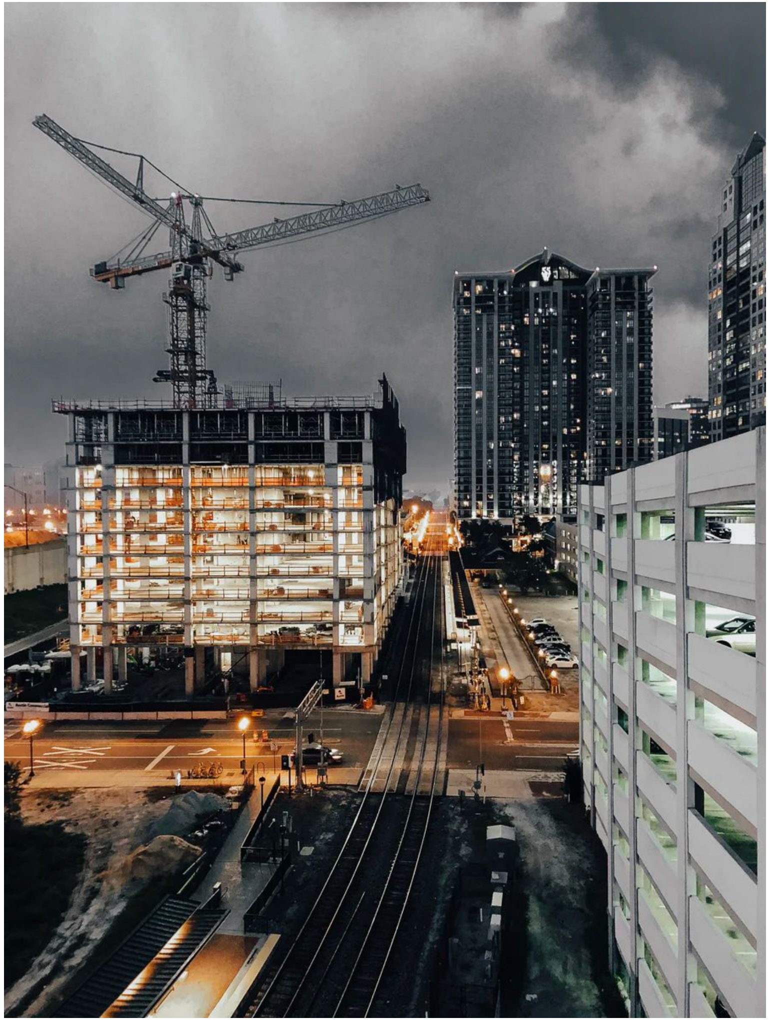
Predicting Median Housing Prices for Chicago Zip Codes Three Years in Advance

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.



Data Sources

- Some factors that can affect housing prices in neighborhood are crime rate and increase newly built and renovated buildings.

- building permits
  - City of Chicago Data Portal

- crime rate
  - City of Chicago Website

- historical housing prices
  - Zillow

Goal

- Build a model that predicts housing prices three years in advance for each zip code.

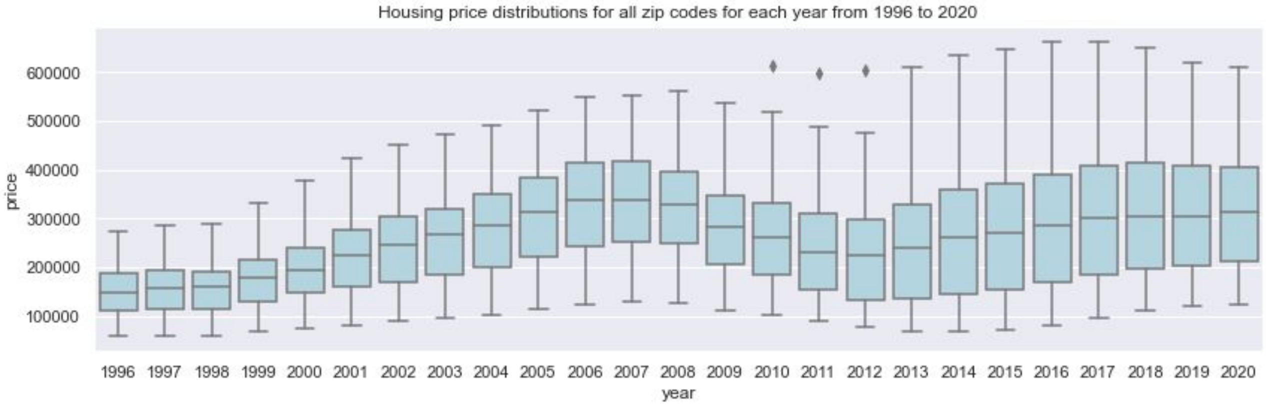
  Approach 1:
  - Univariate time series forecasting using ARIMA. Used historical housing prices since 2006

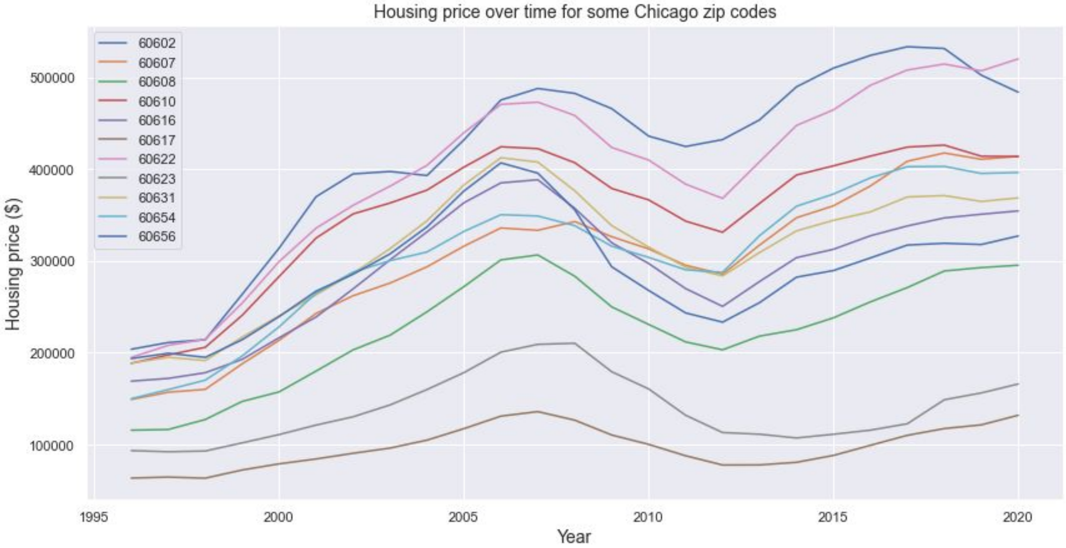
  Approach 2:
  - multivariate time series forecasting using arima models. Used Crime rate and construction count as exogenous variables

Housing Price Over Time

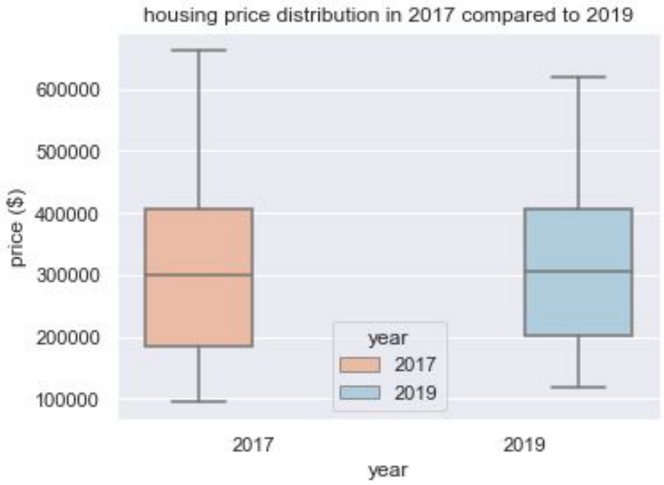
- The gap between the least and most expensive housing prices has increased over time.

