PREDICTING FUTURE HOUSING MARKETS WITH ZILLOW'S DATA

Business Understanding

Overview/Background Information

A real estate company is on a mission to find the hottest zip codes to invest in. They've got historical housing data from Zillow, like a time capsule, and they'll use it to predict future price trends. By analyzing this data, they want to pinpoint the top 5 zip codes with the most promising investment potential.

Problem Statement

Create a data-based recommendation for the top 5 zip codes the real estate investment firm should focus on, considering future price trends, risk factors, and the firm's investment timeframe.

Objectives

1. Build a model to predict future real estate prices for different zip codes.
2. Review the predictions by looking at profit potential, risk, and how long to hold the investment.
3. Suggest the top 5 zip codes for investment based on these factors.

Challenges

1. Determining the "optimal" investment requires balancing the potential for profit (expected price appreciation), risk tolerance (price instability), and investment duration (holding period).
2. Real estate prices are influenced by factors beyond past data, including economic conditions, local development projects, and interest rates.
3. Time series models have limitations; they cannot ensure accurate predictions, and future market behavior is inherently uncertain.

Proposed Solution: Metrics of Success

To address these challenges, we propose the following approach:

1. Apply time series forecasting to the provided historical Zillow data to predict future real estate prices across different zip codes.
2. Analyze the forecasts by evaluating the profit potential, which is the expected price increase in each zip code; the risk, which involves the historical price volatility in each zip code; and the investment horizon, which refers to the planned duration of holding the investment.
3. Prioritizing investment opportunities involves analyzing forecasted prices, profit potential, and risk to identify the top 5 most promising zip codes for real estate investment.

Conclusion

We'll predict future real estate prices and look at how much money can be made and how risky each area is. This will help us recommend the 5 best zip codes to invest in. How well we do depends on how good our predictions are and how carefully we consider everything.

Data Understanding

To understand the data structure, we'll examine its shape (number of rows and columns), list the column names, and identify the data types for each column.

Our data is a csv file and it contains 14,723 rows and 272 columns. The dataset contains the following columns:

- Region ID: Unique identifier for the region.
- Region Name: Zip code of the region.
- City: City name.
- State: State abbreviation.
- Metro: Metro area.
- County Name: County name.
- Size Rank: Rank by size.
- Date columns (from 1996-04 to 2018-04): Real estate prices for each month in this period.

The data types are as follows:

1. 49 columns are of type int64 (mostly identifiers).

2. 219 columns are of type float64 (real estate prices).
3. 4 columns are of type object (text data: City, State, Metro, and County Name).

Statistical Summary:

- There's a clear upward trend in real estate prices. The average price in 1996 was around 118,299 in dollars, and by 2018, it had risen to 288,039 in dollars.
- Prices vary a lot between zip codes, showing the different types of real estate markets in each area.

Data preparation and analysis

We're going to create a function that reads our data and gives us an overall view of it.

From this we can see the various aggregate statistics and we can also see that our data has house prices for various regions from April 1996 to April 2018 we also note that the Region Name column represents the various zip codes. So we'll appropriately rename the column.

We now look into Return on Investment for various cities, since we would like to recommend only the best options for our stakeholders and the outcome is that the best states to invest in is definitely New York. So we'll focus on this city for our project.

We check for unique values, remove the outliers, check for missing values and clean the data.

EXPLORATORY DATA ANALYSIS

Exploratory Data Analysis involves the following;

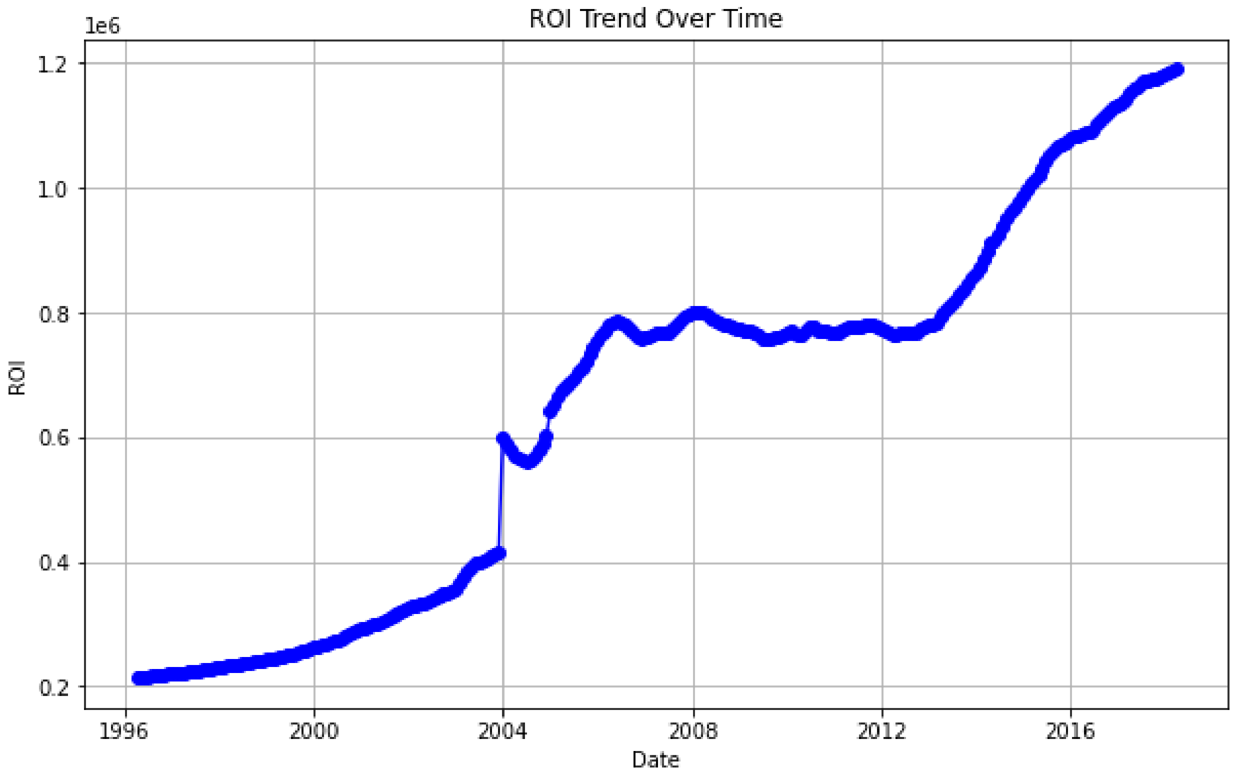
1. Univariate Analysis: Involves the analysis of individual variables to understand their distribution and summary statistics.

