* + Data

The 'human' dataset originates from the United Nations Development Programme. See [their data page](http://hdr.undp.org/en/content/human-development-index-hdi) for more information. For a nice overview see also the [calculating the human development indices pdf](http://hdr.undp.org/sites/default/files/hdr2015_technical_notes.pdf).

Most of the variable names in the data have been shortened and two new variables have been computed. **See the meta file for the modified data** [here](https://raw.githubusercontent.com/TuomoNieminen/Helsinki-Open-Data-Science/master/datasets/human_meta.txt) for descriptions of the variables.

**Data wrangling (max 5 points)**

We will continue to work with the human data wrangled last week. At this point, the joined data should have 195 observations and 19 variables. In case you didn’t finish the data wrangling exercise last week, you can download the data for this weeks wrangling exercise from the link below. The column separator is a comma and the first row includes the column names.

<http://s3.amazonaws.com/assets.datacamp.com/production/course_2218/datasets/human1.txt>

To complete the data wrangling part, continue working on your create\_human.R from last week or create a new script file with that name (include your name, file description and link to original data source as a comment at the top of the file).

* + 1. Mutate the data: transform the Gross National Income (GNI) variable to numeric (Using string manipulation. Note that the mutation of 'human' was not done on DataCamp). **(1 point)**
    2. Exclude unneeded variables: keep only the columns matching the following variable names (described in the meta file above):  "Country", "Edu2.FM", "Labo.FM", "Edu.Exp", "Life.Exp", "GNI", "Mat.Mor", "Ado.Birth", "Parli.F" **(1 point)**
    3. Remove all rows with missing values **(1 point).**
    4. Remove the observations which relate to regions instead of countries. **(1 point)**
    5. Define the row names of the data by the country names and remove the country name column from the data. The data should now have 155 observations and 9 variables. Save the human data in your data folder **including the row names**. You can overwrite your old ‘human’ data. **(1 point)**

Analysis (max 15 points)

In case you did not complete the data wrangling part, below is a link to the 'human' data for the analysis part. The column separator is a comma and the first row includes the column names.

<http://s3.amazonaws.com/assets.datacamp.com/production/course_2218/datasets/human2.txt>

Create a new RMarkdown file and save it as an empty file named ‘chapter5.Rmd’. Then include the file as a child file in your ‘index.Rmd’ file. Perform the following analysis in the file you created.

* + 1. Load the ‘human’ data into R. Explore the structure and the dimensions of the data and describe the dataset briefly, assuming the reader has no previous knowledge of it (this is now close to the reality, since you have named the variables yourself). (**0-1 point)**
    2. Show a graphical overview of the data and show summaries of the variables in the data. Describe and interpret the outputs, commenting on the distributions of the variables and the relationships between them. **(0-2 points)**
    3. Perform principal component analysis (PCA) on the not standardized human data. Show the variability captured by the principal components. Draw a biplot displaying the observations by the first two principal components (PC1 coordinate in x-axis, PC2 coordinate in y-axis), along with arrows representing the original variables. **(0-2 points)**
    4. Standardize the variables in the human data and repeat the above analysis. Interpret the results of both analysis (with and without standardizing). Are the results different? Why or why not? Include captions in you plots where you describe the results by using not just your variable names, but the actual phenomenons they relate to. **(0-4 points)**
    5. Give your personal interpretations of the first two principal component dimensions based on the biplot drawn after PCA on the standardized human data. **(0-2 points)**
    6. Load the tea dataset from the package Factominer. Explore the data briefly: look at the structure and the dimensions of the data and visualize it. Then do Multiple Correspondence Analysis on the tea data (or to a certain columns of the data, it’s up to you). Interpret the results of the MCA and draw at least the variable biplot of the analysis. You can also explore other plotting options for MCA. Comment on the output of the plots. **(0-4 points)**