1.Problem identification

In larger cities old and neglected neighborhoods can quickly transform into a popular and trendy neighborhood with high real estate demand and sharp increase in the real estate prices. If detected early, these areas can be a good real estate investment opportunity. Some signs that can show a neighborhood is transforming are increase in the number of home renovations and new constructions, increase in the number of restaurants and bars in the area, and decrease in the crime rate. These factors all can cause an increase in housing prices.

The goal of this project is to find the next trendy zip codes by predicting the amount of increase in housing prices three years in advance using public data available from city of chicago data portal such as building permits, valid retail food licenses, and crime rate as well as the historical housing prices in each zipcode from Zillow. At the end of the modeling process, the top 10 zip codes that are identified by the model as zip codes with maximum value increase in the next three years will be recommended to feature home buyers or real estate investors to help them to make a better decision.

2.Data Collection

1.Historical Housing Price, 1996 to 2020

Information on housing data is obtained from Zillow in csv format. The data includes historical monthly housing price data from 1996 to 2020 for all zip codes located in the United States. The

data is seasonally adjusted. I only kept the data related to the city of Chicago and dropped the unrelated columns. The data includes information on 54 Chicago zip codes.

2.Construction and Renovation Permits Data

A building permit is required before beginning most construction, demolition, and repair work in Chicago. Permits issued by the Department of Buildings in the city of Chicago from 2006 to the present and the type of the permit is available from the City of Chicago Website. Here, I retrieve the data using Socrata Open Data API (SODA). I only used the data with permit type that showed more significant alteration such as "New Construction and Renovation" which includes new projects or rehabilitations of existing buildings, "Wrecking/Demolition" which includes private demolition of buildings and other structures, and "Renovation/Alteration". I did not include minor repairs such as electric wiring because that probably is not a sign of change in a neighborhood. The data also includes the latitude and longitude of the construction site for each issued permit which I used to find the zip code.

3.Issued Restaurants license Data

Records of business licenses issued by the Department of Business Affairs and Consumer Protection in the city of Chicago from 2002 to the present are available at Chicago Data Portal. I filtered and retrieved the data using Socrata Open Data API to only include valid business licenses with license code 1006 (Retail Food Establishment). A Retail Food Establishment License is required any time perishable food is prepared, served or sold to the public and includes restaurants, cafés, taverns, grocery stores, convenience stores and more. Increase in the number of restaurants and cafés in a neighborhood can be a good sign and showing that the neighborhood is becoming more popular.

4.Crime data

The crime data was obtained from the City of Chicago Data Portal. This dataset reflects reported incidents of crime (with the exception of murders where data exists for each victim) that occurred in the City of Chicago from 2001 to present, minus the most recent seven days. Data is extracted from the Chicago Police Department's CLEAR (Citizen Law Enforcement Analysis and Reporting) system. The data includes location of the crime (latitude and longitude) and crime type.

5.Zip code related data

Some Zip code information was scraped from CYBO, such as all zip codes' population change from 1975 to 2015, and population change from 2000 to 2015, and neighborhoods close to each zip code.

3.Data Cleaning

Problem 1.
There were 6 zip codes that had missing values over some time intervals. To impute the missing values, I looked at the neighboring zip codes and found the one or combination of the ones that best matched the housing price of the zipcode with missing values. Although imputing the missing values using the average price value of all the neighboring zip codes might have been the easier approach but because some neighboring zip codes have drastically different real estate prices that approach was not the most accurate.

Problem 2.
The latitude and longitude information of each construction permit was necessary to find the zip code where the construction happened. There were 476 entries with missing latitude and longitude information. Because there was no other way to find the zip code, I removed the entries with missing location from this data set.
I also had to get the zip code data from the latitude longitude information. To do this I used an open source package from github which allowed me to run a local server which responded with a zip code to requests containing latitude and longitude.

Problem 3.

The Retail Food Establishment license data did not contain the number of valid businesses for each year and zip code, instead it showed when a license for a retail food establishment was issued and when it expired. Looking at the license start date and expiration date columns of the Issued licenses, we can see that issued business licenses are valid for different periods of time. Businesses need to renew their license before the license expiration date. The id column of the data includes two parts. The first part is the unique id, identifying a unique business and the second part is related to the issued license for a given year. I separated the unique business id and saved it as reataurant_id column. For each unique business I found the earliest year in the license start date column and the latest year in the expiration date column to find the period of time for which the business holds a valid license. Then for each year from 1996 to 2024, I added a column to the data frame and for each business I added value 1 for the years at which the business was open and 0 for all the other years. Then I found the number of businesses in each year in each zip code.

Problem 4.

