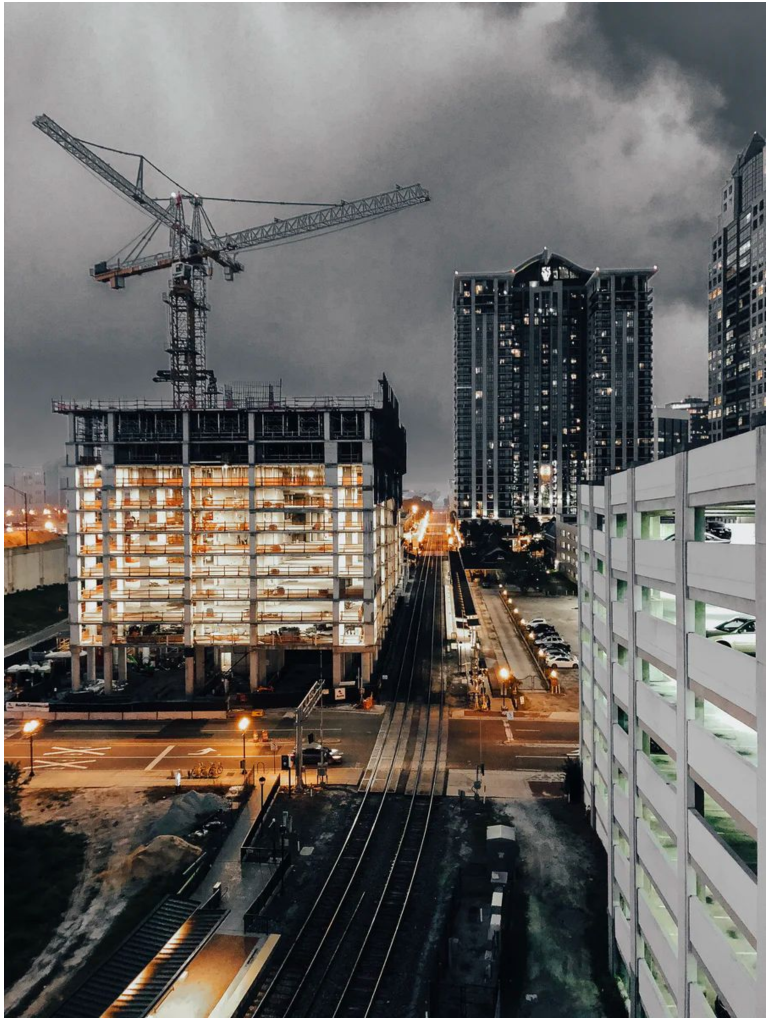
Finding the next trendy neighborhood in chicago

Problem Statement

- Old and neglected neighborhoods can quickly transform into a popular and trendy neighborhood

- This causes 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. The property will see the highest rise in housing price compared to other neighborhoods in the same city.



Goal

- Build a model that identifies next trendy zip codes

  Approach 1:
  - The model predicts the housing price three years in advance using historical housing prices and other factors that can affect the housing price in a neighborhood.

  Approach 2:
  - The model predicts the change in housing price two years in advance using the same features

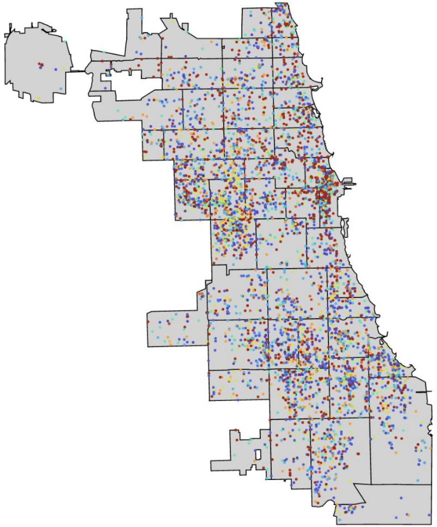
Data Sources

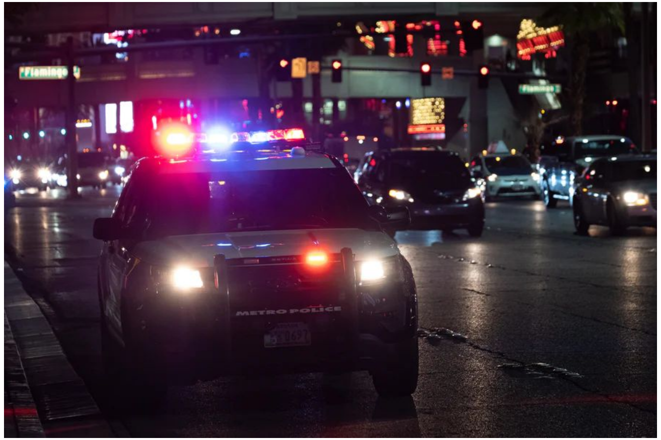
Limitations:

- Data has to be publicly available
- Data has to include location information (zip code, latitude, longitude)

Data:

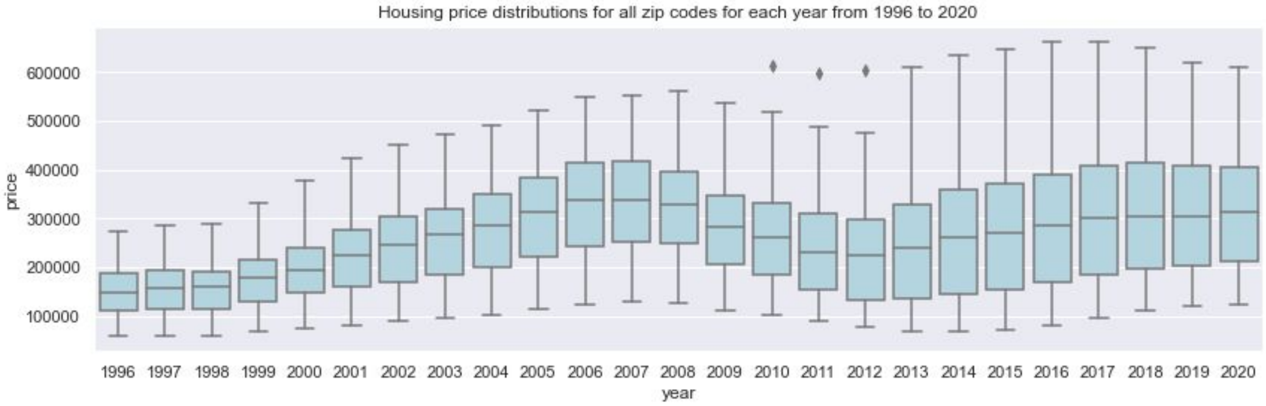
- building permits
  - City of Chicago Data Portal
- valid retail food licenses
  - City of Chicago Data Portal
- crime rate
  - City of Chicago Website
- historical housing prices
  - Zillow
- zip code information
  - Cybo

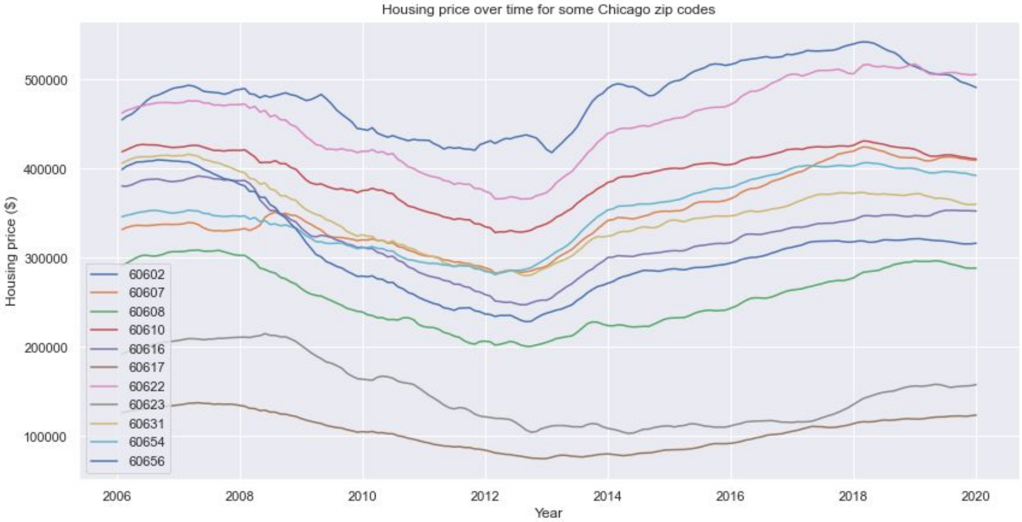




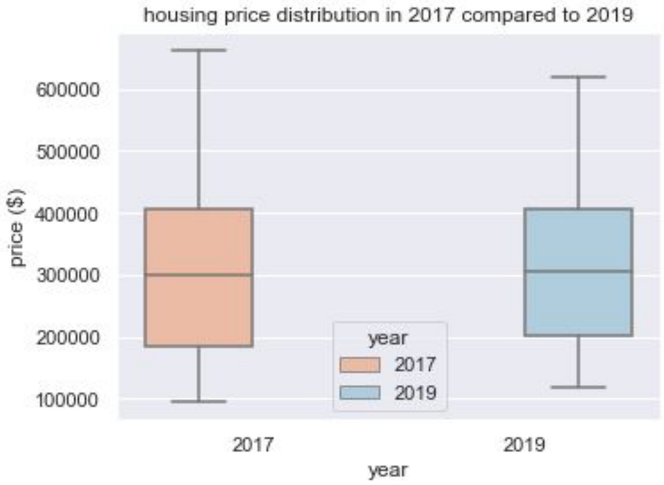
Housing Price Over Time

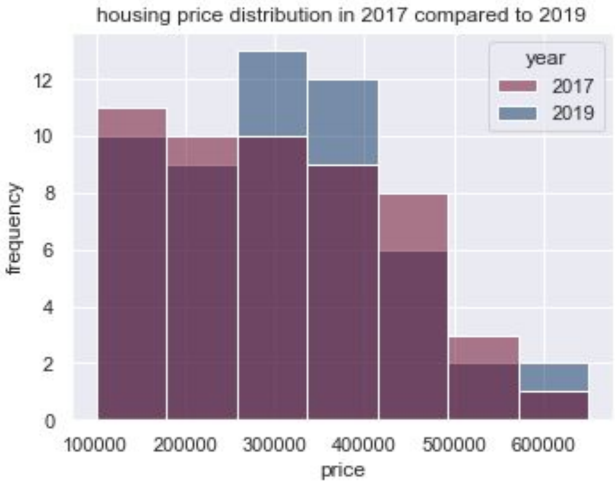
- The gap between the least and most expensive housing prices has increased over time.
- Some zip codes has seen increase in real estate value while some lost value over time.

