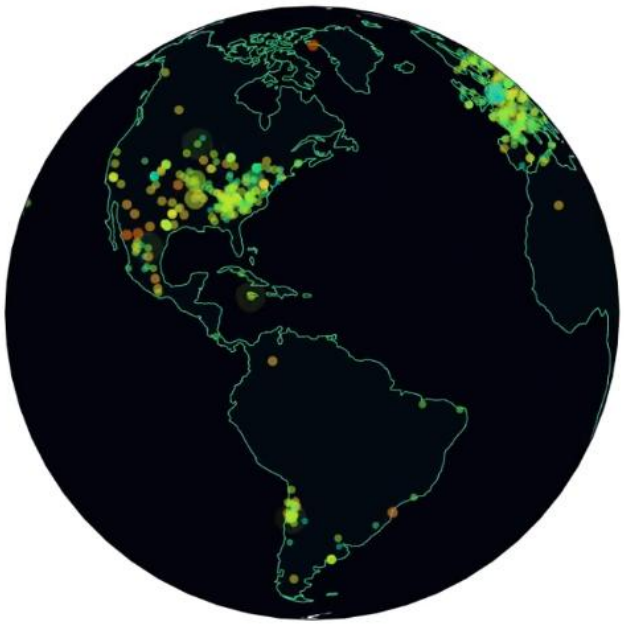
Nathan Herling

INFO 510

Foundations of Data Science

Fall 2025



A Temporal Analysis of Meteorite Findings

Using Data Hosted on the NASA Open Data Portal.

https://data.nasa.gov/dataset/meteorite-landings

Table of Contents

Note: main body of report is from pages 3-5 and meets the required 3-page length.

All appendices are justified in the sense of providing complete reproudceability, all figures are justitified as necessary to tell the story of the analysis. Complete code/scripts are provided in the repository and described in Appendix H – Code Geneology.

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/Fell" meteorites recorded on Earth. Put differently: Do yearly counts of recovered meteorites increase, decrease, or remain stable over time? To investigate the question, this 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 linear temporal trend.

2. Process and Analysis:

2.1 - Data Source and Access

The Meteorite Landings dataset used in this study is publicly available through Data.gov and maintained in collaboration with NASA and the Meteoritical Society. It contains global records of meteorite discoveries spanning several centuries, including meteorite type and recovery year – making it an exceptional choice for this project. Because the dataset contains only scientific observations and no personal identifiers, no privacy or IRB review was required. The Data can be located here: https://catalog.data.gov/dataset/meteorite-landings, and the raw .csv is located in the **Git Hub repository: _code/Meteorite_Landings.csv**. Further Data provenance is discussed in **Appendix J**.

2.2 - Data Preprocessing, Cleaning, and EDA Procedures

Preprocessing focused techniques discussed in [3] – i.e., on isolating information relevant to temporal discovery patterns of "Found/Fell" meteorites with EDA taking place in four phases (**Appendix Figure A1**). From the original dataset of over 45,000 records, only the **year** and **fall** (including "Found" and "Fell") variables were retained. Entries with invalid or missing years, years beyond 2013 (i.e. 2100), or discovery types other than "Found/Fell" were removed. Missing **year** values were imputed via entry removal. The missingness appeared to be **MCAR**, as there was no apparent relationship between missing dates and observed meteorite characteristics. Initial (raw) summary statistics are located in **Table B1**, and the final statistics used for the model in **Table D2**.

The cleaned data were aggregated to annual counts, producing a year-level time series. Exploratory data analysis evaluated distributional characteristics and temporal structure using summary statistics and visualizations. Supporting visuals provided in **Appendices B-E** – summary of statistics are also provided. All Python scripts are located in the repository: **https://github.com/N-Herling-Mk1/INFO_511_FA_25_Final_Proj_Repo.git**, each script is discussed in **Appendix H**.

2.3 - Assessment of Data Quality and Readiness for Modeling

Exploratory analysis indicated that the aggregated annual dataset contained sufficient temporal coverage and variability to support trend modeling. Although skewness (severe right) and outliers were present, these were retained as meaningful features of historical discovery patterns. Diagnostic summaries supporting

data readiness are provided in **Appendices B-E**.

3. Model Specification and Statistical Framework:

3.1 - 3.4 - Model Specification, Hypothesis Testing, and Assumptions

A simple linear regression framework (Figure 1) was used to assess the association between calendar year and annual "Found/Fell" meteorite counts, consistent with course methods for modeling numerical predictor–response relationships [5]. The model estimates a baseline intercept and a slope parameter representing the average yearly change in meteorite discoveries, providing a first-order approximation of long-term temporal trends. Hypothesis testing (Figure 2) was applied to the slope coefficient to evaluate whether the estimated trend differs significantly from zero, following standard regression inference procedures [5].

$$Y_t = \beta_0 + \beta_1 t + \varepsilon_t \quad (1)$$

Y_t: annual "Found" meteorite count | t: calendar year | ε_t: random error

β_0: intercept (expected count at baseline year) | β_1: linear trend per year

Figure 1. Linear Time-Series Regression Model for Annual Meteorite Counts.

Hypothesis Testing Framework

$\mathbf{H_0}$: $\beta_1 = 0$ (no linear temporal trend)

$\mathbf{H_A}$: $\beta_1 \neq 0$ (non-zero linear trend)

The p-value measures the probability of observing a slope as extreme as β_1 under the assumption that $\mathbf{H_0}$ is true.

Results are considered statistically significant when $p < 0.05$, indicating sufficient evidence to reject $\mathbf{H_0}$.

Figure 2. Hypothesis Testing Framework for Temporal Trend Assessment.

Model validity relies on the classical linear regression assumptions: **linearity**, **independence of errors**, **homoscedasticity**, and **normality of residuals**, as discussed in lecture [4]. These assumptions were evaluated using visual diagnostics, including scatterplots, residual-versus-fitted plots, and Q–Q plots, and summarized in **Appendix E**,**F**. Statistical significance was assessed using p-values for the slope parameter under the null hypothesis of no temporal association, with results interpreted in the context of both assumption diagnostics and model fit metrics [5], and evaluated in **Appendix G,F**, with the data only meeting 2 of the 4 required characteristics for a linear model (see section 4.1, Table F1).

