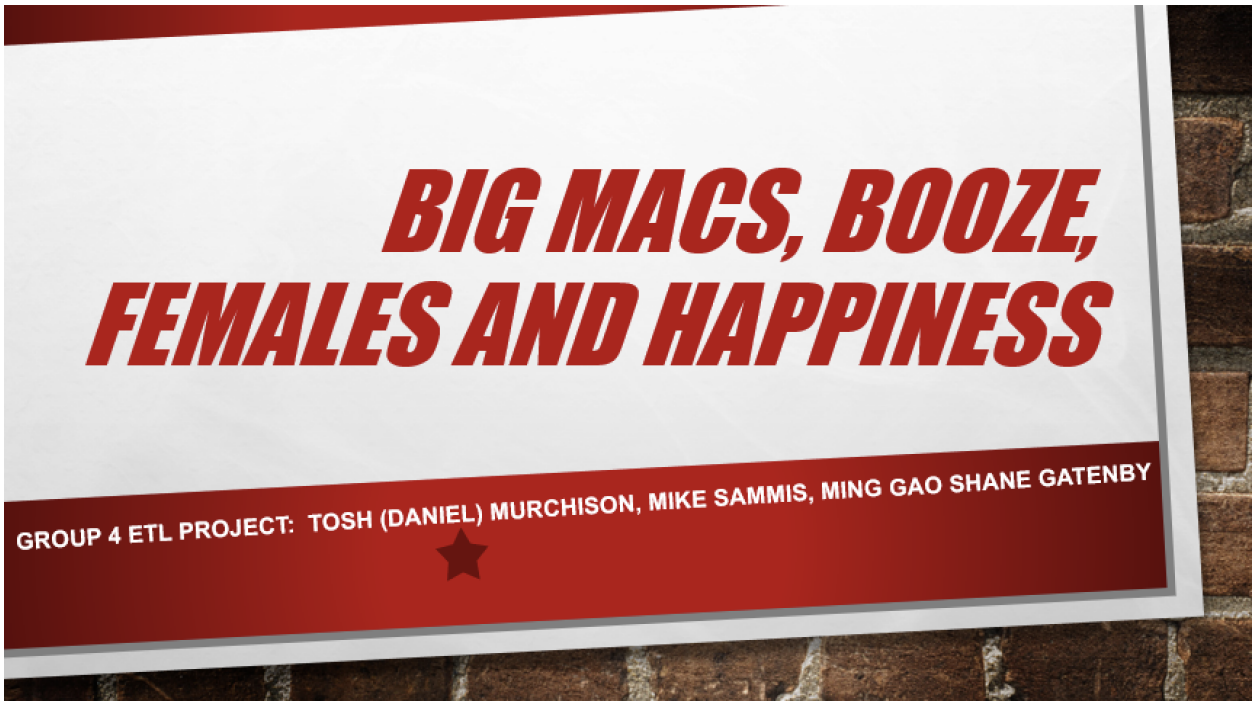
GROUP 4

ETL Project



Introduction

In this project we are utilizing four datasets that give various scores, indices and other demographics, by country across the world for 2016 and then pulling all of those together into one database.

The Datasets (sources of the data and columns used)

- Happiness and Alcohol Consumption (source: kaggle)
    - Description: “Alcohol consumption versus Happiness Score (and another social indexes)“
    - Columns Used:
        - Country
        - Beer consumption per capita
        - Spirit consumption per capita
        - Wine consumption per capita
- The Big Mac Index (source: Kaggle)
    - Description: “The Big Mac Index is published by The Economist as an informal way of measuring the purchasing power parity (PPP) between two currencies and provides a test of the extent to which market exchange rates result in goods costing the same in different countries. It "seeks to make exchange-rate theory a bit more digestible".”
    - Columns Used:
        - Country code (iso_a3)
        - Currency code
        - Local price
        - Exchange rate (local currency to USD)
        - Date
- World Happiness Report (source: Kaggle)
    - Description: “Happiness scored according to economic production, social support, etc.”
    - Columns Used:
        - Country
        - Score
- Gender Statistics (source: The World Bank Data Catalog)
    - Description: “The Gender Statistics database is a comprehensive source for the latest sex-disaggregated data and gender statistics covering demography, education, health, access to economic opportunities, public life and decision-making, and agency.”
    - Columns Used:
        - Country code (iso_a3)
        - Female Population
        - Total Population

Process: Extract, Transform and L(oad

E: Extract

First we downloaded CSVs of each dataset from their source. Those CSVs were then imported into a jupyter notebook using pandas dataframes and transformed in the next step.

T: Transform

Data was cleaned up, removing unneeded columns, filtered to only include 2016 (from datasets that provided data other than 2016) and calculating the female population to total population ratio for the gender data. We also normalized data frames in pandas.

L: Load

Using pandas to_sql() we imported our normalized data frames to postreSQL database. The logical structure of the data, lent itself to a relational database structure and the tables were normalized. Ultimately, our final tables loaded into our database were:

- beer
- bigmac
- country
- exchange_rate
- gender
- happy
- liquor
- wine