Exploratory Visual Analysis of Multivariate Correlations in Global Demographic Data

Alice Kang, Minju Park

1 Introduction

Visual analysis of correlations is relatively straightforward when users have a specific target question in mind, such as examining a relationship in scatterplot or parallel coordinates. However, in exploratory visual analysis—particularly when dealing with multivariate data—the process becomes more complex. When users wish to discover which variables are correlated, they face a combinatorial explosion, where they must manually select pairs of variables and test their relationships one by one. Moreover, multivariate exploratory analysis often involves multiple levels of correlation inspection, where users not only examine relationships between pairs of variables but also explore how these relationships vary across different entities. For instance, when exploring large global datasets with diverse social indicators, discovering meaningful relationships requires extensive manual exploration. Gapminder [1] provides an intuitive interface to visualize such relations across different countries once specific variables are selected, but the discovery process still relies heavily on user intuition, requiring users to manually choose variables to compare.

Previous research has proposed various methods to support this kind of exploratory task. For example, *Voyager* [19] offers visualization recommendations to assist users in exploring multivariate tabular data. Later extensions of *Voyager* [20] have further blended manual and automated chart specification, enabling analysts to engage in both open-ended exploration and targeted question answering. Building on these ideas, our work aims to extend this line of research to focus more directly on discovering correlations and relationships across many variables.

In this project, we aim to develop an interactive visualization tool to support exploratory correlation analysis. The system will help users efficiently identify potentially correlated variable pairs or groups without requiring exhaustive manual trial-and-error, combining computational correlation detection with intuitive visual representations and user interaction. Moreover, we aim to address exploratory tasks that involve multiple levels of correlation analysis, enabling users to examine how one focal attribute co-varies with multiple explanatory attributes across entities and to reveal patterns both among variables and among entities.

In terms of our personal expertise, both of our team members have backgrounds in HCI design software development (e.g., JavaScript, React), though we do not have prior experience specifically with visualization tools like Vega-Lite. Our initial motivation came from an interest in the decline of birth rates in the Four Asian Tigers, and we wanted to explore what other factors might be related. From there, our focus expanded to supporting this kind of exploratory analysis more broadly—enabling users to investigate relationships and uncover interesting patterns across different country groups.

To demonstrate and test our tool, we plan to apply it to global datasets that integrate diverse social indicators such as education level, cost of living, and immigration rate. The tool is designed to visualize meaningful correlations among these indicators and to enable users to further explore and interpret how such relationships vary across countries. This scenario involves both levels of correlation analysis examining relationships among variables (social indicators) and comparing how these relationships differ across entities (countries), making it well-suited for exploratory correlation analysis.

In this project, we focus on this case to demonstrate how the tool supports exploratory correlation analysis in practice. We illustrate specific use cases that address questions such as "What other indicators are correlated with the population decline in the Four Asian Tigers?" or "What other countries show a similar trend with the Four Asian Tigers?" Based on our system design, we will also outline the user flow, showing how users interact with the tool to identify, compare, and interpret potential correlations. Finally, we will conduct a preliminary evaluation with potential users to assess the tool's effectiveness in supporting exploratory correlation analysis and to gather feedback for future improvements.

Manuscript received xx xxx. 201x; accepted xx xxx. 201x. Date of Publication xx xxx. 201x; date of current version xx xxx. 201x. For information on obtaining reprints of this article, please send e-mail to: reprints@ieee.org. Digital Object Identifier: xx.xxxx/TVCG.201x.xxxxxxx

2 Related Work

2.1 Correlation Visualization

Correlation analysis plays a central role in uncovering relationships among variables within multivariate datasets. However, because correlations are defined only between pairs of variables, understanding the overall correlation structure becomes increasingly difficult as dimensionality grows. As the number of variables increases, important relationships can easily be overlooked, motivating the need for visualization techniques that effectively communicate complex correlation patterns.

A wide range of visualization methods have been developed to support correlation analysis, each offering distinct strengths and weaknesses. Scatterplots remain one of the most widely used representations for visualizing pairwise correlations. Numerous studies have examined how users perceive correlations in scatterplots. For instance, perceptual biases can arise from the scaling of axes, where correlations appear stronger when aspect ratios are increased [6]. Building on this, an automatic method was developed to select effective scatterplot aspect ratios, integrating participant preferences to enhance perceptual accuracy [7]. Other studies have compared the perception of correlation across different visualization types, finding that scatterplots generally lead to more consistent subjective judgments than parallel coordinates [9].

Beyond scatterplots, prior work has shown that perception of correlation varies significantly across visualization types, and that visual judgments often differ for positive and negative correlations. For example, studies ranking visualization types using Weber's law demonstrated systematic differences in how people interpret correlation strength depending on the visual encoding and correlation polarity [8]. These findings underscore the perceptual and cognitive challenges involved in designing correlation visualizations that are both accurate and interpretable.

It is also crucial to recognize that correlation does not imply causation. While correlation visualizations help reveal associations among variables, they do not explain the underlying mechanisms or causal directions that produce those relationships. Misinterpreting correlations as causal can lead to misleading insights, particularly when exploring complex, real-world datasets [17]. With these considerations in mind, through this project we aim to develop a visualization system that supports visualization of multivariate correlations, helping users identify and compare relationships across multiple variables while remaining aware of the limits of correlation-based reasoning.

2.2 Exploratory Visual Analysis

A variety of visualization tools have been developed to support data exploration. These efforts align with prior work on designing systems that facilitate exploratory data analysis, a process in which users interactively analyze and summarize datasets to identify patterns, trends, and relationships, as well as research on exploratory search, where users iteratively refine their information-seeking goals.

