**ETL Project**

**Team:** Leah Barbour, Andrew Cummings, Dinesh Nayak

**Project Summary:**

Collect monthly weather and alcohol sales data and load into a relational database for further analysis to find if there is any correlation between the two.

**Data Cleanup & Analysis**

**The sources of data that you will extract from.**

The following data was loaded into two separate CSV files.

* Monthly alcohol sales data in USA for the last 30 years:

<https://fred.stlouisfed.org/series/S4248SM144NCEN>

* Monthly average temperatures in the country:  
  <https://www.ncdc.noaa.gov/cag/national/time-series/110/tavg/1/12/2000-2018?base_prd=true&begbaseyear=1901&endbaseyear=2000>

**The type of transformation needed for this data (cleaning, joining, filtering, aggregating, etc).**

Weather data

The first 4 rows in the CSV file were not actual data, so they were dropped. The data in the last column was not relevant for our analysis, so it was dropped. The remaining columns were renamed to more sensible names.

Sales data

Sales file had only two columns. The columns were renamed to more sensible names.

**The type of final production database to load the data into (relational or non-relational)**

Both data was loaded into two separate tables in a relational database **weather\_db** (Postgres).

**The final tables or collections that will be used in the production database**

Table 1: temperature

CREATE TABLE temperature (  
  id SERIAL PRIMARY KEY,  
  MonthTemp INT,  
  AvgTemp DOUBLE PRECISION

);

Table 2: drinking

CREATE TABLE drinking (  
  id SERIAL PRIMARY KEY,  
  MonthRecorded DATE,  
  MonthlySales INT  
);

Before loading data into the production tables, the data was converted into proper format (Type Date, Int, Double etc.)

Join result:

Once in sql they tables were then joined into a new table  
based on the date (month). This dropped 97 rows from the alcohol sales csv  
(drinking table). The data could be further cleaned, if desired, as  
demonstrated in the final two columns of the jupyter notebook.

CREATE TABLE temps\_alcoholsales AS  
SELECT a.\*, b.monthlysales, b.monthrecorded  
    FROM temperature a, drinking b  
    WHERE MonthTemp = MonthRecorded;

You will be required to submit a final technical report with the above information and steps required to reproduce your ETL process.

## Project Report

At the end of the week, your team will submit a Final Report that describes the following:

\* \*\*E\*\*xtract: your original data sources and how the data was formatted (CSV, JSON, pgAdmin 4, etc).

\* \*\*T\*\*ransform: what data cleaning or transformation was required.

\* \*\*L\*\*oad: the final database, tables/collections, and why this was chosen.

Please upload the report to Github and submit a link to Bootcampspot.