Department of Scientific Computing, Modeling & Simulation

Savitribai Phule Pune University

Mapping Global Biodiversity

Patterns, Trends, and Insights from GBIF Data

Final Story Report

Submitted in partial fulfillment of the requirements for the

Master of Technology (M.Tech)

Submitted By:

Ayush Gaikwad (MT2408)
M.Tech – Scientific Computing, Modeling & Simulation

Under the Guidance of:

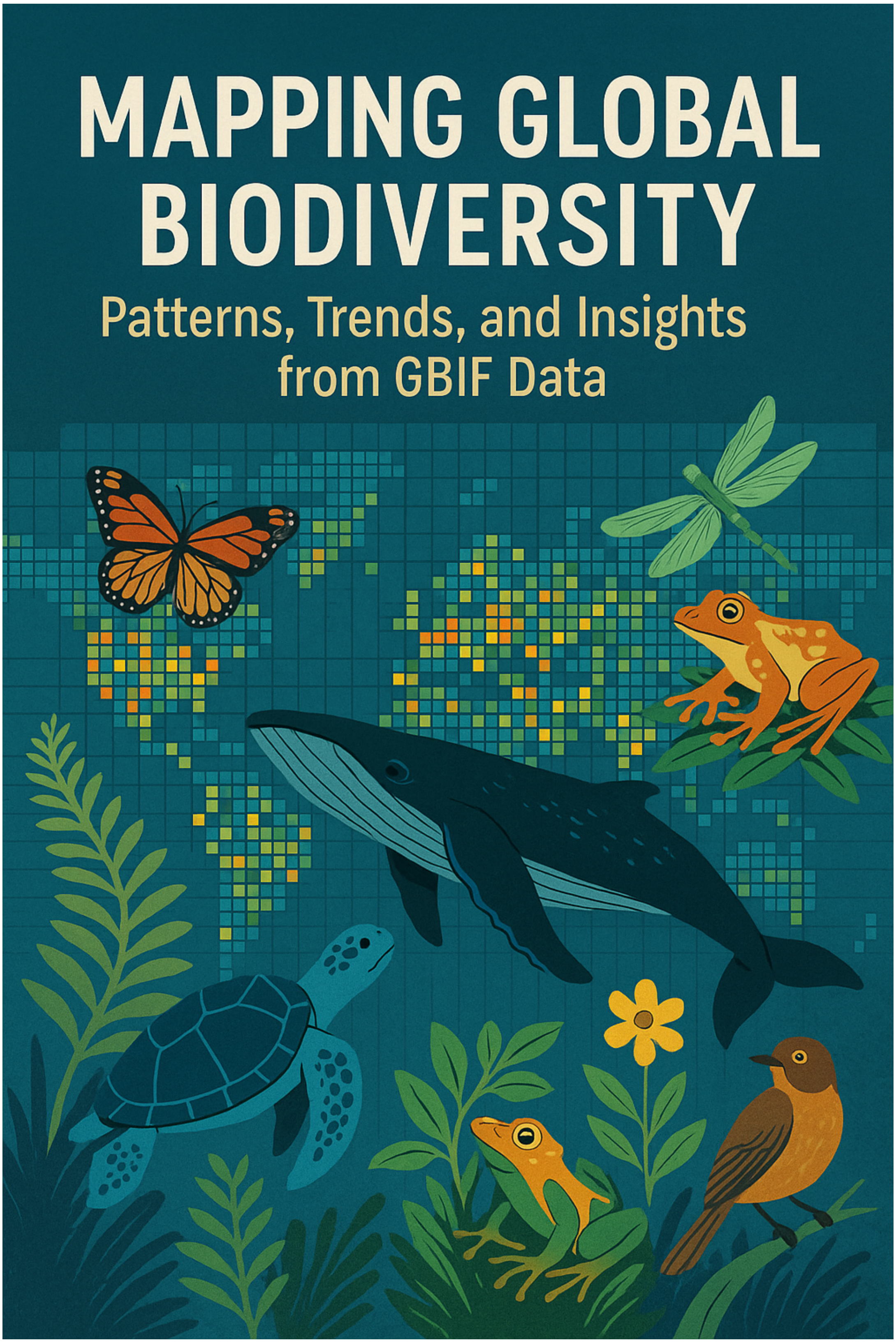
Dr. Bhalchandra Pujari
Professor, Department of Scientific Computing, Modeling & Simulation

Savitribai Phule Pune University

Abstract

This report presents a complete end-to-end analysis of biodiversity occurrence records sourced from the Global Biodiversity Information Facility (GBIF). The study spans the entire workflow from raw data cleaning, exploratory data analysis (EDA), spatial and temporal trend extraction, taxonomic and diversity assessment, and the development of an interactive dashboard.

