**ETL Project**

**Anna Wrobel**

**Nupur Gupta**

# Background

**For this project, we decided to work on two datasets- World Happiness Data (2005-2018) and World Suicides Data (1985-2016).**

# Possible business questions we can answer/ analysis we can perform using these data-sets:

* Are suicide rates higher for men or women?
* Do countries with a high feeling of social support have lower suicide rates and is the relationship different for each gender?
* Is there a relationship between life ladder and number of suicides?
* Do countries with a high GDP per capita have a higher result for generosity?
* And many more!

# **The ETL Process:**

# Extraction

(location: Folder ‘Extracted Raw Files’)

**World-Happiness Dataset:**

(location: ‘Extracted Raw Files/world-happiness-report-2005-2018.csv’)

The World Happiness data set was “extracted” from **Kaggle (**[**link**](https://www.kaggle.com/alcidesoxa/world-happiness-report-2005-2018)**) as a csv file.**

The original source of the data can be traced back to the [World Happiness Report 2018](https://worldhappiness.report/ed/2018/) website. In addition to presenting updated rankings and analysis of life evaluations throughout the world, each World Happiness Report has had a variety of topic chapters, often dealing with an underlying theme for the report as a whole. For the World Happiness Report 2018, special focus was on migration. The report describes in detail, the various variables based on which happiness of each country has been calculated. The following variables were evaluated using some basic questions, and every nation’s results for different years have been tabulated in the dataset:

* Country name
* Year
* Life Ladder or the “Happiness Index”- Please imagine a ladder, with steps numbered from 0 at the bottom to 10 at the top. The top of the ladder represents the best possible life for you and the bottom of the ladder represents the worst possible life for you. This ladder is known as the Cantril Ladder. The Happiness Index - or Life ladder - is simply calculated by averaging the answers to this Cantril Ladder to a single number.
* Log GDP per capita- GDP per capita is in terms of Purchasing Power Parity (PPP) adjusted to constant 2011 international dollars, taken from the World Development Indicators (WDI) released by the World Bank in September 2017. The equation uses the natural log of GDP per capita, as this form fits the data significantly better than GDP per capita.
* Social support
* Healthy life expectancy at birth
* Freedom to make life choices
* Generosity
* Perceptions of corruption
* Positive affect
* Negative affect
* Confidence in national government
* Democratic Quality
* Delivery Quality
* Standard deviation of ladder by country-year
* Standard deviation/Mean of ladder by country-year

Extra reads: [trackinghappiness.com](https://www.trackinghappiness.com/happiness-index-2018/)

**Suicides Dataset:**

(location: ‘Extracted Raw Files/suicide\_data.csv’)

The original data sources for the suicide data set are:

* United Nations Development Program [(link](http://hdr.undp.org/en/indicators/137506))
* World Bank [(link](http://databank.worldbank.org/data/source/world-development-indicators))
* [Szamil] Suicide in the Twenty-First Century [(link](https://www.kaggle.com/szamil/suicide-in-the-twenty-first-century/notebook))
* World Health Organization Suicide prevention ([link](http://www.who.int/mental_health/suicide-prevention/en/))

This data set was extracted from [Kaggle](https://www.kaggle.com/russellyates88/suicide-rates-overview-1985-to-2016) and the format is CSV. This data contains information about global suicide rates, population, gender, age and GPD from 1985 to 2016.

# Transformation

(location: ‘ETL\_World\_Happiness\_Suicide.ipynb)

**In the Jupyter Notebook, both the csv files were read as Pandas Dataframes using Pandas ‘read\_csv ()’ function.**

**World-Happiness Dataset:**

The happiness csv dataset downloaded from Kaggle had most of it’s data in it’s first column, delimited by a ‘ ; ‘. Therefore, after reading the csv into a data frame, the following steps were followed to transform the data:

* Delimited to extract column wise data using str.split() function
* Dropped all empty columns or columns with very less data
* Renamed the columns appropriately

After the above steps, we had with us the 16 majors columns that we can use for analysis and answering different business questions.

* Checked to see the datatypes of each column
* Corrected the datatype of all float/ integer columns from object to numeric to make data manipulation and plotting easier
* Set the index for the data frame to be ‘Country’

**Suicides Dataset:**

Transformation for the suicide data set includes:

* To make the timelines match up for both data sets, information outside of years: 2005-2016 was removed. There was very little data for the year 2016 so data for this year was also removed.
* Columns were renamed appropriately.
* The ‘age’ column had 6 age groups which was causing a lot of duplicates in the data set. We also didn’t need this data for our analysis so the column was removed.
* The following columns were removed because they were not needed for our analysis: ‘country-year’, ‘population’, ‘suicides/100k pop’, ‘generation’.
* The ‘HDI for year’ column was removed because it was missing a lot of data.
* The number of suicides were broken down by age group and after the ‘age’ column was removed the number of suicides was summed up per country and year.
* The commas were removed from the ‘GDP per year’ column and converted to float.

# Loading

(location: ‘ETL\_World\_Happiness\_Suicide.ipynb’ , ‘World\_happiness\_suicide\_query.sql’

To create the **relational database**, we loaded the **‘suicide’** and **‘world\_happiness’** datasets into PostgreSQL database.

1. First created a database named ‘world\_happiness\_suicides’ in PostgreSQL
2. Created tables/schemas for both the datasets- ‘suicide’ and ‘world\_happiness’, using exact same names as their columns. Assigned the ‘Country’ column and ‘Year’ column as composite primary key for ‘world\_happiness’ table, as they together, make every row/ record unique in the table
3. Connected to the local ‘world\_happiness\_suicides’ database using Python
4. Confirmed the connection by displaying names of the tables created under this database in step 2- ‘suicide’ and ‘world\_happiness’
5. Used Pandas to load csv converted ‘world\_happiness’ and ‘suicide’ dataframes into database
6. Ran queries on both, PostgreSQL and Jupyter Notebook, to confirm data addition into the database
7. Performed an **‘Inner join’** on the two tables on ‘country’ and ‘year’ columns, as this would yield the results for all those countries and years, for which data is available in **both** the tables; on PostgreSQL and Jupyter Notebook