PHASE 4 GROUP PROJECT.
PROJECT TITLE: TIME SERIES MODELING FOR ZILLOW REAL ESTATE PRICES.

---

**GROUP 10 MEMBERS**

---


1. Celia Joy Omiah
2. Crystal Wanjiru
3. Kelvin Rotich
4. Miriam Nguru
5. Paul Mbugua
6. Stephen Butiya


# 1 Overview

Real estate constitutes land, structures, and tangible assets, encompassing natural resources like crops, minerals, or water. It is a versatile commodity that can be bought, sold, leased, or rented for various purposes. As a vital element of the global economy, real estate caters to residential, commercial, industrial, and agricultural needs. As an investment class, real estate offers a diverse approach for individuals and businesses to generate wealth and contribute to the economic landscape.

The USA currently has a population of about 331 million. The real estate industry substantially contributes to the nation's economy, constituting around 6% of the Gross Domestic Product (GDP). This diverse industry encompasses residential and commercial real estate, real estate development, property management, and real estate investment trusts (REITs).

The residential real estate sector holds the largest share, dominating real estate transactions. Factors such as population growth, low-interest rates, government policies, and employment opportunities drive the demand for residential properties. The sector, however experiences some challenges:

- Fluctuations in interest rates which influence the affordability of mortgages
- Zoning regulations and local government policies that pose challenges for real estate developers and limit the availability of affordable housing
- Housing affordability is a persistent challenge in many parts of the USA.

The challenges can be solved by:

- Implementation of programs by the government that provide financial assistance or tax incentives for first-time homebuyers or low-income families.
- Implementing monetary policies that maintain stable and affordable interest rates to support housing affordability.
- Review and update zoning regulations to allow for increased density, mixed-use developments, and affordable housing projects

# 2 Business Understanding

This project is designed to enhance the comprehension of real estate investors by employing time series analysis on Zillow data. By offering insights into historical property price trends, the project aims to assist investors in making informed decisions, mitigating risks, pinpointing favorable locations, and accessing forecasts through an intuitive interface. Investors can stay informed in the dynamic real estate market through a continuous improvement strategy, which includes a feedback loop and regular updates. This empowerment enables them to optimize investments and deepen their understanding of the industry.

## 2.1 Problem Statement

To develop a time series model that can be used to predict and help determine the top five zip codes in which to invest

## 2.2 Objectives

i. To identify the top 5 zip codes and states that offer the best investment potential in terms of real estate value. By analyzing historical trends and patterns, the project aims to provide actionable insights to the investment firm, enabling them to make informed decisions on where to allocate their resources.

ii. To analyze the historical real estate value data by looking into the monthly, quarterly, semi-annual, and annual patterns over time.

iii. To create a model that can predict future Real Estate Value.

# 3 Data Understanding

The dataset utilized in this project was gathered from various states in the USA, encompassing historical median house prices spanning from April 1996 to April 2018, covering 22 years. This data was sourced from the Zillow Research Page.

The dataset comprises 14,723 rows and 272 columns in wide format. Among the 272 columns, 4 are categorical, while the remaining columns are numerical.

The categorical columns are:

- City -Specific city name of housing data
- State - represents the state where the region is located.
- Metro - represents the metropolitan area where the region is located.
- County - this is the county name of that region

The numerical columns are:

- Region ID - is a unique ID for the regions
- Region Name - contains the zip code for the region
- Size Rank - this is the ranking done based on the size of that region.
- Date - represents the median home price for the region in months and years

# 4 Data Preparation

This process aims to prepare the data for optimal modeling. We have followed a series of steps:

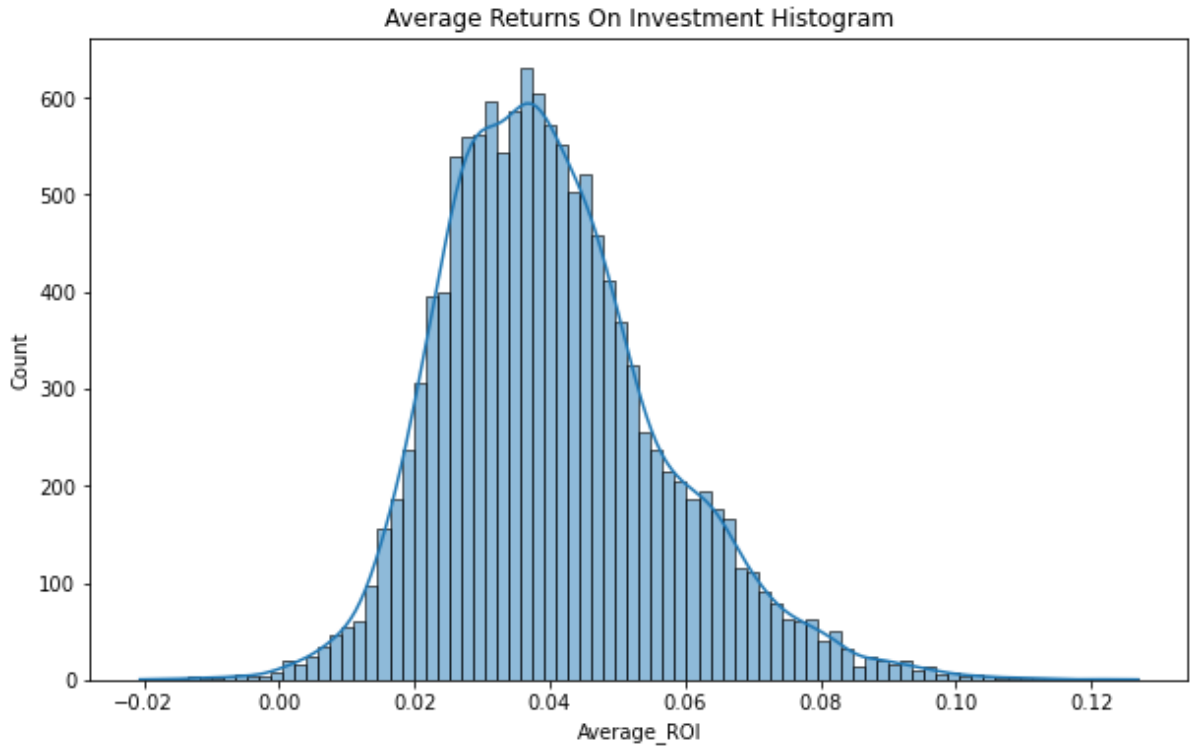
1. Cleaning the Data - Initiating a thorough cleaning process to rectify inconsistencies or inaccuracies within the dataset.
2. Calculating the average returns on investment
3. Handling Missing Values - Identifying and addressing missing values to ensure a comprehensive and complete dataset.
4. Checking for outliers, duplicates, placeholder values
5. Reshaping from Wide to Long Format- Transforming the dataset structure from a wide to a long format. This restructuring enhances the data's suitability for model input, providing a more organized and streamlined arrangement.
6. Compute pertinent time-related attributes, including moving averages and seasonality patterns, to effectively capture and analyze temporal trends.

## 4.1 Exploratory Data Analysis

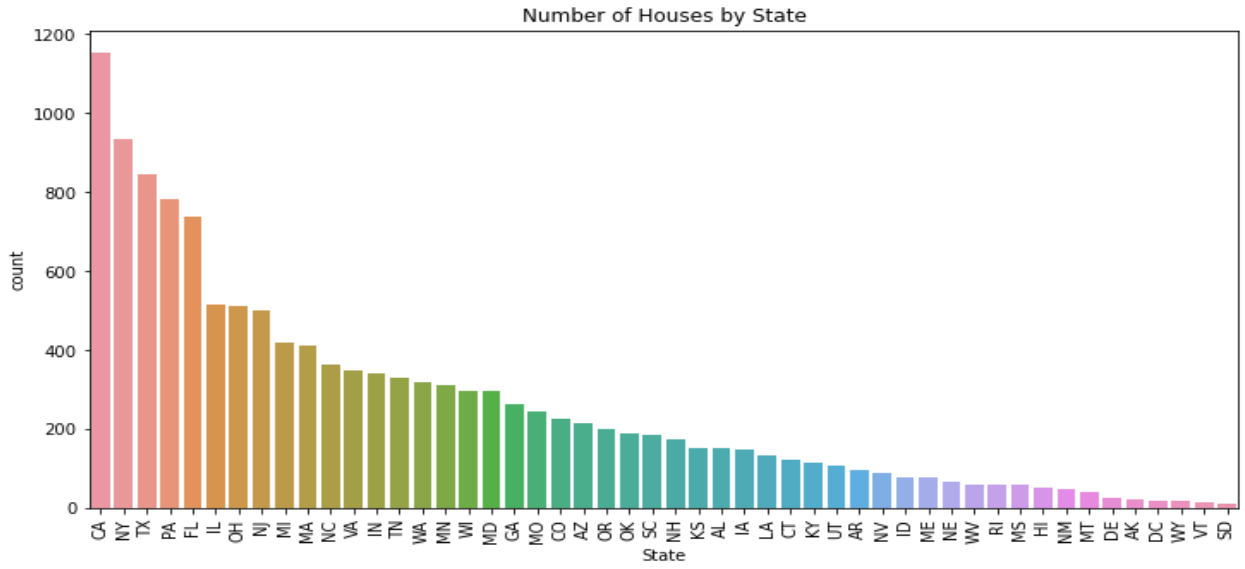
This analysis will involve looking into the dataset visually and statistically, examining and understanding the data before further splitting it for more analysis and modeling.

