**ETL Project**

**Claudia Encinas**

**Tiago Lopes**

**Aldo Salomon**

**Bárbara Sánchez**

What affects countries’ happiness?

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 clearest 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 country has. Using the information available about country indicators we will try to find the most relevant variables that determine happiness.

**Extract**

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

1. **World Development Indicators by the World Bank**

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 extremely useful at analyzing happiness. We extracted one database that contained the Country, indicator, year, and indicator value.

1. **World Happiness Report by the Sustainable Development Solutions Network (SDSN)**

This is the most famous and acknowledged Happiness study and has been used by many for decision-making. Our information ranges for 5 years and for 155 countries. The variables in this database are more complex than the previous one. The participant’s information is transformed into different categories the SDSN previously established. In the end, we have an overall score, and based on that score each country is assigned its ranking. We extracted 5 databases, each one from every year ranging from 2015 to 2019.

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

**Transform**

Steps for the CSV´s for World Happiness Report (2015-2019).

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. Copied the Country and Happiness score columns of 2015 and 2016 datasets to create a new dataset with just this information.
5. Renamed columns “Country” and “Score #Year” so we can have the same names for all columns.
6. Copy the remaining years to our new dataset.
7. Connect to Database

Steps for CSV for World Development Indicators

1. Read CSV Indicators’ file.
2. Get information from data frame for an initial evaluation.
3. Look for duplicate values and remove them from data frame.
4. Create a new data frame to address consistency in data.
5. Get the number of rows available.
6. Group by country.
7. Get the number of years per country.
8. Get the countries where the information per year is different.
9. Get the years of the countries where we don’t have information.
10. Get years where we don’t have information for all countries.
11. Connect to Database

Steps for Integration of Data:

1. Connect to both database

**Load**

For integrate the data we need to make some adjustments due to differences on name country. First we get the country names that were different between Happiness tables & Indicators. We use the ISO name list for the country columns

For the Happiness score csv´s we load a table with the happiness score per year along with the countries so we can compare it with the World Development indicators.

**Findings:**

The main struggle with the tables was to integrate the country names. The integration on country names was by full name according to the ISO country names list. The happiness csv´s country names were not aligned to the ISO list so we have to include several steps on the integration process in order to consolidate both csv´s. After looking for the differences we conclude to eliminate one of the countries

**Sources:**

* <https://www.kaggle.com/worldbank/world-development-indicators>
* <https://www.kaggle.com/unsdsn/world-happiness>
* <https://www.iso.org/iso-3166-country-codes.html>