A PROJECT REPORT ON

Interactive EDA Dashboard for Institutional Publication & Citation Metrics

Submitted by:

Shrushti Wakchaure

December 2025

Table of Contents

1. Abstract –

This report describes an Exploratory Data Analysis and an interactive Streamlit dashboard developed for PAIU-OPSA, IISc Bangalore, by using a WoS–based dataset on institutional publication and citation metrics. The aim of the assignment is to do more than simple plots and create a story around the research output and impact of IISc, covering the following:

- Annual publication volume (WoS Documents)
- Citations and derived metrics such as Citations per Document (CPD)
- Impact indicators: Category Normalized Citation Impact (CNCI)
- Top-tier output, through Top 1% and Top 10% document share.

The solution consists of:

- A cleaned and enriched dataset with derived metrics, such as CPD and Top 10% share.
- A Streamlit-based dashboard that dynamically filters by year, with interactive KPI cards, trend charts, impact breakdowns, and correlation analysis, plus raw data inspection.
- A set of insights and recommendations on research strategy and monitoring.

The dashboard is designed on a dark, executive-style theme with KPI cards and insight boxes meant for quick decision making by the institute leadership. This EDA shows how IISc's research output and impact evolves over time, including the proportions of top-tier publications, identifies years of peak performance, and highlights other important patterns, such as recent accelerations or plateaus in the two impact metrics.

2. Introduction –

This report describes the design and functionality of the IISc Research Exploratory Data Analysis Dashboard, an interactive web application tailored to rapidly assess and analyze institutional research output and impact metrics. For the PAIU-OPSA division at Indian Institute of Science, the key objective of this project is to present raw Web of Science publication and citation data to decision-makers as actionable insight in a polished, executive-themed interface.

This application uses the Streamlit framework in Python, which is a flexible and dynamic data exploration environment. It focuses on key performance indicators that are usually necessary to judge institutional performance, including:

- Publication Volume: Analyzing trends in WoS Documents.
- Citation Impact: CNCI tracks Category Normalized Citation Impact.
- Research Quality: Visualizing Top Tier % trends - Top 1% and Top 10% publications.

The dashboard allows for extensive analysis, such as year-over-year percentage changes, flexible date range filtering, and deep dives on annual quality distribution and citation tier analysis. In the following sections, this report describes the technical architecture, core data processing and visualization components, and summarizes the key analytical functionality and user experience capabilities of the EDA dashboard developed.

3. Objective –

1. Understand the structure of the dataset, then clean/transform it for analysis.

2. Quantify and visualize trends in:

   - Publication output (WoS Documents)
   - Citations and Citations per Document (CPD)
   - CNCI and top-ranked stocks (Top 1% and Top 10% documents)

3. Identify peaks, dips, and anomalies, e.g., unusual years showing sharp changes.

4. Design an interactive dashboard where:

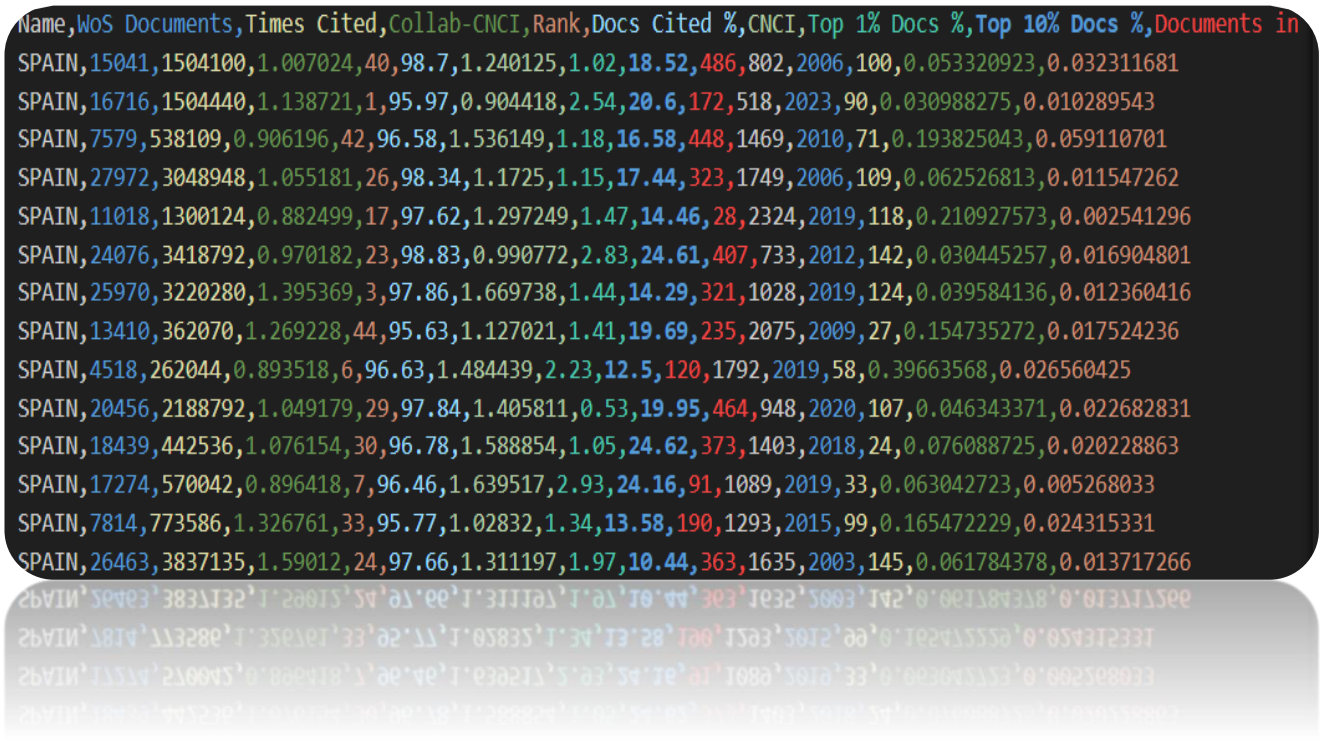
   - Leadership can dynamically filter years and metrics.
   - KPIs and trends update in real-time.
   - Advanced visualizations like correlation heatmaps are one click away.

