# ETL PROJECT

**Group:**

Claudia Encinas

Tiago Lopes

Aldo Salomon

Bárbara Sánchez

## OBJECTIVE: WHICH ECONOMIC, SOCIAL, AND DEMOGRAPHICS INDICATORS AFFECT THE HAPPINESS INDEX?

Happiness is a very complex subject that we, as humans, have always tried to understand. There have been various studies that try to understand if different variables affect happiness. The easiest and most precise way to determine this is by looking at individual countries. For this project, we want to answer the question: **what affects countries' happiness?** There are differences in the opportunities and ways of living that each state has. Using the information available about country indicators, we will try to find the most relevant variables that determine happiness.

## ETL PROCESS

### DATA SOURCES:

For this project, we will be using two different sources.

1. **World Development Indicators by the World Bank - https://www.kaggle.com/unsdsn/world-happiness**

This database contains information about 247 countries. We chose this database because it has over a thousand indicators ranging from many different areas. It also reports for 55 years. Though these indicators have been recollected for a financial purpose, we believe they can be beneficial at analyzing happiness. We extracted one database that contained the country, indicator, year, and indicator value.

2. **World Happiness Report by the Sustainable Development Solutions Network (SDSN) - https://www.kaggle.com/worldbank/world-development-indicators**

This database is the most famous and acknowledged Happiness study and has been used by many for decision-making. Our information ranges for five years and 155 countries. The variables in this database are more complex than the previous one. The participant's data is transformed into different categories the SDSN previously established. In the end, we have an overall score. We extracted five databases, each one from every year ranging from 2015 to 2019.

We believe these two databases will work well together because one database has information that was determined to measure happiness. The other database is much more robust in indicators and, therefore, can also work into returning relevant insights.

## EXTRACT, TRANSFORM & LOAD:

Since we have used different data sources, we have joined the Transformation and Load process under one run in Python.

### MAIN STEPS HAPPINESS INDEX DATA – HAPPINESS.IPYNB:

1. Read each CSV file for happiness scores per year. (From 2015-2019).
2. Get the number of rows available in the dataset.
3. Get the number of countries per year to see if there were no repeated countries.
4. Renamed columns "Country" and "Score #Year" so we can have the same names for all columns. Due to the objective of the project, we disregard all other information (columns) not related to Happiness Index. Those columns had other parameters used to calculate the Happiness Index.
5. Merge information year by year to get an aggregated data frame. Here, we have considered just countries which are in all different years (Inner Join)
6. Including another field called "new_name" to accommodate all country name without spaces and with no capitalization.
7. Connect and upload all information to a Postgres database under the table "happiness."

During the process of validating coherence between data sources, we found there was a difference in the country field information that was retrieved. Therefore we need to create an "intermediary table with the conversion." Thus, we used Country Standards (ISO) information. Below you can see the process of extract and load this new data source.

### MAIN STEPS COUNTRY STANDARD INFORMATION – COUNTRY_CODE.IPYNB:

1. By using pandas, we have read the website: https://www.iso.org/iso-3166-country-codes.html
2. At this data table, we just used the country name and the Alpha 3 digits codes, since all information on World Development Indicators were built around those.
3. Connect and upload all information to a Postgres database under the table "country."

To ensure the correctness and the readability of all data mentioned above, we ran a couple of queries comparing how countries on the two above tables are linked. From the 140 countries mentioned on the *happiness* table, we were only able to find around 120 countries in the *country* table.

To correct this mismatches, we run the below steps under a FOR cycle:

1. Assign the happiness country_name not found in *country* table to a variable
2. Run a "Like" query with the step 1 over the country table assigning the results to a variable
3. Use this variable to update the original country_name under country table

Despite of this general efforts, we still had 6 countries not matching. On those we went into a manual process of identifying those one by one on both databases and do the necessary corrections.

After validating this information we end up with a list of 140 countries. For those we need to get all World Development Indicators. Thus, we entered on the below steps related with the second data source.

MAIN STEPS WORLD DEVELOPMENT INFORMATION – INDICATORS_ETL.IPYNB:

1. Read each Indicators CSV file.
2. Filter point 1 to just accommodate the 140 countries listed above, under the column CountryCode, IndicatorCode,Year and Value.
3. Reduce the range of the years to be evaluated, from 1960-2015 to 2000-2015. This step is important since we want to evaluate how the last generation influences the Happiness Index.
4. To ensure impartiality at the analysis, we are focused to just get the indicators that are presented the same years in all countries. Therefore, we built a group-by analysis to identify those.
5. With point 4, we have furthered filtered the data of point 3, to consider just the same indicators/comparisons all the time
6. Last, we connect and upload all information to a Postgres database under the table “indicators_data.”

Moreover, to keep track of the Indicator Name we created an additional table with IndicatorCode and IndicatorName. The same as uploaded into Postgres under table “indicators”.

Below you can see summary of all tables created in Postgres under ETL_Happiness database.