Predicting Median Housing Prices for Chicago Zip Codes Three Years in Advance

1.Problem Statement

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. The goal of this project is to estimate the median real estate prices for all Chicago zip codes three years in advance. The result would be of interest to future home buyers and real estate investors.

I used ARIMA models to estimate the real estate prices three years in advance. Arima (AutoRegressive Integrated Moving Average) is a popular and widely used method for time series forecasting. I developed two arima models for housing price forecasting. The first model uses a univariate time series in which I only used the historical housing price to forecast the housing price in the next three years. In the second model I used two exogenous variables (number of issued construction permits per month and monthly crime rate) as well as the historical housing prices to predict the future housing price. If crime rate increases in a neighborhood people start to leave and there would be high supply and low demand, this will cause a drop in real estate values, the opposite is also valid.

2.Data Sources:

building permits : City of Chicago Data Portal

crime rate: City of Chicago Website

historical housing prices: Zillow

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.

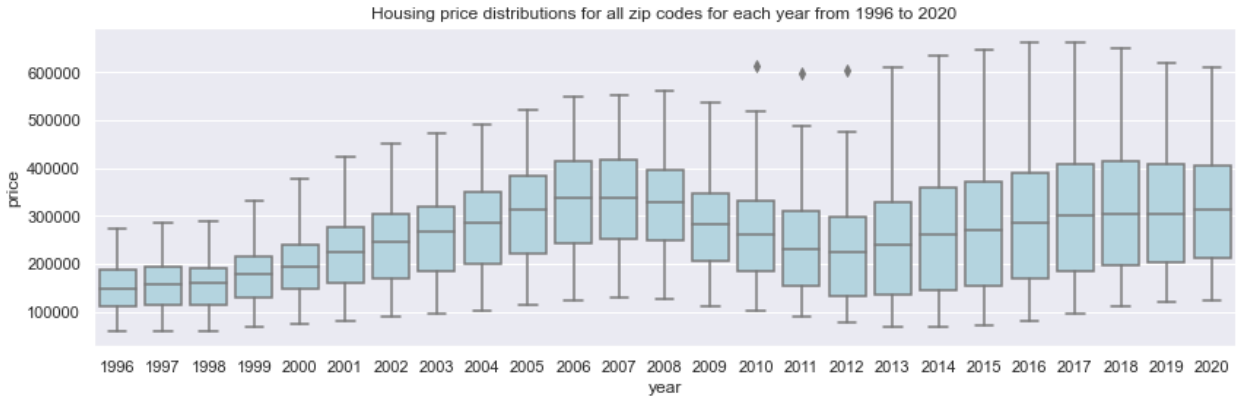
Problem 2. The latitude and longitude information of each construction permit was necessary to find the zip code where the construction happened. There were 476 entries with missing latitude and longitude information. Because there was no other way to find the zip code, I removed the entries with missing location from this data set. I also had to get the zip code data from the

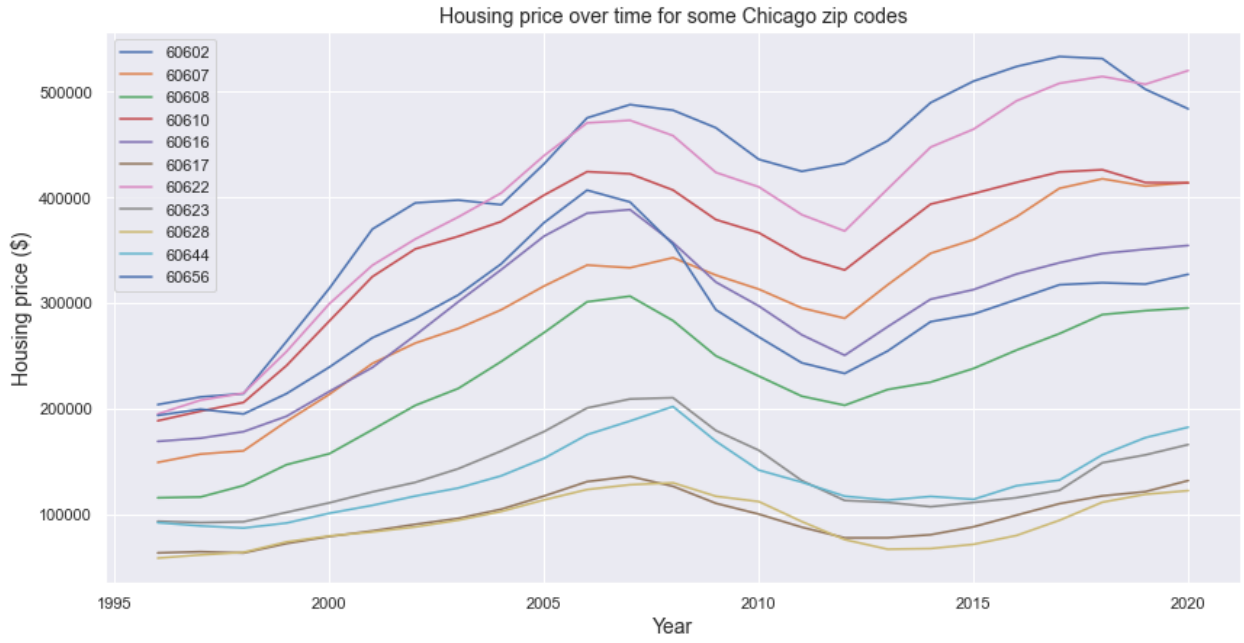
latitude longitude information. To do this I used an open source package from github which allowed me to run a local server which responded with a zip code to requests containing latitude and longitude.

Problem 3. 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.

