Project Proposal Document:

Team: Cesar, Yen, Roma, Bill

Proposal Date: 02Nov2019
Due Date: 11/6/19

Statement of work: The objective of the analysis is to retrieve and clean data sets in an effort to explore correlations between suicide and other factors such as age, gender, geographic location and HDI (Human Development Index). The hope is to identify risk factors in an effort to prevent suicide.

Data sources:

1. World Health Organization

   http://apps.who.int/gho/data/node.main.MHSUICIDE?lang=en)

2. World Bank

   https://databank.worldbank.org/source/gender-statistics/preview/on

3. United Nations Development Programme

   http://hdr.undp.org/en/indicators/137506#

Proposed ETL:

Extract:

1. Identify data sources
2. Upload 3 data files using pandas

Transform:

1. Remove unnecessary columns, null values, and reshape the data using pandas.melt()
2. Review datasets to ensure key fields match up
3. Prepare data for uploading to database

Load:

1. Create ERD and get schema for each table
2. Connect to database in Jupyter Notebook
3. Upload dataframes to database tables (3 tables)
4. Join tables
5. Display final table of summary data

Proposed Final Schema: PostgreSQL

Approved by: Satish Anthony 02Nov2019

Final Project Report (11/6/19 or 11/7/19)

At about 8 PM, your team will submit a Final Report that describes the following:

- **E**xtract: your original data sources and how the data was formatted (CSV, JSON, pgAdmin 4, etc).
- **T**ransform: what data cleaning or transformation was required.
- **L**oad: the final database, tables/collections, and why this was chosen.

Please upload the report to Github and submit a link to Bootcampspot.

Present 3-4 minutes on the project discussing some pain points and how did you resolve them. Only one student from each team should present. 1 min for any Q&A to the class.