Wikipedia Science Articles:

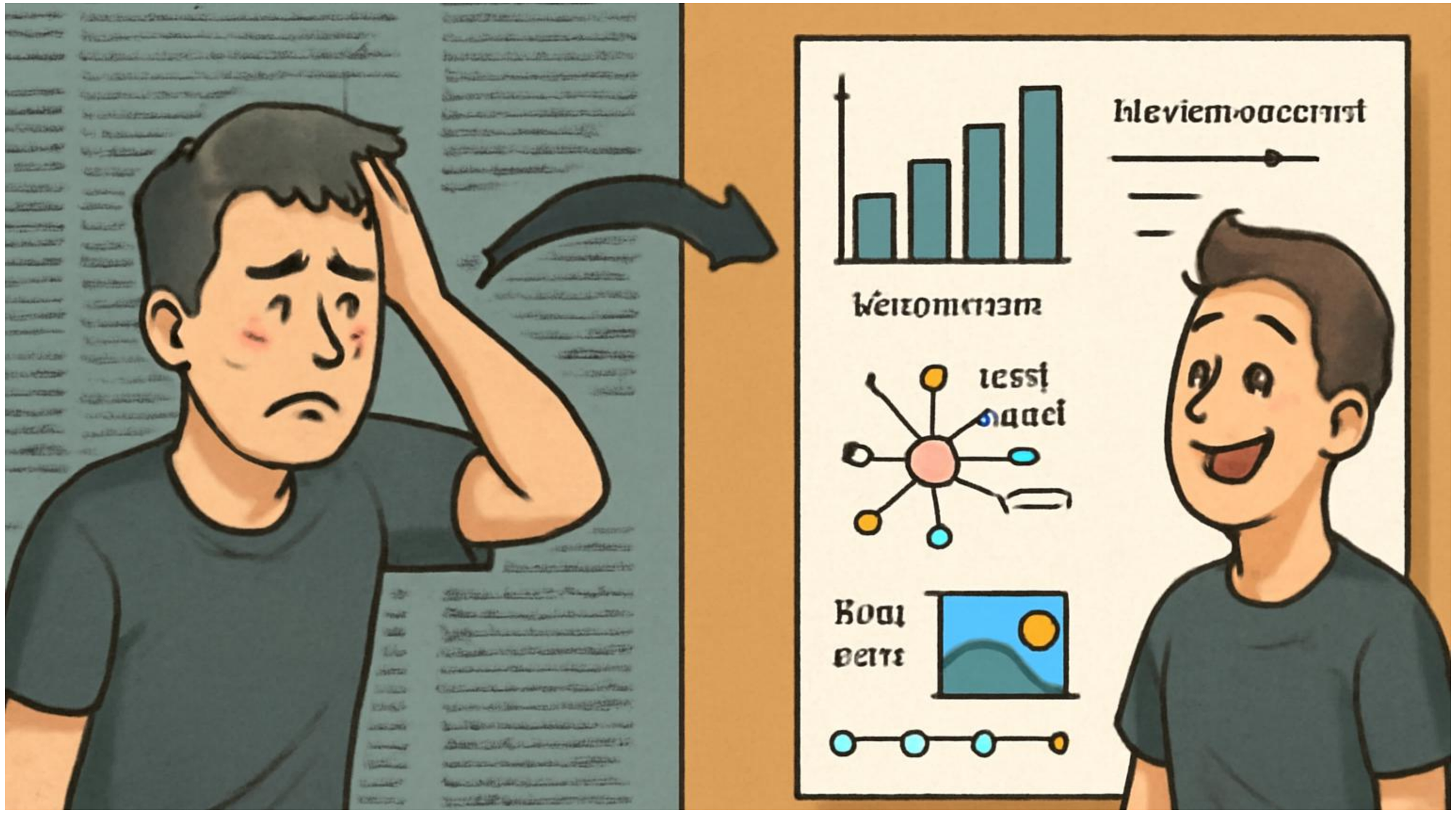
A Comprehensive Data Analysis

Ngo Sy Trung and Tran Trung Duc

03 June 2025

Question need to be answered

- How do quantitative characteristics (such as word count, reference count, link density, etc.) vary among articles within the 'Science' category on Wikipedia ?
- Are there patterns related to article scope or topic revealed through their assigned categories?



Dataset overview

- Crawl from the "Category: Science" on English Wikipedia
- 1000 articles about broad science
- Header include:

Title	Summary	Categories	References	Links	Last Edited	...

