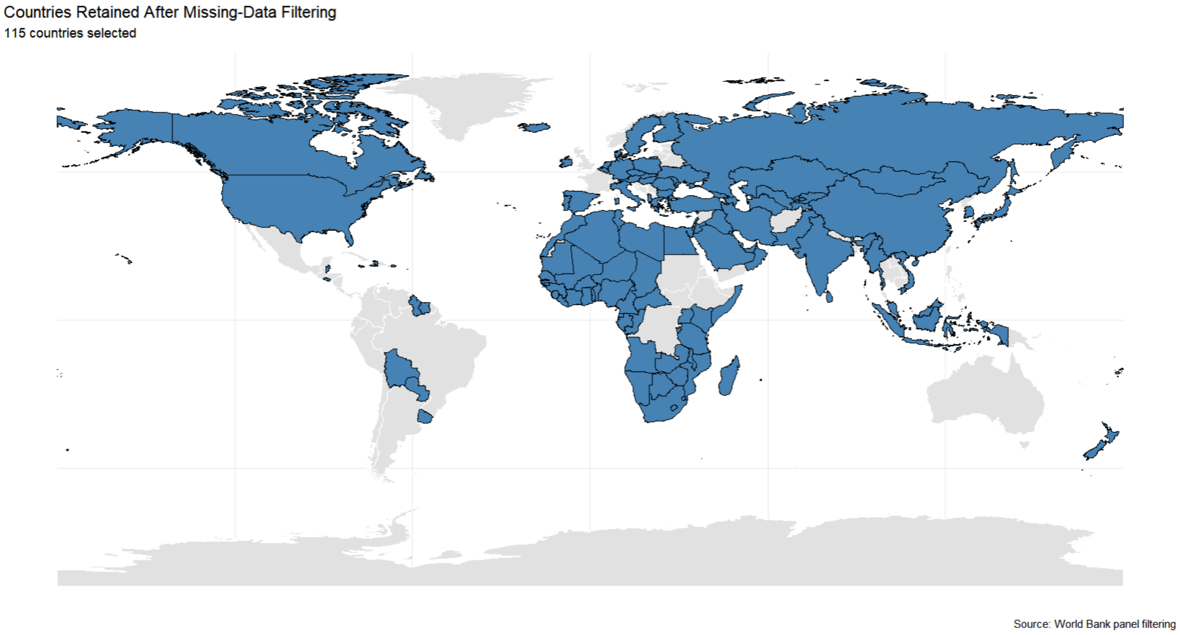
Figure - Dataset Overview After handling Missing Data

All nine “normal” features now have identical dimensions—115 countries by 28 years (plus 4 meta columns)—facilitating manipulation and analysis across multiple indicators. The natural disaster data is reduced to a single column (X2000), representing the 1990–2009 average, and the private investment data likewise consists of a single binary label column, reflecting its recoding from a sparse time series into a presence/absence indicator.

```
> valid_years
 [1] "X1994" "X1995" "X1996" "X1997" "X1998" "X1999" "X2000" "X2001"
"X2002" "X2003" "X2004" "X2005" "X2006"
[14] "X2007" "X2008" "X2009" "X2010" "X2011" "X2012" "X2013" "X2014"
"X2015" "X2016" "X2017" "X2018" "X2019"
[27] "X2020" "X2021"
```

R Console Output - List of retained Years after handling Missing Data

This is the list of retained years across all features after handling missing data. These years form a 28-year span from 1994 through 2021 with no discontinuity, which allows focused analysis on recent decades without any temporal gaps.



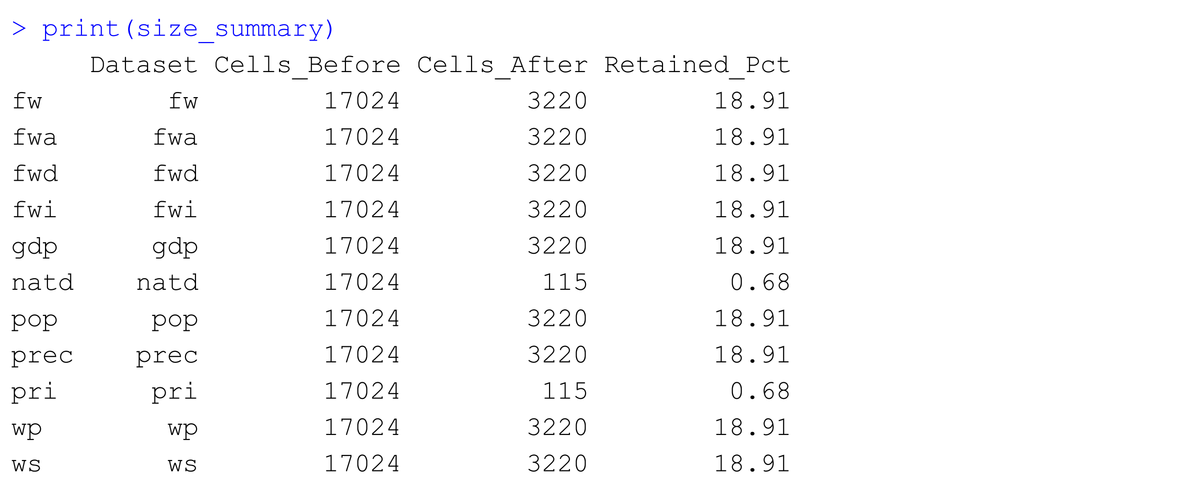
Countries Retained After Missing-Data Filtering

