**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:

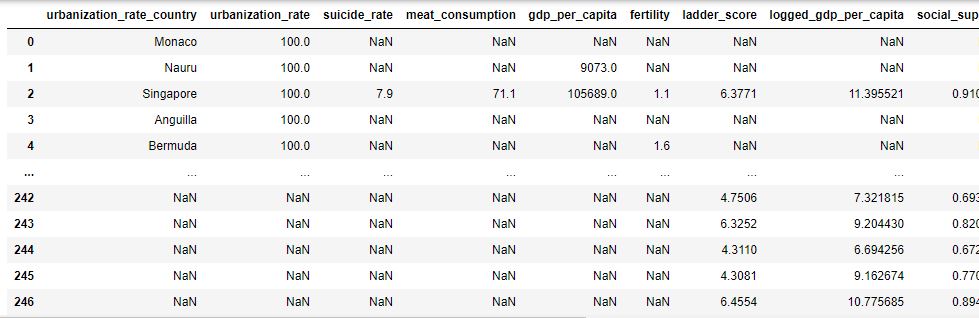
* Fertility rates
* GDP Per Capita
* Meat Consumption
* Suicide Rate
* Urbanization Rate

**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**

Our findings for this project concluded that the happiness level directly relates to GDP per Capita, Fertility, and Urbanization Rate.

Per capita gross domestic product is a metric that breaks down a country's economic output per person and is calculated by dividing its GDP by its population. Countries that have a high happiness level, for the most part, have a significantly higher GDP, with the highest GDP reaching above 100,000. When observing the countries on the lowest end of the happiness ranking, we see most at the 2,500 mark. The lack of opportunity to earn money and afford a decent lifestyle clearly impacts a person's happiness.

Detailing the percentage of the entire population living in urban areas is another indicator of a countries happiness and one that we were very curious about, seeing as how you could argue that urbanization might be good or bad. We found that countries with higher urbanization rates tended to be happier. This makes sense when factoring that people who live in urban/semi-urban areas usually have more access to opportunities, healthcare, and other institutions that make life easier.

Finally, the third indicator that stood out to us was the fertility rate, with happier countries having fewer children (1-2 children) and the lowest recorded countries having significantly more (4-5 children). This fact standing alone does not make much sense but combined with the knowledge we gained on the GDP of countries on the low and high end of the happiness scale, one can conclude that having many children in a country with little access to opportunity can create added stress of how to feed, educate, and house them.

These are just a few factors that result in a country being a happy place to live. We found some were obvious, like GDP, and some less so, like urbanization, and many that we did not analyze. All people are susceptible to the negative ramifications of a bad environment, and the key to being a happy country is creating a land with opportunity and access to resources.