Data Science Boot Camp: Project 1

For this project we explored what factors may influence the value of housing in Chicago (home values and rental prices). We had a few ideas about what conventional wisdom tells us effects housing values, below are the questions we asked and what we discovered.

## Do Median Rent and Median Home Value by Zip Code Follow the Same Trend?

We used Zillow data from the Quandl API to review 2018 median home values and median rental prices in Chicago. Unfortunately, for many zip codes in downtown Chicago, there was no home value information available. We did find that 2018 rental prices were available for all zip codes and that zip codes with higher median home values, also had higher median rental prices. Median home values and median rental prices have a correlation coefficient of 0.95.

A screenshot of a cell phone

Description automatically generated

Given the correlation and the fact that we did not have the available median home value data for every zip code, we decided to use median rental prices for the rest of our analysis.

Do Higher School Ratings and a Better Walkability Score in a Neighborhood Result in Higher Rent Prices?

Schools in Chicago often have a rating associated with them, which indicates if a school has desirable outcomes for its students. We were not able to find an API in time for this project, and some bulk data about schools was blocked by paywalls. Because of this, we did not include school ratings in our analysis.

Walkability is a company that determines the walkability of given neighborhood. However, the data was organized by exact address, not latitude and longitude nor zip code. Given the time constraints of the project we decided not to look into converting and using the data.

Does Having More Amenities Increase the Rental Price in the Area?

Since most of the data we found was organized by zip code, we used the Yelp API to locate the number of business in certain categories. As a group, we selected three categories that we felt would be most likely to affect rent prices: nightlife, restaurants, and grocery stores. The Yelp API only includes business that have at least 1 review, and we normalized the data across zip code by looking at the number of businesses per 1000 people.

A close up of a map

Description automatically generatedA close up of text on a white background

Description automatically generated

A close up of text on a white background

Description automatically generated

None of the three business types had a strong relationship with median rental price. Nightlife, Restaurants, and Grocery stores had a correlation coefficient of 0.20, 0.11, and 0.26 respectively.

Do Higher Crime Rates and Higher Poverty Rates Correlate to Lower Rent Prices?

Crime Rate

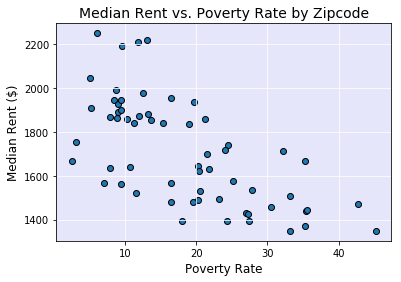
The city of Chicago has a csv file of all reported incidents of crime in 2018. Using the US Zip Code library, we were able to convert the latitude and longitude of each incident of crime, to its corresponding zip code. We normalized the incident of crime per 1,000 people.



There was not a strong correlation between crime per 1,000 people in each zip code and the median rent in each zip code. The correlation coefficient is -0.04.

Poverty Rate

To obtain the poverty rate, we used the 2017 US Census API.



Poverty rate per zip code is inversely related to median rent. The correlation coefficient is -0.56.

Is There a Correlation Between the Median Rent Price and Median Household Income?

Using the median rent price from *Quandl* and the median household income from the 2017 US Census API, we investigated whether median rent is affected by the median household income.

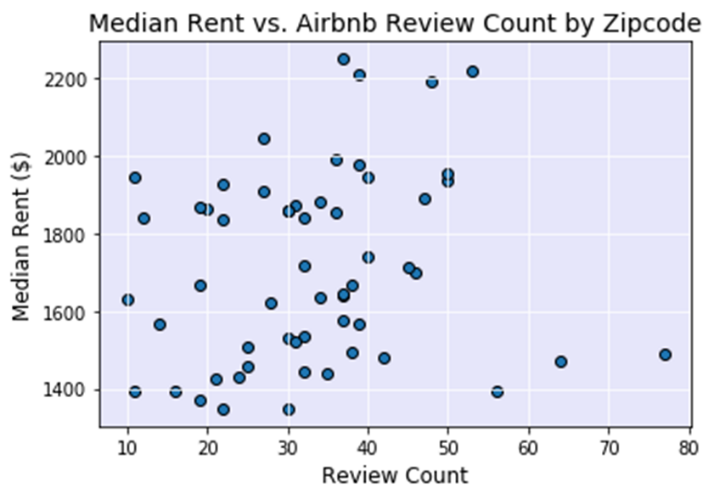
A screenshot of a cell phone

Description automatically generated

Median Household income is positively correlated with median rent. The correlation coefficient is 0.66.

Do Short Term Rentals Such as Airbnb Have an Effect on Rent Prices?

We pulled Airbnb data for Chicago from the Inside Airbnb website. The csv file included the ratings for each listing in Chicago neighborhoods. Latitude and longitude were included for each listing which we converted into zip code. We then calculated the average number of reviews for each zip code.

