Extraction, Transformation, and Load Technical Report

<Point of Living>

Rachael Munyua, Lingzi Xiaoli, Susan Pan

|  |  |
| --- | --- |
| **1.** | **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.

The objective of this project is to identify the best cities to live in in the United States.

# **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 for this demo database includes: Employment rate | Income | Rental rates | Weather

* Sources:
* U.S. Bureau of Labor Statistics
  + [www.bls.gov/web/metro/ssamatab1.txt](http://www.bls.gov/web/metro/ssamatab1.txt)
* Bureau of Economic Analysis
  + <https://www.bea.gov/data/income-saving/personal-income-county-metro-and-other-areas>
* Apartment List
  + [www.apartmentlist.com/rentonomics/rental-price-data/](http://www.apartmentlist.com/rentonomics/rental-price-data/)
* NOAA (National Weather For Environmental Information)
  + <https://www.ncdc.noaa.gov/cdo-web/webservices/v2>

# **1.3 Technologies and resource contributions**

In efforts to obtain the necessary data, the team utilized API & web scraping methodologies and Python, Pandas, Excel, ERD, SQLAlchemy technologies.

* Lingzi Xiaoli: Data source research (five non-economic factors)| Data retrieving and clean up (API weather) |Data import and manipulation example| Final report write-up (2.1, 2.2, 2.3)
* Rachael Munyua: Database creation | Repository creation and maintenance | Each Section Discussion/Brainstorming | Proofreading final report | Final report write-up (1.1, 1.2, 1.3, 1.4)
* Susan Pan: Data source research (five economic factors) | Data clean-up (economic tables) | ERD Graph | Data import and manipulation example | Final report write-up (2.4, 2.5)

# **1.4 Definitions, Acronyms and Abbreviations**

APL - Application Programming Interface

CSV - Comma-Separated Values

ETL - Extract, Transform and Load

ERD - Entity-relationship diagram

MSA - Metropolitan Statistical Area

NOAA - National Oceanic and Atmospheric Administration

TMAX - Maximum

TMIN - Minimum

TAVG - Average

SQL - Structured Query Language

|  |  |
| --- | --- |
| **2.** | **ETL DETAILS** |

# **2.1 Data Import/Extract Sources and Method**

The original **unemployment dataset** covered the unemployment rates and ranks for 389 cities in metropolitan statistical cities in the United States. A csv table was retrieved from the U.S. Bureau of Labor Statistics through **web scraping** (<https://www.bls.gov/web/metro/laummtrk.htm#laummtrk.f.p>).

The original **weather dataset** covered the minimum temperatures (TMIN), maximum temperatures (TMAX), and average temperatures (TAVG) for all weather stations in every metropolitan city shared with the unemployment dataset. We used the following parameters: datatypeid (TMAX, TMIN, TAVG), startdate (2019-01-01), enddate (2019-12-31), units (standard), limit (1000), offset (for loop to exhaust).

Json files were retrieved from the National Oceanic and Atmospheric Administration (NOAA) using **API** with email request for token as the permission. The basic URL is <https://www.ncdc.noaa.gov/cdo-web/api/v2/data>.

The original **apartment rent dataset** covered the rental prices for four different bedroom types (studio, 1 bedroom (br), 2 br, 3br, 4br) for every month of year 2014 to 2020 April in 660 U.S. cities. A csv table was downloaded directly from the Apartment list (<https://www.apartmentlist.com/rentonomics/rental-price-data/>).

The original **income** dataset covered per capita personal income in two consecutive years (2018, 2019) for 384 U.S. cities in metropolitan statistical areas. A csv table was downloaded directly from the Bureau of Economic Analysis (BETA), U.S. Department of Commerce. (<https://www.bea.gov/news/archive?field_related_product_target_id=All&created_1=All&title=>)

**2.2 Data Acquisition**

The **unemployment dataset** used in building our database covered the unemployment rates in March of 2020 for 389 cities in the metropolitan Area in the United States. The **income dataset** covered the Per capita personal income in 383 U.S. metropolitan cities in 2018. The **apartment** **rent dataset** covered rental prices for four different bedroom types (studio, 1 bedroom (br), 2 br, 3br, 4br) for every month of year 2019 in 660 U.S. cities. The **weather dataset** covered average temperatures of 348 cities shared with our Unemployment dataset for each month in year 2019.

The **unemployment and apartment rent** datasets are going to be dynamic based on the global economic and political environments, new industry shifts and trends as well as each city’s long-term plan, therefore the frequency of updating them is aimed to be **monthly.** The **personal income** is also going to be relatively static, so the updating frequency will be **annually**. The weather dataset is going to update **monthly**. Therefore, our dataset will be able to provide the monthly update per request, and the overall comprehensive update will be annually.