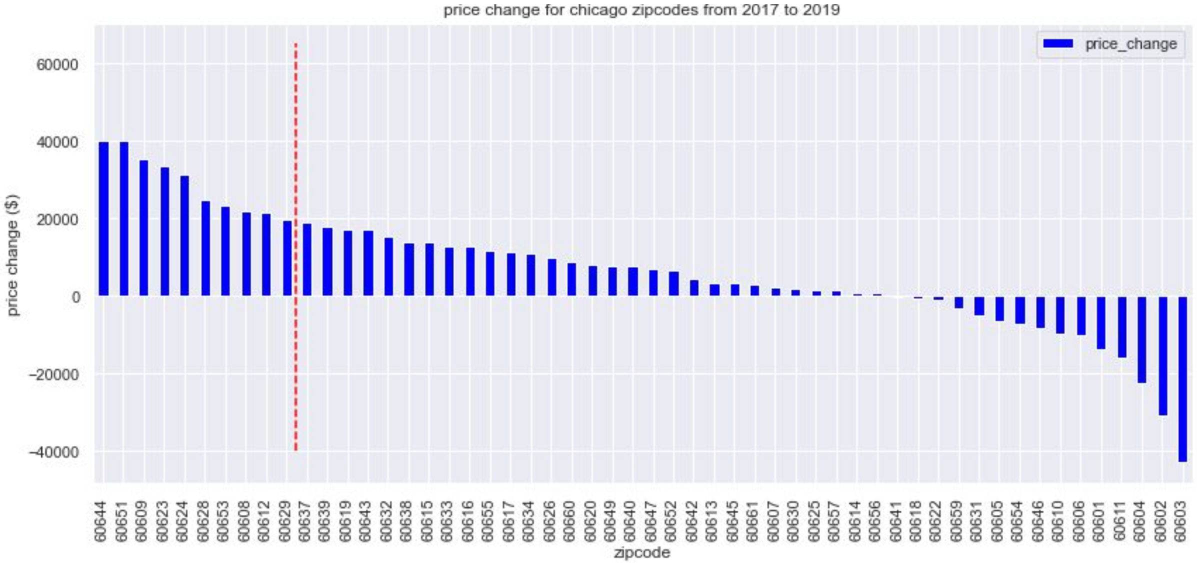




Housing Price Over Time





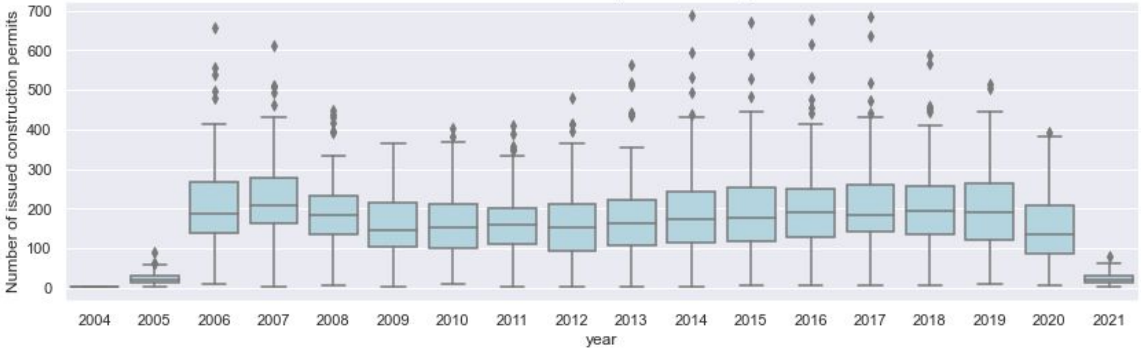


Construction and Renovation Permits Data

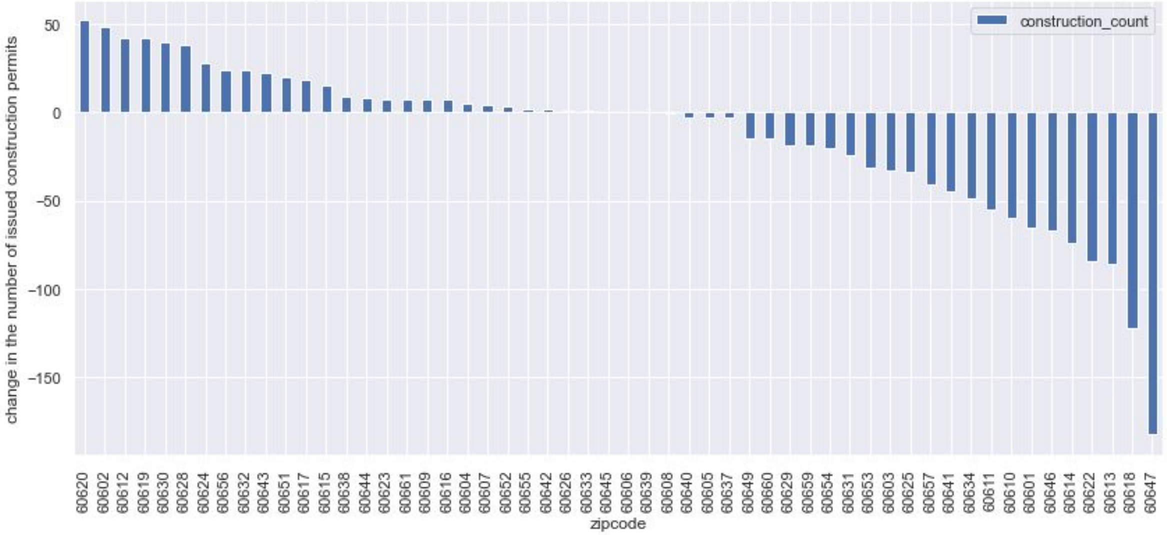
- The median construction count dropped during the housing market crises but slowly increased afterwards.

- In each year there are few zip codes that are outliers and have much more construction counts compared to other zip codes.

Distribution of number of issued construction permits for each year from 2004 to 2021



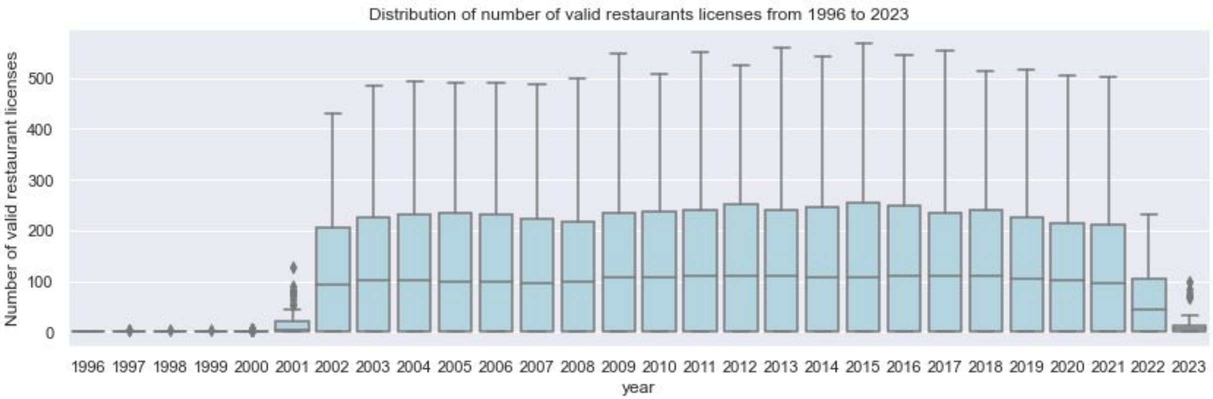
change in the number of construction permits issued for chicago zipcodes from 2017 to 2019