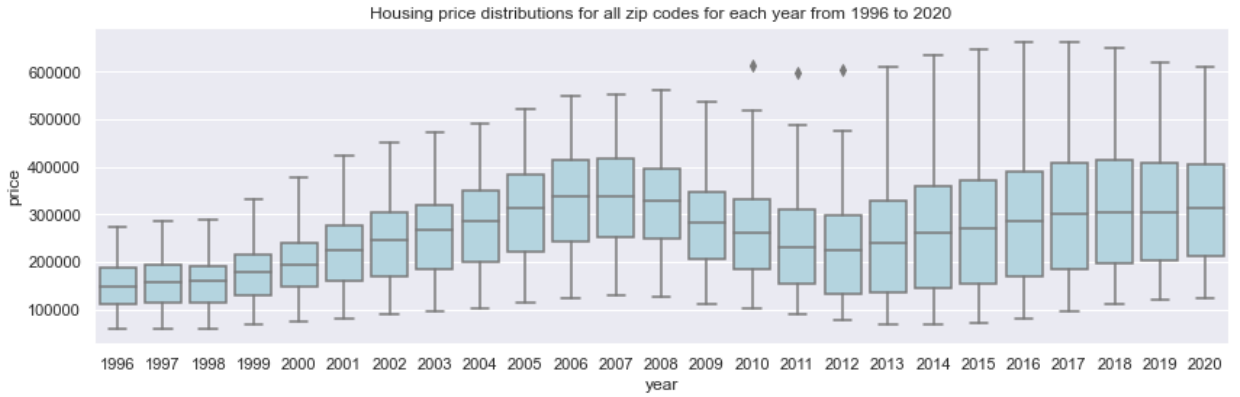
There are 7,301,707 crime entries in the crime dataset of which 71,938 of them had missing location information. Because there was no other way of finding the zip code information for data entries with missing latitude and longitude I dropped those columns.

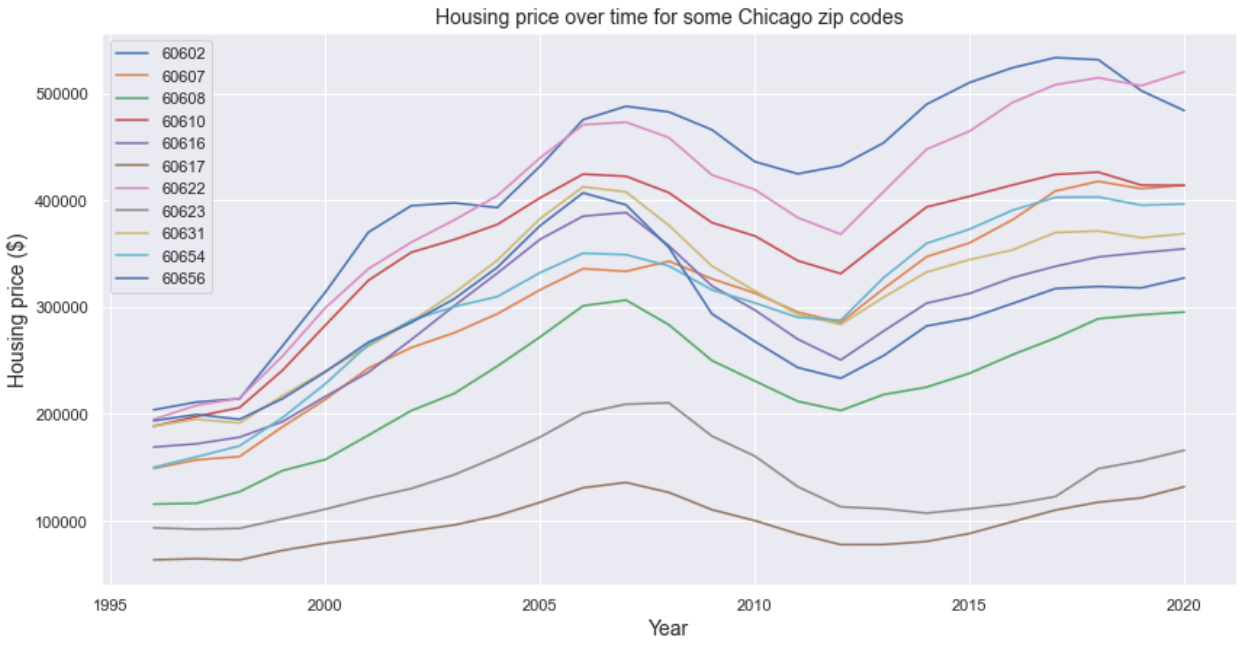
The primary_type column describes the crime type. I grouped the crimes into categories with similar crime types to reduce the total number of crime features in the data. To one hot encode the crime types in the primary_type columns, I first created a dictionary to group all the crime types that are in the same category together and then turned that into a dataframe using DictVectorizer. I then merged that dataframe with the original crime dataframe.