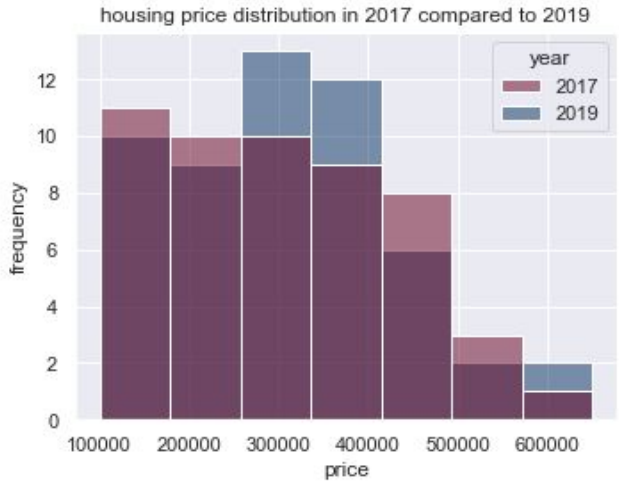
- Some zip codes has seen increase in real estate value while some lost value over time.

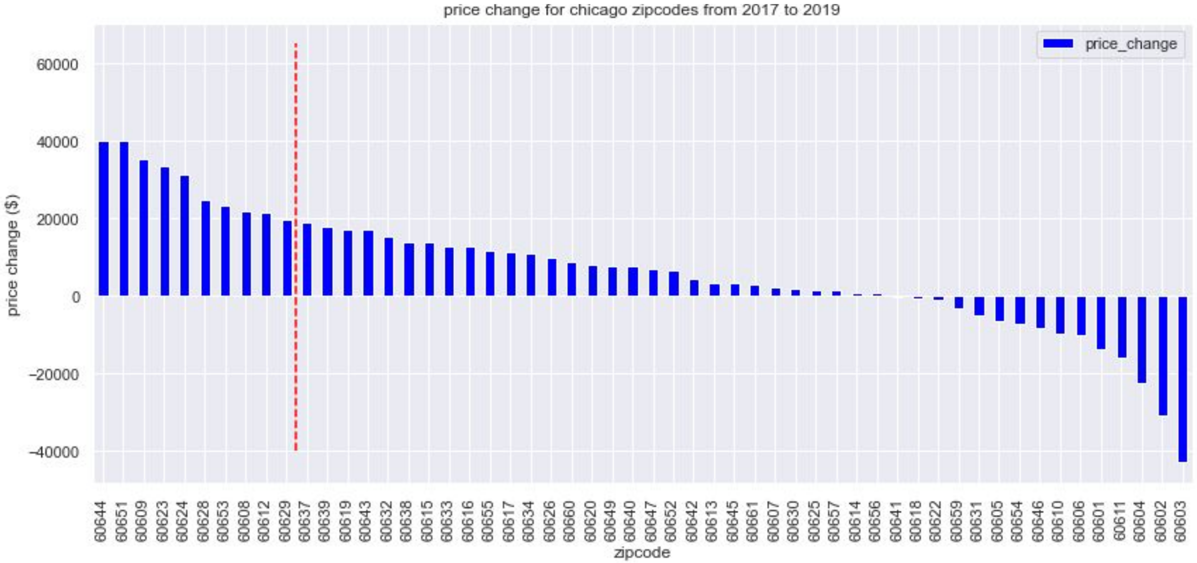




Housing Price Over Time



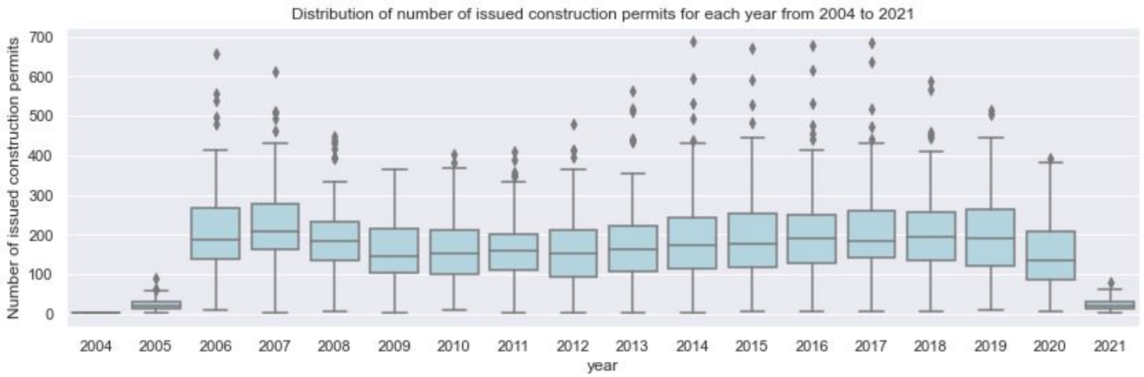


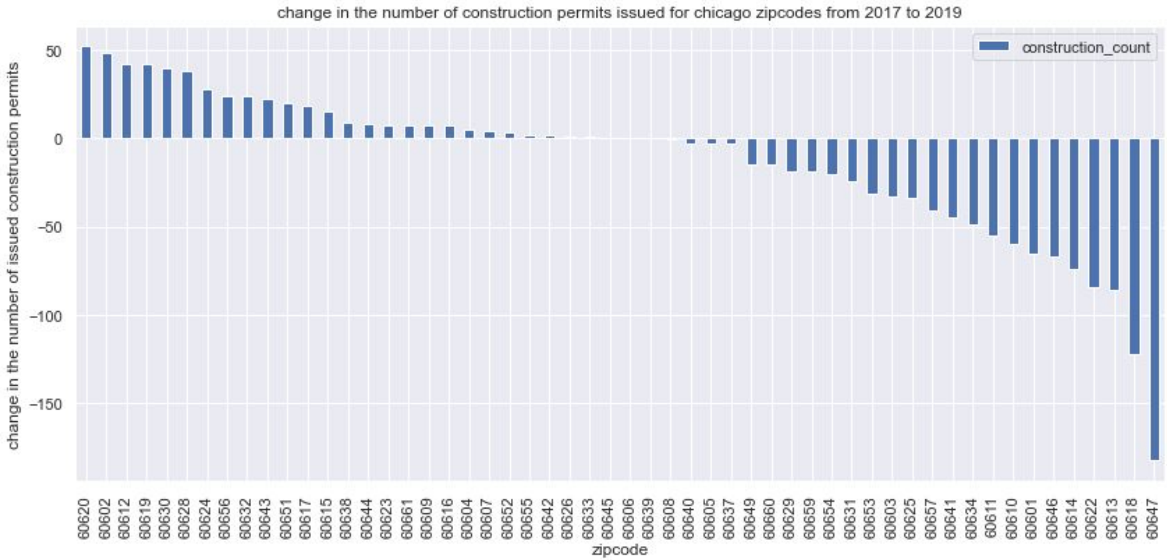


Construction and Renovation Permits Data

- 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.