2. Bivariate Analysis: Involves the analysis between two variables.
3. Multivariate Analysis: Involves the analysis among three or more variables.

So for our case we are going to use univariate alone as we only have 1 column of data against

Univariate Analysis

Distribution of the ROI



From this plot we can see

- Overall Trend:

The ROI values show a clear upward trend over time. This indicates consistent growth in ROI from 1996 to 2018.

- Periods of Rapid Growth:

There are several periods where the ROI increases sharply, such as around the years 2004 and in 2014. These periods of rapid growth could be attributed to various economic factors.

- Plateaus and Stabilization:

Between approximately 2008 and 2013, the ROI appears to stabilize with minor fluctuations. This could suggest a period of market stabilization or maturity.

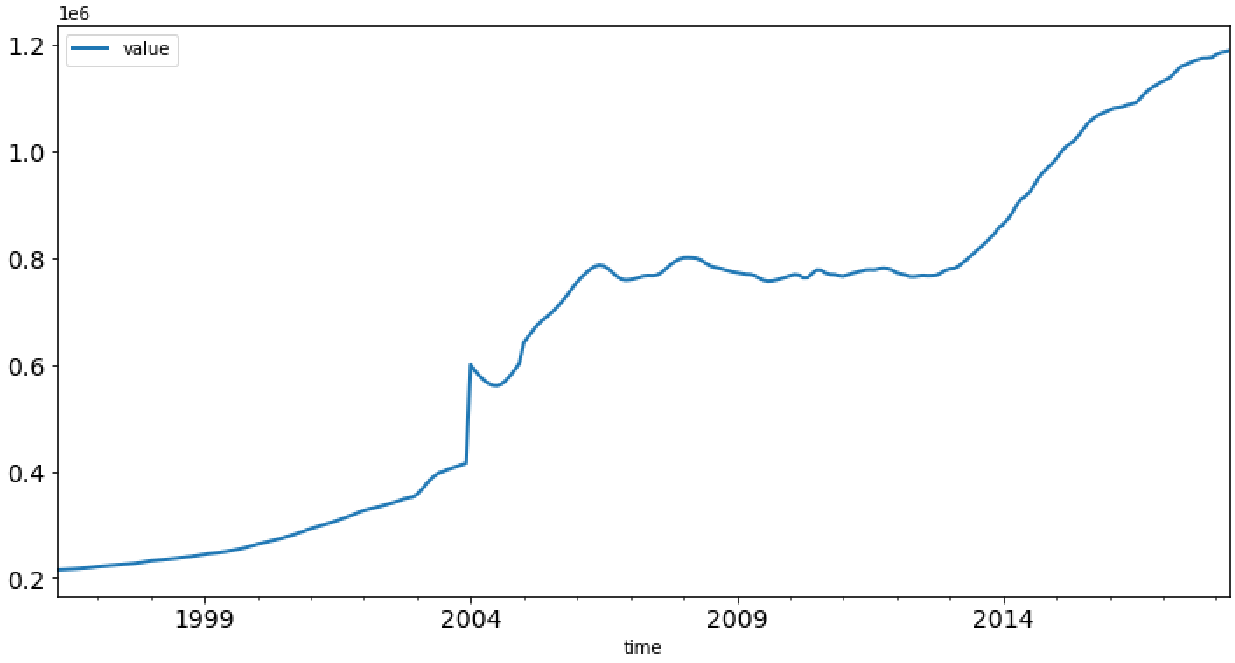
- Overall Volatility:

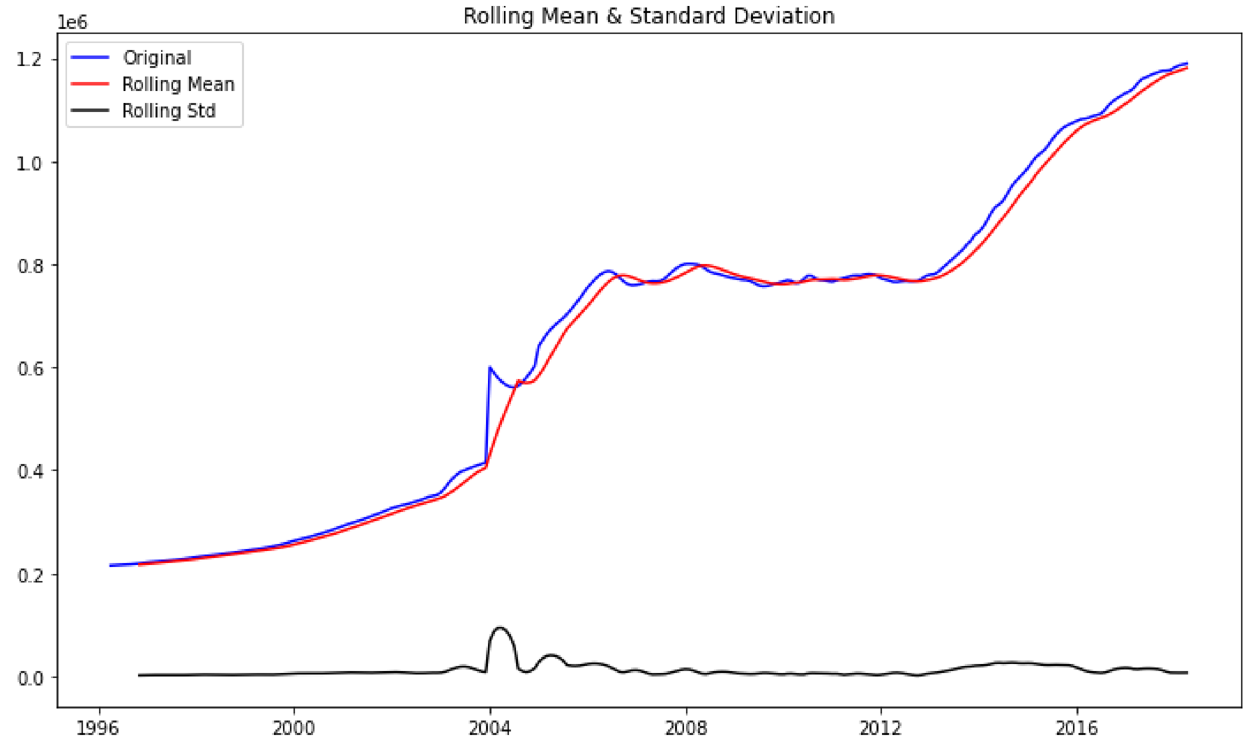
The general trend appears smooth, with a few notable spikes or drops. This suggests that while there is growth, the ROI experiences periods of volatility which could be due to market conditions or specific events impacting the ROI.

- Seasonality or Cyclical Patterns:

There doesn't seem to be a clear cyclical or seasonal pattern from the plot, but further analysis might be needed to confirm this.

TIME SERIES ANALYSIS PREPARATION





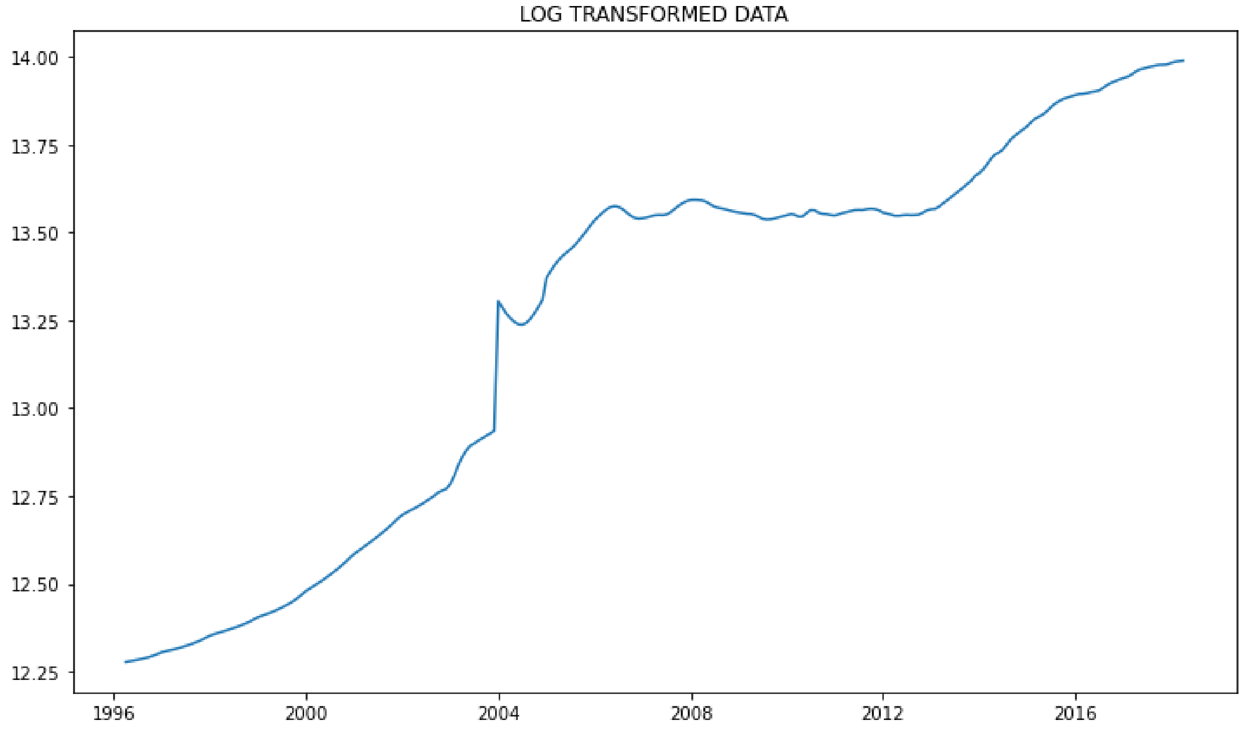
From this we can see that the data is not stationery. Let’s try the ad fuller method.

```
Results of Dickey-Fuller test:

Test Statistic                     0.573334
p-value                            0.986940
#Lags Used                         0.000000
Number of Observations Used      264.000000
Critical Value (1%)               -3.455365
Critical Value (5%)               -2.872551
Critical Value (10%)              -2.572638
dtype: float64
```