4. EDA

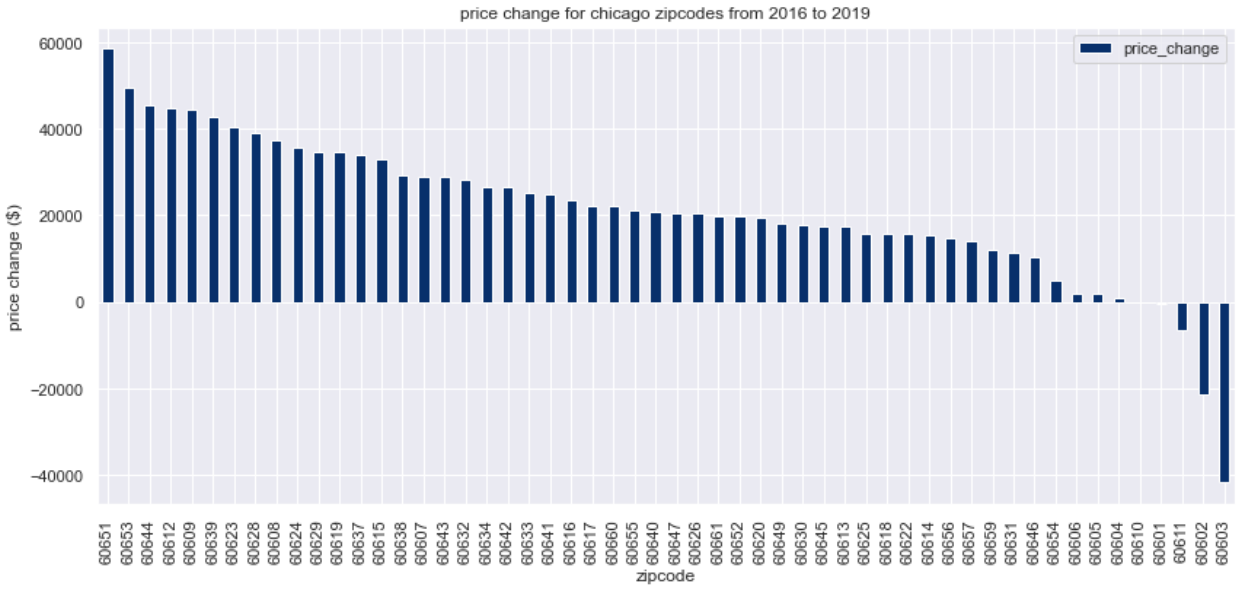
Housing price

Looking at the housing price distribution over time for all the zipcodes in Chicago we can see that the price difference between the most expensive and the cheapest housings increases over time. Showing that although all zip codes saw an increase in housing price since 1996, in some zip codes the housing value increased much more compared to the other zip codes. We also can see the housing market crash starting 2008 and the median housing price has still not recovered to its value before the crash