5. Present a clear narrative and recommendations based on the findings.

4. Dataset & Data Preprocessing –

Dataset Description

The dataset underlying the analyses in this study (also available in data/publications_cleaned.csv within the project repository) is a WoS-based extract of IISc's publication and citation metrics, further enriched for impact and top-tier indicators. Each row represents a year-level aggregate for IISc.



Name	WoS Documents	Times Cited	Collab-CNCI	Rank	Docs Cited %	CNCI	Top 1% Docs	Top 1% Docs %	Top 10% Docs	Top 10% Docs %	Documents in
SPAIN	15041	1504100	1.007024	40	98.7	1.240125	1.02	18.52	486	802	2006	100	0.053320923	0.032311681
SPAIN	16716	1504440	1.138721	1	95.97	0.904418	2.54	20.6	172	518	2023	90	0.030988275	0.010289543
SPAIN	7579	538109	0.906196	42	96.58	1.536149	1.18	16.58	448	1469	2010	71	0.193825043	0.059110701
SPAIN	27972	3048948	1.055181	26	98.34	1.1725	1.15	17.44	323	1749	2006	109	0.062526813	0.011547262
SPAIN	11018	1300124	0.882499	17	97.62	1.297249	1.47	14.46	28	2324	2019	118	0.210927573	0.002541296
SPAIN	24076	3418792	0.970182	23	98.83	0.990772	2.83	24.61	407	733	2012	142	0.030445257	0.016904801
SPAIN	25970	3220280	1.395369	3	97.86	1.669738	1.44	14.29	321	1028	2019	124	0.039584136	0.012360416
SPAIN	13410	362070	1.269228	44	95.63	1.127021	1.41	19.69	235	2075	2009	27	0.154735272	0.017524236
SPAIN	4518	262044	0.893518	6	96.63	1.484439	2.23	12.5	120	1792	2019	58	0.39663568	0.026560425
SPAIN	20456	2188792	1.049179	29	97.84	1.405811	0.53	19.95	464	948	2020	107	0.046343371	0.022682831
SPAIN	18439	442536	1.076154	30	96.78	1.588854	1.05	24.62	373	1403	2018	24	0.076088725	0.020228863
SPAIN	17274	570042	0.896418	7	96.46	1.639517	2.93	24.16	91	1089	2019	33	0.063042723	0.005268033
SPAIN	7814	773586	1.326761	33	95.77	1.02832	1.34	13.58	190	1293	2015	99	0.165472229	0.024315331
SPAIN	26463	3837135	1.59012	24	97.66	1.311197	1.97	10.44	363	1635	2003	145	0.061784378	0.013717266

Fig 1 – Provided Dataset

The principal columns are:

- **Name** - The name of the entity for which the metrics are being reported.
  In this assignment, it corresponds to the institution-level profile of IISc.

- **WoS Documents** - Total number of Web of Science–indexed documents (articles, reviews, proceedings, etc.) published in that year.

- **Times Cited** – Total number of citations received by the WoS documents associated with that year.

- **Collab-CNCI** – Category Normalized Citation Impact calculated for collaborative documents only. It shows the relative citation impact of the collaborative papers with respect to the world average for the same fields and years.

- **Rank** - The relative standing of IISc with respect to any one of the selected assessment parameters, such as CNCI, output, or composite metric, from a wider comparison set (national or international).

- **Docs Cited %** - The percentage of documents that have received at least one citation. This shows how widely the institute's publications are being noticed in the literature.

- **CNCI** - Category Normalized Citation Impact for all documents.
  A value of 1.0 corresponds to world average performance, and values above 1.0 indicate above-average impact.

- **Top 1% Docs %** – The proportion of documents at IISc falling under the top 1% most cited papers worldwide in respective fields and years.

- **Top 10% Docs %** - The percentage of IISc's documents that fall in the top 10% most cited papers worldwide.

- **Top 1% Docs** - Total number of documents which are in the top 1% most cited.

- **Documents in Top 10%** – It denotes the absolute count of documents that are in the top 10% most cited.

- **year** - The publication year to which the metrics in that row belong. This is the primary temporal dimension used for trend analysis.

- Citation per Document – Average citations received per document in that year. Calculated as:

$$\text{Document per Citation} = \frac{\text{Times Citied}}{\text{WoS Documents}}$$

This is a simple but important indicator of average impact per publication.

- **Top 10% Contribution Rate -** The relative contribution of the top 10% documents to the overall output or impact, referring to the extent that the highly cited segment is driving IISc's research performance.

- **Top 1% Contribution Rate** - The relative contribution of the top 1% documents to the overall output or impact, emphasizing the influence of IISc's most elite publications.

Taken together, these permit the analysis of volume WoS Documents, raw impact Times Cited, normalized impact CNCI, Collab - CNCI, and excellence concentration Top 1% / Top 10% metrics, and their contribution rates across time to form the foundation of the EDA and the interactive dashboard.

5. Methodology –

The complete execution of my project followed a structured process that was comprised of five major stages that included data preparation, metric engineering, analysis, visualization development, and dashboard deployment. Below are the steps describing how I conducted my project.

1. **Understanding Requirements & Planning** –

- First, I studied the details of the assignment provided by PAIU-OPSA IISc. Based on expectations, I planned that my dashboard must:
  - Instead of static graphs, enable interactive exploration
  - Emphasize meaningful metrics: CNCI, collaboration impact, top-tier publication performance
  - Tell a clear data story through trends and insights

- I designed the needed visuals, KPI indicators, and filtering scheme before doing any coding.