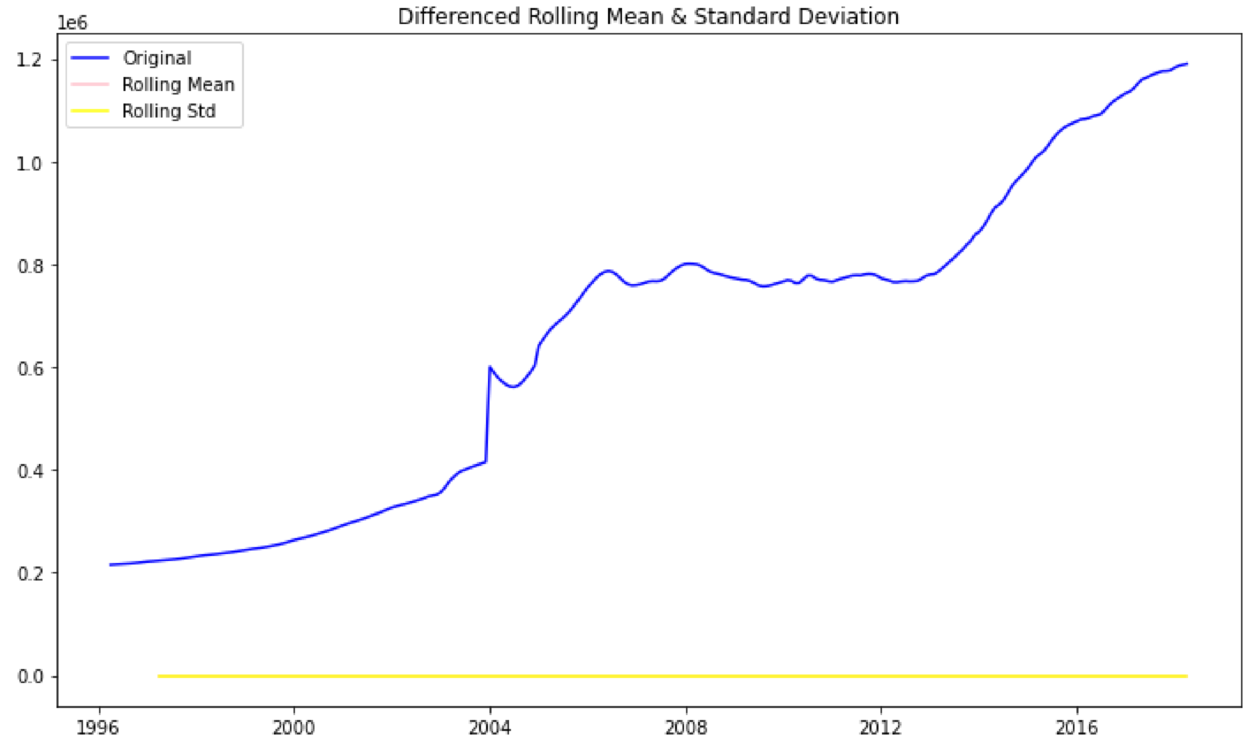
The results of the Dickey-Fuller test indicate that the time series is non-stationary. The test statistic of 0.573334 is higher than the critical values at the 1%, 5%, and 10% significance levels, and the p-value of 0.986940 is significantly greater than 0.05. Therefore, we fail to reject the null hypothesis that the time series has a unit root. This implies that the ROI values over time are not stationary and exhibit trends, which is consistent with the observed upward trend in the data.

Since our data is not stationary we start with the log transformation to try and make it stationery.



Clearly this hasn’t made things better, next let's try differencing method to remove trends and seasonality.

