A black rectangular sign with green and yellow text

AI-generated content may be incorrect.   
 **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 — Phase I  
Phase I examines the raw data for the *Meteorite Landings* dataset.

A table with numbers and text

AI-generated content may be incorrect.**Table 1** summarizes the raw data as it exists in the *Meteorite Landings* dataset, the original data set consists of ~46,000 entries and 10 features per entriy.

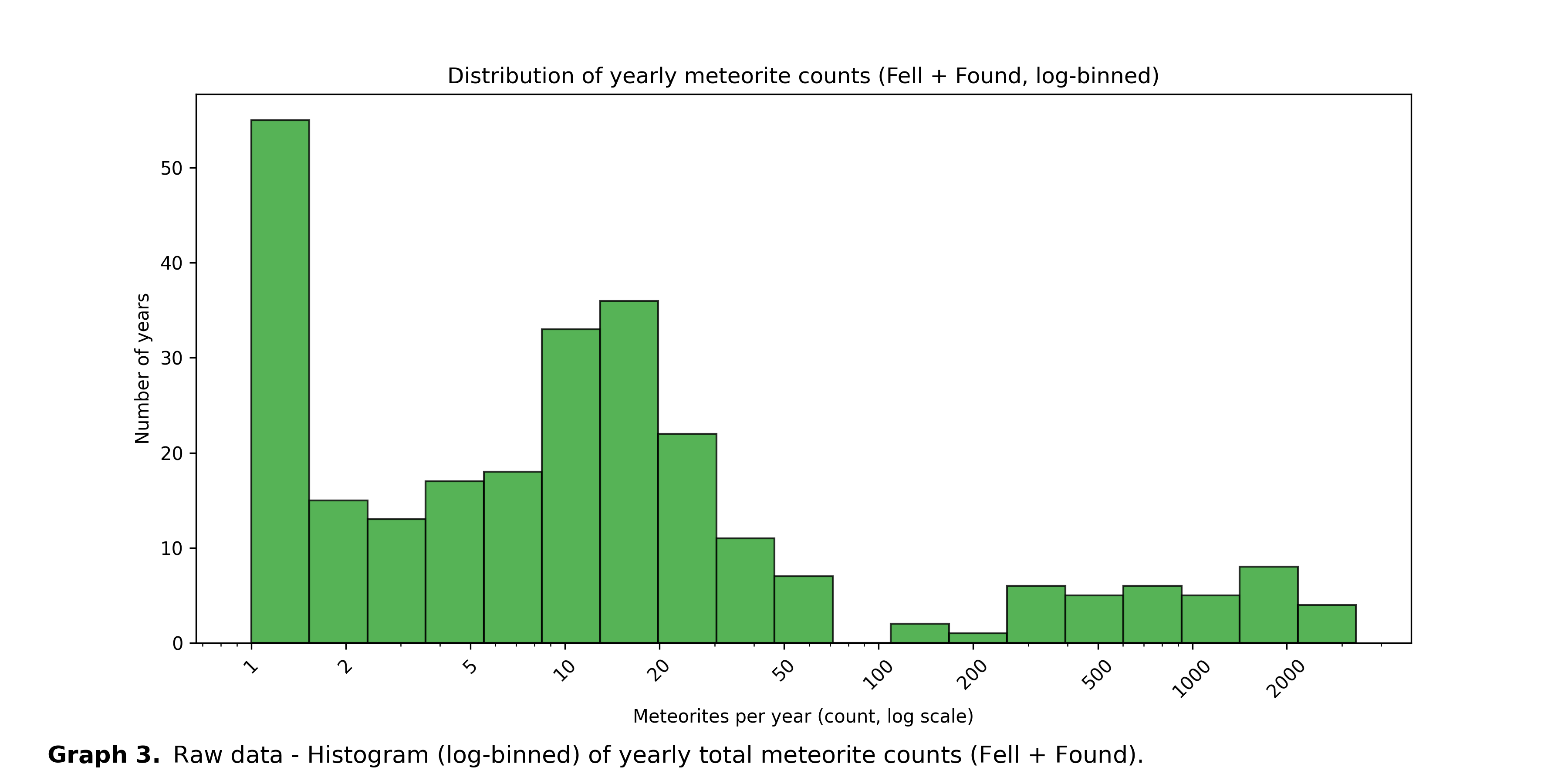
A white paper with black text

AI-generated content may be incorrect.

**Graph 1** shows a scatter plot of the raw data from the *Meteorite Landings* dataset. Further encoding is used to show Fell (observed falling meteorites) vs. Found (no observation of fall).