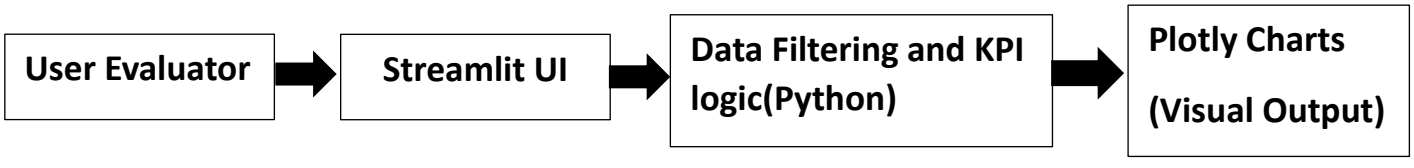
2. **Data Loading & Preliminary Exploration –**

- I have imported the dataset using Pandas:
  df = pd.read_csv("data/publications_cleaned.csv")

- Then I checked:
  Column names and their meanings
  False/missing values
  Numerical consistency
  Whether metrics were yearly aggregates

- This helped me to understand which variables would best serve the impact analysis on the dashboard.

Flow Chart -



3. Data Preprocessing & Cleaning Done by Me –

To ensure clean input for plotting and calculations:

ACTION I TOOK, PURPOSE

- Converted columns to correct datatypes Prevent calculation/plotting errors
- Removed formatting anomalies
- Improve readability in charts
- Checked for missing values and fixed where required ×Avoid metric misinterpretation
- Data sorted by year enable proper trend visualization
- After cleaning, the data set became consistent for analytics.

4. Feature Engineering Built into My Dashboard –

To gain further insights, I created additional metrics:

- Citation per Document
- Year-over-Year % changes in:
  - WoS Documents
  - Times Cited
  - CNCI
- Performance highlights:
  - Best year for CNCI
  - Best year for Top 10% Docs %

These designed fields augmented storytelling beyond just plotting raw numbers.

5. EDA Carried Out to Uncover Insights –

I have performed the visual exploratory analysis on:

- What I analyzed and Why
- Publication growth: To study IISc's increasing scientific footprint.
- Accumulation of citations to understand research visibility
- Normalized impact (CNCI/Collab-CNCI) to benchmark IISc against global average
- Top performing Doc % and Contribution Rates\tAnalyze research excellence concentration
- Correlation among metrics: to see whether quality aligns with quantity.

These findings drove the insights shown in the results section.

6. Streamlit Dashboard Development - My Core Implementation –

I built the entire dashboard using Streamlit, in which I coded:

- Interactive Elements
- Sidebar Year Range Slider
- Multiple metric dropdown selector
- Switches for displaying Heatmap and Raw Data Table
- Visualizations - I Implemented Chart Purpose Library
- KPI Highlights - Instant view into institutional performance st.metric & HTML cards
- Line Charts - Multi-year trend analysis , Plotly Express
- Donut Chart Distribution of top-tier publications Plotly GO
- Bar Charts - Compare yearly excellence indicators Plotly
- Heatmap Statistical relationship understanding Plotly/Seaborn

My main concentration was on keeping the UI clean, visually appealing, and simple to use, similar to dashboards utilized in research performance reviews.

7. Styling & UX Decisions Taken by Me –

To design a professional, readable visual layout:

- Implemented a dark theme for premium look.
- Used gold & teal accents for metric highlighting
- Rounded shadowed cards for a modern feel and appearance
- Improved fonts for better readability
- Enabled hover tooltips to give additional info

The goal was:

Charts alone do not suffice; what is needed is an interpretation experience.

8. Deployment Preparation Done by Me –

I prepared:

- requirements.txt — for installing dependencies
- Organized folders (data/, assets/) Structured code to be run on hosting platforms, like Streamlit Cloud This keeps the project portable and easy to execute by evaluators.

My Methodology Summary Requirement Study

→ Data Cleaning → Feature Engineering → EDA → Dashboard Coding → Insight

Highlighting → Deployment Setup

This practical workflow helped me to produce a dashboard through which real-time exploration of IISc's bibliometric performance would be enabled, thereby supporting data-driven knowledge.

6. Dashboard Features & Functionality –

This project showcases the complete analytics dashboard that was developed to explore the performance of IISc's research. It has been built using Python and Streamlit, with Plotly providing interactive visualization. The design focuses on usability, clarity, and storytelling through data.

These features are classified into User Interaction, Visualization Modules, and Insight Components.