DIFFERENCING 1

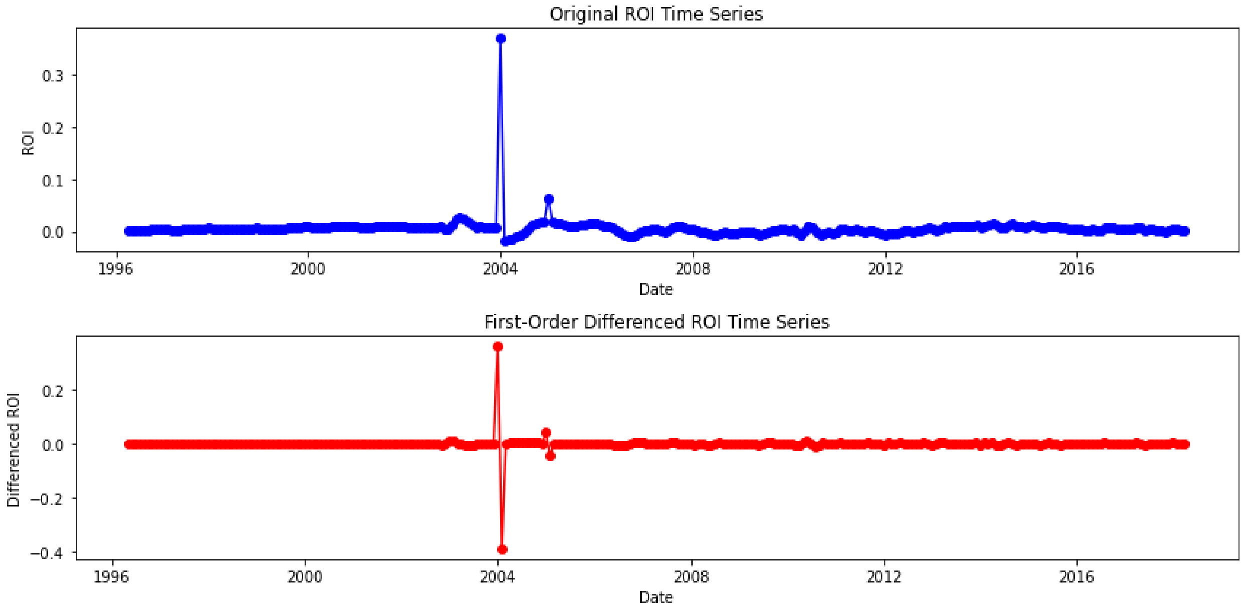


```
Results of Dickey-Fuller test:


Test Statistic                    0.573334
p-value                           0.986940
#Lags Used                        0.000000
Number of Observations Used     264.000000
Critical Value (1%)              -3.455365
Critical Value (5%)              -2.872551
Critical Value (10%)             -2.572638
dtype: float64
```

The results of the Dickey-Fuller test indicate that the time series is non-stationary. The test statistic of 0.573334 is higher than the critical values at the 1%, 5%, and 10% significance levels, and the p-value of 0.986940 is significantly greater than 0.05. Therefore, we fail to reject the null hypothesis that the time series has a unit root. This implies that the ROI values over time are not stationary and exhibit trends, which is consistent with the observed upward trend in the data.

DIFFERENCING 2;