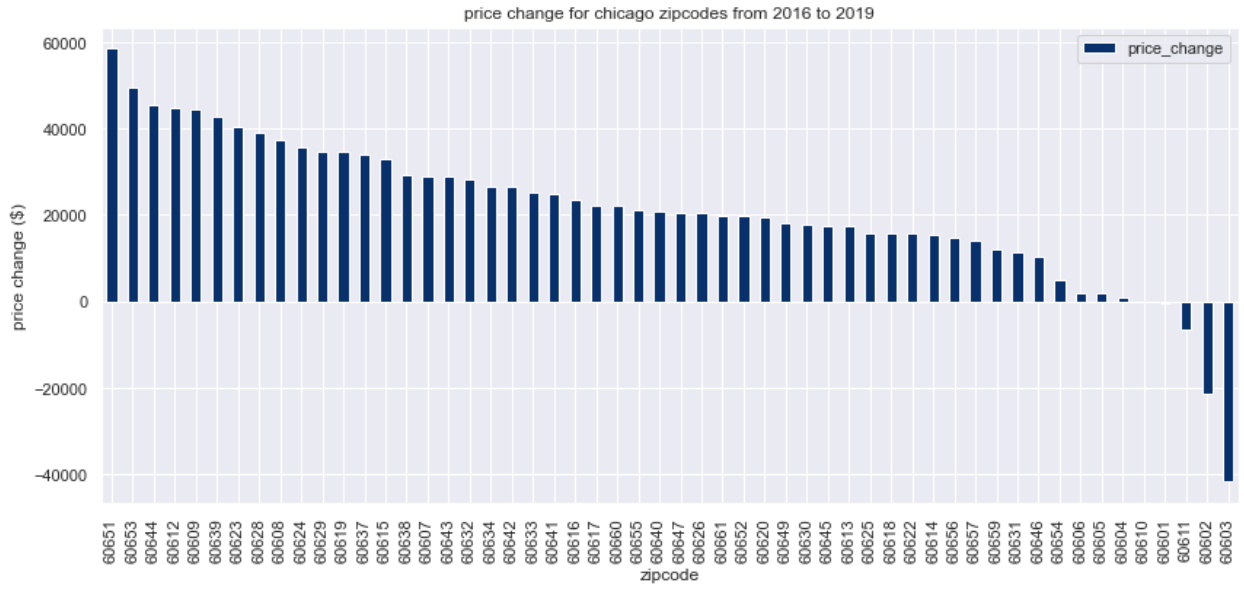
4.EDA

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



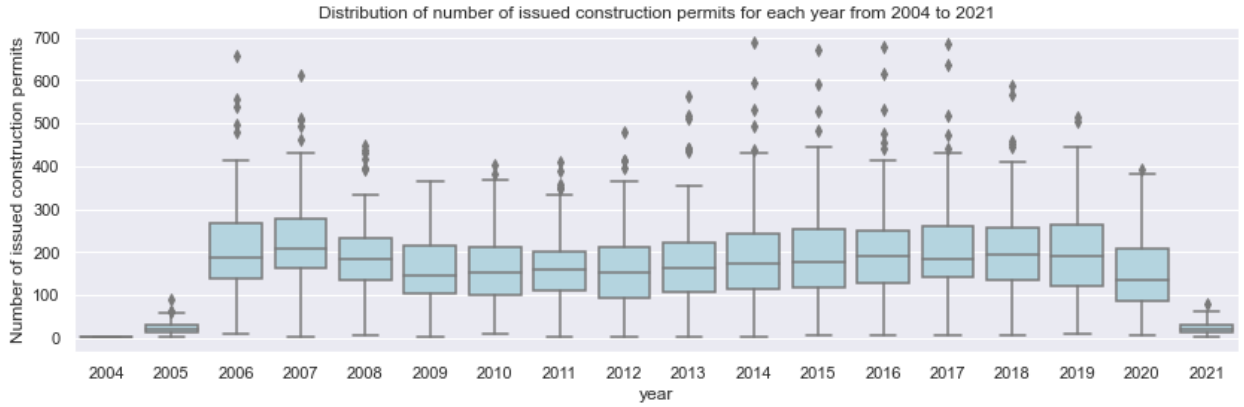


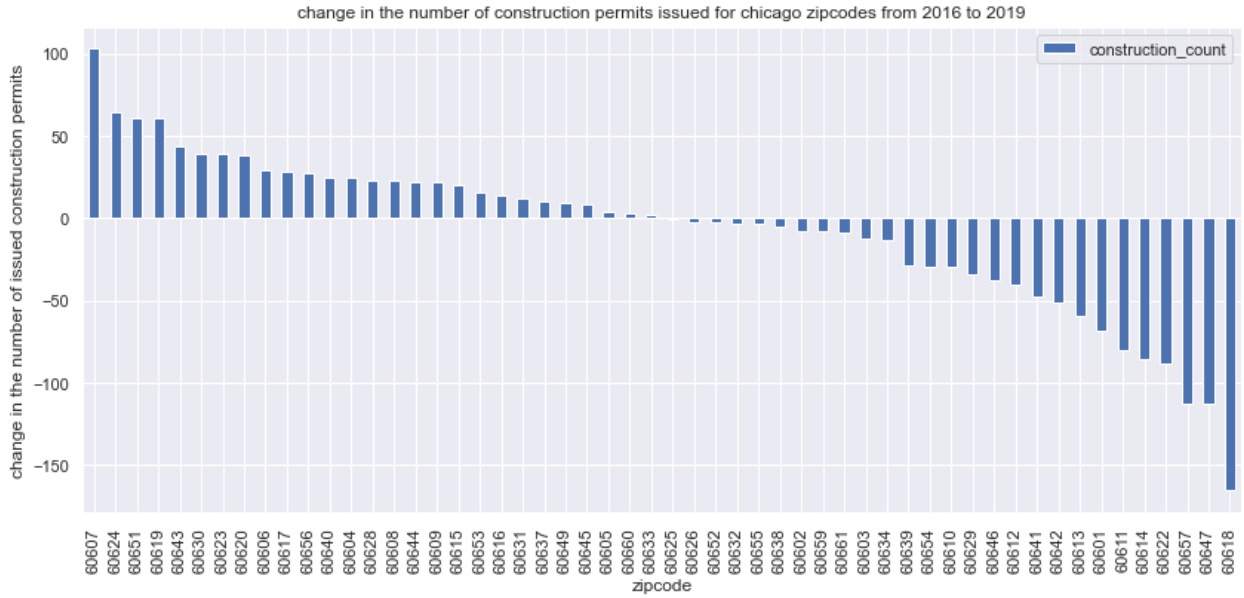
The plot below shows the change in housing price for all Chicago zip codes from 2016 to 2019. We can see from this graph that zip codes 60651, 60653, 60644, and 60612 have the maximum increase in the value over this time period however some zip codes such as 60602 and 60603 saw a decrease in housing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