4. Results:

4.1 – Two competing models

Following data cleaning and year-level aggregation, the analysis produced a complete annual time series of "Found/Fell" meteorite counts spanning multiple centuries. Because the raw count distribution exhibited strong right skew, two variance-stabilizing transformations were considered: a square-root(count) transformation (**Graph D1**) and a log(count + 1) transformation (**Graph D2**). Once the count values were transformed neither candidate model had IQR-outliers (**Graph D3-D4**). Summary statistics and transformed response characteristics for both candidate models are reported in **Appendix D – Table D2.**

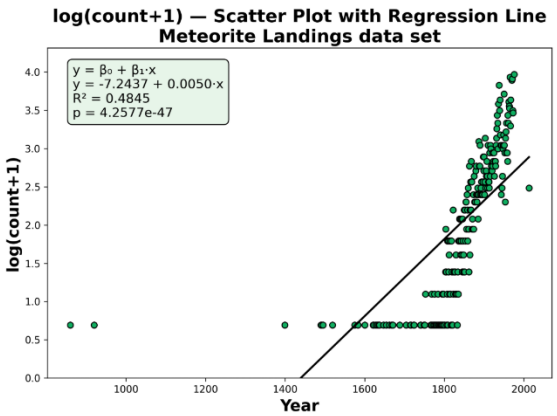
4.2 - Trend Estimation

Both transformed models estimate a positive linear association between calendar year (predictor) and annual meteorite discovery counts (outcome). The log(count + 1) model yields a slope estimate of $\beta_1 \approx 0.0050$, while the square-root model produces a slightly larger slope ($\beta_1 \approx 0.0076$), with both slopes statistically different from zero. Estimated coefficients and corresponding confidence intervals for each model are provided in **Appendix G, Tables G1-G3.**

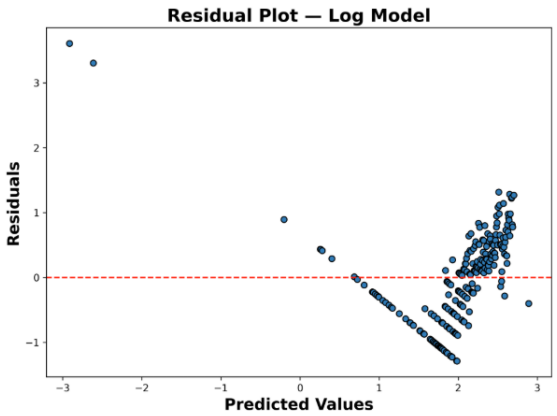
4.3 - Model Fit and Goodness of Explanation

Model performance metrics favor the log(count + 1) specification, which achieves higher explanatory power ($R^2 \approx 0.48$ vs. 0.43), lower prediction error (MAE, MSE, RMSE), and a larger F-statistic relative to the square-root(count) model. These results indicate that the log-transformed model more effectively balances goodness-of-fit and error reduction for the observed data. Full comparative fit statistics are summarized in **Appendix G, Table(s) G1-G3.**

4.4 - Visual Evidence



Graph 1. Annual "Found" meteorite counts over time with fitted linear regression (log(count + 1) model).



Graph 2. Residuals versus fitted values for the log(count + 1) regression model.

Visual inspection supports quantitative results: the log(count + 1) (Graph 1) scatter plot shows a clear upward temporal trend with residuals displaying no strong systematic structure (Graph 2), while complete diagnostics including histograms (**Graph D1**) and **Q–Q** plots (**Graph D5**) are provided in **Appendices D-E**.

4.5 – Interpretation of results

Both transformed models show statistically significant positive associations between calendar year and annual meteorite counts. **These results confirm the research question: yearly counts of recovered meteorites do increase over time. However, the low explanatory power and assumption violations indicate this trend cannot be adequately captured by simple linear regression.** The log(count + 1) model yields $\beta_1 = 0.0050$ ($p = 4.26 \times 10^{-47}$), while the square-root(count) model produces $\beta_1 = 0.0076$ ($p = 7.08 \times 10^{-38}$), with both providing strong evidence against the null hypothesis. However, explanatory power remains limited ($R^2 = 0.4845$ and 0.4282, respectively). Diagnostic results (**Appendix F**) show violations of two core regression assumptions for both models, indicating the relationship is not strictly linear.

5. Conclusions and Next Steps:

This study identifies a statistically significant temporal influence on meteorite recovery counts, though linear regression captures less than half the variability ($R^2 = 0.4845$). The contrast between strong significance ($p < 10^{-40}$) and limited explanatory power exemplifies time-series data where persistent trends yield significant coefficients despite unexplained variance patterns. Findings suggest discoveries reflect complex; nonlinear drivers tied to observations, e.g. - institutional reporting, and scientific priorities. Future research should prioritize nonlinear or time-series models (e.g., ARIMA, exponential smoothing) and incorporate historical and institutional predictors—such as funding levels, technological advances, and field collection priorities—to more characterize the complex dynamics underlying meteorite discovery rates.

References:

[1] NASA, “OSIRIS-REx,” NASA Science, 2024. [Online]. Available: 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

[3] A. Cruze, “Exploratory Data Analysis and Data Preparation,” INFO 511 Lecture 4, University of Arizona, Sep. 18, 2025.

[4] A. Cruze, “Regression Assumptions and Diagnostics,” INFO 511 Lecture 5, University of Arizona, Sep. 30, 2025.

[5] A. Cruze, “Inference and Linear Regression,” INFO 511 Lecture 7, University of Arizona, Oct. 9, 2025.

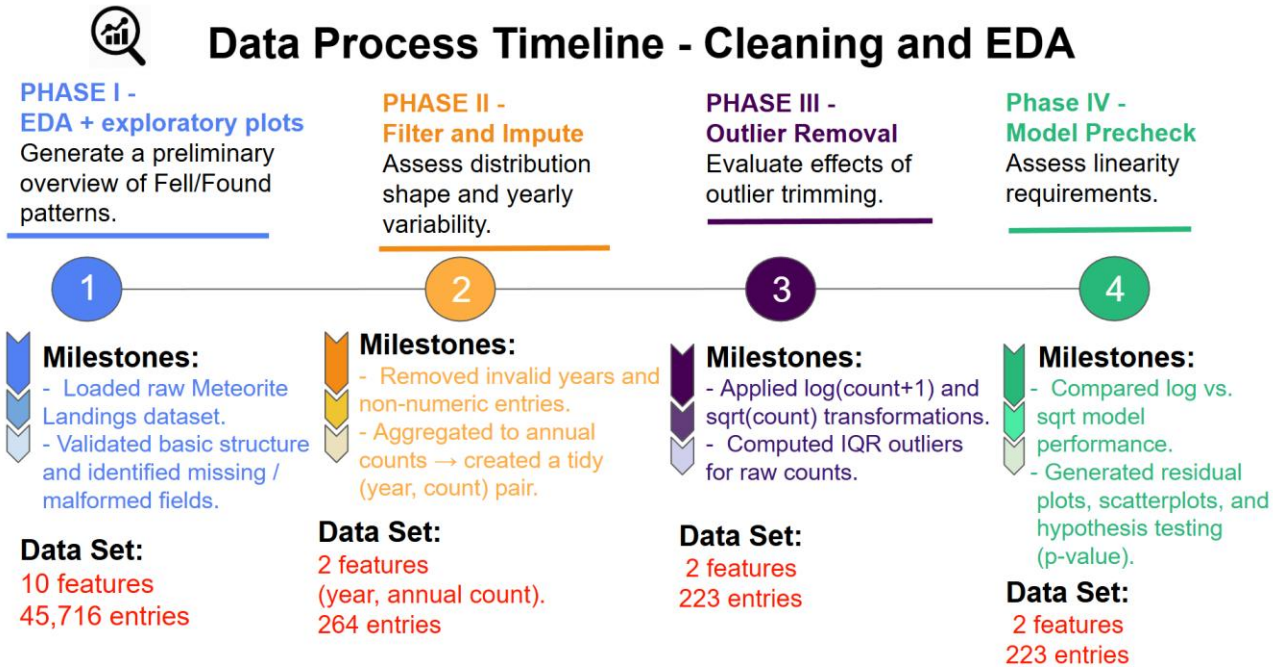


Figure A1. EDA phases – for the project: A Temporal Analysis of Meteorite Findings

Figure A1. summarizes a **Four-Phase EDA plan** data processing timeline for cleaning and exploratory data analysis (EDA) in a temporal study of meteorite findings.