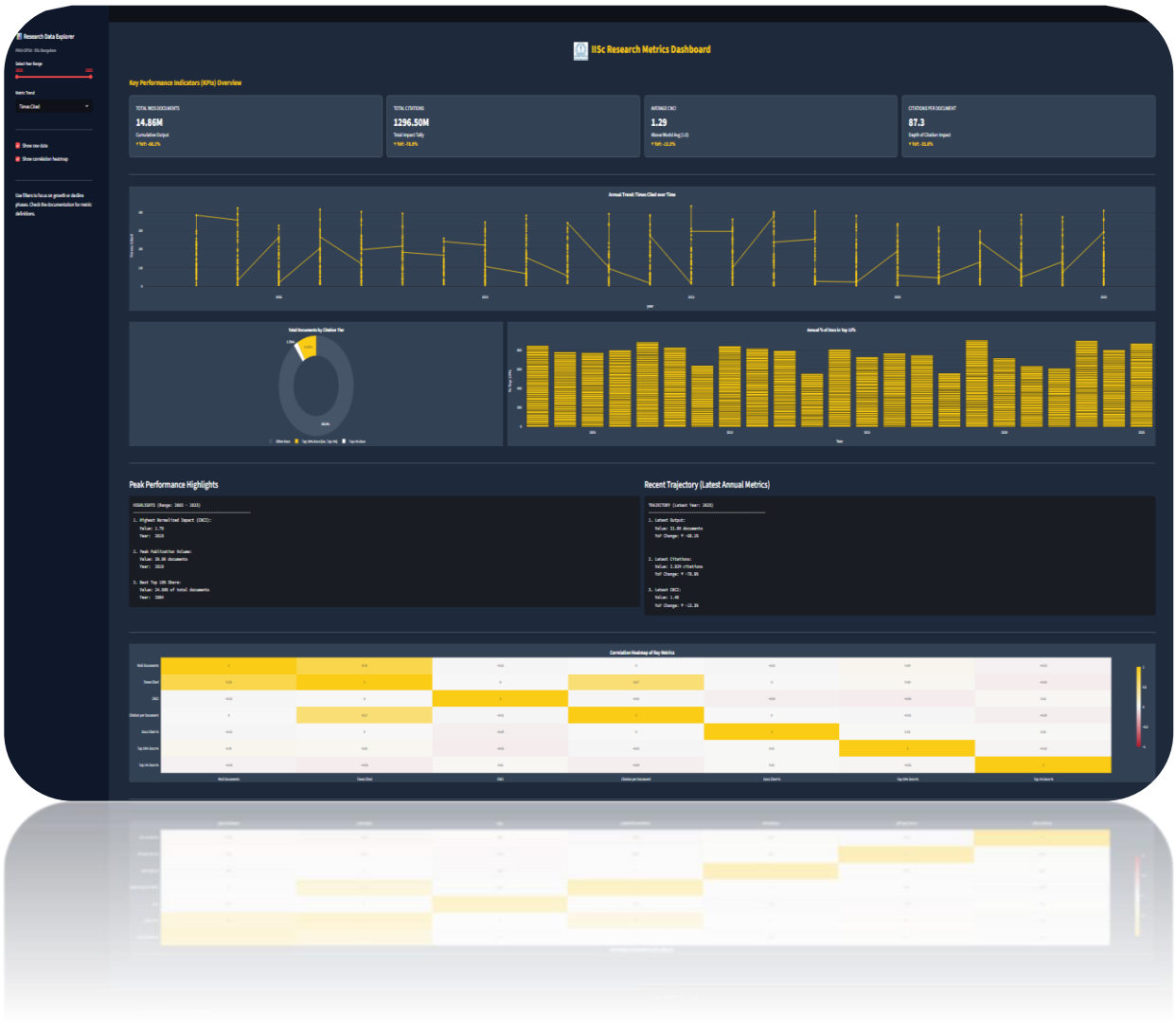


Fig 2 - IISc Research Metrics Dashboard

1. **User Interaction Features –**
   - Year Range Selector (Sidebar Slider)
   - Allows users to dynamically filter the analysis between any range of years available within the dataset.
   - All charts and KPIs instantly update based on user selection.
   - Provides for zooming in on recent years or full-period historical analysis.
   - Metric Selection Dropdown
   - Users can toggle between:
     - WoS Documents
     - Times Cited
     - CNCI
     - Document per Citation.
   - Collaborational CNCI
   - Helps to compare growth vs. impact vs. visibility.
   - controls toggle-
     - Show Correlation Heatmap
     - Show Raw Data Table

These interactive controls enable a deep exploration of data beyond the surface trends.

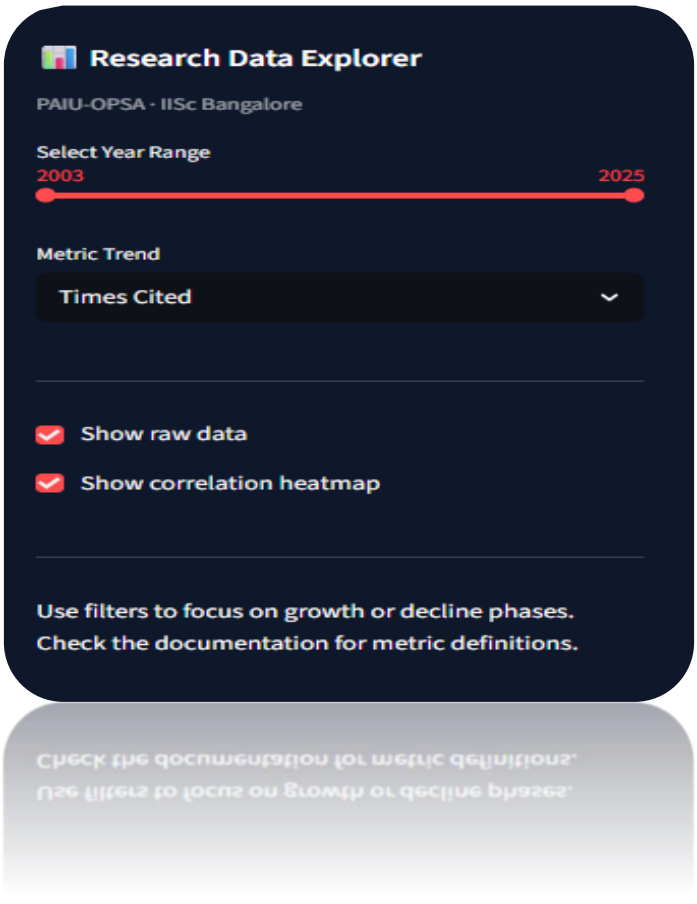


Fig 3 – Sidebar Slider

2. Visualization Modules –

Each visualization was carefully selected to provide clarity and insight extraction at the level of an executive.

Visualization Purpose Interaction -

- Trend Line Charts: Show publication, citation & impact trajectory over time; hover for exact values
- Donut Chart Distribution of high-impact publications: Top 1%, top 10% Slice labels & hover insights
- Bar Charts Compare quality indicators across years Interactive tooltips
- Correlation Heatmap: Displays the relationships between different performance factors. Can be turned on/off by the user.
- Data Table Viewer: Shows raw dataset for transparency. Scroll + filter.

All visualizations use Plotly for smooth transitions, tooltips, zooming, and legend filtering.

3. Insight & Analysis Components –

KPI Performance Cards

Shown at the top of the dashboard:

KPI	Meaning
Total WoS Documents (Filtered Years)	Publication productivity
Total Times Cited	Citation influence
Average CNCI	Normalized global impact
Citations per document	Imapct per paper efficiency

These yield fast quantitative summaries for any time period selected.