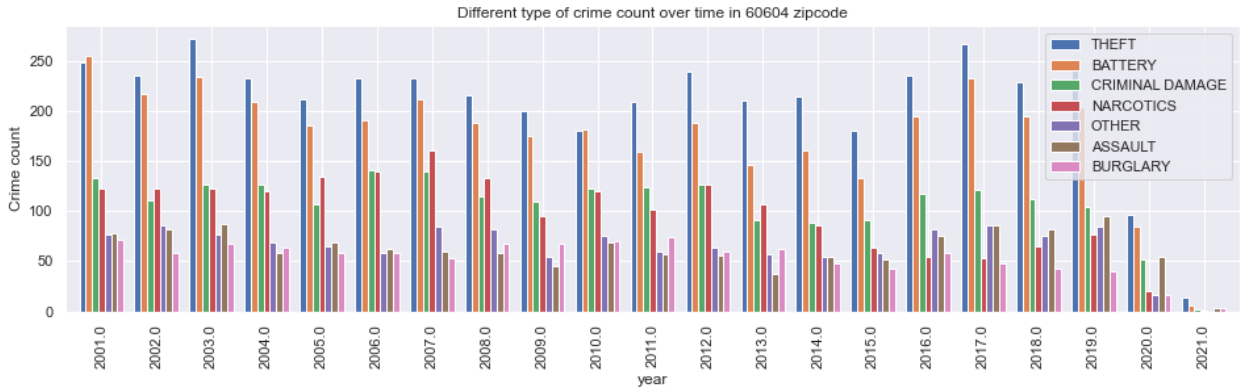
Building permits I plotted the distribution of the number of issued permits for all zip codes from 2004 to 2021. As can be seen, the data before 2006 is not complete and I only used the data after 2006 for modeling. In each year there are few zip codes that are outliers and have much more construction counts compared to other zip codes. We can see that the median construction count dropped during the housing market crises but slowly increased afterwards.

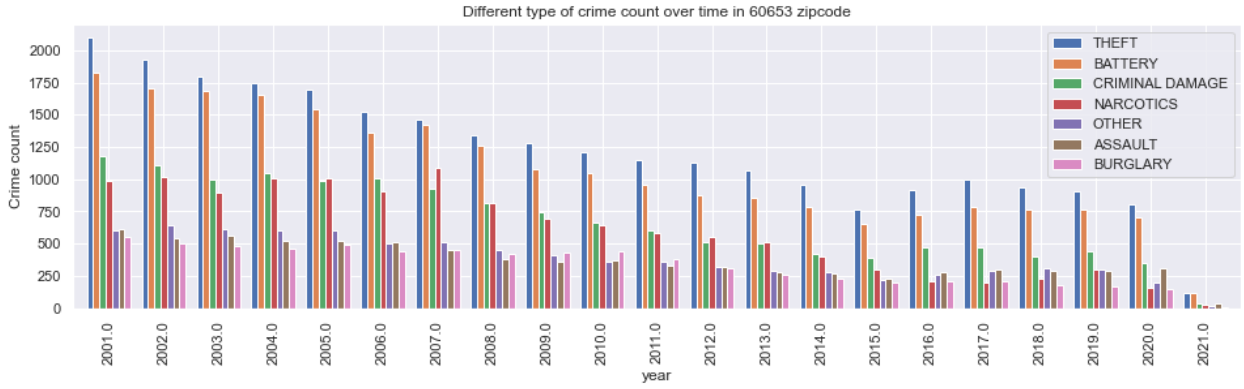
To have a better understanding of how much construction count in a zip code changed over time, I also plotted the change in the number of permits issued for a given time period for all the zipcodes.