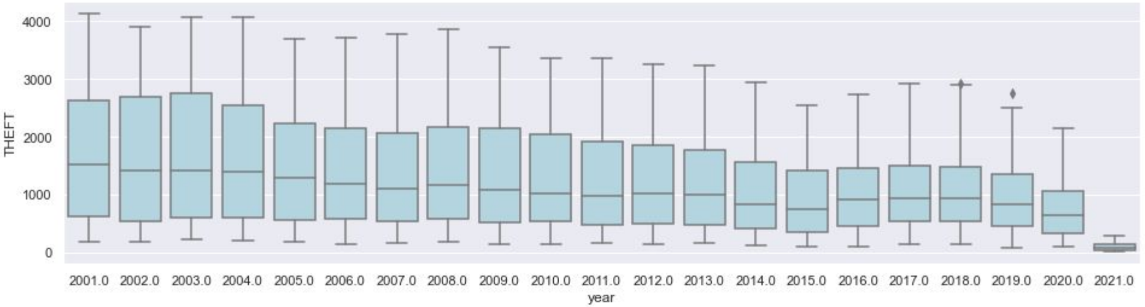


Crime

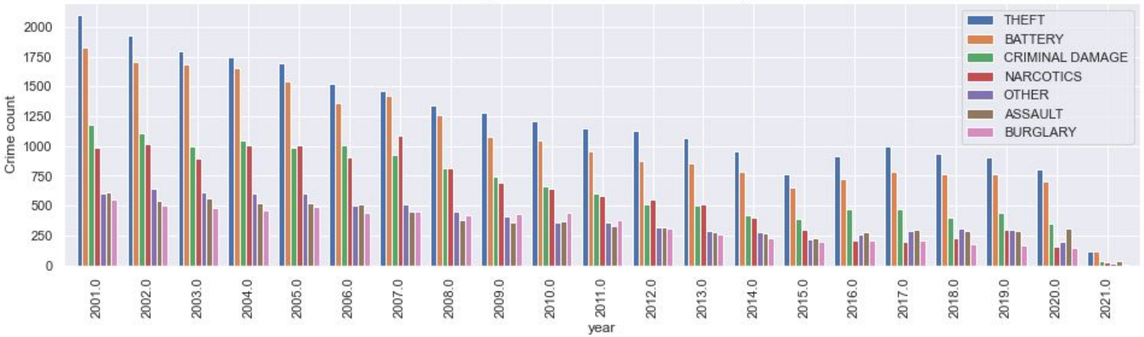
- Crime rate has decreased in Chicago over time.

- 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.

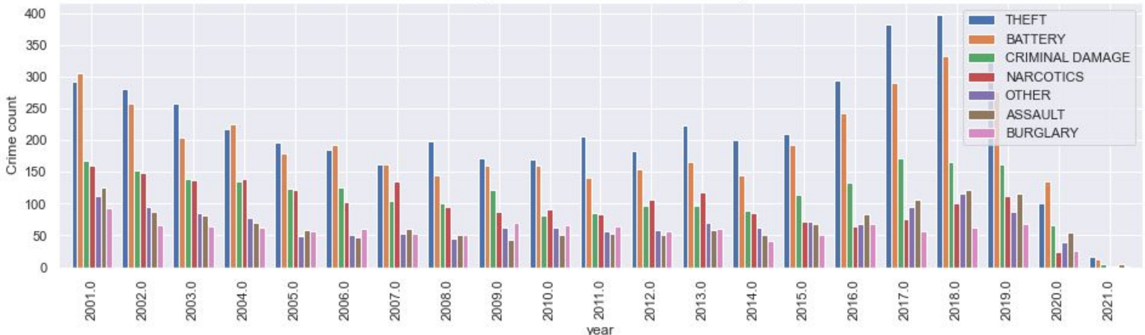
- Zip code 60603 has seen a decrease in housing value over the last 5 years. Crime rate has increased in this zip code over this time period.



Different type of crime count over time in 60653 zipcode



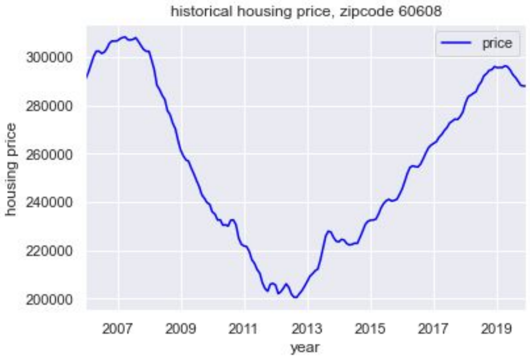
Different type of crime count over time in 60603 zipcode

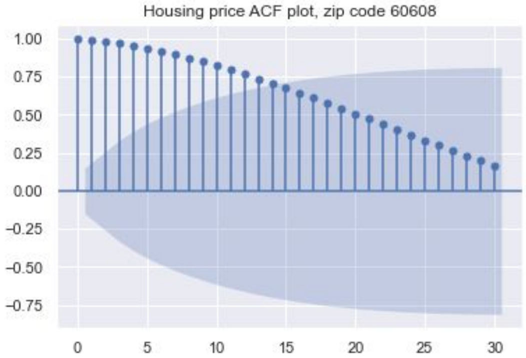


Checking for Stationarity

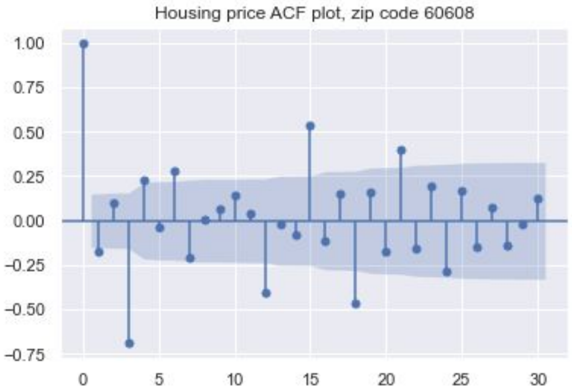
- Estimate the number of differences required to make a given time series for a zip code stationary.

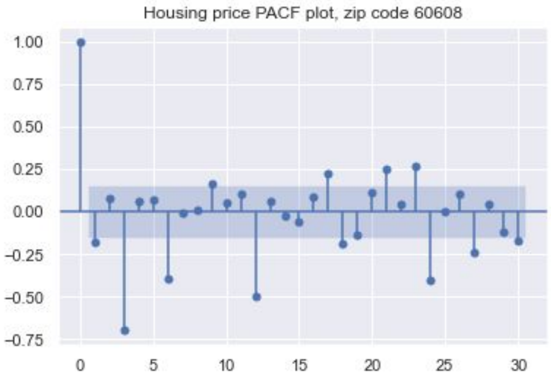
- Used max diff term recommended by python ndiffs function with
  - Kwiatkowski–Phillips–Schmidt–Shin (KPSS) tests
  - Dickey–Fuller test





- ACF and PACF plots for 60608 zip code after two times differencing





Auto-ARIMA

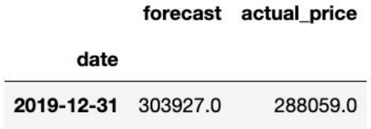
- Used Auto-ARIMA to find the best model for each zip code instead of manually finding the model parameters.