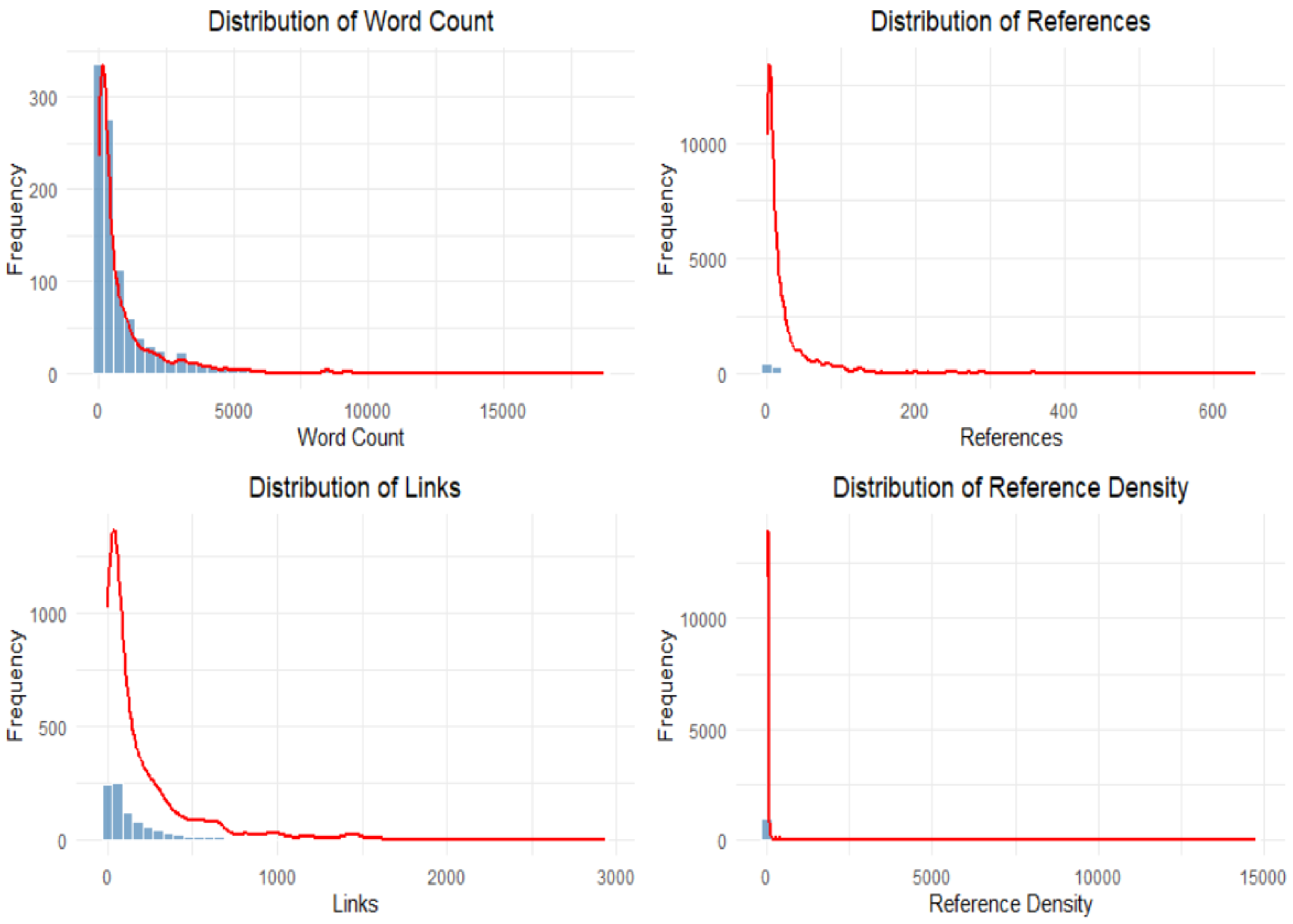
Lists of unsolved problems	The following is a list of unidentified, or formerly unidentified, sounds. All of the NOAA sound files ...	Unidentified sounds;
		Science-related lists;
		Sound-related lists
			14	83	14-05-2025	

History of scientific method	The history of scientific method considers changes in the methodology of scientific inquiry, as distinct from the history of science itself ...	History of scientific method; Scientific method; History of science	120	651	01-06-2025	

Distribution Histograms

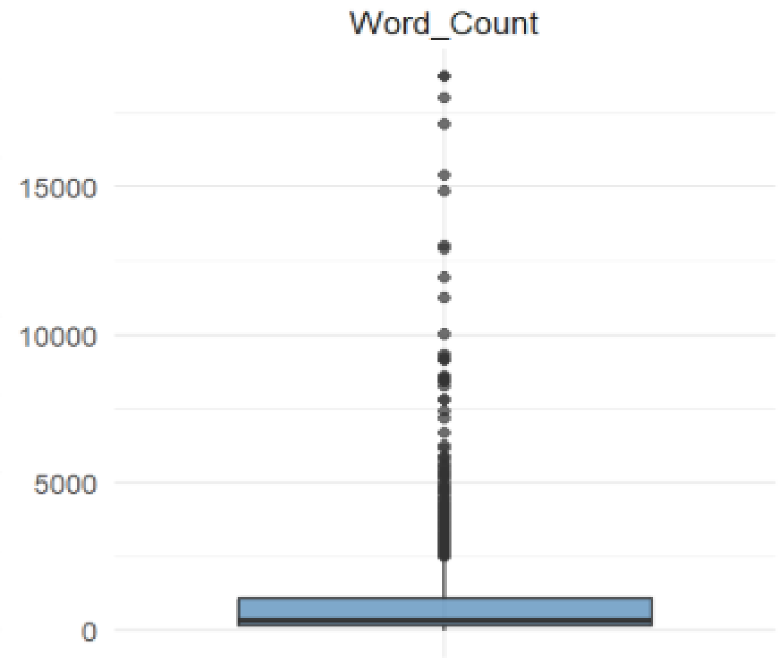
Wide ranges and right-skewed distributions.

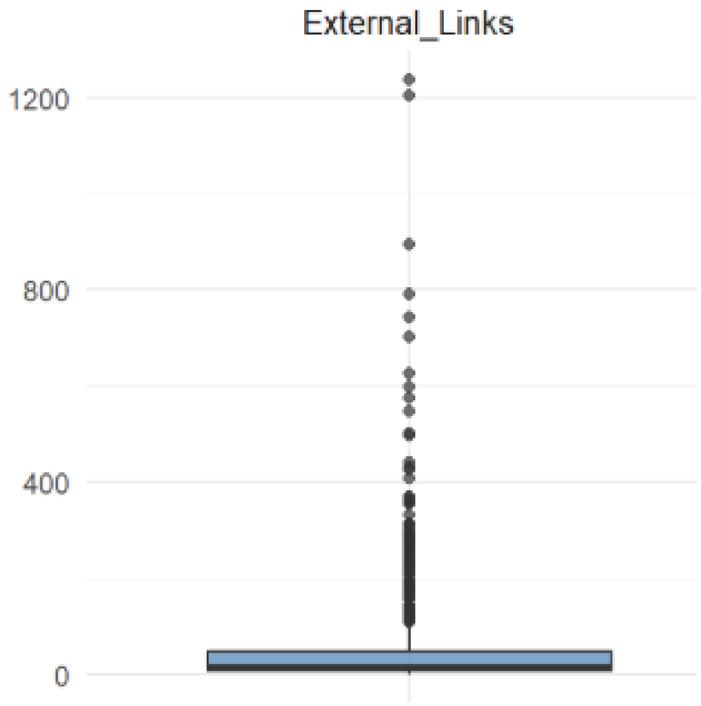
Most articles are relatively short or lightly referenced, but a significant tail includes highly detailed and well-sourced entries.