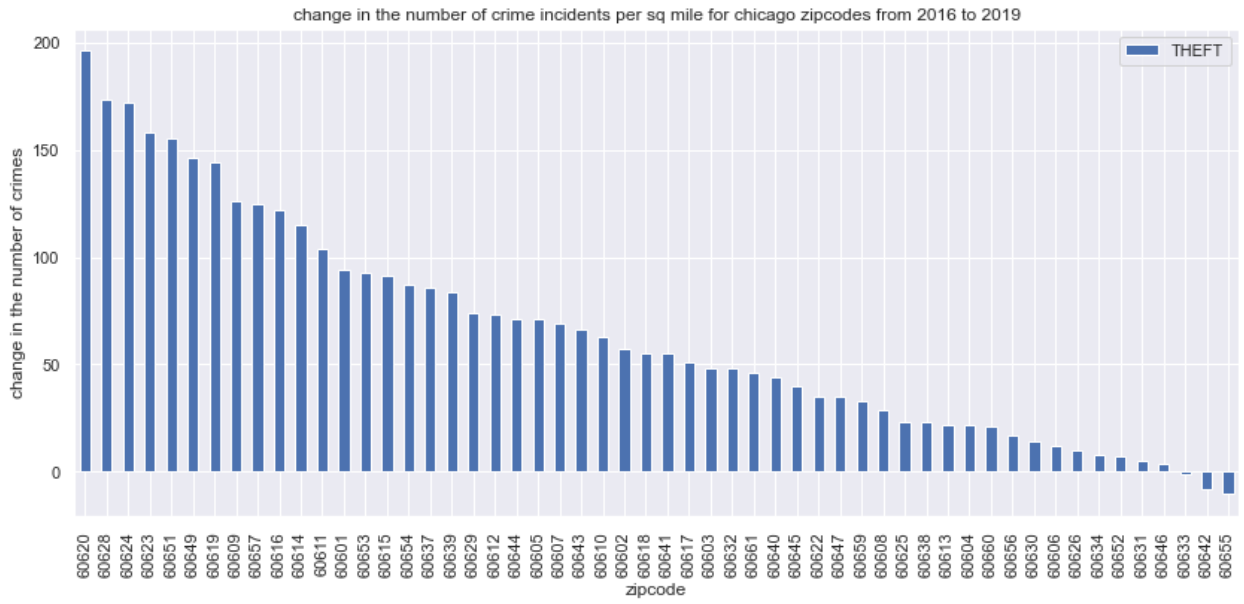


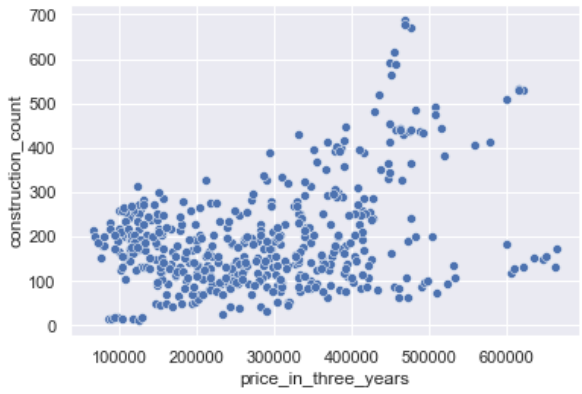


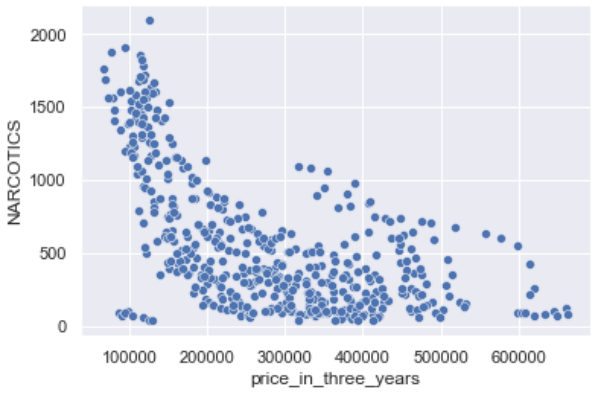
Crime I plotted the theft count distribution over time and we can see clearly that overall the number of theft incidents has decreased over time. To see how different crime types have changed over time for a specific zip code, I plotted bar plots of 7 different crime types for zip codes 60603, 60604, 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.

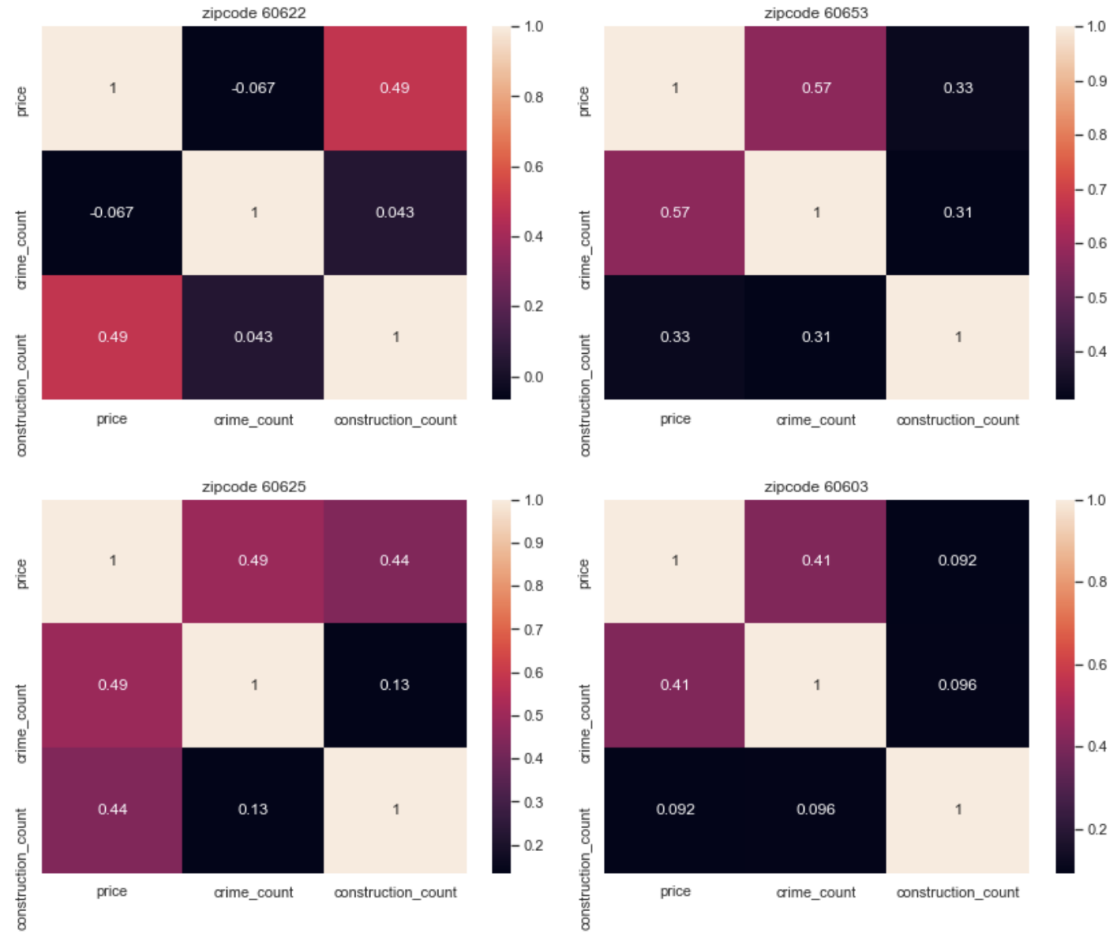

Different type of crime count over time in 60604 zipcode
Crime count
year
THEFT
BATTERY
CRIMINAL DAMAGE
NARCOTICS
OTHER
ASSAULT
BURGLARY


Different type of crime count over time in 60653 zipcode
Crime count
year
THEFT
BATTERY
CRIMINAL DAMAGE
NARCOTICS
OTHER
ASSAULT
BURGLARY


change in the number of crime incidents per sq mile for chicago zipcodes from 2016 to 2019
change in the number of crimes
zipcode
THEFT







5.Modeling

5.1. ARIMA modeling procedure:

Checking for stationarity: To test for stationarity of the time series I used the ndiffs function from pmdarima.arima which estimates the number of differences required to make a given time

series stationary. I used the maximum value between the ndiffs results for Kwiatkowski–Phillips–Schmidt–Shin (KPSS) and Augmented Dickey–Fuller tests (ADF) tests for differencing. After making the time series stationary, I plotted the ACF and PACF to identify the correct model parameters.

Plotting ACF and PACF plots: The ACF plot shows the correlation between a time series and the lagged version of itself. The PACF plot also shows the correlation between a time series and the lagged version of itself after we removed the correlation effects of the smaller lags.

Model parameters:

The parameters of the arima model are:

p: number of lagged observation considered in the model

d: degree of differencing

q: the order of moving average

In general:

1. If the amplitude of the ACF plot tails off with increasing lags and PACF cuts off after p lags, we have an AR(p) model.
2. If the amplitude of ACF cuts off after q lags and PACF tails off, we have a MA(q) model.
3. If both ACF and PACF amplitudes tail off then, we have an ARMA(p,q) model and we can't find the p and q value from the plots.

I used Akaike information criterion (AIC) and Bayesian information criterion (BIC) for model selection. Both BIC and AIC penalize models that are overly complex and have more parameters. Lower AIC indicates a better model.

Model diagnostics:

For model diagnostics I looked at the residual plots. Residuals of an ideal model should be uncorrelated white Guassian noise centered around zero.