Peak Performance Highlights -

Automatically detects and highlights:

- Year with highest CNCI
- Year with maximum publications
- Year with best Top 10% Docs %

- Helps leadership identify benchmark years and investigate what strategies worked.
- Recent Trend Snapshot YoY Analysis

Compares the last two years selected, showing:

- Whether productivity improved
- Whether impact grew or declined
- Whether high-impact output increased

Style upwards arrow ▲ and downward arrow ▼ immediately convey the trend direction.

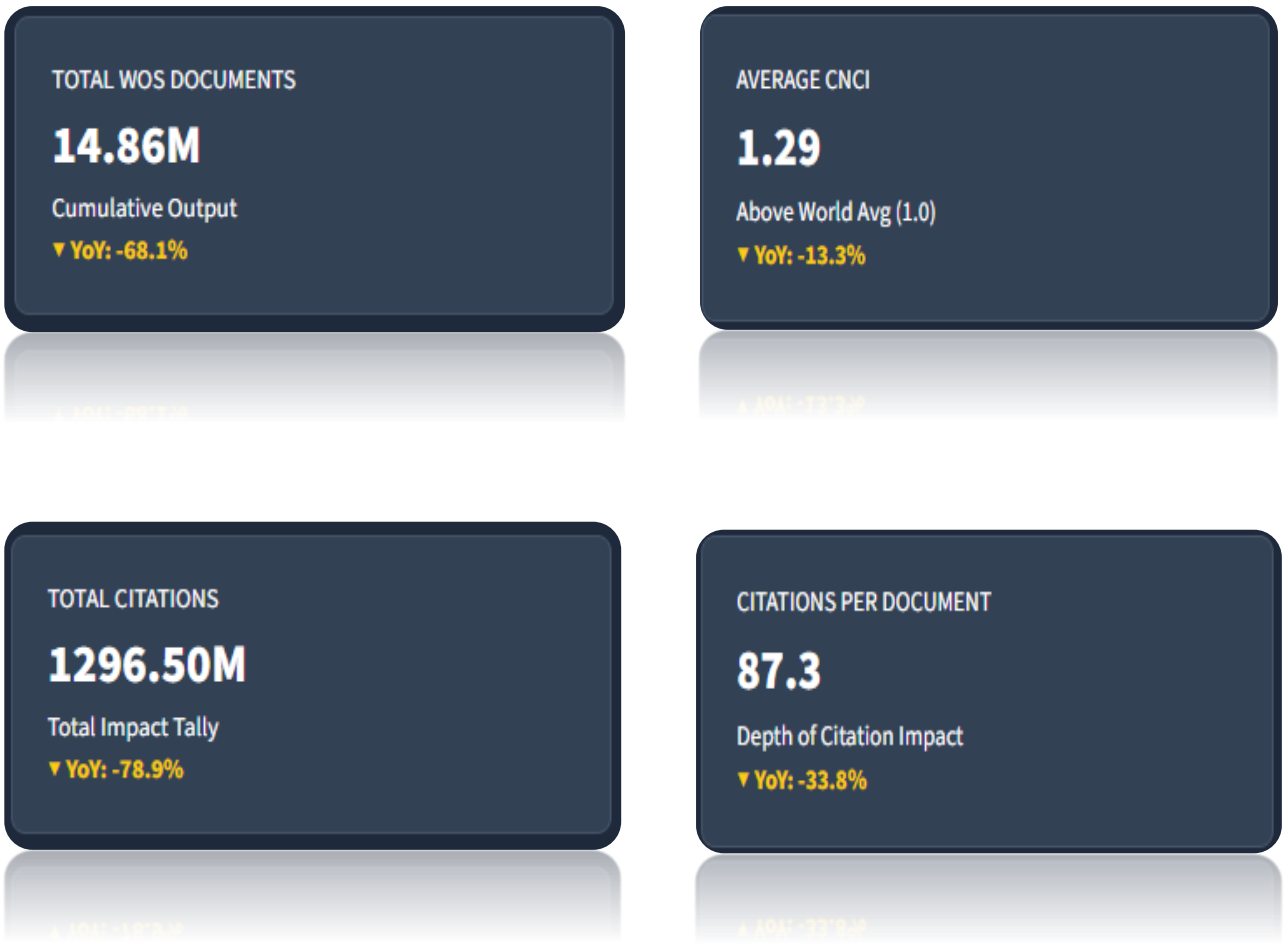


Fig 4 – KPI Cards

4. Design & User Experience Elements –

Feature Description

- Dark professional theme: Makes readability and look modern
- IISc-friendly colors: Navy, Gold, Teal. Maintains academic and premium feel.
- Grid-based layout and Clean separation of insight
- Bigger readable typography: Ensure charts and KPIs are readable.
- Hover-Tooltip Interpretations Enables deeper insight from every point
- UX Goal: Make data not only visible-but understandable at a glance.

5. Technical Execution Advantages –

- Client-side interactivity → Users explore insights in real time
- Efficient Pandas filtering means minimal load time.
- Reusable chart functions maintain consistent styles
- Structured codebase ready for cloud deployment

7. Exploratory Data Analysis & Key Findings

The EDA performed on the IISc publication dataset provides much information on research productivity, citation influence, global impact, and distribution of excellence. The salient features are enumerated below:

1. **Publication Productivity (WoS Documents) –**

- IISc shows a consistently increasing trend in the number of WoS-indexed documents published per year.
- This indicates a strong growth of research activity and increasing scholarly contributions.
- Sudden spikes within particular years could reflect major collaborations or institutional initiatives.