Exploratory visual analysis refers to an open-ended process of engaging with data through visualization, often without a predefined goal [4]. In this stage, users may be unfamiliar with the dataset and uncertain about what to look for. The process can involve identifying questions of interest, inspecting visualized data, and iteratively refining one's questions and hypotheses as new insights emerge. Such exploration is a crucial step that allows users to form an overall understanding of the data and obtain serendipitous insights. However, the open-ended nature of this process also presents challenges for system design, as tools must balance flexibility with effective guidance and avoid overwhelming users with excessive options.

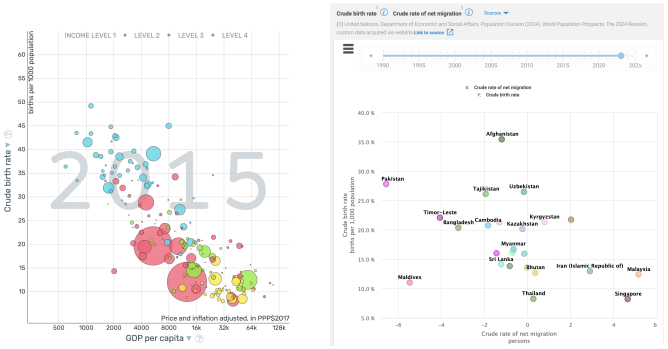
Several studies have investigated how to support open-ended browsing in visual exploration. For example, faceted browsing has been introduced to enable users to navigate large information spaces by progressively filtering data along multiple dimensions [21]. In the information visualization domain, *Voyager* applied this concept to facilitate exploratory visual analysis by recommending a diverse set of visualizations based on statistical and perceptual measures, effectively supporting faceted browsing of charts [19, 20].

Exploratory visual analysis also carries several risks. When users explore a large number of visualizations, the likelihood of encountering spurious patterns increases—a phenomenon known as the multiple comparisons problem [5]. As the number of comparisons grows, users may mistakenly interpret random variation as meaningful structure [22]. Additionally, perceptual biases may lead users to perceive visual patterns that do not reflect the underlying data. To address such risks, graphical inference methods have been proposed to help analysts test whether perceived visual patterns are statistically significant, providing more rigorous evidence for visual findings [18]. With these insights in mind, through this project we aim to enable exploratory visual analysis of multivariate correlations, helping users navigate, compare, and interpret relationships among multiple variables in an open-ended yet structured manner.

2.3 Demographic Data as a Domain for Exploratory Correlation Analysis

Demographic indicators such as birth rate, education level, cost of living, and migration rate are deeply interrelated, often influenced by social, economic, and cultural factors that cannot be explained by a single variable. These multidimensional dependencies make demographic data a particularly suitable context for open-ended exploration, where analysts must iteratively test, compare, and interpret relationships to uncover meaningful insights. Moreover, demographic analysis typically involves reasoning across multiple levels of correlation—examining not only relationships among variables but also how these relationships differ across entities such as countries or regions. Depending on how analysts frame their comparisons or aggregate data, multiple and sometimes contrasting insights can be elicited [13], underscoring the interpretive flexibility inherent in this domain. Such variability highlights the need for visualization tools that help users navigate entangled relationships and identify emerging patterns. In this sense, demographic datasets provide a compelling use case to evaluate how our system supports discovery-driven, multi-level correlation exploration.

A number of visualization systems have been developed to support exploration of demographic and population data. One of the most well-known examples is Gapminder [1], which allows users to select any two indicators as the x– and y–axes and visualize their relationship through animated scatterplots that show how these variables change over time. Gapminder's simplicity and interactivity make it highly accessible, yet it relies heavily on users' intuition in selecting which variables to compare. Discovering meaningful relationships thus often requires trial-and-error, as the system provides little computational or visual guidance for identifying potentially interesting variable pairs.



(a) Gapminder (b) UN Data Portal

Fig. 1: Existing tools for exploring demographic data

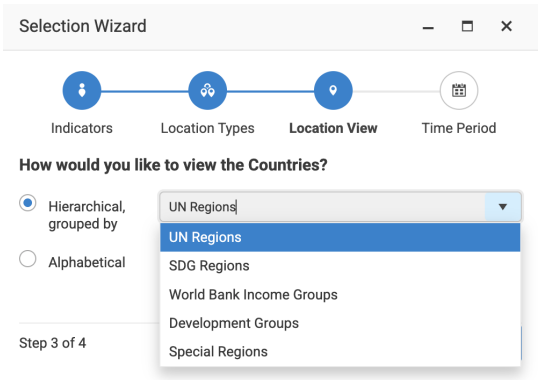


Fig. 2: *Selection Wizard* feature included in the UN Population Division Data Portal

More recently, the United Nations (UN) Population Division Data Portal [16] has introduced a flexible web interface for exploring demographic indicators across countries and regions. Fig 1 shows these two existing tools for demographic data exploration. The portal enables users to choose indicators, location types, and time periods, and even visualization idioms such as line charts, bar charts, and maps. A built-in *Selection Wizard* further guides users by grouping options under thematic categories (e.g., Population, Fertility, Mortality for indicators; UN regions, Development groups for locations), as illustrated in Fig 2. While these structured menus offer some support for navigating the vast space of demographic data, the exploration process can still be overwhelming. Selecting visualization idioms, configuring parameters, and interpreting relationships demand a certain level of visualization literacy and often a steep learning curve.

Given the inherent complexity and interdependence of demographic variables, there remains substantial room to improve how such systems support exploratory correlation discovery. Our project builds on these existing tools but shifts the focus from variable-by-variable inspection to systematic exploration of multivariate correlations, combining computational analysis with interactive visual interfaces to help users uncover meaningful relationships more efficiently and interpret them more effectively.