- **Diagnostics plots**
    a. **Standardized residuals Plot**: In this plot the residuals should be distributed randomly with no obvious structure.

    b. **Histogram plus estimated density plot** shows the distribution of the residuals. In a good model, the KDE plot and the normal plot overlay on each other.

    c. **Normal Q-Q Plot** shows the distribution of a model residual compared to a normal distribution. The data points should lay on the normal distribution line.

    d. **Correlogram Plot** shows the ACF plot of the residuals.

- **test statistics:**I also used the model summary to check the results of Ljung-Box P(Q) and Jarque-Bera P(JB) test statistics.
    a. prob(Q) is the p-value for the test with the null hypothesis that the residuals are not correlated.
    b. prob(JB) is the p-value for the test with the null hypothesis that the residuals are normally distributed.

5.2. Auto-ARIMA

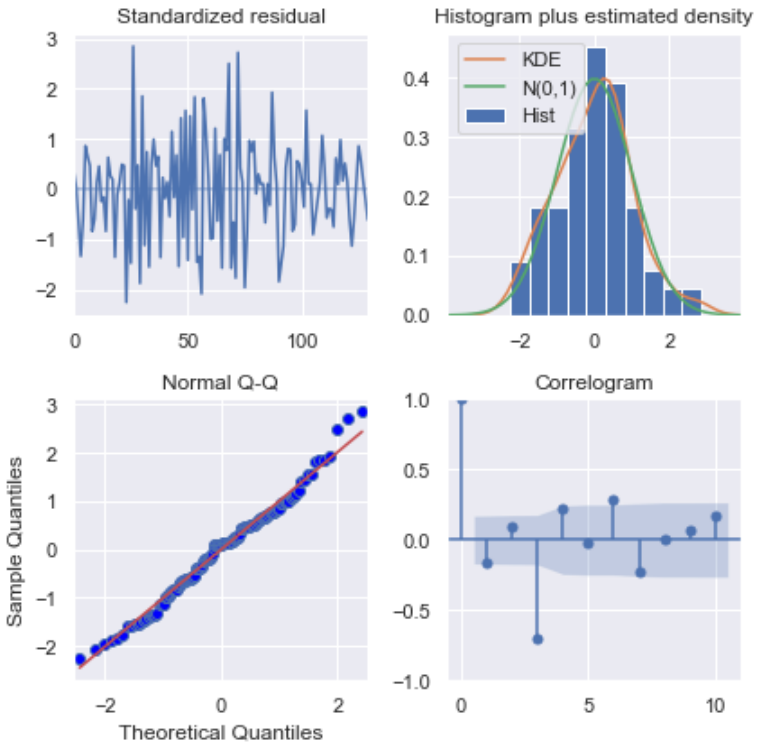
To find the best model for all 52 zip codes in the data set I used auto_arima from pmdarima library to automatically find the best parameters for the ARIMA models for each zip code, Reference.

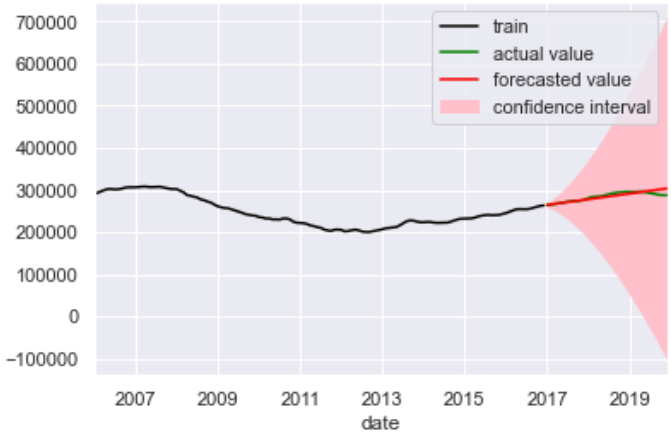
"Auto-ARIMA works by conducting differencing tests (i.e., Kwiatkowski–Phillips–Schmidt–Shin, Augmented Dickey-Fuller or Phillips–Perron) to determine the order of differencing, d, and then fitting models within ranges of defined start_p, max_p, start_q, max_q ranges. If the seasonal optional is enabled, auto-ARIMA also seeks to identify the optimal P and Q hyper- parameters after conducting the Canova-Hansen to determine the optimal order of seasonal differencing, D. In order to find the best model, auto-ARIMA optimizes for a given information_criterion, one of (‘aic’, ‘aicc’, ‘bic’, ‘hqic’, ‘oob’) (Akaike Information Criterion, Corrected Akaike Information Criterion, Bayesian Information Criterion, Hannan-Quinn Information Criterion, or “out of bag”–for validation scoring–respectively) and returns the ARIMA which minimizes the value."

5.3. Model 1: Univariate time series forecasting

Plots below show the forecast real estate prices and diagnostics plots for a few zip codes.