The purpose of this project is to highlight biodiversity patterns across taxa, time, and geography while also examining human contributors responsible for species identification and recording. By integrating scientific analysis with interactive visual tools, this work enables rich ecological insights and supports biodiversity research and decision-making.



MAPPING GLOBAL BIODIVERSITY

Patterns, Trends, and Insights from GBIF Data

Contents

Chapter 1

Introduction

Biodiversity forms the structural and functional fabric of natural ecosystems. Understanding how species are distributed across geographical regions, how their presence changes over time, and how sampling activity varies across countries is fundamental to conservation and ecological science.

The Global Biodiversity Information Facility (GBIF) aggregates biodiversity data contributed by researchers, institutions, and citizen scientists. This project leverages a curated subset of GBIF data and transforms it through the following stages:

- Data cleaning and preprocessing,
- Exploratory data analysis,
- Spatial and temporal pattern extraction,
- Taxonomic structure visualisation,
- Diversity estimation,
- Contributor (metadata) analysis,
- Dashboard development for interactive exploration.

Chapter 2

Dataset Description

The dataset contains species occurrence records with taxonomic, spatial, temporal, and contributor information:

- **Taxonomy:** kingdom, phylum, class, order, family, genus, species
- **Time:** eventDate, year, month
- **Geography:** decimalLatitude, decimalLongitude
- **Location metadata:** Country, stateProvince
- **Contributor metadata:** identifiedBy, recordedBy, rightsHolder

Cleaning resulted in:

- Standardized taxonomic fields,
- Validated coordinates,
- Converted country codes into full names,
- Removal of empty or low-utility columns such as `infraspecificEpithet`.

Chapter 3

Data Cleaning and Pre-processing

3.1 Removing Empty Columns

Columns with 100% missing values were dropped. Highly sparse fields (¿90% missing) that offered no analytical value were removed.

3.2 Date Parsing

`eventDate` was converted into ISO datetime format. Numeric fields such as `year`, `month`, and `day` were type-corrected.

3.3 Coordinate Validation

Invalid coordinate values outside:

$$-90 \leq \text{decimalLatitude} \leq 90, \quad -180 \leq \text{decimalLongitude} \leq 180$$

were removed.

3.4 Country Mapping

ISO country codes were mapped to full names using dictionary mapping.

3.5 Text Normalisation

Whitespace trimming and normalization were applied across key text fields.

The cleaned dataset was stored as:

`APP_final_dashboard_dataset_8_clean.csv`

Chapter 4

Exploratory Data Analysis

4.1 Missing Value Analysis

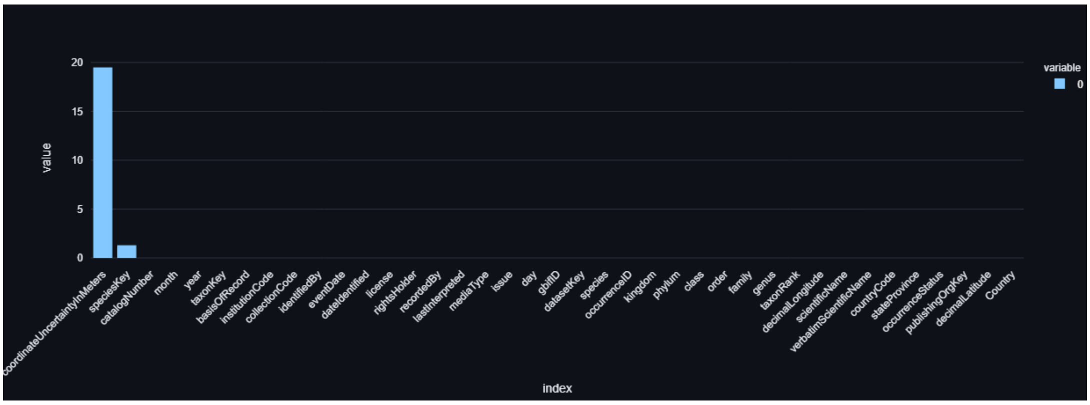


Figure 4.1: Missing value distribution across dataset attributes.

