ETL Project – Final Report

Suicide Rates vs Global Happiness Report

By Luis Bejaran, Genevieve Sloup, Mark Gu

Extract

We used 3 datasets from the public platform Kaggle, all downloaded as csv files.

- WHO Suicide Statistics:
    - Basic historical (1979-2016) data, count of suicides by country, year and demographic groups from the World Health Organization.
- Global Happiness Report:
    - 2015 and 2016 reports
    - A ranking of 155 countries by their happiness score as measured by the Gallup World Poll survey. Survey participants rated the following six factors on a scale of 0-10: GDP per Capita, Family, Life Expectancy, Freedom, Generosity, Trust Government Corruption, and Dystopia – as a countermeasure.

Transform

Global Happiness Report:

We had two csv files that we needed to join to create one dataframe for 2015 and 2016. We had to add a column for year to each csv, being that the year was only indicated in the filename as each file was one year. We also removed columns that did not appear in both years. We then renamed the columns in proper format for SQL.

WHO Suicide Statistics :

Upon examining this dataset, we discovered several countries with missing data, so we performed drop NA to eliminate those datapoints. Since we only used 2015 – 2016 from the Happiness Report, we did not need to load 47 years worth of Suicide Statistics into our database so we filtered for just 2015 and 2016. We then used number of suicides and total population to calculate the suicide rates and added that column.

Load

The data is loaded into two tables in a PostreSQL database called “ETL_Project_”. We decided to use a relational database with each table using a serial ID as the primary key. Both tables have year and country columns that they can be merged on.

Schema

Potential Uses

This data could be used to analyze what external societal factors (as opposed to internal - mental health, for example) may play a role in higher suicide rates. It could also examine whether certain happiness factors present higher risk for various demographic groups.

- Are there significant differences for demographic markers?
    - Male/Female
    - Age
- As yearly Happiness Rank changes, is suicide rate impacted?
- How do the 6 measures of Happiness impact rate of suicide?
    - GDP per Capita
    - Family
    - Health
    - Freedom
    - Govt Trust
    - Generosity
    - Dystopia