### 4.1.1 Univariate Analysis

We started our analysis by looking at the Return on Investment of our data.




From the histogram above, we can conclude that the distribution of the average returns on investment was a normal distribution.

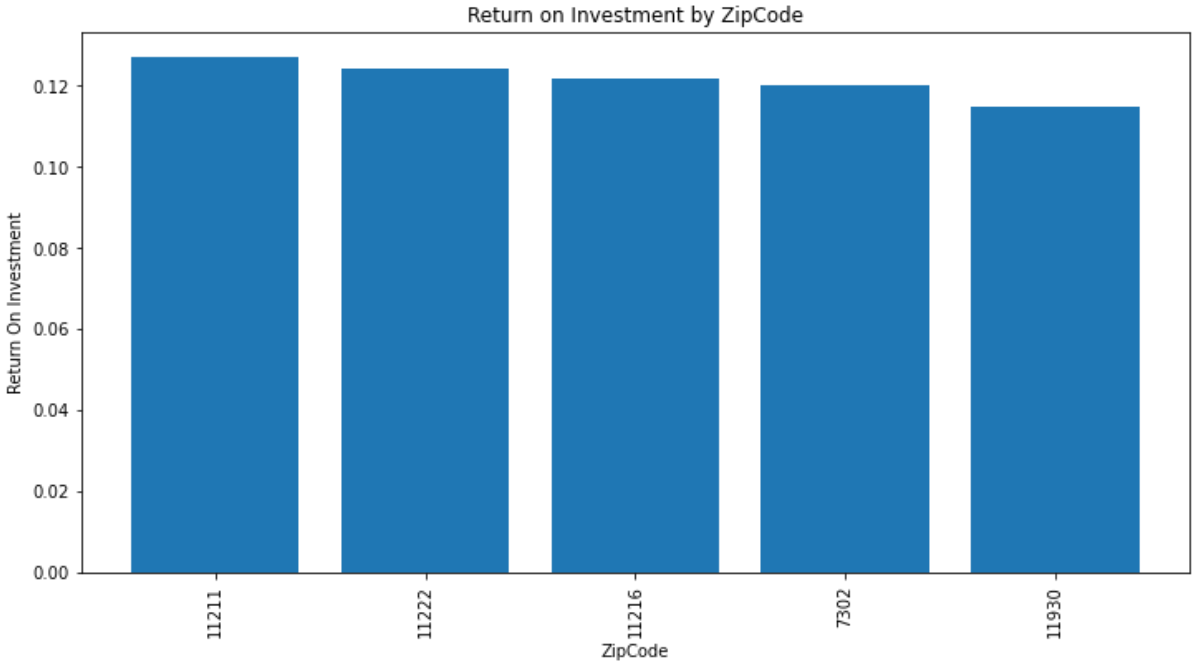



From the visualization above, California, New York, Texas, Pennsylvania, and Florida emerged as the states with the highest count of houses. At the same time, South Dakota, Vermont, Washington DC, Wyoming, and Arkansas registered the lowest number of houses, respectively.