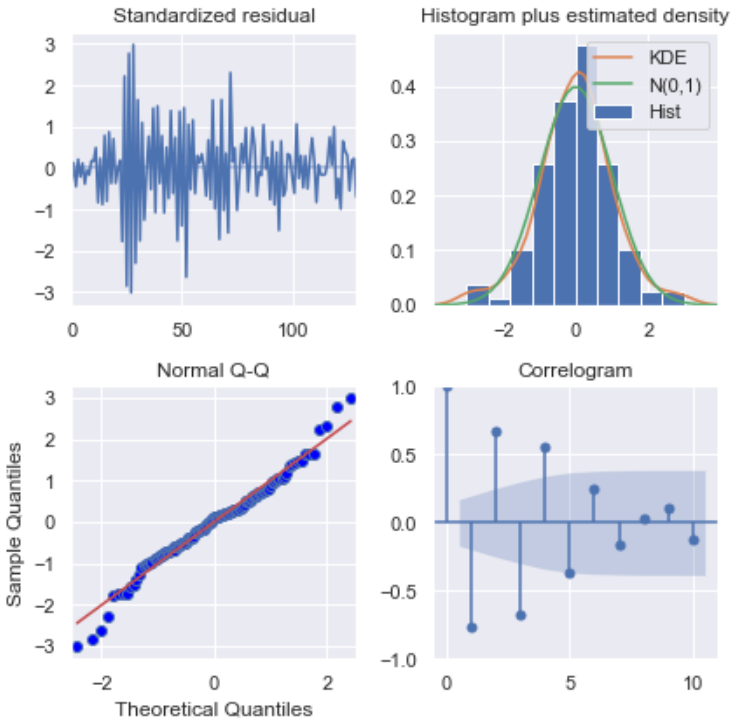
Zip code: 60608

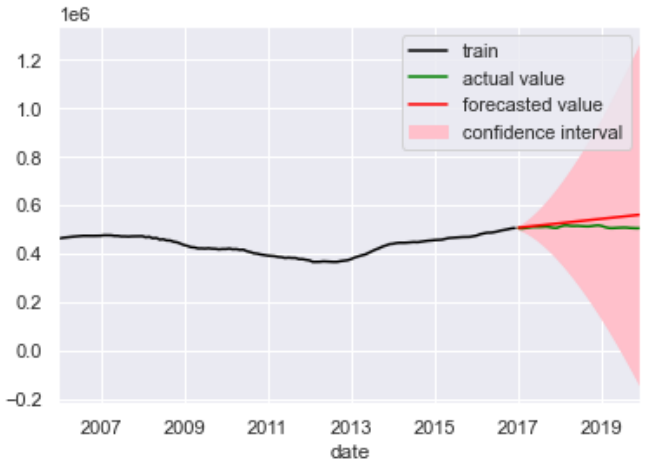




	forecast	actual_price
date		
2019-12-31	303927.0	288059.0

Zip code: 60612



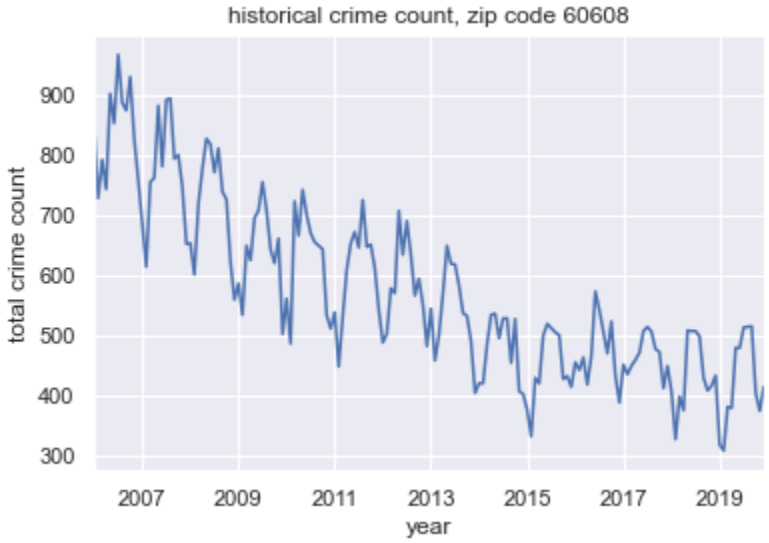


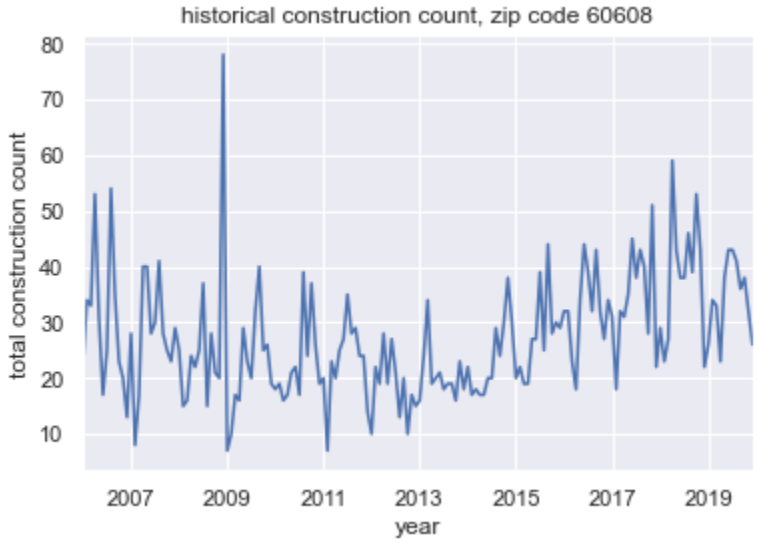
	forecast	actual_price
date		
2019-12-31	338748.0	285758.0

5.4. Model 2: Multivariate time series forecasting

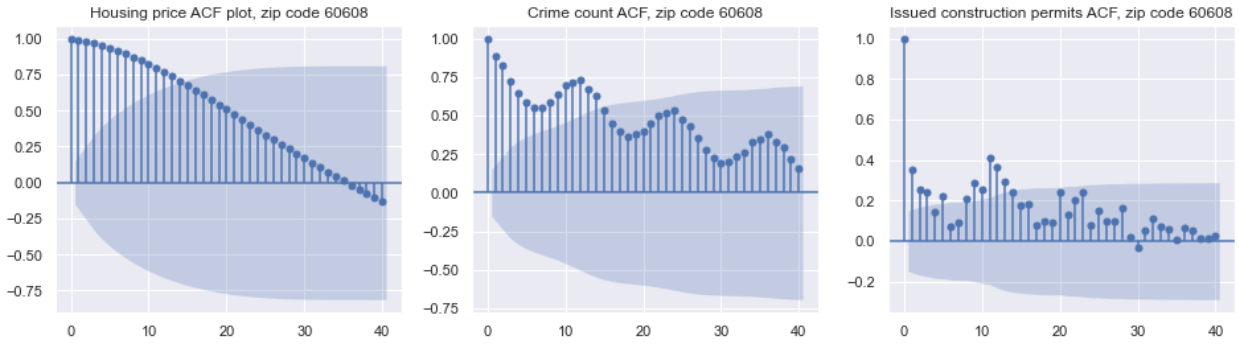
As mentioned before some factors such as crime, construction, and renovation count in a neighborhood can affect the real estate value in a positive or negative way. I used the total crime count and number of issued construction permits for each month as exogenous features in my model. To get a realistic idea of the model performance, I used ARIMA models to first

predict the crime and construction count for the test time period based on the historical data and then used those results as an input to forecast the housing price.
By plotting the crime count for a few zip codes we clearly can see the downward trend and yearly seasonality in the crime data. Crime increases during the summer time and decreases over the winter. The construction data also appears to have yearly seasonality although does not show a clear trend.