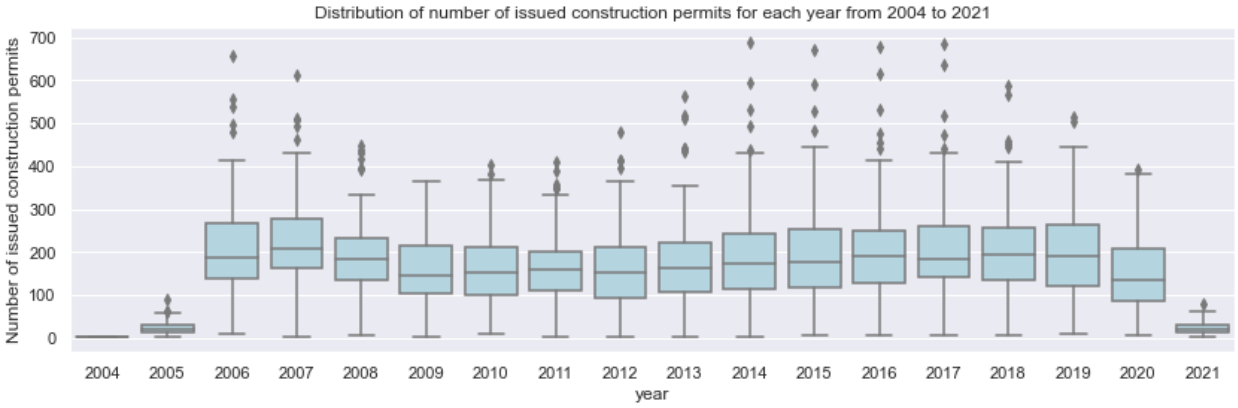


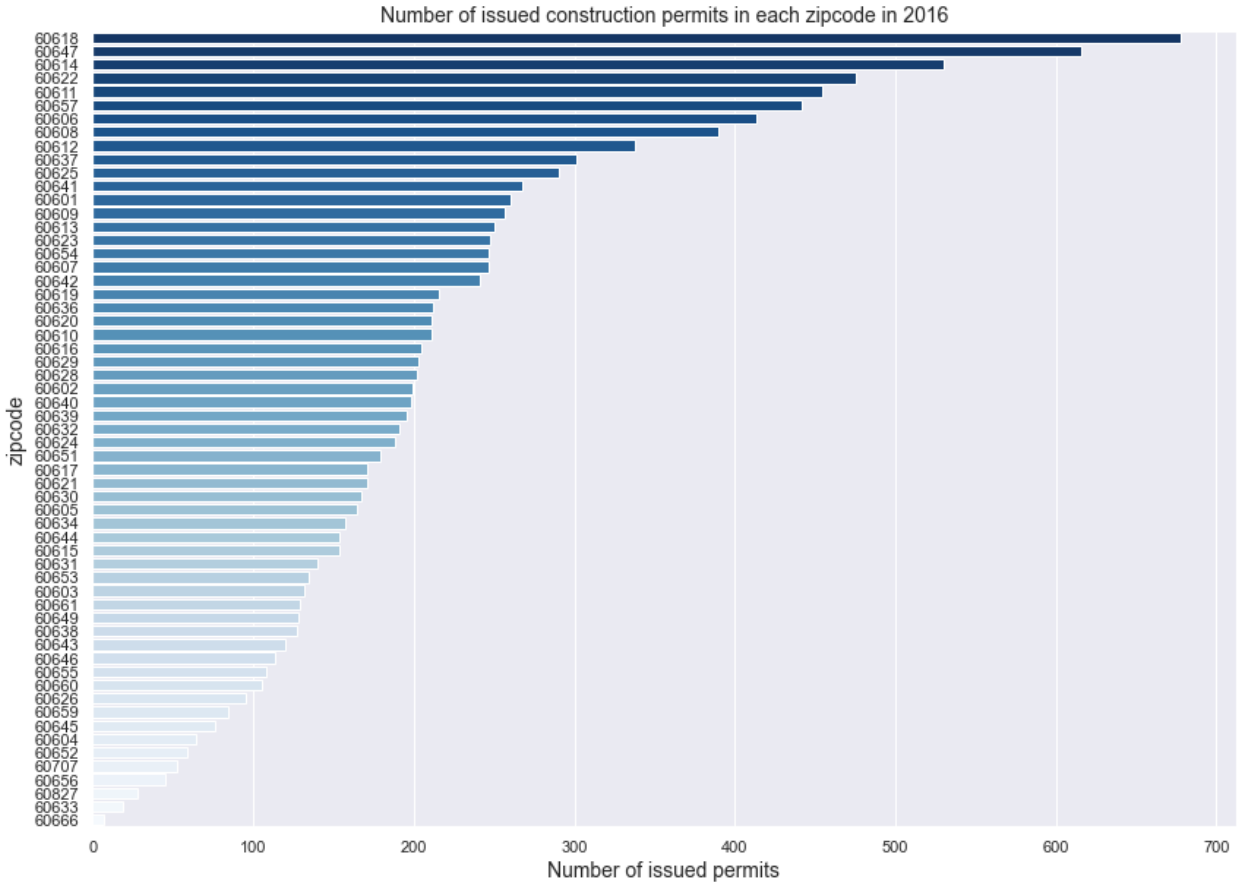
The plot below shows the change in housing price for all Chicago zip codes from 2016 to 2019. We can see from this graph that zip codes 60651, 60653, 60644, and 60612 have the maximum increase in the value over this time period however some zip codes such as 60602 and 60603 saw a decrease in housing prices. The goal of this project is to build a model that can find the top 10 zip codes shown above.



Building permits

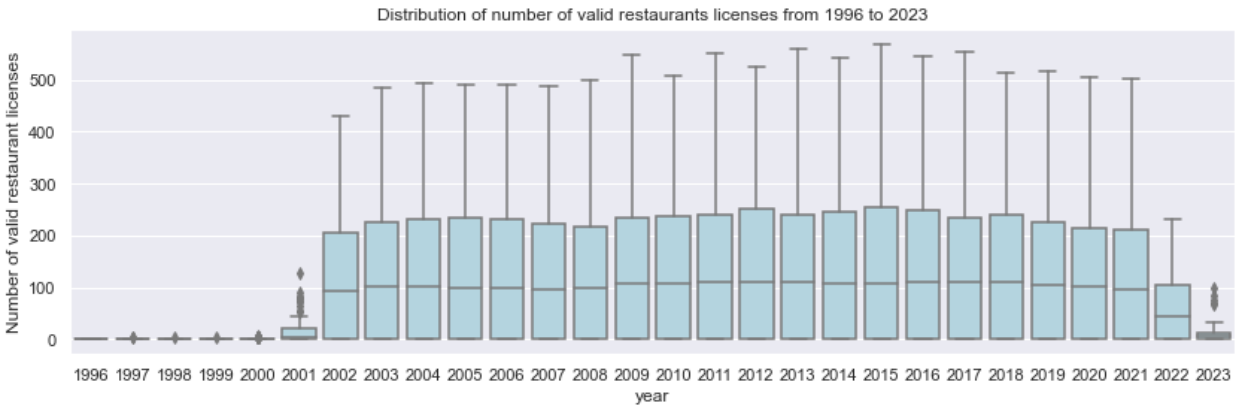
I plotted the distribution of the number of issued permits for all zip codes from 2004 to 2021. As can be seen, the data before 2006 is not complete and I only used the data after 2006 for modeling. In each year there are few zip codes that are outliers and have much more construction counts compared to other zip codes. We can see that the median construction count dropped during the housing market crises but slowly increased afterwards. To have a better understanding of how much construction count in a zip code changed over time, I also plotted the change in the number of permits issued for a given time period for all the zipcodes.