4.2 Correlation Structure

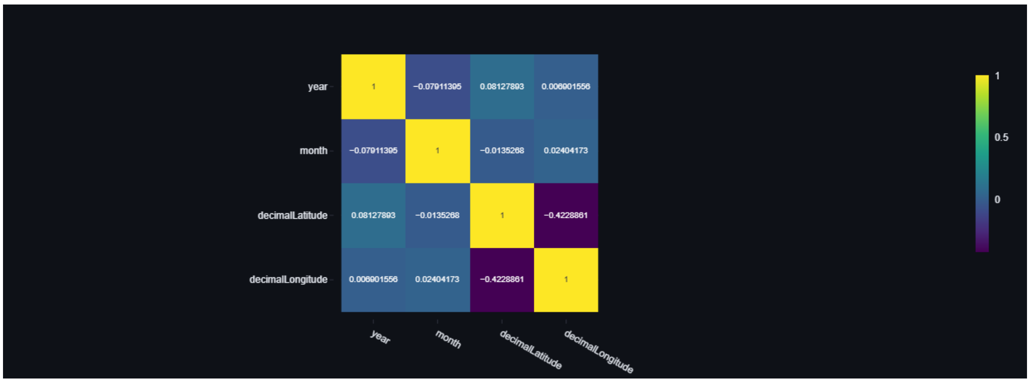


Figure 4.2: Correlation heatmap of selected numerical features.

4.3 Top Phyla

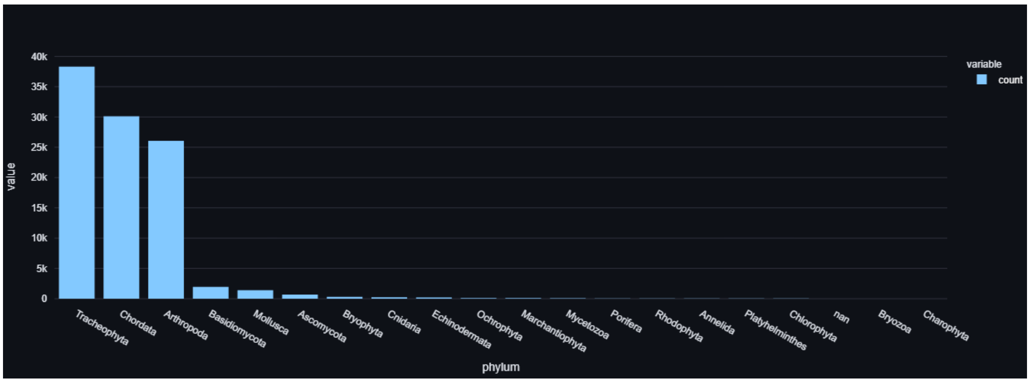


Figure 4.3: Most common phyla in the dataset.

4.4 Top Species

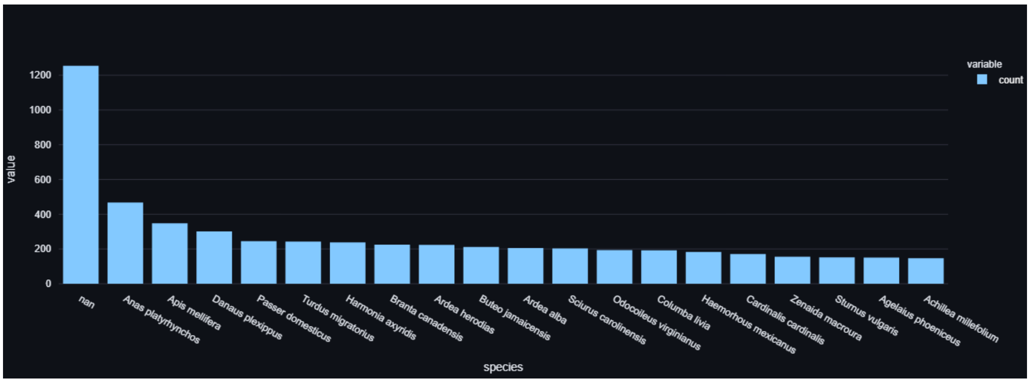


Figure 4.4: Most frequently observed species.