Box Plots & Outlier Identification

- Box plots clearly show the median, quartiles, and outliers for each variable.
- Variables like Word_Count, External_Links exhibit many outliers on the higher end.
- "Word_Count (121 outliers) show many extreme values."



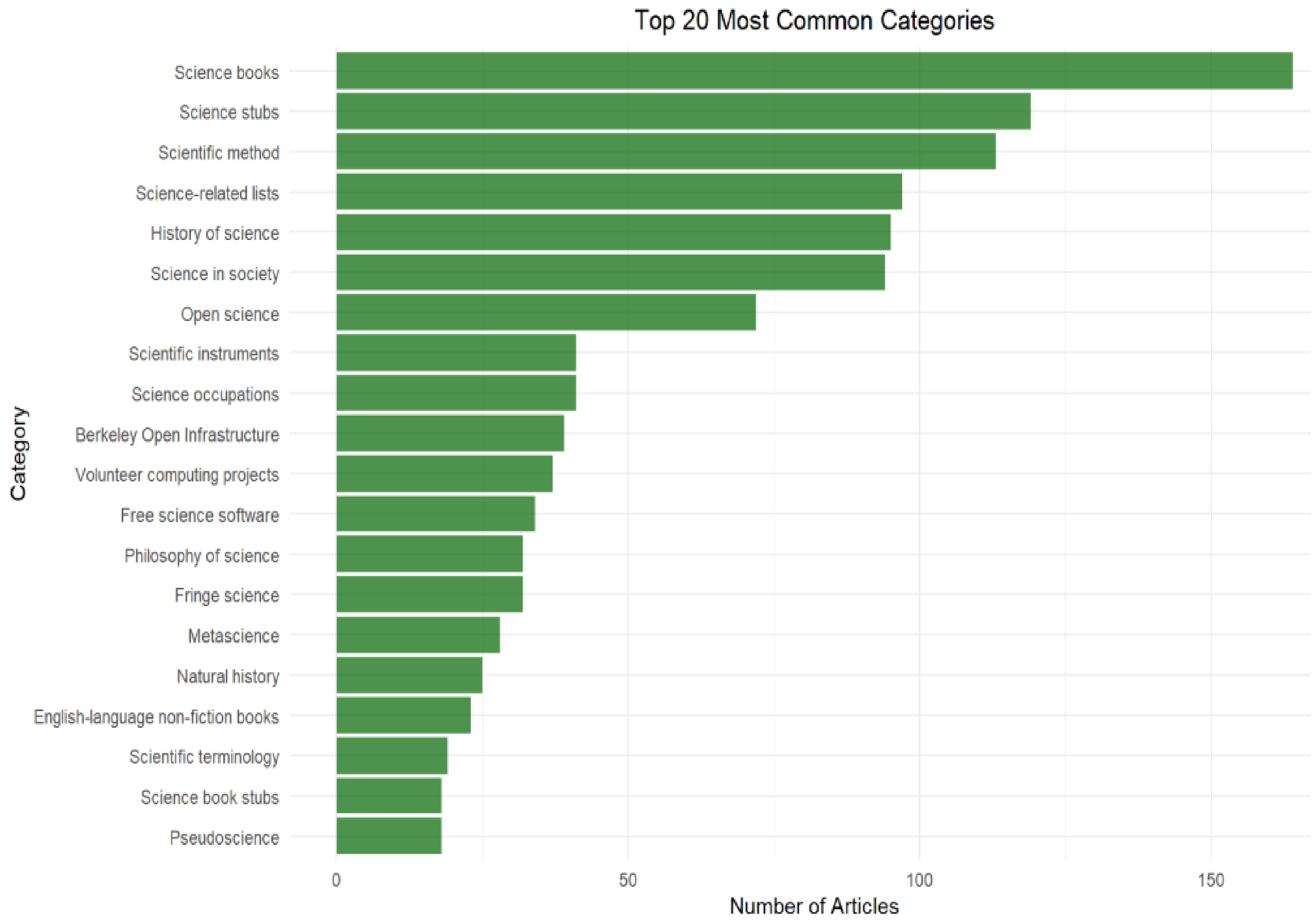


What Topics Dominate Science on Wikipedia?