3 Data and Task Abstraction

3.1 Dataset Information

We leverage open-source data from *Gapminder* [1], a non-profit organization that collects global demographic data from various sources, and compiles it into an interactive tool shown in Figure 1. Most of Gapminder's data is collected from global intergovernmental bodies such as the United Nations, or its affiliates such as the World Health

Organization, and summarized by geographers and engineers on their team. For historic data prior to the date for which it is publicly available, the Gapminder team transition between their own historic estimates and official reports from global organizations [1]. Notably, the team also uses the UN World Population Prospects [15] to generate data up to the year 2100. For data that is not publicly available from global organizations, the team scrapes data from for-profit organizations such as Forbes, for example. As a result of aggregating across these heterogeneous sources, Gapminder provides a rich dataset of not only basic demographic data such as total fertility rate or gross domestic product per capita, but also specific and rather esoteric statistics such as number of people made homeless by natural disasters, or percentage of people who agree/disagree on the necessity of vaccines. Table 1 summarizes the broad categories of data we pull from Gapminder. For each category, we recorded the description, example datasets, and the number of datasets within the category.

3.2 Data Abstraction

Each dataset is formatted as a table, with countries as items and years as columns. Each cell of the table is populated with the value of the given indicator, for its respective country in the respective year. The dataset contains quantitative data, with a mix of continuous (e.g., percentage of population below the poverty line) and discrete (e.g., number of children enrolled in primary education), depending on the indicator. As year is one of the key attributes of the tables, we concluded we are working with time-series data. Due to the large number of datasets we are aggregating, we did not conduct a full data abstraction analysis of all 568 datasets. However, we observed the following three broader types of data across all datasets: **percentages**, **raw numbers**, and **ratios**. We conducted a smaller analysis for each of these three types using example datasets that may represent distinct instances of the type.

Percentage data describing a quantity (e.g. plastic percent of waste composition) fall within a possible range of [0, 100%]. However, when describing the growth/decline of a measure (e.g. population growth), the possible range is [-100, 100]. Realistically, plastic waste or population are not factors that increase drastically over one year. Therefore, their actual ranges are [1.45, 26.2] and [-8.42, 7.32], respectively. The actual range of data is heavily dependent on the indicator, and therefore we emphasize that these are examples of what we classify as percentage data.

Raw number data describes the actual count of the indicator, not compared against a total of any kind. An example dataset of this type is murders, containing absolute numbers of murder cases in countries over time. The range of this dataset is [1.2, 68200], which is quite large given the unique political climates and general safety of each country.

Ratios are like a combination of percentage and raw data, as the measure represents a proportion like a percentage, but is divided by a relative, dynamic quantity given what the indicator aims to represent. Any indicators measured per capita is a good example of this data type, as the raw number of an indicator is divided by the country's population at that year. Like raw numbers, however, the potential range for this type can be drastically different across indicators without a shared unit. GDP per capita, for example, has a range of [373, 925000], measured in international dollars fixed to 2017 prices.

Before moving ahead with the complete collection of datasets, we first decided to conduct a short exploratory data analysis to narrow down the final set of indicators we will use as the source of our visualization.

3.2.1 Exploratory Data Analysis

While examining the data, we noticed that many of the available datasets were quite sparse. As the objective of our tool is to support correlation analysis, we took a rather audacious approach in filtering out sparse datasets. First, we filtered out datasets that contained less than 15 columns, as many of these datasets were hyper-specific indicators that were collected on a select number of years. Examples of these datasets include total number of dollar billionaires, average number of bad teeth per child, and number of battle deaths. Each of the aforementioned datasets only contains data from 2002 up to 2007. For basic demographic indicators such as life expectancy, the average range of years for which data is available is 1800 to 2100. We concluded that a meaningful correlation relationship cannot be drawn between datasets of less than 15 versus 300 attributes, and therefore dropped these datasets completely.

When both the total number and per capita average were available for a single indicator, we discarded the total number dataset. Per capita, meaning "for each head", is an average measure of an indicator per person. As an average, this normalizes the data and allows for the population size to be incorporated when comparing statistics of multiple countries. Selecting only the per capita indicators with more than 15 attributes, we reduced the total number of datasets from 568 to 472.

Next, we calculated the percentage of populated cells for each indicator, and examined the distribution of data if all N/A values were to be filtered out. Figure 3 show the datasets sorted in increasing order of proportion of non-NA values calculated against the total number of values in the table. To our surprise, we found that 267 out of 472 datasets had more than 90% of its cells populated with non-NA values. Based on this observation, we decided to select only the datasets with more than 70% of its cells populated with statistics, leaving us with 319 datasets in total. It is important to note that each dataset may have a different number of cells, based on the availability of the data. This means that not all datasets have the same number of year attributes and country items.

Lastly, we added back some indicators that are known to be correlated with population decline in order to fully investigate our motivating example of population decline in rapidly developed countries, such as South Korea, Hong Kong, Singapore, and Taiwan (otherwise known as the 4 Asian Tigers). Across the world, indicators correlated with population are notably female literacy rates, female education levels, female age at marriage, demand for family planning, and immigration/emigration rates [2, 3, 11, 12, 14]. Given that these indicators are studied to be relevant to our motivating example, we added them out of curiosity of whether their known relevance will also lead them to display a strong correlation with population and fertility rate. After adding these datasets back into our source data, we obtained a resulting data source of 332 datasets.

