**Overview**: Creation of a MySQL database containing state level suicide, drug overdose, and unemployment information for the years 2015-2017.

**Data Sources:**

* Annual Suicides by State (CDC) | <https://www.cdc.gov/nchs/pressroom/sosmap/suicide-mortality/suicide.htm>
* Annual Drug Overdose by State | <https://www.kaggle.com/cdc/vsrr-provisional-drug-overdose-death-counts>
* Unemployment Statistics by State | <https://www.bls.gov/web/laus/laumstrk.htm>
* State Lookup Table | <http://www.fonz.net/blog/archives/2008/04/06/csv-of-states-and-state-abbreviations/>

**Steps**:

EXTRACTION

1. Annual Suicides
   1. Download CSV files from CDC website, separate file for each year
   2. Load csv files into dataframes
2. Annual Drug Overdose
   1. Download CSV file from Kaggle, contains all years
   2. Load csv into a dataframe
3. Unemployment Statistics
   1. Use pandas read\_html function to extract dataframe for each year from html table

TRANSFORMATION

1. Annual Suicides
   1. Merge dataframes on state key
   2. Renamed columns
   3. Output to CSV
2. Annual Drug Overdose
   1. Drop unnecessary columns
   2. Sub-select for rows describing overall drug overdose deaths (and not deaths in general or drug-specific deaths)
   3. Sum up deaths using groupby
   4. Pivot the year column to separate year columns by state for the death totals (one column for each year death total)
   5. Output to CSV
3. Unemployment Statistics
   1. Merge dataframes for each year on state name
   2. Drop NaN
   3. Load State Code Lookup Table into dataframe (States are only listed by name, while our other data sets have states by code (ex. CA))
   4. Merge state dataframe with state code
   5. Output to CSV
4. Merging of Datasets
   1. Load CSV files for all 3 datasets
   2. Merge on State Code
   3. Rename columns

LOAD

1. Create database in MySQL
2. Create engine in python and send merged dataframe (within same script) to MySQL database – called etl\_project