Data_issue_writeup

Joey Herrera

3/20/2021

Data Issue Writeup

Introduction

I came across various obstacles throughout the data cleaning process for the American Community Survey data sample and Zillow Home Value Index data. The purpose of this writeup is to discuss each issue to be addressed before adding a dummy variable for Cultural Districts in Texas regions and the spatial difference in difference analysis. To begin this writeup, I will discuss the purpose of the data cleaning process in the Texas-Metro-Housing-Prices repository. This data cleaning process aims to create a singular dataframe from the ACS data and the ZHVI data. The idea is to use RegionName from the ZHVI data as the key variable to merge the two datasets. The mutated RegionName variable in the IPUMS sample is coded using the corresponding MET2013 variable, which holds identification numbers for each metropolitan area since 2013. Many regions are excluded during this merging process because of different metropolitan regions in the two different datasets, difficulty aggregating potential control covariates, and missing observations for the YEAR variable.

Differing Regions

The ZHVI data's codebook includes values for 68 different metropolitan regions in Texas. However, the IPUMS data sample only includes 33 different metropolitan regions in Texas. Only approximately 25 metropolitan regions overlap between the two datasets. The majority of the other metropolitan regions that the ZHVI data includes that are not included in the merged data frame are relatively small and located near a major metropolitan area. The only exceptions are Del Rio and Eagle Pass, which are cities on the Texas-Mexico border. A potential solution is to leave out these areas or merge the smaller metropolitan regions with the nearest large metropolitan area.

Difficulty Aggregating Potential Control Variables

When grouping the IPUMS data by the YEAR variable, various covariates must be aggregated to ensure the IPUMS data frame rows match the number of rows from the ZHVI data. In order to aggregate other covariates, I must mutate a new variable within the summarise function and aggregate the variable using a transformative technique. The other potential issue is resubmitting a data extract to IPUMS to receive data for additional covariates.

Missing Observations for the YEAR Variable

Within the ZHVI dataset, some Texas metropolitan regions do not have observations for every year from 2006 to 2019. This creates a problem when merging with the IPUMS data from the American Community Survey because it accounts for every year for each region. At the moment, I am excluding these observations but will attempt to implement them soon.