Subject : CSE 564 - Visualization

Topic : World Statistics

PROJECT PROPOSAL

By Melita Saldanha, and Neha Shetty

Background:

Sometimes we need some hard data to fully understand things; it means a lot more to say "1.3 billion people live in poverty" than to say "a lot." World Statistics is all about the data of each country with respect to various categories like Population density, Health Expenditure, Average Income etc. The most important statistic of all is just how many people there are. Population density is a measurement of population per unit area. It is frequently applied to living organisms, most of the time to humans. Another important metric is the wellbeing of the public. The Global Health Expenditure Database (GHED) provides internationally comparable data on health spending for close to 190 countries from 2000 to 2017. This helps ensure health services are available and affordable when people need them.

Few other comparable features are Life expectancy, Mortality rate and Employment rate. Life expectancy at birth reflects the overall mortality level of a population. It summarizes the mortality pattern that prevails across all age groups in a given year – children and adolescents, adults and the elderly. Mortality rate, or death rate is a measure of the number of deaths in a particular population, scaled to the size of that population, per unit of time. Mortality rate is typically expressed in units of deaths per 1,000 individuals per year. Employment rates are defined as a measure of the extent to which available labour resources (people available to work) are being used. They are calculated as the ratio of the employed to the working age population.

World Statistics will help us find out which countries are the largest in the world, have the lowest mortality rate, the highest life expectancy, and more.

Problem:

The dataset consists of various features about each country in the world. Each feature has an attribute: Country and its value for each year in a certain range.

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 percentage of Gross National Income (GNI). [5]
- Mean years in school: Number of years of school attended by people in age group 15-24. [6]
- Average Income: Adjusted National Annual Income. [7]
- Employment Rate: Percentage of population of the country employed during the given year. [8]
- Income Inequality: Gini Coefficient showing Income Inequality in a society. [9]
- Child Mortality Rate: Death of children under 5 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]

Given world statistics data with respect to the above mentioned features we will try to find some insights about the well being of each country, the country with the highest population or which gender has the highest life expectancy in a particular country. We will try to compare data of each country across selected features over the years from 2000 to 2010. We also intend to find correlations between some of the features and try to find trends in the changes like economic rates or average incomes of few countries over the years.

Approach:

We will approach the problem in the following steps:

- **Data Pre-processing:**
  Missing values of each dataset will need to be handled. Since the data is a time-dependent data, we will use forward fill and backfill methods to propagate non-null values forward or backward in each dataset.

- **Merging Datasets:**
  Each row representing a country was replicated to handle the data for different years. Similarly, each row representing a year was replicated to handle gender specific data. Some features whose data is not dependent on gender, will have the same value in both rows irrespective of the value in the gender column (Eg. Population Density).

Country	Year	Gender	Population_Density	Life_Expectancy
United States	2000	Male	30.8	74.1
United States	2000	Female	30.8	79.5
United States	2001	Male	31.1	74.3
United States	2001	Female	31.1	79.6

Sample dataset after merging

- **Filtering Data:**
  The dataset consists of over 30,000 rows which will take very long to process during visualization. Therefore, we will filter the dataset to consider data of each country for the years 2000-2010 only. The data during this time range has very few missing values and hence, it will help us get precise and good insights.

- **Data Analysis:**
  Once we have the cleant 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

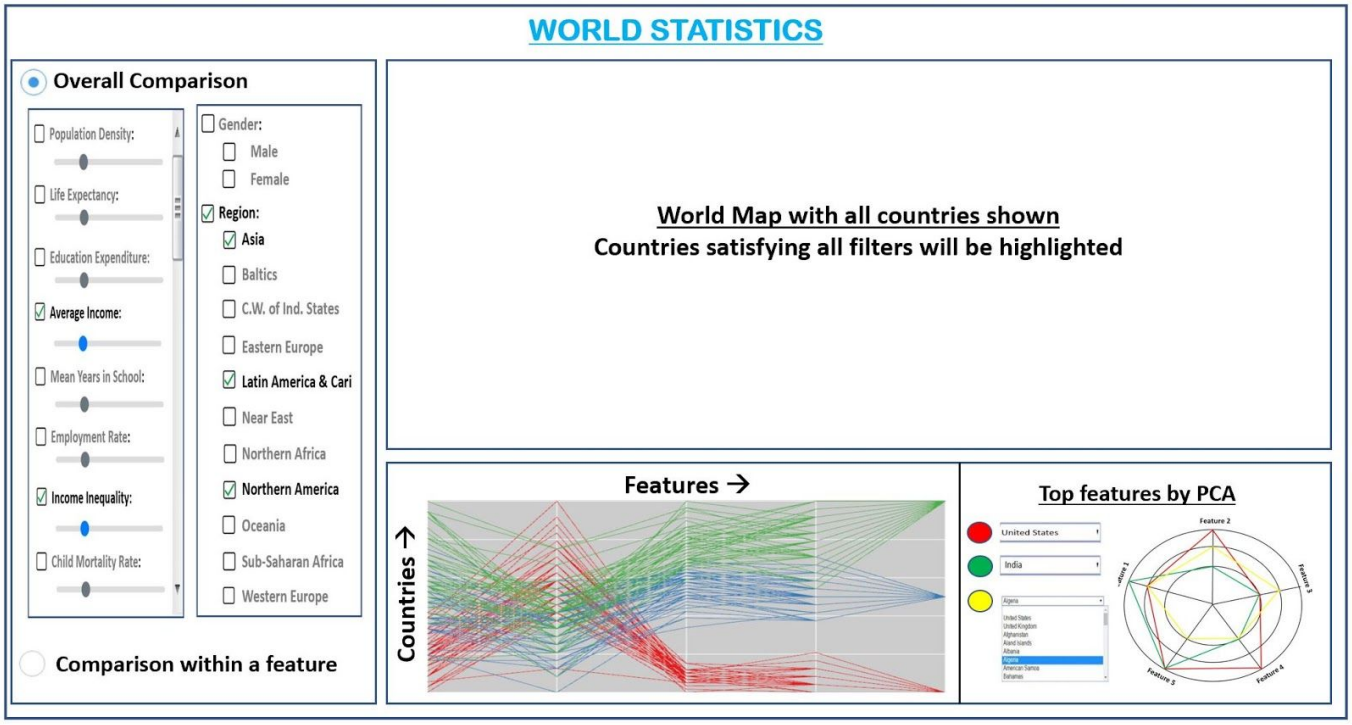
- **Visualization Methods:**
  We will apply some standard visualization techniques to understand and analyze this data, based on its type. Few techniques may be:
  - ➢ Bar / Pie Charts for Categorical Data
  - ➢ Histograms from Numerical Data
  - ➢ Scatter plots to view distributions