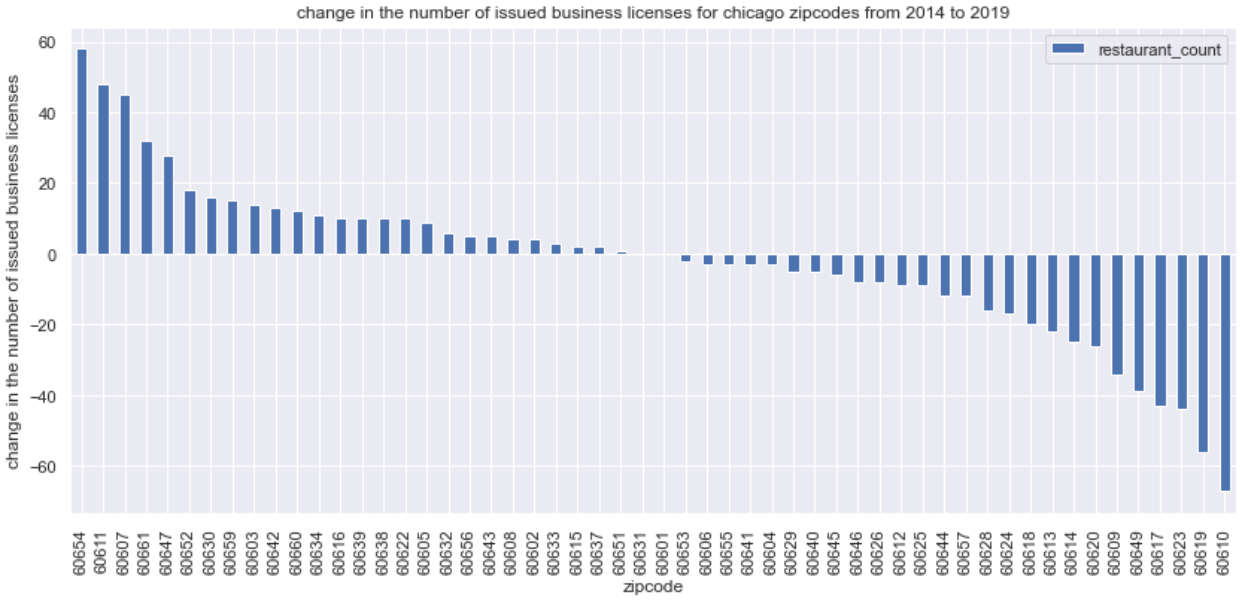




Business licenses

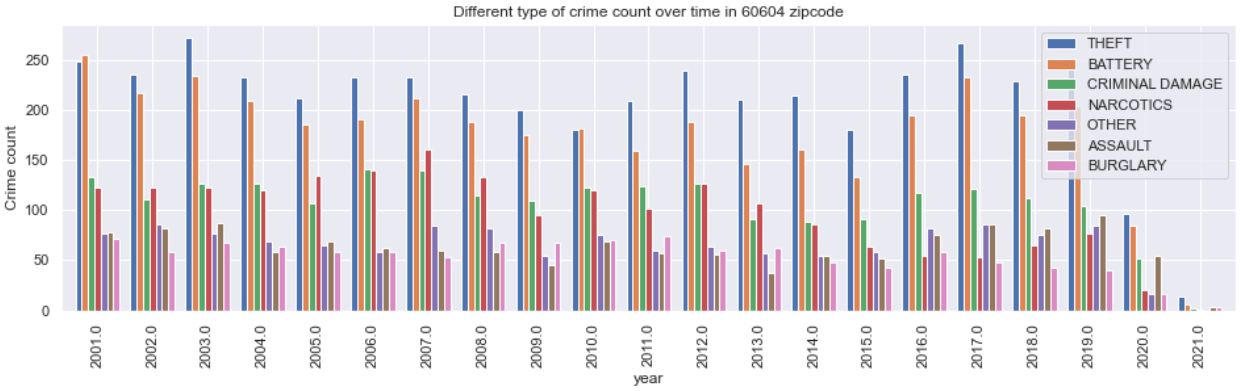
We can see the yearly distribution of valid business licenses is almost uniform over time. The data for years 2021 and later show the number of licenses that will be valid until then but does not represent the total number of valid licenses and hence it's not useful in this project. I plotted the change in the number of valid businesses for a given time period. In this plot we can clearly see that the number of businesses in some zip codes such as 60654 60611 60607 increased from 2014 to 2019 however, some zip codes such as 60610, 60619, and 60623 lost some businesses.

