Data Analytics Case Study

Assignment by – Bhushan Prabhakar Dhawas

My research paper – "**Formulation and Evaluation of Antiseptic Herbal Gel"** (link).

Power Bi report – (link)

Objective

To analyse the data and find insight/patterns from the sample data which could be of interest to a researcher, publisher, university (affiliation) or general public. Your analysis and conclusions could be submitted as a word document(.docx) or python notebook(. ipynb) with relevant chart/graphs that helped you uncover an insight

Data Overview

Below is an overview of the dataset, including the key columns and their descriptions:

Key Columns and Descriptions:

1) **author_countries:**

Description: Contains a list of countries associated with the authors of each research article.

Use Case: Used to analyze collaborations between countries.

2) **affiliation_normalised:**

Description: Provides a distinct list of normalized affiliation names for the authors.

Use Case: Helps identify institutional affiliations and their contributions.

3) **clean_journal_name:**

Description: The cleaned and standardized name of the journal where the article was published.

Use Case: Used for publication trend analysis by journal.

4) **clean_publisher_name:**

Description: The cleaned and standardized name of the publisher that owns the journal.

Use Case: Helps analyze publication trends by publisher

5) **dates_pub:**

Description: Represents the earliest publication date (either online or in print) of the article.

Use Case: Used for year-over-year trend analysis.

6) **citations_research_article_count:**

Description: The total number of citations received by the research article.

Use Case: Used to evaluate the impact of articles and journals. Essential for understanding the research article's influence and recognition.

7) **references_count:**

Description: Represents the number of references cited in the research article.

Use Case: Helps analyze the depth of research and cross-referencing practices.

8) **total_authors:**

Description: The total number of authors who contributed to the article.

Use Case: Useful for identifying large collaborative research projects.

9) **Article_title:**

Description: The title of the research article.

Use Case: Useful for identifying and referencing specific articles.

10) **journal_oa_status:**

Description: Indicates whether the journal operates under Open Access (OA).

Use Case: Used to analyze trends in open-access publishing.

11) **doi:**

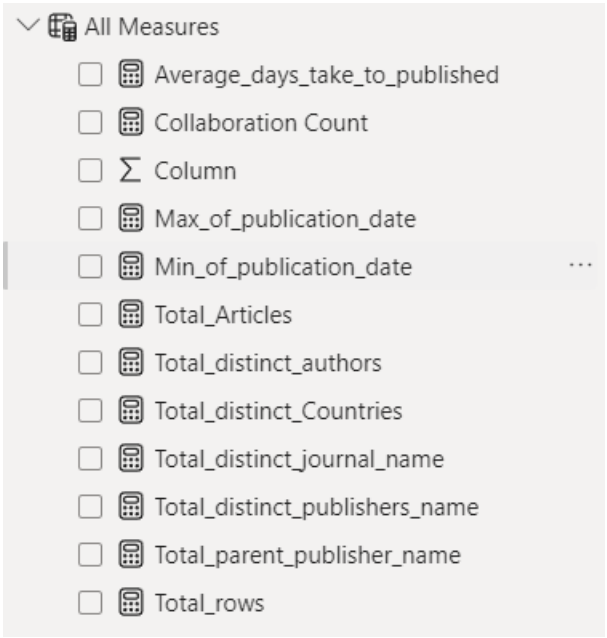
Description: A unique hyperlink (Digital Object Identifier) for the article.

Use Case: Provides a direct link to the article for verification and further reading.

Added Columns -

- **Days take to publish** = DATEDIFF('Publication Data'[dates_publication_history_received],'Publication Data'[dates_pub],DAY)
- **Month of publication** = `FORMAT('Publication Data'[dates_pub],"MMM")`
- **Month no. of publication** = `MONTH('Publication Data'[dates_pub`
- **Year of publication** = `YEAR('Publication Data'[dates_pub])`