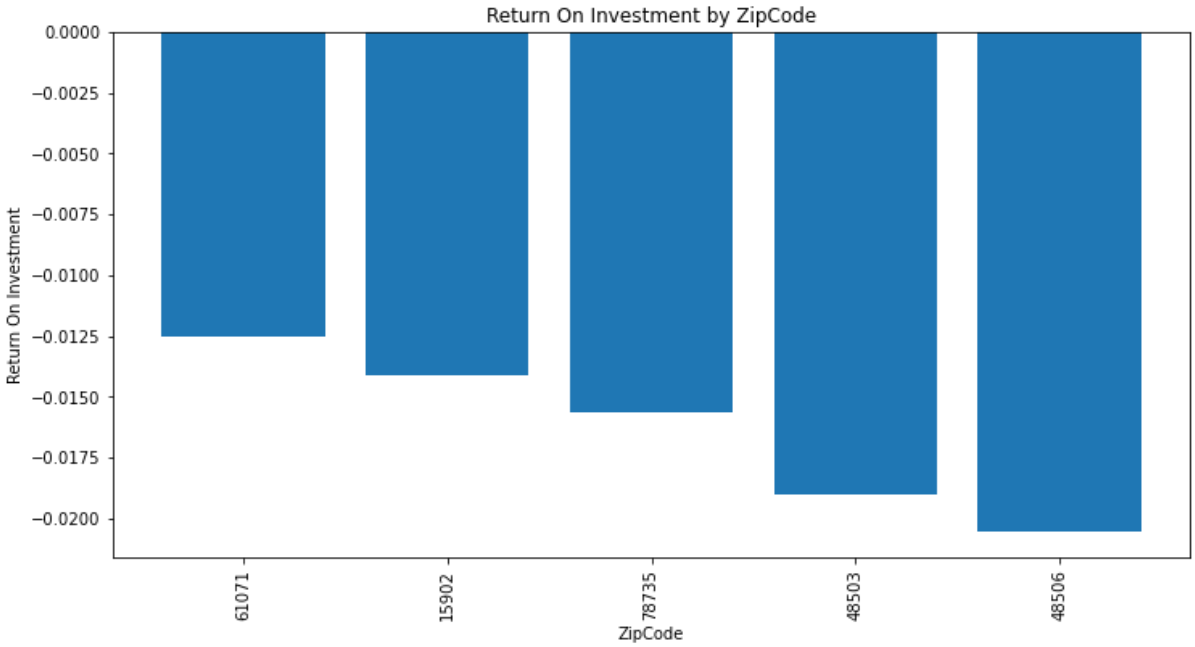
### 4.1.2 Bivariate Analysis

We analyzed two features, the zip code and Return on Investment, to explore their relationship and see how the changes in each feature affect the other.

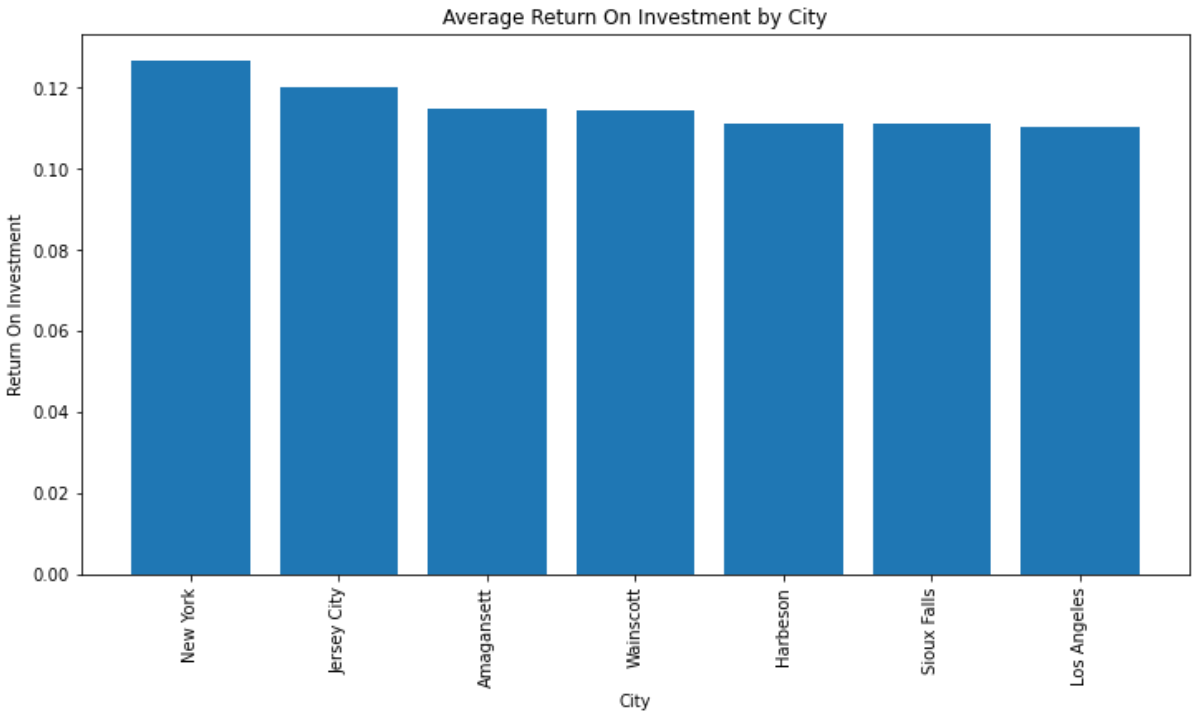
From the bar graph, we can conclude that the region with zip code 11211 in Brooklyn, New York; 111222 in Brooklyn, New York; 11216 Brooklyn, New York; 7302 Jersey City; and 11930 Amagansett, New York, had the highest return on investment.