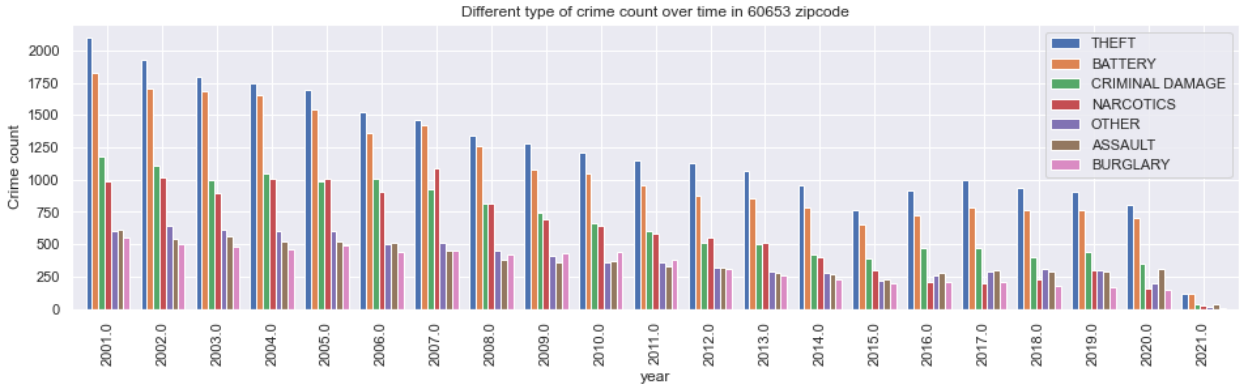


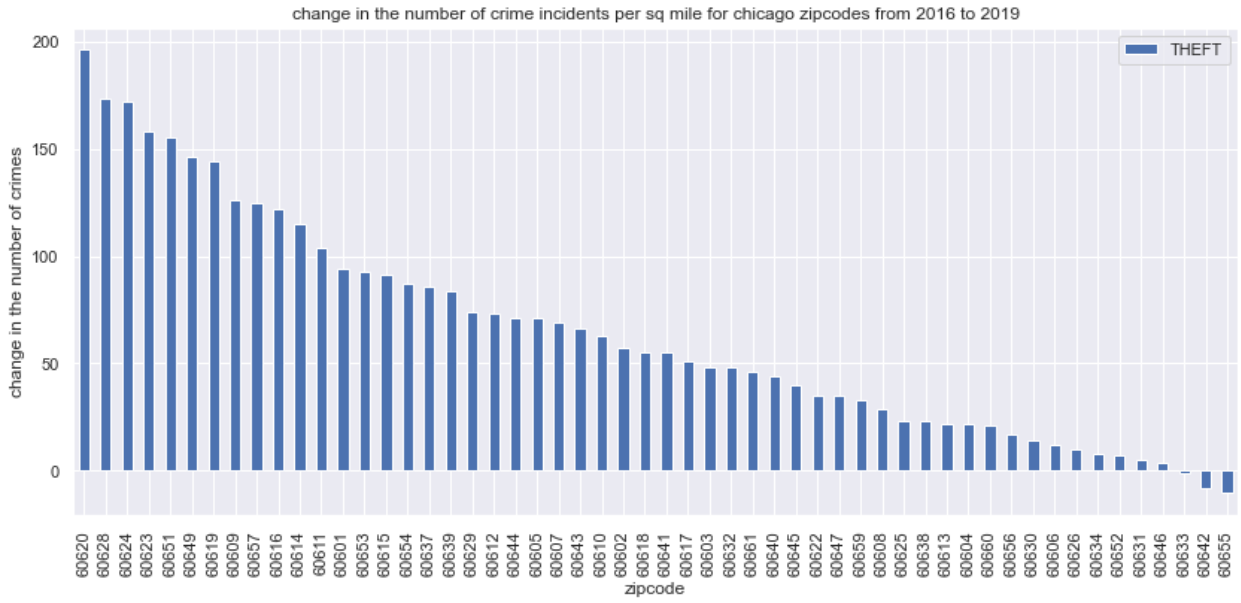


Crime

I plotted the theft count distribution over time and we can see clearly that overall the number of theft incidents has decreased over time. To see how different crime types have changed over time for a specific zip code, I plotted bar plots of 7 different crime types for zip codes 60603, 60604, 60653, and 60612. Zip code 60603 has seen a decrease in housing value over the last 5 years and we can see from the plot below that crime rate has increased in this zip code over this time period. Housing price value increased only slightly over the last 5 years and we can see the crime rate for this zip code has increased as well. Zip codes 60653 and 60612 were among the zip codes that saw the maximum increase in housing value over this time period and we can see that the crime rate has decreased for these zip codes.