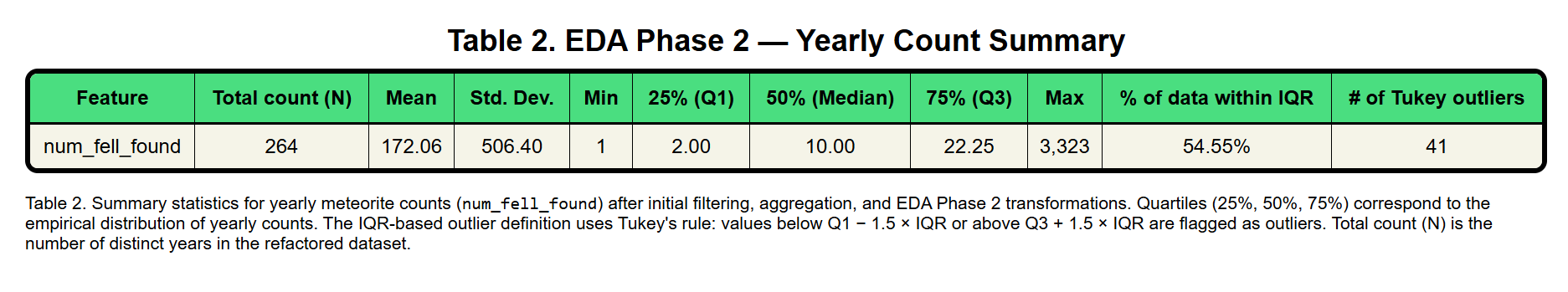
A screenshot of a computer

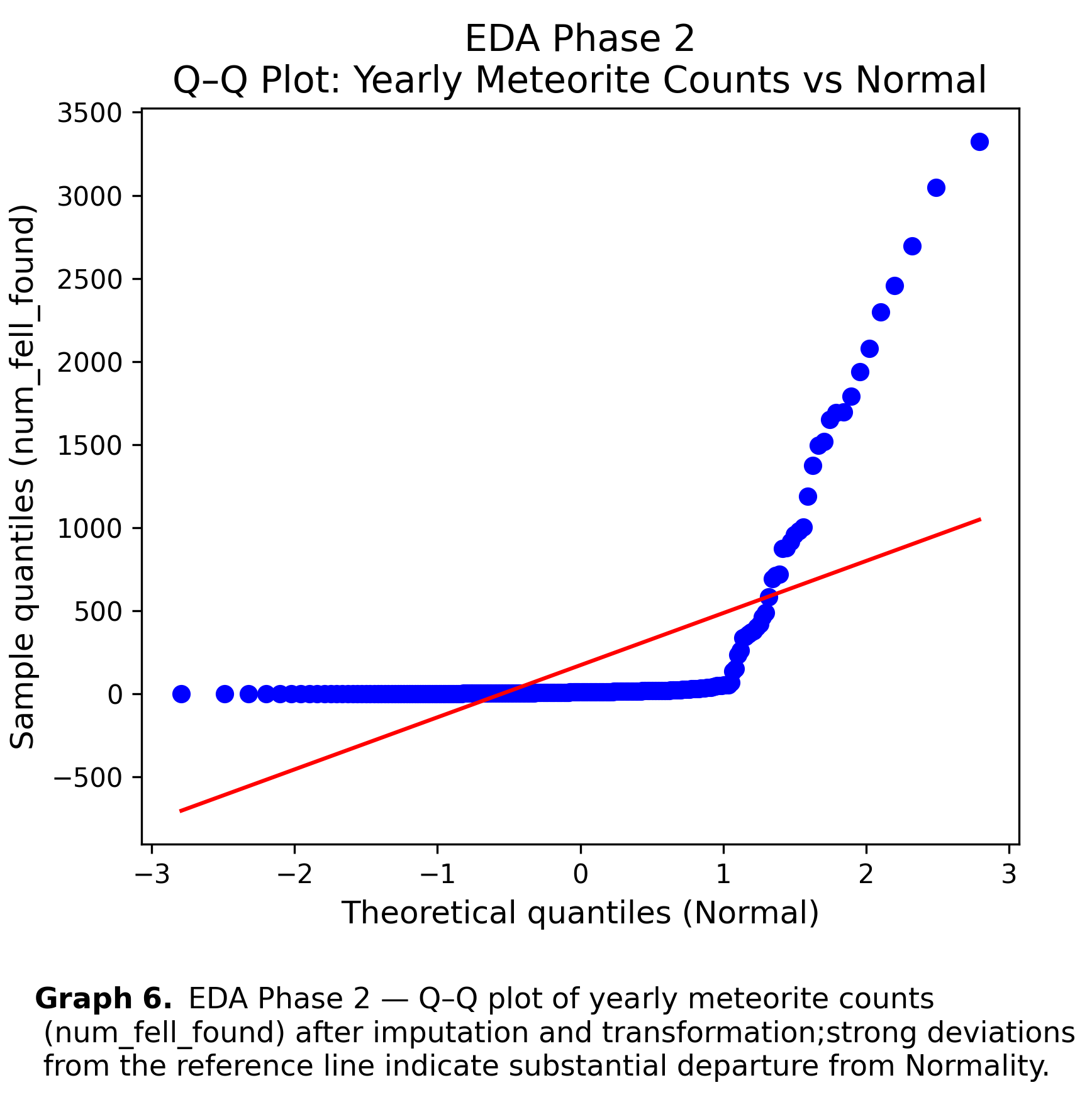
AI-generated content may be incorrect.

