

Exploratory Data Analysis Story – Global Biodiversity Dataset

1. Introduction

Biodiversity plays a crucial role in maintaining ecological balance, and understanding species distribution helps in conservation planning. This project explores a global biodiversity dataset containing species observation records with taxonomic, geographic, and temporal attributes.

The goal of this analysis is to uncover meaningful ecological patterns, evaluate diversity measures, and compare biodiversity across regions, with a particular focus on India, a recognized megadiverse nation.

2. Data Understanding and Preprocessing

The dataset initially contained 50 columns, several of which had extremely high missing values. Columns with **100% missing values** (e.g., elevation, depth, coordinatePrecision, etc.) were removed as they did not contribute any usable information.

Rows missing **class**, **latitude**, and **longitude** values were removed to ensure reliability of spatial and taxonomic analysis. Taxonomy-related columns such as *order*, *family*, *genus*, *species* were imputed with "Unknown" instead of dropping, preserving valuable contextual information.

3. Target Variable: Class

The analysis focused on **class** as the target variable, examining species distribution across biological classes. Dominant classes included **Insecta**, **Aves**, and **Magnoliopsida**, indicating strong representation of insects, birds, and flowering plants.

At the kingdom level, most observations were from **Animalia**, indicating an animal-dominated dataset.

4. Spatial and Temporal Analysis

Spatial Spread

Visual analysis using scatter plots and KDE heatmaps revealed major biodiversity hotspots, especially within India, centered around:

- **Western Ghats**
- **Northeast India**
- **Parts of Central India**

These regions are globally recognized biodiversity hubs, demonstrating alignment between ecological reality and dataset observations.

Temporal Trends

Observation frequency increased dramatically after the year **2000**, correlating with:

- Growth of citizen science platforms (eBird, iNaturalist)
- Digitization of biological records
- Wider accessibility to mobile photography and GIS technology

Seasonal analysis showed **significant peaks in June–September**, corresponding to India’s monsoon season, a period characterized by maximum ecological activity and species visibility.

5. Taxonomic Hierarchy Visualization

A **Sunburst chart** demonstrated hierarchical distribution from **Kingdom → Phylum → Class → Order**, highlighting:

- Dominance of **Arthropoda → Insecta → Lepidoptera / Coleoptera**
- **Chordata → Aves → Passeriformes**
- **Plantae → Magnoliopsida** among plants

This visualization clearly illustrates structural biodiversity patterns and taxonomic richness.

6. Biodiversity Richness Metrics

To quantify biodiversity scientifically, **Shannon and Simpson diversity indices** were computed.

Region	Shannon	Simpson
India	6.15	0.997
USA	Higher than India (Also very high)	

Interpretation:

- India shows **extremely high biodiversity richness and evenness**

- However, USA shows **higher diversity values in this dataset**

Seasonal richness (India only)

Shannon values by month ranged from **3.9 to 4.35**, peaking during monsoon, confirming seasonal biodiversity patterns.

State-wise richness

Highest values corresponded to **Kerala, Karnataka, Assam, Meghalaya**, supporting ecological hotspot theory.

7. India vs USA Diversity Comparison

Although computed diversity indices show **USA having higher observed richness** than India in this dataset, this does **not reflect true biological biodiversity**.

Why?

GBIF is a database of **recorded** data, not a full species census. Diversity values depend strongly on **sampling effort**.

Factor	USA	India
Total observations contributed	~45,000	~800
Citizen science platforms	very active	developing
Digitization history	long-established	less digitized
Research funding & accessibility	high	limited

The USA's higher Shannon index reflects a **greater number of recorded observations**, not higher natural biodiversity.

India is globally recognized as a **megadiverse country**, but is **underrepresented** in GBIF due to sampling bias, lower citizen-science participation, and data digitization gaps.

Scientific viewpoint

The difference shows **dataset bias**, not ecological truth. Therefore, biodiversity conclusions must consider **effort bias**.

8. Key Insights Summary

Aspect	Insight
Taxonomic Patterns	Insecta & Aves dominate globally
Geographic Distribution	Western Ghats & NE India are hotspots
Temporal Trends	Monsoon peak biodiversity
Diversity Metrics	India extremely diverse but underrecorded
Country Comparison	USA appears richer due to sampling bias
Data Quality Insight	High levels of "Unknown" species indicate identification challenges

9. Conclusion

The exploratory data analysis highlights rich biodiversity patterns across geography, taxonomy, and time, emphasizing India's ecological significance despite low dataset representation. While Shannon and Simpson indices show higher recorded species richness in the United States, this discrepancy results from sampling inequality rather than true ecological diversity.

This reinforces the importance of contextual interpretation and awareness of dataset limitations when drawing ecological conclusions.

Future Work

- Integrate climate datasets to analyze environmental drivers
 - Build ML models to predict species class based on location & climate
 - Perform clustering for habitat grouping
 - Normalize richness per sampling effort (rarefaction curves)
-

Final Research Statement

Biodiversity datasets must be interpreted through both analytical and ecological lenses. Observational bias significantly influences perceived diversity, and true biological conclusions require integrating data availability, scientific sampling, and environmental context.