Group ETL-Project Report Data Viz 2019

Clay Franklin

Gabbie Cutaiar

Megan Finley

Proposal and Extraction:

For our analysis we proposed to look at the relationship between happiness, suicide rate, and alcohol consumption based on found data. The first dataset found was a Kaggle resource CSV that related countries, region, happiness, GDP, HDI, and alcohol consumption broken down into beer, wine and spirits. To relate another set of information we looked for suicide rates by country from the World Health Organization (WHO). To reduce the dataset to fit the year range displayed in the first dataset, we used WHO’s website to sort by year and select 2016 data. This dataset was also extracted in CSV form.

Transform:

To transform the data in a form to fit our analysis we imported the csv files into pandas using jupyter notebook. The first step was to identify what was similar in both datasets. This would provide the means on which we could join these two datasets. Country names where chosen. The alcohol data set was sorted by country name and set it to ascend to return country names alphabetically. We then set the index to country. The columns in the suicide dataframe were renamed. We dropped the first line in our table because it was redundant. We then used a pivot function to order the table by country, which then changed the rows for sex to columns: Average, Female, and Male. We then joined the two tables on country. We noticed that some of the countries from WHO did not have corresponding happiness/alcohol data in the Kaggle data frame. We created two datasets. The first dropped all the rows with non-values. The second was the full data join with rows including non-values. Both datasets were exported as CSV files.

Load:

We selected PostgreSQL as the location for data storage. We created two table schemas to import the data. We imported the CSV’s into suicide\_alcohol\_full, and suicide\_alcohol\_reduced tables. We ran quires on the database to ensure the import was successful. We ran additional queries to look for correlations.

We choose this storage type because it uses relational databases and would be easy to expand our proposal by adding addition data sources that may correlate with happiness, suicide rate, or alcohol consumption.