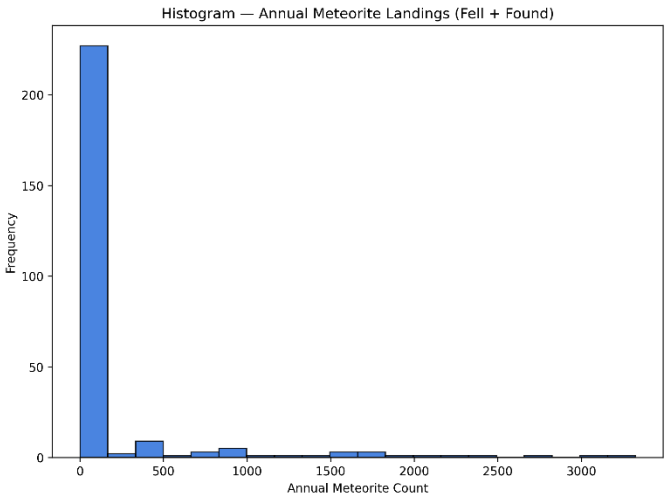
Exploratory summaries and distributional diagnostics were conducted following standard EDA practices for assessing skewness, dispersion, and outliers prior to modeling [3]. **Phase I** establishes an initial understanding through exploratory plots, validating dataset structure and identifying missing or malformed fields in the raw dataset (10 features, 45,716 entries). **Phase II** focuses on filtering and imputation, removing invalid years and non-numeric values and aggregating observations into a tidy annual count format (2 features, 264 entries). **Phase III** addresses outlier removal by applying log transformations and IQR-based methods to assess and mitigate the influence of extreme values, yielding a refined dataset (2 features, 223 entries). Finally, **Phase IV** performs model prechecks by evaluating linearity assumptions, comparing raw versus log-scaled models, and generating residual diagnostics and hypothesis tests, ensuring the cleaned dataset is suitable for downstream statistical modeling.

Meteorite Landings Raw Data Table

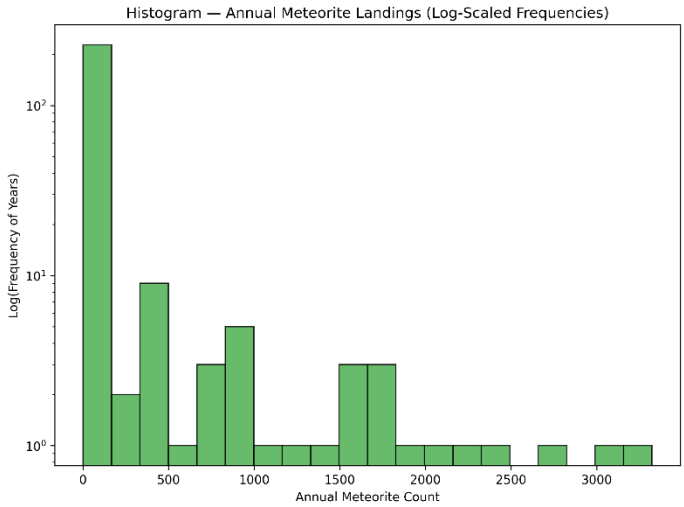
Dataset Size: 45,716 rows × 10 columns

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	float64	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 B1. Meteorite Landings Data Set – unfiltered.

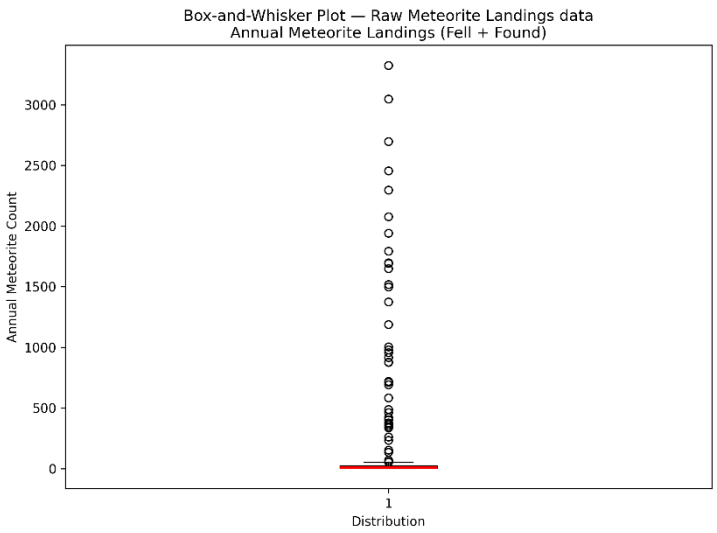


Graph B1. Meteorite Landings Data Set (unfiltered) Histogram.



Graph B2. Meteorite Landings Data Set (unfiltered) Log Transformed Histogram.

Phase I explores the unfiltered Meteorite Landings dataset contains 45,716 observations across 10 variables, encompassing a mix of categorical and numerical attributes with varying levels of completeness. **Table B1** documents the raw structure of the dataset, showing substantial missingness in several spatial and mass-related fields and motivating the restriction of the analysis to variables relevant for temporal modeling. **Graph B1** illustrates the extreme right skew in annual meteorite landing counts (Fell + Found), with most years exhibiting low frequencies and a small number of years containing very large counts. The log-scaled histogram in **Graph B2** further highlights the heavy-tailed nature of the distribution, supporting the use of transformation and aggregation strategies in subsequent modeling steps.



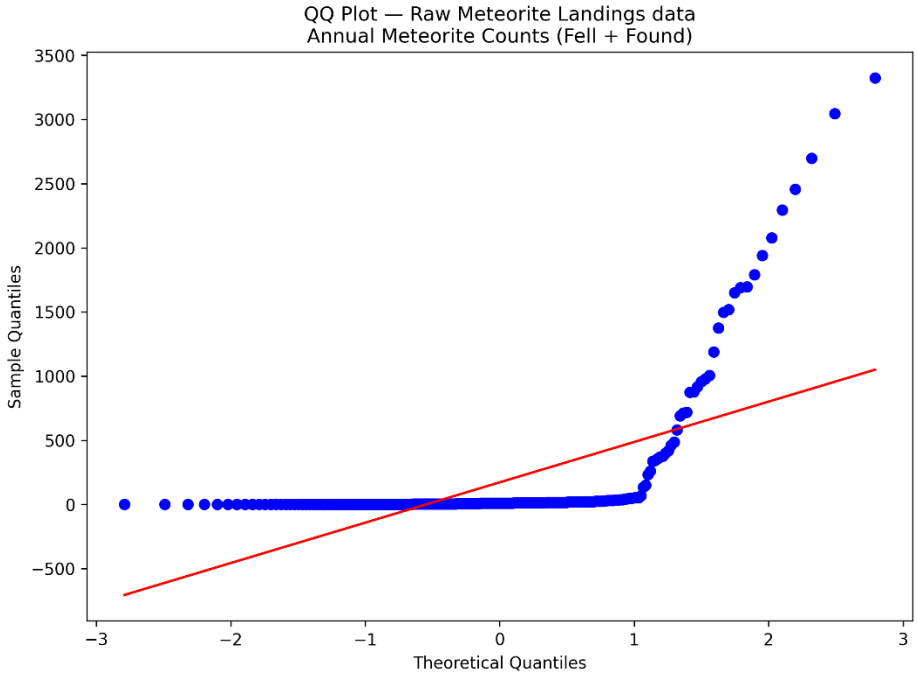
Graph B3. Boxplot of annual meteorite landing counts (raw data).