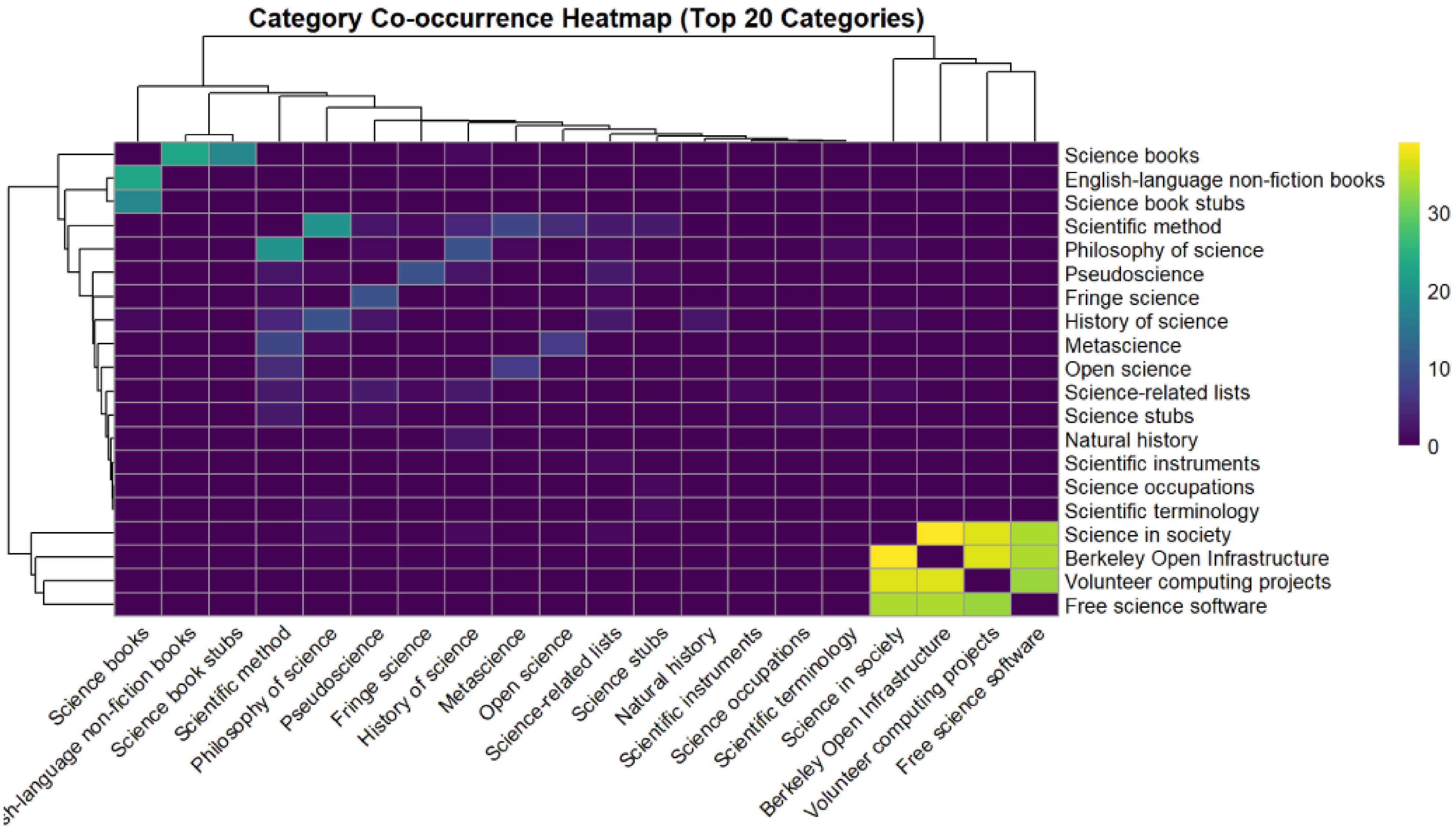
Key Observations:

- "Science books" is the most frequent single category.
- "Science stubs" and "Scientific method" are also prominent.
- Categories like "History of science," "Open science," and specific fields appear.

Insight: Provides a clear view of the most represented subject areas and article types within the science domain.



Category Deep Dive: Co-occurrence


Category Co-occurrence Heatmap (Top 20 Categories)