The data acquisition resources and manipulation tools (including API and web scraping) will be provided. The unemployment/income/apartment rent data can be approached through periodically checking the authoritative websites, while weather data can be accessed through specifying the new start date and end date. For data manipulation/updating, the client needs to have python and excel as the prerequisites if they would like to curate the database by themselves, or we can provide services to update the database per their detailed requirements.

# **2.3 Data Transform**

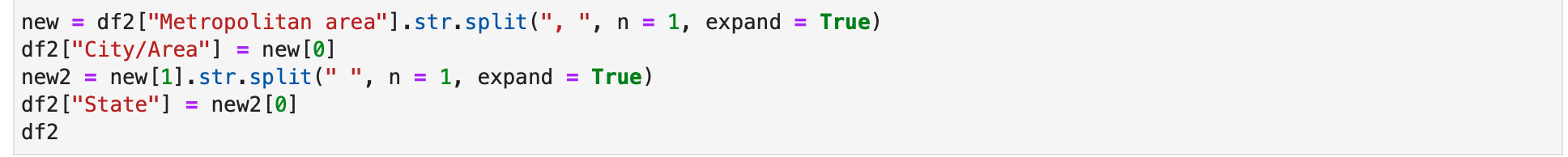
For the unemployment dataset, we dropped N/As and separated the city/area into two columns for city and state. At the same time, we generated a city list named as City\_MSA as a linkage table and the city reference for weather searches.

For the apartment rent dataset, we only kept the most recent info of the year 2019 and 2020 for our demo database, and dropped data from 2014-2018.

For the income dataset, we imported personal income information for year 2018 and 2019 and dropped historical data before 2018.

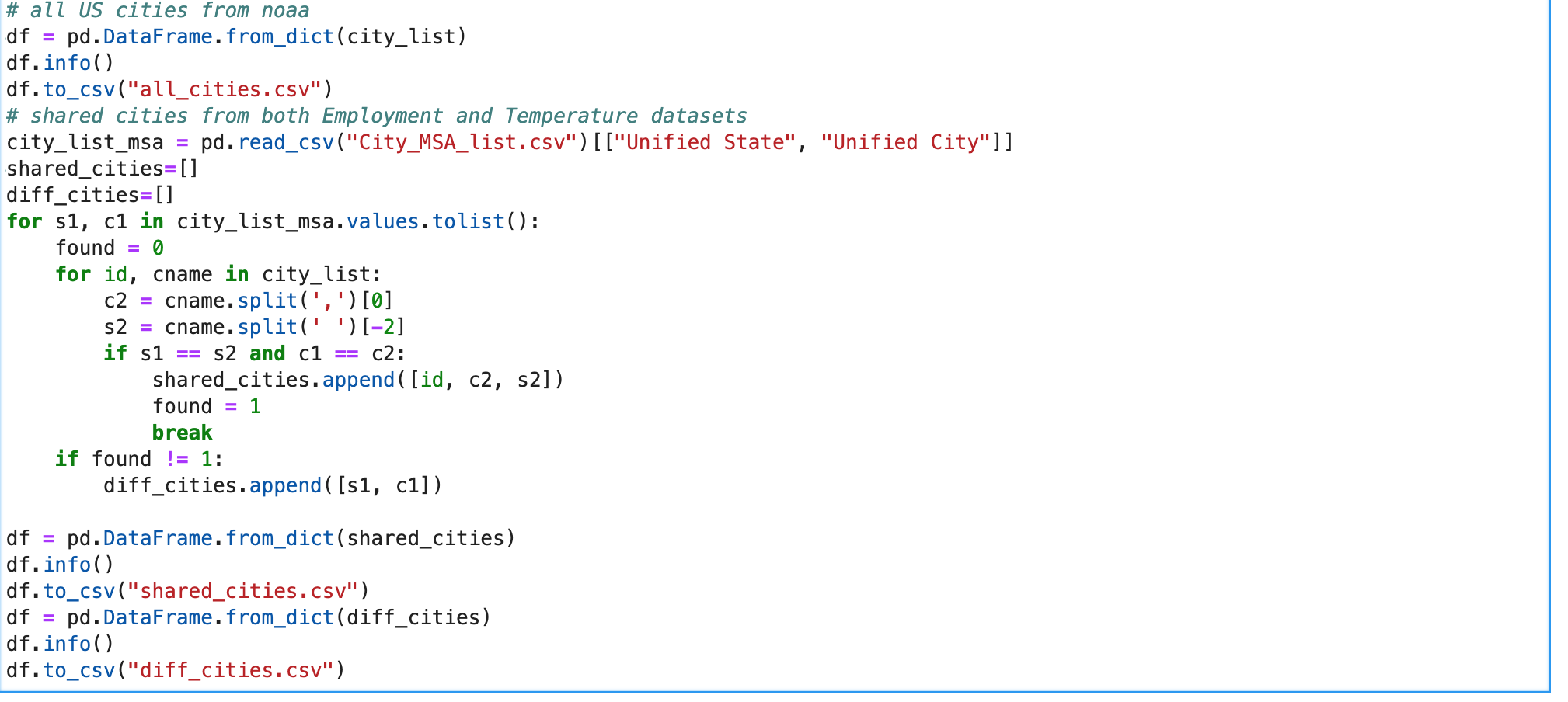
For the weather data, we first found the shared cities between the City\_MSA and city list from NOAA, extracted TAVG for all weather stations in a specific city, then calculated the mean temperature for that month. Further examination revealed four cities (Hilton Head Island, SC; Hammond, LA; Owensboro, KY; Elizabethtown, KY) did not have any temperature record for the year 2019, and six cities (Gadsden, AL; Chico, CA; Ocala, FL; Bloomsburg, PA; Sumter, SC; Lynchburg, VA) had missed a temperature value for one or two months, so we filled in N/A for them. Where uploading all the csv datasets, we changed the N/A as NaN for SQL to recognize.