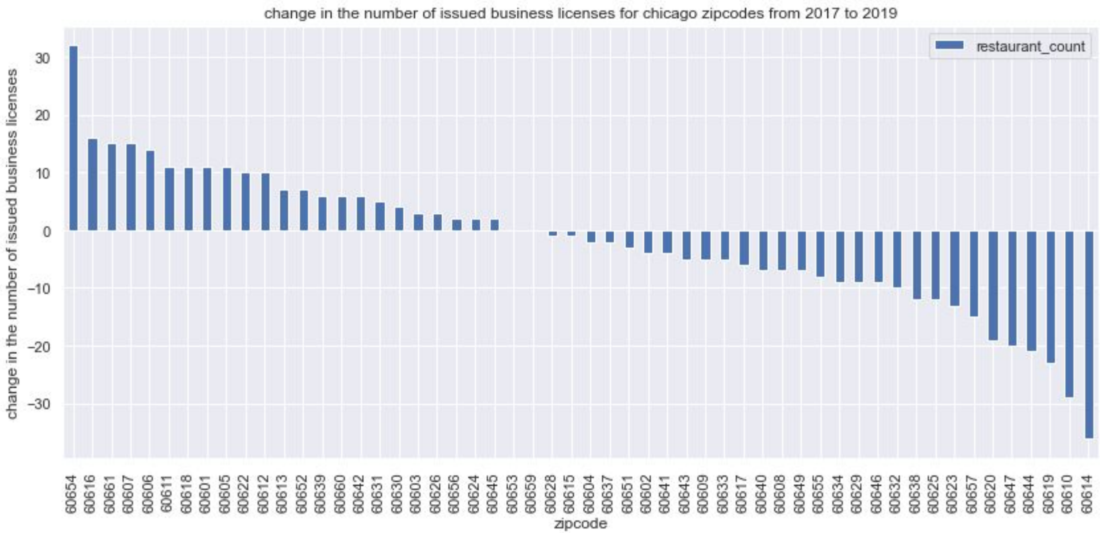


Issued Restaurants license Data

- Yearly distribution of valid business licenses is almost uniform over time.

- 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.



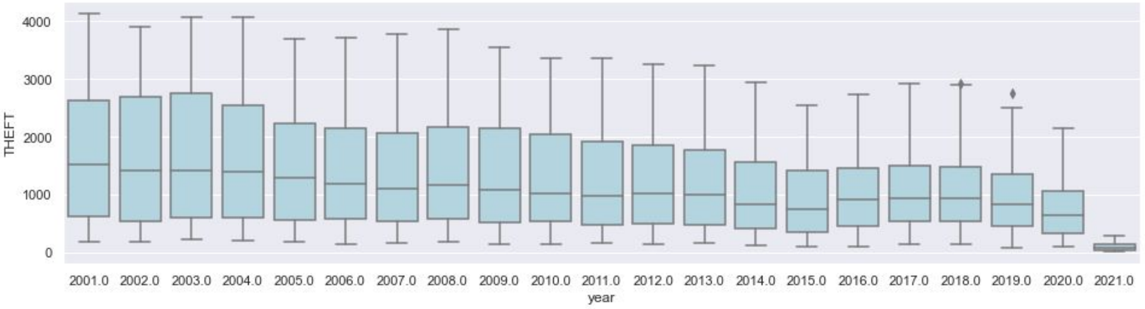


Crime

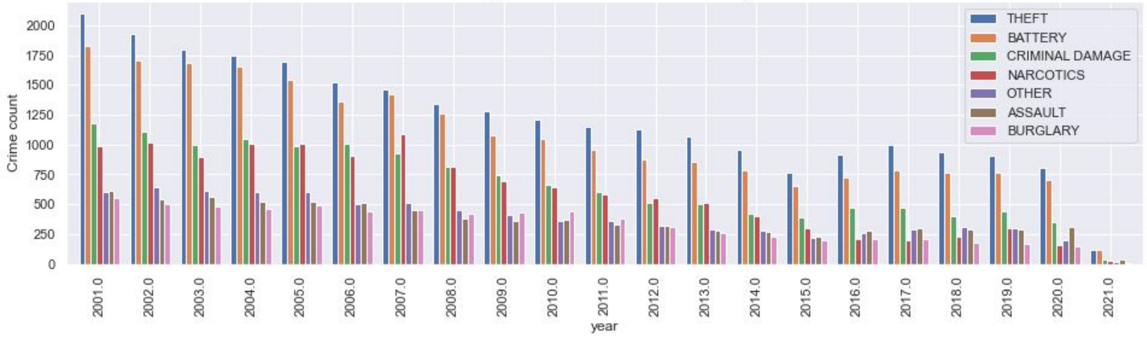
- Crime rate has decreased in Chicago over time.

- Zip codes 60653 saw the highest increase in housing value over this time period and we can see that the crime rate has decreased for this zip code.

