Elizabeth Bagley & Rebecca Wagner

Data Science Final Proposal

Abstract

Do public perceptions or true crime rates drive the housing market? In this project, we will attempt to determine the relationships between housing security as measured by rent prices and eviction rates, public sentiments on safety, and reported crime rates in D.C. neighborhoods. Using a variety of data sources, data collection processes, and analysis methods, we expect to find that public sentiment on neighborhood safety will positively relate to rental/sale prices and negatively relate to eviction rates. We expect that perceived safety will be a better predictor of housing prices than reported crime rates across neighborhoods.

Data Sources & Collection

We will use three main types of data sources in our analysis of the housing market. First, we will use data from the Eviction Lab on eviction rates and Open Data D.C. on crime rates as indicators of housing security and safety. The Eviction Lab is a dataset of formal evictions collected by Princeton University for all fifty states and D.C. that can be grouped by neighborhood. Evictions may be an early indicator of development, housing insecurity, and are often correlated with higher crime rates[1]. Open Data D.C. is part of the Office of the Chief Technology Officer (OCTO) and promotes the use of citywide government data for innovation[2]. Open Data D.C. has crime statistics by geography for the entire city which will help to inform the "true" crime rates for the analysis.

For time series data on cost of housing, we will turn to Zillow and Airbnb. Zillow is a popular platform for users to buy, sell and rent houses and apartments containing a wealth of information about the housing market. We will use the data provided by Zillow Research to hone in on forecasted market values for houses and rental properties as well as the actual market changes in a given region. This data is accessible as a csv file and will be cleaned as part of the data collection process. Airbnb may also be another indicator of the housing market, either by the concentration of rental properties in a neighborhood or by the change in average price of properties. To access this data, we will employ web scraping using Beautiful Soup and Selenium, guided by previous similar work on public github repositories[3]. With the data from the Eviction Lab, Open Data D.C., and housing information provided by Zillow and Airbnb, we will have a robust understanding of true crime and housing values.

The final data collection piece will be getting public sentiment data on different neighborhoods in D.C. We will rely on Reddit's API to scrape for public opinion on the safety of certain neighborhoods in D.C., which will help us to understand perceived safety in each area. Another potential indicator of public sentiment is Yelp reviews, which may include information about the neighborhood's safety. Yelp also has an API we will access to get this information. Each of these six data sources will provide us with robust

[1] https://housingmatters.urban.org/articles/how-eviction-affects-neighborhoods#:~:text=Evidence%20shows%20eviction%20can%20perpetuate,cohesion)%20that%20connect%20community%20members.
[2] https://opendata.dc.gov/pages/about
[3] https://github.com/x-technology/airbnb-analytics/tree/main

information to complete our time-series analysis on perceived vs. “true” crime in neighborhoods and their effects on the housing market.

Methods

We expect to use several methods for preprocessing and analyzing our data. Given that data from Open Data D.C. and The Eviction Lab are already collected and available, we will likely have minimal work to do cleaning this data for analysis. However, we will need to do significant work to clean our data collected from potential sources such as Zillow, Airbnb, Reddit, Yelp, etc. This will look like typical data wrangling.

With all the cleaned data, we will also likely employ a text analysis model for the purpose of identifying and quantifying public sentiments on safety. We will identify words or phrases associated with sentiments of safety or danger in neighborhoods and look for the frequencies of these words/phrases in our data. This will allow us to robustly estimate the public perception of safety, which can then be compared to true crime rates. We may finally employ supervised learning models such as linear or locally weighted regression to attempt to determine a relationship between public sentiments, reported crime rates, and indicators of housing security. The specification of this model will ultimately depend on the quantity and material of the data we are able to collect.

Outcomes

Overall, this project has three goals: 1) collect and clean sufficient data regarding eviction rates, housing security, crime rates, and public public perceptions of safety in D.C. neighborhoods, 2) devise a theoretically sound quantification of these public perceptions and a strong linear model relating housing security and measurements of safety, and 3) report our findings with accessible visualizations.

Through these steps, we would like to show whether there is a difference between perceived safety in a given neighborhood and “true” safety as reported by crime rates. If so, we would also like to show the effect of each on measures of housing security such as eviction rates and rent prices. We are also interested in demonstrating a potential “lag” between public perceptions and crime rates as a neighborhood becomes safer, where public perception takes longer to adjust to this new reality. We would be surprised to find no relationship between any of these variables, or even an inverse relationship between housing indicators and safety (i.e. less safety raises rent prices). Ultimately, this project is aimed at better understanding social processes that contribute to gentrification and housing insecurity.