



# FUTURE HOUSING MARKET'S PREDICTIONS

**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 and by analyzing this data, they want to pinpoint the top 5 zip codes with the most promising investment potential.

## OBJECTIVES

- Build a model to predict future real estate prices by zip code.
- Analyze profit potential, risk, and investment duration.
- Recommend the top 5 zip codes for investment.

## 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.

## CHALLENGES

- Determining Optimal investment balances profit potential, risk, and holding period.
- Real estate prices depend on economic conditions, local projects, and interest rates, beyond past data.
- Time series models have limitations and can't guarantee accurate predictions; market behavior is unpredictable.



## PROPOSED SOLUTIONS

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, preparation and analysis

## DATA UNDERSTANDING



Our data is a csv file and it contains 14,723 rows and 272 columns.

The data types are as follows:

- 49 columns are of type int64 (mostly identifiers).
- 219 columns are of type float64 (real estate prices).
- 4 columns are of type object (text data: City, State, Metro, CountyName).

## Data preparation

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

We now look into Return on Investment for various cities, since we would like to recommend only the best options for our stakeholders

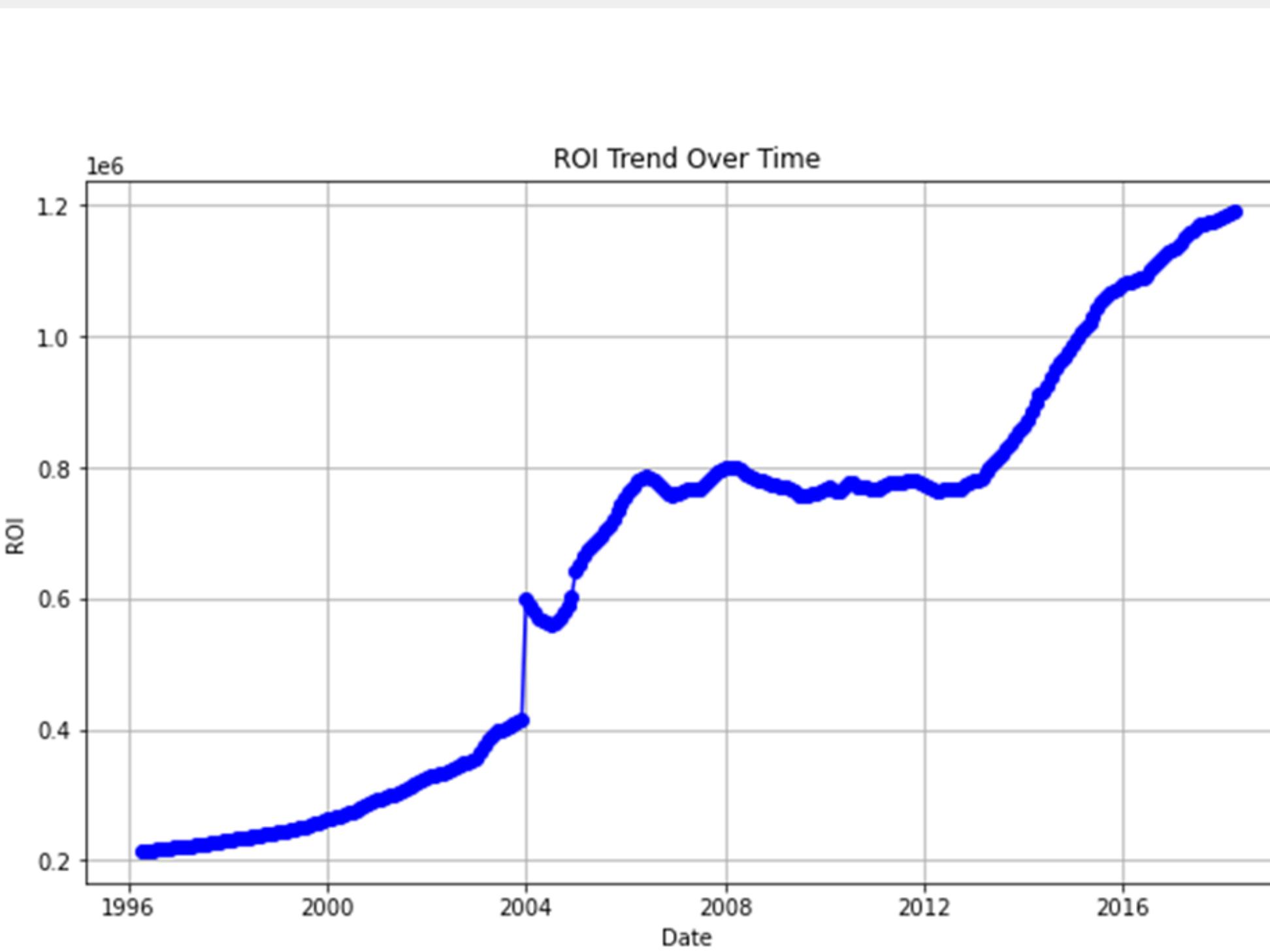
## EXPLORATORY DATA ANALYSIS

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

