House Valuation in Chicago

Jieyi Chen, Hanzhe Zhou, Jaeho Hahm

3/17/2022

Contents

Introduction

Research Question

What factors make a community livable or valuable in Chicago? Conversely, do some living environments or conditions decrease the value of a community? For example, current gun-related violence cases seemingly affect the value of one of the neighborhoods in the city.

The research question is: whether racial composition, public safety, socio-economic, and public provisions status would affect community values in Chicago. The housing value in a zip code area in one year is shown

by the average of twelve quarters to absorb seasonal trends. We measure public safety by the total number of major crimes such as homicide and gun violence. Socio-economic status is measured by four variables: median income, median age, racial composition, and bachelor or higher rate. The number of grocery stores and bus stops are used to measure the level of public services.

We hypothesize that the housing price: 1) negatively correlates with the Black population rate and number of crimes and 2) positively correlates with White and Asian population rate, bachelor holder rate, public provisions.

Data and Approaches

Approach for Research and Coding

Data Selection

As conducting the research in Chicago, we collected crimes, bus stops, grocery stores data from the "Chicago Data Portal". One of the advantages it gives is that we can obtain zip codes for each event or location. Moreover, for crime data, the dates of each crime occurrence are provided. We took housing price data from Zillow in that only the portal has by zip code data for a longer time period. The reason for choosing 2011 as the starting year is that house valuation should differ before and after the 2008 financial crisis.

Demographic characteristics (median income, median age, racial composition, education attainment) are accessed through American Community Survey (ACS). To get zip code level data, we use ACS 5-year data. For example, values of year = 2011 in our dataset are the values of year 2007-2011 in ACS.

Data Wrangling

1. Consistently, we make data frames with zipcode-year format. One of the tasks for cleaning data is pivoting raw data into the one having zipcode-year as an observation unit. We need to alight all data frames with the same format of zip codes since some data has nine digits zip codes instead of five digits zip codes.
2. We drop specific dates and times except the year considering our basic unit of analysis for crimes.
3. Some areas have no records for groceries and crimes in certain years. This could be interpreted as no grocery or no crime at those points. So we substitute these NA with 0 as merging data. Also, grocery data is only available in 2011, 2013, and 2020. Since the number of grocery stores tends to be constant across years, we use 2011 data to permute data for 2000-2010, 2013 data for 2012, and 2020 data for 2014-2021.
4. we access ACS via 'tidycensus' api under zip code level to get demographics data Race composition, the proportion of bachelor or higher degrees are calculated using population division. We handle the issue of missing values for some zip codes in 2015 by replacing NA with the mean of the previous and later years.

The merged data contains 1,232 observations of zipcode-year from 2000 to 2021 and 12 variables (*columns names*): zip code (*zipcode*); year (*year*); average house price (*housing_price*); total number of crimes (*crime*), grocery stores (*grocery*), bus stops (*bus_stop*); median age (*age*); median income (*income*); the proportion of people having a bachelor's degree (*bachelor_rate*); the proportion of the white (*white_rate*), the black (*black_rate*), and the Asian (*asian_rate*).

Static Plots

We investigate basic information for our study by creating four kinds of static plots:
1. box plots to show the trend of housing price and median income from 2011 to 2019

2. maps of housing price changes in Chicago by zip code 2011 v.s 2019
3. treeplot to show distributions for racial composition in Chicago 2011 v.s 2019
4. a correlation plot between time-variant variables

Interactive Plots

In the "Distribution" tab, we show how the distribution of housing prices relate to that of predictors each year across regions. In the "Time Trend" one, we present the time trends for housing prices and each predictor in each zip code area, trying to check the association in one region but across years.

Regression

We conduct a regression with two models - pooled and fixed-effect models - using 'lm' function with 'as.factor' argument for time trends fixed, and 'Stargazer' package for a result table.

Text Analysis

We use two articles before and after Covid for text analysis, presuming the difference in nuance and expectation toward the housing market. Plus, we deal with all the negation words to avoid wrong sentiments. Based on those articles, we conduct sentiment analysis.

Weakness and Difficulties

We have some difficulties regarding collecting data. First, we can only get access to certain year data of bus stops and grocery stores. Second, we cannot find or measure other potential predictors such as libraries, parks, education, pollution, and high-way. Third, due to the limit of ACS data at the zip code level, we only have data for nine years from 2011 to 2019. If we get more observations, the regression should be more precise. Fourth, we did not scrape more articles to show a clear trend of sentiment among Chicago residents. Primarily, many websites like WSJ prohibit web scraping. Fifth, in interactive plots, we intended to have the function of clicking one region and showing the according time trend. But using the "leaflet" library takes too much time that we cannot realize it.

Results of Analyses

Static Plots

1. We compare the housing prices and median income in 2011 to those of 2019 to show the trend. It turns out that housing prices and median income have increased significantly hand in hand.
2. Using heatmap to compare 2011 and 2019 housing prices in Chicago, we can tell from that the housing prices across all areas have risen. The increase of housing prices is most obvious in downtown and north.
3. Race distribution for Chicago changed, the white rate increase by 1.6%, asian rate increase by 1.5% and the black rate decrease by 2% from 2011 to 2019.
4. We use a correlation plot to visualize the correlation between dependent variables and predictors. It shows that our hypotheses are correct except for age.

Implications from Dynamic Choropleths

In the "Distribution" tab, we can see that the areas with higher average housing prices have fewer crimes, higher income, and higher bachelor rate across years. Interestingly, the zip code areas with lower housing prices tend to have more bus stops. We first treat the number of bus stops as indication for convenience and access to public services. But it turns out that bus stop number is related to the size and purchasing power of private cars in one region.

Results of Regression

The table displays the result of pooled and unit fixed-effect linear models. They show similar results and are aligned with our hypotheses. The bachelor or higher rate, White and Asian rate are significant predictors. Those factors positively relate to housing price. Whereas age and Black rate are negatively associated with house values. Notably, income becomes insignificant in the fixed-effect model, showing that income should not affect housing prices after partialling out region specific effects.

Text Analysis

The 'bing' dictionary supports our expectation that the sentiment would change after the Covid. Whereas the other two dictionaries 'afinn' and 'nrc' do not show significant change. This might come from lack of article sources and limited dictionaries.

Discussion and Future Work

Even if our models and static and interactive plots perform in achieving meaningful results as seen above, we acknowledge that insufficient covariates are the first and foremost weakness in our exercise. Variables representing living qualities such as transportation and farmers' markets are not available.
If we have more time, we can dig more articles and find more sentiment dictionaries to confirm our hypothesis of sentiment.
For future work, researchers might use larger scale data based on metropolitan areas, controlling year and unit affect to get more diversified data to test on more interesting variables.