Extracting, Transforming, and Loading Earthquake, Population Size, and GDP Data:
To Predict the Impact of Future Earthquakes

Team Members: Arjun, Christy, Ivy, Shifaa (Team 6)

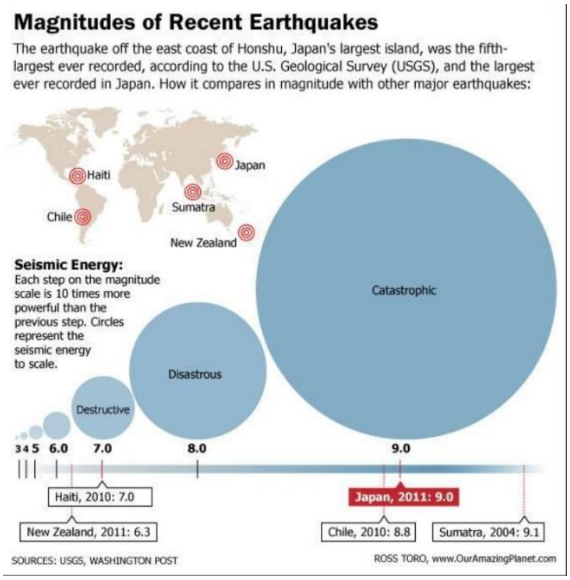
Project Description:

We extracted earthquake data for the last 200 years using the USGS earthquake API and found the nearest city for each earthquake. We also found the current population size of those cities and their country's current GDP. We then transformed and loaded the data into a SQL database so that future analyses could occur.

Earthquake Data

Extract

We extracted earthquake data for the last 200 years using the USGS earthquake API: https://earthquake.usgs.gov/fdsnws/event/1/. We only pulled data for earthquakes with magnitude greater than 6. We chose this cutoff because magnitude greater than 6 is when damage actually occurs. The data from the API was in JSON format.



Transform

First, we only kept the following variables: id, mag, place, timestamp, tsunami (yes/no), eq_lon, and eq_lat. We felt that these variables were important to use when understanding magnitude and location of the earthquake. Then, we identified the nearest city and corresponding country code for each earthquake using citipy. Since id is the primary key in the earthquakes SQL table, we dropped any duplicate rows with the same id.

Load

We created a new SQL database (earthquakes_db) and a SQL table (earthquakes) to hold the extracted and transformed data. Then, we loaded the earthquakes data into the SQL database.

Population Size Data

Extract

We downloaded population size data from a csv file on a Kaggle: https://www.kaggle.com/max-mind/world-cities-database.

Transform

First, we only kept the following variables: city, country, and population. We felt that this data was important to put into our population size SQL table. We dropped any rows with missing data. We dropped any duplicate rows with the same city name.

Load

We created a new SQL table (population_size) to hold the extracted and transformed data. Then, we loaded the population size data into the SQL database.

GDP Data

Extract

We extracted current GDP data for all countries by scraping the UN Statistics website: https://unstats.un.org/unsd/snaama/Index. We extracted data for GDP at current prices for all countries for 2017.

Transform

We converted the GDP into USD in millions and to an integer data type. Then, we only kept the following columns: Country/Area, Year, and Gross Domestic Product (GDP) (Millions). Then, we renamed some countries so that we could use PyCountry to identify country codes. Then, we used PyCountry to identify country names and country codes.

We saved this data in a separate dataframe. Then, we merged both dataframes on country. We only kept the following variables: country name, country code, year, and gdp in usd millions. Lastly, we dropped any duplicate rows with the same country code.

Load

We created a new SQL table (GDP) to hold the extracted and transformed data. Then, we loaded the GDP data into the SQL database.

Future Analyses:

With the data we extracted, transformed, and loaded into this SQL database, we will be able to calculate a risk predication number to determine the impact that an earthquake today would have on the cities based on previous earthquake history. The risk prediction number could factor in the number of people who would be affected (population size) and the resources countries have to rebuild and recover after an earthquake (GDP). As such, the database we created could be used to predict the impact of future earthquakes.