A Comprehensive Analysis of Development Trends and Geographical Distribution in Data Science

STA220 Final Project

Author: Jingkun Yang, Yichao Dong

Date: 2024-03-13

The code for the report has been shared on Github: YichaoDong/STA220-Final-Project

Abstract

Data science is a rapidly evolving field that integrates many domains to extract insights from data. This study analyzes temporal and geographical trends in data science publications using data from the Open Library. This research illustrates the current landscape of data science, demonstrating its significance in interdisciplinary areas.

Introduction

In the current global context, data science is an interdisciplinary field that leverages scientific methods, algorithms, and systems to extract knowledge and insights from structured and unstructured data. It integrates techniques from statistics, machine learning, computer science, and domain-specific expertise to analyze complex datasets and support data-driven decision making. The field encompasses various subdomains, including data mining, predictive modeling, big data analytics, and artificial intelligence (AI), making it a cornerstone of modern technological advancements.

With the exponential growth of data in recent years, driven by the proliferation of the Internet of Things, social media, and enterprise digital transformation, data science has witnessed significant evolution. Meanwhile, the significance of data science extends beyond technological advancements, it plays a crucial role in economic development, scientific research, and policymaking. Therefore, a comprehensive analysis of data science contributes to advancements across various domains, including technology, business, healthcare, and policymaking, by enabling data-driven insights and informed decision-making.

Our study's primary goal is to conduct a detailed analysis of the development trends and geographical distribution in data science. We collected a relatively comprehensive dataset from the Open Library, a valuable online platform that provides access to a vast collection of books.

Our analysis aims to enhance our understanding of the dynamic of data science by concretizing its development trends and geographical characteristics. Based on this, we seek to provide insights into its further advancement.

The challenges encountered during the research process are multifaceted. To begin with, appropriate methods for data collection and processing are required to deal with the extensive and complex datasets obtained from the website. Moreover, standardizing the initial data is essential to ensure its accuracy and relevance for subsequent analysis. In addition, extracting key insights from the analysis results and conducting further in-depth examination and interpretation are crucial for deriving meaningful conclusions.

Data Analysis and Findings

Data Collection and Processing

Our objective is to identify the temporal and geographical publication trends of relevant journals since the emergence of the concept of data science, with a primary focus on all publications related to data science. To achieve this, we extracted information on publication titles and subjects from the Open Library and used this dataset as the foundation for our subsequent analysis.

The process of extracting and cleaning data followed a structured approach to ensure meaningful and well-formatted information for further analysis. Initially, words were extracted from both the title and subject fields of each book entry retrieved from the Open Library. These words were then split into individual components and complied into a raw collection of extracted terms.

Following extraction, the data underwent a refinement process to improve its quality. A predefined set of stops words was used to filter out common, non-informative terms such as “the”, “and”, and “in”. To maintain consistency, all words were converted to lowercase, and punctuation marks were removed. Additionally, non-alphabetical words were discarded, ensuring that word list was produced, reducing noise and enhancing the quality of insights for subsequent analysis.

