carlearn / dsc-phase-4-project Public

forked from learn-co-curriculum/dsc-phase-4-project

0 stars 86 forks

Star Watch

Code Pull requests Actions Projects Wiki Security Insights

main

This branch is 9 commits ahead of learn-co-curriculum/dsc-phase-4-project:main.

Contribute Fetch upstream

carlearn updated README with charts … 37 seconds ago 19

View code

README.md

# Module 4 Final Project

## Overview

Our client, Andersen Investco is a boutique real-estate investment firm based in New York City. They want to have a better understanding of the trends in housing market, and to look for advice on what are the top 5 zipcodes in the state of New York.

Our team is hired to perform a time series analysis using Zillow's historical housing data for the United States. The aim was to provide investors with the best zip codes to buy and develop homes in the state of Texas.

## Objectives

- Select the top 5 zipcodes in NY to invest

- Perform time series modeling with ARIMA to predict the return of investment in 1, 3, 5, 10 year period

# Project Summary

## Exploratory Data Analysis - Zipcode Selection

We followed the preferences of our client and take the five criteria into consideration:

- State: New York State (the home state of our client)
- Urbanization: Zipcode should be in the top 25% according to the SizeRank variable
- Typical Home Value: leverage the ZHVI which reflects the typical value for homes in the 35th - 65th percentile range
- Growth Rate: Considering that our client will invest in mid 2022, the situation will be similar to the post-financial crisis. Therefore, we will select pre-pandemics & post-financial crisis (08/2009 - 02/2020) as our benchmark to measure the growth rate.
- Diversification: Zicodes should be in different county, and their coefficient of variation below the 75 percentile.

Therefore, the five zipcodes we selected are 11235, 14217, 14611, 11414, 10475.

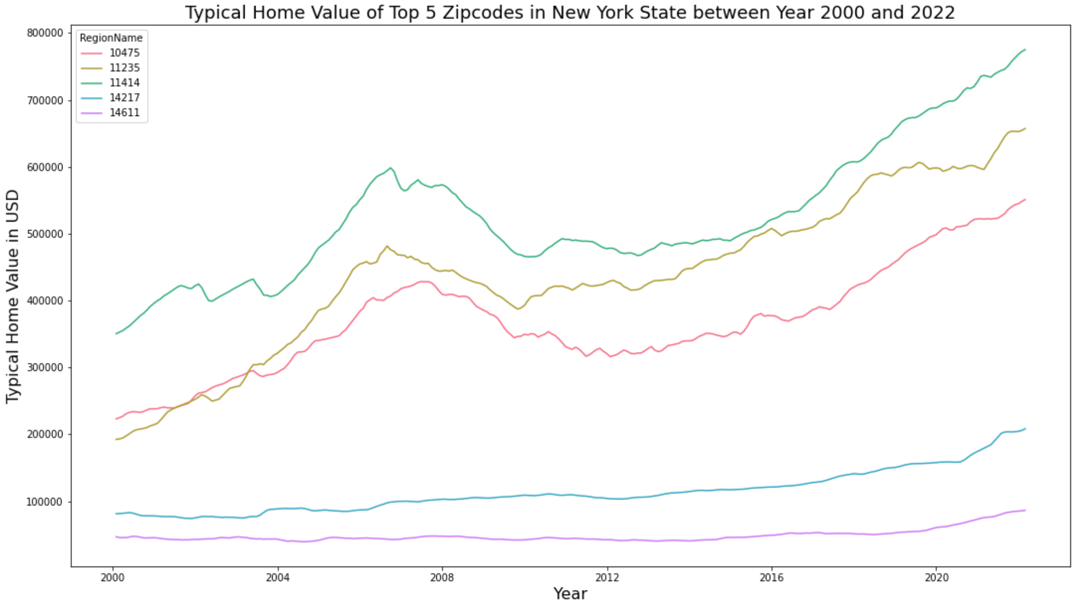
## Reshape Data from Wide to Long Format