Added Measures –



Basic Information About Data

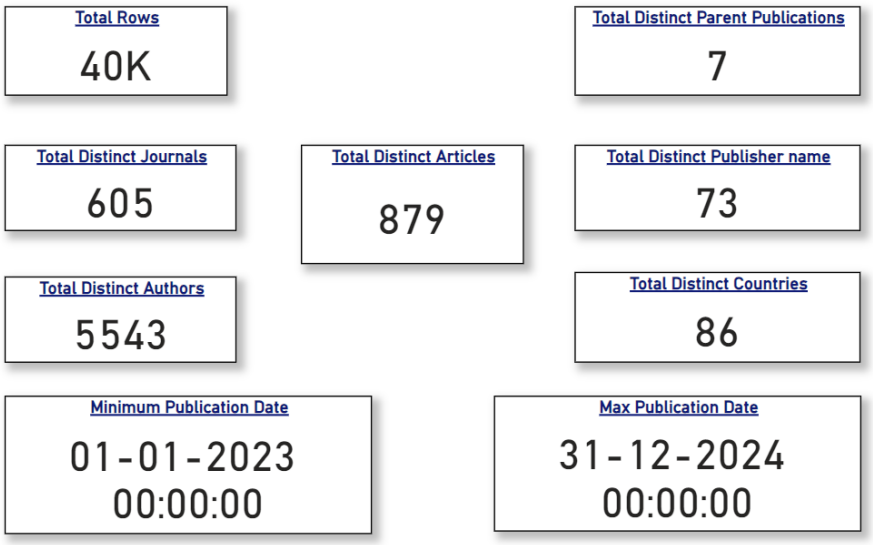
Q1: What does the dashboard reveal about the overall dataset?

Answer:

The dataset contains:

- **40K total rows**, representing detailed records of publications.
- **879 distinct articles**, contributed by **5543 unique authors** from **86 different countries**.
- Publications are spread across **605 journals**, **7 parent publications**, and **73 distinct publishers**.
- The earliest publication date is **01-01-2023**, and the latest is **31-12-2024**. The data shows published articles in the year 2023 and 2024.

This indicates the dataset is comprehensive, covering a wide range of articles and global collaborations over two years.



Year-over-Year Publishing Trends

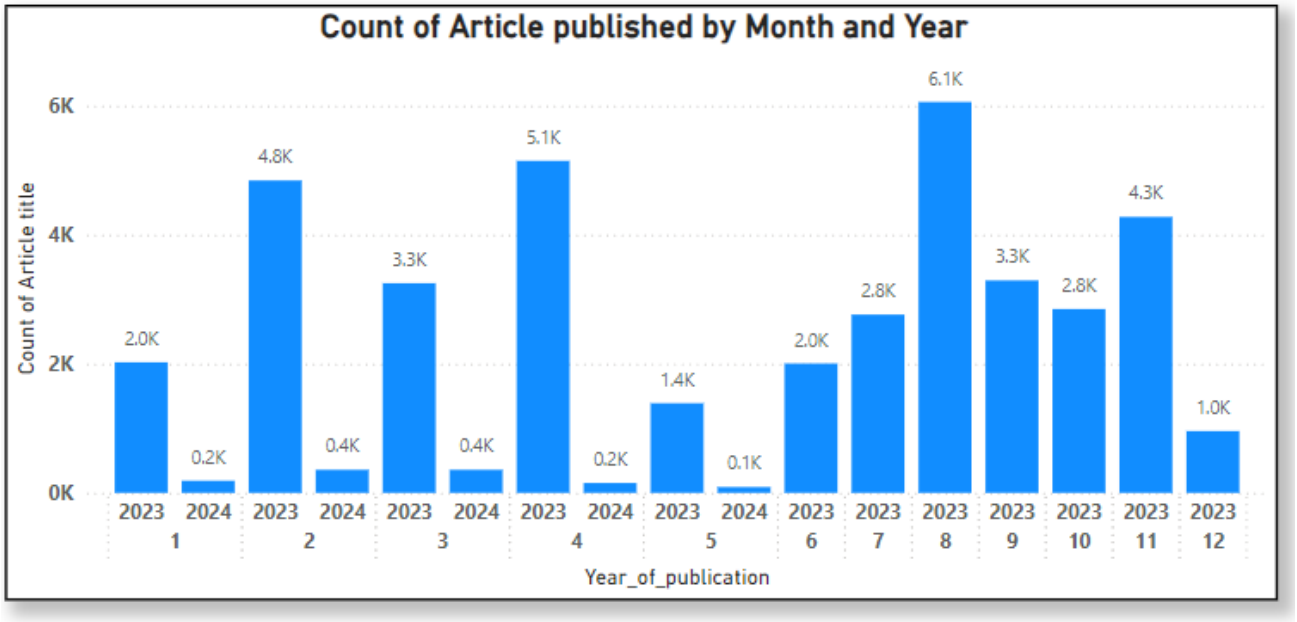
Q2: How are the articles distributed over time?