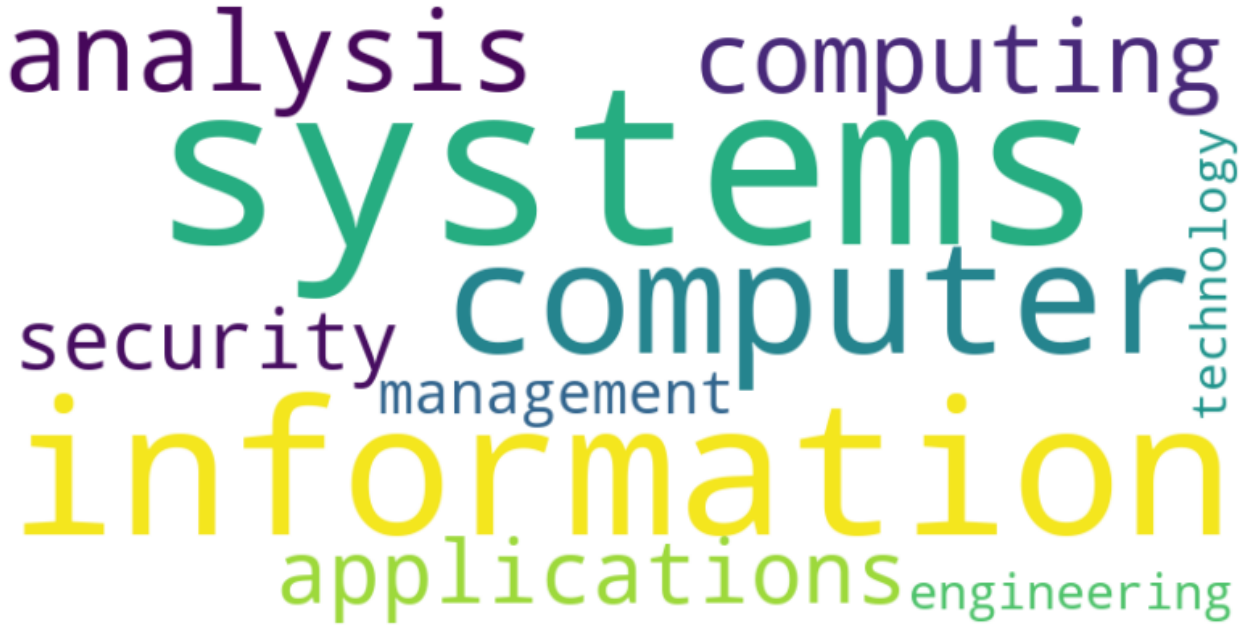
Keyword Frequency Analysis

We utilized the Natural Language Toolkit (nltk) library in Python to generate and display a word cloud to visually represent the most frequently occurring keywords in our dataset.

First, we calculated the frequency of each word and extracted the top 12 most frequently occurring words. To ensure a more diverse and accurate representation, we excluded the two

most frequent terms since they would be “data” and “science”. We retained the words ranked from 3rd to 12th place.

Next, we created a refined word frequency dictionary based on the selected words. Using this data, we constructed a word cloud, where the size of each word corresponded to its frequency in the dataset. Finally, we visualized the word cloud.



From the visualization, key terms such as “systems”, “information”, “computer”, and “applications” stand out, indicating that data science is closely linked to computing systems, information management, and practical applications. Additionally, words like “security”, “technology”, and “engineering” highlight the interdisciplinary nature of the field, encompassing aspects of cybersecurity, technological advancements, and engineering solutions.

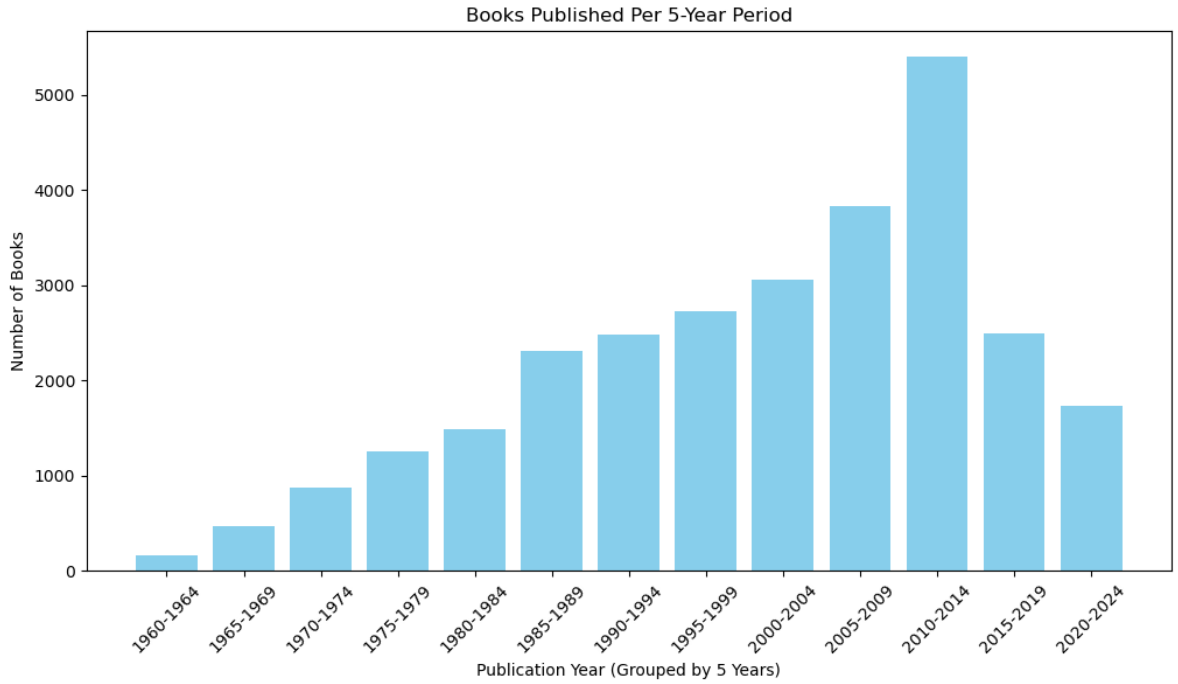
This word cloud provides valuable insights into the core focus areas of data science, reflecting its strong ties to computing, data security, and applied analytics. It serves as a quick and effective way to understand the dominant themes and research directions within the field.