Outlier Summary for Annual Meteorite Counts
Raw data Meteorite Landings data set

Dataset Size: 45,423 rows × 10 columns

Statistic	Value
Total Count (years)	263
Min (annual count)	1
Max (annual count)	3323
Mean	172.711
Standard Deviation	506.286
25th Percentile	2.0
50th Percentile (Median)	10.0
75th Percentile	22.5
% Within IQR Range	84.41%
Number of Outliers	41

Table B2. Outlier summary statistics for annual meteorite counts.



Graph B4. Q–Q plot of annual meteorite landing counts (raw data).

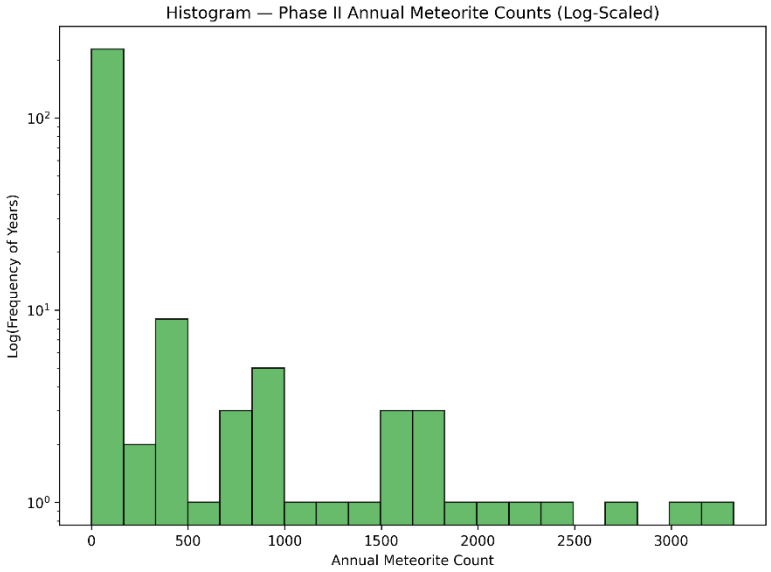
Phase I explores the raw annual meteorite landing counts exhibit substantial dispersion and extreme skewness prior to any filtering or transformation. **Graph B3** shows a highly asymmetric distribution with numerous extreme values, indicating the presence of substantial outliers relative to the central mass of the data. **Table B2** quantifies this behavior, revealing a wide interquartile range, a large maximum value relative to the median, and multiple extreme observations, consistent with heavy-tailed behavior. The Q–Q plot in **Graph B4** further demonstrates strong departures from normality, particularly in the upper tail, reinforcing that the raw annual counts violate standard distributional assumptions and motivating transformation and outlier-aware modeling strategies in subsequent analysis.

Phase II Master Table

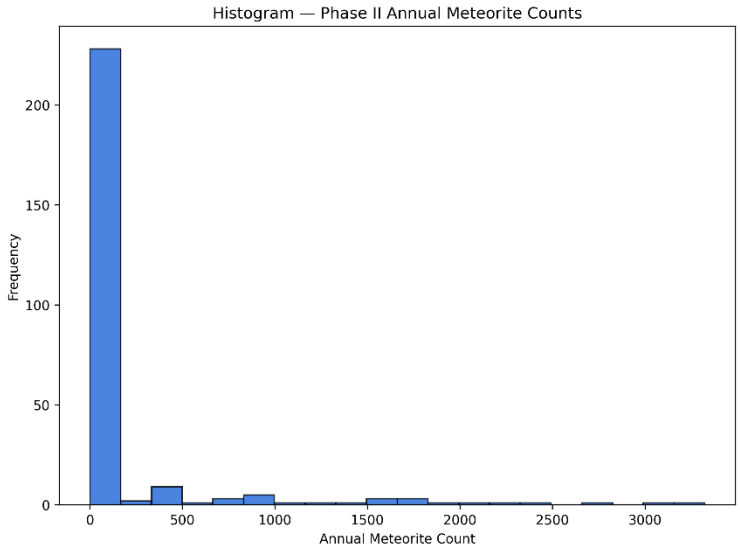
Dataset Size: 264 rows × 2 columns

Feature Name	Pandas dType	Categorical / Numerical	# Unique	% Missing	Description
year	int64	Numerical	264	0.00%	4-digit year indicating the recorded meteorite event.
count	int64	Numerical	81	0.00%	Number of meteorites recorded for that year.

Table C1. Phase II master dataset structure (year-level aggregation valid year filter).

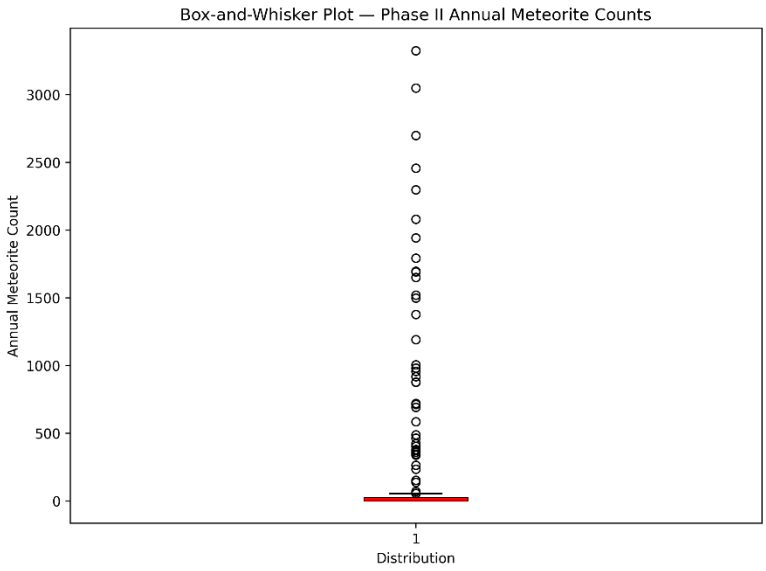


Graph C1. Phase II Histogram of annual meteorite counts (log-scaled).



Graph C2. Histogram of annual meteorite counts (raw scale).

Phase II consolidates the cleaned data into a year-level dataset - binned by year with a valid year filter (800-2013) - consisting of 264 observations and two numerical variables: calendar year and annual meteorite count. **Table C1** documents the resulting structure, confirming the absence of missing values and the suitability of the dataset for regression-based analysis. **Graph C2** shows that annual counts remain highly right-skewed on the raw scale, with a small number of years exhibiting disproportionately large values. The log-scaled histogram in **Graph C1** reduces this skew and reveals underlying structure across lower-frequency years, further motivating the use of variance-stabilizing transformations in subsequent modeling steps.