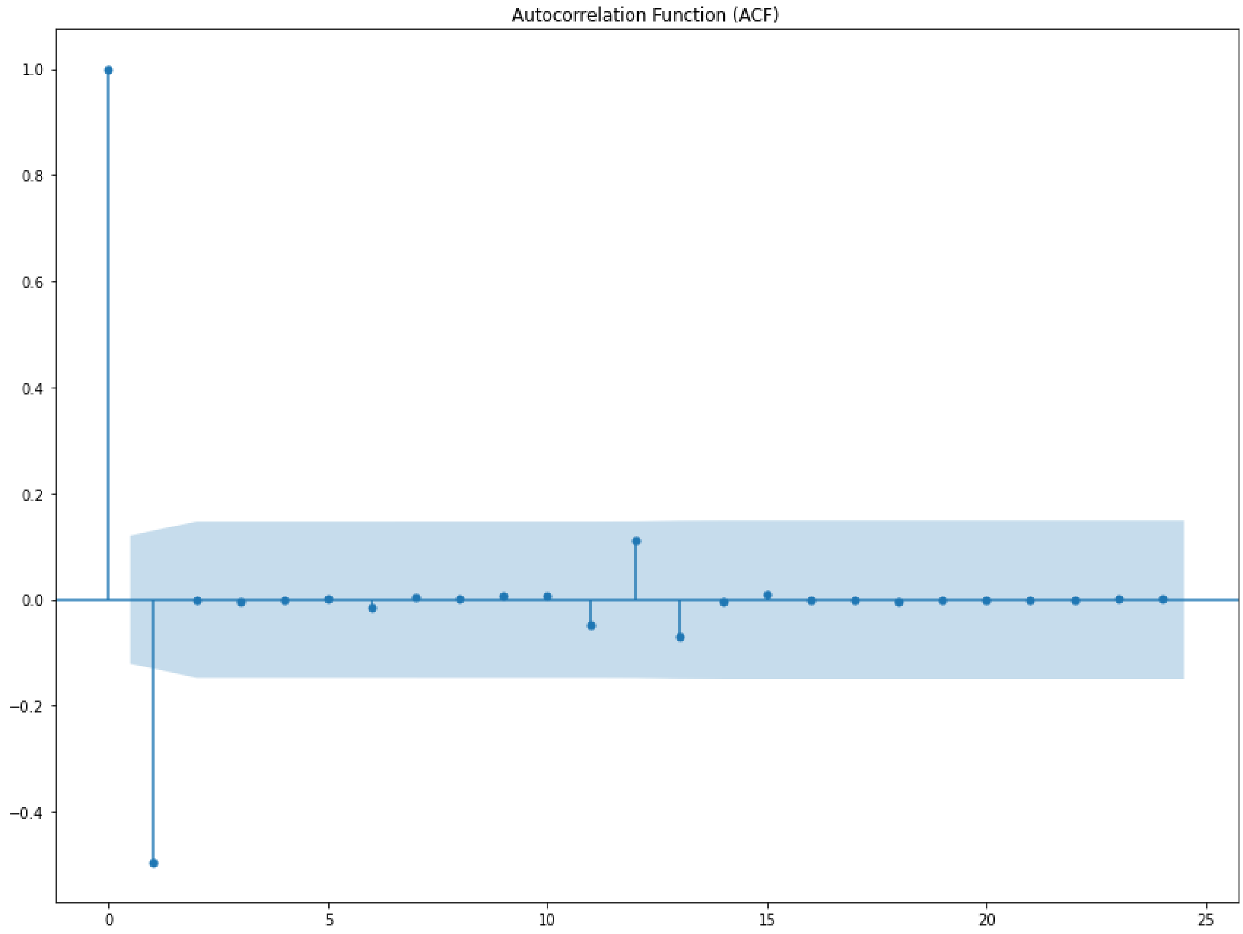


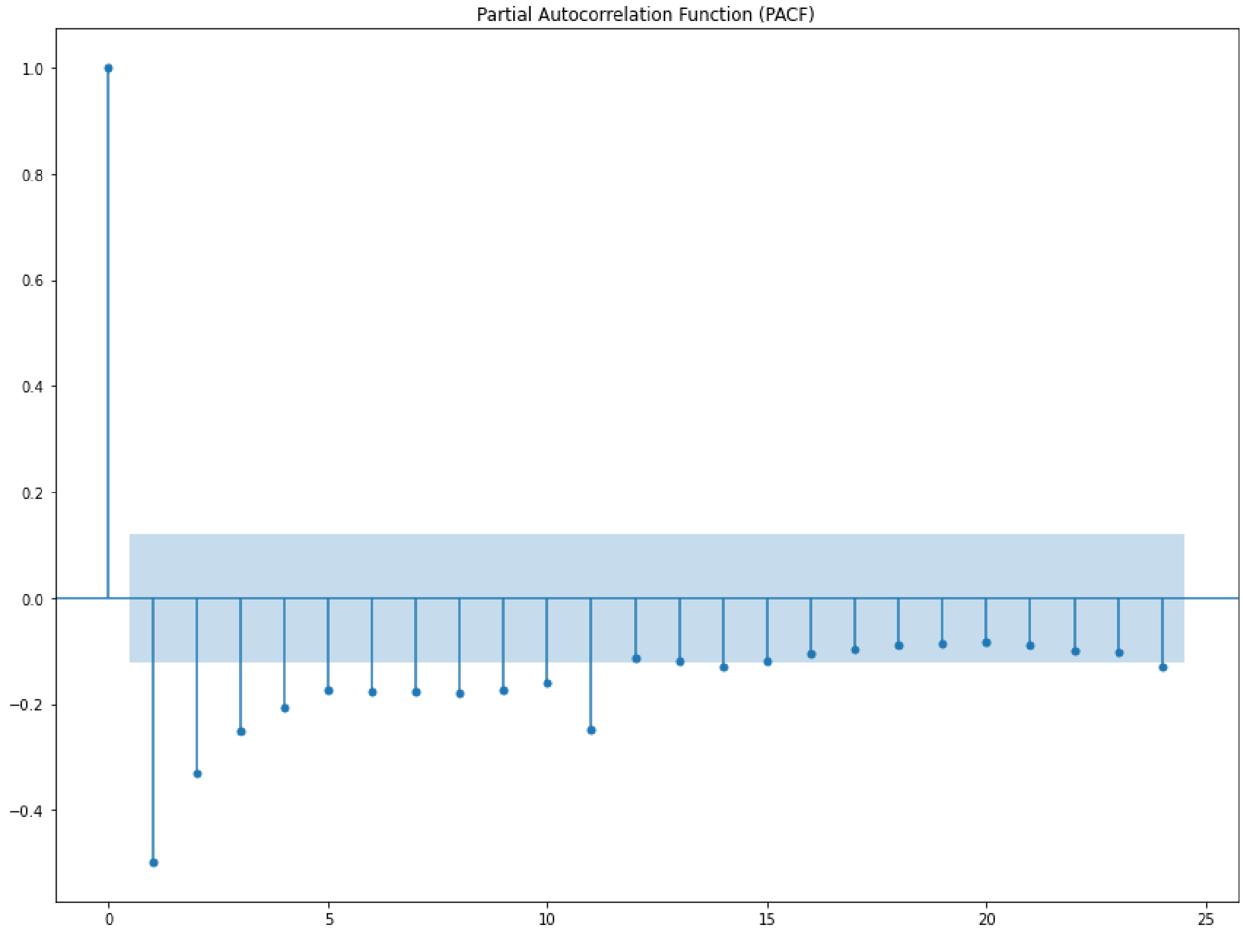
```
Results of Dickey-Fuller Test:

Test Statistic: -10.126381898689557
p-value: 9.15836365445215e-18
#Lags Used: 10
Number of Observations Used: 253
Critical Value (1%): -3.4564641849494113
Critical Value (5%): -2.873032730098417
Critical Value (10%): -2.572894516864816
```

The results of the Dickey-Fuller test on the first-order differenced series indicate that the time series is now stationary. The test statistic of -10.126381898689557 is much lower than the critical values at the 1%, 5%, and 10% significance levels, and the p-value of 9.15836365445215e-18 is significantly lower than 0.05. Therefore, we reject the null hypothesis that the differenced time series has a unit root. This implies that the first-order differencing successfully removed the non-stationarity from the original ROI time series, making it suitable for further time series modeling.

Plotting the ACF and PACF;


Autocorrelation Function (ACF)
1.0
0.8
0.6
0.4
0.2
0.0
−0.2
−0.4
0
5
10
15
20
25


Partial Autocorrelation Function (PACF)
1.0
0.8
0.6
0.4
0.2
0.0
−0.2
−0.4
0
5
10
15
20
25

Based on these plots:

- The significant spike at lag 1 in both ACF and PACF suggests that the time series is likely an AR(1) process, where the current value is primarily influenced by the previous value.
- The quick drop to near zero in both plots for lags beyond 1 indicates that there is no significant autocorrelation or partial autocorrelation at higher lags.