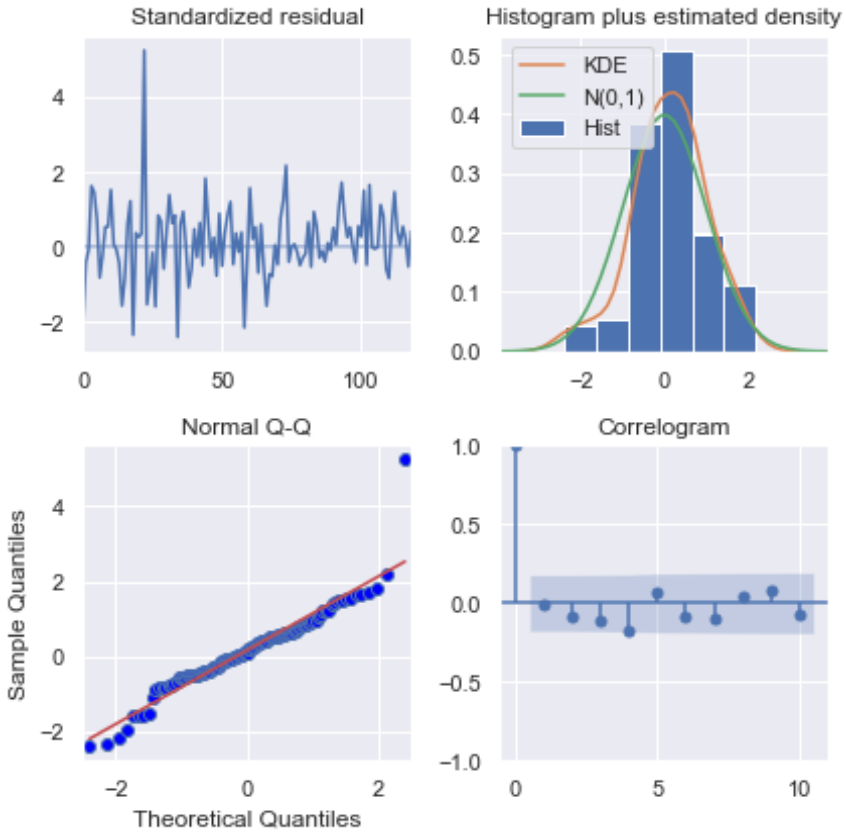


To better understand the seasonality of the data, I plotted the ACF for all three time series.

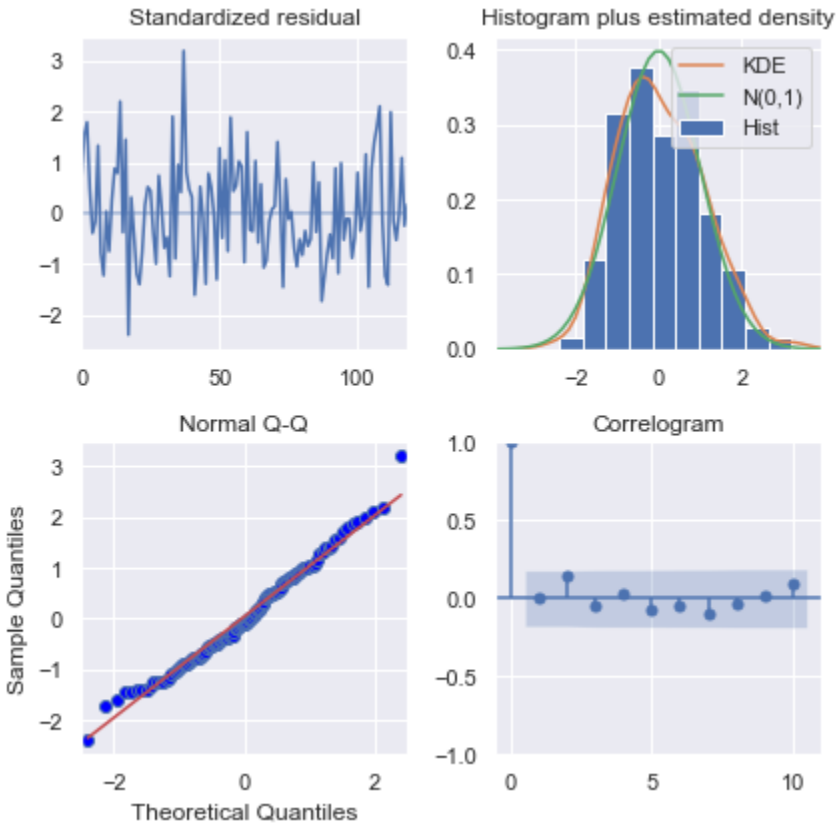


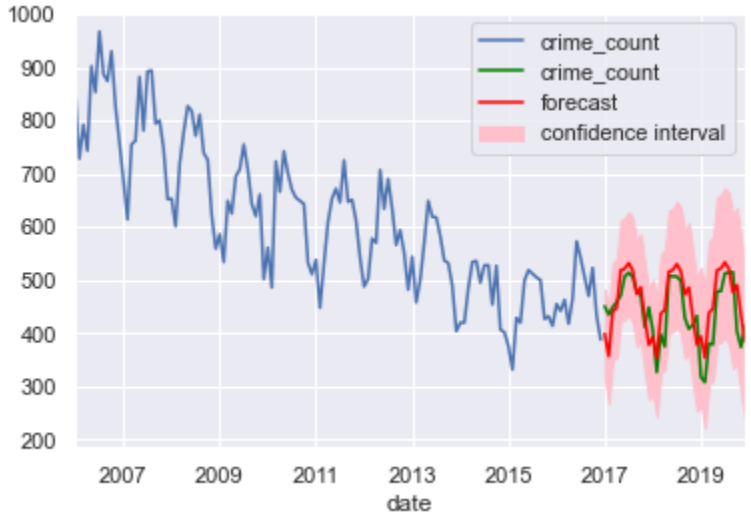
Plots below show the prediction results using auto-arima for the construction and crime count and the real estate price forecast based on crime and construction count predictions for zip codes 60608 and zip code 60612.