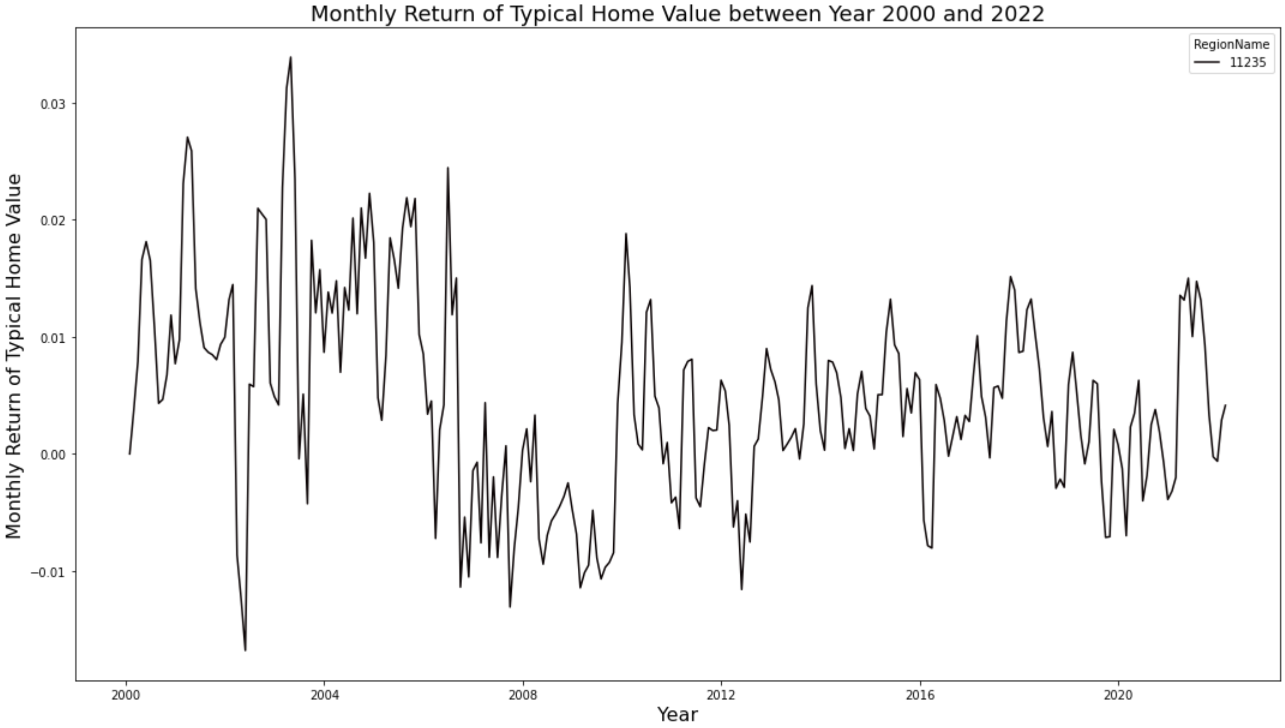
We created a melt_data(df) fuction to transform the data from wide to long format. It is for the purpose of time series modeling.

## Time Series Modeling

### Visualization

We visualized the typical home value of these five zipcodes in the review period (2000-2022). The typical housing prices had a positive trend and the prices are probably not stationary because the next period prices depended on the previous period price. The ultimate goal of investing is to achieve the high ROI. Therefore, we will focus on the monthly returns and build a model for the selected 5 zipcodes to forecast their monthly returns.