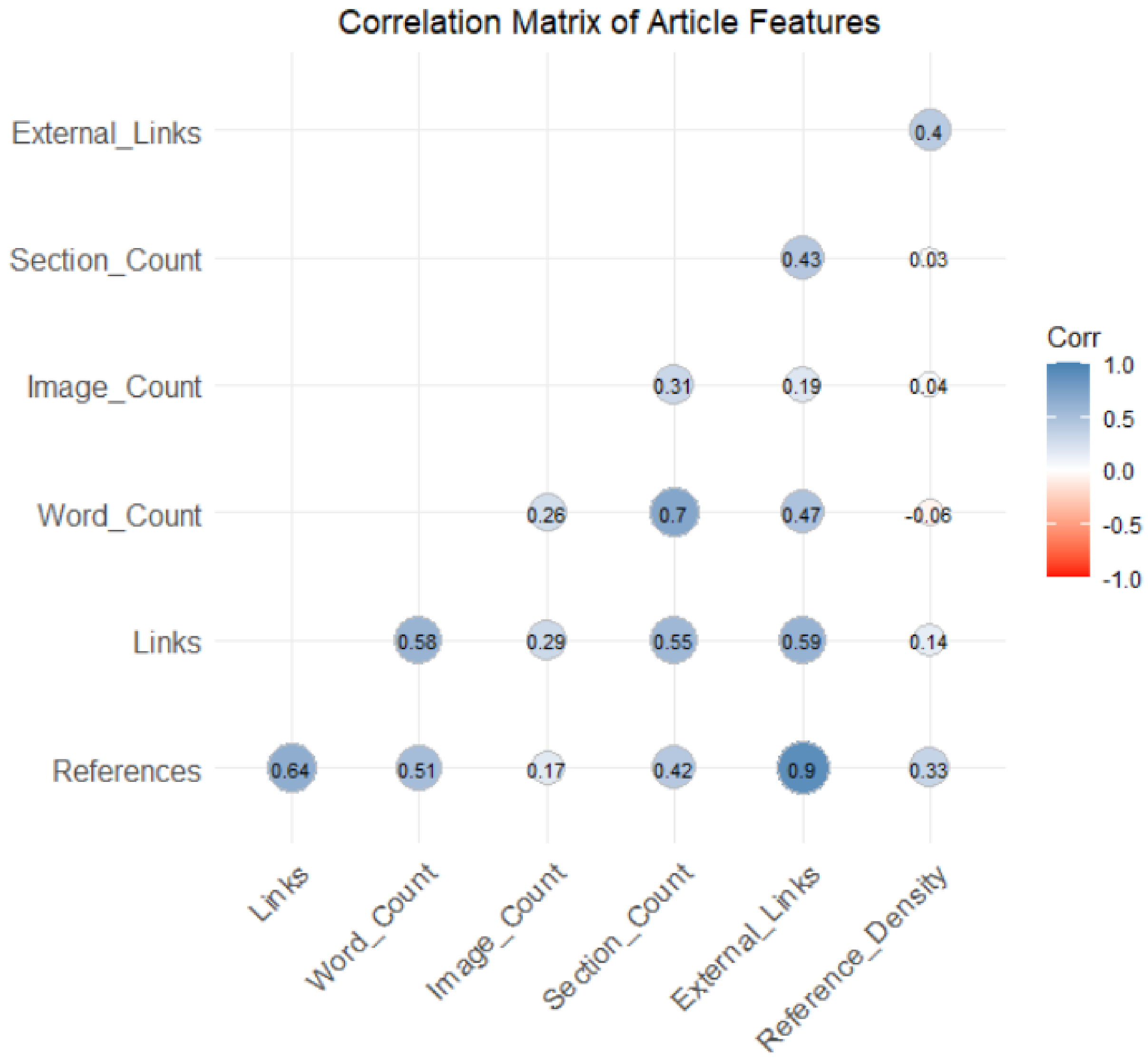
Correlation Matrix

Strong Positive:

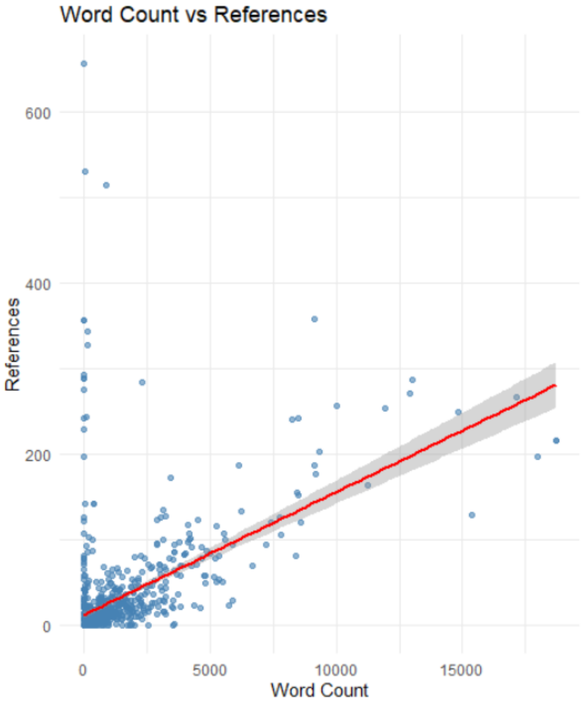
- References are likely External_Links (0.9)
- Word Count and Section_Count (0.7) – Longer articles strongly tend to have more section.

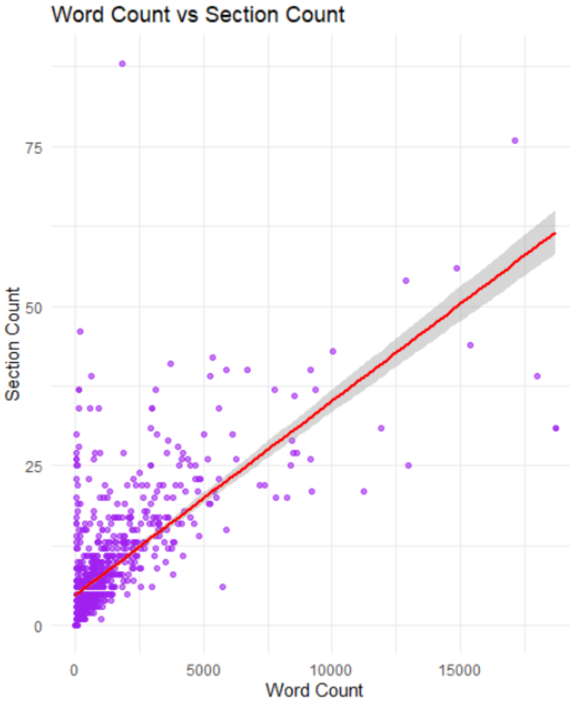
Moderate Positive:

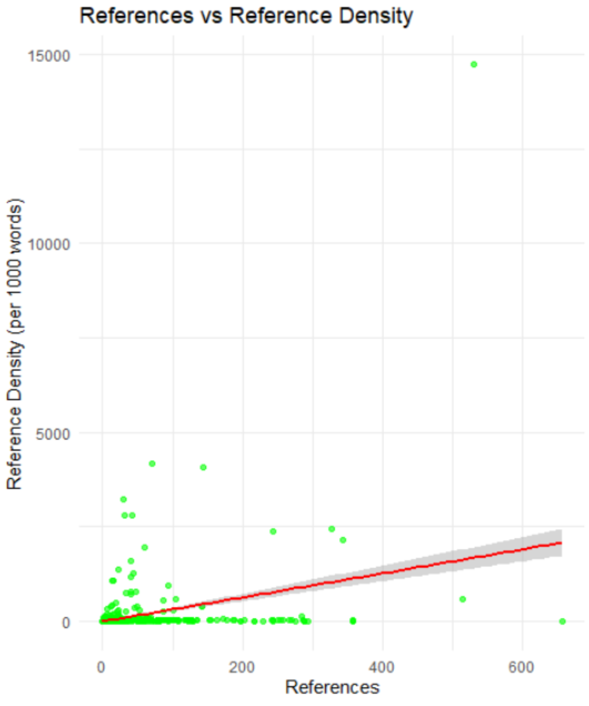
- Links are likely External_Links (0.59)
- References are likely Links (0.64)