4.5 Taxonomic Hierarchy

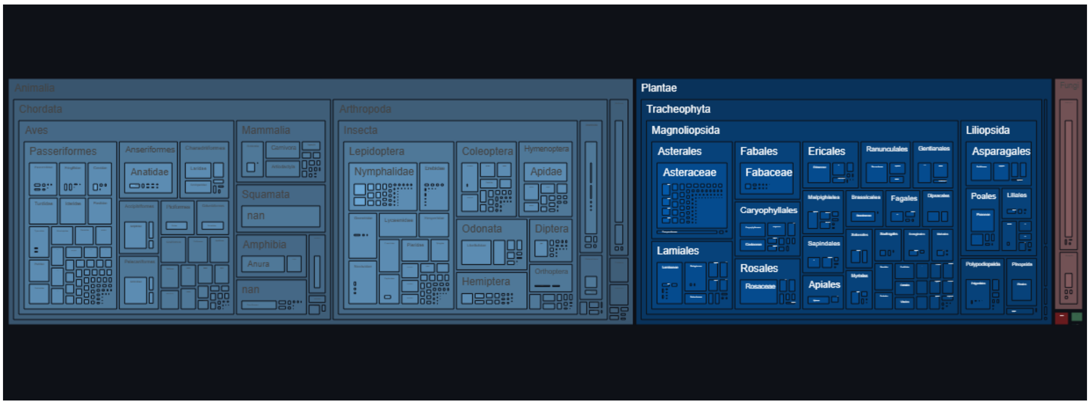


Figure 4.5: Taxonomic tree hierarchy from kingdom to species.

Chapter 5

Temporal Analysis

5.1 Yearly Trends

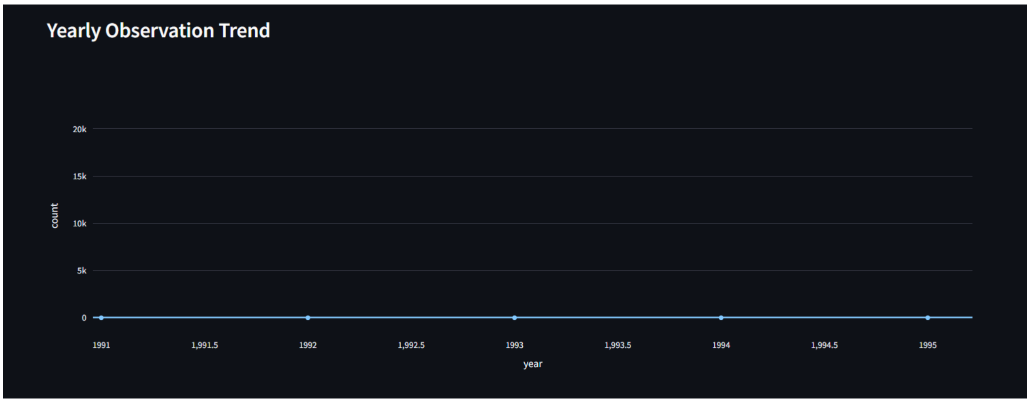


Figure 5.1: Observation trends over the years.

5.2 Monthly Trends

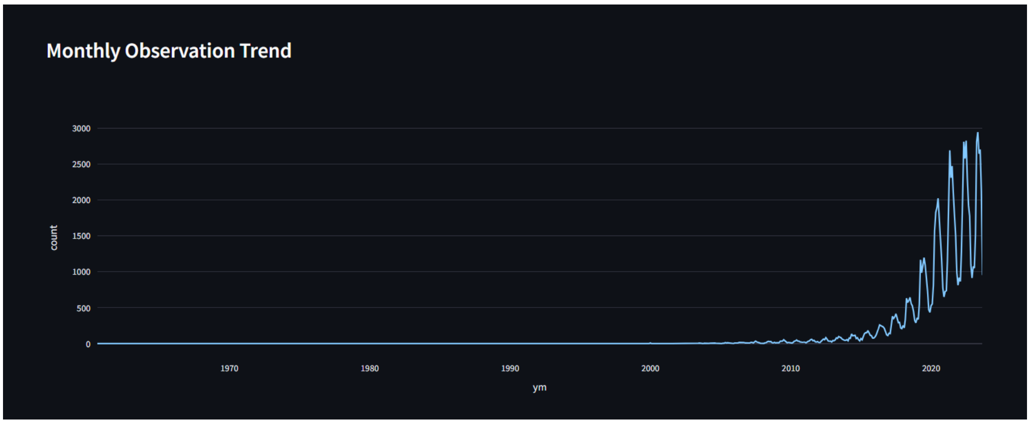


Figure 5.2: Seasonal/Monthly observation variations.