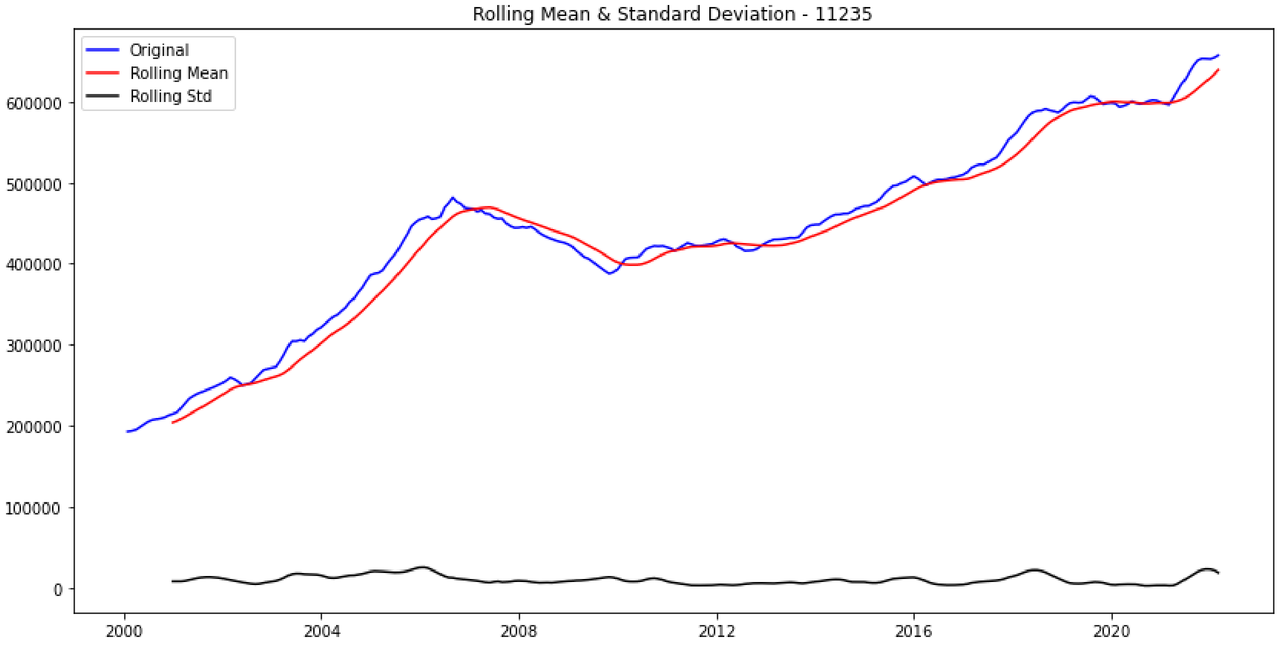
## Stationarity Check

Our approach is to plot the rolling mean and standard deviation against the original data to visually see if there are any trends and by using the Dickey-Fuller Test. The null hypothesis for Dickey Fuller is that the data is not stationary.

For the five zipcodes:

- The rolling mean shows a trend.

- The Dickey-Fuller test cannot reject the null hypothesis of non-stationarity (p-value > 0.05).