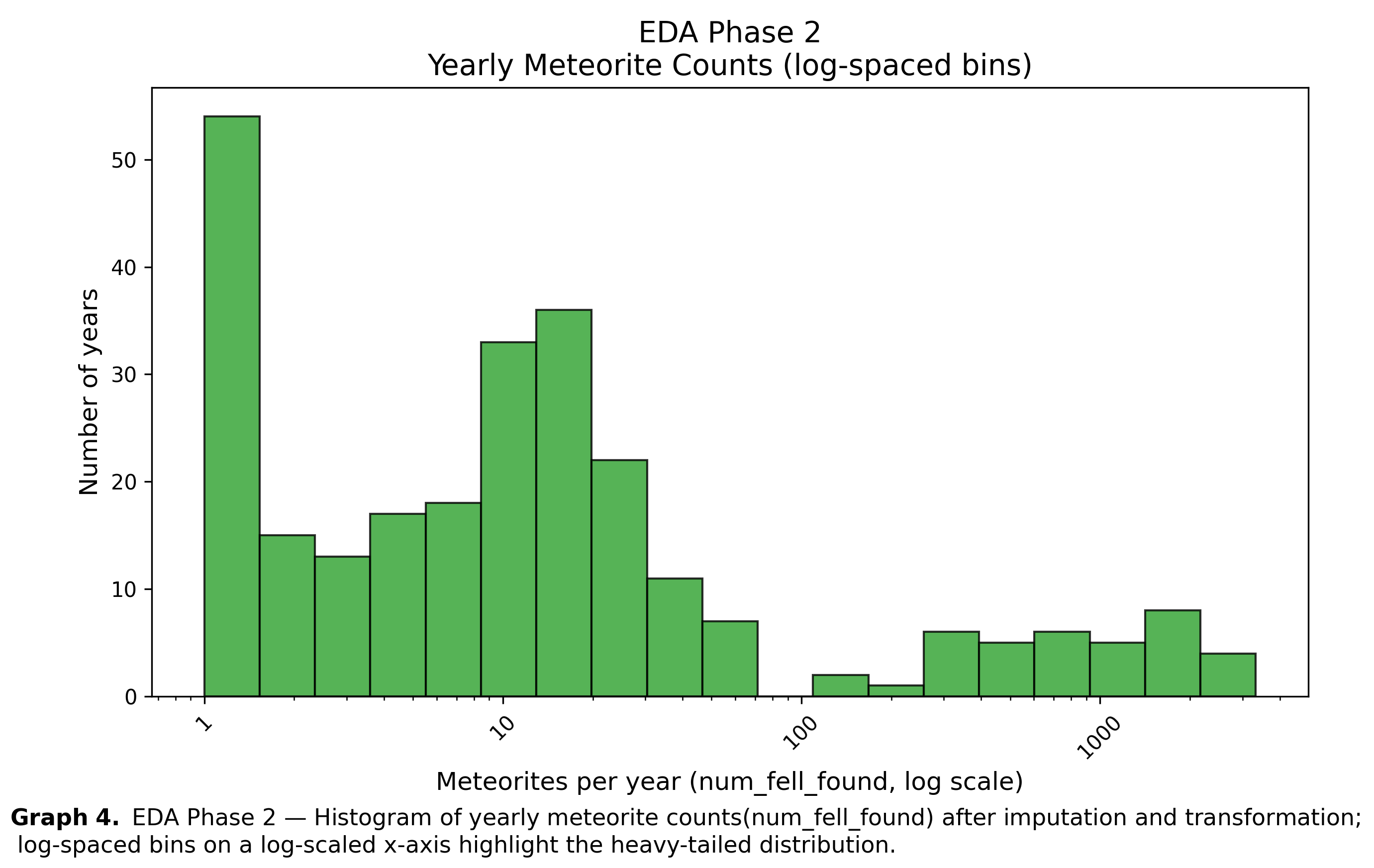
**Graph 2** shows the same data as Graph 1 – with a log(time) independent variable axis. This was done as an exploratory step to see if transforming the data would help with visualization.

**Graph 3** shows the frequency data for the raw *Meteorite Landings* dataset. Bins represent number of meteorites found in a year, and ‘Number of years’ represents the count for that bin. The independent variable (bin axis) has been graphed using a log scale for better data presentation.

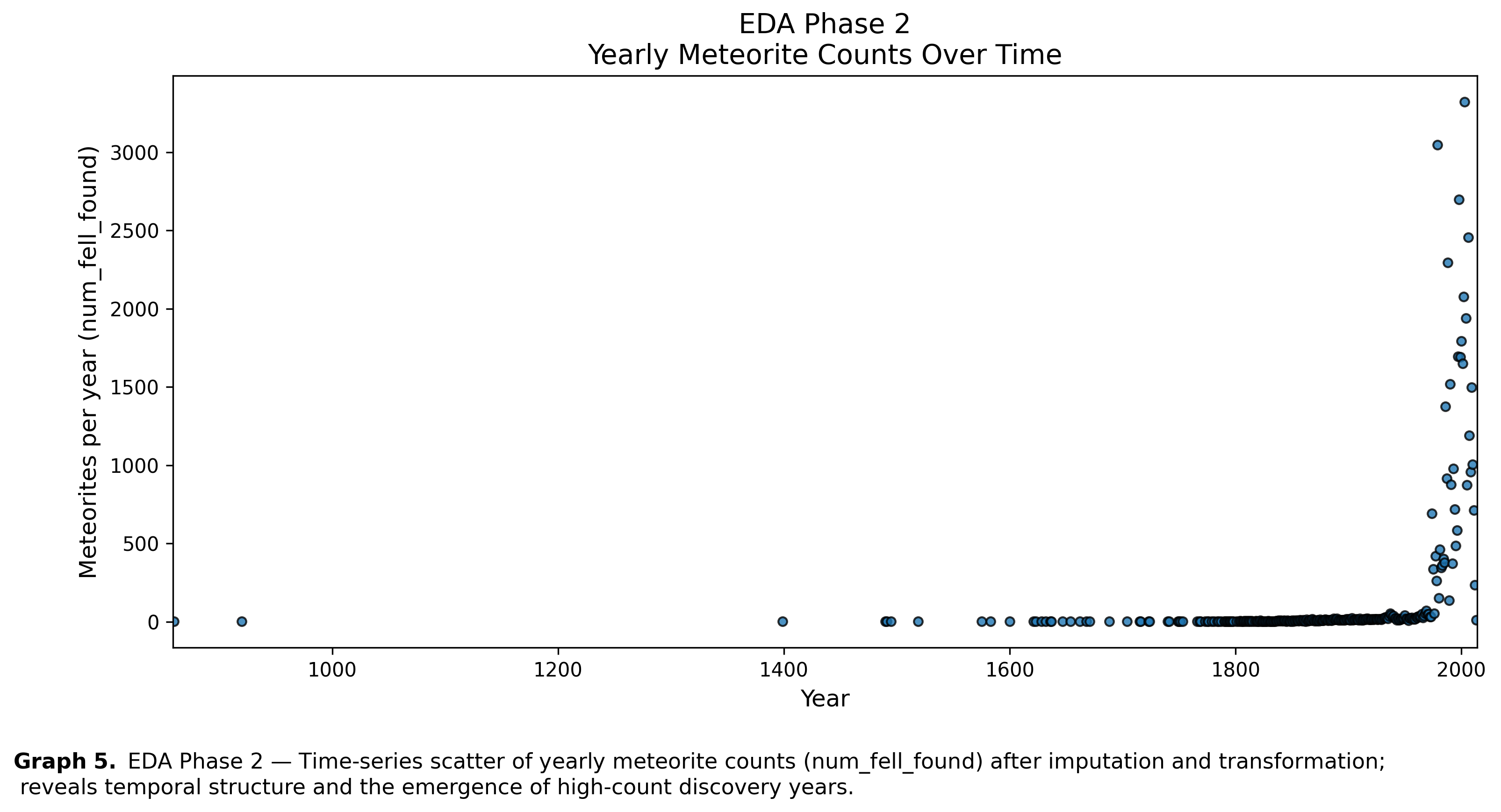
**Appendix A — EDA — Phase II**  
Phase II begins the process of imputation, normality exploration, and data transformation.