Answer:

The bar chart titled "Count of Articles Published by Month and Year" shows a clear trend in publication activity:

- Peak publication months include **August 2023 (6.1K articles)** and **April 2024 (5.1K articles)**.
- Significant dips in publication numbers occur in **January 2023 (2K articles)** and **June 2024 (1.4K articles)**.

This suggests seasonal variations in publication patterns, possibly linked to academic schedules or conferences.



Articles Published By Countries

Q3: Which countries contribute the most to article publications?

Answer:

The "Top 10 Countries with Most Articles Published" table highlights:

- **China** as the top contributor with **345 articles**, followed by the United States (168 articles) and Germany (48 **articles)**.
- Other countries like **India, France, and Italy** also play significant roles, though at a smaller scale.

This demonstrates that publications are highly concentrated in certain countries, reflecting their research infrastructure and investment.

Top 10 countries with most Articles published

author_countries	Count of Article title
China	345
France	17
Germany	48
India	45
Italy	35
Netherlands	14
Portugal	9
Spain	19
United Kingdom	48
United States	168
Total	665

Q4: What do these findings indicate about global research collaboration?

Answer:
The diversity of **86 distinct countries** participating in publications signifies a high level of international collaboration. Countries like **China** and the **United States** dominate, but contributions from Europe, Asia, and other regions underline a global network.

Analysis of Citation Counts

Q5: How does the high percentage of open access journals (72.89%) contribute to the increased citation counts in comparison to non-open access journals?

Answer :

Key Observations and Insights:

1. **Highest Citation Count by Journal:**
    - **Journal Name:** *Parkinson's Disease*
    - **Total Citations:** 3,936
    - **Significance:** This journal leads with the highest citation count, indicating its substantial impact in the

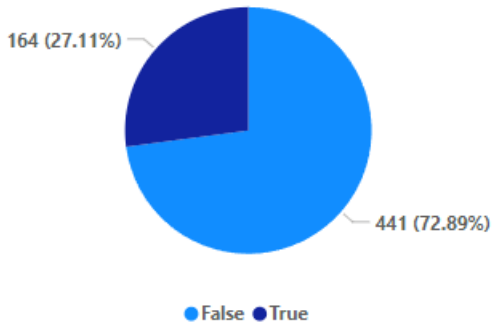
research community, particularly in the field of neurodegenerative diseases.

2. **Distribution of Citations Across Journals:**
    - A total of **40,000 citations** are distributed among **605 journals**.
    - High-impact journals like *Journal of Hepatology* (2,106 citations) and *European Journal of Heart Failure* (2,048 citations) contribute significantly to the overall citation pool, showing their prominence in their respective domains.
3. **Open Access Journals and Citations:**
    - **72.89% of journals** are open access. These journals tend to have a broader reach, enabling higher accessibility and potentially driving up citation counts.
4. **Role of Parent Publishers:**
    - Major publishers like Elsevier, with **27 child publishers**, play a dominant role in disseminating highly cited research.

