ETL Project

Cody Braun and Sunitha Ramachandran

For this project we chose to analyze global Happiness rankings and their relationship to other information about a given country. Everyone is always trying to find happiness, so now they can just query our database!

**The sources of data that you will extract from:**

We are using data from the following two links:

* <https://www.kaggle.com/unsdsn/world-happiness>
* <https://github.com/iancoleman/cia_world_factbook_api#data>

The world happiness data is delivered in 3 .csv files, one for each of the years from 2015-2017. These files each contain the following columns:

* Country (Name of the country.)
* Happiness Rank (Rank of the country based on the Happiness Score.)
* Happiness Score (A metric measured in 2015 by asking the sampled people the question: "How would you rate your happiness on a scale of 0 to 10 where 10 is the happiest.")
* Whisker high
* Whisker low
* Economy (GDP per Capita) (The extent to which GDP contributes to the calculation of the Happiness Score.)
* Family (The extent to which Family contributes to the calculation of the Happiness Score.)
* Health (Life Expectancy) (The extent to which Life expectancy contributed to the calculation of the Happiness Score.)
* Freedom (The extent to which Freedom contributed to the calculation of the Happiness Score.)
* Trust (Government Corruption) (The extent to which Perception of Corruption contributes to Happiness Score.)
* Generosity (The extent to which Generosity contributed to the calculation of the Happiness Score.)
* Dystopia Residual (The extent to which Dystopia Residual contributed to the calculation of the Happiness Score.)

The CIA Factbook data is delivered in a JSON format and contains a large amount of data points about every country in the world. Because this data set is so large, we chose to select a few interesting data points to bring into our database.

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

We decided to only use the happiness data from 2017 because that is the date of the data from the CIA World Factbook. This data did not require any cleaning except for dropping the columns “Whisker high”, “Whisker low”, “Generosity”, and “Dystopia Residual”. We decided to drop these columns because they are not important to the analysis we want to do.

In order to pare down the World Factbook we filtered the large JSON document to a smaller data frame and joined this to the World Happiness data. From the World Factbook data we captured the following data points for each country:

* Population
* Population global rank
* Inflation rate
* Unemployment rate
* Population without electricity
* Type of government
* Percent of population with access to internet

The World Factbook data had the country names formatted differently than the Happiness data. In order to resolve this issue, we had to re-format the country names from the Happiness data by adding underscores in place of spaces in the multi-word country names. We also converted the country names to lowercase. Finally, some countries had completely different names between the datasets and we created a dictionary of these relationships so we could convert the names before filtering the World Factbook JSON file.

There were 2 null values in our dataset because the Happiness data contained 2 countries that the World Factbook does not recognize. These null values have been left in the database since they still have Happiness data.

These data points were then joined to the happiness data frame to create one data frame with all of the information. We then had to rename the columns to match the fields of our SQL database.

**The type of final production database to load the data into (relational or non-relational).**

MySQL database with one table directly mirroring our data frame.

**The final tables or collections that will be used in the production database.**

The final SQL database table has the following columns:

* + id int primary key,
  + country varchar(30),
  + happiness\_rank int(6),
  + family decimal(7,5),
  + economy decimal(7,5),
  + health decimal(7,5),
  + freedom decimal(7,5),
  + generosity decimal(7,5),
  + population\_total int(10),
  + population\_global\_rank int(3),
  + inflation\_rate decimal(5,2),
  + unemployment\_rate decimal(5,2),
  + population\_without\_electricity int(10),
  + type\_of\_gov varchar(60),
  + percent\_population\_internet decimal(5,2));

Some example queries that can be performed on this table are listed below. The results of these queries can be found in the iPython Notebook on our GitHub site.

* Selected all the rows and columns with all country data
* Selected only the top 10 happiness ranking countries
* Selected countries with Parliament of Republic