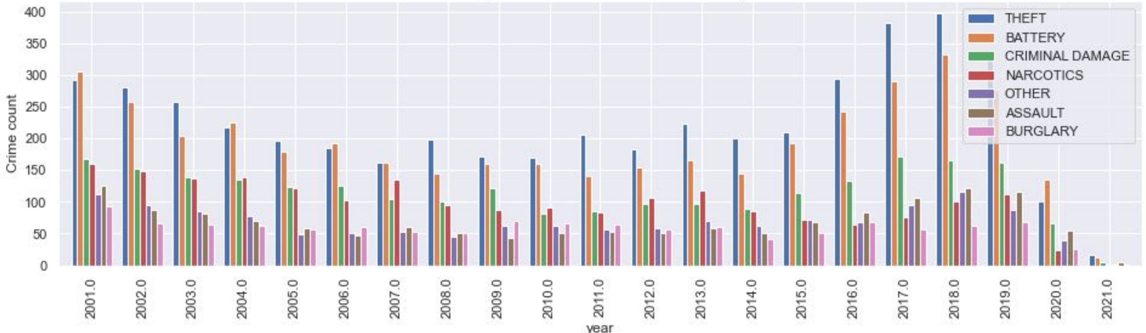
- Zip code 60603 has seen a decrease in housing value over the last 5 years. Crime rate has increased in this zip code over this time period.



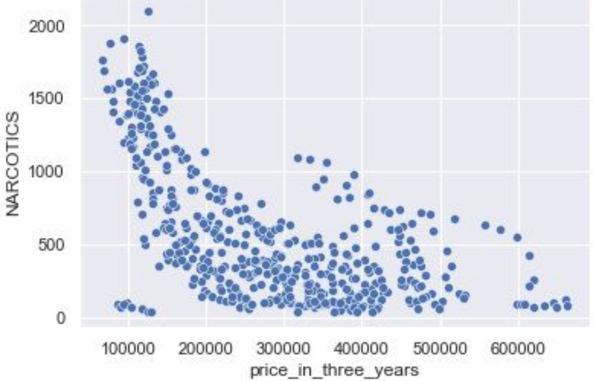
Different type of crime count over time in 60653 zipcode

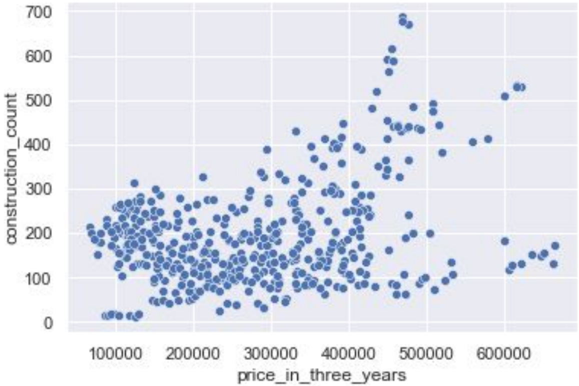


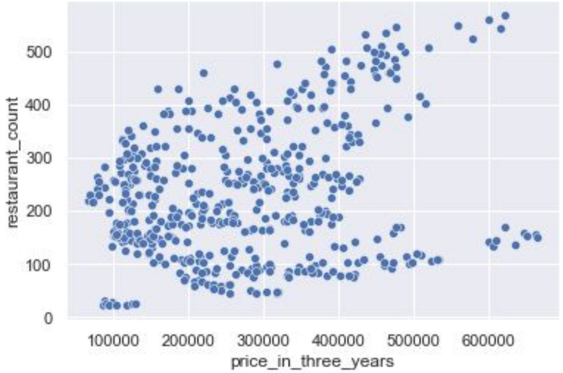
Different type of crime count over time in 60603 zipcode

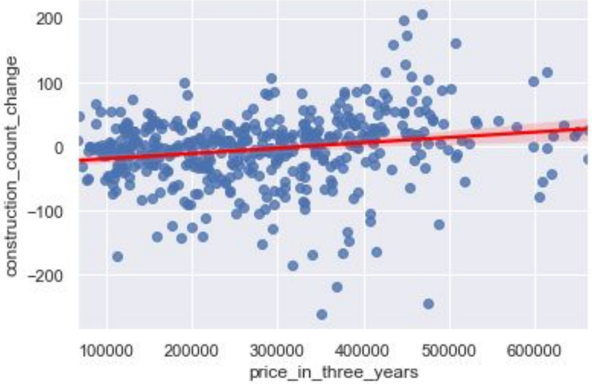


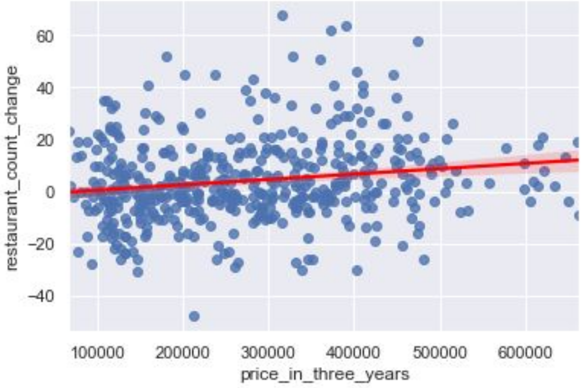
Relation Between Target and Features











Machine Learning - preprocessing

- **Preprocessing**
  - One hot encoding categorical values
  - Scaling numerical values

- **Feature engineering**
  - Lagged values
  - change in the lagged values
  - squared
  - Inversed