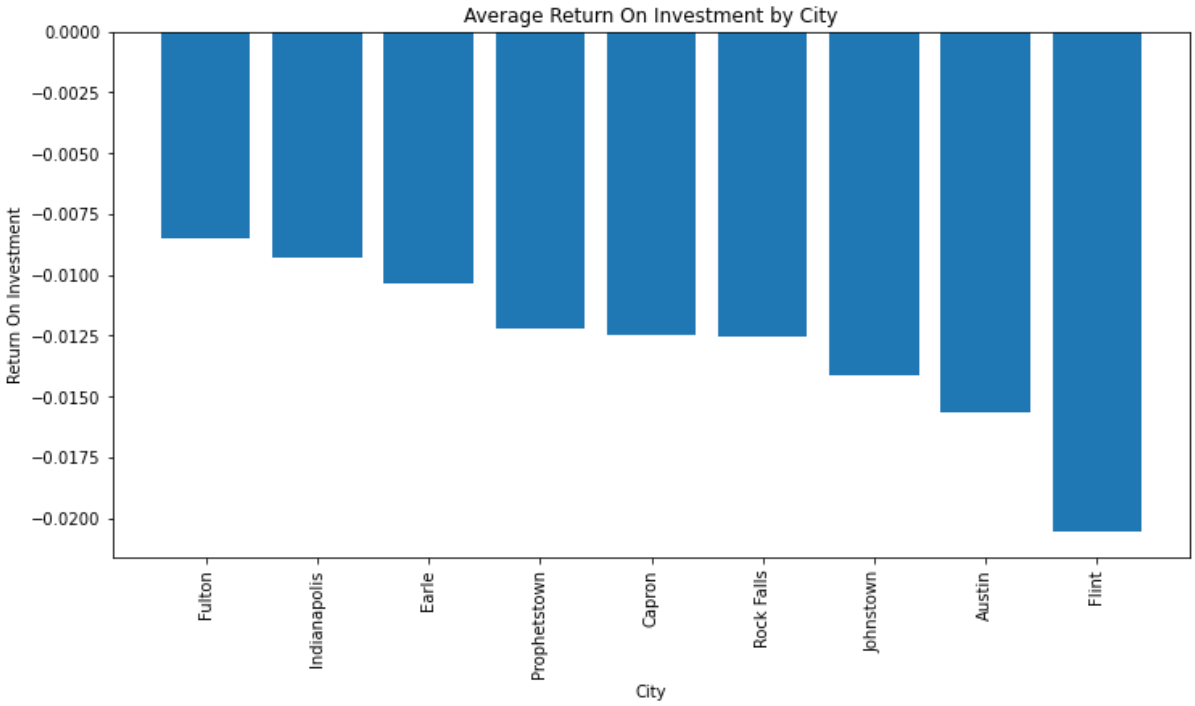

On the other hand, zip codes 48506 Flint, Michigan, 48503 Flint, Michigan, 78735 Austin, Texas, 15902 Johnstown, Pennsylvania, and 61071 Rock Falls, Illinois had the lowest ROI.



