Meteorite Landing Data Workflow and Analysis

S. Palis

Udacity AI Programming Foundations Capstone

February 7, 2026

Overview

This project implements a reproducible data workflow for cleaning, exploring, and visualizing meteorite landing records. The NASA Meteorite Landings dataset was used as a proxy for noisy scientific observation logs, allowing structured data hygiene, anomaly handling, and exploratory analysis. The workflow demonstrates how raw tabular data can be transformed into a transparent analytical pipeline suitable for future machine learning tasks.

Dataset Description

The dataset contains over 45,000 meteorite records with attributes such as name, classification, mass, year of discovery, and geographic coordinates. Several fields contain missing or corrupted values, including incomplete spatial information and invalid timestamps. The analysis focused on mass, temporal trends, meteorite class distribution, and recovery bias between observed falls and discovered fragments.

Because the dataset mixes reliable records with corrupted entries, it is an appropriate case study for structured data cleaning and reproducible workflow design.

Workflow Description

The workflow followed a standard data science lifecycle: ingestion, cleaning, exploratory analysis, visualization, and interpretation.

The dataset was ingested using Pandas and preserved in its raw form before any transformation. Cleaning functions were implemented to remove incomplete geographic records and invalid year values. Exploratory analysis used reusable functions to summarize numeric variables and categorical distributions.

Visualizations revealed detection bias, class concentration, temporal trends, and skewed mass distribution. The notebook concludes with an interpretive summary linking patterns to real-world constraints in scientific data collection.

Key Decisions and Assumptions

Two primary cleaning decisions shaped the workflow. Records with missing coordinates were removed because spatial analysis requires valid geographic information. Records with year values outside the documented valid range were excluded because the dataset documentation identifies these as parsing errors.

These choices prioritize reliability over completeness. While some historical information may be lost, the resulting dataset supports consistent analysis.

The exploratory focus on mass distribution, meteorite classification, and recovery type was chosen to highlight physical processes and human detection bias embedded in the dataset.

Transparent cleaning rules and function-based transformations support reproducible research practices, which are central to modern data science education (Danchev, 2022).

Results and Interpretation

The cleaned dataset reveals several important structural patterns. Most meteorites were discovered after landing rather than observed falling, indicating strong detection bias (see Figure 1). Recovery depends heavily on human presence and search effort, meaning the dataset reflects observational behavior as much as physical events.