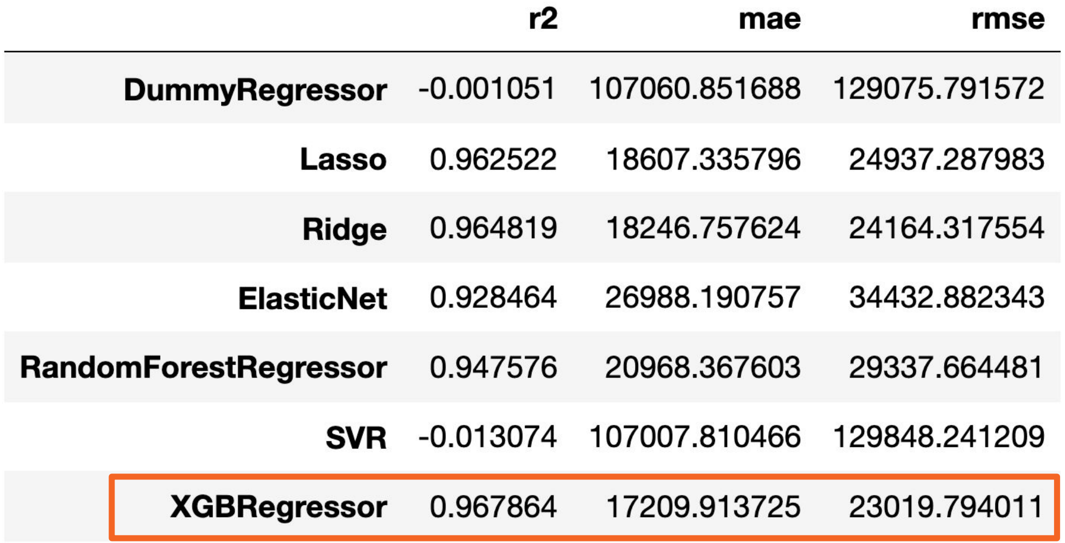
- **Train and Test split**

Machine Learning

model 1 - predicting housing price three years in advance

Trying vanilla models

- Trying few different vanilla models with cross validation to see which ones have a better initial performance
- compared model performance to a baseline model.
- The XGBRegressor model had the best performance.



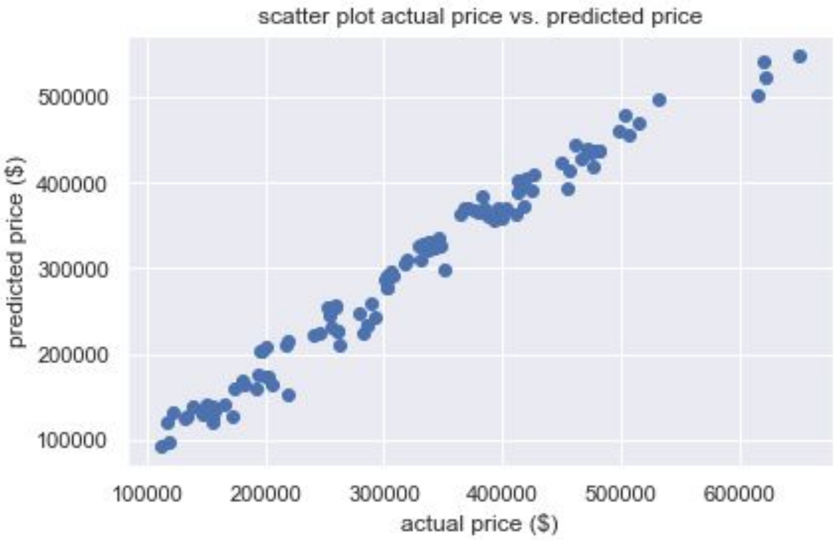
	r2	mae	rmse
DummyRegressor	-0.001051	107060.851688	129075.791572
Lasso	0.962522	18607.335796	24937.287983
Ridge	0.964819	18246.757624	24164.317554
ElasticNet	0.928464	26988.190757	34432.882343
RandomForestRegressor	0.947576	20968.367603	29337.664481
SVR	-0.013074	107007.810466	129848.241209
XGBRegressor	0.967864	17209.913725	23019.794011

MODEL 1 - predicting housing price three years in advance

XGBOOST

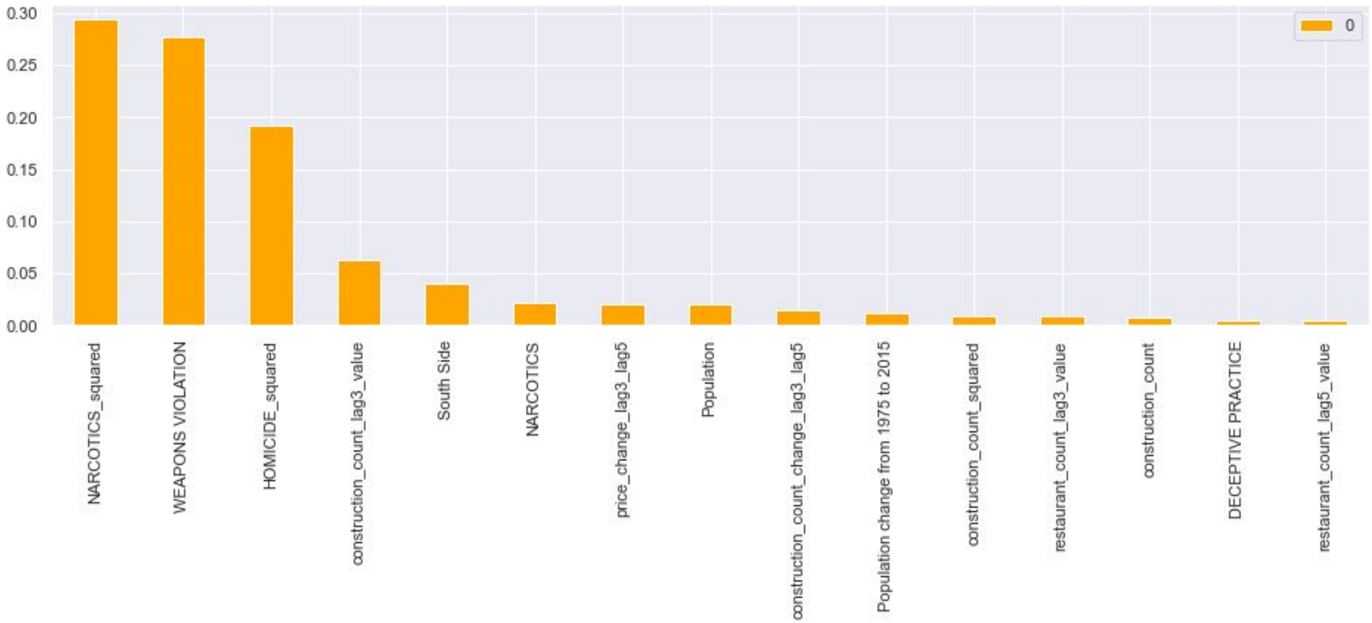
- Used mean absolute error as the scoring metric because it’s less sensitive to outliers compared to root mean squared error.