Temporal Analysis of Data Science Publications

To better understand the publication trends in data science, we analyzed the number of books published over time, grouping them into five-year periods. By extracting publication years from the dataset, we identified trends in academic and industry-related works, highlighting periods of significant growth and decline. Our approach ensured a structured analysis, allowing us to observe patterns in data science publications over the past decades.

We began by extracting publication years from the dataset while filtering out records outside the range of 1968 to 2024 to maintain relevance. The extracted years were then categorized into five-year intervals, ensuring a more structured and interpretable distribution. By aggregating the number of books published within each period, we created a dataset suitable for trend analysis.

To facilitate visualization, we sorted the grouped years in chronological order. Each interval was formatted as a range (e.g., “2000-2004”) for clarity. The total number of books published within each interval was computed, and these values were used to construct a bar chart that visually represents the trends over time.



The visualization reveals several key insights regarding the temporal distribution of data science publications. First, there is a clear upward trend in the number of books published over time, with a particularly significant increase after 2000. The publication count reached its peak during 2010-2014, indicating a period of heightened academic and industry interest in data science.

Following this peak, we observe a decline in the number of books published from 2015 onwards, with a noticeable drop in the 2020-2024 period. Despite this recent decline, the overall trend suggests that data science has experienced substantial growth over the past few decades, particularly in the early 21st century. These findings align with the increasing demand for data-driven technologies, machine learning advancements, and the integration of AI across various industries (Lee, J., Bagheri, B., & Kao, H. A. (2015).).

Our temporal analysis of data science publications demonstrates that the field has undergone significant growth, especially since the early 2000s. The observed peak in 2010-2014 suggests a

period of intense research activity and publication efforts. While the decline in recent years may indicate other reasons rather than a loss of interest.

This analysis provides a historical perspective on the evolution of data science research and its dissemination. Future work can expand on this by incorporating broader sources and examining the impact of specific technological breakthroughs on publication trends.

Geographical Distribution of Data Science Publications

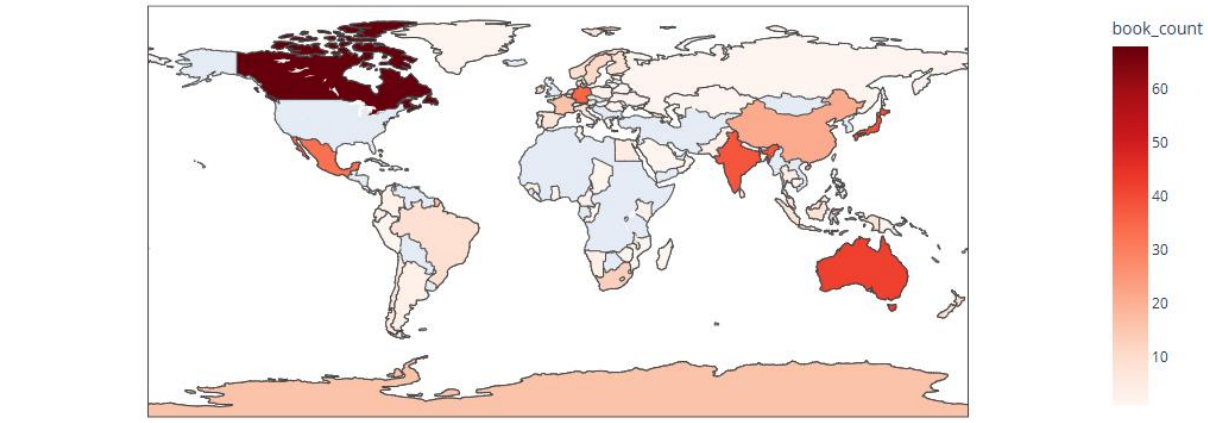
To gain insight into the geographical distribution of data science-related publications, we conducted an analysis of book counts across different countries and areas. By extracting location-based publication data, we aimed to visualize which regions contribute the most to data science literature. This analysis helps identify the global spread of research efforts and potential disparities in knowledge production.