Zip code 60608
Construction count prediction

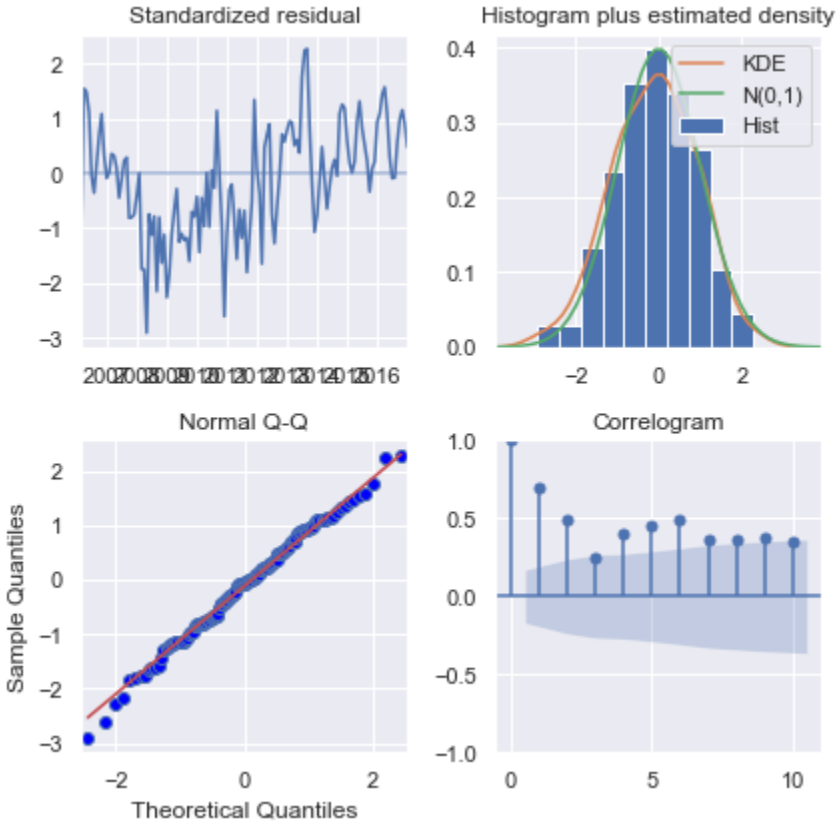


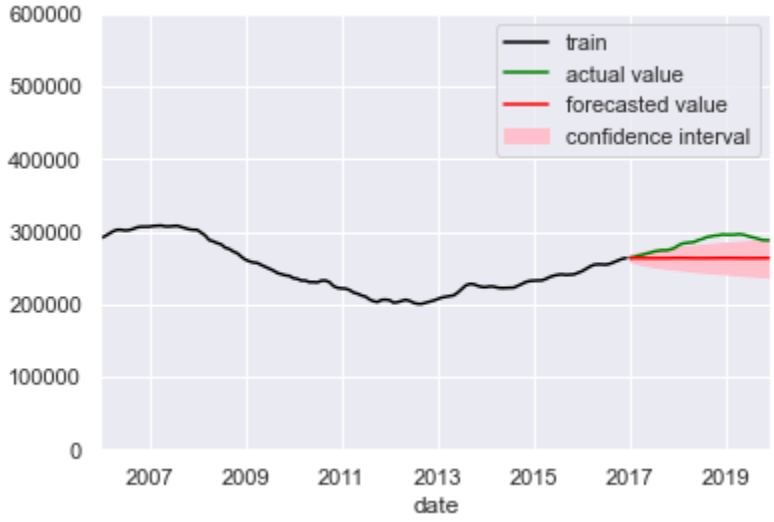
Zip code 60608
Crime prediction





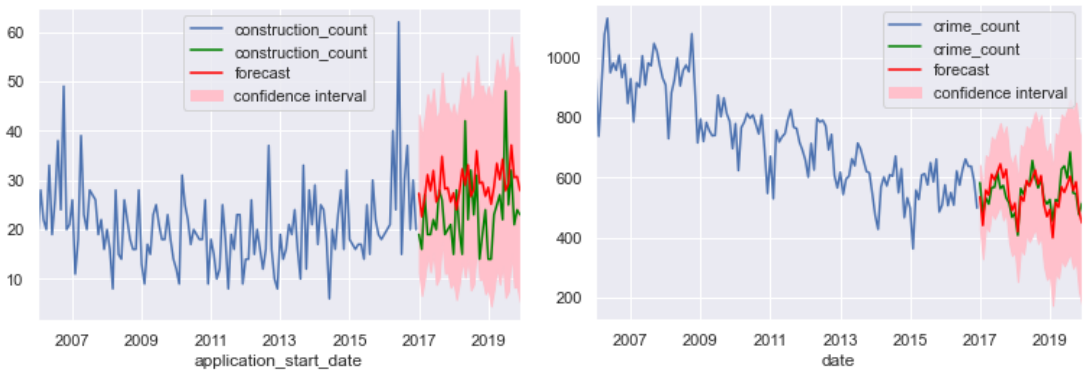
Zip code 60608
real estate price prediction



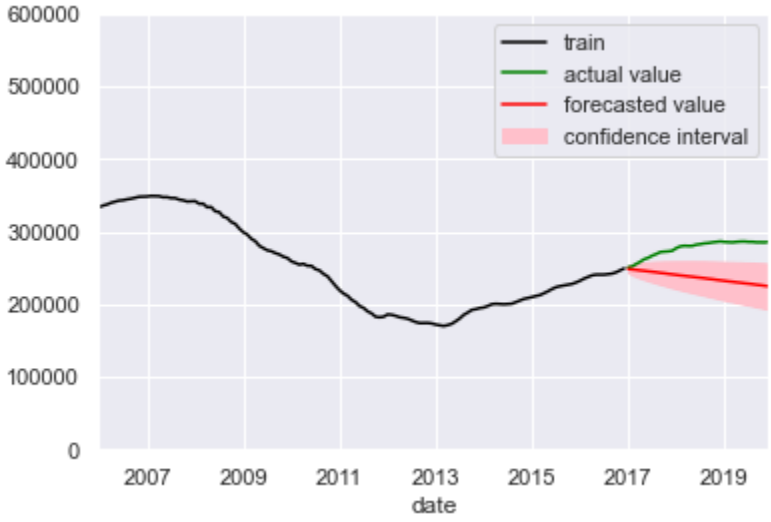


date	forecast	actual_price
2019-12-31	263631.32876	288059.0

Zip code 60612
Construction and crime count predictions



Zip code 60612
real estate price prediction



	forecast	actual_price
date		
2019-12-31	224710.572487	285758.0

6. Results

To compare the performance of the two models, I found the mean absolute error and root mean squared error values of the models for 40 zip codes.

	RMSE	MAE	R squared

Model 1	45169	37834	0.87
Model 2	37290	30369	0.91

The results show that model 2 (ARIMA model based on multivariate time series) is more successful in forecasting the housing prices. Model 2 takes much longer to run because it needs to predict the crime and construction count to be able to predict the housing prices, but since we only need to run it once every few months this will not be an issue. Home buyer can decide based on the model results 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

7. Future Improvements

Trying different exogenous variables such as population, GDP might be helpful.