Open Ended Capstone – Market Research

Yeh, Margaret

Dataset

American Community Survey 5-Year Data (2009-2022)

- https://www.census.gov/data/developers/data-sets/acs-5year.html

- https://api.census.gov/data/2022/acs/acs5/profile/variables.html

Groups: https://censusreporter.org/topics/table-codes/

- Group DP02 – Citizenship, Education, Household composition

- Group DP03 - Selected Economic Characteristics: County level economic characteristics, income of various households, retirement, social security, health insurance

- Group DP04 – Rent and Housing information

- Group S0804 – Means of Transportation to Work by Selected Characteristics for Workplace Geography

- Group S0701 – Geographic Mobility by Selected Characteristics in the United States

- Group S2303 – Work status in the past 12 months

- Group CP03 - Comparative Economic Characteristics

Target Questions to Answer

- Where would it be good to invest a new retail business?
  - Locations of rapid increase in income, increased people moving in, consider housing costs. Employment statistics
- Where are the areas that can benefit from improved public transport?
  - Locations of high population and lots of people commuting by car.
- Is there anything interesting that can correlate with increasing or decreasing housing prices?
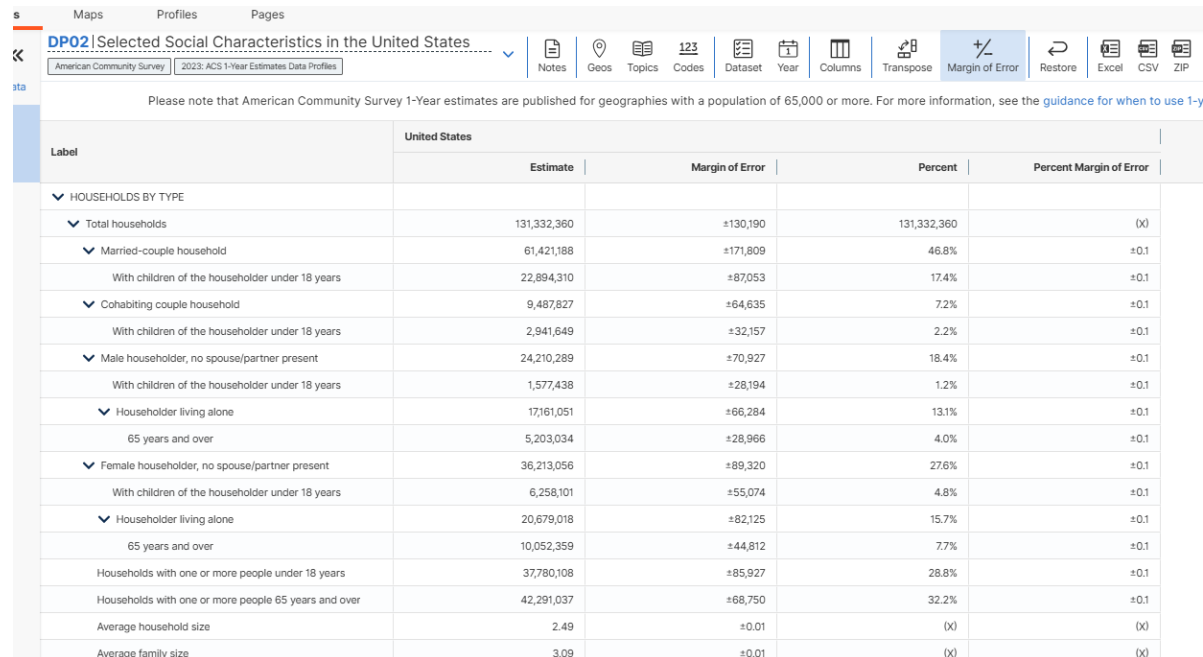
Data Exploration

- For some groups, there are years that the API request failed to get. It seems to happen at random

- Sample image of group DP03 (Rent and Housing information)

	DP03_0001E	DP03_0001EA	DP03_0001M	DP03_0001MA	DP03_0001PE	DP03_0001PEA	DP03_0001PM	DP03_0001PMA	DP03_0002E	DP03_0002EA	...	DP03_(
0	1367247	None	837	None	1367247	None	-888888888	(X)	922490	None	...	-888
1	1084	None	187	None	1084	None	-888888888	(X)	575	None	...	-888
2	34850	None	147	None	34850	None	-888888888	(X)	16238	None	...	-888
3	178693	None	318	None	178693	None	-888888888	(X)	103769	None	...	-888

Data Exploration

- Data.census.gov hosts their data and shows some of the variables in table format on their website. It is useful for initial exploration and checking for interesting datasets to add. For example, for the group DP02: https://data.census.gov/table/ACSDP1Y2023.DP02?q=dp02

Data Exploration

- There are some missing columns, and the headers are coded. Descriptions of the codes are included in the dataset documentation but not in the dataset itself

- 1. Is the data homogenous in each column?
  - Yes, within the column. If there are empty values, then the whole column is empty.