We started by retrieving data science-related books from the Open Library, filtering them based on their associated publication locations. For each country, we queried the dataset to count the number of books published in that region. The results were stored in a dictionary, mapping each country to its corresponding book count.

Since raw location data from Open Library may contain inconsistencies, we standardized country names using a predefined country list. Additionally, to improve clarity in visualization, we excluded the United States, as its publication volume was significantly higher than other countries, which could overshadow global patterns in a comparative analysis.

To create a meaningful geographical visualization, we converted country names into ISO Alpha-3 codes, a standardized three-letter country code format. This conversion ensured compatibility with mapping tools and allowed for accurate data representation.

Using Plotly’s choropleth mapping, we assigned colors to each country based on the number of books published, with darker shades representing higher publication volumes. The color ranged from light to dark red, effectively highlighting regional disparities.



The resulting geographical heatmap reveals that Canada, India, and Australia exhibit a high number of publications, suggesting that strong research activities in these regions. Meanwhile, European countries such as Germany, France, and the Netherlands also show a notable number of contributions. However, developing countries, including parts of Africa and South America, display significantly lower book counts, indicating potential gaps in research output or accessibility to publication platforms.

By mapping the geographical distribution of data science publications, we identified regional disparities in research output. The concentration of publications in developed countries indicates strong institutional support and technological infrastructure for data science research. However, the lower representation in some regions suggests a need for increased investment in research initiatives, open-access publishing, and international collaboration.

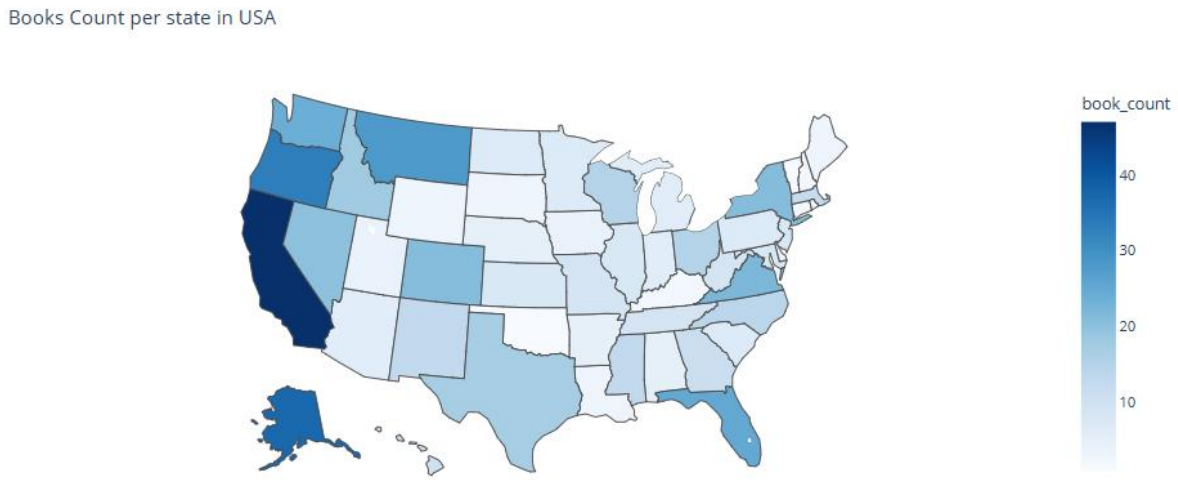
To further examine the geographical distribution of data science-related publications within the United States, we analyzed the number of books published across different states. This investigation helps to identify regions with significant research activity and highlights potential disparities in publication output at the state level. By visualizing these variations, we gain insight into how data science research is distributed across the country.

We began by retrieving the number of data science-related books published in each U.S. state. For each state, we queried the Open Library to determine the count of relevant books and stored the results in a structured format. States with zero publications were excluded from the final dataset to ensure meaningful representation.

To facilitate visualization, we standardized state names by mapping them to their respective state abbreviations, ensuring compatibility with geographic visualization tools. The data was then

formatted into a structured table, linking each state's abbreviation to its corresponding book count.