we are going to use univariate alone as we only have 1 column of data against

# Univariate Analysis

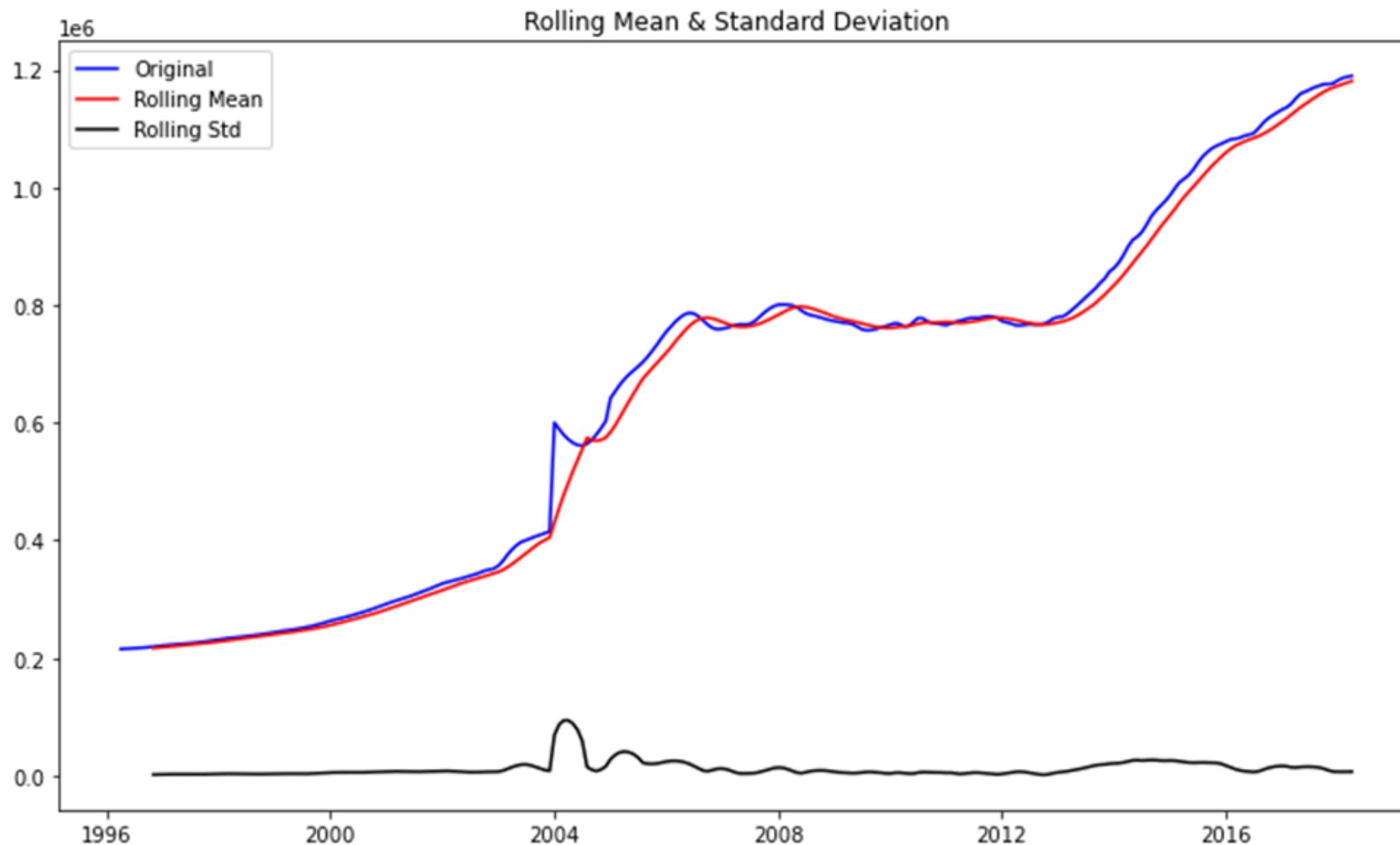
## Distribution of the ROI



# OBSERVATIONS

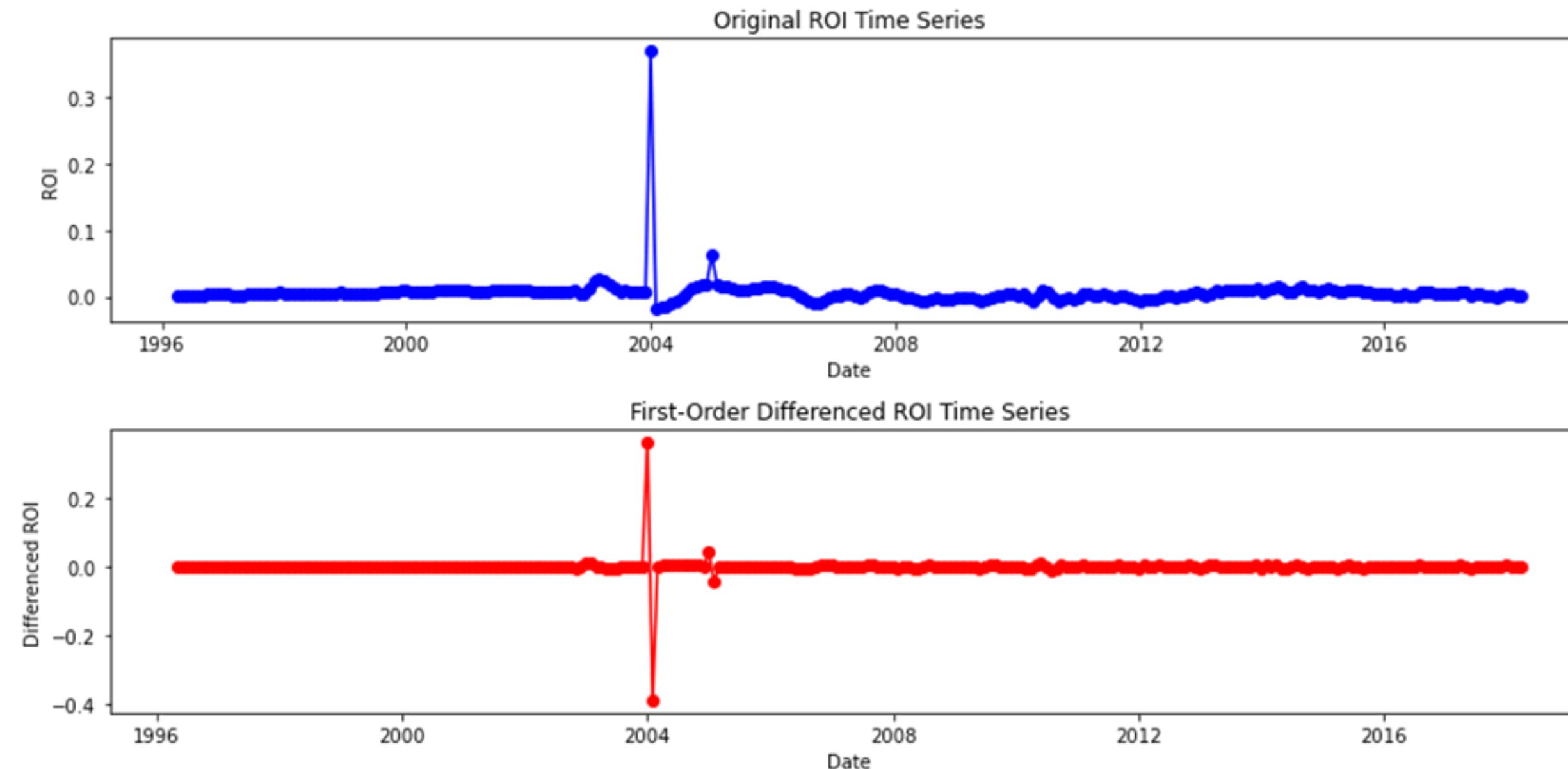
- Overall Trend: The ROI values show a clear upward trend over time
- Periods of Rapid Growth; There are several periods where the ROI increases sharply
- Plateaus and Stabilization: Between approximately 2008 and 2013, the ROI appears to stabilize with minor fluctuations.
- Overall Volatility: The general trend appears smooth, with a few notable spikes or drops.
- Seasonality or Cyclical Patterns: There doesn't seem to be a clear cyclical or seasonal pattern from the plot

# TIME SERIES ANALYSIS PREPARATION



From this we can see that the data is not stationary.