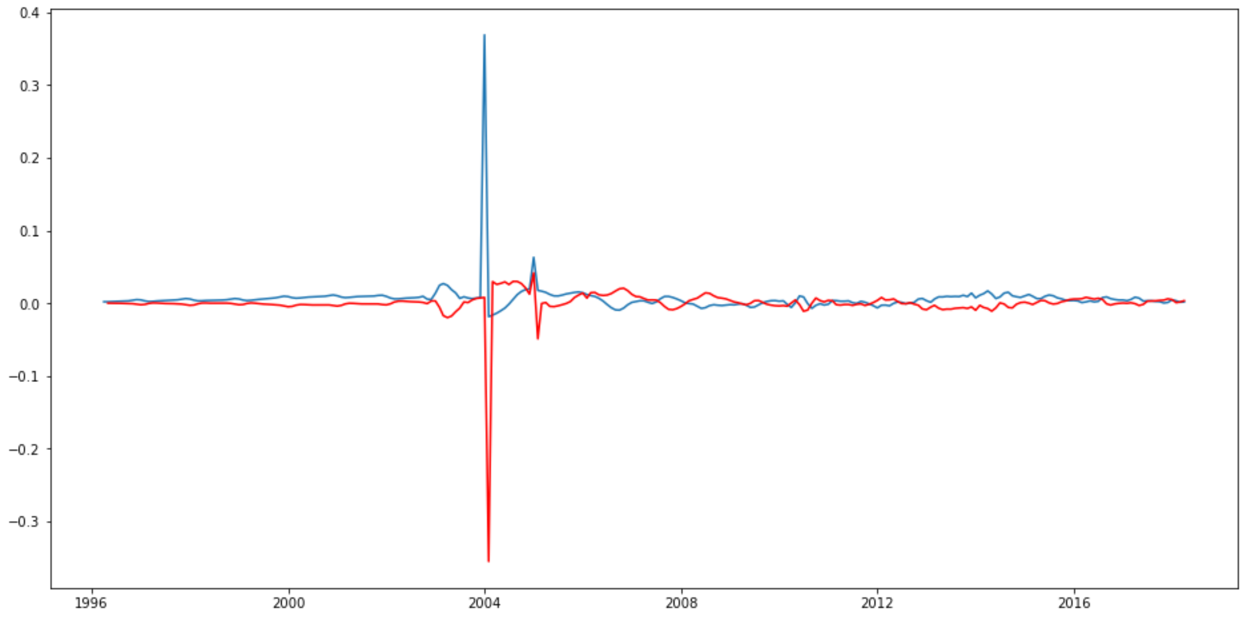
Modelling

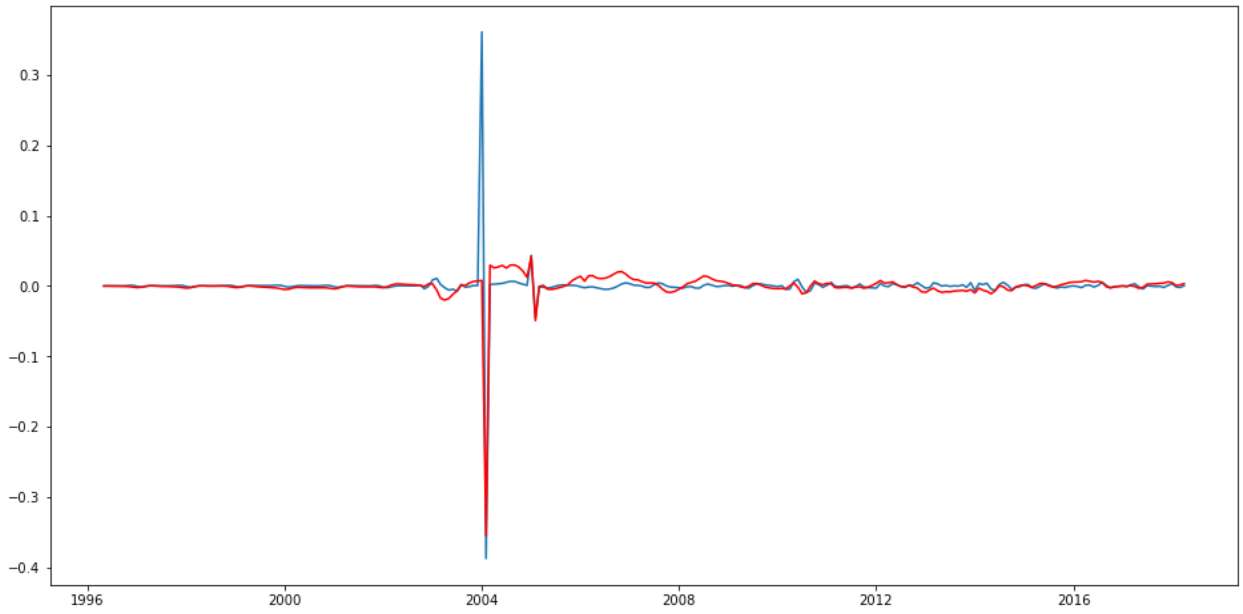
Now that we have done all the necessary preprocessing, we can go to the fun part and build a model that helps our investors in determining future prices of houses in New York. NY has the best returns and that is why we want to help investors get the best return for their investments and value for money

The Model we are going to build is ARIMA and try and tune it to get the best predictions for future prices.