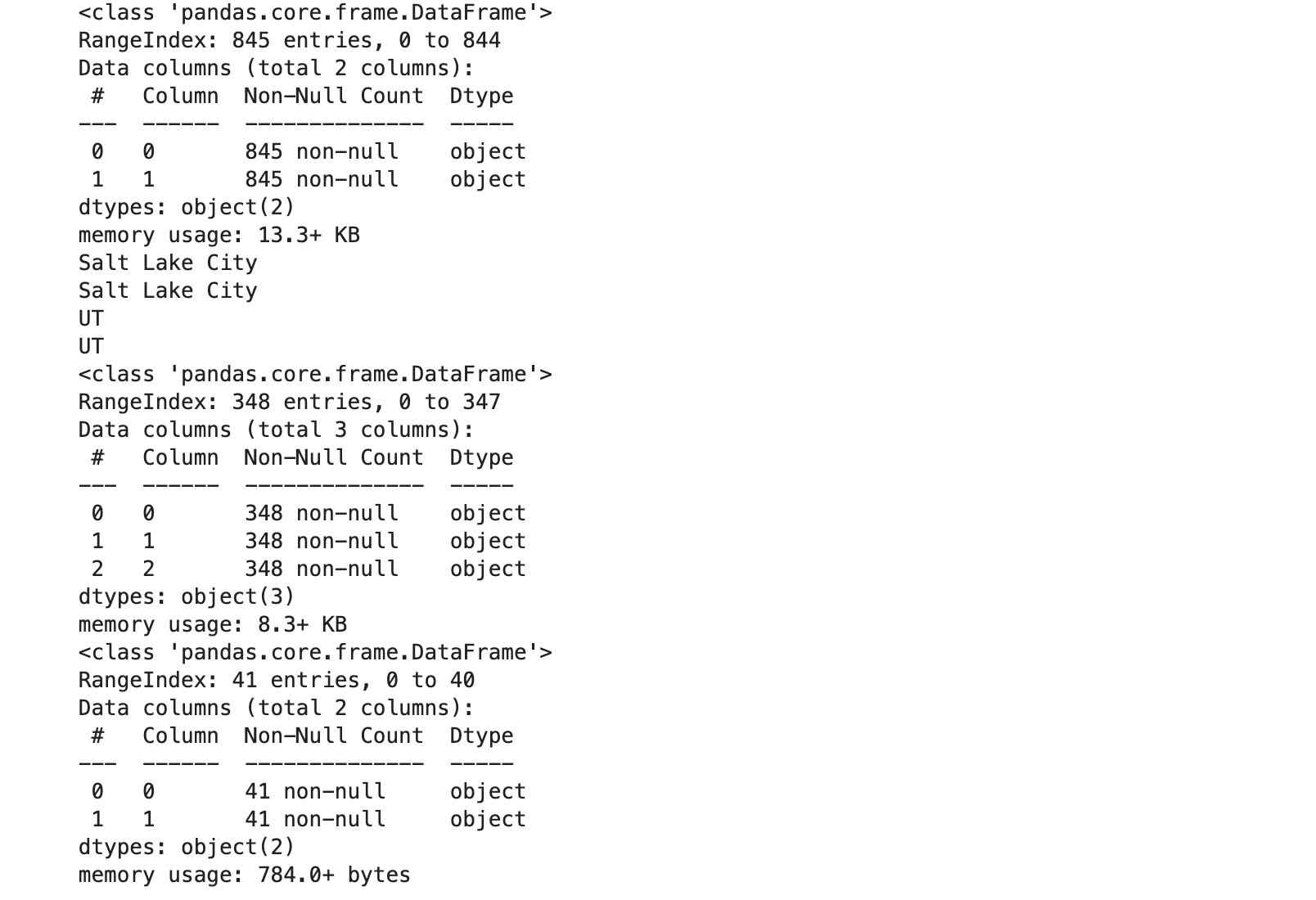
**Example of data clean-up (split “Metropolitan area” into “City/Area” and “State”)**