5.3 Animated Time Map

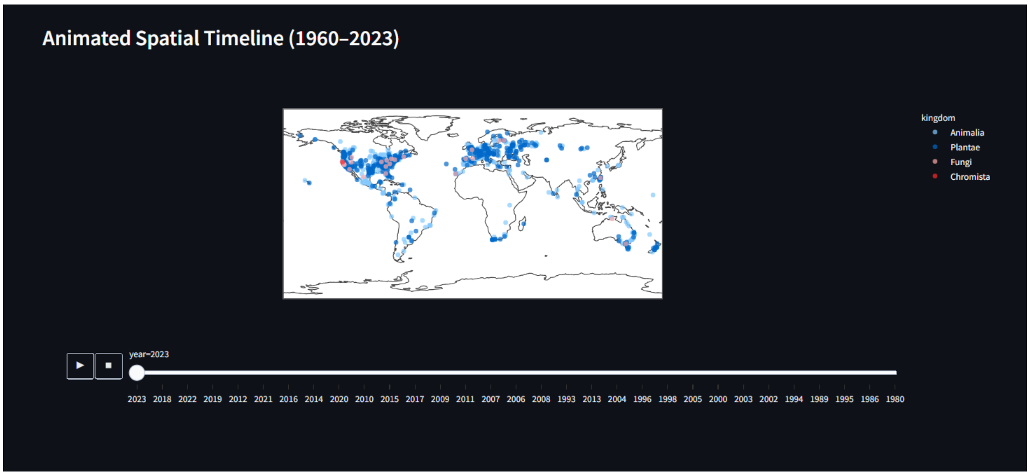


Figure 5.3: Global spatio-temporal animation (1960–2023).

Chapter 6

Spatial Analysis

6.1 Density Map

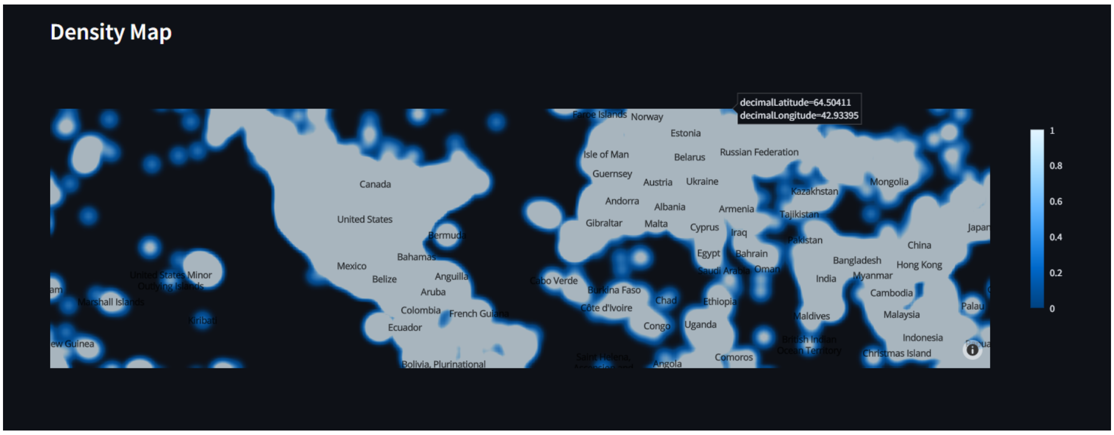


Figure 6.1: Smoothed global biodiversity density map with country overlays.