**Table 2** shows the reduction of the original raw data set by collapsing all meteorites found/fell into the same year category. Further statistics are shown for the data, including the number of outliers.



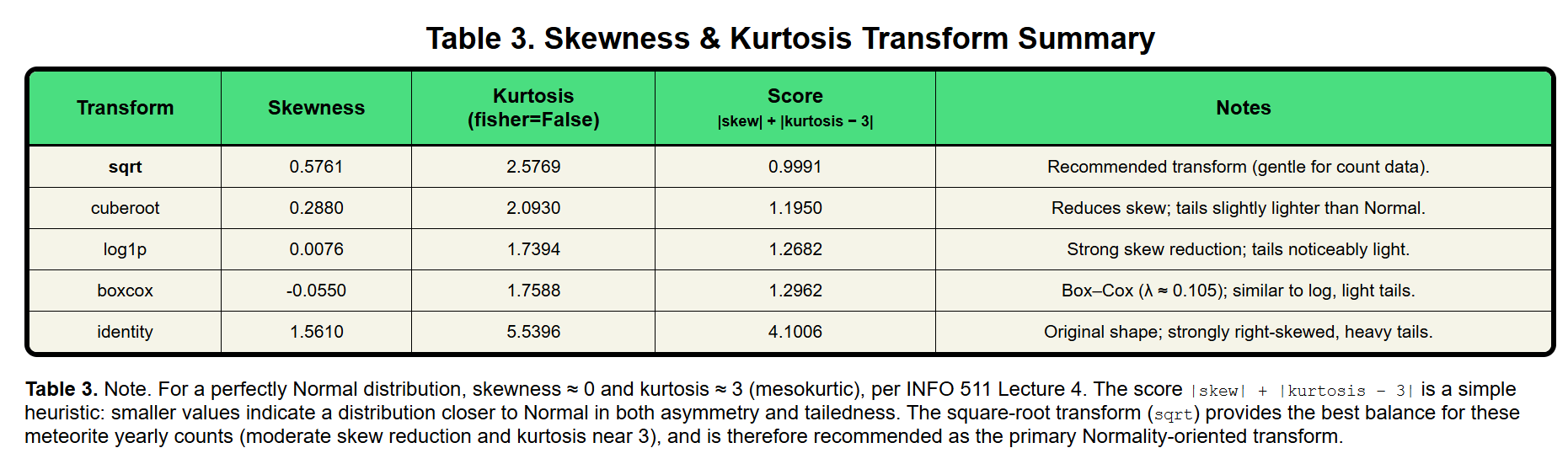


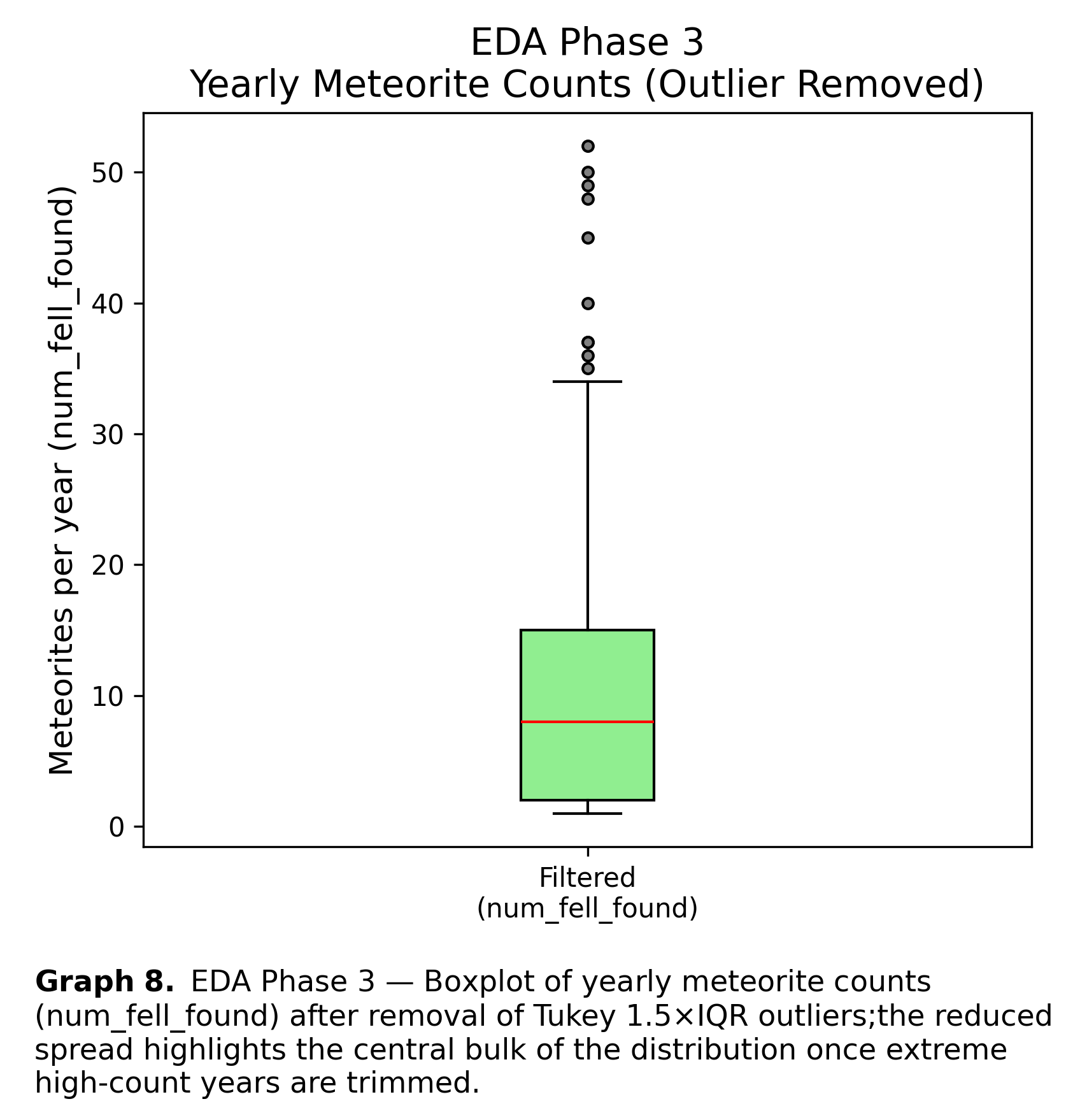