Train MAE	13577
Test MAE	26247
Test RMSE	27053
R squared	0.93



Model 1 Results

- Feature importance by the XGBOOST model

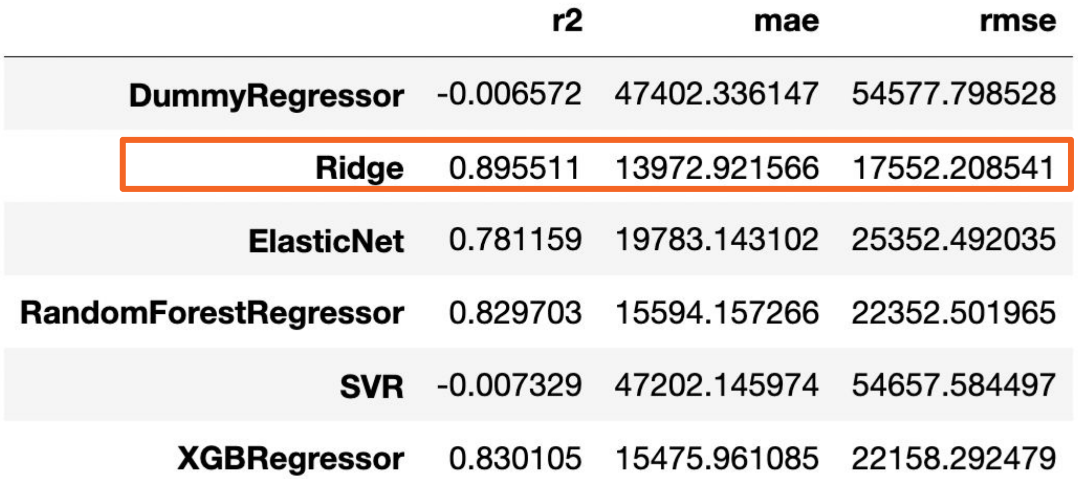


- Successful in identifying only three zip codes in top ten zip codes with maximum value increase from 2017 to 2019

60623, 60628, 60624

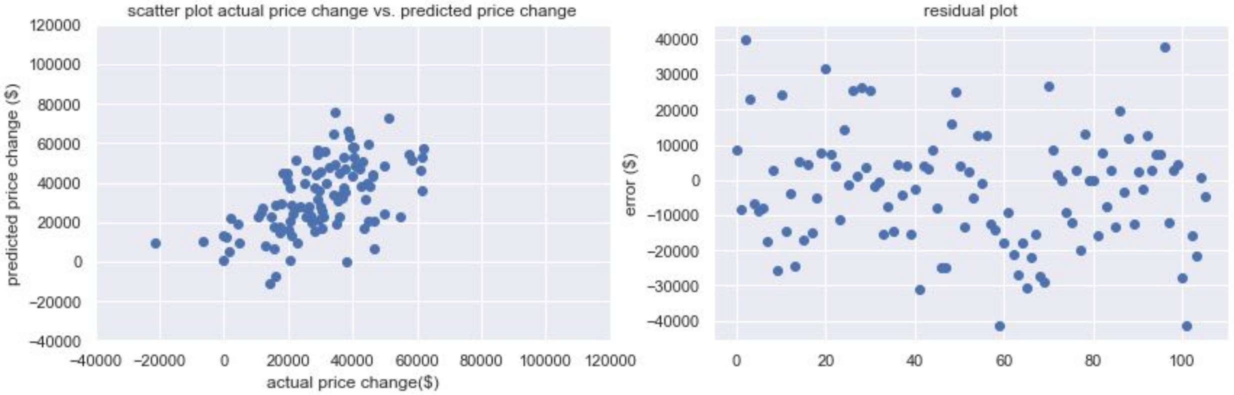
model 2 - predicting change in housing price two years in advance

- Change the target from housing price to change in housing price over the period of two years.