3.3 Task Analysis

3.3.1 Who

Our target users are geographers, global studies scholars, economists and researchers in the social sciences. The purpose of our tool is to narrow down the broad use cases of Gapminder's interactive tool to allow for a more specific correlation analysis among various demographic indicators provided by Gapminder. While Gapminder offers a great tool for exploring the large amounts of data their team has collected over time, we are specifically interested in visualizing the correlation between demographic indicators, which can provide social scientists with a starting point to new research questions, studies, and hypotheses. Outside of these core users, we can imagine uses cases of our tool in applications of machine learning algorithms. Understanding the basic correlation between demographic indicators can surface global phenomenons in which new machine learning algorithms can be leveraged in investigating. Lastly, the everyday user may find our tool interesting, especially when paired with Gapminder's pre-existing interactive tool. Our solution may open doors to new ways of playing around with global demographic data, and enhance the understanding of relationships between and across different countries and indicators for the everyday user.

3.3.2 Action

For many users, their user goals are within the *Analyze* level, based on the levels of actions defined in [10]. In consuming the available data, they may *discover* new correlations between specific indicators of a single country, and different countries with similar correlations between those indicators. As mentioned in Section 3.3.1, this may introduce the grounds to new hypotheses and studies for social science researchers. Everyday users may have the *enjoy* goal, as they interact with the tool driven by their own curiosity. Relationships identified by

Category Name	Category Description	Number of Datasets
Communication	Data related to a population's use of wired, online, and physical communication channels (i.e., number of cell phones per capita, number of Internet users, number of missing journalists).	15
Economy	Data related to a country's economy (i.e., amount of annual aid given/received, poverty rates, and Gross domestic product).	88
Education	Data related to a population's education (i.e., children/teen/adult literacy rate, mean years in school across men versus women, and schooling cost).	38
Energy	Data related to a country's energy sources (i.e., coal consumption per capita, oil production per person, and number of natural gas reserves).	25
Environment	Data related to a population's relations to climate change (i.e., number of oil spills, deaths from natural disasters, and forest products removal).	62
Health	Data related to a population's health (i.e., child mortality rates, life expectancy, number of HIV cases).	130
Infrastructure	Data related to a country's public infrastructure (i.e., basic sanitation access rates, number of roads paved, basic water source access rates).	10
Population	Data about the population itself (i.e., population density, median age, sex ratio across different age groups).	82
Society	Data related to the societal structure of a population (i.e., corruption perception index, happiness score, number of immigrants/emigrants).	60
Sustainability	Data related to the perceived sustainability of a country (sustainable development index).	1
Work	Data related to the work done by the population (i.e., employment rate by sector, sex, forced labour rates, unemployment rates).	57

Table 1: Categories into which all Gapminder data is organized

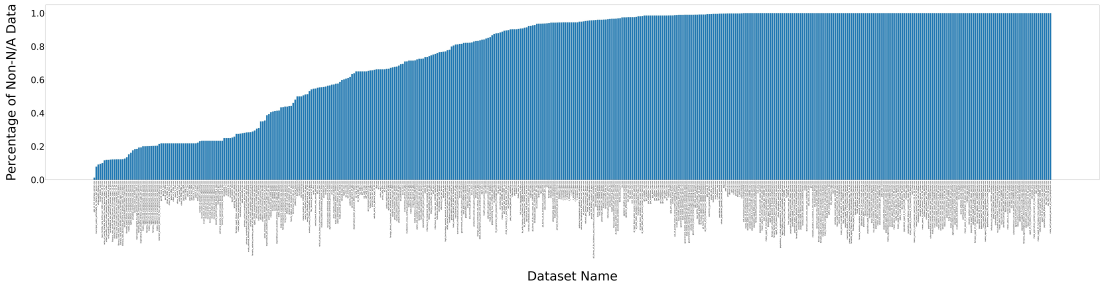


Fig. 3: Distribution of non N/A data contained in each dataset

the visualizations may then pique their curiosity to continue exploring different countries and/or indicators using the tool.

At the *Search* level, the user may have a *lookup* goal. For users with a specific question in mind, as narrow as "what demographic indicators are correlated with the population decline in the 4 Asian Tigers?", they will directly populate the country and indicator of interest to answer their question. Alternatively, the user may perform a *browse* in the case where they have determined only the country/group of countries or indicator(s) of interest. If the previous user was interested in general demographic correlations in the Asian Tigers but do not know which indicators they are specifically interested in, they may be more suited to a *browse* goal. If neither of the two variables are determined, the user may *explore*, and start with a broad overview of the data used in the tool. As in many cases, a *browse* or *explore* goal may easily blend into a *lookup* goal as the user starts to narrow down their search parameters.

Finally, users may have *identify* and *compare* goals at the *Query* level. After the user performs a *lookup* related to a specific question in mind, they can then further understand the characteristics of the resulting correlation (such as its strength) through *identify*. The *compare* goal can be used between correlations between indicators of interest across different countries/groups of countries, to generate insight not limited to a single search space.

3.3.3 Targets

The main targets of our tool are *trends* and *correlations*. Like in Gapminder's interactive tool, we will display the trends in a single demographic indicator across time following the user's selections. For users with goals in the *Analyze* level, this may identify specific indicators that show a clear increasing/decreasing trend across the years, which may then prompt them to further question other indicators that are in correlation with this variable of interest. *Correlations* are the main target we wish to highlight in our tool, which also addresses the gap in pre-existing data visualizing tools. Our main objective is to display various correlations between user-inputted demographic indicators in select countries, so the users can easily understand the relationships between global demographic statistics.

3.3.4 Attributes

The three broader types of data described in Section 3.2 define the attributes which we will focus on in our task. Additionally, we will include a derived attribute of *Similarity Score* to compute the Pearson correlation between a pair of individual countries or indicators, or between a single country/indicator to a group of pre-defined countries/indicators.

