# BUS212f: Analyzing Big Data II Spring 2018

## Case # 3: Clustering for World Happiness

### Introduction

Leo Tolstoy began his epic novel *Anna Karenina* with this famous line: "All happy families are alike; each unhappy family is unhappy in its own way." Does the same logic apply to nations? Are all happy countries alike?

Since 2012, a group of independent researchers has issued the *World Happiness Report*, an effort to measure and understand human happiness around the world. According to the 2017 version,

### The first World Happiness Report was published in April, 2012, in support of the UN High Level Meeting on happiness and well-being…. Happiness is increasingly considered the proper measure of social progress and the goal of public policy. In June 2016, the OECD committed itself “to redefine the growth narrative to put people’s well-being at the centre of governments’ efforts.” In a recent speech, the head of the UN Development Program (UNDP) spoke against what she called the “tyranny of GDP”, arguing that what matters is the quality of growth.“ Paying more attention to happiness should be part of our efforts to achieve both human and sustainable development” she said*[[1]](#footnote-1)*.

The research team uses surveys to gather opinions from citizens of 156 countries and develop a Happiness index called the “Cantrill Life Ladder”. The researchers also rely on data that may be correlated with happiness, and in this assignment, we want to use some of those variables to attempt to discover meaningful clusters of similar countries.

### Variable Definitions:

The data for this exercise come from the World Happiness Report website (<http://worldhappiness.report/#happiness2018>) and some other sources. I have placed a csv file with the most recent 2018 data on GitHub. We will only work with this subset without partitions for training and test sets. The specific columns are these:

* **country**: name of country—for identification only.
* **region**: region of the world
* **LifeLadder**: National mean value of respondents’ responses to the “Cantrill Life Ladder” question, averaged for 2015-2017. The question asks respondents to value their lives today on a 0 to 10 scale, with the worst possible life as a 0 and the best possible life as a 10. This is the basic overall happiness score, and because these are national averages, none are as low as 0 or as high as 10.
* **change7:** 7-year change in LifeLadder score between 2008-2010 and 2015-2017.
* **forn\_ladder:** Life ladder score for foreign-born residents
* **local\_ladder** Life ladder score for locally-born residents
* **SE\_life:** standard error of life-ladder scores (a measure of variability among respondents)
* **LnGDPpc** : Log GDP per capita (2015-2017). The equation uses the natural log of GDP per capita (see below).
* **GDPpc**: GDP per capita (2015-2017). GDP per capita is measured in terms of Purchasing Power Parity (PPP) adjusted to constant 2011 international dollars, taken from the World Development Indicators (WDI) released by the World Bank divided by population.
* **LifeExp**: Healthy life expectancy at birth, based on slightly adjusted data from the World Health Organization (WHO) and WDI. These figures are in years. See the WHR2018 report for more details (optional).
* **SocSupp:** Social support is the national average of the binary responses (either 0 or 1) to the Gallup World Poll (GWP) question “If you were in trouble, do you have relatives or friends you can count on to help you whenever you need them, or not?” (can range from 0 to 1).
* **SEsoc:** standard error of SocSupp
* **LifeChoice**: Freedom to make life choices is the national average of binary responses to the GWP (Gallup World Poll) question “Are you satisfied or dissatisfied with your freedom to choose what you do with your life?” (can range from 0 to 1).
* **SEChoice**: Standard error of LifeChoice.
* **Generosity:** Generosity is the residual of regressing the national average of GWP responses to the question “Have you donated money to a charity in the past month?” on GDP per capita. Hence it measures a country’s deviation from global donation patterns. Can be positive or negative, and mean is approximately 0.
* **SEGen:** Standard error of Generosity
* **Corruption:** Perceptions of corruption are the average of binary answers to two GWP questions: “Is corruption widespread throughout the government or not?” and “Is corruption widespread within businesses or not?” Where data for government corruption are missing, the perception of business corruption is used as the overall corruption- perception measure. (can range from 0 to 1).
* **SECorr:** Standard error of corruption
* **OECD:** 0-1 dummy variable, with 1 indicating that country is a member of the OECD (Organisation for Economic Co-operation and Development)
* **Power:** The British newspaper *The Independent* publishes a list of the most powerful nations on earth. This 0-1 dummy identifies the powerful nations.