6.2 Scatter Map

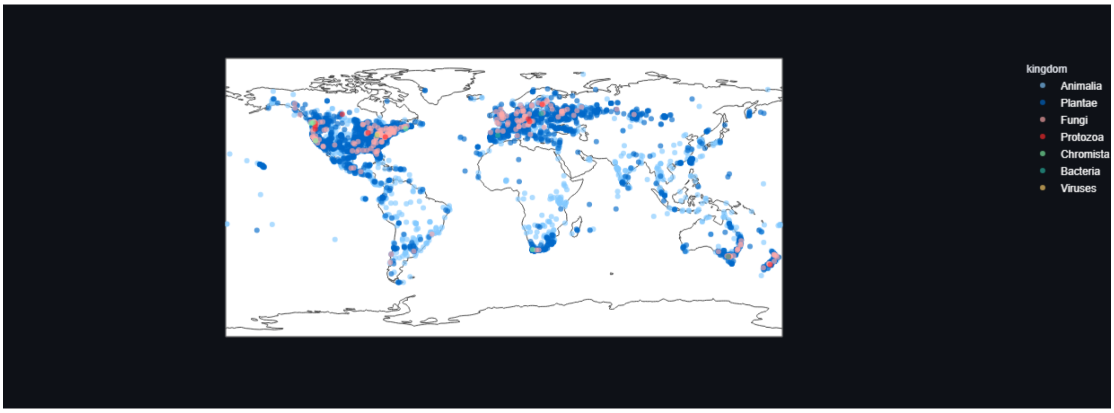


Figure 6.2: Geospatial scatter plot of species occurrences.

Chapter 7

Diversity and Ecological Insights

7.1 Shannon Diversity Index

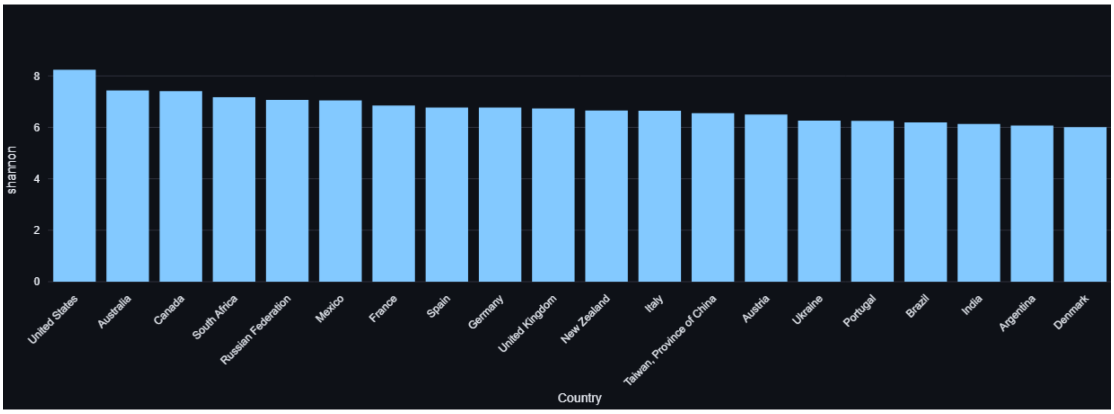


Figure 7.1: Shannon diversity across different geographic categories.

Chapter 8

People Behind the Observations

8.1 Top Identifiers

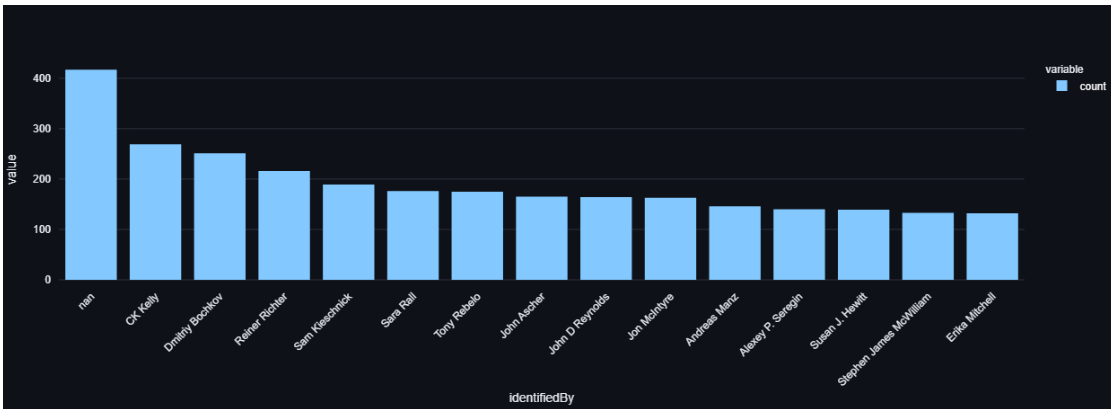


Figure 8.1: Most active species identifiers in the dataset.