Percentages Percentages represent the amount of an indicator compared against a total of some kind. This attribute is convenient for expressing the changing trends or distributions of an indicator over time. Re-iterating the example from Section 3.2, plastic percentage of waste composition is an example of this attribute, where 0% indicates no plastic in the total waste produced by a country, and 100% indicates the entirety of the waste is composed of plastics.

Raw Numbers Raw numbers represent the actual measures of a given indicator, through which users can easily identify the outliers in a given dataset. Especially within a sparse dataset, high values of this attribute highlight a specific country or country group to stand out within the global context. In the example from Section 3.2, it is straightforward for users to identify the countries experiencing higher numbers of murder cases, such as Brazil with numbers consistently in the thousands, among countries like Slovenia with numbers in the tens.

Ratios Ratios describe a proportion of a given indicator against a relative and dynamic total. Specifically within this dataset, many indicators are expressed as a ratio *per capita*, which represents the proportion of an indicator in a country against the population of that country in a given year. Like percentages, this attribute is suitable for expressing trends and distributions, but is also a good comparative measure to highlight extremes across the globe. The example used in Section 3.2 is GDP per capita. Following a single country over time, users can clearly read the trends in its economy's growth/decline. When comparing two countries, users can easily identify which countries have a much greater GDP compared to others, but will need additional information to understand whether this difference in GDP is due to the difference in their population sizes, or the actual market values of their products.

Similarity Score To help users interpret existing correlations and explore new ones, we compute similarity scores that measure how closely countries and indicators align. These scores support users to identify countries that exhibit similar indicator trends or indicators that behave similarly across countries. We define two levels of similarity: (i) between two entities, and (ii) between an entity and a group of entities.

We operationalize similarity using the Pearson correlation coefficient, where higher positive correlation indicates greater similarity.

- **Pairwise similarity.** The similarity between two countries or two indicators is measured by the Pearson correlation between their respective value sequences:

$$sim(X,Y) = \text{corr}(X,Y) \tag{1}$$

  where corr denotes the Pearson correlation coefficient computed over corresponding data points (e.g., yearly values):

$$\text{corr}(X,Y) = \frac{\sum(X-\bar{X})(Y-\bar{Y})}{\sqrt{\sum(X-\bar{X})^2}\sqrt{\sum(Y-\bar{Y})^2}} \tag{2}$$

- **Entity–group similarity.** Given a user- or system-defined country group $G_C = \{C_1, C_2, \ldots, C_N\}$ and indicator group $G_I = \{I_1, I_2, \ldots, I_M\}$, we compute the similarity between a *country of interest* C and its group G_C, as well as between an *indicator of interest* I and its group G_I.

$$sim(C, G_C) = \text{corr}\left(C, \text{mean}(G_C)\right) \tag{3}$$

$$sim(I, G_I) = \text{corr}\left(I, \text{mean}(G_I)\right) \tag{4}$$

4 Solution

4.1 Proposed Solution

Below is our proposed solution, allowing a flexible exploration of both countries and indicators of interest through 4 different visualization views. See Figure 4 for a lo-fi mockup of the solution.

4.1.1 Basic User Flow

The 4-view interface (shown in Figure 4) allows users to freely add and remove countries and indicators of interest. First, the user can select a pre-defined country group (to which they can add individual countries) or create their own country group in the *Country View* shown in Figure 5a. Pre-defined groups are suggestions of interesting starting points for the user based on existing literature, such as the 4 Asian Tigers. The user may only select country groups using the drop-down menu, but they can select individual countries using either the drop-down or on the zoomable map.

Once the user has selected a country group of interest, the *Indicator View* is activated. As shown in Figure 5b, indicators are grouped in higher-level categories seen in Table 1 for easier look-up and organization. If the user wishes to look at all indicators of a given category, they can select that category to add all its indicators into the group of interest. The user may also delete individual indicators from this group after selection.

After both the country and indicator groups are selected, the main page is activated. By activating subsequent views only after the previous selections have been made, we impose an order of operation on the user's first interaction with the tool—first selecting a country group, then an indicator group, and finally exploring trends and correlations. The main page consists of the two aforementioned views (*Country* and *Indicator*), along with the *Trends* and *Correlation* views (see Figure 4a).

The *Trends View* displays a simple line graph for each chosen indicator, with every country's line superimposed on top of each other.The hue channel encodes the country using a categorical colourmap.

The *Correlation View* offers two options through which a user can understand the similarity scores of the selected indicator groups. One option is the correlation matrix, which displays the pairwise similarity between two indicators (averaged across the selected country group) within the corresponding cell of a conventional matrix. The other option is a chord diagram, which displays each indicator as a node on a radial graph, and uses thickness of a line mark connecting any two indicators to encode its correlation. Each line is also annotated with the pairwise similarity value between the two end node indicators (same value used in the correlation matrix). A node can be selected within the diagram to highlight its outgoing connections. We imagined the chord diagram to be a more creative and exploratory tool to visualize how the indicators are connected to each other, whereas the matrix is better for understanding raw similarity scores to those who are already familiar with the concept of a correlation matrix. The user can switch between these two options using a toggle within the *Correlation View*.

4.1.2 Adding a New Country and/or Indicator

When the user hovers over a new country (see Figure 4b, a written description of the country and the country's entity-group similarity (i.e., $sim(C, G_C)$) with the rest of the selected countries are displayed on a pop-up view. Once they select a new country (see Figure 4c, a new line is added to each line graph in the *Trends View* to represent the new country's data. The *Correlation View* updates with new correlation scores computed with the additional country.