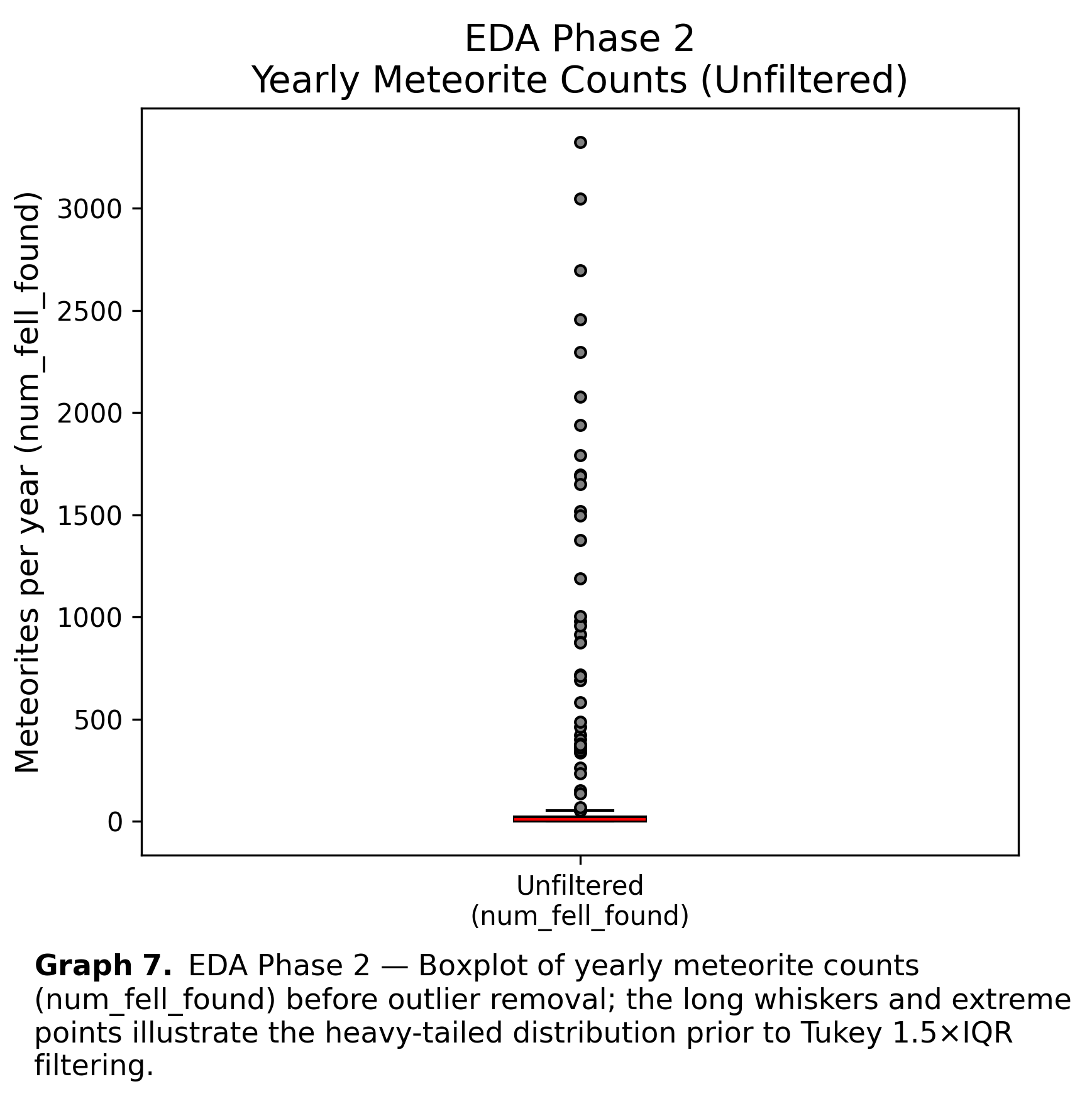
**Graph 4** examines the transformed data distribution from Table 2, while **Graph 6** shows a QQ-plot of the histogram data. The data is both multi-modal and skewed. Interestingly each data grouping in the histogram shows its own kurtosis and skewness.



