Visualization Tool using CIA Dataset

Harshavardhan Dharman
Eindhoven University of Technology
h.dharman@student.tue.nl

Surya Kannan
Eindhoven University of Technology
s.kannan@student.tue.nl

Shashank Venkatesha
Eindhoven University of Technology
s.venkatesha@student.tue.nl

Leela Karthikeyan Haribabu
Eindhoven University of Technology
l.k.haribabu@student.tue.nl

1 Introduction

The visualization tool developed for this project is intended for students, analysts, and professionals who want to explore global patterns across multiple national indicators. These users may be interested in understanding how countries differ in terms of economic stability, demographic trends, transportation infrastructure, communication access, governance characteristics, and energy capacity. The goal of the tool is to enable exploration and analysis of an integrated interface built from several CIA-sourced datasets, supporting users in identifying meaningful patterns and communicating insights effectively.

Because the dataset spans diverse dimensions such as GDP growth, literacy rates, transportation networks, communication penetration, land characteristics, and energy resources traditional tabular formats make comparison difficult. A visualization tool is therefore essential for presenting this wide-ranging information in a coherent and comprehensible way. Users can interact with the data by focusing on the domains most relevant to their questions, enabling targeted exploration that is not easily achievable through static tables or isolated statistical outputs.

By bringing multiple datasets together into a unified interface, the tool supports cross-domain exploration even though the datasets remain separate at the data level. Visualization is thus a natural and effective approach for understanding complex, multidimensional global data and supporting exploratory, analytical, and presentation-oriented tasks.

2 Data Analysis

2.1 Domain Data Specification

The visualization tool is built around seven independent CIA-derived datasets, each representing a distinct domain of national-level information: economy, demographics, transportation, communications, geography, government and civics, and energy. Every dataset consists of country-level attributes describing different aspects of global structure and development. Since each domain contributes its own theme such as economic performance, population characteristics, infrastructure availability, land distribution, technological access, or political organization the datasets provide rich material for domain-specific exploration. In the tool, users select the domain of interest, and the corresponding dataset is loaded directly into the interactive visualizations. This design allows focused and uncluttered analysis within each domain and combining attributes across multiple datasets for multi-domain analysis.

2.1.1 Preprocessing

Preprocessing in the tool primarily consisted of standardizing the datasets into a format suitable for visualization. The CSV files were loaded individually, and several columns were cleaned or renamed to ensure consistency across the domain views. In particular, attribute names were converted into a uniform lowercase, underscore-separated format to avoid mismatches during filtering and plotting. Numeric fields such as GDP measurements, population values, transportation distances, pipeline lengths, and energy production figures were verified to be numeric and converted when required so they could be used directly in scatter plots and correlation matrices.

No merging of datasets was performed. Instead, each dataset remained separate and was accessed independently depending on the user's chosen domain. This preprocessing strategy preserves the native structure of each dataset while ensuring they integrate cleanly with the callbacks used for filtering, map visualization, scatter plots, and correlation analysis.

2.1.2 Missing Values

Several attributes across the datasets contain missing values, which is expected given the uneven availability of global country-level statistics. Missing values were not imputed or removed globally; instead, the tool handles them dynamically according to the specific visualization. ~~For example, correlation matrices drop rows with missing numerical values only for the matrix being computed, while scatter plots omit a data point only if it lacks either of the selected comparison attributes. Countries with missing values in non-numeric fields, such as government type, capital, or coastline length, remain in the dataset and continue to appear in visualizations that do not require those attributes. Because the map visualization relies solely on the country name for location matching, countries with partial data are retained and still appear on the world map. Each dataset is therefore used in as complete form as possible without removing entire countries unnecessarily~~.