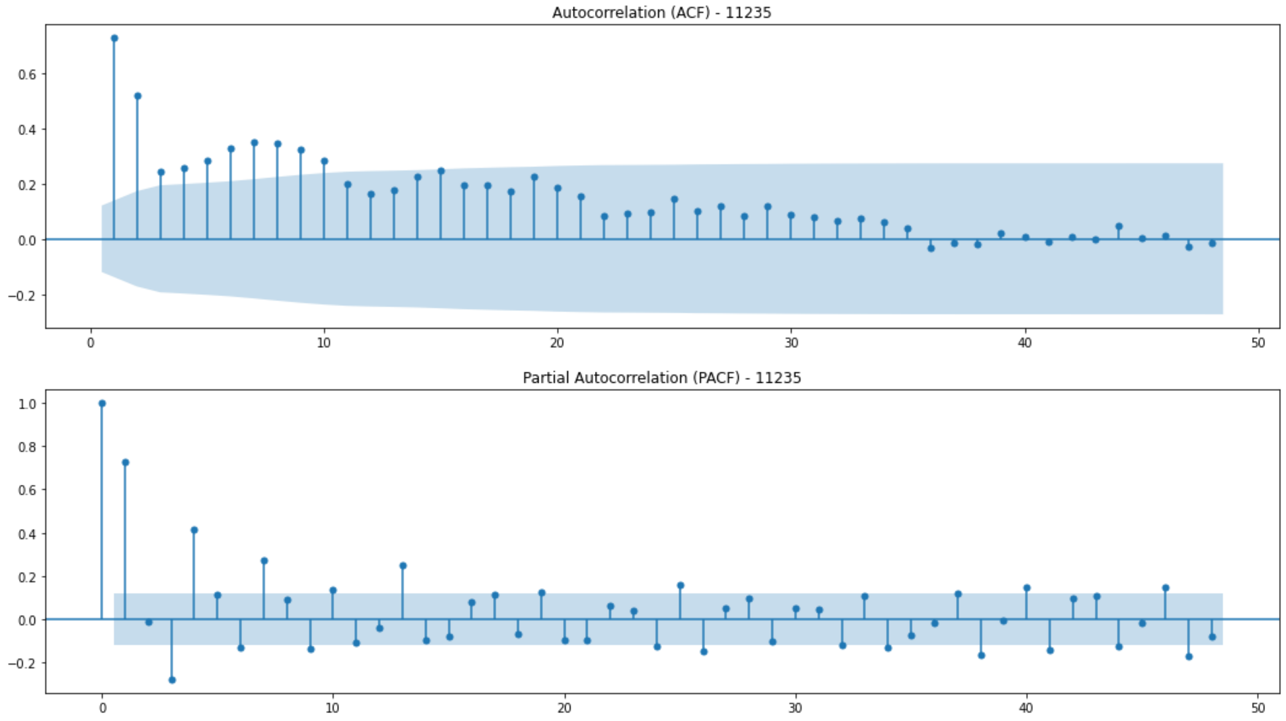


From the results above, two of the five zip codes resulted in non-stationary at a 95 % confidence level. Therefore, these two zipcodes will need the 'I' parameter in the ARIMA model is going to be set to 1, but we will look at the aic result for further determination.

## ARIMA Modeling and Forecasting

### ACF and PACF

The AR and MA parameters can be estimated using the autocorrelation function (ACF) and partial autocorrelation function (PACF) plots of the stationary time series.