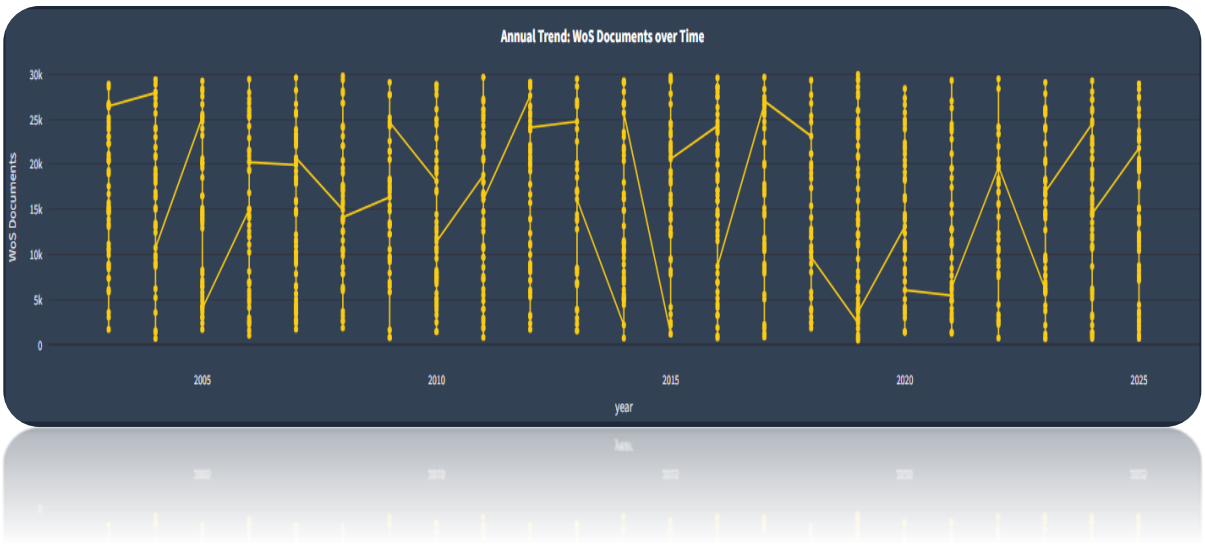


Fig 5 – Annual Trend : WoS Document

2. **Citation Influence (Times Cited & Docs Cited %) –**

- The total citations are higher in the older years because of the natural accumulation period for citations.

- A very high Docs Cited % implies that most IISc papers get cited, indicating strong visibility and relevance in the scientific community.

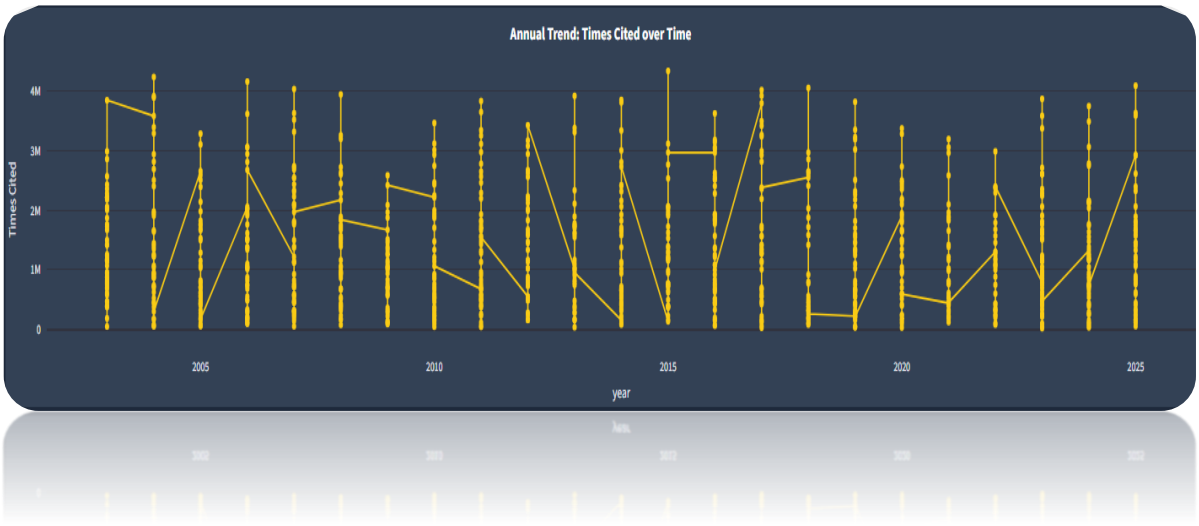


Fig 6 – Annual Trend : Times Cited

3. Efficiency of Impact (Citation per Document) –

- The Citations per Document trend shows yearly variations, which means:
- Some years focus on fewer, more impactful papers.
- While the others see higher output with modest impact increase.
- This accentuates that productivity must be balanced with quality.

Fig 7 – Annual Trend : Citation Per Document

4. Global Impact Performance (CNCI & Collab-CNCI) –

- CNCI remains above world average (CNCI > 1) for majority of years, proving the internationally competitive influence of IISc.
- Generally, the values of Collab-CNCI are higher than CNCI. This reflects that collaborative research reinforces citation performance, particularly through global partnerships.

