ETL Project – Final Report

Houston Statistics by Year

By David Nassif, Philip Wirth, Michael Dowlin

# Background

The group wanted to explore some of Houston’s oil history, mainly how correlated the oil price was with: population, employment and housing cost.

# Data

We pulled from two main sources for all of the data, the Federal Reserve Bank of St. Louis (FRED) and Macrotrends (website). FRED has lots of valuable datasets for various economic and demographic statistics. We used Macrotrends to get the oil price history table.

# Extraction

We downloaded the FRED datasets as csv’s. These csvs are included in the project, in the raw data folder. The Macrotrends data is scraped using Pandas “read\_html” function.

# Transform

The data transformation include aggregation (to 1 row per year), converting some #’s from thousands to millions (population) and renaming columns/indexes.

# Load

The data is loaded into five tables in a PostreSQL database called “houston\_statistics\_db”. We decided to use a relational database and have “year” be the primary key for each table. The data can then be combined in multiple ways, i.e. “all 5 tables together, or population and oilprice, etc.”

# Flask API

Our Flask program, app.py, can be used to pull a combined dataset (all five tables), or each table separately. The six calls can be accessed from the local host address on any browser. The index.html site lists out the addresses for the calls.

# Potential Uses

The data could be used to look for trends and correlations among any combination of the five datasets. Potential uses:

* Does a rising oil price mean: increased population, increased employment and an increase in house cost?
* Will a big decrease in oil price lower the housing cost? If so, is there a one year lag, or two?
* While we are pretty sure the Houston population has no impact on oil price, you could test that theory too.
* Which changes faster in reaction to large oil price swings: employment, housing cost or population?