Global Research Impact Analysis Dashboard - Technical Report

1. Application Overview

Purpose: Interactive web application for analyzing global research publication and citation data (2003-2025)

Technology Stack:

- Framework: Streamlit
- Visualization: Plotly (Express & Graph Objects)
- Data Processing: Pandas, NumPy
- Statistics: SciPy, Scikit-learn

Key Features: 7 analysis modules with 40+ interactive visualizations

2. Data Architecture

Input Data Requirements

- File: publications.txt (tab-separated)
- 8 core columns: Name, year, Web of Science Documents, Times Cited, CNCI, Collab-CNCI, % Top 1%, % Top 10%

Derived Metrics (5)

1. **Citations_per_Doc** = Times Cited / Documents
2. **Elite_Ratio** = % Top 1% / % Top 10%
3. **Impact_Score** = CNCI × % Top 10%
4. **H_Index_Proxy** = √(Citations × Documents)
5. **Productivity_Index** = Documents / Years Active

3. Filter Controls

Filter	Type	Purpose
Country	Multi-select	Geographic focus
Year Range	Slider	Temporal scope
Citation Threshold	Numeric input	Minimum impact level
Primary Metric	Dropdown	Analysis focus
CNCI Range	Slider	Quality filtering
Show Outliers	Checkbox	Include/exclude anomalies

4. Analysis Modules

Tab 1: Data Overview

- **Data Quality:** Missing values, duplicates, data types
- **Top 15 Rankings:** Dynamic by selected metric
- **CNCI Distribution:** Histogram with statistical annotations
- **Correlation Matrix:** 8 metrics, heatmap visualization
- **Data Table:** Top 100 records with gradient coloring

Tab 2: Geographic Analysis

- **Country-Year Heatmap:** CNCI performance across time
- **Efficiency Scatter:** Documents vs Citations/Doc (bubble size = citations)
- **Elite Output:** Top 1% vs Top 10% comparison (top 15 countries)
- **Regional Insights:** Most consistent, most improved, best collaboration

Tab 3: Temporal Trends

- **Multi-Axis Time Series:** 3 subplots (volume, quality, efficiency)
- **YoY Growth:** Document and citation growth percentages
- **Trend Analysis:** Linear regression with R^2 and significance testing

Tab 4: Quality Metrics

- **Quality vs Quantity Matrix:** Log-scale scatter with CNCI baseline
- **Collaboration Impact:** Collab-CNCI vs CNCI with OLS trendline
- **Distribution Comparison:** Box plots and violin plots for top 10 countries

Tab 5: Advanced Analytics

- **Statistical Distributions:** Skewness, kurtosis for 4 key metrics
- **Composite Scoring:** Weighted multi-metric ranking (5 metrics, normalized)
    - Weights: Impact Score (30%), CNCI (25%), Top 10% (20%), Citations/Doc (15%), H-Index (10%)
- **Percentile Analysis:** 7 percentiles for CNCI and Citations/Doc

Tab 6: Outlier Detection

- **IQR Method:** 1.5 × IQR threshold for 4 metrics
- **Z-Score Method:** $|z| > 3$ for extreme outliers
- **Visualizations:** Box plots, scatter plots with outlier highlighting
- **Outlier Characteristics:** Geographic and temporal distribution

Tab 7: Statistical Summary

- **Descriptive Stats:** 11 measures for 7 key metrics (includes variance, skewness, kurtosis)
- **Country-Level Aggregation:** Sum, mean, std, min, max by country (top 20)
- **Hypothesis Testing:** One-sample t-test (CNCI vs world baseline 1.0)
- **Correlation Analysis:** Pearson r with significance testing (4 pairs)
- **Trend Testing:** Linear regression on yearly CNCI with R^2 and p-values

5. Statistical Methods

Descriptive Statistics

- Central tendency: Mean, median, mode
- Dispersion: Std, variance, IQR, range
- Shape: Skewness, kurtosis

Inferential Statistics

- **T-Test:** Tests if CNCI differs from 1.0 ($\alpha = 0.05$)
- **Pearson Correlation:** Measures linear relationships with significance
- **Linear Regression:** Trend detection with R^2 and p-values

Outlier Detection

- **IQR:** Q1 - 1.5×IQR to Q3 + 1.5×IQR
- **Z-Score:** $|z| > 3$ threshold

Normalization

- **Min-Max Scaling:** Transforms metrics to 0-1 range for composite scoring

6. Key Visualizations (40+)

Chart Types Used:

- Bar charts (rankings, growth rates)
- Scatter plots (relationships, efficiency)
- Line charts (temporal trends)
- Heatmaps (country-year patterns, correlations)
- Box plots (distribution comparisons)
- Violin plots (detailed distributions)
- Histograms (frequency distributions)
- Area charts (cumulative trends)

Color Strategies:

- Sequential: Blues, Greens, Reds

- Diverging: RdYlGn, RdBu
- Qualitative: Turbo, Plasma, Viridis

7. Performance Optimizations

- **Caching:** @st.cache_data for data loading
- **Sampling:** Max 500-1000 records for scatter plots
- **Lazy Loading:** Visualizations render on tab access
- **Efficient Aggregation:** Pandas groupby operations

8. Key Metrics Explained

Metric	Baseline	Interpretation
CNCI	1.0	>1.0 = above world average impact
Top 10%	~10%	% of highly cited publications
Citations/Doc	Varies	Citation efficiency per publication
Impact Score	Composite	CNCI × % Top 10%
H-Index Proxy	N/A	√(Citations × Documents)

9. Executive Dashboard

6 KPIs Displayed:

1. Total Records (with % of total)
2. Total Documents (in millions)
3. Total Citations (in millions)
4. Average CNCI (with delta from 1.0)
5. Average Top 10% (mean %)
6. Countries (unique count)