Point of Living Magazine
ETL Brief

Provided by Point of Living Magazine

Rachael Munyua, Lingzi Xiaoli, Susan Pan

Introduction

1.1 Background

Point of Living is a globally recognized magazine based in the United States. The mission of Point of Living (POL) is to inspire readers in making a positive impact in their lives. One of the main events for the magazine is to publish the city rankings in terms of economic health and living conditions.

1.2. Client Request

POL is interested in compiling a database of United States cities, which will include most of the indicators of economic status and living conditions. By querying through the database, POL will be able to calculate the city scores in terms of “best cities to raise a family”, “ best cities to start a career”, “best cities for retirement” to provide valuable information to different customers (e.g. young professionals, real estate investors) who are willing to explore different cities.

Parameters of the database include: 1) Employment rate, 2) Income, 3) Real Estate Price, 4) Rental Price, 5) Crime, 6) Weather, 7) Education, 8) Healthcare, 9) Transportation 10) Index Table*

*Index Table is the formula table which contains the score weights for different ranking, this data will be provided by POL. The ETL team only needs to build the table to pull the data into the database.

1.3 Database Demo

By this Saturday, the ETL team will deliver a demo database for testing purposes, which will include three data inputs retrieved by different approaches (API, web scraping and direct csv loading). We are specifically interested in the following three parameters: **Weather, Real Estate Price and Employment/Unemployment Rate** and hope to obtain a list of top ten cities with best employment rates.

ETL Design

All the datasets will be connected at city/county level. Each of the following datasets include information of the data description, source, format, size and updating frequency.

2.1 Economic Data

2.1.1 Employment/Unemployment Rate

Description: Employment/Unemployment Rate is a strong indicator of economic health, a city/area with a high employment rate or low unemployment rate would be attractive to people.

Source: www.bls.gov/web/metro/ssamatab1.txt

Data Format: html (data retrieved through web scraping)

Updating Frequency: The Employment/Unemployment Rate data is released by the United States Department of Labor by Month, the data should be updated accordingly.

Data size: Per every update about 400 entries

2.1.2 Personal Income/Household Income

Description: Strong indicator of economic health for a city/area, the current available data is the median personal income data.

Source: U.S. Bureau of Economic Analysis

Data Format: downloadable CSV file

Updating Frequency: yearly or quarterly

Data size: Per every update about 400 entries

2.1.3 Real Estate Price

Description: Real Estate Price means housing affordability within the areas .

Source:www.kiplinger.com/article/real-estate/T010-C000-S002-home-price-changes-in-the-100-largest-metro-areas.html

Data Format: html (data retrieved through web scraping)

Updating Frequency: yearly or quarterly

Data size: Per every update about 400 entries

2.1.4 Rent Price

Description: Main component of living cost.

Source: www.apartmentlist.com/rentonomics/rental-price-data/

Data Format: downloadable CSV file

Updating Frequency: quarterly or monthly

Data size: Per every update about 4000 entries

2.2 Living Conditions

2.2.1 Crime

Description: It reported the number of violent crimes in 2019 (Jan-June) and 2018 (Jan-June) in each city with a population over 100,000. The violent crime category includes murder, rape (revised definition), robbery, and aggravated assault. Property crimes include burglary, larceny-theft, and motor vehicle theft; the last one is arson.

Source: FBI (https://ucr.fbi.gov/crime-in-the-u.s/2019/preliminary-report); Table 4

Data format: csv

Updating frequency: yearly

Data size: 96 KB

2.2.2 Weather

Description: Open weather provides historical weather data for 37,000+ cities including min/max temperature, humidity, feels_like.

Source: Open weather (https://openweathermap.org/history#name);

Data format: json (retrieved through API)

Updating frequency: daily

Data size: upon request

2.2.3 Education (top universities)

Description: This dataset includes the university ranking, city and state information for top 168 universities in the United States.

Source: Times Higher Education's World University Rankings 2020. (https://www.timeshighereducation.com/student/best-universities/best-universities-united-states)

Data format: csv

Updating frequency: yearly

Data size: 17KB

2.2.4 Healthcare (Hospitals)

Description: This dataset contains locations of Hospitals and number of beds by county level in 50 states in the United States, Washington D.C., US territories of Puerto Rico, Guam, American Samoa, Northern Mariana Islands, Palau, and Virgin Islands.

Source: Kaggle (https://www.kaggle.com/carlosaguayo/usa-hospitals)

Data format: csv

Updating frequency: every five year

Data size: 3.7 MB

2.2.5 Transportation (Accidents)

Description: This dataset contains a countrywide traffic accident dataset, which covers 49 states in the United States from February 2016 to December 2019. The traffic data captured by a variety of entities, such as the US and state departments of transportation, law enforcement agencies, traffic cameras, and traffic sensors within the road-networks. Currently, there are about 3.0 million accident records in this dataset.

Source: Kaggle (https://www.kaggle.com/sobhanmoosavi/us-accidents)

Date type: csv

Updating frequency: yearly

Data size: 1.2 GB

2.3 Data Relationship

The main data tables will be organized by MSA (Metropolitan Statistical Area). Data that is only available at normal City/County level will require a MSA-City and a MSA-County table to make the linkage.