2.1.3 Interpretation of Attributes

The datasets collectively describe a wide range of meaningful domain-specific attributes. Economy-related data includes GDP estimates, unemployment rates, budget indicators, trade figures, and poverty levels, enabling users to explore national economic structures. The demographics dataset contains population size, birth and death rates, literacy levels, fertility rates, and median age, offering insight into population and social characteristics. Transportation attributes quantify infrastructure such as roads, railways, airports, pipelines, and waterways, while communications data reflects the technological penetration of telephone lines, mobile subscriptions, internet usage, and broadband access. Geographic attributes describe land area, water area, elevation extremes, agricultural land, forest coverage, and coastline length, providing environmental and spatial context. Government and civics data identifies capital cities, coordinates, government systems, and suffrage information. Finally, energy attributes describe electricity access, generation capacity, fossil fuel production, natural gas quantities, and carbon emissions. These variables define the analytical scope of each domain and guide the types of patterns users can meaningfully explore.

2.2 Data Abstraction

Each dataset is abstracted into general data types to support visualization and exploratory analysis. Numerical attributes, such as GDP growth, infant mortality, railway length, mobile subscriptions, land area, and CO_2 emissions, are treated as quantitative variables suitable for comparison, correlation analysis, and multivariate views. Categorical attributes, including government type, capital city, and internet country code, are used for filtering or grouping tasks. The unifying identifier across all datasets is the country name, which allows consistent interaction with the geographical map even though the datasets are not merged.

These abstractions enable the tool to support a wide range of exploratory tasks such as retrieving values for a specific country, comparing countries within a domain, identifying outliers, detecting broad trends, and examining numerical relationships. Because each dataset is explored independently, users gain a clear and domain-focused understanding of global patterns without the complexities that arise from forced dataset integration. This abstraction layer ensures clarity, flexibility, and coherence across all domain views.

3 Task Analysis

3.1 Domain Specific Tasks

The main purpose of our visualization tool is to provide analysts an effective means to view CIA statistic datasets through a visualized medium and turning them into clear insights. The visualization program will address three levels of data analysis for seven different domains of data (Economy, Energy, Demographics, Communications, Geography, Transportation, and Government & Civics).

The low-level analytical task is characterised by the primary goal of obtaining simple data retrieval and identifying geographic locations. When using low-level analytical tasks, an analyst would access a Choropleth Map based on a single metric, such as the percentage of electric energy services in the world (by Energy) and begin to analyze the global distribution of that metric. The most basic level of interaction with the Choropleth Map would be to select a country (e.g., Angola) and initiate the Details Panel. The Details Panel will provide instant validation of data across multiple domains by showing the numerical information for several of the country's key indicators (e.g., the country's Total Population (by Demographics), the length of its roadways km (by Transportation), and the Official Exchange Rate GDP (in billions of USD) (by Economy) instantaneously, giving analysts the opportunity to quickly fact-check the information and understand it within a larger contextual framework.

Mid-level tasks involve comparing two or more metrics to identify similarities, trends, or correlations. A scatter plot helps reveal relationships between two variables. For example, plotting total unemployment rate against youth unemployment rate to understand how overall economic conditions affect young people. When a third factor is important, a bubble chart extends this comparison by encoding an additional variable through bubble size. This allows analysts to see not only how two measures relate, but also how they vary with population size, GDP, or any other key attribute. Together, scatter and bubble charts provide an intuitive way to explore multi-variable relationships and uncover deeper patterns in the data.

High-Level tasks involve sophisticated Synthesis and Clustering to uncover holistic national profiles within a single, complex domain. These tasks utilize multivariate analysis to combine multiple indicators from the same dataset. For example, using the PCA + K-Means analytical view, an analyst could understand and feed in purely about the Economy indicators like Real GDP Growth Rate percent, Unemployment Rate percent, and Public Debt percent of GDP to mathematically group countries into underlying clusters (e.g., "High-Growth Developing Markets" or "Highly Indebted Fragile States") based on their overall economic signature. Alternative multivariate views, such as Parallel Coordinates, allow for the simultaneous comparison of specific Demographics metrics like Birth Rate, Death Rate, and Total Fertility Rate across a manually selected set of countries, enabling a rapid, nuanced comparison of population dynamics and trends.

