Global Research Impact Analysis Dashboard - Technical Documentation

Overview

This is a comprehensive Streamlit-based web application designed for exploratory data analysis (EDA) of global research publications and citation metrics. The dashboard provides interactive visualizations, statistical analysis, and advanced analytics for understanding research impact patterns across countries and time periods.

System Requirements

Required Libraries

- streamlit: Web application framework
- pandas: Data manipulation and analysis
- plotly.express: High-level plotting interface
- plotly.graph_objects: Low-level plotting interface
- numpy: Numerical computing
- scipy: Scientific computing and statistics
- sklearn: Machine learning utilities (MinMaxScaler)

Data Requirements

- Input file: publications.txt (tab-separated format)
- Required columns:
  - Name (country name)
  - year
  - Web of Science Documents
  - Times Cited
  - Category Normalized Citation Impact
  - Collab-CNCI
  - % Documents in Top 1%
  - % Documents in Top 10%

Application Structure

Page Configuration

- Title: "Global Research Impact Analysis"
- Layout: Wide mode
- Sidebar: Expanded by default

Custom Styling

The application includes custom CSS for:

- Gradient headers
- Metric cards with gradient backgrounds
- Insight boxes (blue theme)
- Warning boxes (yellow theme)
- Success boxes (green theme)

Core Features

1. Data Loading and Processing

Function: load_data()

- Loads data from tab-separated file
- Computes derived metrics:
    - Citations_per_Doc: Citations divided by documents
    - Elite_Ratio: Top 1% percentage divided by Top 10% percentage
    - Impact_Score: CNCI multiplied by Top 10% percentage
    - H_Index_Proxy: Square root of citations times documents
    - Productivity_Index: Documents per year since publication
- Uses Streamlit caching for performance

2. Interactive Filters (Sidebar)

Country Filter

- Multi-select dropdown

- Allows selection of specific countries
- Default: All countries included

Year Range Filter

- Slider with min/max values
- Filters data by publication year range

Citation Threshold Filter

- Numeric input
- Sets minimum citation count
- Step size: 10,000

Primary Metric Selector

- Dropdown menu
- Options:
    - Times Cited
    - Web of Science Documents
    - Category Normalized Citation Impact (CNCI)
    - Citations per Document
    - Impact Score
    - H-Index Proxy

Advanced Filters

- Show Outliers: Checkbox toggle
- CNCI Range: Slider for filtering by citation impact range

3. Executive Summary Dashboard

Six key metrics displayed:

1. Total Records: Count with percentage of total dataset
2. Total Documents: Sum in millions
3. Total Citations: Sum in millions

4. Average CNCI: With delta from baseline (1.0)
5. Average Top 10%: Mean percentage
6. Countries: Unique country count

4. Analysis Tabs

Tab 1: Data Overview

Data Quality Report

- Missing value detection and reporting
- Duplicate row identification
- Data type summary (numeric vs categorical)

Key Findings

- Top performer by selected metric
- Most productive year
- Highest average CNCI country
- Most citation-efficient country

Distribution Analysis

- Top 15 countries bar chart (customizable by metric)
- CNCI histogram with statistical annotations (mean, median, baseline)

Correlation Analysis

- Heatmap of correlation matrix
- Includes all key numeric metrics
- Color scale: Red-Blue diverging

Detailed Data View

- Top 100 records by Impact Score
- Color-coded by Impact Score and CNCI
- Sortable columns

Tab 2: Geographic Analysis

CNCI Heatmap

- Countries (rows) vs Years (columns)
- Color intensity indicates citation impact
- Red-Yellow-Green color scale

Research Efficiency Scatter Plot

- X-axis: Document count
- Y-axis: Citations per document
- Bubble size: Total citations
- Top 15 countries displayed

Elite Output Comparison

- Grouped bar chart
- Top 1% vs Top 10% document percentages
- Top 15 countries

Regional Performance Insights Three key metrics:

1. Most Consistent: Lowest standard deviation in CNCI
2. Most Improved: Largest CNCI increase over time
3. Best Collaboration: Highest Collab-CNCI average

Tab 3: Temporal Trends

Multi-axis Time Series Three subplot panels:

1. Publication and Citation Volume
   - Documents (primary y-axis)
   - Citations (secondary y-axis)
2. Quality Metrics Evolution
   - CNCI (primary y-axis)
   - Top 10% percentage (secondary y-axis)

3. Citation Efficiency Trend
    - Citations per document

Year-over-Year Growth Analysis

- Document growth percentage (bar chart)
- Citation growth percentage (bar chart)
- Zero baseline reference line

Trend Patterns

- Linear regression analysis
- Slope and R-squared values
- Trend direction indicators

Tab 4: Quality Metrics

Quality vs Quantity Matrix

- Scatter plot with logarithmic x-axis
- X-axis: Document count
- Y-axis: CNCI
- Bubble size: Citations
- Color: Top 10% percentage
- World baseline reference line (CNCI = 1.0)

Collaboration Impact Analysis

- Scatter plot with OLS trendline
- X-axis: Collab-CNCI
- Y-axis: Category Normalized Citation Impact
- Sample size: 500 records maximum

Distribution Comparison

- Box plots for CNCI distribution (top 10 countries)
- Violin plots for Citations/Document distribution

- Outlier points displayed

Tab 5: Advanced Analytics

Statistical Distribution Analysis Four-column layout displaying:

- CNCI statistics (mean, std, skewness, kurtosis)
- Top 10% statistics
- Citations per Document statistics
- Impact Score statistics

Overall Impact Score Ranking

- Composite score calculation using normalized metrics
- Weights:
  - Impact Score: 30%
  - CNCI: 25%
  - Top 10%: 20%
  - Citations per Doc: 15%
  - H-Index Proxy: 10%
- Top 20 countries displayed
- Detailed breakdown table with color gradients