Autocorrelation (ACF) - 11235

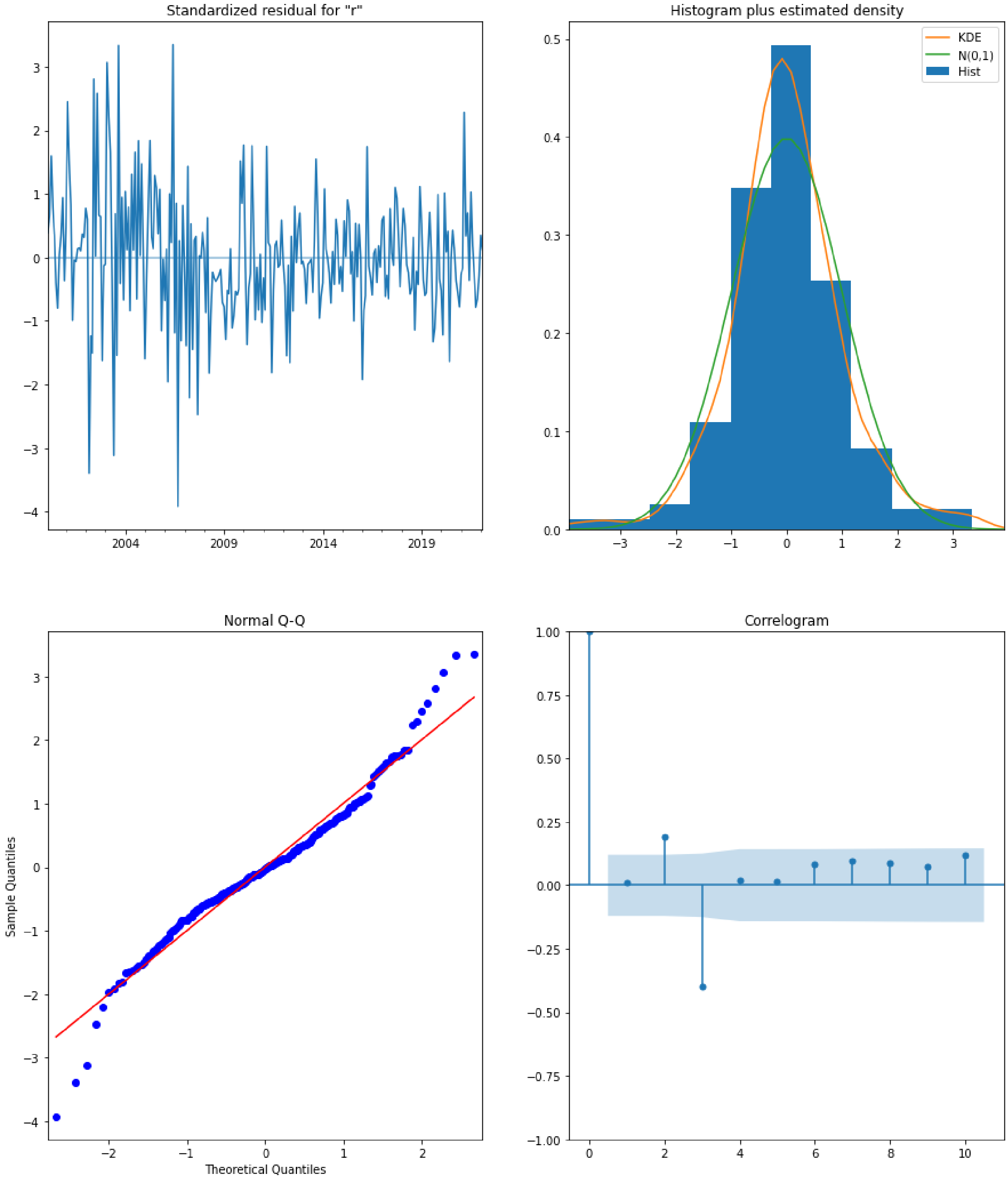
Partial Autocorrelation (PACF) - 11235

### Parameter Selection for the ARIMA Time Series Model

We used the AIC (Akaike Information Criterion) as Regularization Measure. We find the lowest AIC and select the parameters for further prediction analysis.

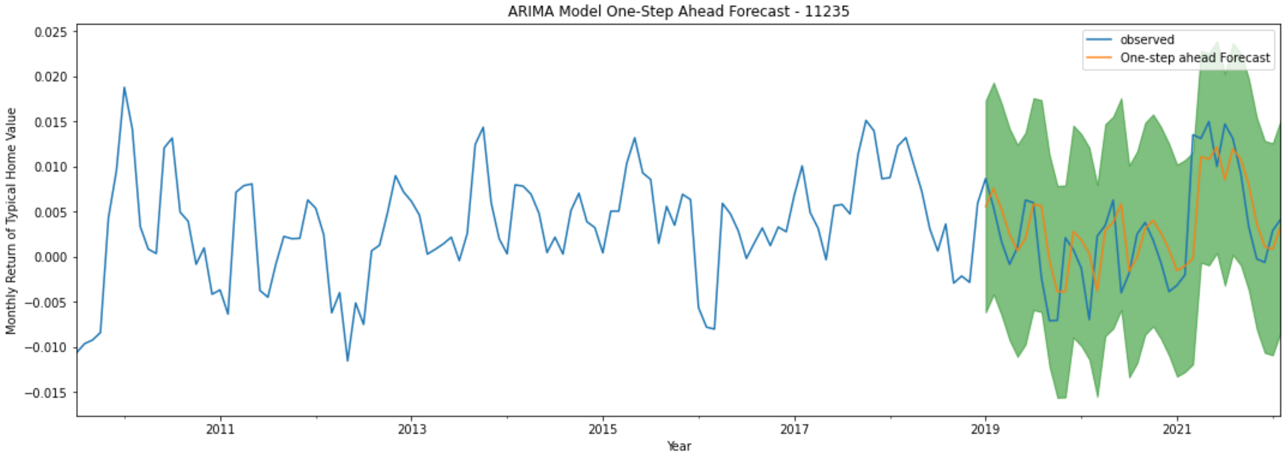
### Fitting an ARIMA Time Series Model:

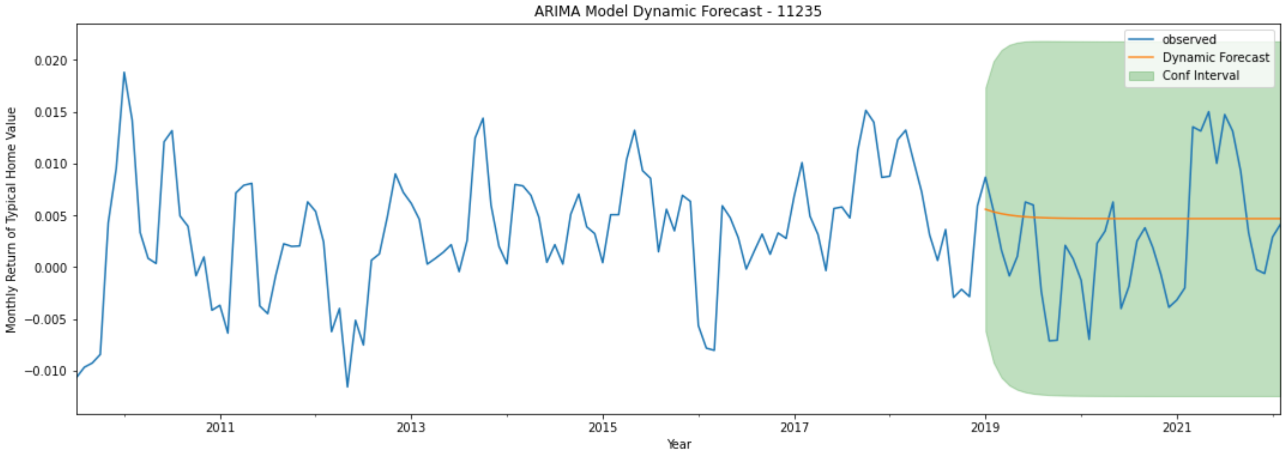
We fit the ARIMA model for each zipcode, whose p-value of the AR and MR parameters is below 0.05 threshold. We fit the model and plot the diagnostics with model summary and plots to ensure that residuals remain uncorrelated, normally distributed having zero mean. In the absence of these assumptions, we can not move forward and need further tweaking of the model.



