INFO 511 Foundations of Data Science – Fall 2025
Milestone 3 – project proposal
Nathan Herling
Thursday-November-13-2025
Title: A Temporal Analysis of Meteorite Findings

1. Introduction:

NASA's OSIRIS-REx mission to the asteroid Bennu—including the detection of organic, potentially prebiotic molecules in the returned samples—has generated renewed public and scientific excitement about small bodies, human space exploration, and our place in the universe [1]. Motivated by this context, this project examines whether a linear association exists between calendar year and the annual number of "Found" meteorites recorded on Earth. Put differently: **Do yearly counts of recovered meteorites increase, decrease, or remain stable over time?** To investigate this question, the analysis uses the publicly available *Meteorite Landings* dataset hosted through NASA/Meteoritical Society sources on Data.gov [2]. This project uses exploratory data analysis and statistical modeling to evaluate whether long-term meteorite recovery patterns reveal a meaningful temporal trend.

2. Process and Analysis:

2.1 - Data Source and Access

The dataset used in this project is the *Meteorite Landings* dataset hosted on Data.gov and maintained by the U.S. General Services Administration in collaboration with NASA and the Meteoritical Society [2]. This publicly available CSV includes global records of meteorite observations, classifications, discovery types, and coordinates spanning several centuries. Because it contains no personal identifiers and only scientific observations, no IRB or privacy review was required. Its scope and standardized structure make it well suited for analyzing long-term meteorite discovery trends.

2.2 - Data Preprocessing, Cleaning, and EDA Procedures

After downloading the raw *Meteorite Landings* CSV from Data.gov, preprocessing focused on retaining information relevant to the research question: the temporal pattern of "Found" meteorites. The dataset includes more than 45,000 entries with varying completeness across fields such as mass, coordinates, and discovery type. Table 1 in *Appendix A – EDA visuals* contains the initial EDA metrics. Because this project examines discovery counts over calendar years, the primary variables used were **year** and **fall** (categorizing entries as "Fell" or "Found"). For the cleaning process a valid-year filter was applied to retain only years between the earliest recorded observation and 2013, the latest complete year listed by the source. Records with invalid or missing years were imputed via removal, duplicate records were removed, as were entries not labeled "Found." No imputation was performed because year is a structural variable.

Following filtering, the data were aggregated to a year-level dataset, with each row representing a calendar year and the number of "Found" meteorites reported in that year. Exploratory data analysis (EDA) assessed the distribution of annual counts using histograms and boxplots, while scatterplots of count versus year provided an initial view of potential linear patterns. These summaries supported proceeding with linear regression.

2.3 - Assessment of Data Quality and Readiness for Modeling

EDA showed that the cleaned, year-aggregated dataset was suitable for linear modeling. Annual "Found" meteorite counts displayed moderate skewness, but no transformation was applied because the goal was to estimate a simple linear trend over time. Outlier years were retained, as they likely reflect real variation in search activity. The dataset contained a complete sequence of valid years, and scatterplots and summary statistics indicated sufficient variability and an approximately linear pattern, supporting the use of simple linear regression.

3. Model Specification and Statistical Framework:

3.1 - Model Selection and Rationale

To evaluate whether annual "Found" meteorite counts change over time, a simple linear regression model was used. This aligns with course methods for assessing the association between a numerical predictor (Year) and numerical outcome (annual discovery count). Exploratory scatterplots indicated an approximately linear pattern, supporting this choice.

3.2 - Formal Model Definition

3.3 - Hypothesis Testing Framework

3.4 - Regression Assumptions

Model validity relies on the assumptions reviewed in class (lecture 5 and 7):

* Linearity -
* Independence of Errors -
* Homoscedasticity -
* Normality of Errors -

Diagnostics supporting these assumptions appear in Appendix C.

4. Results:

4.1 - Descriptive Statistics

4.2 - Trend Estimation

4.3 - Model Fit and Goodness of Explanation

4.4 - Visual Evidence

5. Discussion:

5.1 - Interpretation of Findings

5.2 - Limitations

5.3 - Connection to Motivation

6. Conclusions:

6.1 - Summary of Main Findings

6.2 - Next Steps and Future Research

References :

[1] NASA, “OSIRIS-REx,” NASA Science, 2024. [Online]. Avalable: https://science.nasa.gov/mission/osiris-rex/

[2] U.S. General Services Administration, “Meteorite Landings,” Data.gov, 2024. [Online]. Available: https://catalog.data.gov/dataset/meteorite-landings

