# ETL Project Proposal

Team Members:

Tasneem Bhaijee , Hung Nguyen

Extract, transform and load data to answer below questions:

1. Literacy rate impact on suicide rate in **European** countries
2. Literacy rate, adult total (% of people ages 15 and above) on suicide rate

Sources:

|  |  |
| --- | --- |
| **Data File Name** | **Sources** |
| Master | <https://www.kaggle.com/russellyates88/suicide-rates-overview-1985-to-2016?select=master.csv> |
| Population by educational attainment level, sex and age (%) - main indicators [edat\_lfse\_03] | Eurostat |
| Literacy rate, adult total (% of people ages 15 and above) | <https://data.worldbank.org/indicator/SE.ADT.LITR.ZS> |
| datasets\_23752\_30346\_countries of the world | Worldbank |
| Country rank order | https://www.cia.gov/library/publications/the-world-factbook/rankorder/rawdata\_2127.txt |

Create PostgreSQL database name suicide rate.

Possible transformation:

Join multiple data sources to link related information based on place, time for each country.

Data Transformation:

1. Master files:

List of columns: Country, year, sex, age, suicides\_no, population, suicids/100k pop, country-year,

HDI for year, gdp\_for\_year, gdp\_pr\_capita, generation

Columns to keep: country, year, sex, age , suicides\_no, population

1. Edat\_lfse\_03\_1\_Date: For European countries keep the date for Tertiary education
2. Education Stat Data : Drop column data before year 1985

Drop column country code

Drop column Indicator code

Select records for Barro-Lee indicator name and only total % for each age group

Keep data only upto 2014

Use melt function to convert year from column to row for each country

1. Created logical Data model using QuickDBD software
2. 6 Tables, Names: country, sex, age, year, suicide data and literacy data design to store related information for transform data.
3. Created PostgreSQL data base name ETL\_db . created all 6 tables and import data from csv files.
4. Data