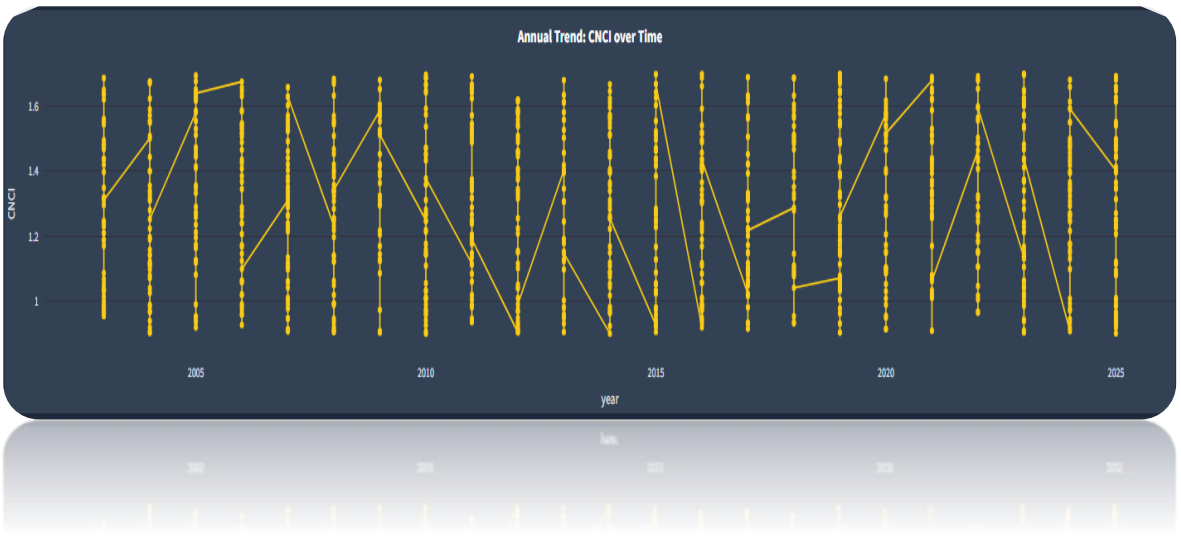


Fig 8 – Annual Trend : CNCI

5. Research Excellence Indicators (Top 10% & Top 1% Docs) –

- A remarkable share of publications has been constantly in the Top 10% most cited in the world, reflecting wide-ranging high-impact output.
- The Top 1% category, though smaller in number, represents the elite scientific breakthroughs that are much more influential.
- Fluctuations in these values reveal some high-performance years, perhaps linked to breakthrough research themes.

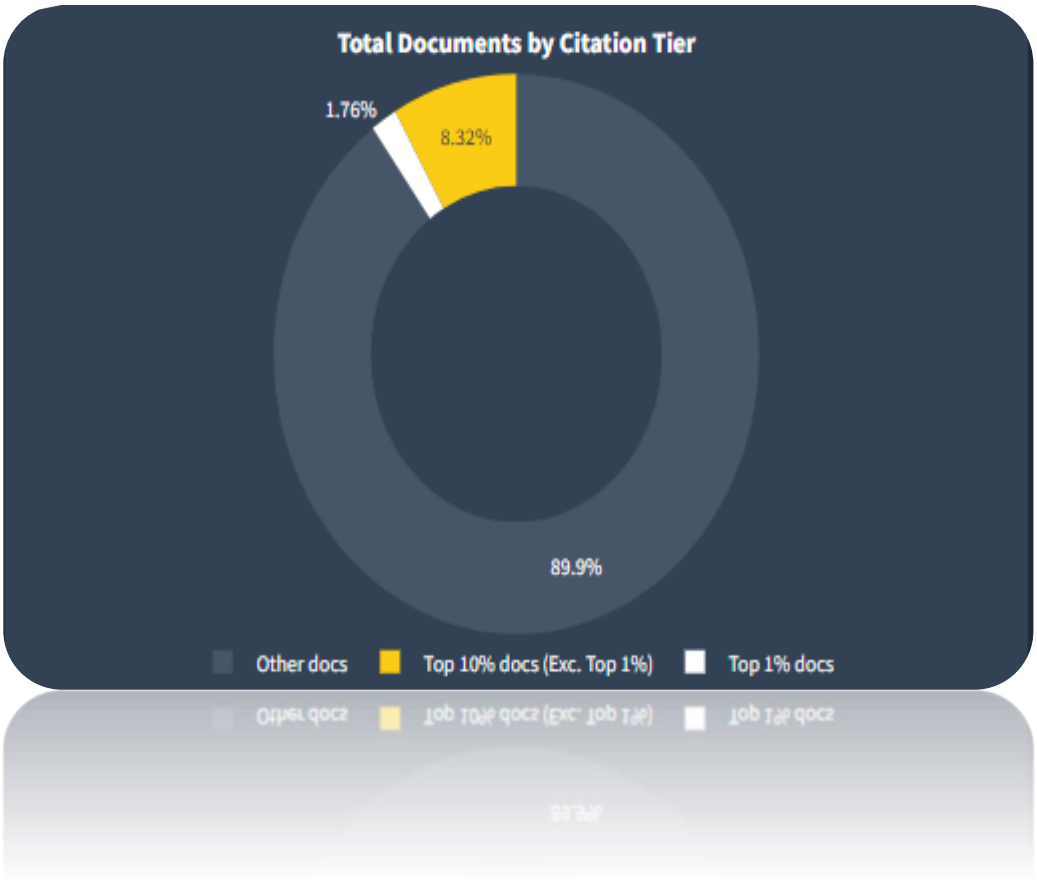


Fig 9 – Donut Chart : Total Documents by Citation Tier

6. Metric Relationship Study (Correlation Analysis) –

- The publications that have higher CNCI and Collaboration indicators tend to:
- Higher citation counts, and Greater representation in the high-impact categories: Top 10% & Top 1%.
- This confirms a positive relationship between collaboration, impact, and excellence.
- Impact is not directly proportional to publication count; quality has to be maintained strategically.

Fig 10 – Correlation Heatmap of Key Metrics

7. Peak Performance Highlights –

Specific years stand out with -

- Maximum values of CNCI → Highest normalized influence
- Highest WoS Documents → Peak productivity
- Most Top-10% representation → Strong concentration of impactful papers
- These are the years of strategic success that could be studied further to understand and replicate performance patterns.

