Project: Deaths by Natural Disaster

Collaborators: Bonnie-jo Barnaby & Charles Lindner

Sources:

* World Health Organization Global Health Observatory
  + <http://apps.who.int/gho/data/node.sdg.13-1-data?lang=en>
    - WHO\_Disaster\_Data.xls
    - WHO\_Disaster\_Data\_Summary.csv
* International Red Cross and Red Crescent Federation
  + <http://data.ifrc.org/fdrs/data-download>
    - RC\_Disaster\_Data.xml
    - RC\_Disaster\_Data.csv
    - RC\_Disaster\_Codes.xlsx

Data Cleanup & Analysis: (items in yellow are TBD)

* Sources of data extracted from: xls, xlxs & csv
* Type of transformation needed
  + Cleaning
  + Joining
  + Filtering
  + Aggregating
* Type of final production database to load into
  + Relational - Postgres
* Final tables or collections used in production database
  + TBD

For Tables in Postgres - Columns – REDCC:

1. Country
2. Years – 2012 to 2016
3. GDP
4. LifeExp
5. Population
6. Child Mortality
7. Maternal Mortality

Cleanup process for REDCC

* Created dataframe from CSV file
  + List of headers
  + Selected columns we wanted
* Filtered out data to only include years that match WHO data (2012-2016)
* Grouped by Country and averaged data columns
  + GDP
  + LifeExp
  + Population
  + Child Mortality
  + Maternal Mortality
* Added a new column “Year” which represents the averages of 2012-2016
* Exported as CSV file

For Tables in Postgres – Columns – WHO:

* Country
* Mortality
* Land Area
* Region
* Gender
* Years – 2012 to 2016

Cleanup process for WHO

* Determine sheets we wanted to pull data from (data-text and Country sheets)
* Created two dataframe from excel file (data-text and Country)
  + List of headers for each file
  + Selected columns we wanted
  + Renamed column headers
* Merged both dataframes on “Country” name column
* Filtered column “Gender” to get combined morality rate for both sexes by Country
* Exported as CSV file

Project Report

* E (xtract): Original data sources and how the data was formatted: xls, xlxs & csv
* T (ransform): What data cleaning or transformation was required
* L (oad): The final database, tables/collections and why this was chosen
* Both statistical files reporting at 1 per 100,000 people

Upload

* Upload final report to Github and
* Submit a link to Bootcampspot under ETL Project (Due Tuesday, August 27, 2019 – end of class)