- 2. How do you anticipate this data will be used by data analysts and scientists downstream?
  - Data analysts can draw correlations between the data of different groups. For example, with the population group, people can separate out demographics in each county and see how they correlate with the target they are studying.

- 3. Does your answer to the last question give you an indication of how you can store the data for optimal querying speed and storage file compression?
  - It hints that people might want to be able to access datasets from multiple groups at the same time. Maybe storing with a cloud service and using cloud analytics is the best method of optimizing both the speed and storage.

- 4. What cleaning steps do you need to perform to make your dataset ready for consumption?
  - Removal of empty columns and columns that provide no information. Replace the coded headers or put the description of the coded headers somewhere it is easy to look up.

- 5. What wrangling steps do you need to perform to enrich your dataset with additional information?
  - The datasets can be joined so that not so many files need to processed. Maybe in each group, join all the years into one table.
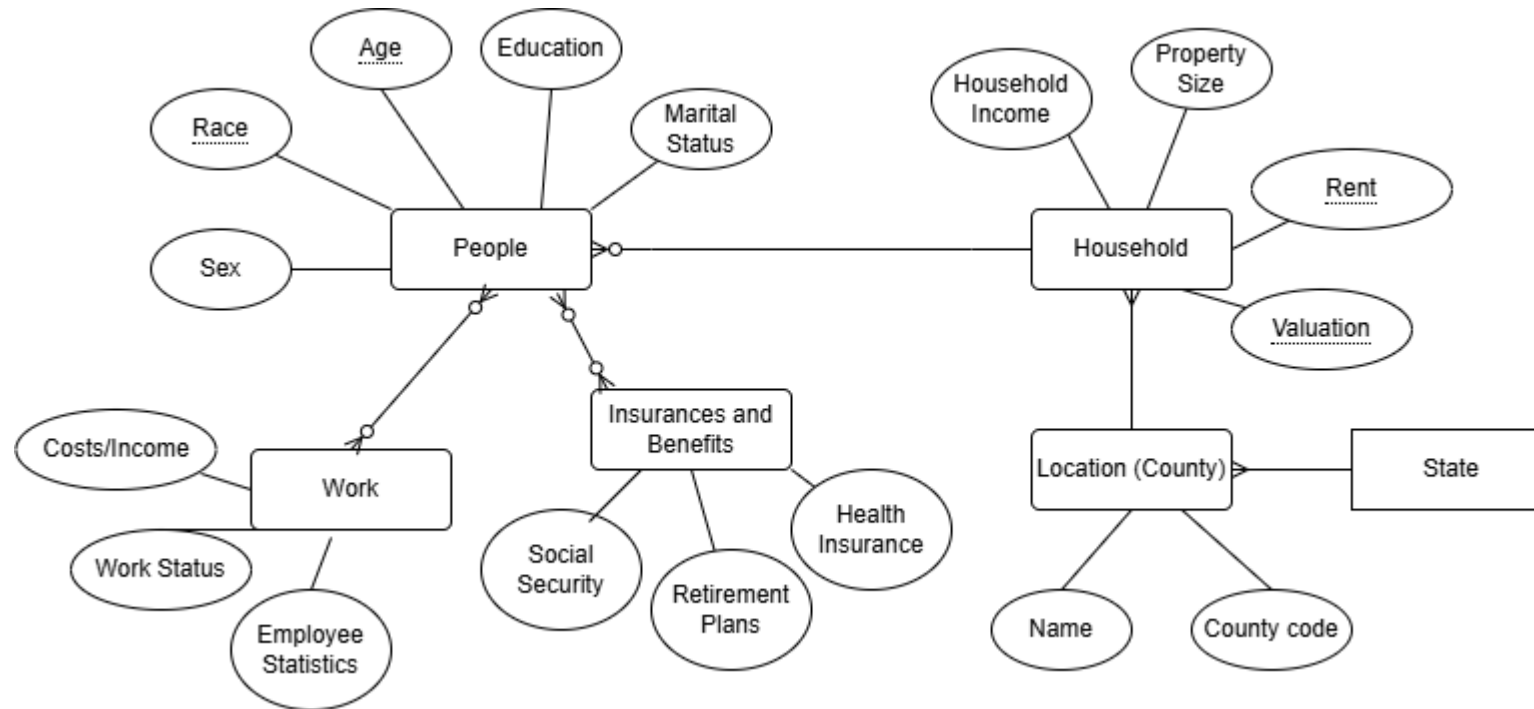
Data Exploration: Entity-Relationship Model

Figure made using Draw.io

Data Exploration:
Goals with transforming the dataset

- Keep most of the information untouched.

- Combine most of the separate datasets together, see if its possible to have all the years in one table (sliding window? Originally the data for each year is a pooling of the previous 5 years). I want to be able to graph the data for a specific county over time.

- Maybe include the json that holds information that corresponds the header code to descriptions, put that in a table for easy lookup