ETL Project

Topic Proposal

—

Omari Blockton

Nicholas McCarty

Data Sources

Federal Housing Finance Agency (FHFA) house price index data (source: https://www.fhfa.gov/DataTools/Downloads/Documents/HPI/HPI_PO_state.txt)

Per capita income data from the United States Bureau of Economic Analysis (source: BEA API)

FIPS codes (source: http://lenkiefer.com/img/charts_feb_20_2017/region.txt)

Milestones

1. Obtain and prepare data for analysis
2. Perform any necessary data transformations
3. Create final production database
4. Produce technical report

Hypothetical Use Case

In order to compare property value by region, it is useful to create an index. One such index could be the ratio of property value to per capita income. That ratio can be easily calculated once we are able to get all of our datasets in a production database.

Columns/Keys Used

From the FHFA housing price index dataset, we will use the following columns:

- place_name (state code)
- yr (year)
- period (quarter)
- index_sa (seasonally-adjusted housing index values)

From the BEA Regional Income dataset, we will use the following columns:

- GeoName (state name)

- TimePeriod (year + quarter)
- DataValue (regional income in USD)

From the FIPS Codes dataset, we will use the following columns:

- statecode (state code)
- statename (state name)

The keys we will use to join the datasets will be **state code/abbreviation** and **state name**.

Transform Strategy

- Read both data sets into pandas
- Clean data (i.e., dropping columns, removing duplicates, removing nulls, stripping strings, etc.)
- Combine year and quarter in FHFA housing index dataset, then format it to match how TimePeriod is in BEA Regional Income dataset (e.g., 2018Q1)
- Create the MySQL database and table(s) to receive data.

Rationale for DB Selection

With the availability of common keys (state code and state name, via the FIPS codes dataset), as well as that of clean, well-structured government data, the creation of a NoSQL database is unnecessary. Instead we will create a relational production database using MySQL.