- Model parameters
  - **p**: number of lagged observation considered in the model
  - **d**: degree of differencing
  - **q**: the order of moving average

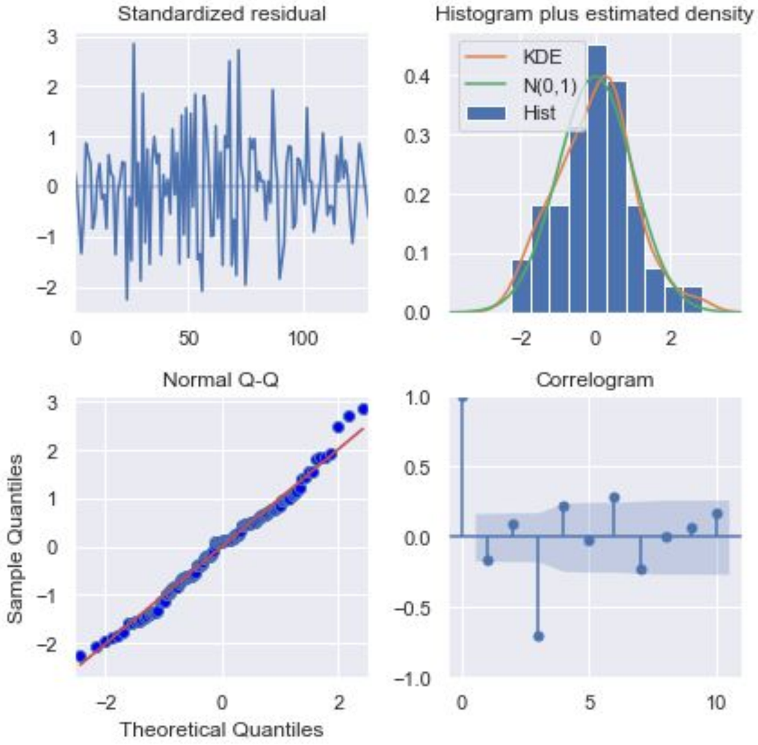
- Auto-ARIMA steps:
  - Conducts differencing tests (i.e., Kwiatkowski–Phillips–Schmidt–Shin, Augmented Dickey-Fuller or Phillips–Perron) to determine the order of differencing, d,
  - Fits models within ranges of defined start_p, max_p, start_q, max_q ranges.
  - Finds the best model by that provides the least Akaike Information Criterion value

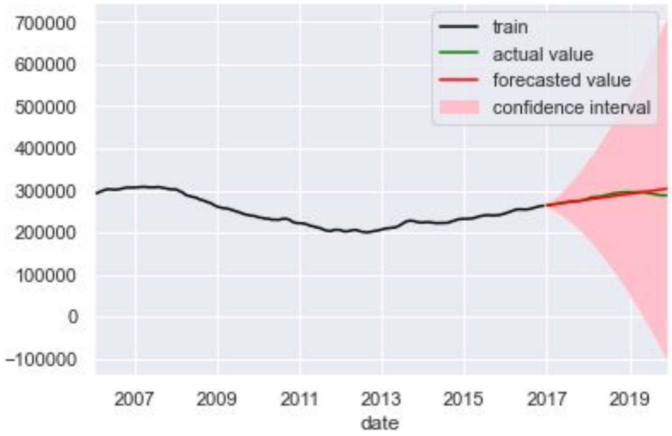
Model 1 - Housing Price Prediction

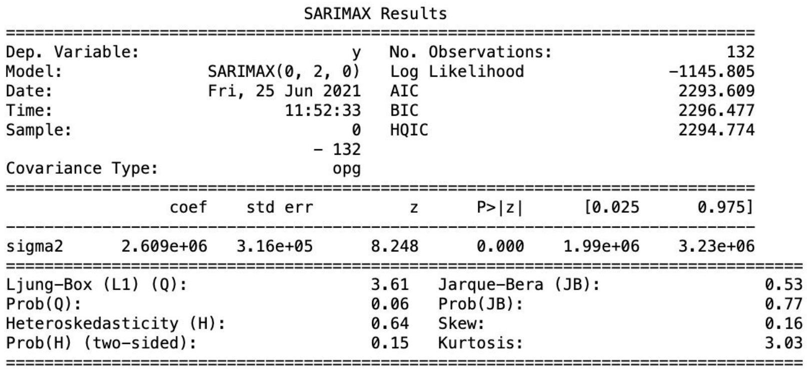
Zip Code 60608



	forecast	actual_price
date		
2019-12-31	303927.0	288059.0



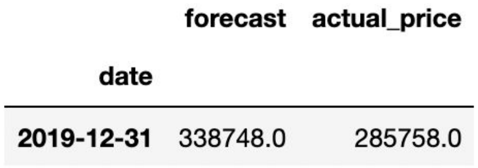




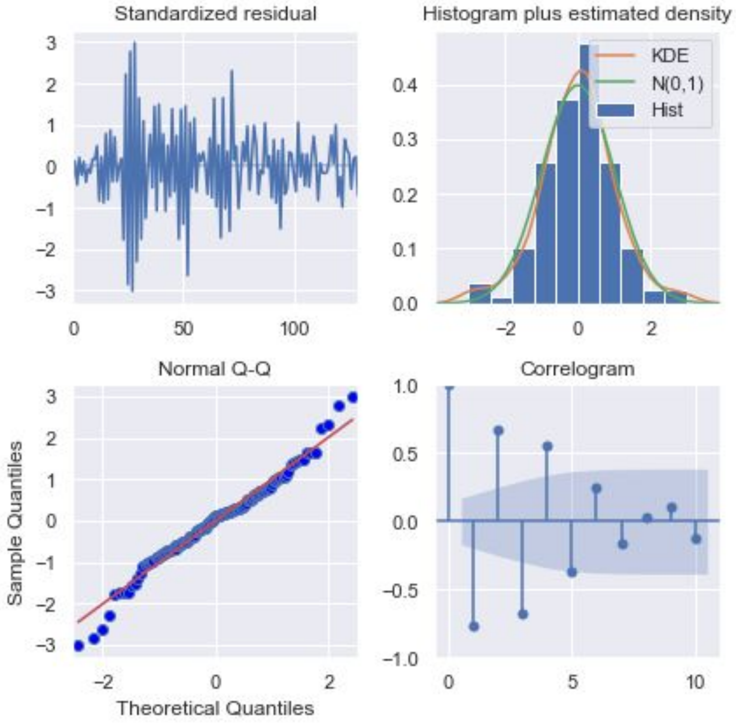
```
                               SARIMAX Results
==============================================================================
Dep. Variable:                      y   No. Observations:                  132
Model:               SARIMAX(0, 2, 0)   Log Likelihood               -1145.805
Date:                Fri, 25 Jun 2021   AIC                           2293.609
Time:                        11:52:33   BIC                           2296.477
Sample:                             0   HQIC                          2294.774
                                - 132
Covariance Type:                  opg
==============================================================================
                 coef    std err          z      P>|z|      [0.025      0.975]
------------------------------------------------------------------------------
sigma2      2.609e+06   3.16e+05      8.248      0.000    1.99e+06    3.23e+06
===================================================================================
Ljung-Box (L1) (Q):                   3.61   Jarque-Bera (JB):                 0.53
Prob(Q):                              0.06   Prob(JB):                         0.77
Heteroskedasticity (H):               0.64   Skew:                             0.16
Prob(H) (two-sided):                  0.15   Kurtosis:                         3.03
===================================================================================
```