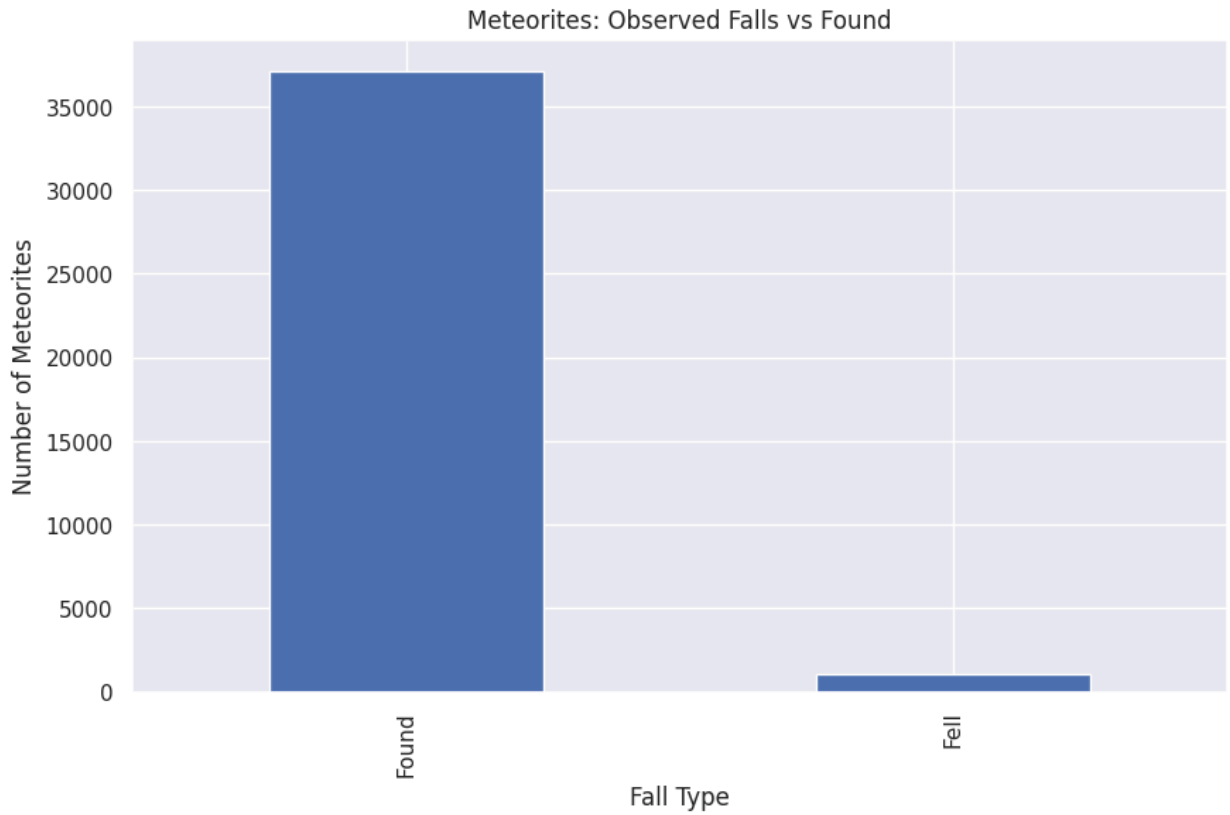

Figure 1

Distribution of observed meteorite falls versus discovered meteorites.

Meteorite mass follows a pronounced long-tail distribution, with many small fragments and a few extreme events (see Figure 2). This skew is consistent with fragmentation physics and

demonstrates why logarithmic scaling is necessary for meaningful interpretation of scientific magnitude data.

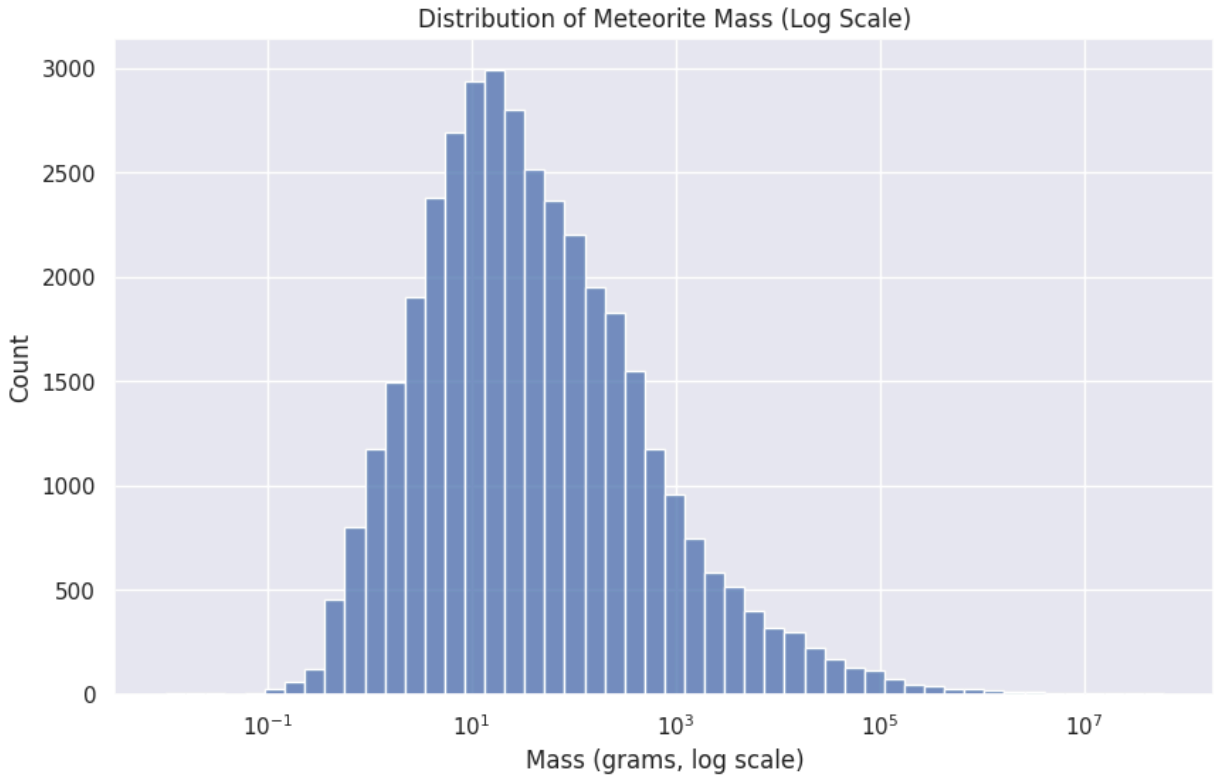

Figure 2

Log-scaled distribution of meteorite mass showing long-tail skew.

Discovery counts increase dramatically in recent decades (see Figure 3). This trend likely reflects improved reporting infrastructure, scientific coordination, and global data collection rather than an increase in impact frequency. Together, these patterns illustrate how scientific datasets encode both natural phenomena and human observation constraints.

