**Are We Really Fat, Drunk and Happy?**

**Project Overview**

Our ETL project focuses on collecting and creating an SQL database that serves as an alternative to the World Happiness Data and Report published by the Sustainable Development Solutions Network. The World Happiness Report ranks 158 countries by how happy their citizens perceive themselves to be. The rankings are calculated using seven indexes derived from Gallup World Poll Data. Rather than relying on seven indexes that constitute the World Happiness Dataset, our project offers four new datasets, all organized by country, that we believe, without any scientific proof, may be related to aggregate national happiness calculations. This database could be used to further analyze happiness and perhaps answer the age-old question: *are we really fat drunk and happy, or just fat and drunk?*

**Extraction**

Our baseline data set is the World Happiness Rankings for the year 2015. This data was retrieved from Kaggle as a .csv file. The data is organized by country and region (mostly sub-continental geographic areas). Our extractive efforts focused on collecting the five additional datasets: described below:

1. Alcohol Consumption: collected as a .csv file from Kaggle published by the World Health Organization (WHO).
2. Obesity: collected as a .csv file from the WHO website published by the WHO.
3. Gross Domestic Product: collected as a .csv file from Kaggle published by the World Bank
4. Suicide Rates: collected as a .csv file from Kaggle published by the World Bank
5. Estimates of Mental Health Disorder Prevalence: collected as a .csv file from Our World in Data published by the Institute for Health Metrics and Evaluation.

**Transformation**

Our data transformation efforts largely centered on converting each csv file to Pandas data frames and cleaning up the primary key column, country, of each file. All indexes were reset to the primary key. We also spent time converting all other column titles to the same lower-case string format, performed a join of the GDP and Happiness dataframes, and dropped null values and other extraneous values and columns from our data sets. One interesting transformation was necessary when cleansing the obesity data from the WHO. In this case the column in the .csv file that represented the percent of the population that was considered obese also contained supplemental data that we felt was unnecessary. We used a split function to separate the relevant data from the superfluous data in this column and then completely removed the superfluous strings from our list.

**Loading**

We decided on creating a relational database for all the data we collected. Using the create engine tool from sqlalchemy, we loaded each data frame as a table into an SQL database. We verified that the data loaded correctly and since we could join the data based on a common key of country, we decided to utilize a relational database in order to discover a deeper understanding of how different factors could lead to varying levels of perceived happiness in different countries.