**Graph 5** is a scatter plot of   
 the transformed data from   
 **Table 2**.

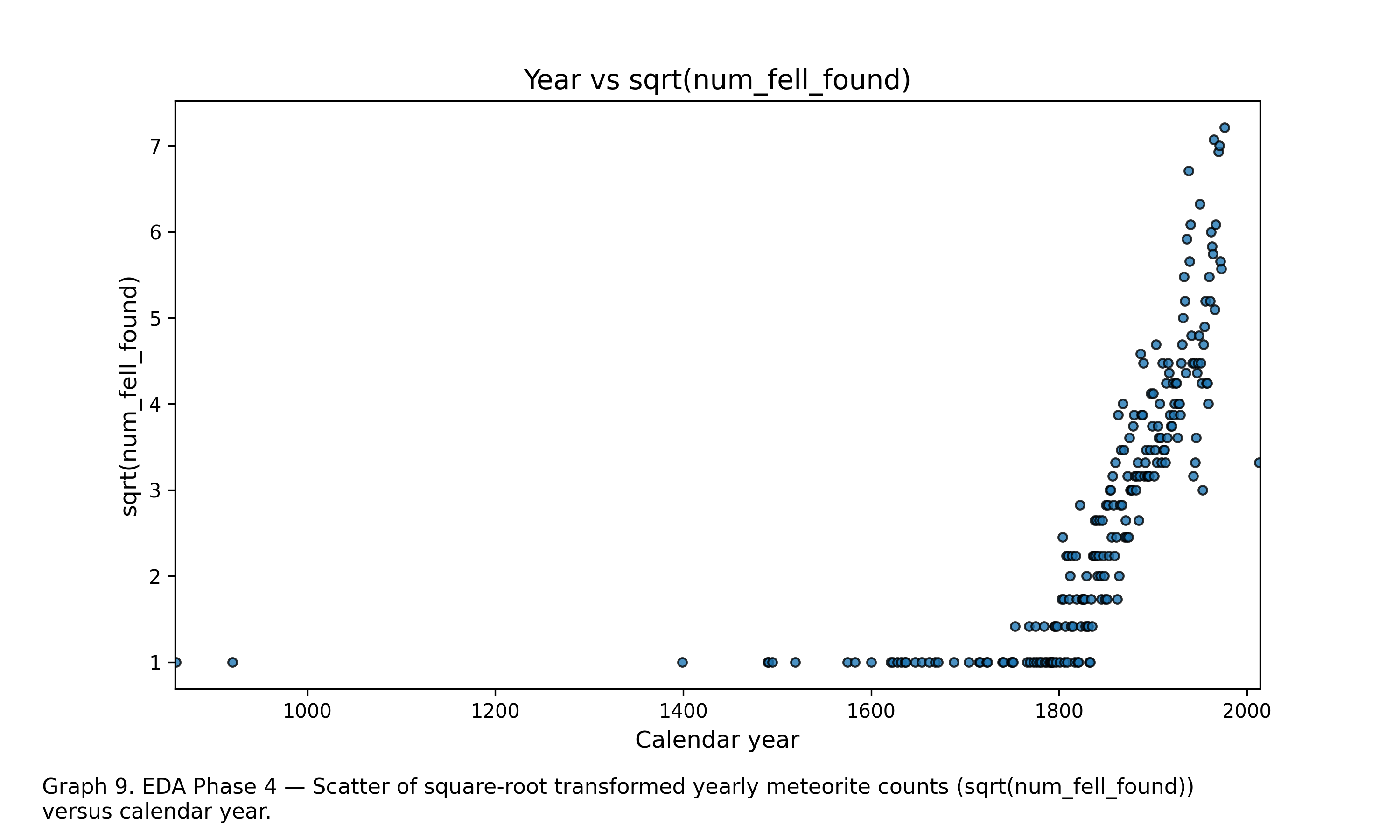
**Appendix A — EDA — PHASE III**   
Phase III examines data set topology and possess the question: Will a transform help? And which one would be best?