Percentile Distribution Analysis

- CNCI percentiles (10th, 25th, 50th, 75th, 90th, 95th, 99th)
- Citations per Document percentiles
- Bar charts with world baseline reference

Tab 6: Outlier Detection

IQR-Based Outlier Detection

- Identifies outliers using Interquartile Range method
- Threshold: 1.5 × IQR from Q1/Q3
- Metrics analyzed:

    - Category Normalized Citation Impact
    - Citations per Document
    - % Documents in Top 10%
    - Times Cited
- Displays count and percentage for each metric

Visualization

- Box plot with outliers highlighted
- Scatter plot showing outlier distribution
- Color coding for outlier vs normal points

Z-Score Based Anomaly Detection

- Identifies extreme outliers (|z-score| > 3)
- Displays top 20 extreme outliers
- Outlier characteristics analysis:
    - Countries with most outliers
    - Temporal distribution of outliers

Tab 7: Statistical Summary

Descriptive Statistics

- Complete statistical summary for all key metrics
- Includes: count, mean, std, min, 25%, 50%, 75%, max
- Additional metrics: variance, skewness, kurtosis
- Color-coded heatmap format

Country-Level Statistics

- Aggregated statistics by country
- Multiple aggregation functions (sum, mean, std, min, max)
- Top 20 countries by total citations

Hypothesis Testing

- One-sample t-test against world baseline (μ = 1.0)
- Null hypothesis: CNCI equals 1.0
- Alternative hypothesis: CNCI differs from 1.0
- Displays t-statistic, p-value, and conclusion
- Significance level: α = 0.05

Correlation Analysis with Significance Tests correlation between pairs:

- Documents vs Citations
- CNCI vs Top 10%
- Citations per Doc vs CNCI
- Collab-CNCI vs CNCI
- Reports Pearson correlation coefficient and p-value
- Classifies strength (Strong/Moderate/Weak)

Trend Analysis Over Time

- Linear regression on yearly CNCI averages
- Reports slope, R-squared, and p-value
- Scatter plot with OLS trendline
- Trend significance assessment

Key Statistical Insights Two-column summary:

1. Distribution Characteristics
   - Skewness interpretation
   - Kurtosis interpretation
   - Coefficient of variation
2. Performance Metrics
   - Percentage above baseline
   - Median CNCI
   - 90th percentile threshold

Key Metrics Explained

Category Normalized Citation Impact (CNCI)

- Baseline: 1.0 (world average)
- Values > 1.0: Above-average citation impact
- Values < 1.0: Below-average citation impact

Impact Score

- Composite metric: CNCI × % Documents in Top 10%
- Combines quality and elite output

H-Index Proxy

- Calculated as: √(Citations × Documents)
- Approximates h-index concept

Elite Ratio

- Ratio of Top 1% to Top 10% document percentages
- Indicates concentration of highly cited work

Productivity Index

- Documents per year since publication
- Normalized by time span

Visualization Types Used

1. Bar Charts: Country rankings, growth rates
2. Scatter Plots: Quality vs quantity, efficiency analysis
3. Heatmaps: Country-year patterns, correlation matrices
4. Time Series: Temporal trends, multi-metric evolution
5. Box Plots: Distribution comparison, outlier detection
6. Violin Plots: Detailed distribution shapes
7. Histograms: Single variable distributions

Statistical Methods

Descriptive Statistics

- Mean, median, standard deviation
- Percentiles and quartiles
- Skewness and kurtosis

Inferential Statistics

- One-sample t-test
- Pearson correlation with significance testing
- Linear regression for trend analysis

Outlier Detection

- IQR method (1.5 × IQR rule)
- Z-score method (|z| > 3)

Data Normalization

- Min-Max scaling for composite scores
- Standardization for comparison

Performance Optimization

- Data caching using @st.cache_data decorator
- Sample-based plotting for large datasets (max 500-1000 records)
- Efficient pandas operations for aggregation

User Interface Elements

Color Schemes

- Primary: Blue gradients (#1f77b4, #3b82f6)
- Secondary: Orange (#ff7f0e), Purple (#667eea, #764ba2)
- Status: Green (success), Yellow (warning), Red (error)

Layout Patterns

- Multi-column layouts (2-6 columns)
- Tabbed interface for organized content

- Collapsible sidebar for filters

Interactive Elements

- Hover tooltips on all charts
- Clickable legends
- Zoomable plots
- Downloadable visualizations

Footer Information

Displays:

- Total records analyzed
- Number of countries
- Number of years covered
- Technology stack
- Key feature list

Error Handling

- Try-catch block for data loading
- Graceful error messages
- Application stops if data loading fails
- Null value handling in calculations

Data Flow

1. Load data from file
2. Calculate derived metrics
3. Apply user-selected filters
4. Aggregate data by country/year
5. Generate visualizations
6. Perform statistical analyses
7. Display results in organized tabs

Best Practices for Use

1. Start with full dataset to understand overall patterns
2. Use country filter to compare specific regions
3. Adjust year range to focus on relevant time periods
4. Experiment with different primary metrics
5. Review outliers to identify exceptional cases
6. Check statistical significance before drawing conclusions
7. Use composite scores for holistic assessment

Limitations and Considerations

- Assumes clean, properly formatted input data
- CNCI baseline of 1.0 represents world average
- Derived metrics depend on data quality
- Statistical tests assume normal distribution where applicable
- Outlier detection is sensitive to threshold settings
- Composite scores use predefined weights

Future Enhancement Possibilities

- Export functionality for filtered data
- Custom metric weight adjustment
- Additional statistical tests
- Machine learning-based predictions
- Real-time data updates
- Comparative benchmarking tools
- Citation network analysis
- Geographic mapping visualizations