8.2 Top Rights Holders

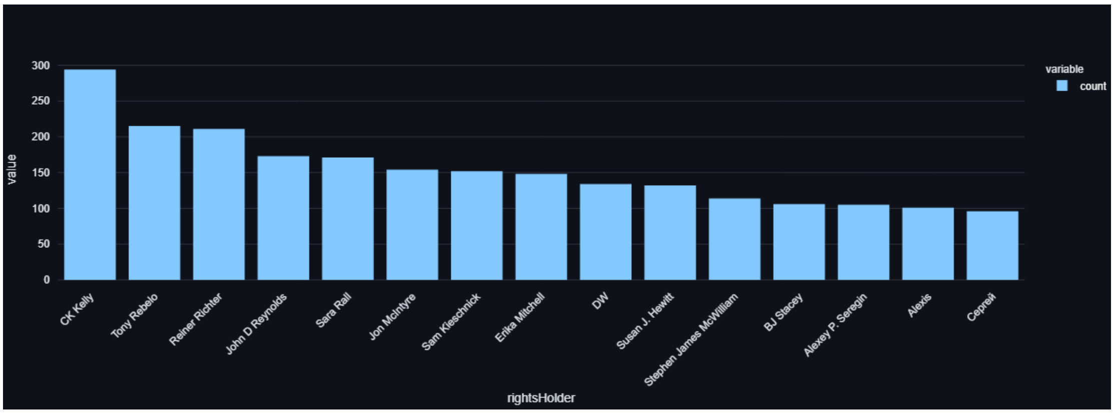


Figure 8.2: Top rights holders associated with the dataset entries.

8.3 Top Recorders

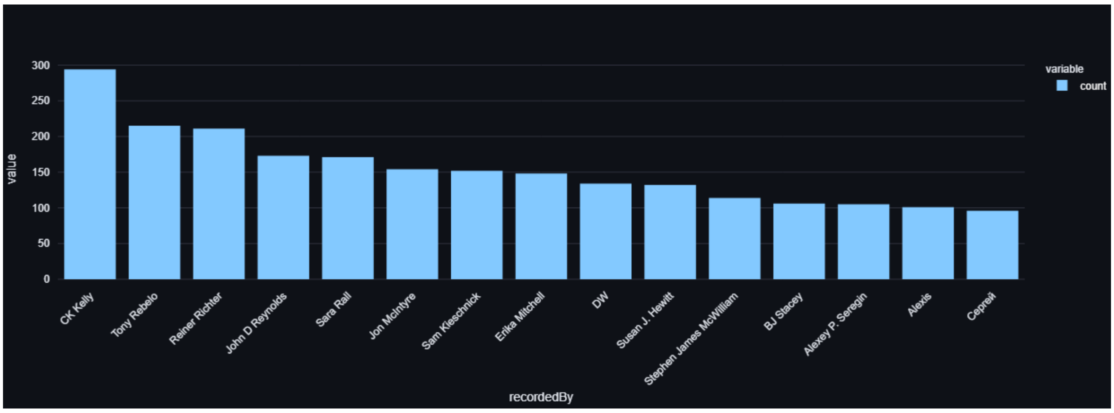


Figure 8.3: Most active field recorders contributing observations.

Chapter 9

Discussion and Conclusion

The combined analyses reveal:

- Regions with higher sampling effort show higher observed diversity.
- Temporal analysis indicates significant biodiversity documentation after 2000.
- Spatial density maps highlight global biodiversity hotspots.
- Taxonomic distribution shows uneven sampling across taxa.
- Contributor metadata indicates strong participation from a small group of experts and institutions.

This project establishes a complete analytical pipeline from raw biodiversity data to interactive visual analytics. It showcases how biodiversity data can be structured, analyzed, visualized, and interpreted to support ecological research and conservation planning.

Chapter 10

Future Work

Future enhancements may include:

- Linking climate data with species occurrence,
- Machine learning models for species distribution prediction,
- Advanced anomaly detection for taxonomic data,
- Automated dashboards with real-time GBIF API integration.

References

- GBIF Global Biodiversity Facility: https://www.gbif.org
- Plotly Python Graphing Library: https://plotly.com/python
- Streamlit Documentation: https://docs.streamlit.io