**Homework 2**: Exploring Data

**Objective: Explore and analyze data without using an information visualization system.**

**First, download the UN Countries Development Indicators dataset. There are two ways to do this:**  
  
**1) Copy and paste the following into your web browser (clicking may or may not work):**  
[**https://hdr.undp.org/sites/default/files/data/2020/2020\_statistical\_annex\_table\_7-15\_excluding\_table14.xlsx**](https://hdr.undp.org/sites/default/files/data/2020/2020_statistical_annex_table_7-15_excluding_table14.xlsx)

**or**

**2) Download the file from our class** [**Canvas files (under homeworks).**](https://gatech.instructure.com/courses/391510/files/folder/Homeworks)

Data Source: United Nations, Human Development Report 2020. Check out more at the link below:  
<https://hdr.undp.org/data-center/documentation-and-downloads>

Directions: Take 15 minutes to explore this interesting dataset. You can use Excel or **any** other environment / programming language (Python, R, etc.), ***but avoid the visualization and charting functionalities of these tools.*** (You will get points off if you include charts). Your goal is to perform an exploratory analysis of the dataset, to learn about countries and UN Sustainable Development Goal (SDGs) indicators, and to develop insights about the world.

While and after you explore, please answer the following:

**PART 1: Focus on Holistic Dataset (55%)**

1. **(15%) List (bullet list of items)** of **three** **queries** **or** **questions** that someone may have about this PARTICULAR data set. These would be questions that an analyst examining the data might be pondering. These questions are **not** about the metadata (who collected it, etc.). Pretend the dataset is ‘perfect’ already---so these questions should be about the countries and indicators. Do not make them all correlation questions, or all min-max questions, etc.

(*Example question: Are countries with high human development indicators near each other?)*(Not a good example question for this exercise: *What other data could we collect? Why is data missing? What does ‘Current health expenditure’ mean?)*

1. **(20%) 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**”, chunks of knowledge about the countries and indicators that you gained while exploring the data. An insight could be some understanding of the data and its characteristics that is not initially obvious or intuitive.

*(Example insight: In the Russian Federation, the suicide rate for men is more than 6 times the rate for women.)*

1. **(20%) List (bullet list of items)** of at least **four steps / tasks** you performed as partof the exploration and analysis. For each, what did you do in Excel/whatever software you used? (No step is not important!) *and* why did you do these/how did they help?

*(**Example step: I* ***sorted*** *the life expectancy column ascending, I did this to see the countries with the lowest life expectancy.)*

**(keep going)**

**PART 2: Focus on Attributes (45%)**

1. **(25%) Some questions could really benefit from a visualization, while others may be easy to answer *without* one:**For which of these questions would a visualization be MOST helpful: 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).

**-**You don’t need to answer the actual questions posed!

**-**You also don’t really need to look at the table for these, but you can.

*(Example: Which country has the most people (highest population)? [ex. “Low: because…”])*

1. Do countries in cold regions have more *Female Internet Users*?
2. Globally, what is the *Average Life Expectancy* for females (one global value)?
3. What is the *Unemployment Rate* in Thailand?
4. Do countries with high *Maternal Mortality* have low *Mean Years of Schooling*?
5. How many missing values are there for *Literacy Rate*?
6. Which attributes (columns) follow a normal distribution and which follow a bimodal distribution?

**Bonus Question (5%): How many countries have ratified the *CRC: Convention on the Rights of the Child* (Table 15) by each year? (aka “As of a particular year, how many countries have officially accepted this convention?”)**

1. **(20%) Data classification** Look at the distributions of values in the columns of data and recommend how you may want this data classified into groups.
2. 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?
3. 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?
4. Which is a better classification for ***Female youth literacy rate*** (Table 9): “natural breaks” or “quantiles” and why?

1. Name a column in Table 10 where you would recommend a logarithmic scale binning method (e.g., a base 10 log scale).

**To turn into Canvas: A word document or PDF with the answers above.**