**World Suicides Report**

**ETL Project**

**Team Members: Guillermo Huertas, Ashish Karki, Cheryl Hornung**

**E**xtract:

* scrape.ipynb
* <https://www.kaggle.com/russellyates88/suicide-rates-overview-1985-to-2016>
* <https://www.kaggle.com/nikitagrec/world-capitals-gps>
* Pandas used to read CSV files and create Dataframes
* world\_suicides\_df
  + master.csv -> 1985-2016.csv
* lat\_long\_df
  + world\_capitals.csv -> concap.csv

**T**ransform:

* scrape.ipynb
* Columns removed from world\_suicides\_df
  + country-year
  + HDI for year
  + gdp\_for\_year ($)
  + gdp\_per\_capita ($)
* Columns removed from lat\_long\_df
  + CountryCode
* Changed the names of some countries in our world\_suicides\_df to match the lat\_long\_df before loading them into pgAdmin so it did not result in a foreign key violation within the database
  + 'Cabo Verde': 'Cape Verde'
  + 'Republic of Korea': 'South Korea'
  + 'Russian Federation': 'Russia'
  + 'Saint Vincent and Grenadines': 'Saint Vincent and the Grenadines’
* Changed the names of columns in world\_suicides\_df to match database schema
  + 'country': 'country\_name'
  + 'suicides\_no': 'suicide\_number'
  + 'suicides/100k pop': 'suicides\_per\_100k'
* Changed the names of columns in lat\_long\_df to match database schema
  + 'CountryName': 'country\_name'
  + 'CapitalName': 'capital\_name'
  + 'CapitalLatitude': 'capital\_latitude'
  + 'CapitalLongitude': 'capital\_longitude'
  + 'ContinentName': 'continent\_name'

**L**oad:

* scrape.ipynb
* QuickDBD used for creating database schema and ERD diagram
  + schema.sql
  + ERD.JPG
* Loaded schema into world\_suicide\_db within pgAdmin
* Loaded the Dataframes containing the CSV file data using Pandas into pgAdmin
  + Updated schema created using QuickDBD to allow null values in capital\_name field
* Used pgAdmin to query the tables to validate functionality
  + queries.sql