PROJECT #2 - Proposal - ETL
Members - Abla, Alex, Jordy, Viral

We can dissolve all the data to show the relation between “World Happiness” and the indicators alluding to the results.

Our data will be retrieved from Kaggle, and the World Bank – Data Catalog.

The sources of data that you will extract from.

- https://data.worldbank.org/indicator/AG.LND.ARBL.ZS?view=chart
- https://www.kaggle.com/unsdsn/world-happiness/version/2
- https://data.worldbank.org/indicator/SE.XPD.TOTL.GD.ZS?view=chart
- https://data.worldbank.org/indicator/BM.GSR.ROYL.CD?view=chart
- https://data.worldbank.org/indicator/MS.MIL.XPND.ZS?view=chart

The type of transformation needed for this data (cleaning, joining, filtering, aggregating, etc).

- CSV The table charges of the use of intellectual property: columns to be used are 2016 (which will be renamed to “intellectual property charges”). We are joining on country code. Renaming the table name to “Intellectual_Property_Charges”. Reading the table into python using pandas pd.read_csv function
- CSV Table of “World development indicators” : columns to be used are country code, short name (which will be renamed to country name), and 2016 (which will be renamed to “Arable_Land) columns. We are joining on country code. Reading the table into python using pandas pd.read_csv function. Renaming the table name to “Arable_Land”.
- XML Table of “Government expenditure on education, total (% of GDP)”: columns to be used are country name, government expenses on education, and year 2016 (which will be renamed to “Government expenditure on education, total (% of GDP)”). We are joining on country name. Renaming the table name to “Government_Expenses_on_Education”.
- Excel table of “Military expenditure (% of general government expenditure) ”, columns to be used are country name, 2016 (which will be renamed to “Military Expenditure”) columns. We are joining on country name. Renaming the table name to “military_expenses”
- CSV table “World Happiness Report”: columns to be used are country, happiness rank, happiness score, economy GDP per capita, health life expectancy, trust government corruption. Renaming the columns by removing the dots in the column names.

The type of final production database to load the data into (relational or non-relational).
The final database will be relational. Each of the data sets we are using has data for countries. We will load each on to mySQL and join them based on the name of country.

The final tables or collections that will be used in the production database.
The final table is a combination of multiple datasets and formats ranging from CSV, Excel, and XML. We are joining each dataset on Country Name, to show the relation between “World Happiness” and the indicators alluding to the results. The indicators are selected by each member which are found within the datasets indicated above. The results include “Country”, “Country Code”, “Happiness Rank”, “Happiness Score”, , “Arable Land”, “Economy GDP Per Capita”, “Government expenditure on education, total (% of GDP)”, “Health Life Expectancy”, “Income Group”, “Intellectual Property Charges”, and “Military Expenditure”, “Trust Government Corruption”.