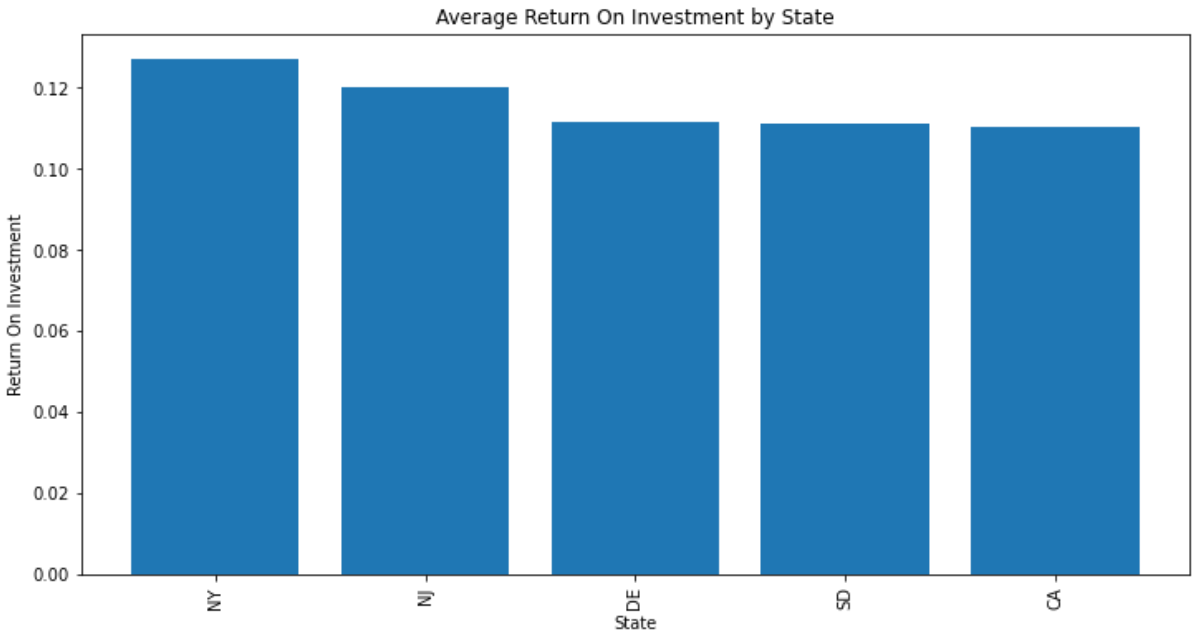

Additionally, the graph below shows the Average Return on Investment by city. New York City had the highest return on investment, followed by Jersey  City, Amagansett, Wainscott, Harbenson, Siouxx Falls, and Los Angeles had the highest returns on investment.