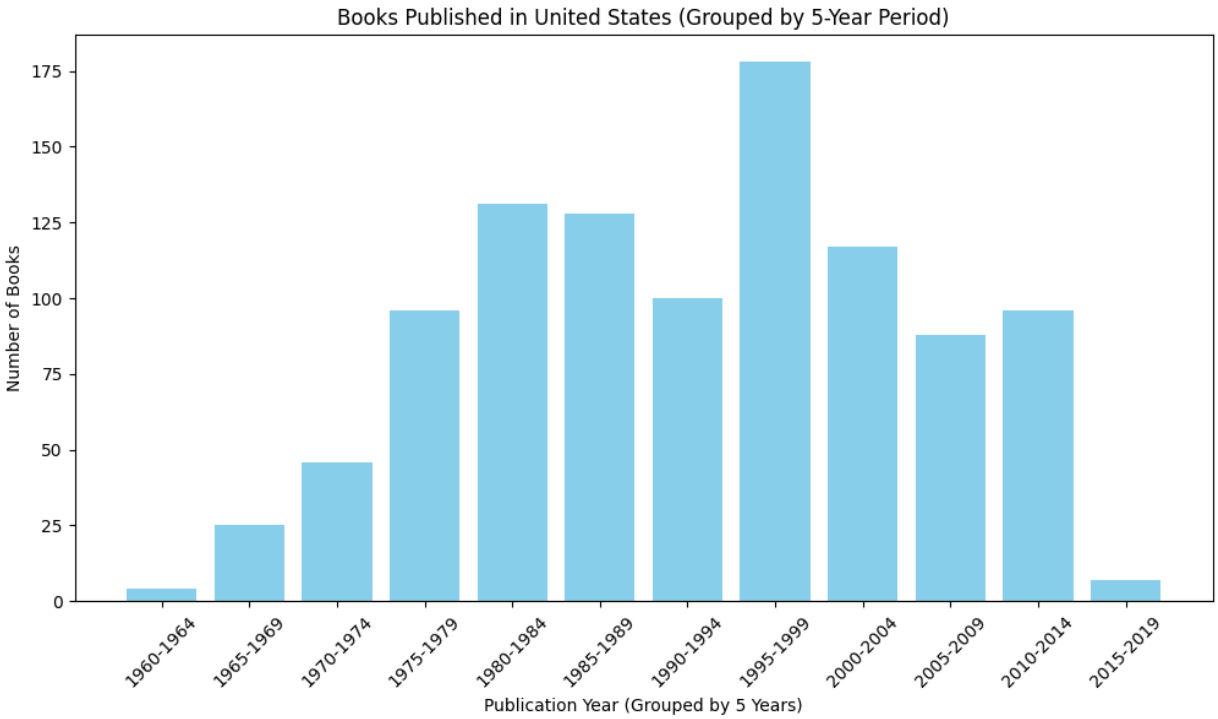
Similar to the global map, we utilized choropleth mapping to illustrate state-level variations. This visualization provides an intuitive understanding of how research output is distributed within the United States.

