**ETL Project Final Report**

[Naomi Scully](https://app.slack.com/team/U013YEDM6LX), Nicole Langford, Jiaping Shen

**Extract:**

How Happy are you? We decided to find out by choosing the [World Happiness Report](https://www.kaggle.com/unsdsn/world-happiness) from kaggle.com

This data was formatted in 4 CSV files (1 for each year, 2015-2018) and houses several categories/factors used to determine a happiness rating from 1-10 (10 being the highest).

We compared this data to the Gross National Income (xlsx file) in the [Income by Country Report](https://www.kaggle.com/frankmollard/income-by-country), also found on kaggle.com, to determine if higher income areas see a higher level of happiness.

Next we must transform the data.

**Transform:**

For the income dataset, we transformed the xlsx file into one DataFrame and selected columns of countries and income information of each country in years 2015-2018. Then, we converted column values into rows values to make the DataFrame more clear, by which we got a DataFrame with three columns “country”, “year” and “gni\_per\_capita”. After that, we sorted the rows by country and year.

For the happiness datasets we used a loop and transformed those 4 csv files into DataFrames with three columns “country”, “year” and “happiness\_score”, and then combined the four DataFrames together as one DataFrame and sorted the rows by country and year.

Finally, we joined the income and happiness dataframes together depending on the country and year, and a combined DataFrame with columns of “countries”, “year”, “gni\_per\_capita” and “happiness\_score” was obtained. Additionally, four DataFrames, with information in each year, were retrieved from the combined DataFrame.

**L**oad:

We chose to use a SQL database because our data was relational by date and country. We split our data into four different tables, by year, this way if we wanted to search for a particular country’s happiness score in a given year we could select the “country”, “year”, “gni\_per\_capita”, and “happiness score” columns to display after running the command. The tables are: combined\_2015, combined\_2016, combined\_2017, and combined\_2018. By creating 4 tables, we were able to use the country as the primary key, instead of creating composite keys for use in a single table. A non-relational database like MongoDB would not have been the best option to use because the variables change each year and it would have been more complicated to display that in a non-relational database.