ETL project report

Mina BB Abad, Reyna zhang

Global Footprint Network

**Extract:**

The following three data sets were used to study the ecological footprint of countries taking into account the population and size of each country.

1. List of world’s countries by land was scraped form [Wikipedia](https://en.wikipedia.org/wiki/List_of_countries_and_dependencies_by_area) page into pandas data frame.
2. The ecological footprint data set of countries was downloaded from [Kaggle.com](https://www.kaggle.com/footprintnetwork/national-footprint-accounts-2018) in CSV format.
3. World carbon dioxide emission data by country was download from [theguardian.com](https://www.theguardian.com/news/datablog/2011/jan/31/world-carbon-dioxide-emissions-country-data-co2#data) in xml format.

**Transform:**

Data exploration was done by looking at the name of the columns (decoration of each table) to identify variables and finding missing values in each data frame.

The three main variables in this study that connect the three data sets are country names, country code, and year. In the Co2 emission data frame, year information was stored in row so before anything, we reshaped that data frame.

While all the data frames had one common column named “country”, some countries were named differently in each data. In addition, one data frame was missing ISO-code, while the other two were using alpha2 and alpha3 ISO-codes. The name of those countries were changed manually in Co2 emission and NFA data frames and then the ISO-code column was merged into them.

Some countries in the data frames that no longer existed were dropped including world that contained total values.

Name of some columns were changes for the sake of clarity.

Multiple sub data frames were created to be saved in mySQL.

**Load**:

A Mysql engine was created. In mySQL all the tables along with their primary key and foreign keys before importing the data.