This tool enables users to examine how globally located metrics are dispersed across the world, compare two different metrics against each other in order to identify unique or unusual relationships, and also determine the primary trends across all metrics at the same time. Essentially, it is a complete method for transitioning back and forth between an overall view of the world to specific data related questions about how national metrics interact with one another.

3.2 Task Abstraction

The Task Abstraction for this project systematically translates the analyst's high-level goals into the minimal, generalized operations the visualization can support. The Task Abstraction framework is based upon the Action (what users do) and Target (what data they are examining) of users and has ensured that the visualization design is robust. The highest priority for the Task Abstractions are to allow users to Correlate data to discover previously hidden dependencies between two different data fields. Additionally, users should have the ability to derive new information by consolidating the complete multivariate distribution into different clusters of countries (using PCA) and finally, every visual element must support interaction tasks (i.e. Filter (to filter the dataset), Locate (to search for countries on a map), and Retrieve Detail (to get the actual number details)), this means that users can seamlessly transition from high-level overviews to detailed inspections of data within the same tool.

Table 1: Task Abstraction based on Munzner's Visualization Model

Action	Target	Justification / View Implication
Correlate	**Dependency**	~~Primary for **Scatter Plot**; links two attributes to find relationships~~.
Derive	**Clusters**	~~Supported by **PCA** mode; calculates new variables to **summarize** data~~.
Compare	**Extremum**	~~Supported by **Bar charts**; assesses quantitative differences between countries~~.
Filter	**All Data**	~~Supported by **Map selection**; reduces visible set based on~~ criteria.
Locate	**Items**	~~Supported by **Map View**; finds position of a country or its coordinates~~.
Retrieve Detail	**Single Item**	~~Supported by **Details Panel**; provides raw numerical values for one country~~.

4 Current Solution (HOW)

The current solution for this project is a linked, interactive dashboard that allows analysts to explore seven diverse datasets from the CIA World Factbook, moving beyond simple data lookups to complex analytical discovery. The tool integrates several coordinated visual components: a world map, a scatter plot, a PCA view with clustering, a cross-domain comparison view, and a bubble chart. These views are linked through interaction so that selections and filters in one component are reflected in the others.

The world map provides a quick overview of a single quantitative attribute in its geographic context (e.g., where land area or population is most concentrated). A choropleth map is used because it directly leverages spatial position (country location) and color intensity to reveal regional patterns and spatial clusters, which are

important for many of the domain-specific tasks. For pairwise analysis, the scatter plot is the core view: it enables users to compare two quantitative indicators (such as economic output vs. mobile phone subscriptions) using position on a common scale on both axes, one of the most accurate channels for judging magnitude and correlation. This supports tasks such as detecting trends, correlations, and outliers defined in the task analysis.

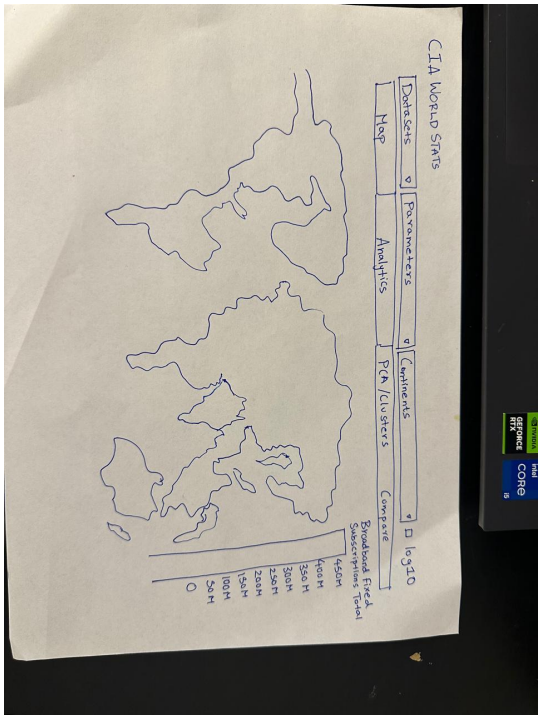
The PCA (Principal Component Analysis) mode synthesizes selected attributes from the datasets into a two-dimensional embedding that summarizes global differences between countries. In this view, countries that have similar multivariate profiles appear close together, which helps users identify natural clusters or atypical countries. Clustering (e.g., via k-means) further groups countries into interpretable categories that support high-level tasks such as characterizing "development profiles" or comparing groups of countries. The *Compare* functionality is supported by a dedicated ranking view (e.g., a bar chart), which uses aligned length and position to allow precise side-by-side comparison of selected countries on a chosen metric.