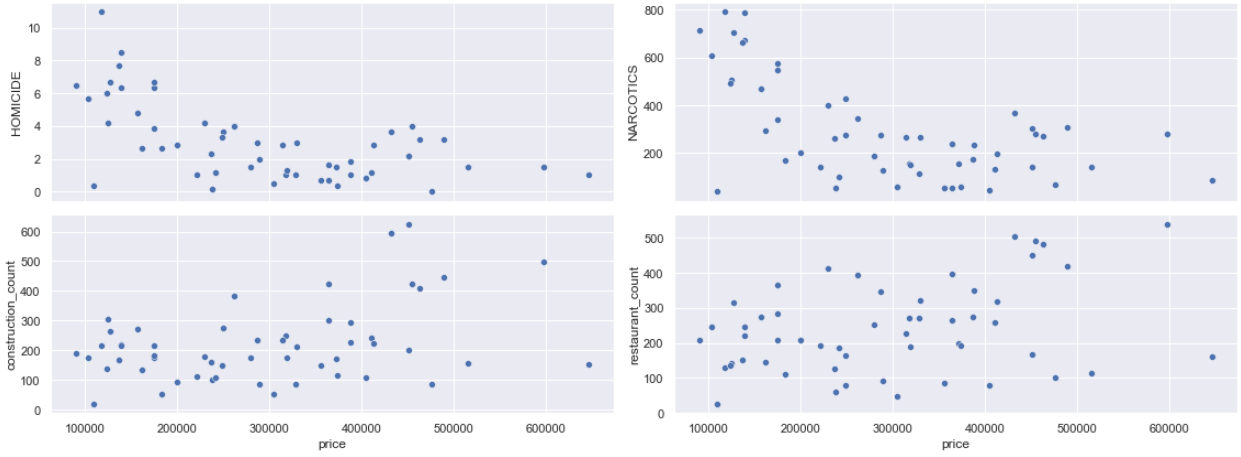


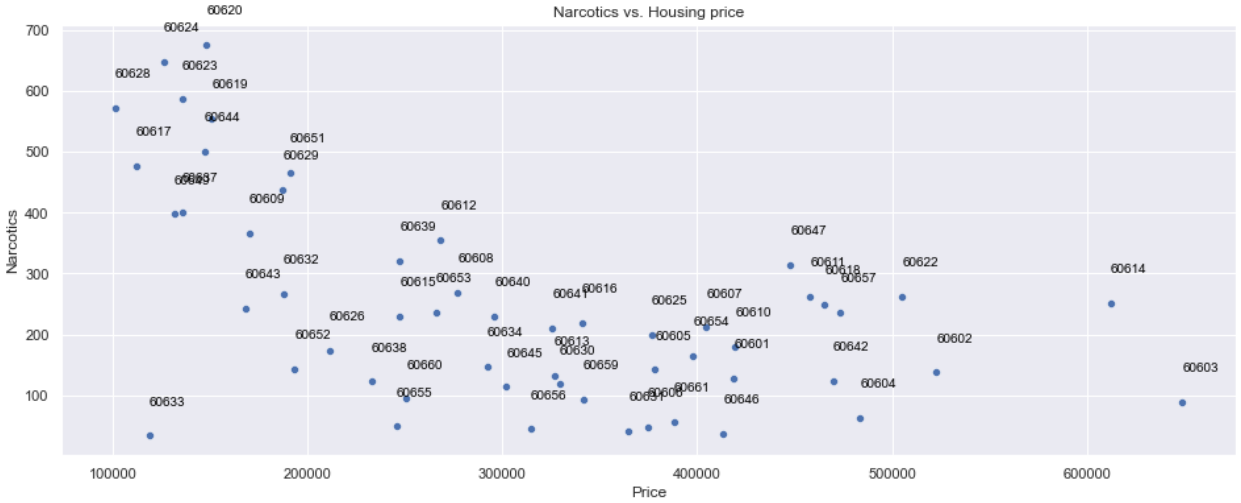


I merged all the yearly data for each zip code and kept it in a dictionary with zipcode as a key and the time series data as value and used that to find the relation between price and other features for a given time period

Below I plotted the scatter plot of the price and some of the other features from 2014 to 2019 for all zip codes. Each point represents a zip code. We can see that as price increases the crime rate decreases. However, after the median house price increases above 400K we see a slight increase in crime rate as well. This can show that for the most popular neighborhoods such as the downtown area, due to the higher population theft and robbery increases as well.

Looking at the annotated price vs. Narcotics plot, we can see that zip code 60633 is an outlier. Although it has one of the lowest housing prices the crime rate is pretty low.





To identify the trendy neighborhood in advance, I found the lagged feature values for the previous 3, 4, and 5 years and added them as a feature to the dataframe. I also found the year by year change in features and added them as new features to the data as well. The model will look at the last 3 consecutive years information and will estimate the change in housing price over the next 3 years

5. Splitting data into test and train

Because here we have time series data, the order matters and we can't randomly split the data between test and train. For each zip code I kept the first 85% of the data for training and the rest for the test dataset. Because the goal is to find the housing price three years in advance, for each observation the feature includes the change in housing price and the other features for three consecutive years and the target value is the change in housing price three years after that. For example using features in 2014, 2015, and 2016 to predict the change in housing price from 2016 to 2019.