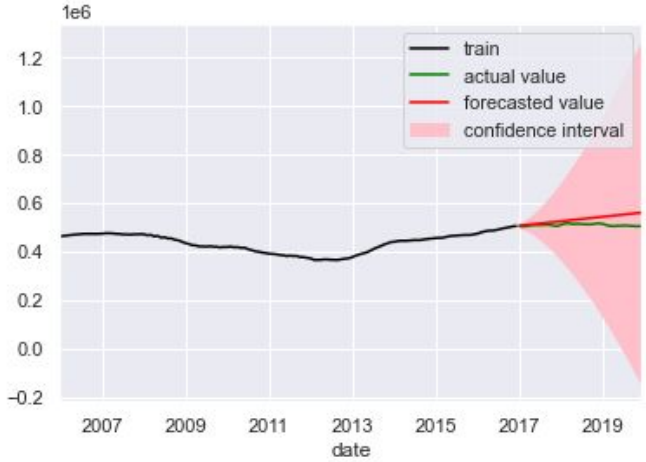
Model 1 - Housing Price Prediction

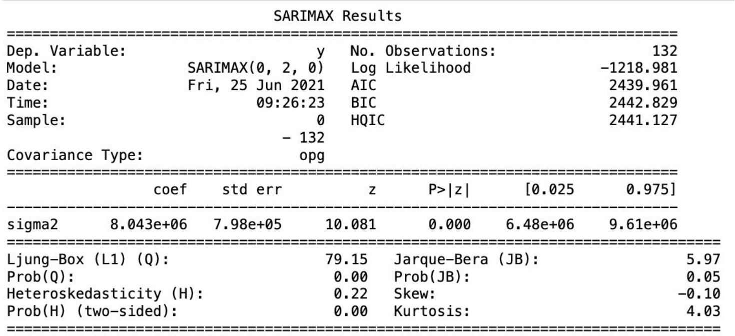
Zip Code 60612



	forecast	actual_price
date		
2019-12-31	338748.0	285758.0



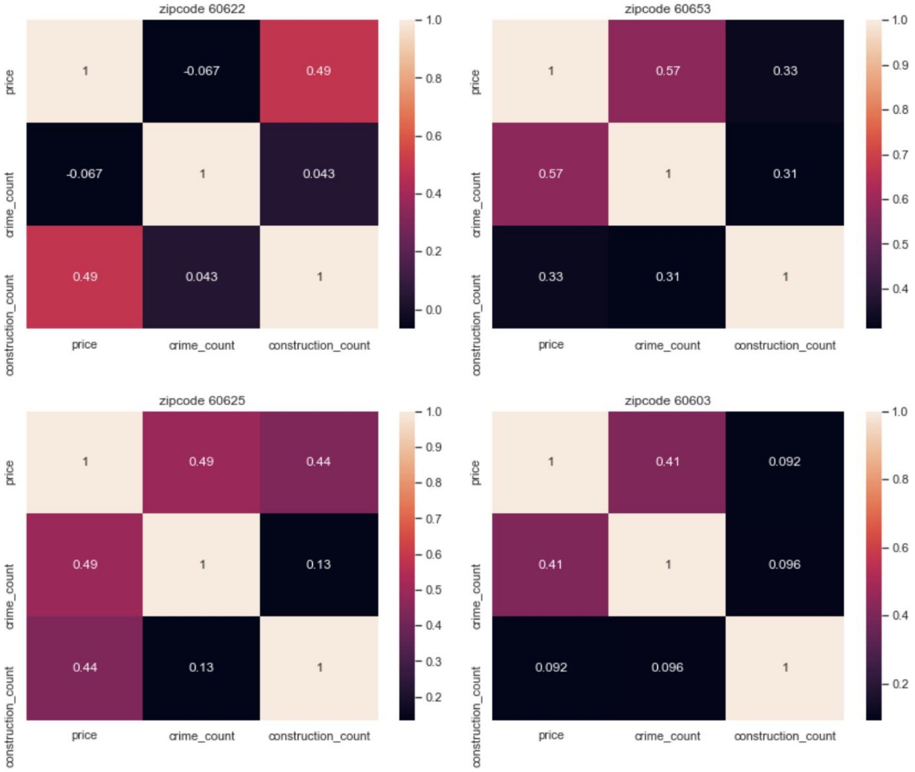




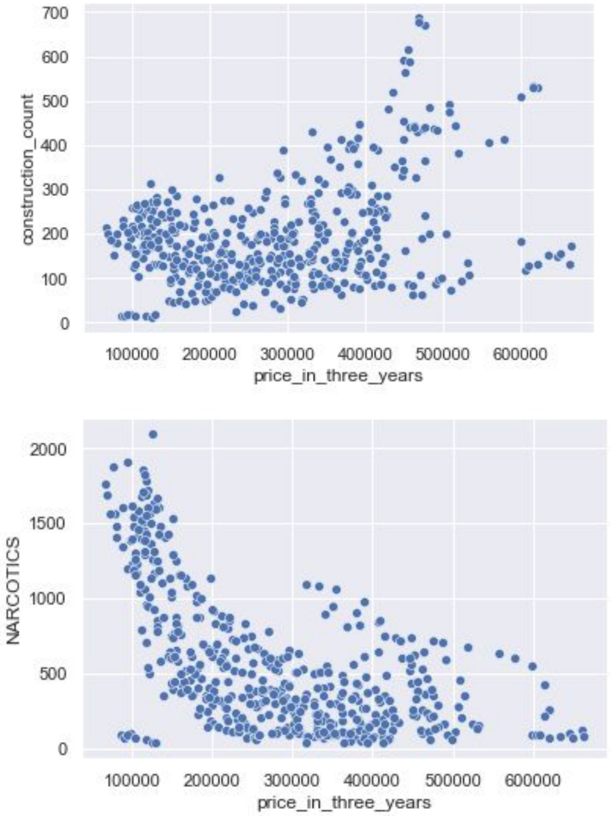
```
                                     SARIMAX Results
==============================================================================
Dep. Variable:                      y   No. Observations:                  132
Model:               SARIMAX(0, 2, 0)   Log Likelihood               -1218.981
Date:                Fri, 25 Jun 2021   AIC                           2439.961
Time:                        09:26:23   BIC                           2442.829
Sample:                             0   HQIC                          2441.127
                                - 132
Covariance Type:                  opg
==============================================================================
                 coef    std err          z      P>|z|      [0.025      0.975]
------------------------------------------------------------------------------
sigma2      8.043e+06   7.98e+05     10.081      0.000    6.48e+06    9.61e+06
===================================================================================
Ljung-Box (L1) (Q):                  79.15   Jarque-Bera (JB):                 5.97
Prob(Q):                              0.00   Prob(JB):                         0.05
Heteroskedasticity (H):               0.22   Skew:                            -0.10
Prob(H) (two-sided):                  0.00   Kurtosis:                         4.03
===================================================================================
```