## Validating the Models

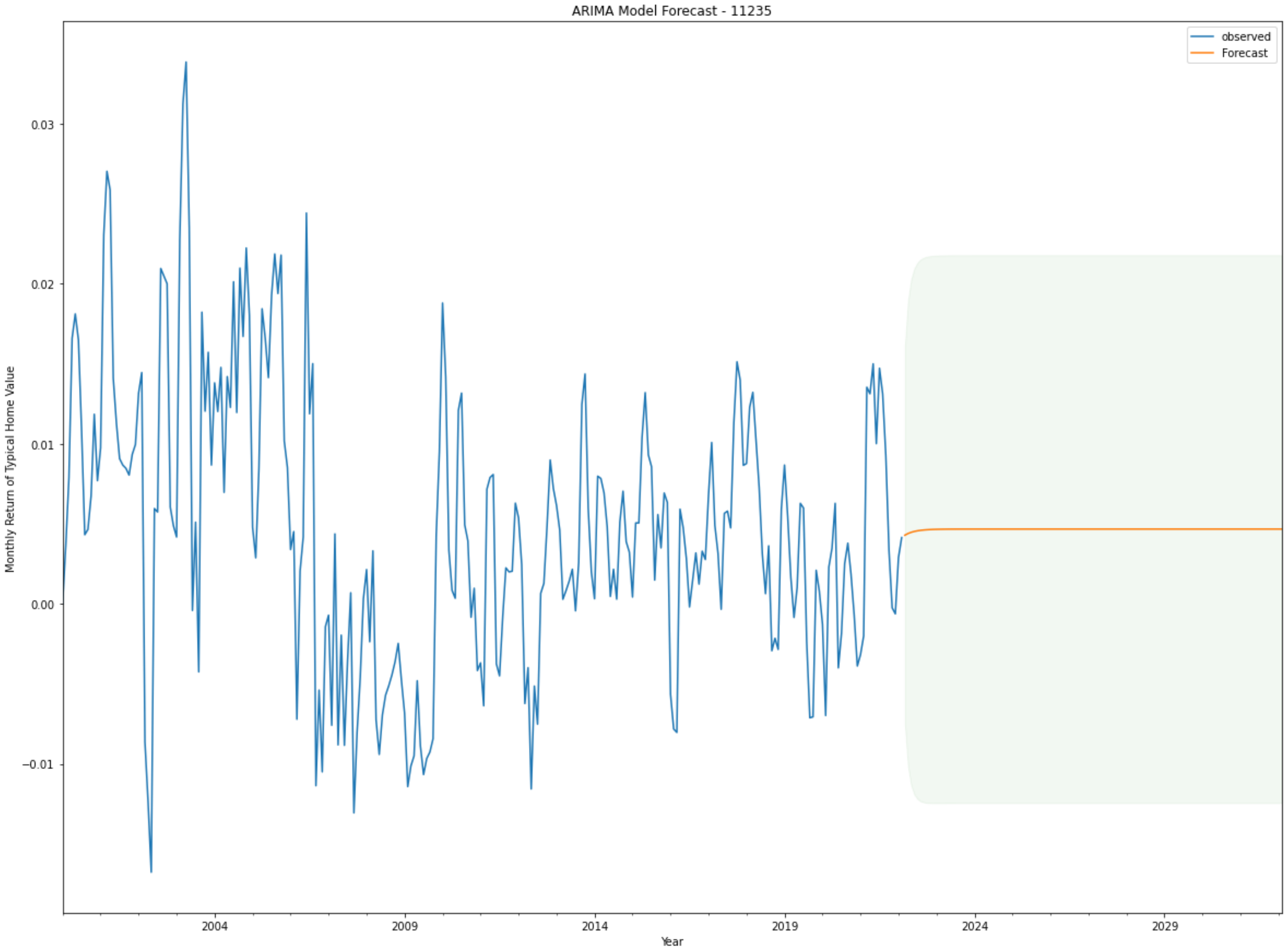
We performed one-step ahead forecasting and dynamic forecasting to predict the model and validate the model. We further checked the accuracy of the forecasts using MSE which are all close to 0.0.





## Producing and Visualizing Forecasts

We leveraged the .get_forecast() to obtain the forecast of the return in 1 year, 3 years, 5 years and 10 years.



We summarized our findings below:

Zipcode 11235, Brooklyn NY - 18.1% ROI in 3 years Zipcode 14217, Kenmore NY - 16.5% ROI in 3 years Zipcode 10475, Bronx NY - 13.5% ROI in 3 years Zipcode 11414, Queens NY - 11.7% ROI in 3 years Zipcode 14611, Rochester NY - 10.3% ROI in 3 years

## Summary and Next Steps

Based on the five criteria (State, Typical Home Value, Urbanization, Total Growth Rate and Diversification) of investing in real-estate market, we selected 5 zipcodes in New York State with the highest growth rate between year 2009 and 2020 (which excluding the financial crisis and pandemic, these two large events). Upon the selection, we performed a Time Series modeling with ARIMA to predict the future ROI.

Forecasts are solely based on historic monthly returns, and past performance does not necessarily predict future results. We excluded the financial crisis and pandemic when we conducted the preliminary selection. However, when our client start the investment, it will be at the time period of post-pandemic. The typical home value in these areas will be changed.

Therefore, as next steps, we need to take into account of other external factors to improve the models and ultimately the quality of the forecasts.

The factors to be further taken into consideration are:

- historic event: financial crisis / pandemic
- macro-economics: interest rate (mortgage rate, our client will apply leverage to their investment)
- house market safety score: given the hate crime is severe in the New York City, the safety will be a key concern when people buy the houses.

## Releases

No releases published
Create a new release

## Packages

No packages published
Publish your first package

## Languages

● **Jupyter Notebook** 100.0%