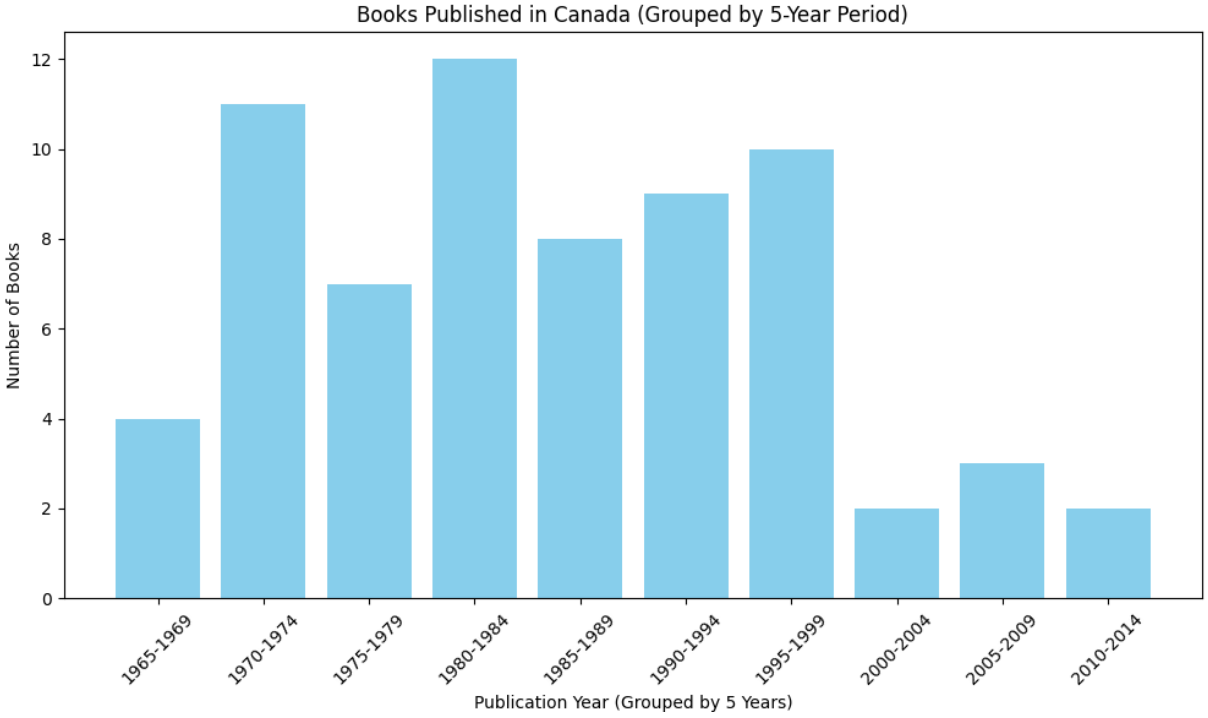


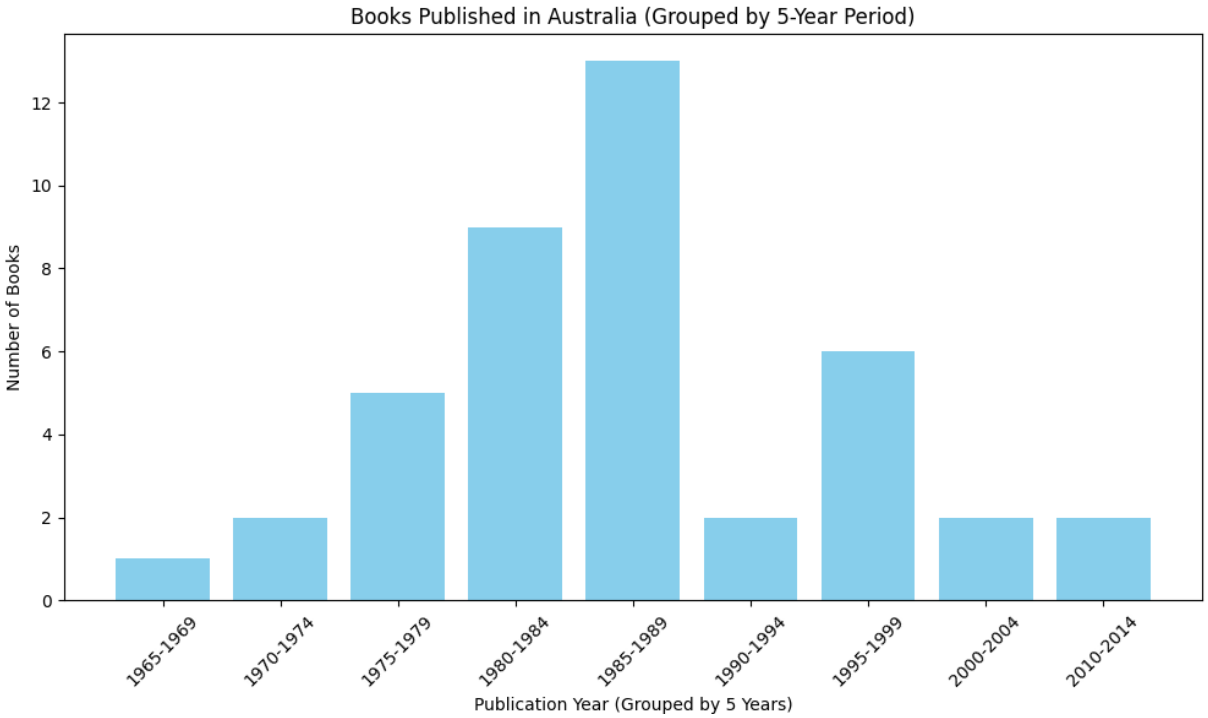
The analysis of state-level data science publications in the United States highlights regional disparities in research output. States with strong academic and technological infrastructures, such as California and New York, have higher publication volumes, while others contribute less to the field. This aligns with the presence of leading universities, technology companies, and research institutions in the states. These findings suggest that institutional resources, funding availability, and industry presence play a crucial role in shaping research activity at the state level.

