INTR 4051/SOSS 7091
Lab 4: Calculating the Human Development Index

Do this assignment in an R Markdown file called *yoursurname*`A05.Rmd`.

This assignment involves analyzing the data used to calculate the Human Development Index (HDI) of the United Nations Development Program (UNDP). We will be using the 2018 Statistical Update of the HDI, which you can read about at http://hdr.undp.org/en/2018-update.

I have put the data spreadsheet and the "technical note" that explains how the UNDP calculates the HDI on Sakai. (You could also use the "download data" button at http://hdr.undp.org/en/composite/HDI and get the technical notes from http://hdr.undp.org/sites/default/files/hdr2018_technical_notes.pdf.

You will need two R packages for the assignment: `readxl` and `countrycode`. If you need to install them, you can do so by running the following once in the console, with internet access:

```
install.packages(c("readxl", "countrycode"))
```

1. Open the data spreadsheet with a spreadsheet editor – MS Excel is fine, or you can install Libreoffice (free, open-source, available at https://www.libreoffice.org/download) and use Calc.

   Identify the top-left and bottom-right cells of the rectangular range that contains all the data needed:

   - **Columns:** must include everything from the name of the country through GNI per capita;
   - **Rows:** must include all countries for which HDI has been calculated, but exclude "other countries or territories" and aggregates at the bottom.

   Note the identifiers of the top-left and bottom-right cells – (for example, if you wanted the very top-left cell of the spreadsheet, you would identify it as "`A1`"). Also note the ways (there is more than one) that "missing values" are reflected.

   Read the rectangular range of data into R using the `read_excel` function in the `readxl` package, assigning it to an object that you call `HDI`.

   (Hints: you will need to use the `path` (filename), `range`, `col_names=FALSE`, and `na` arguments. Check the help page for the function to see how these work. Don't forget to include `library(readxl)` in your code chunk to load the package.)

   Use functions like `dim`, `str`, `head`, `tail`, and `summary` to inspect the new object `HDI`. There should still be some "junk" to be cleaned. But if you notice that any essential data is missing, check the spreadsheet again, modify your call to `read_excel`, and retry.

2. "Cleaning" the data should now only require:

   - Removing (entire) unneeded columns from the data frame;
   - Removing (entire) unneeded rows from the data frame;
   - Adding informative column names.

   If you notice other problems with the data frame, it is probably easier to go back and "fix" the `read_excel` step.

   Use indexing or functions to remove columns and rows that do not contain country data. Then add these column names:

   - "Country" = `country`
   - "HDI value" = `hdi`
   - "Life Expectancy at Birth" = `lifexp`
   - "Expected Years of Schooling" = `school.exp`
   - "Mean Years of Schooling" = `school.mean`
   - "Gross National Income (GNI) per capita" = `gni.pc`

Use R functions to check your work.

3. Countries are identified by their names. It is sometimes convenient to be able to refer to them by abbreviations – ISO-3C being a common set (International Standards Organization three-character).

   Use the `countrycode` function in the `countrycode` package to try to match the country names to their ISO-3C codes, putting the result in a new column called `HDI$code`.

   You should only need three arguments, in this order: `sourcevar` (the vector with the country names), `origin` (the coding scheme of `sourcevar`, in this case `"country.name"`), and `destination` (the coding scheme for the new variable, in this case `"iso3c"`).

4. Graph the univariate distributions. For each of the five numeric vectors:

   - First, use `hist` with the argument `prob=TRUE` to get a histogram scaled by "density";
   - Second, immediately add a density curve to the graph with `lines(density(x), col="red")`.
     (where `x` is the name of the vector; if there are `NA`s, include `na.rm=TRUE` in your arguments to `density`).

     (If you want to save each graph for later reference, you can "sandwich" these two plot commands between `pdf(file="`*`filename`*`.pdf")` and `dev.off()` — one file for each graph.)

5. Which of the five distributions looks most:
   (a) Symmetrical?
   (b) Left-skewed (i.e., long left "tail")?
   (c) Right-skewed (i.e., long right "tail")?

6. A distribution is *unimodal* if it has one distinct "hump," *bimodal* if it has two distinct "humps," and *multimodal* if it has more than two distinct "humps."

   In these terms, how would you describe the distribution of `HDI$hdi`. Briefly, what does this tell you substantively about "human development" (as measured by the HDI) in the contemporary world?

7. The HDI's income indicator uses the natural logarithm of GNI per capita, rather than GNI per capita itself.
   (a) Run a histogram with density curve (as above) for *the natural logarithm of* GNI per capita (remember the `log` function for logarithms);
   (b) Compare the shape of the distribution with the shape of the distribution of GNI per capita (from the previous graph). How do they differ?

8. Follow the instructions from the UNDP's "Technical Note 1" to calculate the component indexes of the HDI and then the HDI itself. (Hint: Use vector arithmetic to do the calculations for all countries at once.)

   (Be careful to note the table with "Goalposts for the HDI in this Report" and its relevance to the calculations. Also note the worked example of the HDI for Cyprus, which should help clarify the calculations.)

   (Finally, note that a *geometric mean* differs from the more familiar *arithmetic mean* (average). A geometric mean of n elements is calculated by *multiplying* all of the elements and then taking the nth root (i.e., raising to the $1/n$ power).)
   (a) The life expectancy index (call it `HDI$lifexpI`);
   (b) The education index (call it `HDI$educI`);
   (c) The income index (call it `HDI$incomeI`);
   (d) The overall HDI (call it `HDI$myHDI`)

   (The result of each of these calculations should be a vector, with a value for each country. Run `summary` on each vector to check that the values are plausible.)

9. Compare your HDI values (`hdi$hdi`) with the UNDP's values (`hdi$hdi`). Are there any discrepancies too large to be the result of rounding error?

A central rationale for the HDI is to provide a measure of "human development" that goes beyond just per capita income. As the UNDP puts it:

> Human Development is a development paradigm that is about much more than the rise or fall of national incomes. It is about creating an environment in which people can develop their full potential and lead productive, creative lives in accord with their needs and interests. People are the real wealth of nations. Development is thus about expanding the choices people have to lead lives that they value. And it is thus about much more than economic growth, which is only a means —if a very important one — of enlarging people's choices.

10. Calculate a "non-income HDI" (`hdi.nonincome` as the geometric mean of `lifexpI` and `educI`.
    (a) Make a scatterplot (`plot(x, y)`) of the non-income HDI against the income index (`HDI$incomeI`), with income on the horizontal axis and UEHDI on the vertical axis.
    Is the association between the two variables positive, negative, or is there no clear association?
    (b) Create a vector containing the differences between non-income HDI and income for each country.
    Which country's non-income HDI exceeds its income index by the largest margin?
    Which country's income index exceeds its non-income HDI by the largest margin?
11. The HDI is a "multidimensional" measure of human development by design. Comment briefly on its validity and reliability. How does it compare with a purely income-based measure? What do you think the HDI's main limitations are?