115 countries selected

Source: World Bank panel filtering

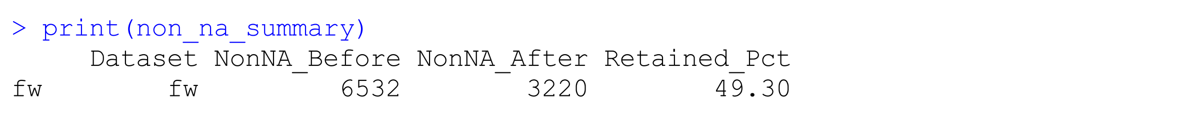
Figure - Global Coverage After handling Missing Data

This world map shows the countries retained across all indicators after handling missing data. While some exclusions occur in Asia, Europe, Africa, and Oceania, the greatest concentration of drop-outs is in Central and Latin America. This pattern may reflect limited statistical capacity, resource constraints, or challenges in data collection and dissemination of that region. Therefore, analyses focusing on Central or Latin America should be interpreted with caution or supplemented by additional sources, as they rely on a very narrow subset of countries and may skew conclusions.



```
> print(size_summary)
     Dataset Cells_Before Cells_After Retained_Pct
fw        fw        17024        3220        18.91
fwa      fwa        17024        3220        18.91
fwd      fwd        17024        3220        18.91
fwi      fwi        17024        3220        18.91
gdp      gdp        17024        3220        18.91
natd    natd        17024         115         0.68
pop      pop        17024        3220        18.91
prec    prec        17024        3220        18.91
pri      pri        17024         115         0.68
wp        wp        17024        3220        18.91
ws        ws        17024        3220        18.91
```

R Console Output - Cell-Count Retention After Filtering



```
> print(non_na_summary)
     Dataset NonNA_Before NonNA_After Retained_Pct
fw        fw         6532        3220        49.30
```

```
fwa      fwa        6121        3220        52.61
fwd      fwd        6106        3220        52.74
fwi      fwi        5985        3220        53.80
gdp      gdp        8266        3220        38.95
natd    natd         168         115        68.45
pop      pop       16930        3220        19.02
prec    prec       10268        3220        31.36
pri      pri         405         115        28.40
wp        wp        8263        3220        38.97
ws        ws        6024        3220        53.45
```

R Console Output - Non-NA Cell Retention After Filtering

These two summaries illustrate different aspects of information loss from the filtering procedure. The first R console output shows the loss in total cell slots by the formula

$$retention\,percentage \;=\; \frac{countries \times years\ (\,after\ filtering)}{countries \times years\ (\,before\ filtering)}$$

, making natural disaster and private investment appear heavily diminished because they were collapsed to a single column. However, most of these removed slots were empty from the beginning, and therefore a second measure of information loss was deemed necessary. The second R console output quantifies loss of actual observations by the formula

$$retention\,percentage \;=\; \frac{non{-}NA\ entries\ after\ filtering}{non{-}NA\ entries\ before\ filtering}$$

. It shows that “normal” features typically retain 30–55% of their initial non-NA values, while natural disaster and private investment data preserve 68.5% and 28.4%, respectively. The population data is an exception: since its original missing rate was exceptionally low (under 1%), there is little difference between structural loss (cell-count reduction) and observational loss (non-NA reduction) for that feature.

Balancing the trade-off between dataset completeness and information loss guided the threshold choices. Looser thresholds allow more cells to remain but at the cost of higher overall missingness; stricter thresholds reduce missingness but eliminate more data. Intermediate cutoffs were chosen to maintain missing-data rates within reasonable bounds while preserving as many cells as possible. However, these thresholds were based on subjective judgment and can be revisited if further analyses call for a different balance. For example, if a classification task exhibits high variance, suggesting the dataset is too small, the filters can be loosened to include more cells even if that increases the missing-data rate.

