CS4460 HOMEWORK 2: EXPLORING DATA

KALANI DISSANAYAKE

Part 1: Focus on Holistic Dataset

1. List (bullet list of items) of three queries or questions that someone may have about this PARTICULAR data set

- What countries have the longest healthy life expectancy, and how do their economic factors compare?
- Are there any countries that have a high Human Development Index (HDI) but low education levels?
- How does female involvement in the workforce correlate to gender equality in different countries?

2. Not using your list from question1 at all (do not respond to the queries you wrote in question 1), please list (bullet list of items) of five "insights"

- A strong correlation exists between maternal mortality rates and low education levels.
- Countries with a high Gender Inequality index (GII) often have lower female education/labor rates.
- Countries with the highest healthcare spending don't necessarily have the best health outcomes
- Many countries with low unemployment rates are associated with high education levels
- In several South Asian and African countries, there are notable differences in the literacy rates between men and women

3. List (bullet list of items) of at least four steps / tasks you performed as part of the exploration and analysis

- Sorted "Healthy life expectancy at birth" from highest to lowest to identify trends within health outcomes
- Filtered the data to compare the countries with the highest GII and lowest GII to see how gender inequality may correlate to other indicators.
- Focused on countries with the lowest unemployment rates and compared it to literacy rates and time in school to see if there was a correlation.
- Calculated the difference between male and female literacy rates in each country, to see which countries had the largest gender gaps.

Part 2: Focus on Attributes

4. Rate them as low, medium, or high (low is least priority for a visualization) and your reasoning (you could include a thought for a helpful visualization, if applicable).

- Do countries in cold regions have more Female Internet Users?

    - high priority, a map would be great as it can show the locations and data, making it more simple to spot trends
- Globally, what is the Average Life Expectancy for females (one global value)?
    - Low priority, a simple statistic is enough and a visualization is unnecessary
- What is the Unemployment Rate in Thailand?
    - Low priority, again a simple statistic is enough, and a visualization is not needed.
- Do countries with high Maternal Mortality have low Mean Years of Schooling?
    - High priority, a scatterplot could display this relationship very clearly.
- How many missing values are there for Literacy Rate?
    - Medium priority, a bar chart could help visualize but it is not essential to see the gaps.
- Which attributes (columns) follow a normal distribution and which follow a bimodal distribution?
    - High priority, histograms are the best way to visualize data and show the patterns/distributions

5. Data classification Look at the distributions of values in the columns of data and recommend how you may want this data classified into groups.

- Which of the following is best classified using “equal interval” classification: Refugees by country of origin (Table 12) or Youth not in school or employment (Table 11) and why?
    - “Youth not in school or employment” is best classified using equal interval because it allows for a straightforward way to categorize countries with similar levels.
- Do you think a defined interval would be a good data classification method for Dependency ratio: Young age (0–14) (Table 7)? Why or why not?
    - No because a defined interval might not capture the wide range of values across countries, instead we should use something more flexible.
- Which is a better classification for Female youth literacy rate (Table 9): “natural breaks” or “quantiles” and why?
    - Natural breaks is better because it helps to cluster similar values together, making the data more meaningful and concise.
- Name a column in Table 10 where you would recommend a logarithmic scale binning method (e.g., a base 10 log scale).
    - GDP per capita, since it highly skewed with high values for some countries.