Journal Names and Count of citation

clean_journal_name	Count of citations_research_article_count
Parkinson'S Disease	3936
Journal of Hepatology	2106
European Journal of Heart Failure	2048
American Journal of Human Genetics	1984
Medicina Intensiva (English Edition)	1728
Sports Medicine and Health Science	1309
Journal of Biochemical and Molecular Toxicology	630
Europace	600
Health Science Reports	599
Communications Biology	588
Indian Chemical Engineer	560
Acta Anaesthesiologica Scandinavica	528
Scientific Reports	505
Nature Communications	476
Journal of Molecular Structure	376
Advanced Biology	360
Biomedicines	348
Agricultural and Forest Meteorology	336
International Journal of Molecular Sciences	322
Journal of Alloys and Compounds	316
European Journal of Paediatric Neurology	308
Animals	306
Angewandte Chemie - International Edition	275
Journal of the American College of Cardiology	264
Fusion Engineering and Design	245
Cancer Science	241
Chemical Engineering Journal	240
Lwt	239
Value in Health	222
Resources Policy	217
Total	40000

Journals Open Access Status



Parent Publication and Child publications

publisher_family_parent_name	Total_distinct_publishers_name
Elsevier	27
Springer Nature	20
Wiley-Blackwell	12
Informa PLC	8
Oxford University Press	4
MDPI	2
SAGE	1
Total	**73**

Submission Time and Published Time

Q5: What does the average number of days between the requested date and the published date reveal about the typical publication timeline?

Answer:

With an average of **158** days between the requested date and the published date, it suggests that, on average, publications take approximately five months to go from the submission or request stage to final publication. This indicates a relatively moderate publication timeline, which could involve review, revisions, and other necessary processes before a paper is officially published. It could also point to a balanced workflow for journals, allowing for thorough evaluation without significant delays.

Q6: What are the key trends in journal influence, accessibility, and publishing dominance?

Answer:

- *Highest Citation Count:* The journal **"Parkinson's Disease"** leads with 3,936 citations, highlighting its prominence in research.
- *Open Access Trends:* **72.89% of journals** are open access, showcasing a shift towards wider research accessibility.
- *Publishing Dominance:* **Elsevier** has the highest number of child publishers (27), emphasizing its pivotal role in the academic publishing ecosystem.

Q7: What relationships can be observed between journal accessibility, citation counts, and publishing diversity?

Answer:

- *Open Access and Citations:* Open access journals' broader reach may contribute to their higher citation counts, though further analysis would solidify this observation.
- *Child Publishers and Diversity:* Leading publishers like **Elsevier** and **Springer Nature** leverage their numerous child publishers to foster research diversity, thereby dominating citation metrics and expanding influence across multiple fields.

Reference Analysis

Q8: What patterns emerge from authorship, collaboration, and article references?

Answers:

- *Collaboration in Authorship:* The article **"Prevalence and Factors Associated with Drooling in Parkinson's Disease"** has the most authors (**48**), reflecting the collaborative effort behind high-impact.
- *Reference Depth:* Articles like **"Prevalence and Factors Associated with Drooling in Parkinson's Disease"** exhibit high reference counts (**49.00**), indicating a robust foundation of prior research and impactful contributions.
- *Themes in Research:* Among articles with maximum authors, research themes revolve around pressing issues such as COVID-19, genetics, and obesity-related outcomes, underscoring the relevance of current global health challenges.

Top 5 Average Reference count for Articles

Article title	Average of references_count
Prevalence and Factors Associated with Drooling in Parkinson's Disease: Results from a Longitudinal Prospective Cohort and Comparison with a Control Group.	49.00
Genome sequencing and comprehensive rare-variant analysis of 465 families with neurodevelopmental disorders	48.00
Anticoagulation improves survival in patients with cirrhosis and portal vein thrombosis: The IMPORTAL competing-risk meta-analysis	36.00
Mineralocorticoid receptor antagonist use and the effects of empagliflozin on clinical outcomes in patients admitted for acute heart failure: findings from EMPULSE.	21.00
Association of obesity on the outcome of critically ill patients affected by COVID-19	20.00

Conclusion –

This analysis highlights key trends in the evolving landscape of academic publishing, offering valuable insights for researchers, publishers, and policymakers aiming to improve research dissemination, collaboration, and impact. Further exploration into citation correlations with global breakthroughs and subject-specific trends can provide deeper understanding and enhance future research efforts.