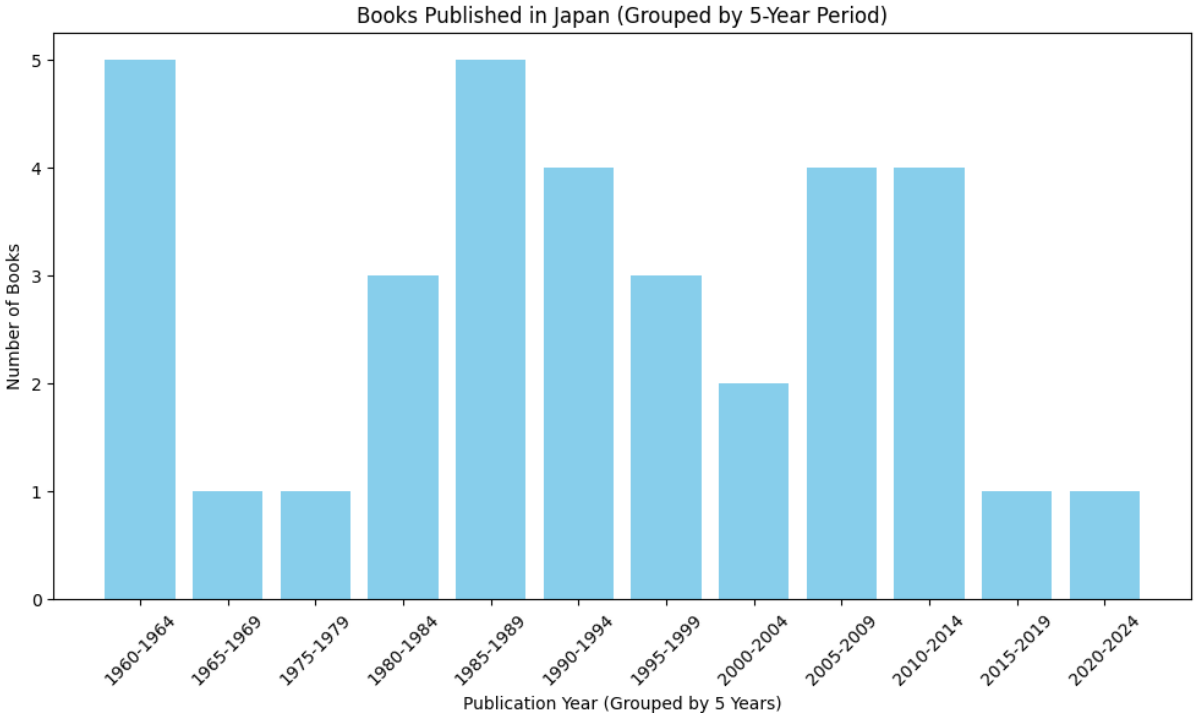
Temporal Analysis of Data Science Publications in the Top Five Countries

