ETL Project – Hadijat, Tyler, Paul

8/14/2019

EXTRACT

Data Sources/How formatted

1. Population data

From IDPH (Illinois Department of Public Health)

<http://www.dph.illinois.gov/data-statistics/vital-statistics/illinois-population-data>

Data obtained from html by scraping above website using beautiful soup and splinter.

2. Employment data

From USDA (US Department of Agriculture)

<https://www.ers.usda.gov/data-products/county-level-data-sets/download-data/>

Data obtained from specific link to data in csv format:<https://www.ers.usda.gov/webdocs/DataFiles/48747/Unemployment.csv?v=9115.7>

3. Poverty data

From USDA (US Department of Agriculture)

<https://www.ers.usda.gov/data-products/county-level-data-sets/download-data/>

Data obtained from specific link to data in csv format:<https://www.ers.usda.gov/webdocs/DataFiles/48747/PovertyEstimates.csv?v=2529>

TRANSFORM

Cleaning/Transformation

- The three data sources were transformed into Pandas data frames. Unneeded data was removed such as non-Illinois counties from the USDA sources, and percentage change columns from the census data. The USDA data were merged to create one data frame.

LOAD

The final database was loaded into postgress for future queries. As an example, we queried the county and count columns from the census table, and the unemployment rate data from the poverty table. We then joined the data to search for unemployment data in cook county, since this is our county of residence.

We observed that Cook County only had about a 14% unemployment which is the median rate in Illinois despite having a population that is 10times the 75th percentile of the state population.

- Data loaded into PG Admin to allow queries on the merged data.

- Results attached: .ipynb script

.sql codes

Exported csv files (exported after scarping via python)