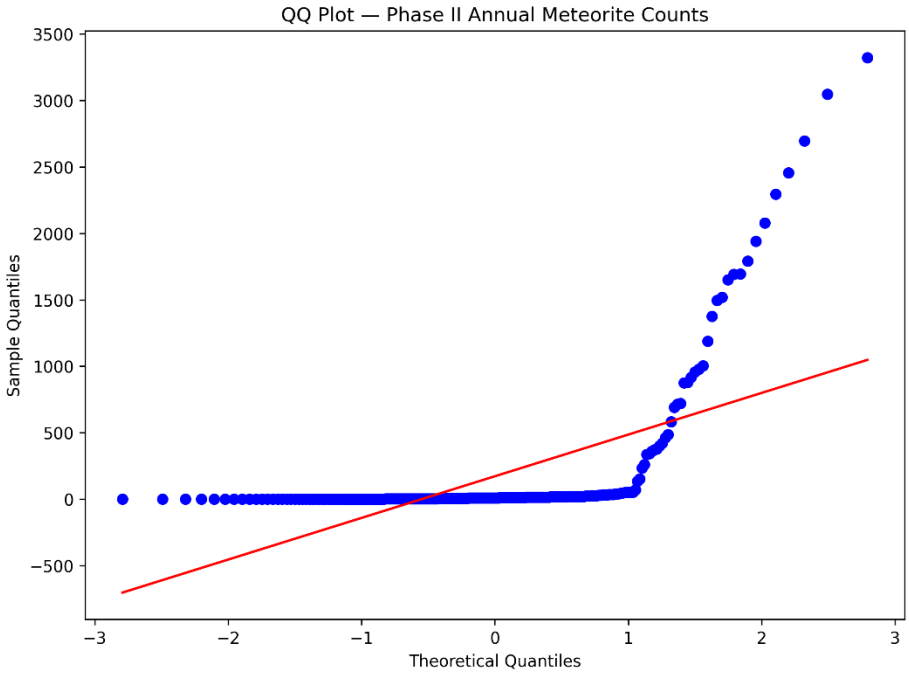
Graph C3. Boxplot of annual meteorite counts (Phase II, untransformed).

Phase II Outlier Summary

Dataset Size: 264 rows × 2 columns

Statistic	Value
Total Years in Dataset	264
Min (annual count)	1
Max (annual count)	3323
Mean	172.061
Standard Deviation	505.437
25th Percentile	2.0
50th Percentile (Median)	10.0
75th Percentile	22.25
% Within IQR Range	84.47%
Number of Outliers	41

Table C2. Outlier summary statistics for annual meteorite counts (Phase II).



Graph C4. Q–Q plot of annual meteorite counts (Phase II).

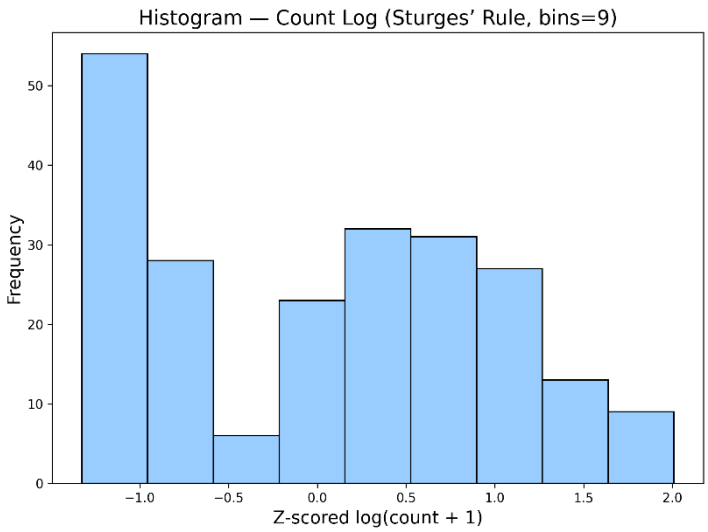
Phase II examines the distributional characteristics and outlier structure of the aggregated annual meteorite count data prior to transformation. **Graph C3** shows a highly skewed distribution with a large concentration of low-count years and numerous extreme values, indicating substantial dispersion relative to the central tendency. **Table C2** quantifies this behavior, reporting a median of 10 counts per year, a maximum exceeding 3,300, and 41 observations identified as outliers under the IQR criterion, with only 84.47% of values falling within the interquartile range. The Q–Q plot in **Graph C4** reveals strong departures from normality, particularly in the upper tail, confirming that the untransformed annual counts violate standard distributional assumptions and motivating the use of variance-stabilizing transformations in subsequent EDA phases.

Phase III — Master Feature Table

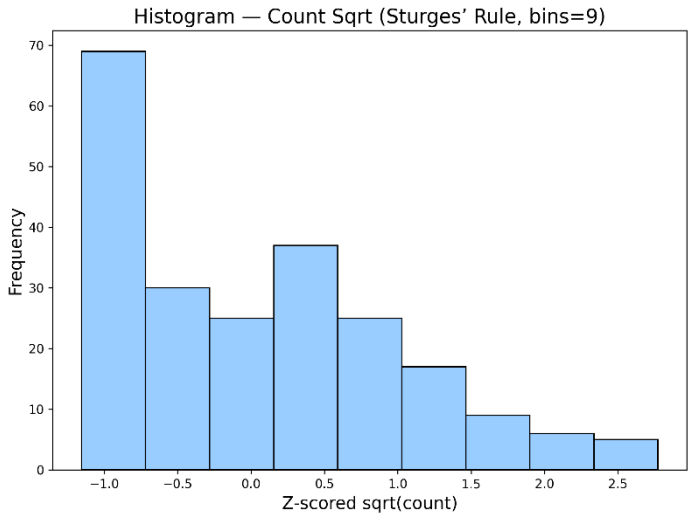
Dataset Size: 223 rows × 4 columns

Feature Name	Pandas dtype	Categorical / Numerical	# Unique	% Missing	Description	Transform Summary
year	int64	Numerical	223	0.00%	Four-digit calendar year of meteorite record.	Not transformed
count	int64	Numerical	41	0.00%	Number of meteorites recorded for the given year.	Not transformed
count_log	float64	Numerical	41	0.00%	Log-transformed annual meteorite count: log(count + 1).	mean=1.997, std=0.985, min=0.693, max=3.970
count_sqrt	float64	Numerical	41	0.00%	Square-root-transformed annual meteorite count.	mean=2.826, std=1.585, min=1.000, max=7.211

Table D1. Phase III master feature table with transformation summary.



Graph D1. Histogram of z-scored log(count + 1) transformed annual counts.



Graph D2. Histogram of z-scored square-root–transformed annual counts.