Crime and Construction Count Analysis

- Correlation plot for various zip codes

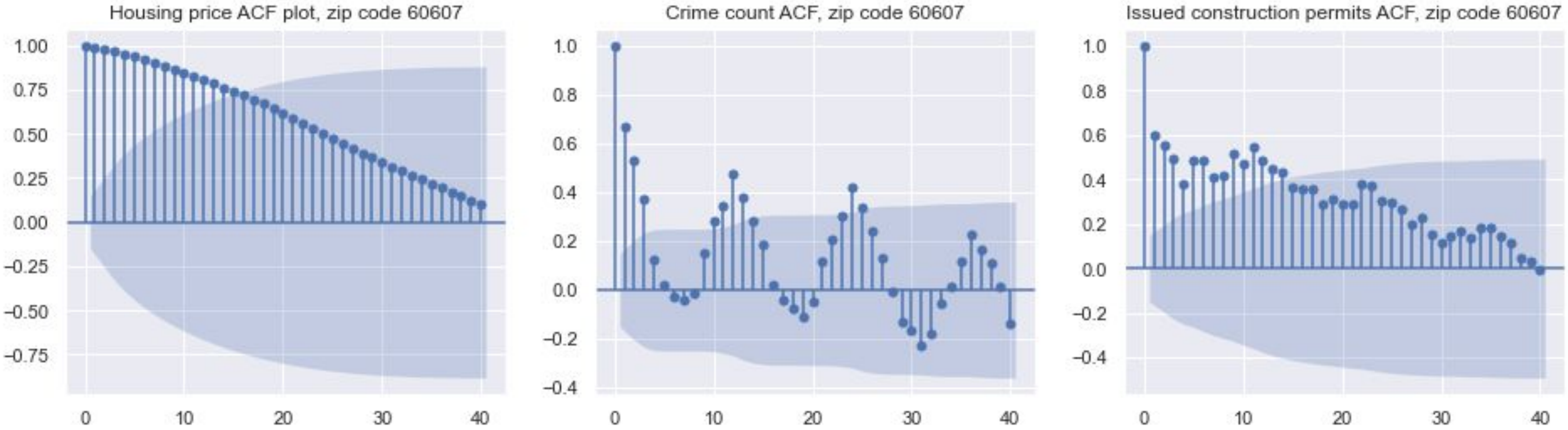


- Scatter plot showing price vs. construction count and price vs. narcotics relationships



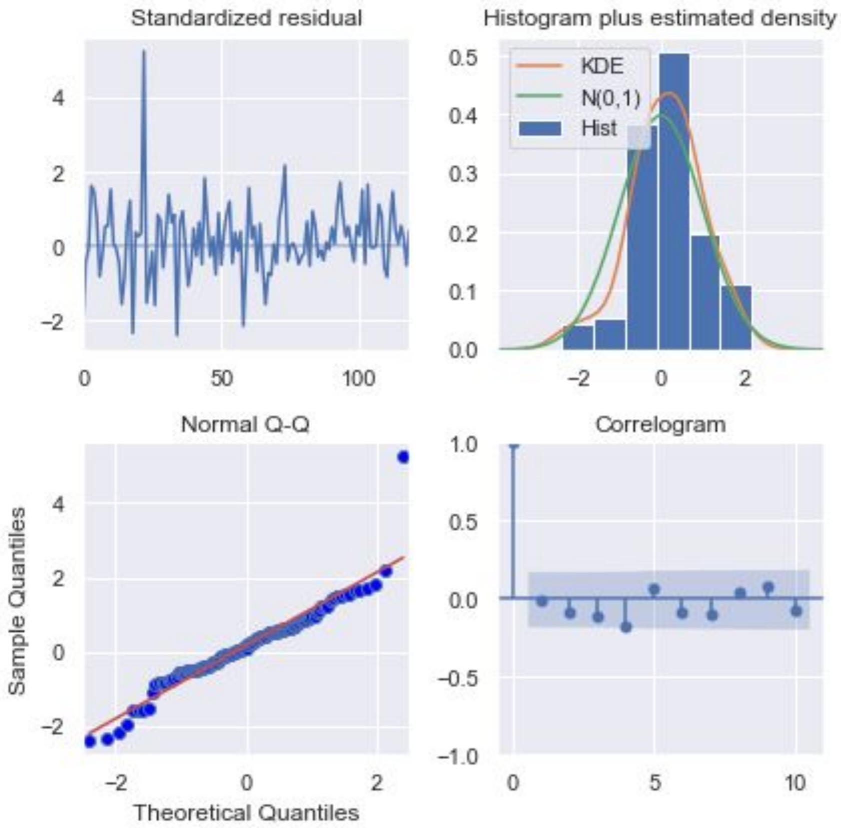
Autocorrelation Function Plots

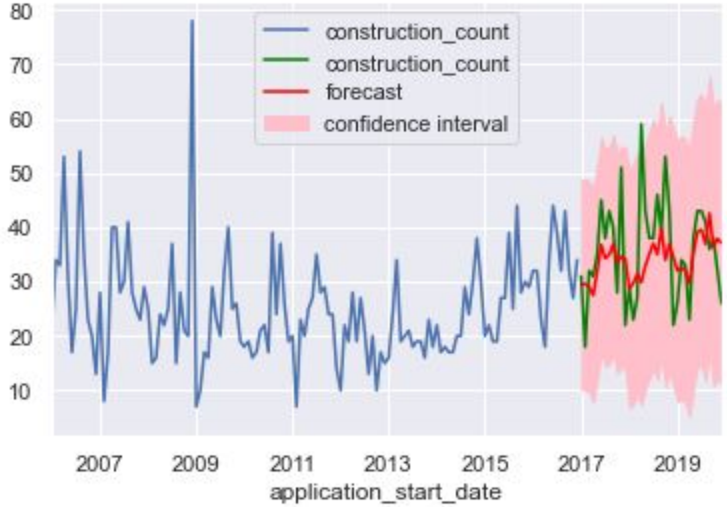
- ACF Plots for housing price, crime, and construction count