# DIFFERENCING

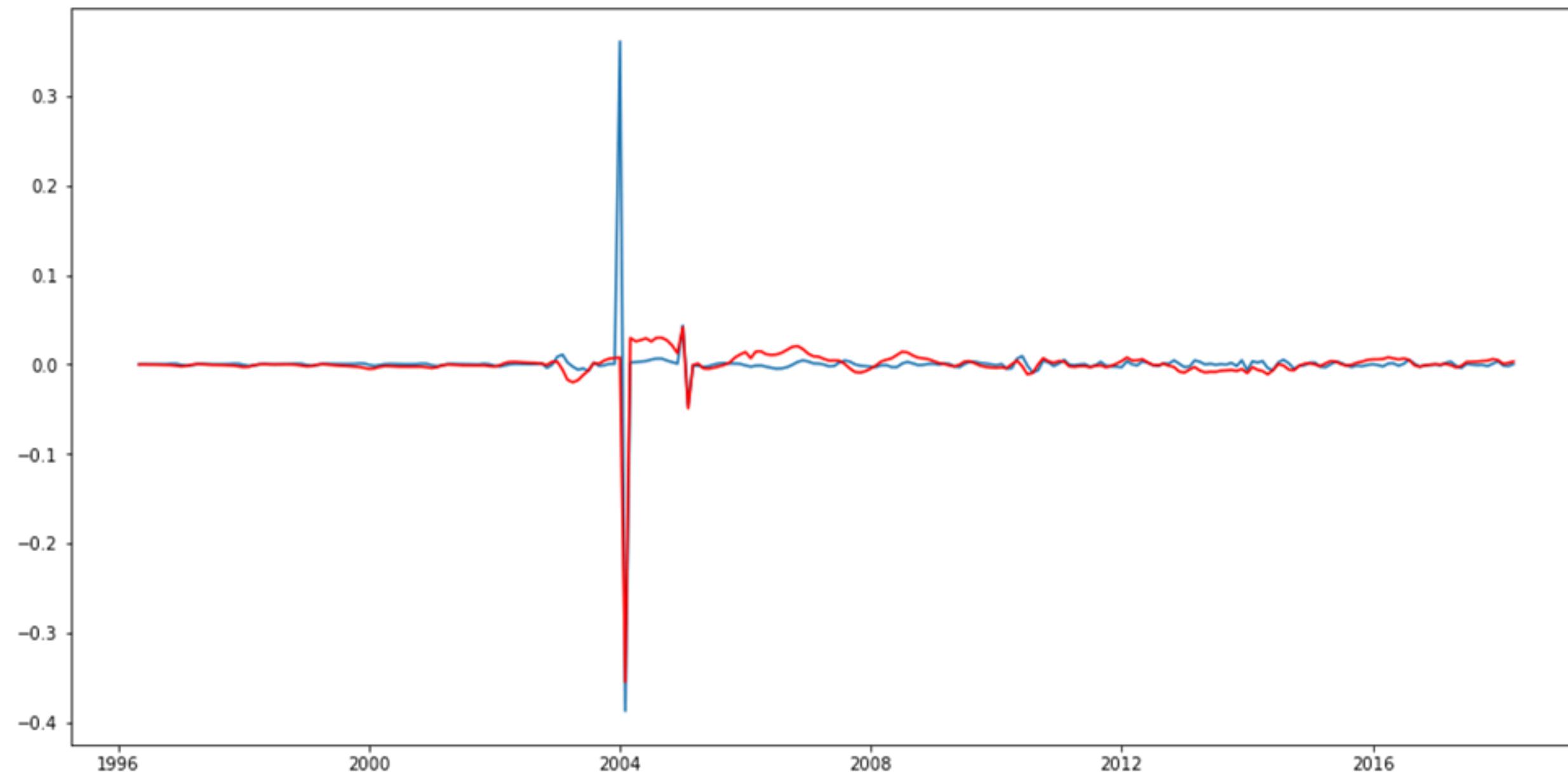


This implies that the first-order differencing successfully removed the non-stationarity from the original ROI time series, making it suitable for further time series modeling.

# MODELLING



We build models to help our investors in determining future prices of houses in new york.The Model we are going to build is ARIMA and try and tune it to get the best predictions for future prices.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. The 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.

## 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.

## 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



# Thanks for listening!

group members;

- Calvine Dasilve
- Jack otieno
- Sandra Kiptum
- Salahudin Salat
- Hellen Samuel