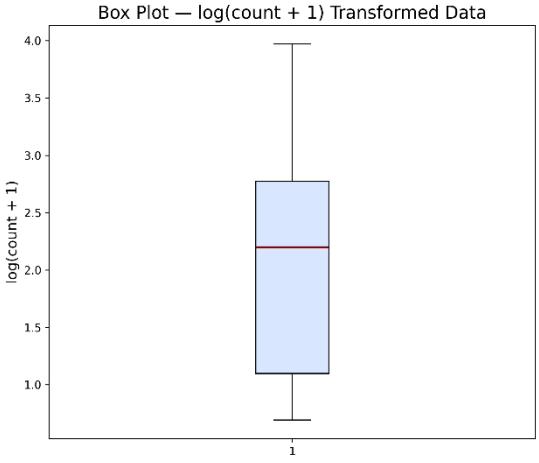
Phase III consolidates the transformed variables used for outlier-aware modeling and documents their statistical properties. **Table D1** summarizes the master feature set, indicating that the raw annual count variable was retained alongside two variance-stabilizing transformations, log(count + 1) and square-root(count), while the year variable remained untransformed. The histograms in **Graphs D1 and D2** show the standardized (z-scored) distributions of the transformed counts, demonstrating reduced skewness and more balanced spread relative to the raw data. These results confirm that both transformations improve distributional symmetry and comparability, supporting their use in subsequent regression and model diagnostic analyses – while neither transforms the data into the desired normal curve.

Phase III Outlier Summary

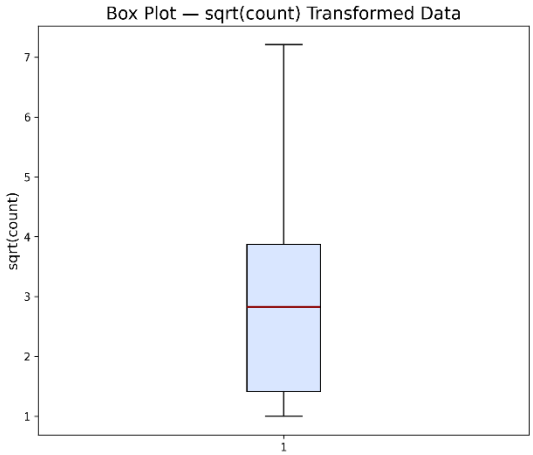
Dataset Size: 223 rows × 4 columns

Column	Total Count	Min	Max	Mean	Std Dev	25th %ile	50th %ile	75th %ile	% Within IQR	Outliers
year	223	860.0	2013.0	1835.883	135.896	1804.500	1861.000	1916.500	93.27%	15
count	223	1.0	52.0	10.489	10.736	2.000	8.000	15.000	95.52%	10
count_log	223	0.6931471805599453	3.970291913552122	1.997	0.983	1.099	2.197	2.773	100.00%	0
count_sqrt	223	1.0	7.211102550927978	2.826	1.581	1.414	2.828	3.873	100.00%	0

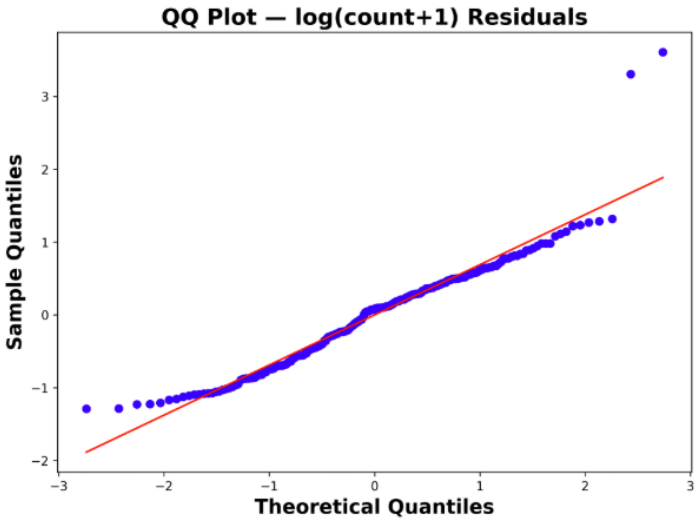
Table D2. Outlier summary statistics for transformed annual meteorite counts.



Graph D3. Boxplot of log(count + 1) transformed annual counts.

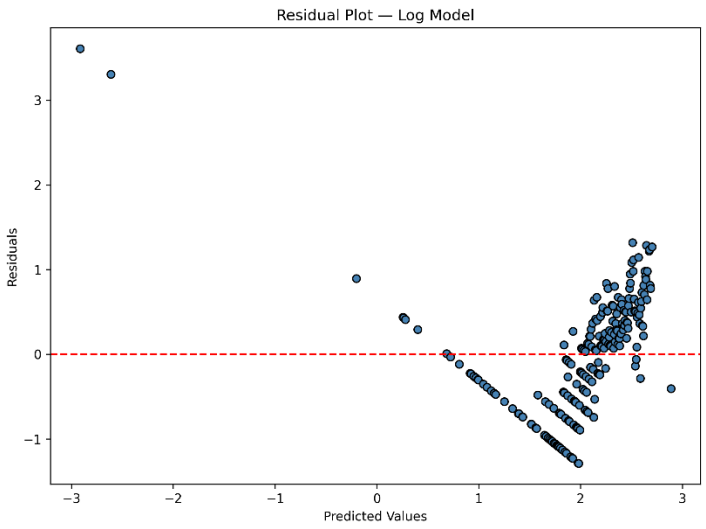


Graph D4. Boxplot of square-root–transformed annual counts

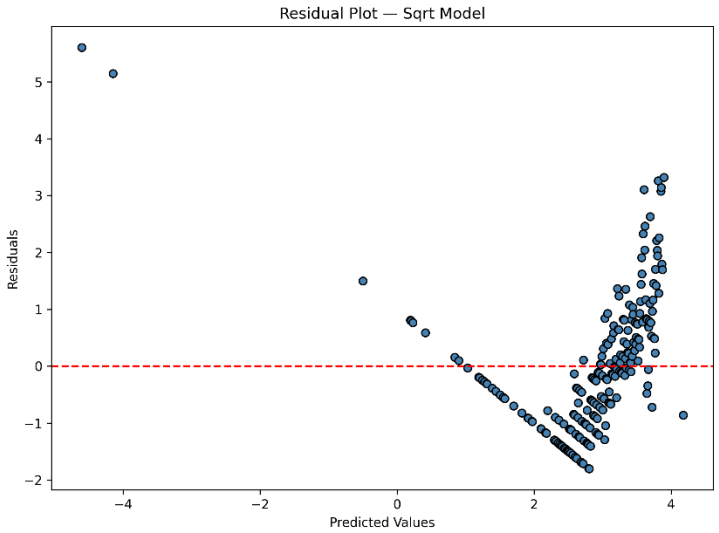


Graph D5. Q–Q plot of log(count + 1) model residuals.