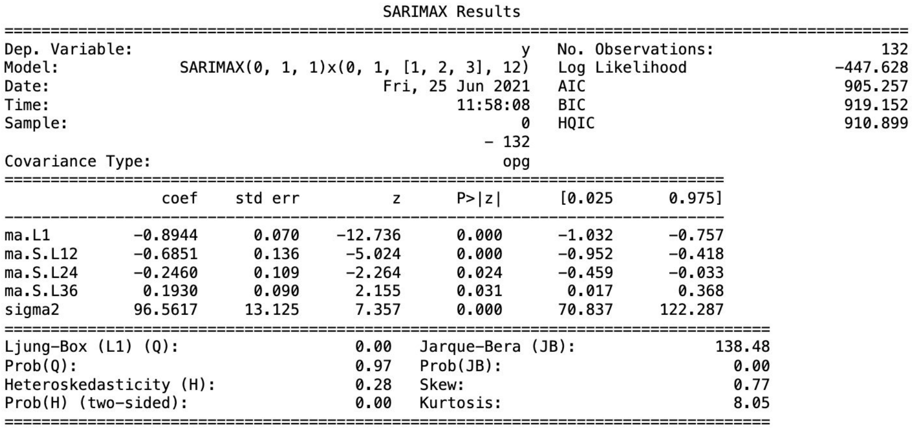


Building Construction Count Prediction

Zip code 60608



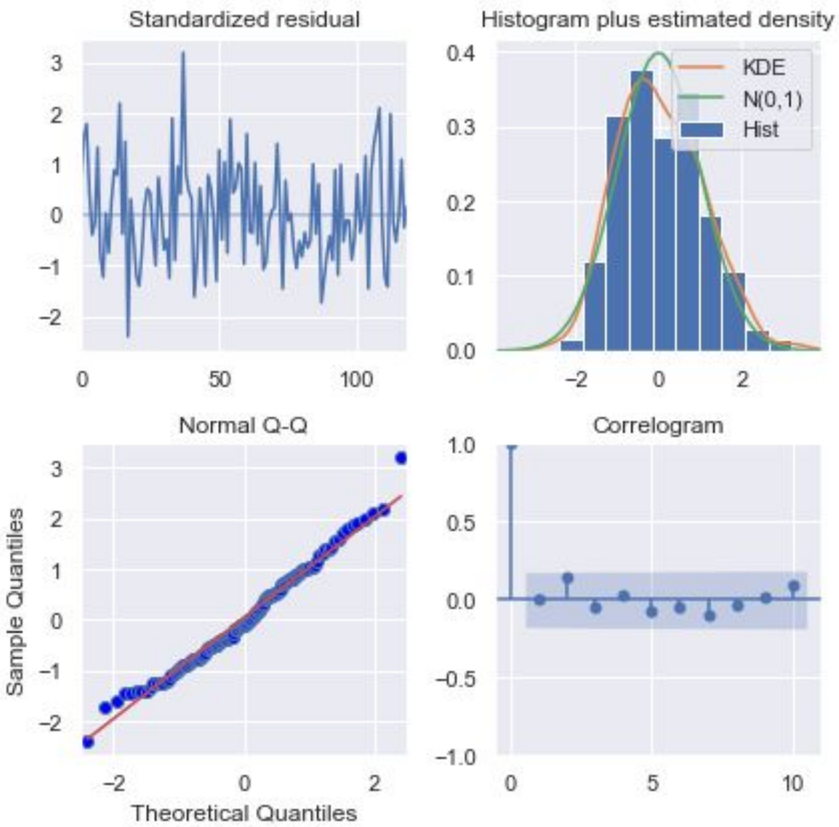


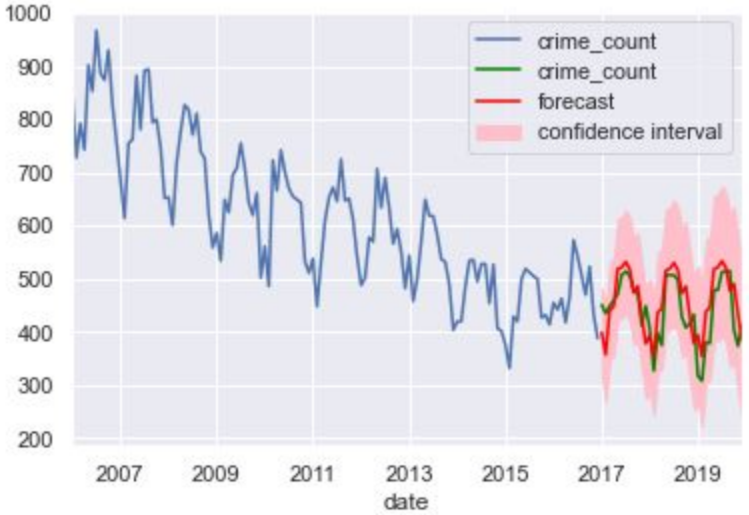


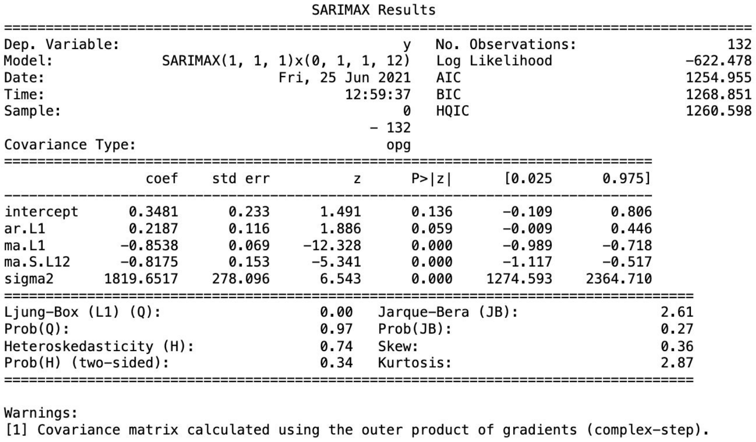
```
                                     SARIMAX Results
==========================================================================================
Dep. Variable:                                 y   No. Observations:                  132
Model:             SARIMAX(0, 1, 1)x(0, 1, [1, 2, 3], 12)   Log Likelihood             -447.628
Date:                           Fri, 25 Jun 2021   AIC                            905.257
Time:                                   11:58:08   BIC                            919.152
Sample:                                        0   HQIC                           910.899
                                           - 132
Covariance Type:                             opg
==============================================================================
                 coef    std err          z      P>|z|      [0.025      0.975]
------------------------------------------------------------------------------
ma.L1         -0.8944      0.070    -12.736      0.000      -1.032      -0.757
ma.S.L12      -0.6851      0.136     -5.024      0.000      -0.952      -0.418
ma.S.L24      -0.2460      0.109     -2.264      0.024      -0.459      -0.033
ma.S.L36       0.1930      0.090      2.155      0.031       0.017       0.368
sigma2        96.5617     13.125      7.357      0.000      70.837     122.287
===================================================================================
Ljung-Box (L1) (Q):                   0.00   Jarque-Bera (JB):               138.48
Prob(Q):                              0.97   Prob(JB):                         0.00
Heteroskedasticity (H):               0.28   Skew:                             0.77
Prob(H) (two-sided):                  0.00   Kurtosis:                         8.05
===================================================================================
```