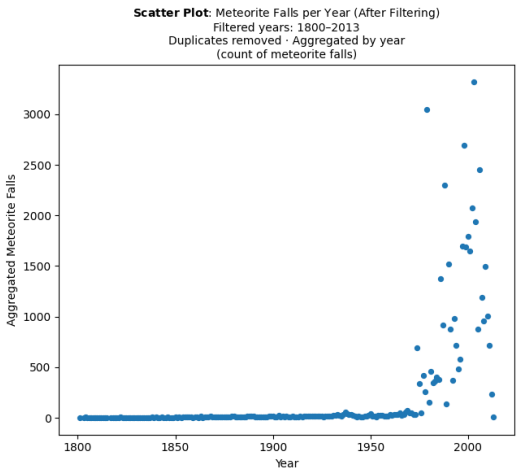
Appendix A — EDA Visuals

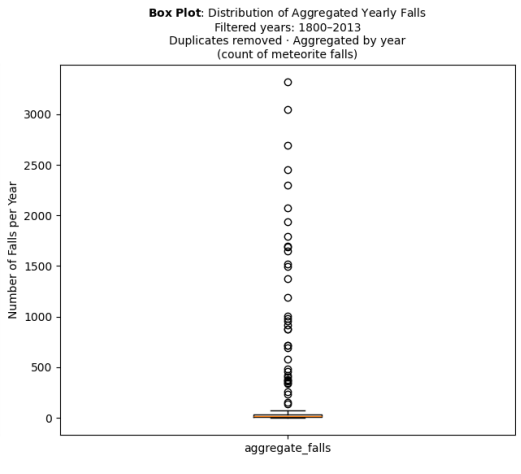
- Histogram of variable dist.
- Missing data heatmap
- initial scatter plots
- Boxplots showing outliers

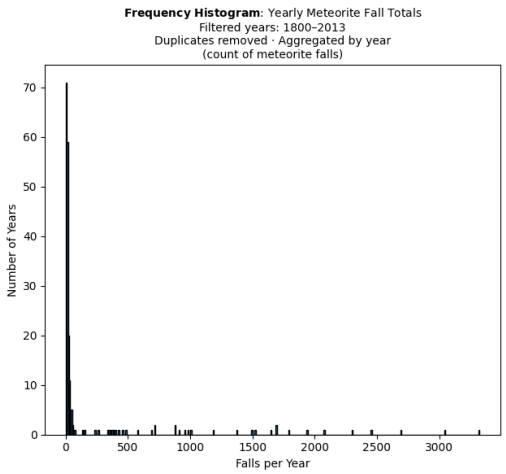
Table 1. Initial EDA Exploration

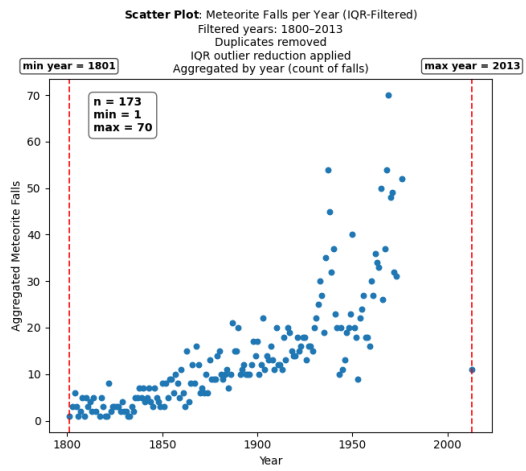
Feature Name	Pandas dtype	Categorical / Numerical	# Unique	% Missing	Description
name	object	Categorical	45716	0.00%	Name of the meteorite as recorded in the catalog.
id	int64	Numerical	45716	0.00%	Unique numeric identifier assigned to each meteorite record.
nametype	object	Categorical	2	0.00%	Indicates valid meteorite names ('Valid') or paired/duplicate names ('Relict').
recclass	object	Categorical	466	0.00%	Classification based on chemical and petrological type.
mass (g)	float64	Numerical	12576	0.29%	Reported mass of the meteorite in grams.
fall	object	Categorical	2	0.00%	Indicates whether the meteorite was 'Fell' (observed fall) or 'Found'.
year	datetime64[ns]	Numerical	265	0.64%	Year the meteorite was found or fell.
reclat	float64	Numerical	12738	16.00%	Latitude of the recovery site.
reclong	float64	Numerical	14640	16.00%	Longitude of the recovery site.
GeoLocation	object	Categorical	17100	16.00%	Coordinate pair representing the recovery location (latitude, longitude).