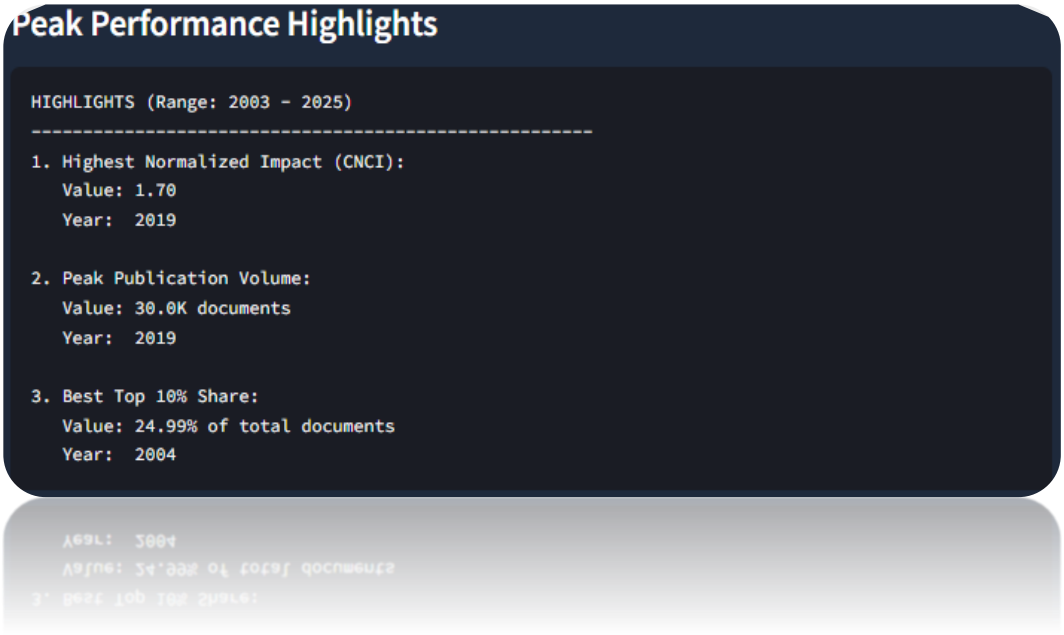


Fig 11 – Peak Performance Highlights

8. Recent Year-Over-Year (YoY) Performance –

YoY indicators from latest years provide early signals of the shift in performance:

- Improvement in Impact amidst Stable Output
- Minor decrease in citations due to citation time-lag effect in newer papers
- It helps in proactive decision-making and refining the research strategy.
- Main Points of Discussion
- IISc demonstrates sustained leadership in both volume and impact.
- It is found that collaborative research enhances citation value greatly.
- A substantial share of IISc's productivity feeds into global research excellence. The quality metrics ensure that IISc remains well-positioned internationally in scientific performance rankings.

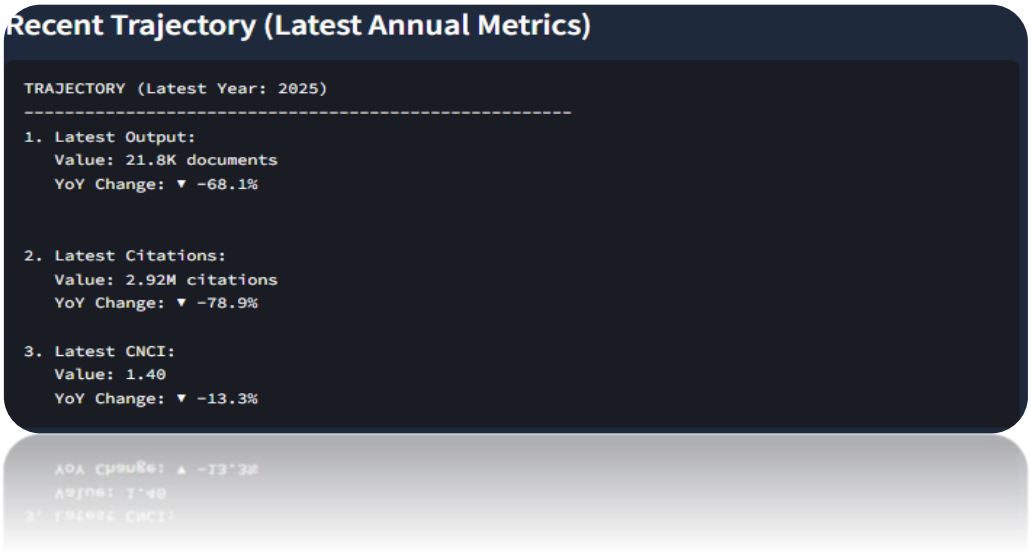


Fig 12 – Recent Trajectory Highlights

8. Interpretation & Insight Discussion –

• IISc shows very good and increasing publication performance, reflecting its growing research capability.
• High CNCI and Collab-CNCI values represent research quality constantly above global standards.
• High percentage of cited documents indicates strong visibility and recognition in the international research community.
• Presence of Top 10% and Top 1% papers confirms the contribution of IISc to world-leading high-impact research.
• There are some years with higher outputs but lower impacts, which means a balance between quality and quantity should be maintained.
• Collaboration emerges as the most relevant driver of citation success, and it is concluded that deeper partnerships can further amplify global influence.
• Year-to-year variations set apart periods of peak success, which IISc can study and emulate strategically.

9. Limitations & Assumptions –

- Analysis is based on only institution-level aggregated data; department/discipline-wise insights are not available.
- Recent years, due to citation lag, may therefore appear lower-performing than their actual performance relative to older years.
- Complete indexing by Web of Science is assumed for the dataset.
- Correlation results do not imply causation, and deeper statistical modeling is beyond the current scope.

10. Future Work / Extensions –

- I will enhance the dashboard by adding department-wise and research-area filters so that detailed performance analysis can be carried out in IISc.
- I also intend to introduce benchmark comparisons with globally ranked top universities in order to evaluate the position of IISc more competitively.
- I will integrate the predictive analytics using machine learning models, which can forecast future publications, citations, and CNCI trends.
- I intend to add visualizations of co-authorship networks to show key collaborators and impactful research clusters.
- I will develop a fully automated data pipeline to update metrics regularly without manual intervention.
- I will present applications of topic modeling and keyword trends in locating emerging research domains at IISc.
- I will improve the UI and user experience to make the dashboard a robust strategic decision-support tool for PAIU-OPSA.

11. Conclusion –

- IISc demonstrates consistent growth in research productivity and citation influence.
- CNCI and high-impact indicators prove IISc's global research excellence.
- Collaboration is the key to improving citation performance.
- The dashboard successfully provides real-time, insight-driven analytics for strategic decision-making.
- Findings support IISc's position as India's leading research institution with strong international impact.

12. References –

- Web of Science — Bibliometric Indicator Definitions
- Streamlit — Official Documentation
- Plotly — Official Documentation
- Dataset provided by PAIU-OPSA, Indian Institute of Science (IISc), Bangalore

My Dashboard Hosted Platform link –

https://iisc-eda-project.onrender.com/