**ETL Project**

By: Elizabeth Vander Vorst, Tyler Vaughn, Alexis Speliotis, Puja Verma, Andrew Statz

**Background**

The 2020 world happiness report has just been released, and your boss at the international NGO where you work is not satisfied by the data points that are included in the report. She wants you to add other data points to see if there is any correlation.

From Kaggle.com you download the world happiness report and see that besides the happiness score, it includes:

* Social Support
* Healthy Life Expectancy
* Freedom to make life choices
* Generosity
* Perceptions of corruption

You return to Kaggle.com and find a collection of world datasets that have been scraped from Wikipedia and download 5 that you think your boss would want to see. They are:

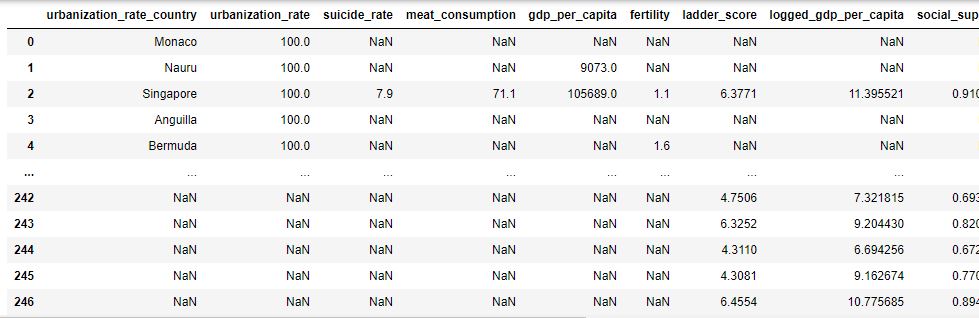
* Fertility rates
* GDP Per Capita
* Meat Consumption
* Suicide Rate
* Urbanization Rate

**Extract**

You open PGAdmin and create a table schema to hold the data from the csv files, each file in its own table. You then write a query to join all the tables together by country name.

**Transform**

Now that all the data are in one table, you need to make sure that the data are clean. You load the file into a Jupyter notebook and see that there are a lot of NaN values.



You start by removing all rows with NaN values. You then see that some of the column titles are confusing, so you rename ladder\_score to be happiness\_score, and urbanization\_rate\_country to be just Country. Finally, you decide that some of the scores are not easy to read, so you round them all to 2 decimal points.

**Load**

The last step is to load all this new data into a database so you can easily use it again. Since it is a list of scores grouped just by country, it makes sense to load it into a non-relational database like MongoDB.

**Conclusions**