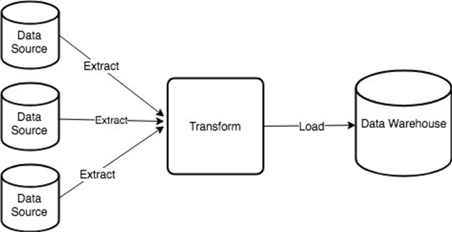
ETL Report

Megan Lane, Maria Iglesias, Rob Giese

Alcohol Consumption & Happiness Around the World



1. Extract
   1. Dataset Sources: We selected two datasets from the Kaggle data sharing website. Kaggle is a data sharing website that supports a variety of dataset publication formats where dataset publishers are strongly encouraged to share their data in an accessible, non-proprietary format. The two datasets we selected were:
      1. **World Happiness Report, 2016** (published by the Sustainable Development Solutions Network): The World Happiness Report is a landmark survey of the state of global happiness. The reports review the state of happiness in the world today and show how the new science of happiness explains personal and national variations in happiness.
         1. *Variables*: Country name, Region, Happiness Rank, Happiness Score, Lower Confidence Interval, Upper Confidence Interval
            1. Happiness Variables: extent to which these factors contribute to the calculation of the Happiness Score

Economy (GDP Per Capita)

Family

Health (Life expectancy)

Freedom

* + - 1. <https://www.kaggle.com/unsdsn/world-happiness>
      2. File format: CSV
    1. **Happiness and Alcohol, 2016** (published by Marcos Pessotto): The World Happiness Report is a landmark survey of the state of global happiness. The reports review the state of happiness in the world today and show how the new science of happiness explains personal and national variations in happiness.
       1. Variables: Country name, Region, Hemisphere, Happiness Score, Human Development Index, GDP Per Capita, Beer Per Capita, Spirit Per Capita, Wine Per Capita
       2. <https://www.kaggle.com/marcospessotto/happiness-and-alcohol-consumption>
       3. File Format: CSV
  1. The two datasets were downloaded as CSVs and saved to a local repository. A quick glance of the tables showed that there were 122 countries listed in the alcohol file and 157 countries listed in the happiness file. I also noticed that some of the country names were different between files, e.g. Russian Federation vs Russia. I used VLOOKUP in Excel to see which countries were on both lists with the same name. I then looked at unmatched countries and changed the names to match if there were matching countries. Altogether, this required 4 country name changes.
  2. The updated CSVs were then saved. A Jupyter notebook was started and the Pandas library was imported. During the extract phase, the data was converted from csv files to pandas dataframes and then merged together using country as the merging variable. This resulted in a dataset with 21 columns and 122 rows.

1. Transform
   1. Because there was overlapping data (i.e. the happiness score) in both datasets, some columns with duplicate data were then eliminated and the columns were reordered in a more intuitive way.
   2. Specific columns were picked to create the data frame desired, and columns were renamed to eliminate spaces. Columns were also renamed to make them compatible with PgAdmin.
   3. Country was set as the index for the data frame.
   4. The happy\_alcohol\_db database was created in PgAdmin, which was then followed by creating a connection from Pandas to PgAdmin.
   5. Created the table and then loaded the dataframe to SQL.