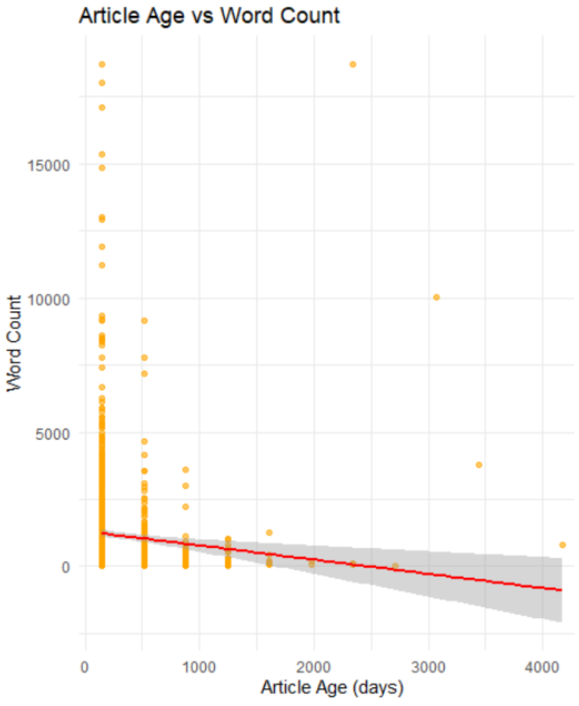


Word Count vs. References


Word Count vs References


Word Count vs Section Count

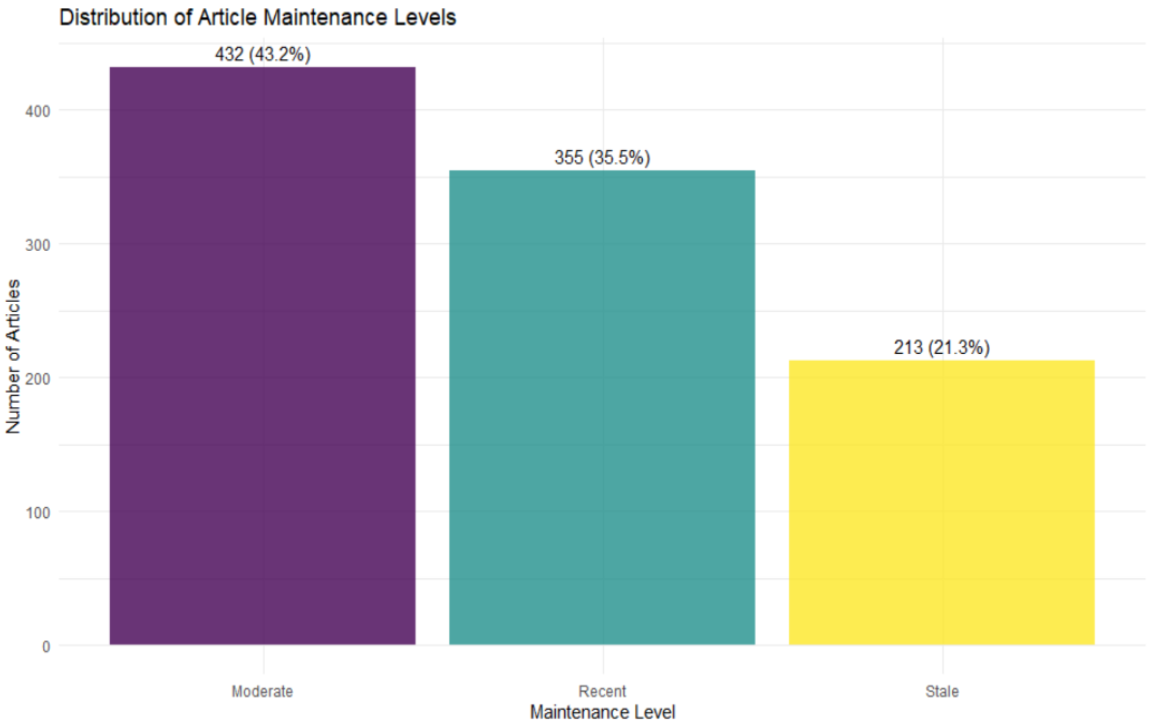

References vs Reference Density

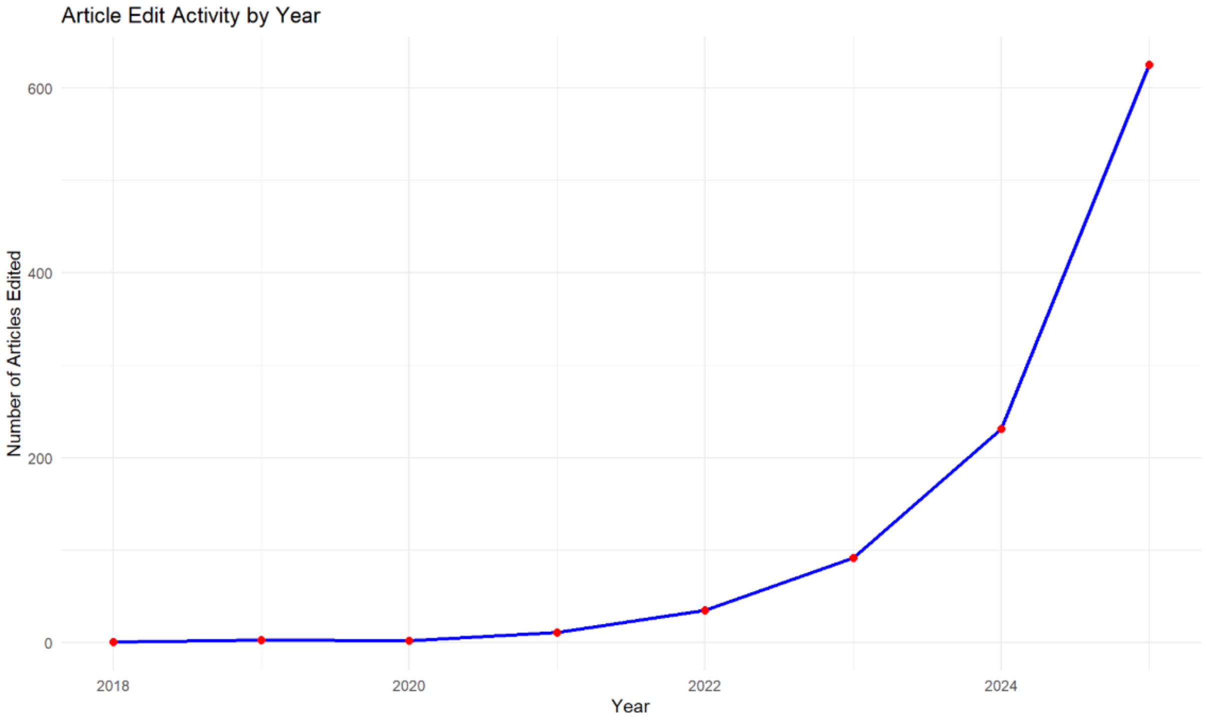

Article Age vs Word Count

Strong positive between number of references, Section