**Example of data clean-up (find the shared cities in City\_MSA and NOAA city list) :**

****

****

# **2.4 Data Integrity & Update Frequency**

Please refer to Image 2.4.1 for the database’s ERD.

The unemployment\_rate data is from the U.S. Bureau of Labor Statistics. This data is updated by month, and current month data will be released by the end of next month, the data table can be updated accordingly.

The personal\_income data comes from the U.S. Bureau of Economic Analysis, which is released annually, the data table can be updated accordingly.

The rent\_rate table should be updated monthly from the source of [www.apartmentlist.com](http://www.apartmentlist.com).

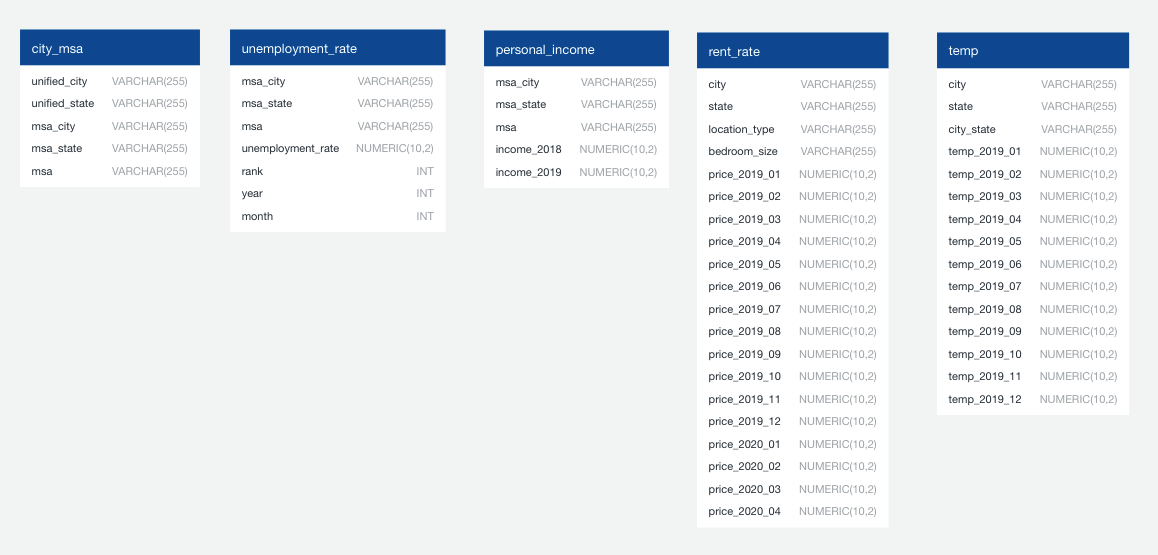
The temp date should be updated by month through API query to NOAA.

Both unemployment\_rate data and personal\_income data are surveyed at MSA\* level, so we created the city\_msa table to make the linkage between two geographic systems.

All data tables can be queried/joined by “city” and “state”. But each data source has a slight difference in the location pool. So we didn’t set up the primary key and foreign key system. But this won’t impact the demo database’s usage.

*\*MSA: ( metropolitan statistical area) a geographical region with a relatively high population density at its core and close economic ties throughout the area.*

**Image 2.4.1**



# **2.5 Data Loading and Availability**

This database is built on Amazon Web Service, the client can access it by logging in their account and querying the data through the platform SQL tools.

The client can export the data as a CSV file for data manipulation, or they can import the data directly to data manipulation tools like Python for further exploration.

For future use, this database should continue importing datasets like university list by city, crime data by city and so on to enrich the content and serve its purpose. The client can eventually query the data and use their self-developed formula to calculate scores for each city in terms of living environment, economic potential.

**Example of data loading into Python:**



**Example of data calculation:**

****