ETL PROJECT

By

SOUMYA PATIL

JERIN BLACK

# Extraction : Original Data Sources , data formatting

We used three different data sources for our Project.

* + - * Data Source#1 <https://www.kaggle.com/londeen/world-happiness-report-2020>
      * Data Source#2 <https://www.worldometers.info/world-population/population-by-country/>
      * Data Source#3 <https://www.worldometers.info/demographics/life-expectancy/#countries-ranked-by-life-expectancy>
* Python’s BeautifulSoup library was used to pull data out of the websites. The websites were parsed to navigate and fetch the tables , which were later-on converted to dataframes.

# Transform : What data transformation was required

* We renamed the columns .
* Filtered the specific list of columns that would be worked on.
* Converted datatypes to maintain consistency across all Tables.
* Truncated symbols in the cells , and kept data like percentages at their face value.

# LOAD : DATABASE CONNECTIONS, AND CHOICE OF DATBASE

We chose Postgres as our Database because we didn’t have to deal with a lot of unstructured and variety of data.

A total of 3 tables were created , and record insertions was done using pandas.