For comparing cross-domain attributes of countries, a bubble chart is implemented. It shows three indicators per country at once: the x-axis, y-axis, and bubble size. It extends a standard scatter plot by adding a "magnitude" dimension via bubble area, enabling users to examine trade-offs (for example, between economic, health, and population features) in a single visual. This directly supports tasks that involve balancing multiple criteria and identifying countries that stand out along several dimensions at once.

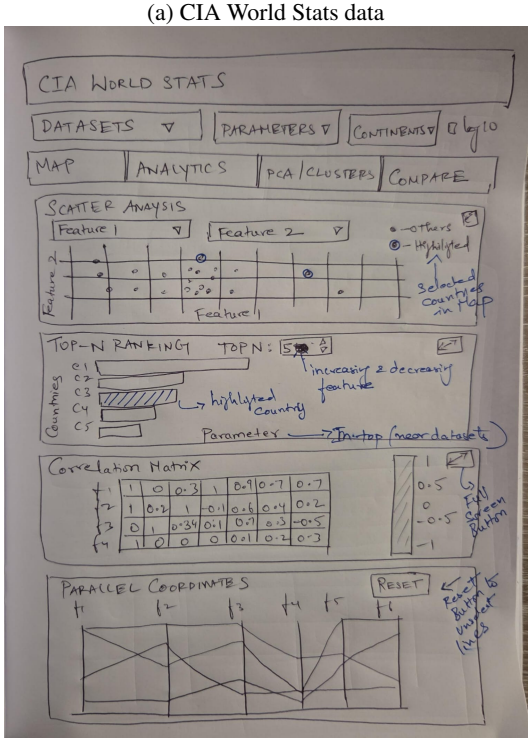
During the design process, we considered alternative encodings such as a large heatmap matrix (countries $\times$ indicators) and radar charts for country profiles. We discarded these options because a heatmap with hundreds of countries becomes visually cluttered and difficult to interpret, and radar charts are known to hinder accurate comparison of values across items. Instead, we chose maps, scatter plots, bar charts, and PCA-based views, as these better align with our abstract tasks (compare, correlate, cluster, filter, retrieve detail) and make use of perceptually effective visual channels for quantitative data.

5 Implementation

The tool is implemented entirely in Python, using Dash for web-based UI construction and Plotly for interactive visualizations. Data cleaning employs Pandas and regular-expression–based preprocessing. ISO code and continent inference use external libraries such as pycountry, pycountry convert, and gapminder. To achieve Multi View coordination, Dash creates Callbacks that link the state (i.e., clickData, relayoutData, and dcc.Store) of the respective components together. The technical challenges we encountered while developing the system included inconsistent Country names from the CIA database; Missing or Non-numeric; Handling large heterogeneous datasets; Maintaining a uniform display of a stable map (as it would be used from interaction through either brushing methodology); Linking map and scatter selection using ISO-code matching; Designing a Dark-Themed Responsive UI with animated transition, Full-Screen modal, and Vertical Scrollable Containers. The final system integrates all components cohesively while remaining modular for future extension.