When the user hovers over a new indicator (see Figure 4d, an updated chord diagram containing the new indicator and the indicator's entity-group similarity (i.e., $sim(I, G_I)$) with the rest of the selected indicators are be displayed in a pop-up view. This design leans into the strengths of chord diagrams as an exploratory tool, and uses them as a preview to effectively support users with their tasks at the *Analyze* level (*discover* and *enjoy*). The pop-up window shows the user what changes they can expect to the correlations if they were to choose the indicator they are hovering over.

When the user selects a new indicator (see Figure 4e), a new line graph of the selected indicator is added to the *Trends View*. For the correlation matrix in the *Correlation View*, an additional row and column is populated in the matrix to account for the new indicator. From here, if the user toggles the switch in the view to display the chord diagram (see Figure 4f), the same chord diagram they would have seen when they hovered over the given indicator is displayed. The user may

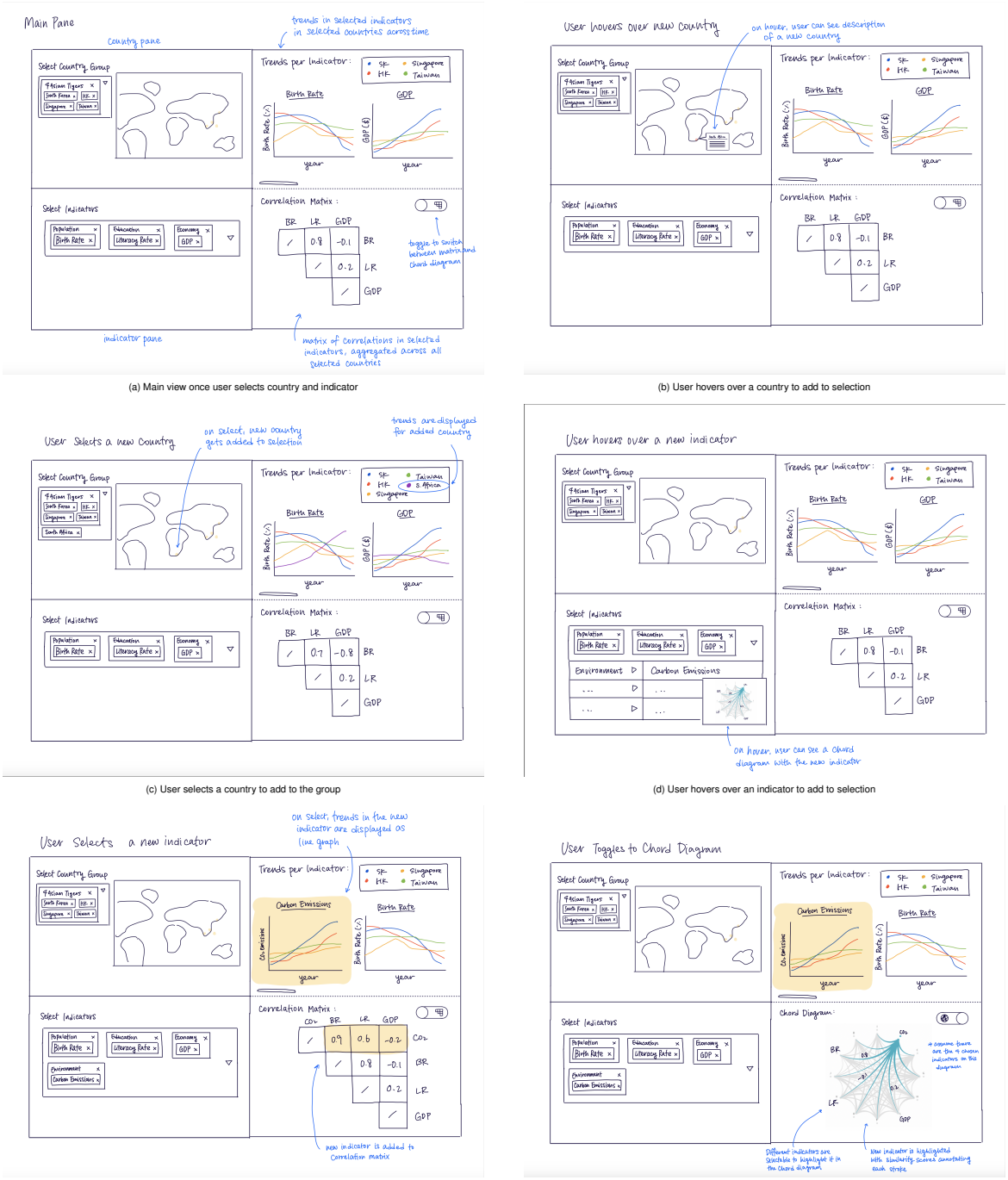


Fig. 4: Mockup of Main Interface

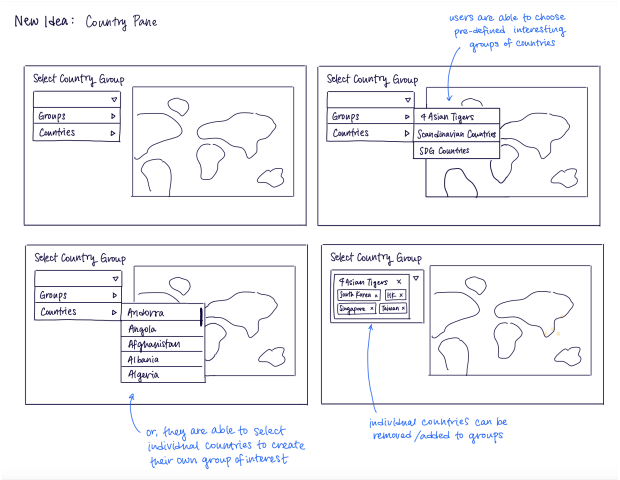
interact with the diagram at this point to highlight different indicators and assess their pairwise similarities.