Crime Prediction

Zip code 60608





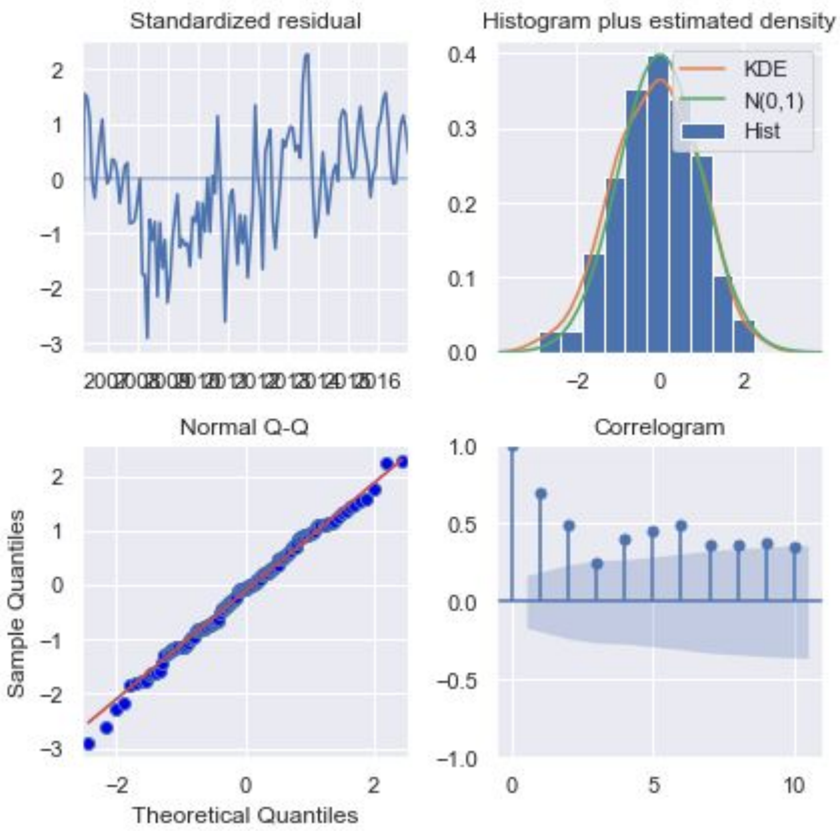


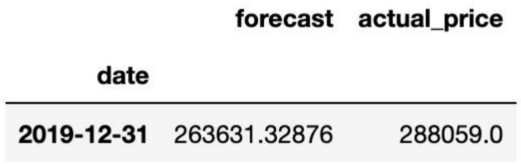
```
                                     SARIMAX Results
==========================================================================================
Dep. Variable:                                  y   No. Observations:                  132
Model:             SARIMAX(1, 1, 1)x(0, 1, 1, 12)   Log Likelihood                -622.478
Date:                            Fri, 25 Jun 2021   AIC                           1254.955
Time:                                    12:59:37   BIC                           1268.851
Sample:                                         0   HQIC                          1260.598
                                            - 132
Covariance Type:                              opg
==============================================================================
                 coef    std err          z      P>|z|      [0.025      0.975]
------------------------------------------------------------------------------
intercept      0.3481      0.233      1.491      0.136      -0.109       0.806
ar.L1          0.2187      0.116      1.886      0.059      -0.009       0.446
ma.L1         -0.8538      0.069    -12.328      0.000      -0.989      -0.718
ma.S.L12      -0.8175      0.153     -5.341      0.000      -1.117      -0.517
sigma2      1819.6517    278.096      6.543      0.000    1274.593    2364.710
===================================================================================
Ljung-Box (L1) (Q):                   0.00   Jarque-Bera (JB):                 2.61
Prob(Q):                              0.97   Prob(JB):                         0.27
Heteroskedasticity (H):               0.74   Skew:                             0.36
Prob(H) (two-sided):                  0.34   Kurtosis:                         2.87
===================================================================================

Warnings:
[1] Covariance matrix calculated using the outer product of gradients (complex-step).
```

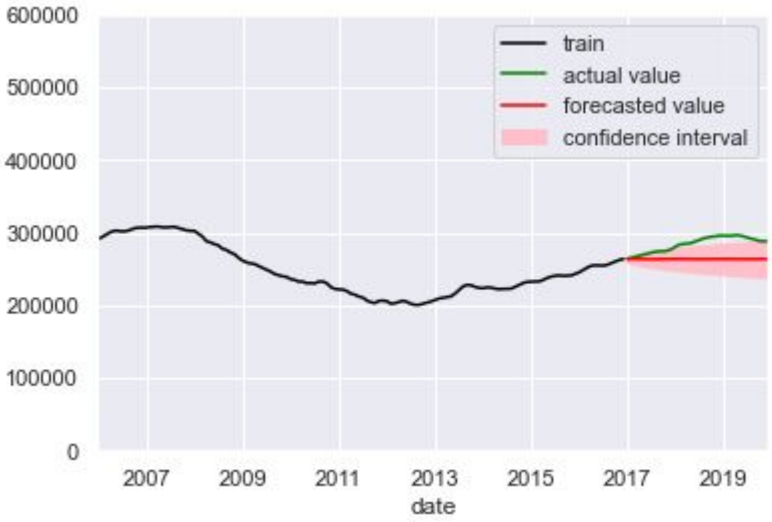
Model 2 - Housing Price Prediction

Zip code 60608



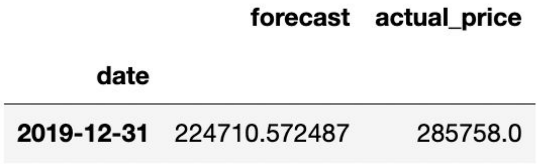


	forecast	actual_price
date		
2019-12-31	263631.32876	288059.0

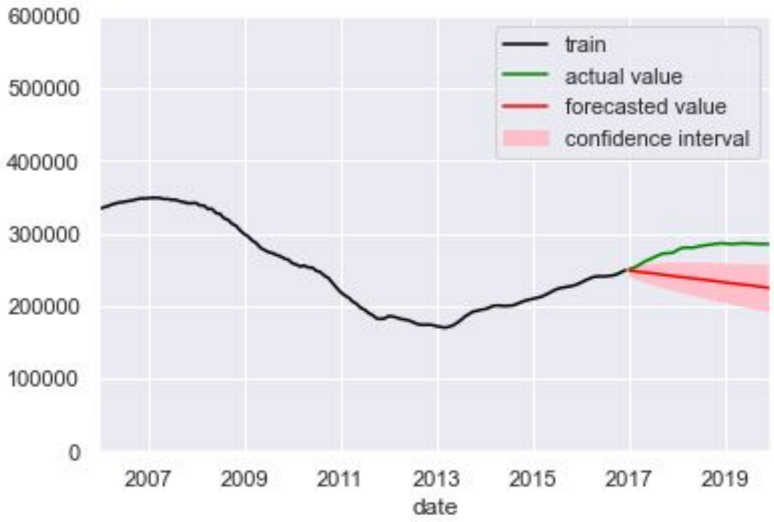


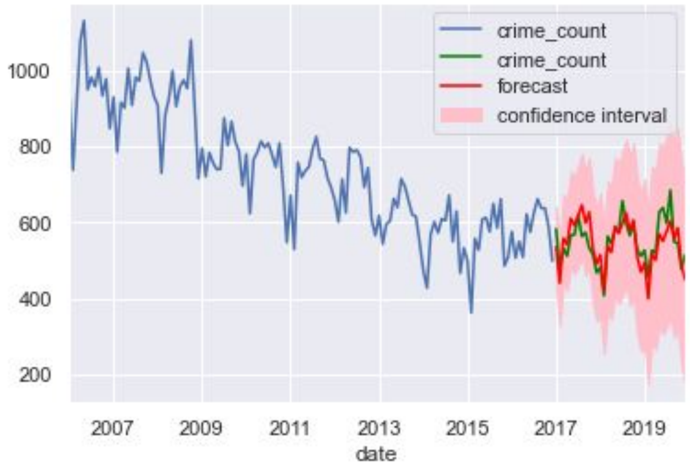
Model 2 - Housing Price Prediction

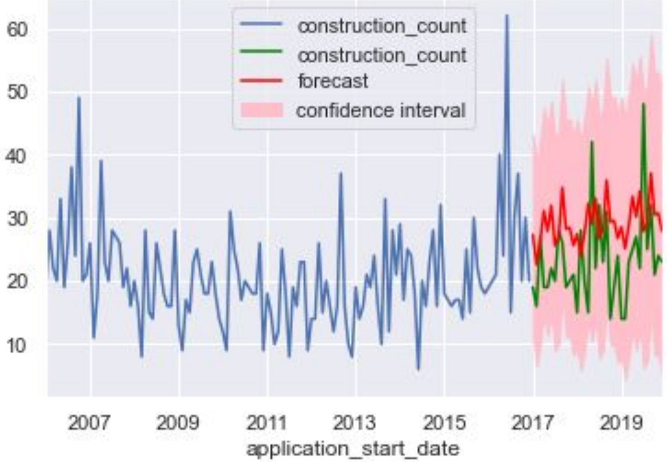
Zip code 60612



date	forecast	actual_price
2019-12-31	224710.572487	285758.0







Comparing model 2 and model 1 performances

Models performance for all zip codes:

	RMSE	MAE	squared
Model 1	45169	37834	0.87
Model 2	37290	30369	0.91

Model 2 has a better performance compare to model 1

Results

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

http://www.aag.com/

Improvements

- Trying different exogenous variables :
    - Zip code specific data such as demographics and population over time
    - Adding other type of data (economy, GDP...)