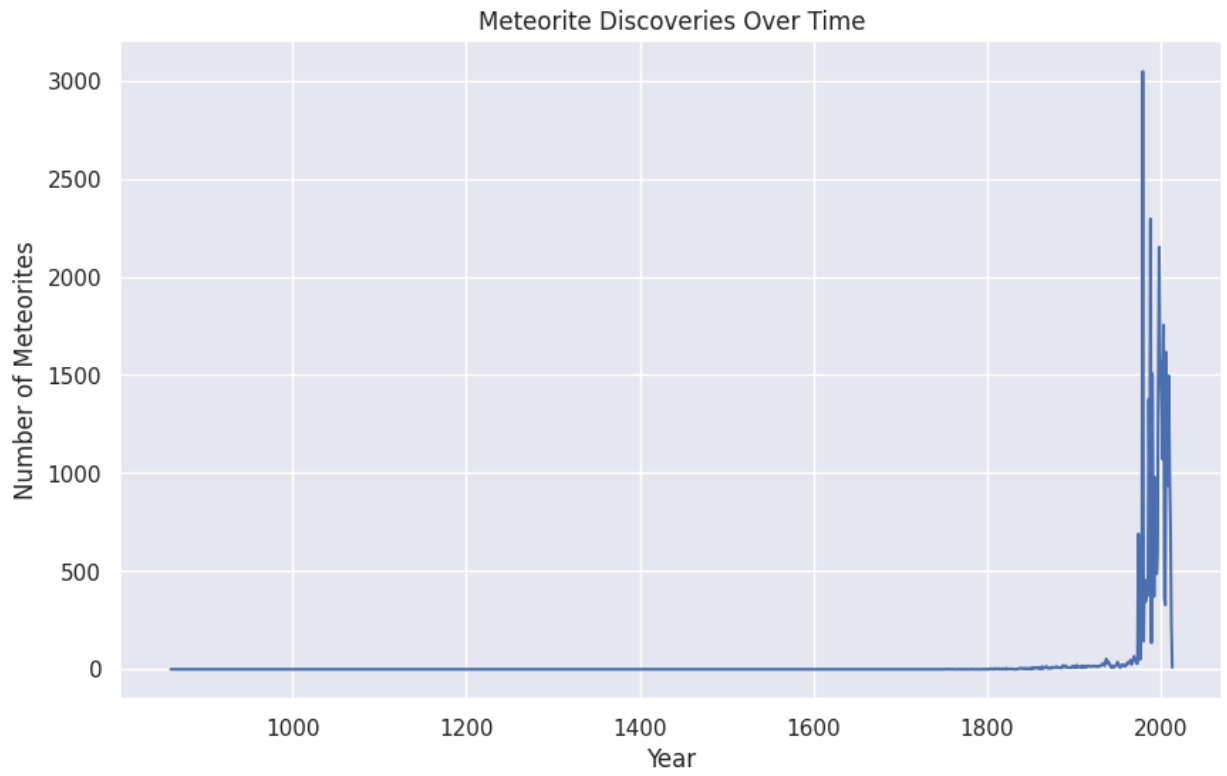

Figure 3

Meteorite discovery counts over time reflecting reporting growth.

Responsible Practice (Bias)

Cleaning decisions can introduce bias. Removing incomplete geographic records may disproportionately exclude regions where precise location data was never recorded. Similarly, filtering invalid years assumes corruption rather than rare historical events. These steps improve analytical consistency but reduce representation of uncertain records.

Recognizing how preprocessing alters dataset composition is essential for ethical and responsible data science practice (Rule et al., 2019). Future work could explore imputation strategies or metadata recovery to mitigate cleaning bias.

Reproducibility

The notebook is structured to run top-to-bottom without hidden state. All transformations are implemented as documented functions, and a requirements.txt file captures the software environment. Git version control tracks workflow evolution through multiple commits and branching.

This design follows reproducible data science principles by integrating code, documentation, and narrative into a single executable artifact (Danchev, 2022). Reproducibility enables other researchers to inspect, rerun, and extend the analysis.

References

Danchev, V. (2022). Reproducible data science with Python: An open learning resource. *Journal of Open Source Education, 5*(56), 156. https://doi.org/10.21105/jose.00156

Rule, A., Birmingham, A., Zuniga, C., Altintas, I., Huang, S.-C., Knight, R., Moshiri, N., Nguyen, M. H., Rosenthal, S. B., Pérez, F., & Rose, P. W. (2019). Ten simple rules for writing and sharing computational analyses in Jupyter notebooks. *PLOS Computational Biology, 15*(7), e1007007. https://doi.org/10.1371/journal.pcbi.1007007

NASA. (2015). *Meteorite landings dataset*. Kaggle. https://www.kaggle.com/datasets/nasa/meteorite-landings