### Data Preparation:

The raw data file (WHR2018.csv) contains all of the data we’ll use. You will find the file on GitHub in our course repository. In addition, for the convenience of the team, I have created separate private repos through GitHub Classroom. Your team should again take advantage of version control to manage your work on this case.

Prior to clustering, you’ll want to explore the data and consider the extent of missing observations, correlations among columns, and make decisions about log and/or power transformations.

### Your Challenge:

Using the data provided, you will develop two pairs of hierarchical and *k-*means clustering models (four models in all) to group countries according to these guidelines:

1. **EXCLUDE** the LifeLadder column, and build clusters using some or all the remaining columns.
   1. Create a hierarchical **and** a *k-*means model with between 3 and 8 clusters. The team should use its judgement about the selection of input variables and about a suitable number of clusters. For the purposes of comparing models, select the **same** number of clusters for both.
   2. After you choose the clusters you prefer, report on the number of countries in each cluster, and prepare either a heatmap (see text Fig. 15.4) or centroid profile plot (see Fig. 15.5) to summarize the attributes of each cluster.
   3. For each cluster, list up to five countries assigned to that cluster.

With 156 rows of data, you could have clusters with a relatively large number of members. For highly populated clusters, simply show the names of a handful of countries to communicate the nature of the cluster.

* 1. Finally, write a short description of the main characteristics of each cluster. In other words, describe what the countries within a cluster seem to share.

1. **INCLUDE** the LifeLadder column, and build clusters using some or all of the remaining seven columns.
   1. Create a hierarchical and a *k-*means model with between 3 and 8 clusters. The team should use its judgement about the selection of input variables and about a suitable number of clusters. For the purposes of comparing models, select the **same** number of clusters for both, but you are free to vary the number from part 1.
   2. After you choose the clusters you prefer, report on the number of countries in each cluster, and prepare either a heatmap (see Fig. 15.4) or centroid profile plot (see Fig. 15.5) to summarize the attributes of each cluster.
   3. For each cluster, list up to five countries assigned to that cluster.
   4. Finally, write a short description of the main characteristics of each cluster. In other words, describe what the countries within a cluster seem to share.
2. Briefly compare the impact of including the happiness Life Ladder in the clustering analysis. Did it alter the composition of the clusters? Comment on the extent to which the happiness score captures something different from the other variables.
3. You now have 4 sets of clusters. Choose 1 of the four that you find most meaningful and explain which one you chose. Devise descriptive labels or titles for the clusters, analogous to the textbook introductory example of the public utilities as (*e.g.* “High Sales/Low Cost”) .
4. Write a very short paragraph addressing this question: Finally, is it accurate to generalize Tolstoy’s assertion to **countries** as well as families? Explain your thinking.

### Deliverable:

You will submit 2 files: a Word document and a PowerPoint file.

Your team should prepare a Word document created with R Notebook (markdown) reporting on your analysis and discussing your conclusions. In addition, prepare a two-slide PowerPoint (no title slide – just the content) presentation that summarizes the clusters you ultimately defined. One slide should show either a heatmap or a centroid profile plot, and the other should list the names of a few representative countries in each cluster. (Some clusters could have many countries—just select 5 or 6)

**In all you will upload two files: one from the knitted notebook, and one from PowerPoint.**

The R Markdown document must include:

1. a short introductory discussion of how the team elected to choose variables, to deal with missing data, and whether there were any other data preparation steps taken.
2. Relevant R code that you used for the four models you report. Suppress warnings, errors, and other messages. Use comments and prose liberally to explain the code as well as the output.
3. Some of the output will influence your final model selection. Be sure that you include output that was influential to the team’s decision-making.
4. In conclusion, offer a few bullet points with your own reflections about the strengths and weaknesses of the two main approaches to clustering.

1. Helliwell, J., Layard, R., & Sachs, J. (2017). World Happiness Report 2017, New York: Sustainable

   Development Solutions Network. Full text available at http://worldhappiness.report/#happiness207. [↑](#footnote-ref-1)