  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

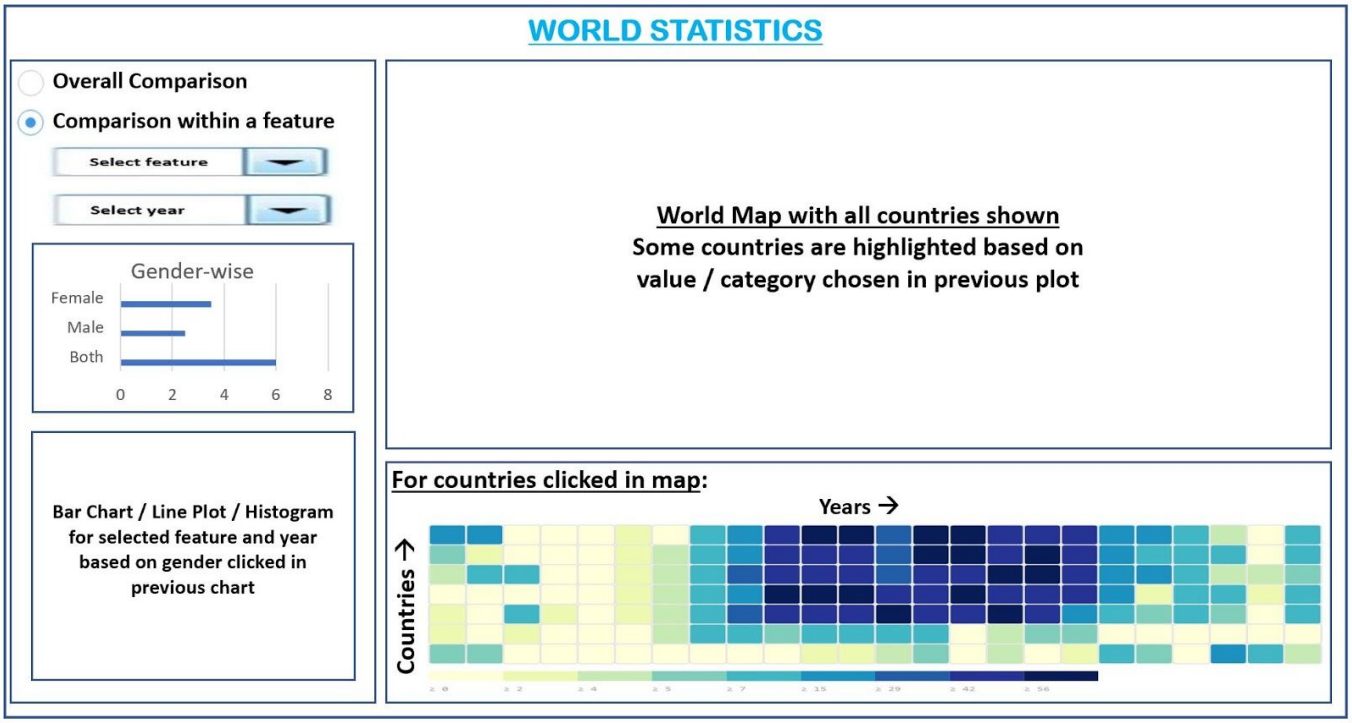
- **Implementation Details:**
  - ➢ **Python Flask Framework:** On server side for Data Processing
  - ➢ **D3:** Javascript library on client-side for visualization
  - ➢ **HTML and CSS:** For design and beautification
  - ➢ **AJAX HTTP requests:** Used to communicate between client and server for asynchronous service calls
  - ➢ **JSON Framework:** Used to respond to the service calls and return the required data

Dashboard Design Template:

The dashboard will have two radio buttons to select either ‘Overall Comparison’ or ‘Comparison within a feature’. The visual elements of analysis will change based on which button is checked. Following are the proposed templates:



Overall Comparison based on filters provided for all features



Comparison of selected feature

Conclusion:

With this project we will be able to understand various aspects of each country in the world using data analysis and visualization. It will help us find insights and trends in this data in a meaningful way. We will create an interactive dashboard that will enable the user to easily identify correlations among several features, how some features might possibly impact others thereby gaining more information about this data.

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.ZS