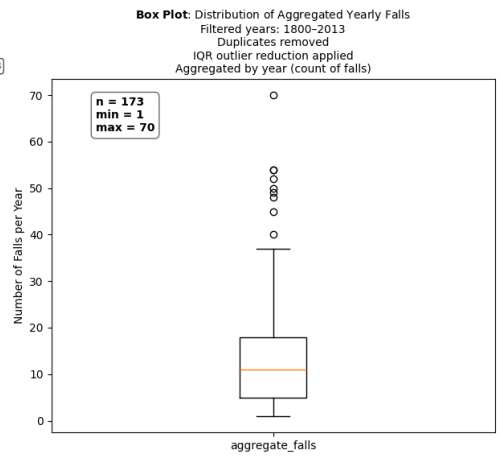
Table 1. Summary of the primary features in the NASA Meteorite Landings dataset, including pandas data types, inferred feature type, cardinality, missingness, and brief semantic descriptions used for subsequent exploratory analysis.

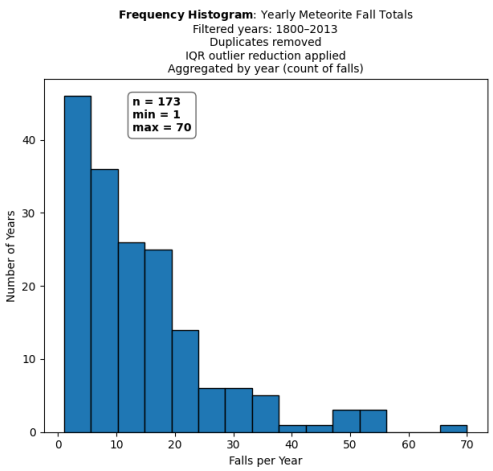












3.4 Regression Assumptions

Consistent with the regression assumptions outlined in Lectures 5 and 7, model validity relies on four conditions:

1. **Linearity** — The relationship between year and annual discovery count should be approximately linear. This was supported visually by the scatterplot.
2. **Independence of Errors** — Each year's count should be independent of others. Because observations represent distinct calendar years, this assumption is reasonable.
3. **Homoscedasticity** — The variability of residuals should be roughly constant across fitted values. This is evaluated through the residual-versus-fitted plot.
4. **Normality of Errors** — Residuals should be approximately normally distributed. A histogram and Q–Q plot of residuals were used to assess this.

As covered in class, these assumptions ensure valid inference for slope estimates and associated confidence intervals.

Appendix B — Data Cleaning Evidence

- Table of missing values before/after cleaning
- Description of rows removed/imputed
- Variable transformation examples

Appendix C — Model Diagnostics

- Residual Plot
- Q-Q Plot
- Influence/leverage plot

Appendix D — Extended Tables

- Full summary statistics
- Correlation matrix
- Categorical variable level counts

Researcher Bio-Sketch:

Nathan Herling is a first-year Master's student in Data Science at the University of Arizona and the lead contributor on this project. He holds Bachelor of Science degrees in Molecular Biology, Physics, and Electrical & Computer Engineering, with additional minors in Computer Science, Chemistry, and Mathematics. His interdisciplinary training spans computational modeling, machine learning, experimental physics, and full-stack software development. Nathan has conducted research in high-energy particle physics, serving as a Research Assistant in the Ken Johns group affiliated with CERN, where he contributes to muon spectrometer calibration and machine-learning–driven analyses for Long Lived Particle searches. His previous work includes developing reinforcement learning models for cognitive radio systems, security automation tools in industry, and supervised machine learning pipelines for engineering applications. Across academic, research, and industry roles, Nathan brings a leadership-driven, technically diverse, and data-focused perspective to the project.

Generative AI Tool Use Acknowledgment:

Generative AI tools, including **OpenAI's ChatGPT** and **Microsoft Copilot (image generation)**, were used to support this project. ChatGPT assisted with clarifying statistical concepts, refining written sections, organizing report structure, and generating explanatory text, while all analytical decisions, coding, and interpretation of results were performed independently by the author. The use of these tools followed an iterative prompting process, where multiple refinements were required to reach accurate, context-appropriate outputs; no single prompt produced a complete or final solution. Microsoft Copilot was used solely for generating illustrative images that supported conceptual understanding. All final methodological choices, analyses, and conclusions reflect the author's own work and judgment.