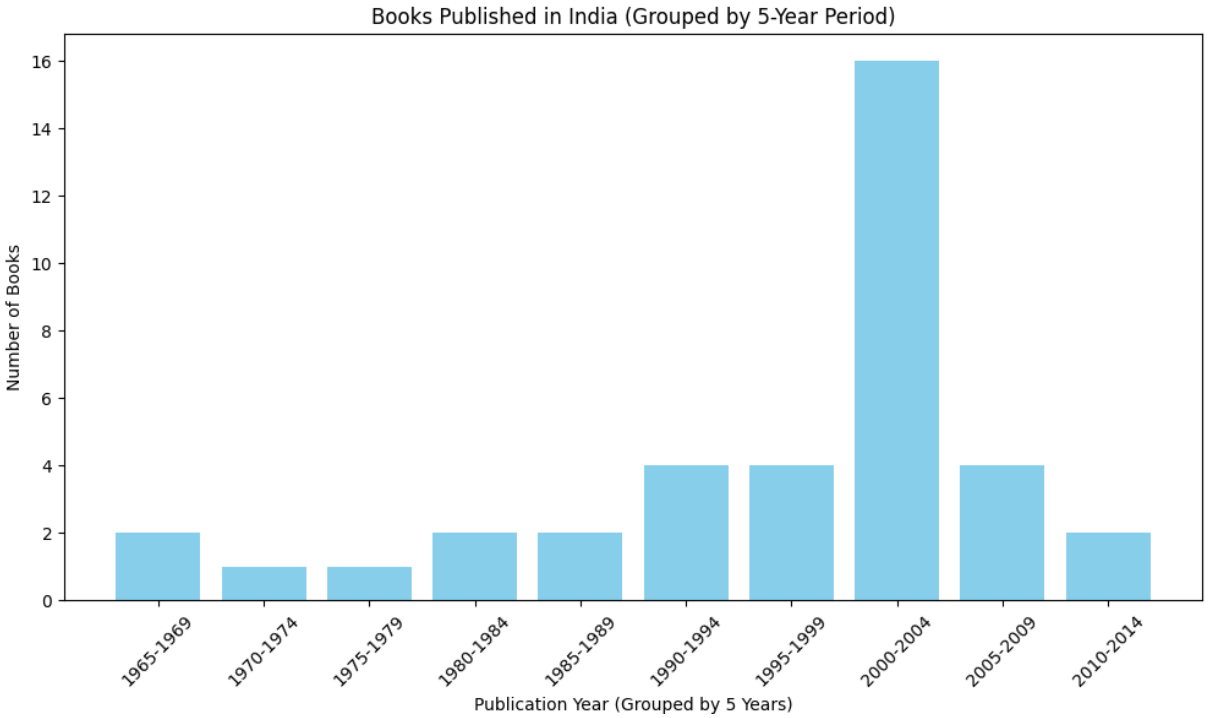
We focused on the top five countries with the highest number of publications. This analysis helps us understand how research and literature production have evolved in these leading nations over time. By grouping book publications into five-year intervals, we aim to identify patterns of growth, peaks in publication activity, and potential shifts in research focus.


Books Published in United States (Grouped by 5-Year Period)
Number of Books
0
25
50
75
100
125
150
175
1960-1964
1965-1969
1970-1974
1975-1979
1980-1984
1985-1989
1990-1994
1995-1999
2000-2004
2005-2009
2010-2014
2015-2019
Publication Year (Grouped by 5 Years)


Books Published in Canada (Grouped by 5-Year Period)
Number of Books
0
2
4
6
8
10
12
1965-1969
1970-1974
1975-1979
1980-1984
1985-1989
1990-1994
1995-1999
2000-2004
2005-2009
2010-2014
Publication Year (Grouped by 5 Years)


Books Published in Australia (Grouped by 5-Year Period)
Number of Books
0
2
4
6
8
10
12
1965-1969
1970-1974
1975-1979
1980-1984
1985-1989
1990-1994
1995-1999
2000-2004
2010-2014
Publication Year (Grouped by 5 Years)


Books Published in Japan (Grouped by 5-Year Period)
Number of Books
0
1
2
3
4
5
1960-1964
1965-1969
1975-1979
1980-1984
1985-1989
1990-1994
1995-1999
2000-2004
2005-2009
2010-2014
2015-2019
2020-2024
Publication Year (Grouped by 5 Years)



Our analysis of data science publications across the top five countries reveals distinct trends in each region. The United States and Canada have long-established histories of data science research, while India's significant rise in the 2000s highlights its emerging role in global data science contributions. The differences in publication trends across these countries reflect variations in research funding, academic focus, and technological adoption. Future research could explore the underlying factors driving publication trends.

Discussion and Conclusion

This study examined data science publication trends, revealing key thematic, temporal, and geographical patterns. The field experienced significant growth, peaking in 2010-2014, with a decline in recent years. Geographically, developed nations dominate data science research, while developing regions exhibit lower output.

In conclusion, this study not only captures the current state of data science but also projects future trajectories. Factors such as technological infrastructure, research funding, and economic policies play crucial roles in shaping data science development, which requires efforts from all aspects to ensure a promising development future for data science.

References

Lee, J., Bagheri, B., & Kao, H. A. (2015). A cyber-physical systems architecture for industry 4.0-based manufacturing systems. *Manufacturing Letters*, 3, 18-23. https://doi.org/10.1016/j.mfglet.2014.12.001