count and reference density and number of words.

Long articles tend to be written more recently.

Maintenance Level Distribution

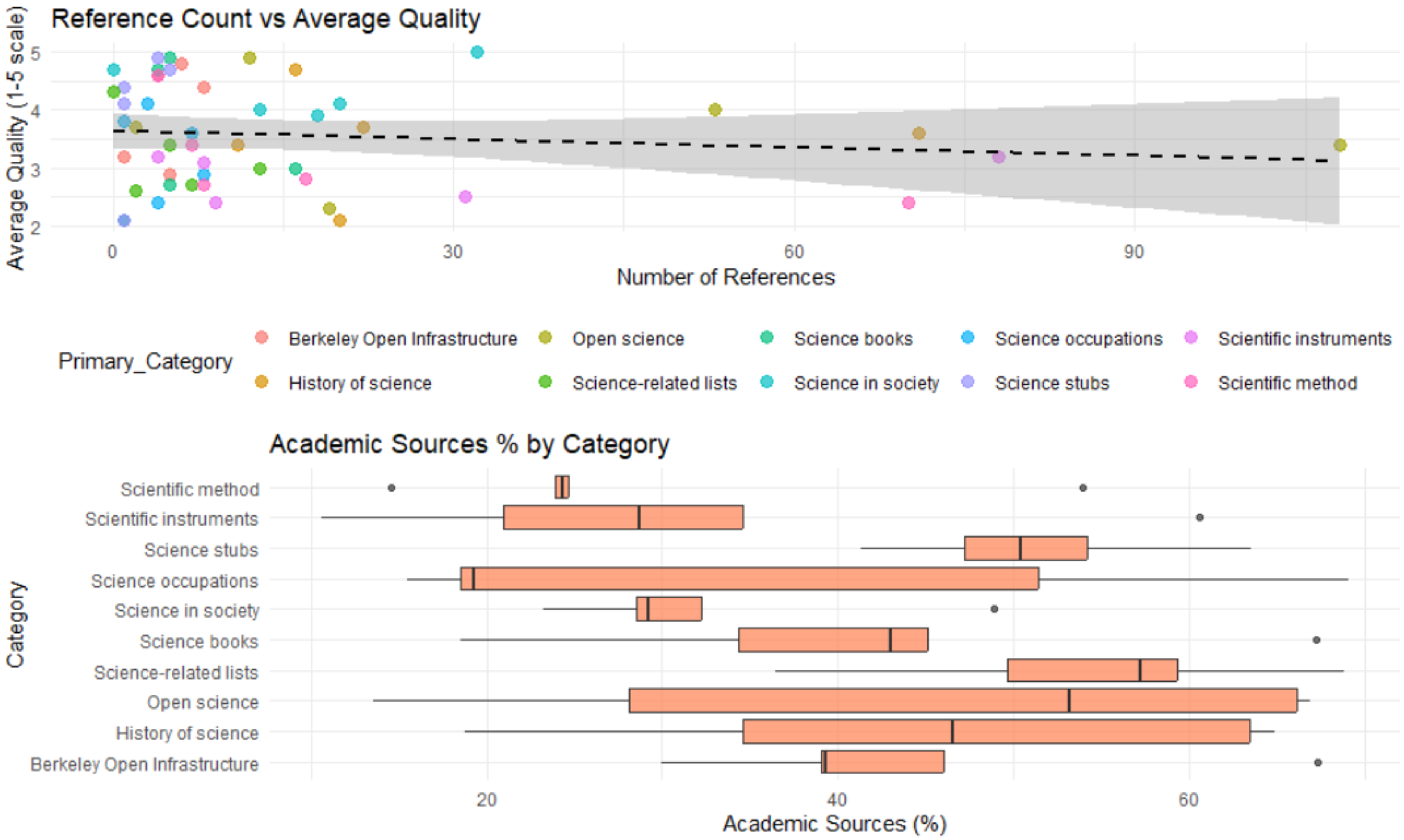




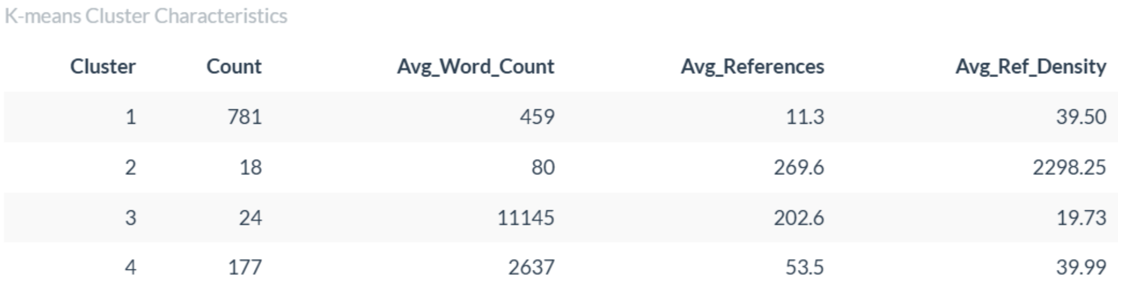
About 20% of articles haven't been edited in over a year, potentially indicating outdated information.