Currently, we are hopeful that the interface would support selecting as many countries and indicators the user would be interested in. This means that there is technically no limit to how many countries and indicators the user can select at the same time. However, we may impose a limit on this value as we continue to implement the tool and see slower performance or illegible visualizations due to pixel availability. If so, the imposed limit will be symmetrical between countries and indicators, meaning that the maximum number of countries a user may select is equal to the maximum number of indicators they may select. If the user does not add additional countries to a pre-defined country group, however, the limit to the number of indicators they can select may be bigger than the fixed amount of countries in a given group.

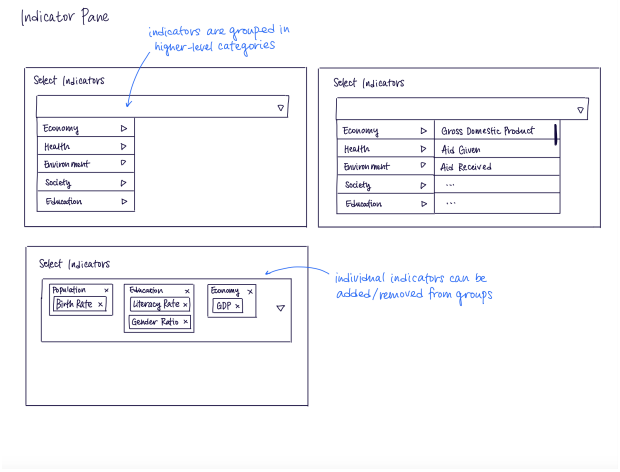
Every update to either the *Trends View* or *Correlation View* will be highlighted to allow users to easily identify the changes made to the main page. Given that there are many different visualizations displayed together, this highlight will reduce the user's cognitive load to manually track changes across the views.

4.1.3 User Scenario

The user begins with a predefined group of countries and indicators. In this scenario, this group include four countries—Four Asian Tigers (South Korea, Hong Kong, Singapore, and Taiwan)—and three indicators: birth rate, literacy rate, and GDP. With these selections, the *Trends View* displays line charts showing each indicator's trend across



(a) Country selection view



(b) Indicator selection view

Fig. 5: Mockup of (a) Country View and (b) Indicator View

the four countries, while the *Correlation View* visualizes correlations among indicators. The user can first examine the precise correlation coefficients using the correlation matrix. Then, they can choose to toggle to a chord diagram that visualizes the correlation along a radial graph with line marks to encode the strength of correlation between two given nodes (indicators).

Next, the user can explore whether other countries share similar patterns with the Four Asian Tigers. By selecting a country on the map or in the country list, the system dynamically updates the visualizations. The user first hovers over South Africa to see its group similarity with the 4 Asian Tigers. A group similarity of 0.8 suggest that South Africa exhibits trends similar to the 4 Asian Tigers. The user selects South Africa. A new line for South Africa is superimposed in the *Trends View*, and in the *Correlation View*, the correlation visualizations update with values computed with South Africa added to the group.

The same interaction applies to indicators. The user hovers over "C02 Emissions" to preview an updated chord diagram and its group similarity to the rest of the three indicators in a pop-up view. This allows the user to gauge how C02 Emissions may effect the correlations between birth rate, literacy rate, and GDP before committing to add it to the group. The user selects C02 Emissions. A new line chart is generated in the *Trends View* and an additional node is created for the chord diagram, and an additional row/column is generated for the correlation matrix in the *Correlation View*.

4.2 Tools

To implement our design, we will aim to use Vega-Lite due to its flexibility and potential for it to be hosted as a web application. We will aim to leverage Vega-Lite instead of vanilla D3, as we are not implementing complex visualization idioms but rather exploring innovative ways in which simple idioms can convey a deeper correlation analysis. For data cleaning and exploratory analysis, we used Python.

5 Milestones

Table 2 summarizes the proposed milestones, organized by project phase, task, due date, estimated time, and assigned team members. The timeline can be iteratively refined to adjust the project scope and tool requirements as the work progresses.

References

[1] Gapminder. https://www.gapminder.org/, 2025. Free material from www.gapminder.org. 1, 2, 3

[2] R. J. Aitken. Population decline: where demography, social science, and biology intersect. *Reproduction*, 168(1), 2024. 3

[3] M. Aradmehr. Socio-demographic and Religious Factors Affecting Fertility Rate among Childbearing Women in Easter Iran: A Population-based Study. *Reproductive health*, 7(1):1553–9, 2019. 3

[4] L. Battle and J. Heer. Characterizing Exploratory Visual Analysis: A Literature Review and Evaluation of Analytic Provenance in Tableau. In *Computer graphics forum*, vol. 38, pp. 145–159. Wiley Online Library, 2019. 2

[5] Y. Benjamini. Simultaneous and selective inference: Current successes and future challenges. *Biometrical Journal*, 52(6):708–721, 2010. 2

[6] W. S. Cleveland, P. Diaconis, and R. McGill. Variables on Scatterplots Look More Highly Correlated When the Scales are Increased. *Science*, 216(4550):1138–1141, 1982. 1

[7] M. Fink, J.-H. Haunert, J. Spoerhase, and A. Wolff. Selecting the Aspect Ratio of a Scatter Plot Based on Its Delaunay Triangulation. *IEEE Trans. Visualization and Computer Graphics (TVCG)*, 19(12):2326–2335, 2013. 1