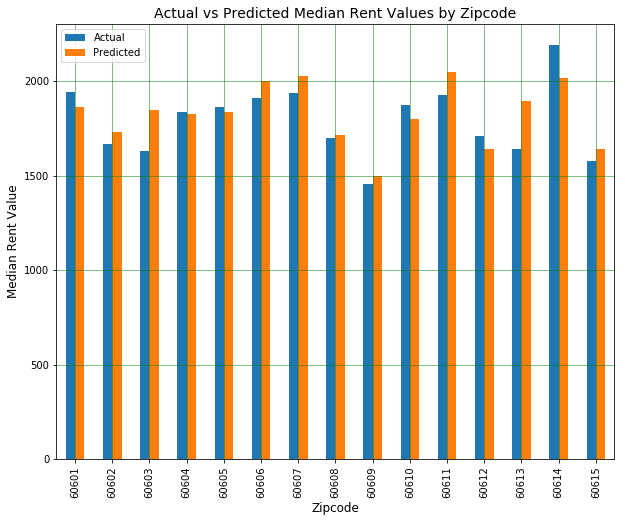


The average number of reviews per Airbnb listing is not strongly correlated with median rent. The correlation coefficient is 0.14.

Multi-linear Regression

In order to verify how well the above variables predicted the rent prices in Chicago, we ran multi-linear regression.

We ran multiple iterations of the regression model. The final model included the following variables: median household income, poverty rate, total crime count, average Airbnb reviews, total amenities (grocery, nightlife, restaurants).



The final R squared value was 0.66 and all variables that we included had a p-value under 0.05.

A screenshot of a cell phone

Description automatically generated

During the iteration process, we discovered a few insights:

1. As previously shown, the per 1000 people metrics were not highly correlated to rent prices. Our initial assumption was to normalize the total counts by population to better compare across zip codes. This proved to be ineffective in predicting rent prices. The relationship between population and the absolute number of crimes or amenities actually has much more explanatory power in the model than the absolute numbers alone. This can be seen in both the below correlation matrix and the p-values.

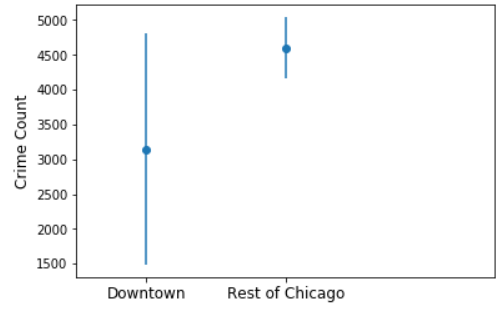
A screenshot of a cell phone

Description automatically generated

1. We found that the total nightlife and restaurant counts were highly correlated (coefficient = 0.87) which violates one of the core assumptions in multi-linear regression. We decided to combine these counts together in order to get rid of the correlation. After doing so, we still found that some of the variables had p-values higher than 0.05. We did one final test combining all amenity variable together into a total amenity count and this resulted in p-values that were all below 0.05.
2. Our multi-linear regression analysis yielded two groups of input variables:
   1. The first group measures affordability or income requirements to rent in each zip code. (median household income and poverty rate)
   2. The second group measures popularity of each zip code. (amenity count, Airbnb, and to some extent crime) This could also be thought of as measuring demand for each zip code.
   3. We hypothesize there is a third group that measures the supply of available units to rent in each zip code. This is essentially a basic model where supply and demand determine the rent price.

T-Test

We noticed a pattern in the crime data where downtown zip codes seem to have higher total crime count than the rest of Chicago. We used the t-test to determine whether these zip codes were statistically different.

****

We found the p-value was 0.45 and concluded that these two groups were not statistically different.

Summary

For this project we reviewed whether certain neighborhood characteristics effect the rent prices for a given zip code in Chicago. The number of amenities and crime per 1000 people and average Airbnb reviews were not strongly correlated with median rent price. However, there is an inverse correlation of median rent price with the poverty rate and total crime count. There is also a positive correlation of median rent price with total amenities and median household income. When these variables are combined in a multi-linear regression model, we are able to account for 66% of the variance in median rent price.

Implications

Chicagoans do care about amenities and crime when looking for housing but median rent price is also driven by other factors. More testing needs to be done to identify them.

Lessons Learned

* Although home values and rent prices were aligned, we may want to analyze how these variables relate to home values separately.
* It may also be helpful to perform a time series analysis to capture how these variables change over time.
* Zip code may be too large of a geographical area to assess indicators of rent. It may be better to look at census blocks or community areas.
* There may be better variables to measure amenities. We only included quantity measures but no qualitative measures such as number of reviews, ratings, etc.
* Our scope was defined on easily accessible data. If given more time, we would have tested public transportation access, school ratings, walkability score, and supply variables such as building permits, number of available units, etc.