From the auto correlation function and partial auto correlation functions determined above, the best p, d, q arrangement is (12, 1, and 12).

We are going to split our data into train and test split in order to gauge the performance of the model in predicting the future values





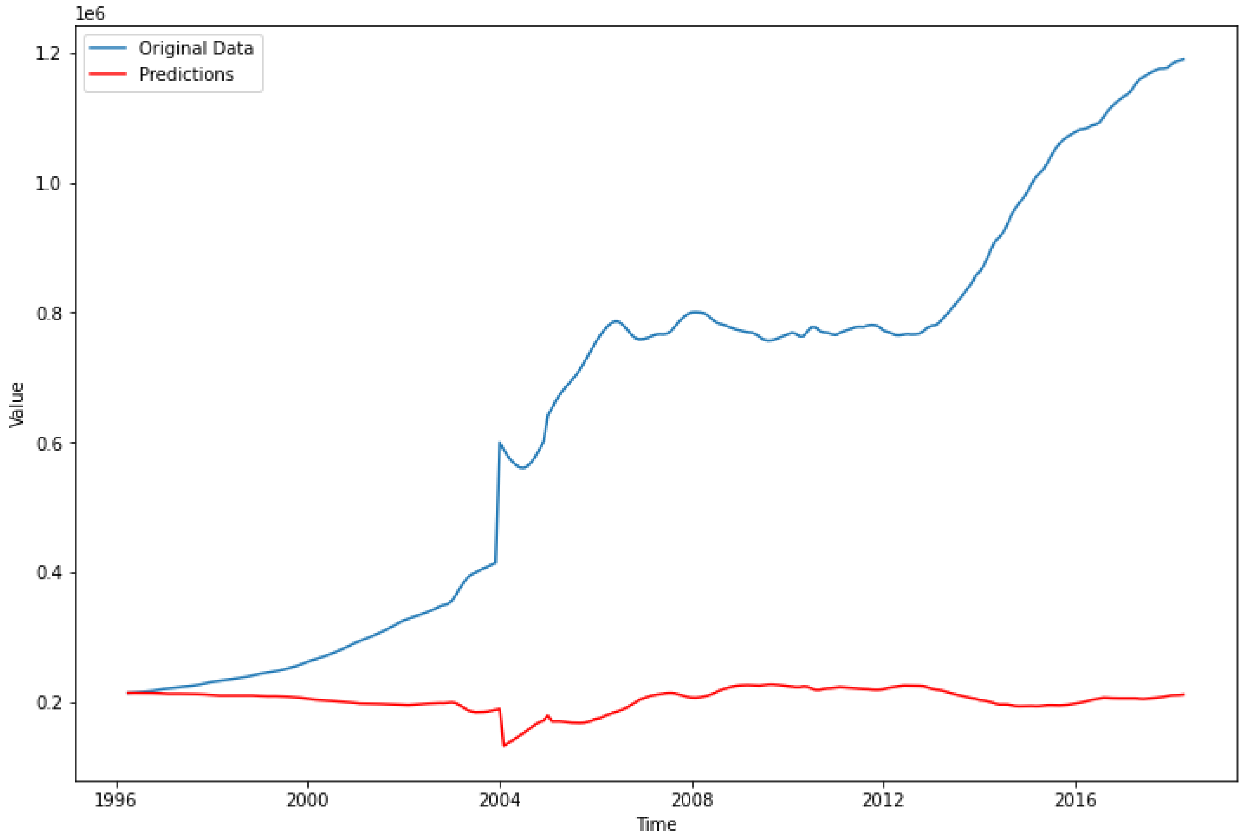
This is how our model is fitting into the data, the red is the prediction from the model while the light blue one is the actual values.

These plots typically compare the residuals of a fitted model to those of the original time series.

Residuals at higher lags appear randomly distributed and small, resembling white noise.

Both models seem to handle higher lags well, with residuals appearing random and small, suggesting they are well-fitted for those lags.

From here we will work backwards in order to get the actual values being predicted by our model



The blue line represents the "Original Data" and shows a significant upward trend, particularly after 2004. The values rise sharply, reaching around 1.2 million by the end of the period.

The red line represents the "Predictions" and remains relatively flat throughout the entire period, fluctuating slightly around 200,000

The original data shows a sharp increase around 2004 and continues to rise steadily.

The predictions do not capture the upward trend observed in the original data and remain largely unchanged.

RECOMMENDATIONS

For the base model the ARIMA captured the upward trend in predicting the values but for better performance we recommend to use a grid search or auto ARIMA to get better performance by selecting the best parameters

FURTHER STEPS

1. .Implement grid search or auto ARIMA: Using grid search or auto ARIMA to find optimal parameters (p,d,q) could improve model performance. This would help fine-tune the model beyond the initial (10, 0, 10) configuration.

2. Validate model assumptions: Check if the residuals of the ARIMA model meet key assumptions like normality and no autocorrelation. This can be done using diagnostic plots and statistical tests.

3. Perform out-of-sample forecasting: Split the data into training and test sets to evaluate how well the model performs on unseen data. This will give a better indication of its predictive power.

CONCLUSION

The ARIMA model has proven to be a powerful tool for forecasting real estate prices. However, it is crucial to continuously monitor the market and update the model with new data to maintain its accuracy. By leveraging these predictions, investors can make informed decisions and maximize their returns in the dynamic real estate market.