[8] L. Harrison, F. Yang, S. Franconeri, and R. Chang. Ranking Visualizations of Correlation Using Weber's Law. *IEEE Trans. Visualization and Computer Graphics (TVCG)*, 20(12):1943–1952, 2014. 1

[9] J. Li, J.-B. Martens, and J. J. Van Wijk. Judging Correlation from Scatterplots and Parallel Coordinate Plots. *Information Visualization*, 9(1):13–30, 2010. 1

[10] T. Munzner. *Visualization Analysis and Design*. CRC Press, 2014. 3

[11] P. G. Nair. Decline in Birth Rate in Kerala: A Hypothesis about the Inter-Relationship between Demographic Variables, Health Services and Education. *Economic and Political Weekly*, pp. 323–336, 1974. 3

[12] P. K. Rai, S. Pareek, and H. Joshi. Regression Analysis of Collinear Data using r-k Class Estimator: Socio-Economic and Demographic Factors Affecting the Total Fertility Rate (TFR) in India. *Journal of Data Science*, 11(2):323–342, 2013. 3

[13] H. Ritchie and L. Rodés-Guirao. Peak global population and other key findings from the 2024 UN World Population Prospects. *Our World in Data*, 2024. https://ourworldindata.org/un-population-2024-revision. 2

[14] J.-E. Song, J.-A. Ahn, S.-K. Lee, and E. H. Roh. Factors related to low birth rate among married women in Korea. *PLoS one*, 13(3):e0194597, 2018. 3

[15] United Nations Department of Economic and Social Affairs, Population Division. World Population Prospects 2024, 2024. Accessed: 2025-10-17. 3

[16] United Nations Department of Economic and Social Affairs, Population Division. UN Population Division Data Portal, 2025. Accessed: 2025-10-17. 2

[17] T. Vigen. spurious correlations. https://www.tylervigen.com/spurious-correlations, 2025. Accessed: 2025-10-17. 1

[18] H. Wickham, D. Cook, H. Hofmann, and A. Buja. Graphical Inference for Infovis. *IEEE Trans. Visualization and Computer Graphics (TVCG)*, 16(6):973–979, 2010. 2

[19] K. Wongsuphasawat, D. Moritz, A. Anand, J. Mackinlay, B. Howe, and J. Heer. Voyager: Exploratory Analysis via Faceted Browsing of Visualization Recommendations. *IEEE Trans. Visualization and Computer Graphics (TVCG)*, 22(1):649–658, 2015. 1, 2

Phase	Task	Due Date	Duration	Team Member	Status
Ideation	Pitch	Sep 25	2hr	Minju, Alice	Completed
Proposal	Brainstorming, Pre-proposal meetings	Oct 9	3hr	Minju, Alice	Completed
	Dataset Collection	Oct 16	1hr	Alice	Completed
	Preliminary EDA	Oct 17	2hr	Alice	Completed
	Related Literature Review	Oct 19	3hr	Minju	Completed
	Define Data/Task Abstraction	Oct 19	2hr	Alice	Completed
	Solution Brainstorming	Oct 19	1hr	Minju, Alice	Completed
	Report Write-up	Oct 19	3hr	Minju, Alice	Completed
Update	Finalizing Solution Idea	Oct 30	1hr	Minju, Alice	Completed
	Tool Familiarization (Vega-Lite, D3)	Oct 29	10hr	Minju, Alice	Completed
	Data Cleaning	Nov 5	4hr	Alice	Completed
	Create Initial Implementation	Nov 6	6hr	Minju	Completed
	Update Writeup	Nov 12	2hr	Minju, Alice	Completed
Peer Review	Peer Review	Nov 13	3hr	Minju, Alice	Not started
Post Update	Post Update Meeting with Tamara	Nov 20	3hr	Minju, Alice	Not started
Implementation	Country Selection Map View (Viewl 1)	Nov 13	4hr	Minju	In Progress
	Indicator Selection View (Viewl 3)	Nov 13	4hr	Alice	In Progress
	Individual Chart View (Viewl 2)	Nov 20	4hr	Minju, Alice	Not started
	Correlation View (Viewl 4)	Nov 27	6hr	Minju, Alice	Not started
	Integrate Views and Implement Interactions	Dec 4	10hr	Minju, Alice	Not started
Evaluation	Evaluation Planning	Dec 5	3hr	Minju, Alice	Not started
	Evaluation	Dec 9	5hr	Minju, Alice	Not started
	Analyze Results	Dec 10	2hr	Minju, Alice	Not started
Presentation	Presentation Prep	Dec 10	4hr	Minju, Alice	Not started
	Presentation	Dec 11	30min	Minju, Alice	Not started
Final Report	Report Writeup	Dec 15	6hr	Minju, Alice	Not started

Table 2: Tentative timeline

[20] K. Wongsuphasawat, Z. Qu, D. Moritz, R. Chang, F. Ouk, A. Anand, J. Mackinlay, B. Howe, and J. Heer. Voyager 2: Augmenting Visual Analysis with Partial View Specifications. In *Proc. ACM Human Factors in Computing Systems (CHI)*, pp. 2648–2659, 2017. 1, 2

[21] K.-P. Yee, K. Swearingen, K. Li, and M. Hearst. Faceted Metadata for Image Search and Browsing. In *Proc. SIGCHI Conf. Human Factors in Computing Systems*, pp. 401–408, 2003. 2

[22] E. Zgraggen, Z. Zhao, R. Zeleznik, and T. Kraska. Investigating the Effect of the Multiple Comparisons Problem in Visual Analysis. In *Proc. ACM Human Factors in Computing Systems (CHI)*, pp. 1–12, 2018. 2