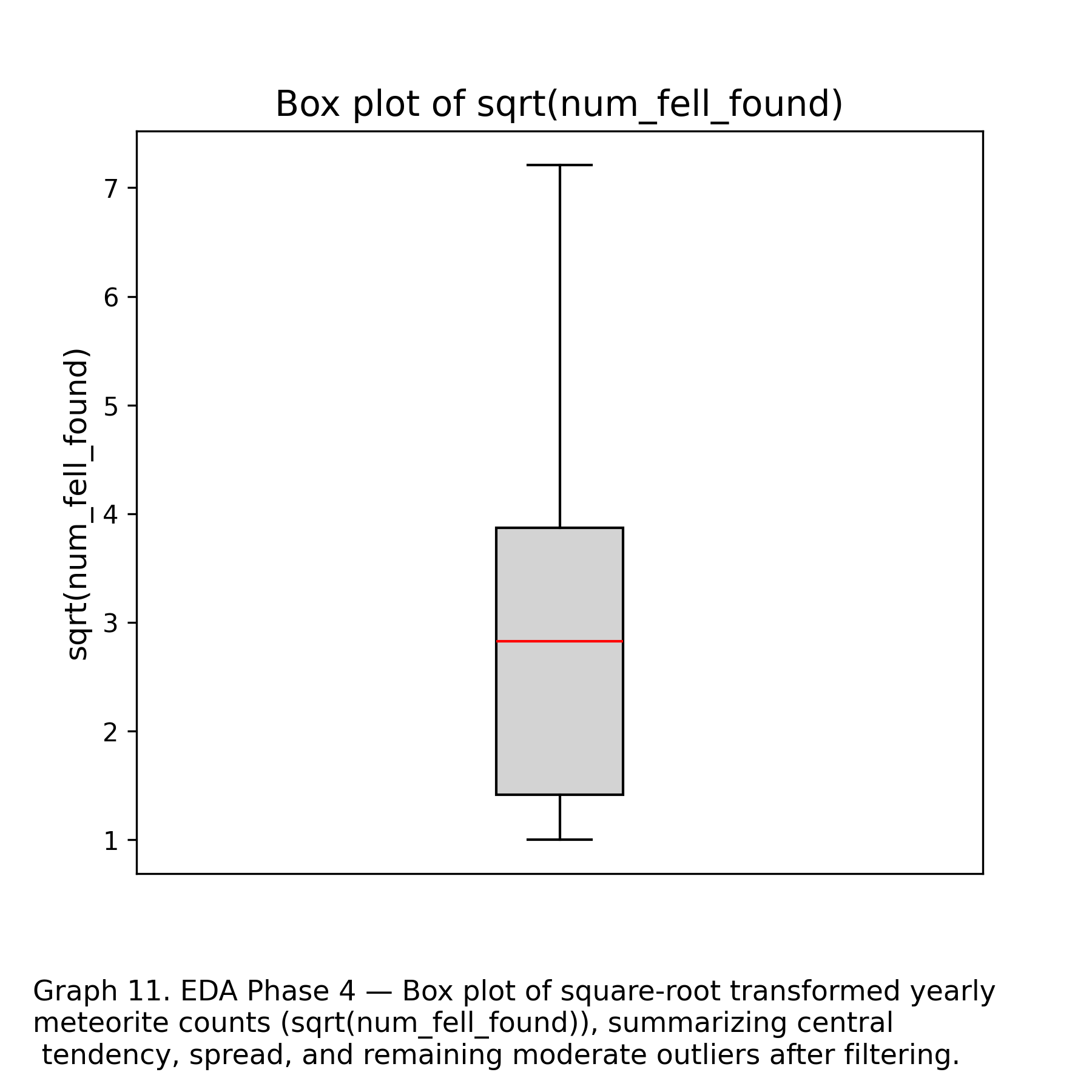
**Table 3** shows the result of a python script (EDA\_phase3\_Normality\_Check.py) - the analysis explores which types of transforms may be beneficial to the data in order to transform the data to a normal distribution. The program reported that a ‘sqrt’ transformation will be the best option in an attempt to transform the data in order to meet normality assumptions.