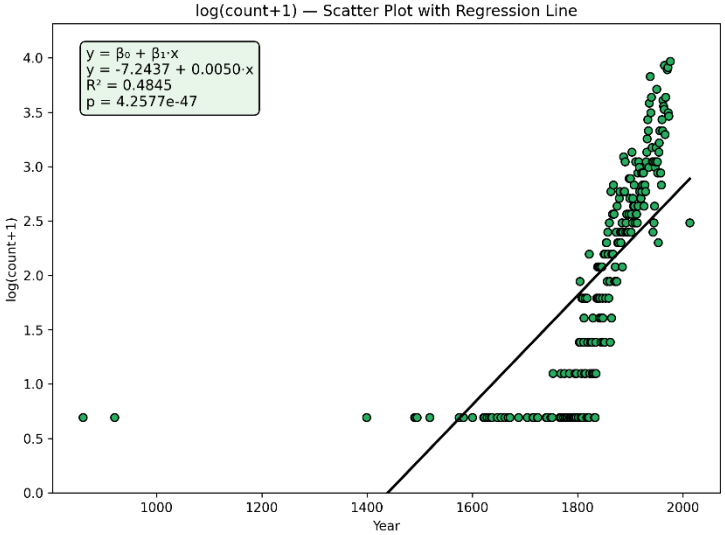
Phase III evaluates the impact of outliers and variance-stabilizing transformations on the annual meteorite count data. **Table D2** summarizes distributional properties before and after transformation, showing that both log(count + 1) and sqrt(count) transformations substantially reduce the influence of extreme values, with all observations falling within the interquartile range after transformation. The boxplots in **Graphs D3 and D4** illustrate improved symmetry and reduced dispersion relative to the raw counts, indicating greater suitability for regression analysis. However, the Q–Q plot in **Graph D5** reveals persistent departures from normality in the residuals, particularly in the upper tail, confirming that while transformations mitigate outlier effects, they do not fully satisfy linear model assumptions – in the model that appears to be the best of the options: log(count+ 1).



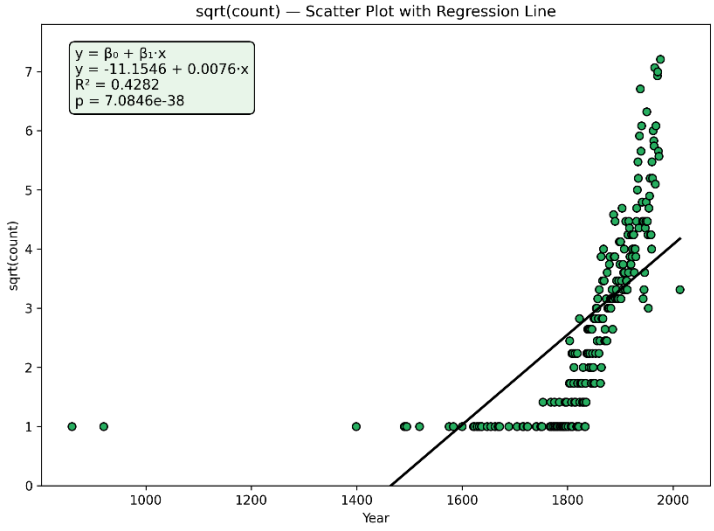
Graph E6. Residuals versus fitted values for log(count + 1) model.



Graph E7. Residuals versus fitted values for square-root(count) model.



Graph E8. Scatter plot of log(count + 1) with fitted linear regression line.



Graph E9. Scatter plot of square-root(count) with fitted linear regression line.

Phase IV evaluates diagnostic behavior and linear trend structure for both transformed models. After both the square-root and log(count_1) transforms – no outliers remained (**Table E2**). **Graph E6-E7** show residual plots for the log(count + 1) and square-root(count) specifications reveal non-random structure and heteroscedasticity-violation, while scatter plots (**Graphs E8-E9**) show a clear upward temporal trend with substantial dispersion. The log(count + 1) model is presented as the primary specification because it achieves **slightly higher explanatory power** than the square-root model (R^2 = 0.4845 vs. 0.4282; see **Appendix F – Model Metrics and Results**), not because it satisfies all or more linear regression assumptions as compared to the sqr-root(count) model.

Appendix F – Model Description and Assumptions

Assumption	Diagnostic Used	Assessment	Conclusion
Linearity	Scatter plot of log(count+1) vs. year with regression line	Approximately linear upward trend	**Assumption met**
Independence of Errors	Study design (yearly counts as independent observations)	Each observation corresponds to a distinct calendar year	**Assumption met**
Homoscedasticity	Residuals vs. fitted values plot	Clear fan-shaped structure and increasing variance	**Assumption violated**
Normality of Errors	Histogram and Q–Q plot of residuals	Deviations in tails; departure from reference line	**Assumption violated**

Table F1. Regression assumptions for linear independence

The Assumption of Data Independence

Individual meteorite fall events are independent physical phenomena, but the *annual counts* used in this analysis form a time-indexed series, which may exhibit autocorrelation due to historical, observational, and reporting effects.

Regression Assumptions – Table F1.

Linearity – Linearity was assessed using a scatter plot of **log(count + 1) versus year** with a fitted regression line. After transformation, the relationship exhibits an approximately linear upward trend, despite some curvature driven by early and late outliers. Overall, the linearity assumption was judged to be reasonably satisfied for the transformed model.

Independence of Errors - was evaluated based on the study design, where each observation represents a **distinct calendar year**. While individual meteorite fall events are independent physical phenomena, aggregating them into annual counts introduces a time-indexed structure. For modeling purposes, the yearly observations were treated as independent, and this assumption was considered met

Homoscedasticity - was assessed using a **residuals versus fitted values plot** from the log(count + 1) model. The plot showed a clear fan-shaped pattern, with residual variance increasing as fitted values increase. This indicates non-constant variance, and therefore the homoscedasticity assumption is violated. See Graph D6.

Normality of Errors **-** Normality of errors was evaluated using both a **histogram** (Graph D1) and a **Q–Q plot** (Graph D5) of the residuals. Deviations were observed in the tails, with residuals departing from the reference line, indicating heavier-than-normal tails. As a result, the normality assumption was not fully satisfied.

Model	R^2	Adj R^2	MAE	MSE	RMSE	F-statistic
Log(count+1)	0.4845	0.4822	0.5528	0.4977	0.7055	207.7481
Sqrt(count)	0.4282	0.4256	0.9037	1.4301	1.1959	165.5081

Table G1. Comparative model fit and error metrics for transformed linear models.

Model	Intercept β_0	Slope β_1	Intercept CI [low,high]	Slope CI [low,high]	Equation
Log(count+1)	-7.2437	0.0050	[-8.5037, -5.9837]	[0.0043, 0.0057]	log(count+1) = -7.2437 + 0.0050·year
Sqrt(count)	-11.1546	0.0076	[-13.2904, -9.0188]	[0.0065, 0.0088]	sqrt(count) = -11.1546 + 0.0076·year

Table G2. Estimated regression coefficients and confidence intervals for candidate models.

Model	p-value	Decision	Interpretation
Log(count+1)	4.2577e-47	✓ Reject H_0	Evidence of a statistically significant linear trend.
Sqrt(count)	7.0846e-38	✓ Reject H_0	Evidence of a statistically significant linear trend.

Table G3. Hypothesis test results for temporal trend significance.

Appendix G - summarizes the quantitative performance and inferential results for the two competing linear specifications. **Table G1** shows that the log(count + 1) model achieves modestly higher explanatory power (R^2 = 0.4845 vs. 0.4282), lower prediction error, and a larger F-statistic than the square-root model, motivating its selection for presentation despite overall limited fit. **Tables G2 and G3** indicate that both models yield extremely small p-values ($p \approx 10^{-47}$ and 10^{-38}), providing overwhelming evidence that the slope differs from zero. In statistical terms, this combination of moderate R^2 and very low p-values implies that calendar year explains only a limited fraction of the total variance in meteorite counts, yet still has a **real, measurable effect** on the response. Such behavior is common in time-series feature spaces, where persistent directional trends and large sample sizes can produce strong statistical significance even when substantial unexplained variability remains, reinforcing that significance does not imply strong predictive performance.