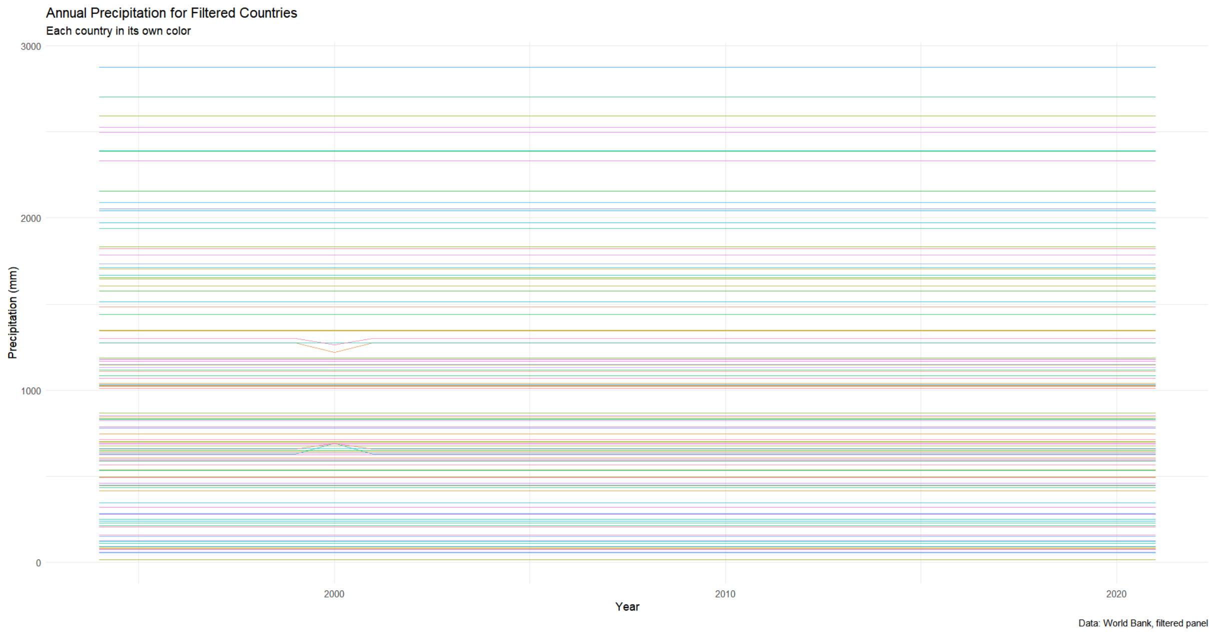
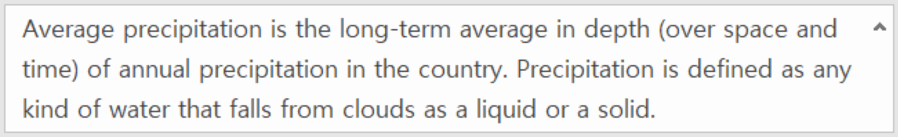


Figure - Annual Precipitation Time Series after Filtering

This is the time-series plot of the filtered panel of annual precipitation, with each country drawn in its own color. A striking feature is that the lines are almost flat, with little year-to-year variation.



Average precipitation is the long-term average in depth (over space and time) of annual precipitation in the country. Precipitation is defined as any kind of water that falls from clouds as a liquid or a solid.

Figure - Detailed Description of Precipitation stated in the Metadata

The reason for the near-flat lines became clear once the metadata was consulted. By definition, “average precipitation” is not the annual total for each specific year but the long-term spatial and temporal average depth of precipitation in the country. In other words, each “year” in the panel simply repeats the same summary value rather than reflecting true year-to-year variability. This characteristic may limit the usefulness of the precipitation data for time-series analysis or for any other approach that requires year-to-year variability.

2.4. Additional Feature Derivation

By definition, several original features are functionally dependent. To recover possible latent information from these relationships, three additional features have been derived:

- Available resources:
  Derived from
  $$water\ stress = \frac{total\ withdrawals}{available\ resources} \Rightarrow available\ resources = \frac{total\ withdrawals}{water\ stress}$$

- GDP (USD):
  Derived from

$$water\ productivity\ =\ \frac{gdp\ (usd)}{total\ withdrawals}\ \Rightarrow\ gdp\ (usd)\ =\ water\ productivity\ \times\ total\ withdrawals$$

- GDP (PPP):
  Derived from
  $gdp\ per\ capita(PPP)\ \times\ population\ =\ gdp\ (ppp)$

Introducing these level-based features would uncover variation in absolute resource endowment and economic scale that percentage- or ratio-only metrics does not capture. However, for any given analysis, either the original ratios or the new levels—but not both—would have to be excluded to avoid multicollinearity while retaining the most relevant information for the task at hand.

3. Summary

The original World Bank dataset encompassed 266 countries over the period 1960–2023 but suffered from an extensive amount of missing values. By applying a filtering step, all years and countries with high missingness were removed to yield a continuous panel of 115 countries from 1994 to 2021, ensuring every core indicator could be directly compared across both space and time. Regression based imputation was then used to fill the remaining holes in the natural disaster series.

To handle the nearly empty private investment series, countries were recoded into a simple presence/absence label, enabling the creation of a complete panel from an otherwise sparse time-series dataset. Finally, three derived features—available water resources and GDP in both USD and PPP terms—were calculated to expose scale effects that ratios alone would not reveal. Normalization was deliberately postponed in order to allow each subsequent task to apply the most appropriate scaling.

Together, these steps produce a clean, analysis-ready dataset that supports exploratory analysis, statistical inference, and machine-learning applications. Even though these preprocessing steps are generally time-consuming and technically demanding, they are crucial for ensuring the accuracy and reliability of all subsequent analyses. If any errors or logical flaws are detected at this stage, the entire subsequent workflow would need to be revisited and reevaluated, which is an even more time-consuming endeavor.