ETL Project Final Report

Jill Peloquin, Rick Soljour, Tova Donahue

**EXTRACT**

* We sourced all of our data from Kaggle.
  + “World Happiness Report”
    - This is a series of csv files by year with several factors that accumulate into a happiness rating by country.
    - <https://www.kaggle.com/unsdsn/world-happiness?select=2018.csv>
  + “Ramen Ratings”
    - This is one csv file that lists the ratings of different brands of ramen
    - <https://www.kaggle.com/residentmario/ramen-ratings>

**TRANSFORM**

* World Happiness Report
  + For all three csv files there were minor differences in the wording of names of columns and countries. For example “USA” and “United States” are referring to the same country, so we just streamlined this information to make it more clear and readable.
  + We excluded columns that did not exist in all three csv files, and trimmed down the amount of information available. Our decision to keep only the country, rank, health life expectancy, and happiness score for all three sheets was to tell a clearer story with the data.
* Ramen Ratings
  + There were a few columns on this sheet that added no value – top ten rated and review number were dropped. Again, countries were listed out and compared to ensure that there were no similar duplicates.
  + There were a few unrated listings, which we deleted.

**LOAD**

* Each csv was given it’s own table in our database.
  + We chose to keep the Happiness years separate because there is not a one to one relationship to the Ramen data.
* We joined the Ramen table to each year of World Happiness to ensure the joins were effective.