Real Estate and Housing – 2012 and 2022

Project Overview:

Housing is a topic that is frequently in the news here in the U.S. Many people have opinions about this topic without consensus. What are the issues, who is faced with the greatest issues because of those issues, and how do we resolve them to name but a few.

This project will explore static data from the years 2012 and 2022 to determine what has changed, and what has remained the same (or similar).

Questions:

- What is the percentage increase (or decrease) in total housing units from 2012 to 2022.
- What is the difference in percentage of owner-occupied vs. renter-occupied housing units from 2012 to 2022.
- Which geographic markets have seen the greatest increase and decrease from 2012 to 2022.
- Is the perceived increased difficulty in housing affordability based on fact, or is it based on paradigm?
    - "I want to live in a safe area I can afford" vs. "I want to live in New York or San Francisco and can't afford a safe area there".
- Has the population density per unit increased in terms of percentage between 2012 and 2022?
- Are there differences in the total number of rooms per occupied unit between 2012 and 2022?
- Is it possible to determine whether remote work has had an impact on housing affordability and, if so, what is that impact?

Data Source:

- 2022 data.census.gov and 2012 data.census.gov
    - I used the DP04 tables with 5-year estimates.
    - The U.S. Census Bureau is a government agency; the data source is considered trustworthy.
    - The data was obtained through surveys conducted by the U.S. Census Bureau.
    - It is possible that the two datasets will be combined based on a key field to make analysis easier.

Data Limitations:

- This is survey data. Even when collected by the U.S. Census Bureau, the data is subject to nonreporting by some entities.
- It appears that data for Puerto Rico is included, in addition to data for all 50 states and the District of Columbia. This will be addressed as part of data wrangling.

Ethics and PII:

There was no PII included in either the original datasets, or the created subset. The data was made publicly available by the U.S. Census Bureau. There are no apparent ethical or privacy concerns associated with the use of this data.

Data Profile:

- The original 2022 dataset was in CSV format and was comprised of 574 columns and 3,222 rows, with an additional 2 header rows. The original 2012 dataset was in CSV format and was comprised of 573 columns and 3,221 rows, with an additional 2 header rows.
- One header row in each dataset was non-descriptive, while the other held very long and somewhat unclear descriptions.
- There were multiple duplicated columns and columns in each dataset that held the results of calculations performed by the Census Bureau.
- Due to the relatively small size of the dataset, the complexity of the header rows, and the volume of duplicate and unverifiable data fields, the best tool to use for the first pass through data cleaning was Excel.
    - From each dataset I was able to remove 489 columns, eliminate one header row, and make the remaining field labels more descriptive.
- The resulting 2022 dataset was 85 columns and 3,222 rows. The 2012 dataset was 84 columns and 3221 rows. Additional exploration and cleaning will be conducted using Python (one notebook for each dataset until such time as they are joined).
    - The preliminary assumption due to expectations based on the source and collection methodology is that there was nonreporting by some entities. Since this project is designed to compare two points in time, we are only interested in a full join of the two datasets for rows (entities reporting). Any differences will be addressed as part of data wrangling.

Field Review:

The “GEO.ID” field contains 14-digit codes that identify the summary level of data, the geographic component of the data and FIPS codes that uniquely identify the data. Since this

serves as both as unique identifier and a location code, it will be retained. This field is a candidate as a join criterion.

"us_county" is an object field that contains strings to indicate the county and state/territory name for each county. This field is a candidate as a join criterion. I will need to create two new fields from this field: county and state as part of the wrangling phase.

"Total housing units" is int64 and contains the total number of units in the survey for each county. This includes owned, rented, occupied, and vacant units. An additional two fields, "Occupied housing units" and "Vacant housing units", are int64 and contain the counts of each subset of Total housing units.

Nine int64 fields break down physical structures and conveyances (boats and mobile homes) used for housing by the number of units, type of conveyance, and whether single units are attached or detached.

Ten int64 fields break down the "Total housing units" field by the year of construction in the 2022 dataset. There are only nine fields used to break this down in the 2012 dataset. The categories are not identical for each dataset, so some wrangling may be necessary.

Nine int64 fields break down the "Total housing units" field by the total number of rooms in each. This includes bedrooms, bathrooms, kitchens, family rooms, and any other room types contained in each unit.

Six int64 fields break down the "Total housing units" field by number of bedrooms in each unit.

Two int64 fields were used to distinguish the owner-occupied and renter-occupied units.

Six int64 fields break down the "Total housing units" field by move-in year for each unit. The categories are not identical for each dataset, so some wrangling may be necessary.

Four int64 fields were used to denote the number of vehicles available to each unit.

Eight int64 fields were used to relate the heating fuel used for each unit.

Four fields were used to relate the number of units in each county with significant deficits, such as units with no heat, limited bathroom and/or kitchen, and no landline telephone available. Three fields are int64, "No landline available" is object.

Three int64 fields were allocated to describe the population density in terms of occupants per room in each unit.

Eight int64 fields break down the purchase price ranges for owner-occupied units. The categories are not identical for each dataset, so some wrangling may be necessary.

Two int64 fields break out owner-occupied units with and without mortgages.

One int64 field provided a count of rent producing units.

The final eight int64 fields were used to present gross rent ranges for renter occupied units. The categories are not identical for each dataset, so some wrangling may be necessary.

Descriptive Statistics:

Due to the shape of the datasets, all descriptive statistics are included in Jupyter Notebooks and not repeated here.

Additional Notes:

It is possible that additional data may be required to address questions. If so, any such data will be sourced from sources similar to the U.S Census Bureau. Such sources may include data.gov, the Internal Revenue Service, or other agencies of the U.S. Government.