6.Machine Learning

preprocessing

I previously one hot encoded all the categorical variables so at this step we only have numerical variables. I used a standard scaler to make sure all the numeric values have the same scale.

Feature engineering
Other than the lagged features and the change in the lagged values, based on the EDA results I added the squared value of some of the features to the feature dataframe as well.

Trying vanilla models
I tried a few vanilla models shown below with cross validation to see which ones have a better initial performance and compared their performance to a baseline model. I looked at R-squared, mean absolute error and mean squared error metrics. The Ridge model had the best performance.

	r2	mae	rmse
DummyRegressor	-0.016509	45179.678083	52563.660426
LinearRegression	0.809325	17815.507418	22532.591576
Ridge	0.848368	15792.808242	20253.714657
ElasticNet	0.604103	25140.266433	32711.531128
RandomForestRegressor	0.663020	21418.122908	30205.066289
SVR	-0.056677	45034.171238	53645.347122
XGBRegressor	0.638123	22133.815950	31167.823736

Model selection (Hyperparameter tuning)

I chose xgboost, random forest and ridge for hyperparameter tuning and further analysis. I used mean absolute error as the scoring metric because it's less sensitive to outliers compared to root mean squared error. The results for each model is shown below:

Ridge:

train MAE: 18613, test MAE: 19864, test RMSE: 25827

	feature_importance
price_lag3_value	0.111911
price_change_lag3_lag4	0.072490
construction_count	0.068175
construction_count_lag5_value	0.056521
NARCOTICS	0.049741
NARCOTICS_squared	0.043937
HOMICIDE_squared	0.043780
construction_count_change_lag3_lag5	0.043260
construction_count_lag4_value	0.039280
construction_count_lag3_value	0.032380

Random forest :

train MAE: 11060, test MAE: 25141, test RMSE: 28415

XGBOOST:

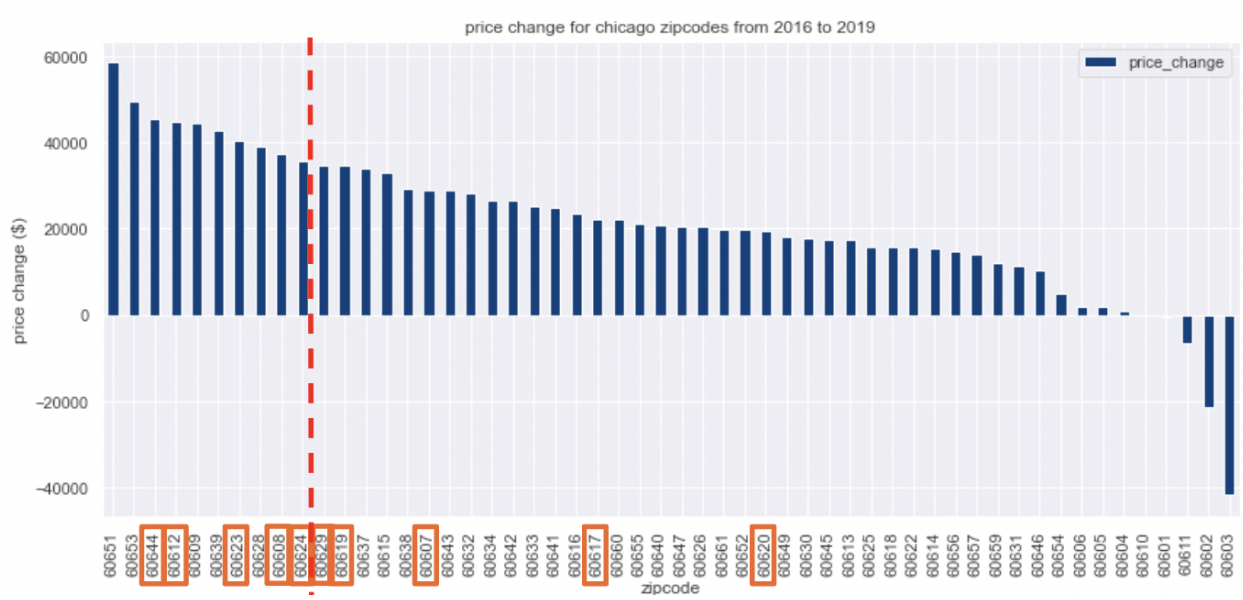
train RMSE: 19993, test RMSE: 26080, test MAE: 20470

7.Results

I used all three trained models to predict the change in housing price from 2016 to 2019. Then I compared the top 10 predicted zip codes by each model to the zip codes with the maximum value increase based on the housing price data. The results show that the modes were successful in finding four of the top 10 zip codes.

Correctly identified zip codes by Ridge model

60651, 60623, 60624, 60628, 60608, 60612



Correctly identified zip codes by xgboost model

60608

7. Future Improvements

1. Adding more features could help improve the model performance. Some features like the population for each zip code in each year, and the demographics information for each zip code could have been useful. Part of this information can be added to the model after 2020 census data is released.
2. More feature engineering also could make the model results more accurate