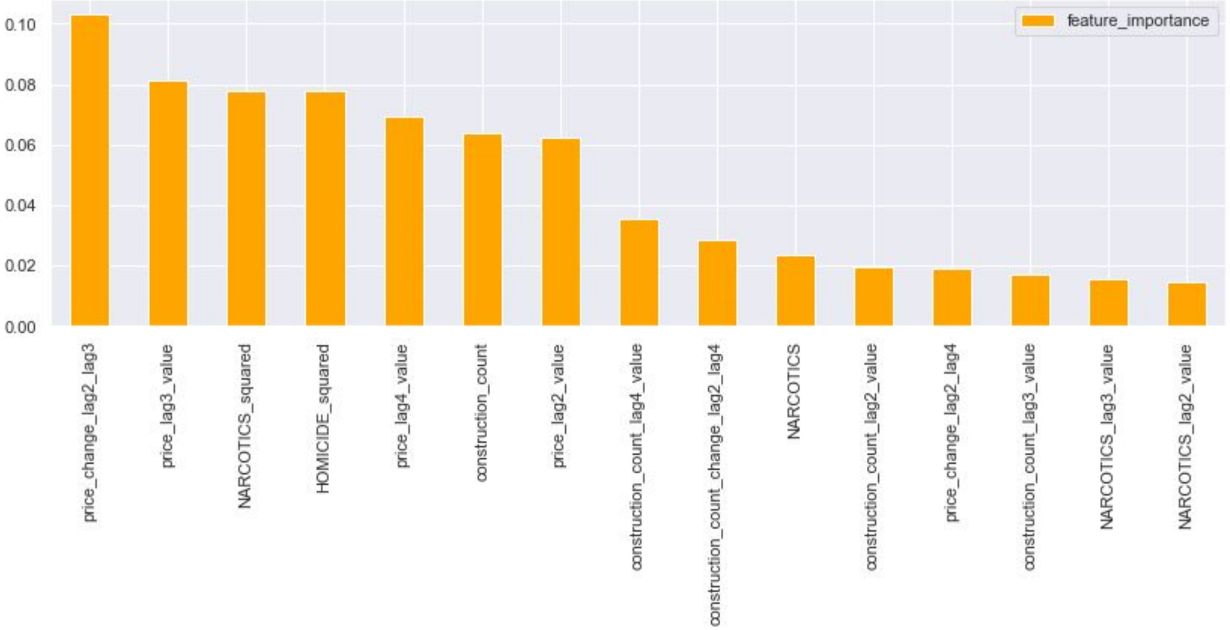


	r2	mae	rmse
DummyRegressor	-0.006572	47402.336147	54577.798528
Ridge	0.895511	13972.921566	17552.208541
ElasticNet	0.781159	19783.143102	25352.492035
RandomForestRegressor	0.829703	15594.157266	22352.501965
SVR	-0.007329	47202.145974	54657.584497
XGBRegressor	0.830105	15475.961085	22158.292479

Model 2 Results



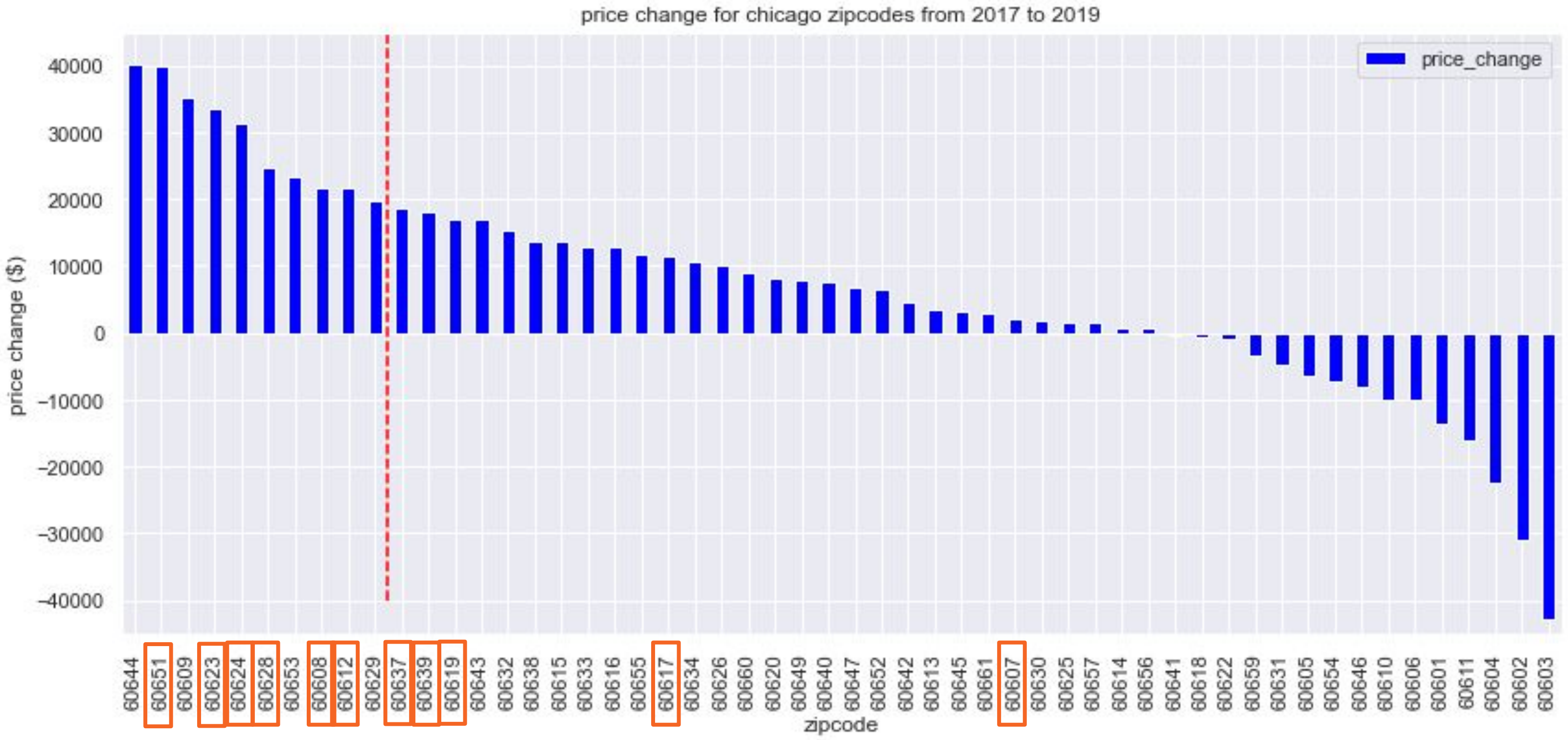
Train MAE	13000
Test MAE	13075
Test RMSE	16529
R squared	0.05



Results

- Top 10 zip codes recommended by the model:
  60619, 60637, 60628, 60612, 60617, 60608, 60623, 60651, 60607, 60639

- Zip codes identified correctly by the model:
  60651, 60623, 60624, 60628, 60608, 60612



Results

Home buyer can decide based on the top 10 recommended zip codes as well as other factors such as the

- budget,
- type of real estate (house, condo,..)
- number of schools in the area,...

to make a more informed decision

http://www.aag.com/

Improvements

- More data:
  - Zip code specific data such as demographics and population over time
  - Adding other type of data (economy, GDP...)
- More feature engineering