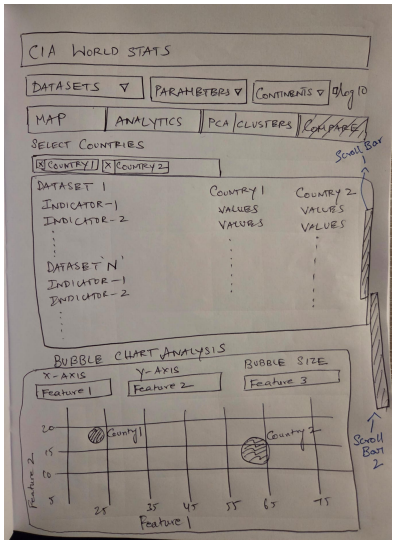


(a) CIA World Stats data

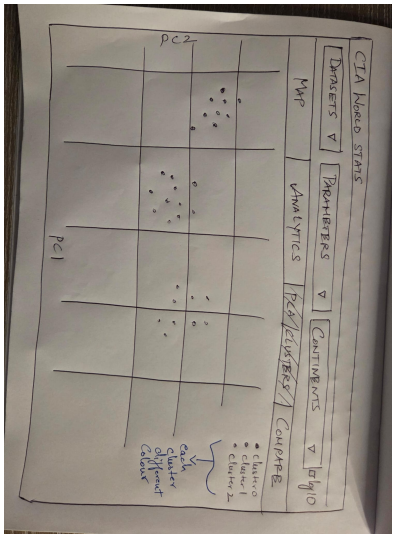


(b) Analytics Attribute

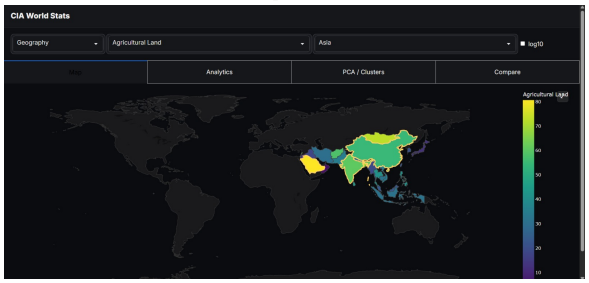
Figure 1: Design Sketches for the Visualization Tool



(a) Compare Attribute



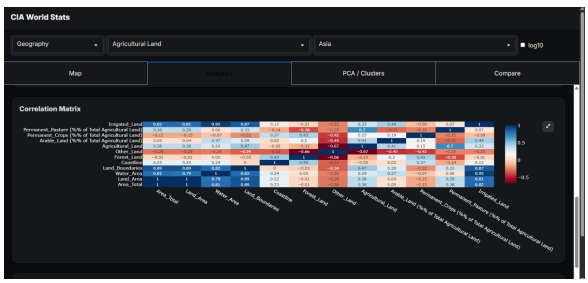
(b) PCA Clusters Sketch



(c) World Map View

(d) Scatter Plot Analysis

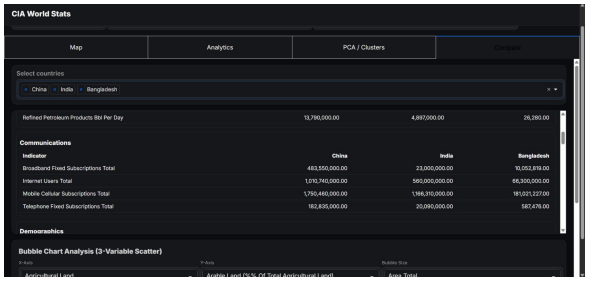
(e) Top N Ranking



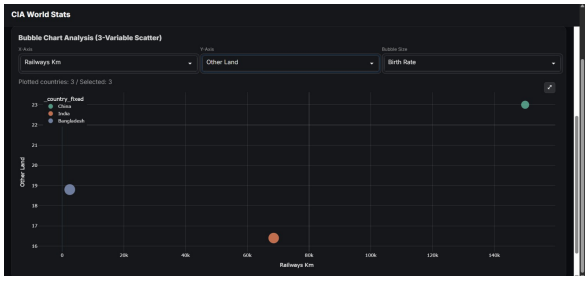
(f) Correlation Matrix

(g) Parallel Coordinates

(h) PCA Clustering



(i) Country Comparison



(j) Bubble Chart

Figure 2: Project Workflow Screenshots