A majority are edited at least annually

Quality of articles

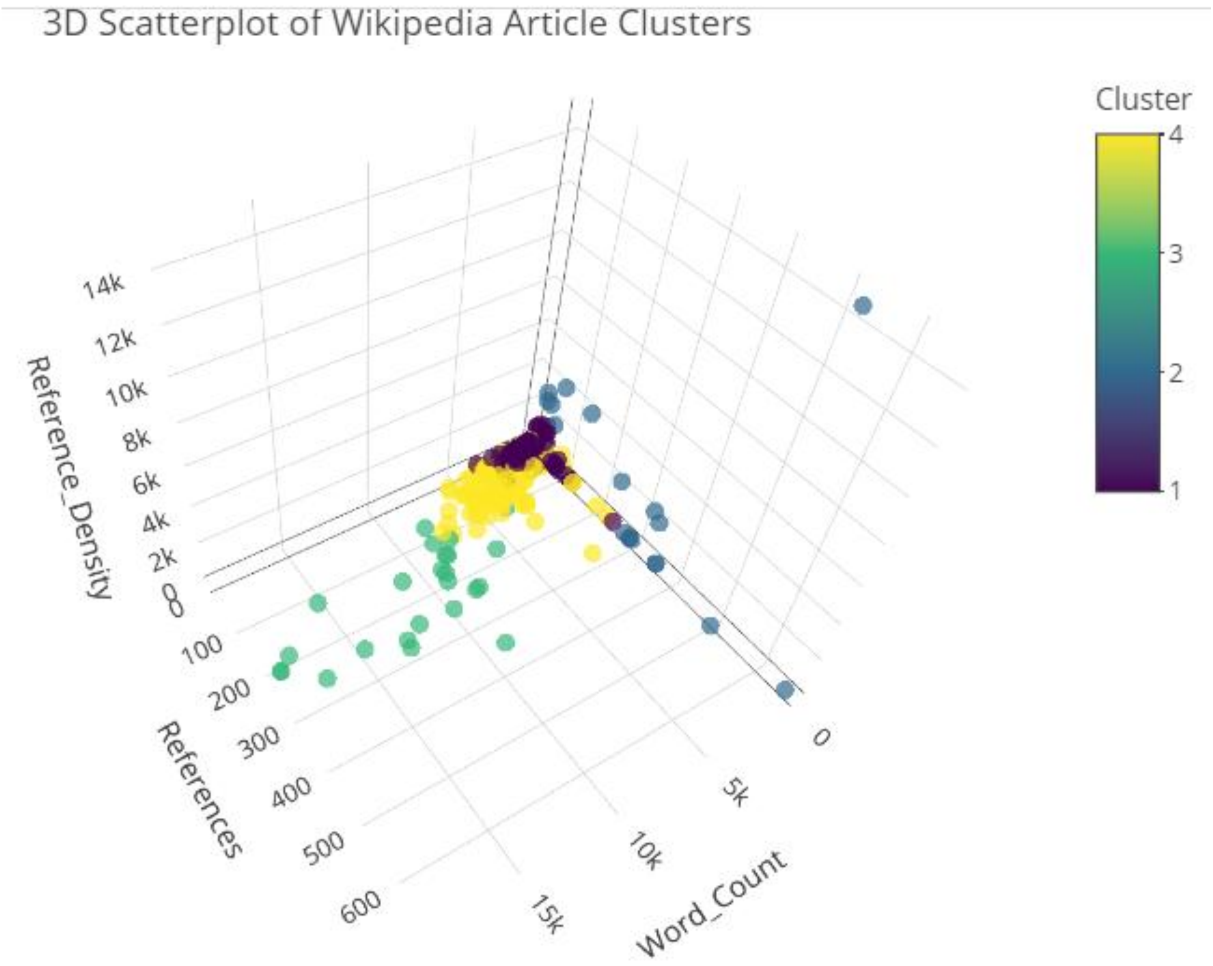


Clustering Visualization



K-means Cluster Characteristics

Cluster	Count	Avg_Word_Count	Avg_References	Avg_Ref_Density
1	781	459	11.3	39.50
2	18	80	269.6	2298.25
3	24	11145	202.6	19.73
4	177	2637	53.5	39.99



3D Scatterplot of Wikipedia Article Clusters

- Four clusters
  - Medium word count, low No. Refs
  - Low word count, great No. Refs
  - Very high word count, medium high No. Refs
  - Medium high word count, medium No. Refs

Key take-away

Finding	Value
Average Word Count	1094 words
Average References	28 references
Most Common Category	Science books
Strongest Correlation	Word Count & References
Articles Needing Maintenance	213 articles
Reference Quality (Pilot)	3.6 / 5.0
Identified Clusters	4 distinct article types

Summary

- Conclusion
  - Our analysis has provided a multifaceted view of Wikipedia science articles, highlighting their characteristics, relationships, and engagement patterns.
  - Data visualization and statistical analysis can uncover significant insights into large-scale collaborative knowledge bases like Wikipedia, identifying both strengths and areas for potential improvement.

- Limitations and Future work:
  - Dataset Scope: 1000 articles; results may not generalize to all Wikipedia science content.
  - Feature Set: other features (Readability scores, editor statistics, etc.) could provide deeper insights.

Q
&
A