Rodrigo Ivan Rodriguez DSC-540 Winter

Milestone 1

Three Data Sources & Relationships

Instructions:

3 data sources, along with a description of each one (links to each are fine, no need to submit the actual data)

The relationships between them, or the relationship you will make between them

Answer:

API

- **Name**: House canary API
- **URL**: https://api-docs.housecanary.com/#getting-started
- **Description**: House Canary API has detailed real-estate information at the house level. Data includes details about an individual house as well as market-evaluation and sale price.
- **Relationship**: MSA is the Foreign Key (FK) that ties this dataset to the other two dataset, with a many-to-one relationship. Ex: One 'home_address' from the API can only be matched to many 'MSAs' from the Flat File or Web Table.

Flat File

- **Name**: Zillow Research – Sales – Median Sale Price (Smoothed & Seasonally Adjusted, All Homes, Monthly)
- **URL**: https://www.zillow.com/research/data/
- **Description**: This dataset is provided by Zillow Research and is comprised of the median price at which homes across various geographies were sold.
- **Relationship:** RegionName is the FK that ties this dataset to the other two datasets, with a one-to-one relationship. Ex: One RegionName from this dataset can be matched to One MSA from the API or Web Table.

Web Table

- **Name**: Standard Metropolitan statistical areas by PCPI, adjusted by regional price parity
- **URL:** https://en.wikipedia.org/wiki/List_of_U.S._cities_by_adjusted_per_capita_personal_income
- **Description:** This dataset is obtained from Wikipedia and has the top 50 U.S. MSA's by adjusted per capital personal income.
- **Relationship:** Metropolitan Areas is the FK that ties to the other two datasets, with a one-to-one relationship. EX Metropolitan Areas from this dataset can be matched to One MSA from the API and Flat File datasets.

Accomplishing all 5 Milestones

Instructions:

What you believe you will have to do to the data to accomplish all 5 milestones and what your interpretation is of what the data means (you could provide a data dictionary or a summary of what the data is) – should be at least 250 words

Answer:

The data will help accomplish all 5 milestones functionally and contextually. What I mean by contextually is that after the three datasets are cleaned, prepped, and joined, they can be used to conduct an in-depth analysis on the housing market.

The milestones involve finding the three datasets, cleaning and prepping each different type of data source, joining the datasets, and finally creating visualizations from the data. All datasets have multiple columns and relationships between them. Thus, the datasets can all be cleaned and joined together. Additionally, because the datasets all share relationships with each other, creating visualizations will be efficient and insightful.

The flat file has 68 columns with the majority of columns consisting of month columns. This dataset has the median sale price by MSA by month, with each month having its own column.

The API has a large number of variables. There are variables that have details at a lower granularity than the other two datasets. Data has information about individual house as well as market-evaluation and sale price. The data from this API can be rolled up at the MSA level to match with the other two datasets.

The Web data is a table similar to a flat file but hosted online on Wikipedia. The web table has data at the MSA level and holds information on adjusted per capital personal income.

The three datasets can be joined together by the MSA variable and each data source will be able to satisfy the requirements in each of the 5 milestones.

Project Subject

Instructions:

Project Subject Area: Describe your project in 1-2 sentences.

Answer:

The project focuses on cleaning, prepping, joining, and visualizing data. My project will aim to visualize housing market data at the MSA level by various attributes. The intent is to derive insight on what variables affect the housing market in different MSAs.

Relationships

Instructions:

- Describe how the data from each source is connected (see example below).
- If there isn't an obvious relationship, explain how you will make ones.

Answer:

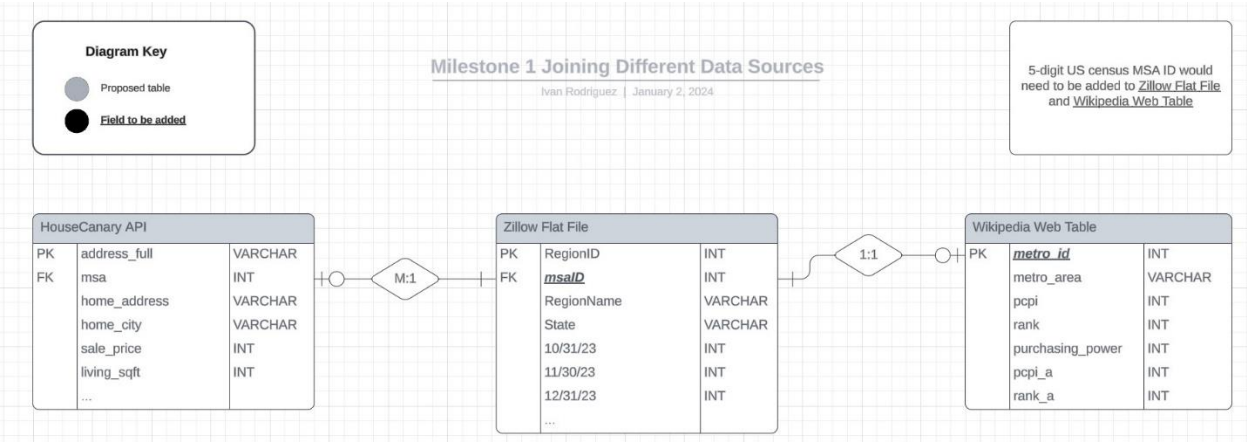
API Relationship: The 'MSA' variable is the Foreign Key (FK) that ties this dataset to the other two dataset, with a many-to-one relationship. This variable is an ID that uniquely identifies the MSA.

Flat File Relationship: 'msaID' is the FK that ties this dataset to the other two datasets, with a one-to-one relationship with the Web Table & one-to-many relationship with the API.

Web Table Relationship: 'metro_id' is the FK that ties to the other two datasets, with a one-to-one relationship with the flat file and one-to-many relationship with the API.

While both the Flat File and Web Table are at the MSA level, both data sources need the inclusion of an 'msa' variable that matches that of the API. This means that, a 5-digit US census MSA ID would need to be added to Zillow Flat File and Wikipedia Web Table

Here is an ERD diagram that shows these relationships:



Description

Instructions:

250 Words describing how you plan to tackle the project, what the data means, the ethical implications of your project scenario/topic, and what challenges you might face.

Answer:

My project revolves around finding three different data sources that have relationships between them. Then cleaning, prepping, joining, and creating visualizations from the final joined data set.

I chose real-estate data from the HouseCanary API at the house level, median sale price data from a Zillow flat file at the MSA level, and personal income data from a Wikipedia Web Table at the MSA level. The focus is to use the variables available in these three data sources to identify any trends or insights in the housing market for some of the top MSAs in the country.

I have reviewed the three data sources and identified that both the Zillow and Wikipedia datasets need the inclusion of a Census MSA ID field to match with the 'msa' field in the API dataset. After doing this I can begin cleaning and prepping each dataset to be joined.

To avoid ethical concerns, I will ensure that any data used for this analysis does not contain any Personal Identifiable Information and is aggregated where possible to protect privacy. Additionally, some potential challenges I could face would be around data integration. I mentioned that I will need to add variables to the flat file and web table that will act as foreign keys. I will be using the census MSA IDs to match these IDs to the MSA primary keys in the respective data sources. I will need to ensure that this step is taken accurately in order to get accurate and consistent results.