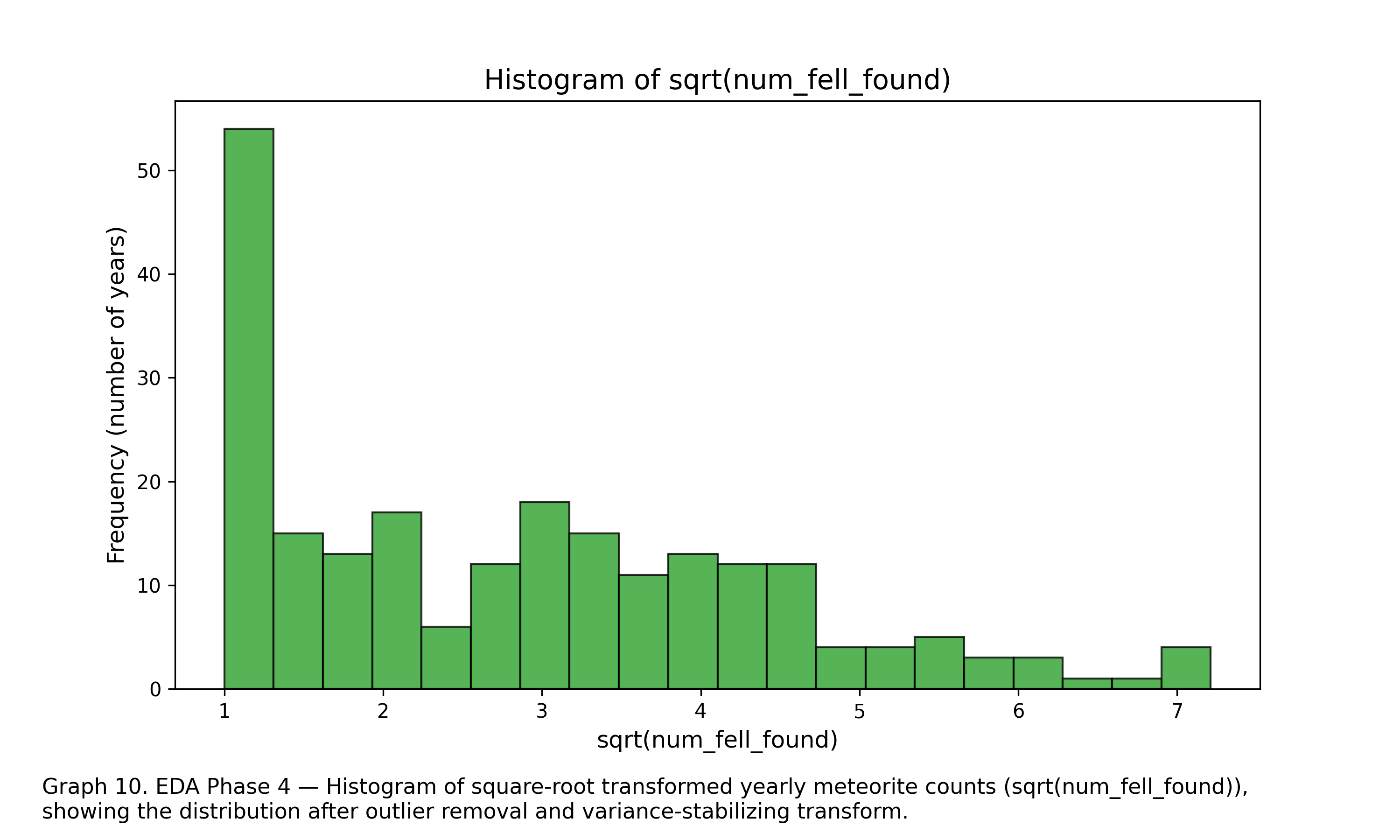


**Graph 7** shows the original raw data set (no filter, no transformation) as a Box plot, while **Graph 8** (IQR outlier removal)shows the dataset after outliers have been removed.

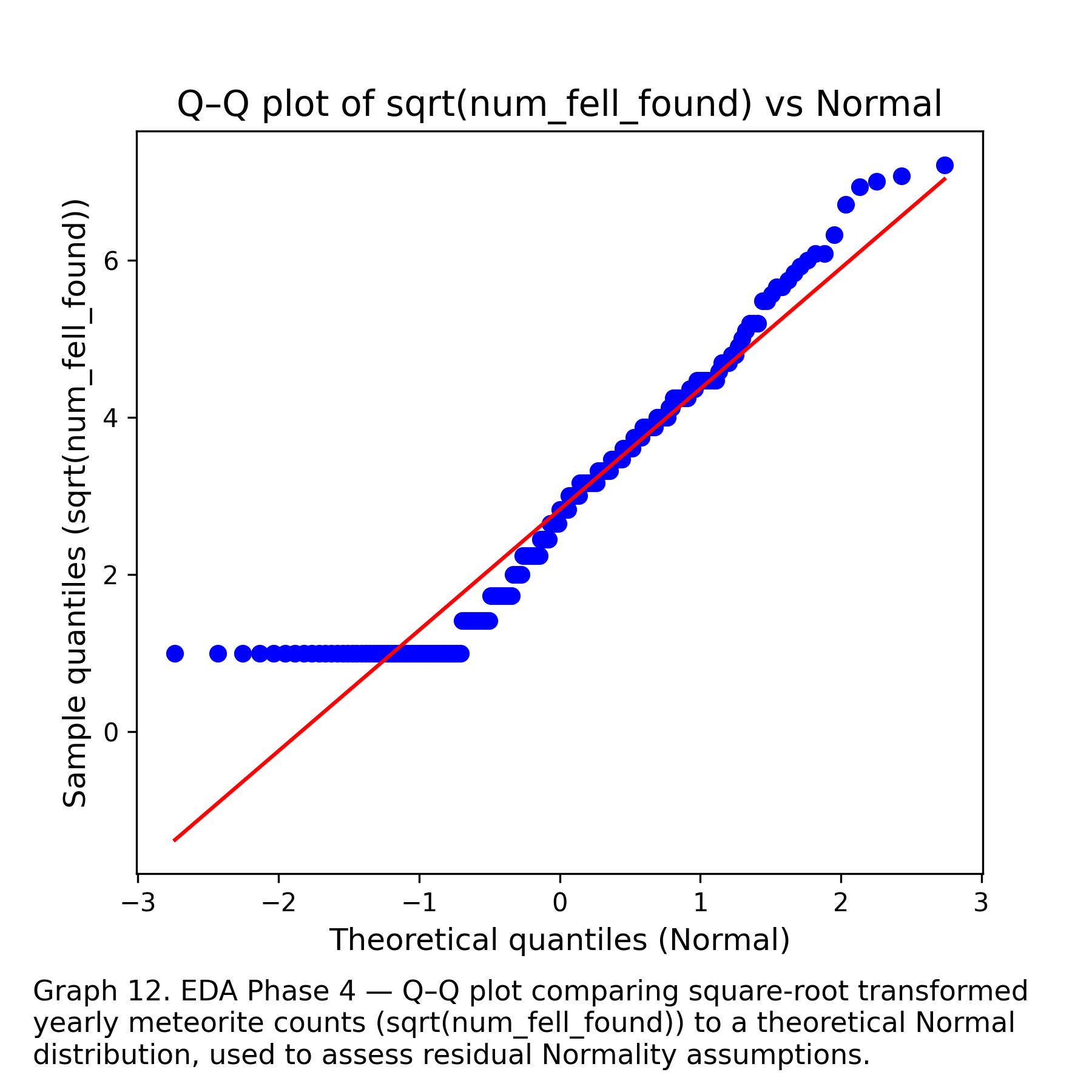
**Appendix A — EDA — PHASE IV**







Appendix B — Model Diagnostics  
- Residual Plot



Appendix D — Extended Tables  
- Full summary statistics  
- Correlation matrix  
- Categorical variable level counts

A black and white picture of a person's profile

AI-generated content may be incorrect.**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.

A black and white logo with a robot and check mark

AI-generated content may be incorrect.**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.

A close-up of a sign

AI-generated content may be incorrect.**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**](http://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.