CS661: Big Data Visual Analytics

Visualization of World Statistics

Group Members -

Anjanesh Rakesh (200133)
Manan Kabra (200552)
Manas Kumar (200555)
Sachin Kumar (200832)
Siddharth Sharma (200981)
Yash Barjatya (201138)
Yashwardhan Singh (201157)
Yogesh Nain (201165)
Yudhveer (201166)

Introduction:

Global statistics are more than just a collection of data; they serve as an insightful prism through which we may see the lives of billions of people. These indicators provide a clear picture of the status of nations and the well-being of their residents, ranging from the extreme density of the population to the resources allocated to healthcare. Beyond the mere numerical values, world statistics possess actual power. Population density illustrates where people have chosen to live, while information on healthcare spending shows how committed a country is to providing affordable healthcare. Rates of mortality and life expectancy serve as indicators of the health of a society, and employment data offers insight into the heart of a country's economy.
We explored the realm of statistics and embarked on an important exploration journey. We are starting to decipher the intricacies of our worldwide community, acquiring a more profound comprehension of the obstacles and prospects that will mold our shared destiny.

Problem:

The dataset encompasses diverse features for each country worldwide, with each feature associated with an attribute: the country and its corresponding values over a specified range of years.

The following features are being used:

- Country: List of all countries.
- Region: Region the country belongs to. [1]
- Year: The year for which the row contains data.
- Gender: Some attributes have gender-specific values.
- Population Density: Average number of people on each square km of land in the country. [2]
- Health Expenditure: Percentage of health expenditure paid by: [3]
    - Government
    - Private Firms
    - Individual
- Life Expectancy: Average number of years a newborn child would live. [4]
- Education Expenditure: In terms of Gross National Income (GNI) percentage. [5]
- Mean years in school: Number of years of school attended by people in the age group 15-24. [6]
- Average Income: Adjusted National Annual Income. [7]
- Employment Rate: Percentage of the country's population employed during the year. [8]
- Income Inequality: The Gini Coefficient shows income inequality in society. [9]
- Child Mortality Rate: Death of children under five years as per 1000 live births. [10]
- Adult Mortality Rate: Death of adults 1000 live adults. [11]
- Suicides: Mortality due to self-inflicted injury as per 100,000 standard population. [12]

● Murders: Mortality due to interpersonal violence as per 100,000 standard population. [13]
● Military Expenditures: In terms of percentage of GDP. [14]

We'll analyze global statistical data on various features to gain insights into the well-being of each country, determine which country has the highest population, identify which gender has the highest life expectancy in specific countries, and compare data across selected features from 2000 to 2010. Additionally, we'll explore correlations between certain features and trends in changes, such as economic rates or average incomes, across different countries over the years.

Features of the dashboard:*

- **Filters:** This section will allow users to apply filters to the dataset based on various features. Users will be able to select specific categories or items to analyze.

- **Choropleth Map**: The map will display all countries worldwide, with each country's population indicated by a color scheme. When filters are applied, countries matching the selected criteria will be highlighted, while others will be disabled or displayed in grey. Additionally, clicking on a country will highlight its data in the parallel coordinate plot.

- **Parallel Coordinates Plot:** This plot will present data for each country across multiple dimensions simultaneously. Filters applied in the dashboard, such as category selection, year, and region, will be linked to this plot, displaying only the filtered data. Users will be able to brush a selected portion of a dimension, scale axis values, and exclude specific data points. Additional features will include axis manipulation, inversion, and data export.

- **Lollipop Bar Chart:** This chart will illustrate the number of countries in each year that meet the applied filters. Users will be able to filter data by year by clicking on the corresponding bar to enable or disable it.

- **Regions Bar Chart:** This chart will display the number of countries in each region. Clicking on a region will toggle filtering for that region. Each region will be color-coded, corresponding to the colors used in the parallel coordinate plot.

- **Radar Plot:** This plot will showcase the intrinsic dimensions obtained through PCA (Principal Component Analysis). Users will be able to analyze dimension values for specific countries like the United States, India, and China by selecting the desired year.

*Note: The above features are subject to change based on refinement and feedback

Approach:

We will approach the problem in the following steps:

- **Data Pre-processing:**
  We will use both forward and backward-filling methods to handle missing data, which is typical in time-series datasets. This propagates the values already present in each dataset to fill the gaps.

- **Merging Datasets:**
  To manage the data for various years, duplicate rows for each nation will be created. To manage data particular to gender, a duplicate row will be created for every year. Certain features (such as population density) whose data is gender independent will have the same value in both rows regardless of the value in the gender column.

- **Filtering Data:**
  The dataset will take a long time to process during visualization because it has over 30,000 rows. As a result, we will filter the dataset so that it only includes information for each nation from 2000 to 2010. We can obtain accurate and insightful information from the data throughout this period because there aren't many missing numbers.

- **Data Analysis:**
  Once we have the cleaned and filtered data, we will analyze it in the following way:
  - ❖ Based on Country, Year and Gender
  - ❖ Correlation between features
  - ❖ Find the best features
  - ❖ View trends in features over the years

- **Visualization Methods:**
  We will apply standard visualization techniques to understand and analyze this data based on its type. A few methods may be:
  - ❖ Bar / Pie Charts for Categorical Data
  - ❖ Histograms from Numerical Data
  - ❖ Scatter plots to view distributions

  We will also apply some non-standard visualization techniques to gain more powerful insights from the data. Few techniques we may use:
  - ❖ PCA
  - ❖ Parallel Coordinates
  - ❖ Heatmaps

- **Implementation Details:**
  - ❖ Python Flask Framework: On the server side for Data Processing
  - ❖ D3: Javascript library on client-side for visualization

- HTML and CSS: For design and beautification
- AJAX HTTP requests: Used to communicate between client and server for asynchronous service calls
- JSON Framework: Used to respond to the service calls and return the required data

Conclusion:

This project aims to discover trends and obtain insights by analyzing and visualizing data from different nations. Users can investigate relationships between various features and comprehend prospective repercussions using an interactive dashboard.

References for Datasets:

[1] Region: Kaggle - https://www.kaggle.com/fernandol/countries-of-the-world/version/1
[2] Population Density: https://population.un.org/wpp/
[3] Health Expenditure: https://www.who.int/gho/en/
[4] Life Expectancy: https://population.un.org/wpp/
[5] Education Expenditure: UNESCO
[6] Mean years in school:
http://ghdx.healthdata.org/record/global-educational-attainment-1970-2015;
http://www.healthmetricsandevaluation.org/
[7] Average Income: http://gapm.io/dgdppc
[8] Employment Rate: https://www.ilo.org/ilostat/
[9] Income Inequality: http://gapm.io/ddgini
[10] Child Mortality Rate: https://www.gapminder.org/data/documentation/gd005/
[11] Adult Mortality Rate: (1) United Nations Population Division. World Population Prospects 2017 Revision, (2) University of California, Berkeley and Max Plank Institute for Demographic Research. Human Mortality Database.
[12] Suicides: https://www.healthdata.org/
[13] Murders: https://www.healthdata.org/
[14] Military Expenditure: https://data.worldbank.org/indicator/MS.MIL.XPND.GD.Z