**ETL Project Report**

For the ETL project, we aimed at extracting crime and education data variation across counties in the USA to facilitate possible assessment of the relationship between crime and education. The following are our data resources:

<https://www.kaggle.com/mikejohnsonjr/united-states-crime-rates-by-county>

<https://www.ers.usda.gov/data-products/county-level-data-sets/download-data/>

<https://www.kaggle.com/laptou/us-fips-codes-for-states>

<https://www.kaggle.com/stansilas/us-state-county-name-codes#US%20Countiess.csv>

Crime data was extracted from [kaggle.com](http://www.kaggle.com) in comma-separated values (csv) file and while education from usda.gov in excel file which were uploaded onto pandas data frames. The data was cleaned by droping irrelevant notes present from the excel files of education data and important columns were extracted. From both data files FIPS (Federal Information Processing Standard) codes were common which we assumed as primary key linking both data sets.

The data was uploaded into pandas data frames and four tables including state, county, crime and education were created by extracted important columns from main data frames.

A SQL database (Crimes\_db) was created on MSQL and four tables including us\_states, us counties, us\_crimes, us\_education were made in the database. The choice of tables was because we aimed at making a database where the relationship between crime and education can be analyzed. Also, crime variation per county or state all over USA can be pulled out of our dream database. The transformed data was loaded unto SQL database using MySQL.Connect function in pandas.

**Challenges**

* Cleaning education data was challenging as it was available in excel with irrelevant notes at the top.
* To extract the state and county codes from FIPS code and create two new columns was a major challenge.
* Create a new header variable from the first row of the data frame
* Replacing the data frame with a new one which does not contain the first row
* Rebuild the header and reset the indexing was another one
* Dropping rows where the columns are null for column 2003 Rural-urban Continuum Code
* Extract the state and count from FIPS code and create two new cols
* Refomat the education data frame to the rows that will be loaded into the education table

The figure beloow presents EER diagram for the database (crimes\_db)