Code Genealogy — EDA Phases I–IV

Python files and their purpose – included in the code repository.

Phase I — Raw Data Profiling [File name: File Description]

Repository Location: _code/0_EDA_Phase_I

(1) - **EDA_Phase_1.py**: Driver script for Phase I exploratory analysis on raw Meteorite Landings data.
(2) - **0_EDA_phase_I_Master_Table.py**: Generates master metadata table for raw dataset.
(3) - **0_EDA_phase_I_Histogram.py**: Plots raw annual count distributions.
(4) - **0_EDA_phase_I_BoxPlot.py**: Creates box-and-whisker plot for outlier visibility.
(5) - **0_EDA_phase_I_QQplot.py**: QQ plot assessing normality of raw annual counts.
(6) - **0_EDA_phase_I_outLier_Table.py**: Computes IQR-based outlier summary for raw counts.

Phase II — Filtering & Aggregation [File name: File Description]

Repository Location: _code/0_EDA_Phase_II

(1) - **0_EDA_phase_II_df_maker.py**: Cleans raw data and aggregates yearly meteorite counts.
(2) - **0_EDA_phase_II_Master_Table.py**: Metadata summary of year-level dataset.
(3) - **0_EDA_phase_II_Histogram.py**: Histogram of aggregated annual counts.
(4) - **0_EDA_phase_II_BoxPlot.py**: Box plot highlighting heavy-tailed behavior.
(5) - **0_EDA_phase_II_QQplot.py**: QQ plot of annual counts pre-transformation.
(6) - **0_EDA_phase_II_OutlierTable.py**: Outlier statistics for aggregated counts.

Phase III — Transformation & Topology [File name: File Description]

Repository Location: _code/0_EDA_Phase_III

(1) - 0_EDA_phase_III_data_transform.py: Applies log(count+1) and sqrt(count) transformations with IQR trimming.
(1) - **0_EDA_phase_III_master_table.py**: Feature-level summary including transformed variables.
(2) - **0_EDA_phase_III_Histogram_plot.py**: Histograms of transformed count distributions.
(3) - **0_EDA_phase_III_Box_plot.py**: Box plots comparing transformed distributions.
(4) - **0_EDA_phase_III_QQ_plot.py**: QQ plots for transformed variables.
(5) - **0_EDA_phase_III_Data_Topology_check.py**: Quantifies skewness and tail behavior to justify transformations.
(6) - **0_EDA_phase_III_OutlierTable.py**: Outlier comparison across raw and transformed scales.

Phase IV – Final Model Decisions.

Repository Location: _code/0_EDA_Phase_IV

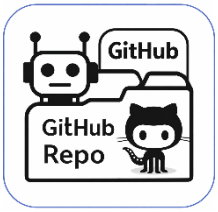
(1) - **0_EDA_phase_IV_analysis.py**: Automated regression assumption testing across candidate models.
(2) - **0_EDA_phase_IV_compare_contrast_test.py**: Compares log vs sqrt models using metrics, CIs, and hypothesis tests.
(3) - **0_EDA_phase_IV_presentation_table.py**: Generates final presentation-ready feature table.

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.



Git Hub Repository:

https://github.com/N-Herling-Mk1/INFO_511_FA_25_Final_Proj_Repo.git

This repository presents a four-phase exploratory data analysis of global meteorite discovery records from 800 to 2013, completed for **INFO 511 – Foundations of Data Science at the University of Arizona – Fall 25**. The analysis pipeline progresses systematically from raw data profiling through aggregation, statistical transformation (logarithmic and square-root), and regression diagnostics. Each phase generates modular Python scripts that produce distribution plots, outlier analyses, and normality assessments to guide modeling decisions. The project features an animated geospatial visualization encoding meteorite mass through color mapping, revealing two centuries of discovery patterns – examining the years 800-2013 of meteorite records. This work demonstrates structured statistical analysis with reproducible workflows and effective visual communication of temporal and spatial trends in meteorite recovery data.

Where did the data come from?

The data comes from The Meteoritical Society and is hosted on the NASA Open Data Portal (data.nasa.gov). The dataset was originally collected and curated by Javier de la Torre as a Fusion Table and has been maintained and updated by NASA Public Data.

Why this data?

This dataset was selected to explore temporal and spatial patterns in meteorite discoveries over multiple centuries (800-2013). The comprehensive nature of the dataset—containing ~45,000 meteorite records with geospatial coordinates, mass measurements, and discovery classifications—enables statistical analysis of trends in scientific discovery efforts and the distribution of meteorite falls versus finds across time and geography.

How was it accessed?

The data is publicly accessible through NASA's Open Data Portal at https://data.nasa.gov/dataset/meteorite-landings. It is available in multiple formats (CSV, JSON, XML, RDF) for direct download without authentication or special permissions. The dataset was accessed via direct download of the CSV file.

Privacy/Security/IRB Considerations

No IRB required: This is publicly available, non-human subject data

Security: The data is hosted on a secure government portal (data.nasa.gov)

Privacy: Not applicable—data concerns geological/astronomical objects, not individuals

Public & Anonymized: Yes, the data is fully public domain with no personal information

Who owns the data and how was it originally collected?

The data is owned by The Meteoritical Society and maintained by NASA. The dataset represents a comprehensive catalog of known meteorite landings compiled from historical records, scientific expeditions, and verified meteorite recovery reports submitted to The Meteoritical Society. The original collection spans multiple decades of scientific observation and documentation, with the most recent update occurring on May 14, 2013 (with subsequent maintenance updates through 2023).

Citation

NASA Public Data. (2023). Meteorite Landings [Data set]. NASA Open Data Portal. https://data.nasa.gov/dataset/meteorite-landings (Original data from The Meteoritical Society; last modified January 31, 2023)

Peer review recommendations response page:

Instructions: You must list the recommendations that your peer made and respond to their comments/recommendations.

(1) Recommendation:

Milestone 3 report largely unfinished, fill in current gaps.

(1) Response:

The Gaps for milestone 3 have largely been filled in. While still not quite the final document, major gaps in EDA, analysis, and validation have been addressed.

(2) Recommendation:

Slide show largely unfinished, fill in current gaps.

(2) Response:

Gaps will be filled in as the milestone and final project are finished. It's noted that the slide show is due:

(3) Recommendation:

For main report conclusion and next step - flush out ideas

(3) Response:

This has been addressed in milestone 3. Future steps include ideas for clustering algorithms to see geographical patterns in meteorite finds or fell observations, checking parallel data sets – such as population density and meteorite locations, and establishing a research question to see if any spike in 'found' meteorites can be ascribed to sociological causes.

(4) Recommendation:

For git hub repository, make sure the updated repository structure is reflected in the **README.md file.**

(4) Response:

This will be handled during construction of the final draft/repository construction – for the Milestone 4/Final Report – due December-16-2025.