Final Project Progress Report

1. Project scope update

The project scope remains consistent with the original proposal. I have successfully implemented data loading from all three primary data sources. The main enhancement is the implementation of automated data cleaning and merging procedures that handle different data formats and FIPS code standardization across datasets.

2. Data sources

2.1. Data obtained in Python code

Zillow Home Value Index: Successfully loaded county-level housing data with median home values for December 2022. The dataset includes 3,073 counties with RegionName, State, and FIPS codes.

US Census Bureau API: Retrieved comprehensive socio-economic data including:

- Median household income (B19013_001E)
- Population counts (B01003_001E)
- Educational attainment (bachelor's, master's, professional, doctorate degrees)
- Poverty statistics

Bureau of Labor Statistics: Loaded unemployment data with labor force statistics for 3,225 counties, including unemployment rates and labor force size.

2.2.API used

Census Data API: Used the census Python library to access American Community Survey 5-year estimates (2022). The API provides structured JSON responses that are automatically converted to pandas DataFrames. Authentication is handled via API key stored in configuration files.

3. Issues / difficulties

The main technical challenge was standardizing FIPS codes across different datasets. Zillow provides separate StateCodeFIPS and MunicipalCodeFIPS columns, while the Census API has separate 'state' and 'county' fields. I solved this by using .str.zfill method to create uniform 5-digit FIPS codes that work for merging.

Understanding the Census API variable names required some research since they use codes like B19013_001E for median income and B15003_022E for bachelor's degrees. I had to look up what each code represented and calculate percentages from some of the raw count data.

Merging the datasets was tricky because they contained different numbers of counties. I used inner joins to ensure data quality, which meant the final dataset became smaller than the original sources. I had to carefully check that the FIPS codes matched correctly across all datasets during the merge process.