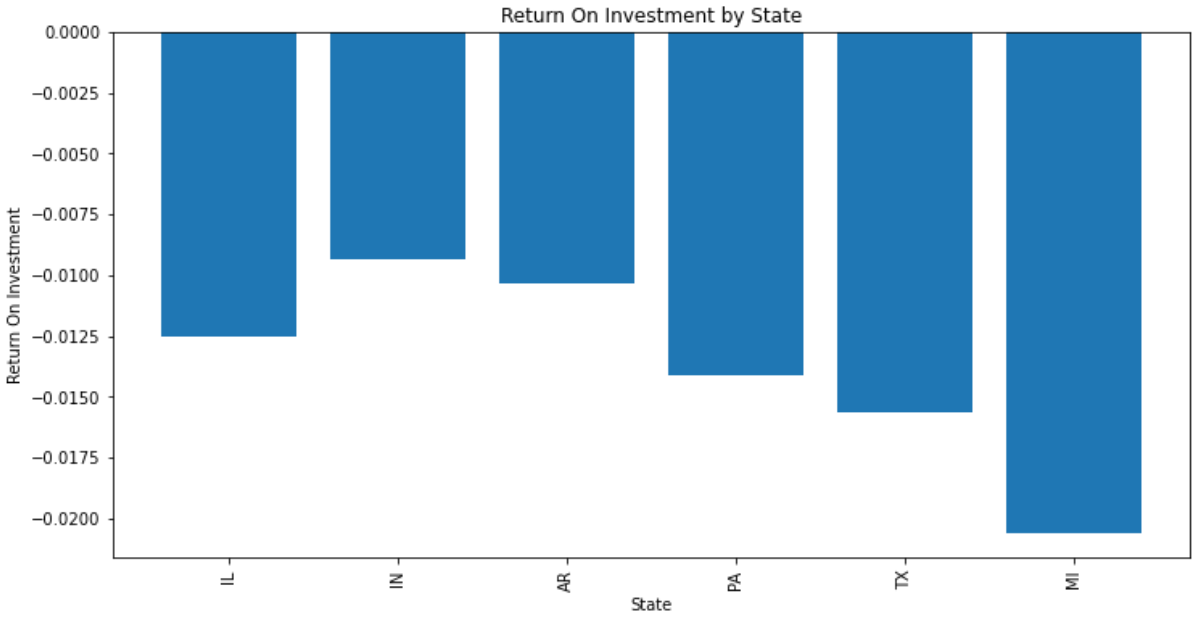



However, Fulton, Indianapolis, Earle, and Prophetstown had the lowest average ROI.




Furthermore, the states with the highest average returns on investment are New York, New Jersey, Delaware, South Dakota, and California. On the other hand, the states with the lowest returns on investment are Michigan, Texas, Pennsylvania, Arkansas, Indiana, and Illinois.

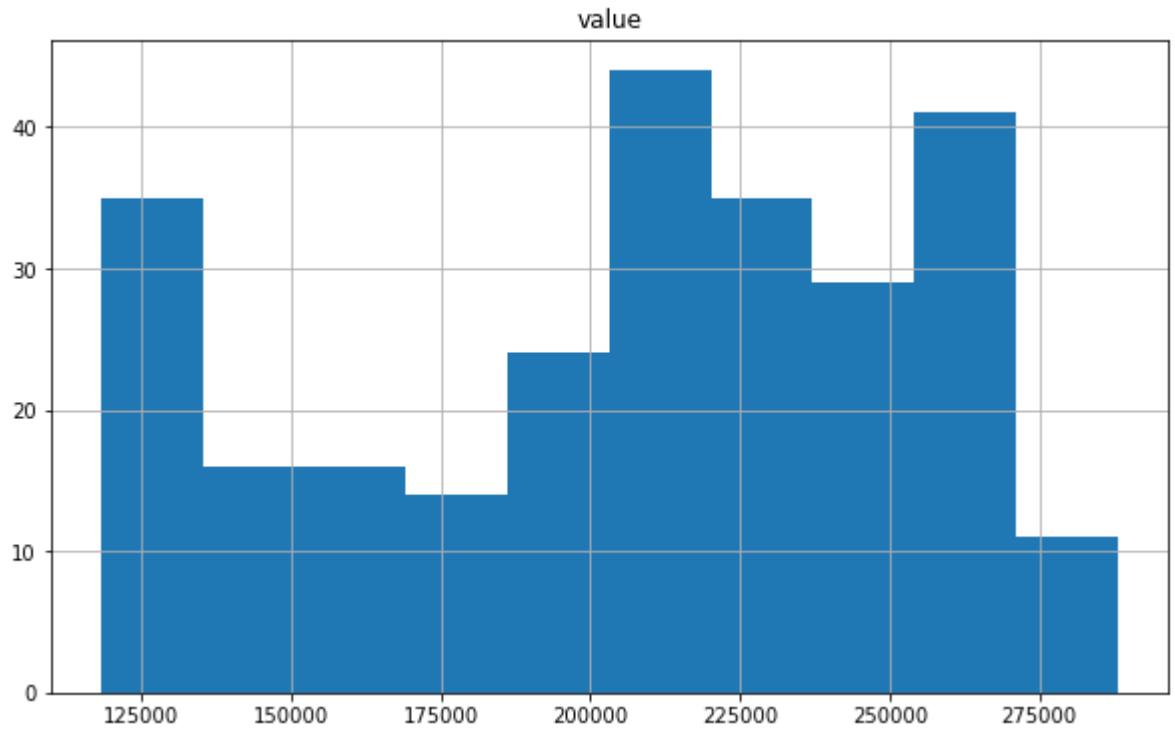




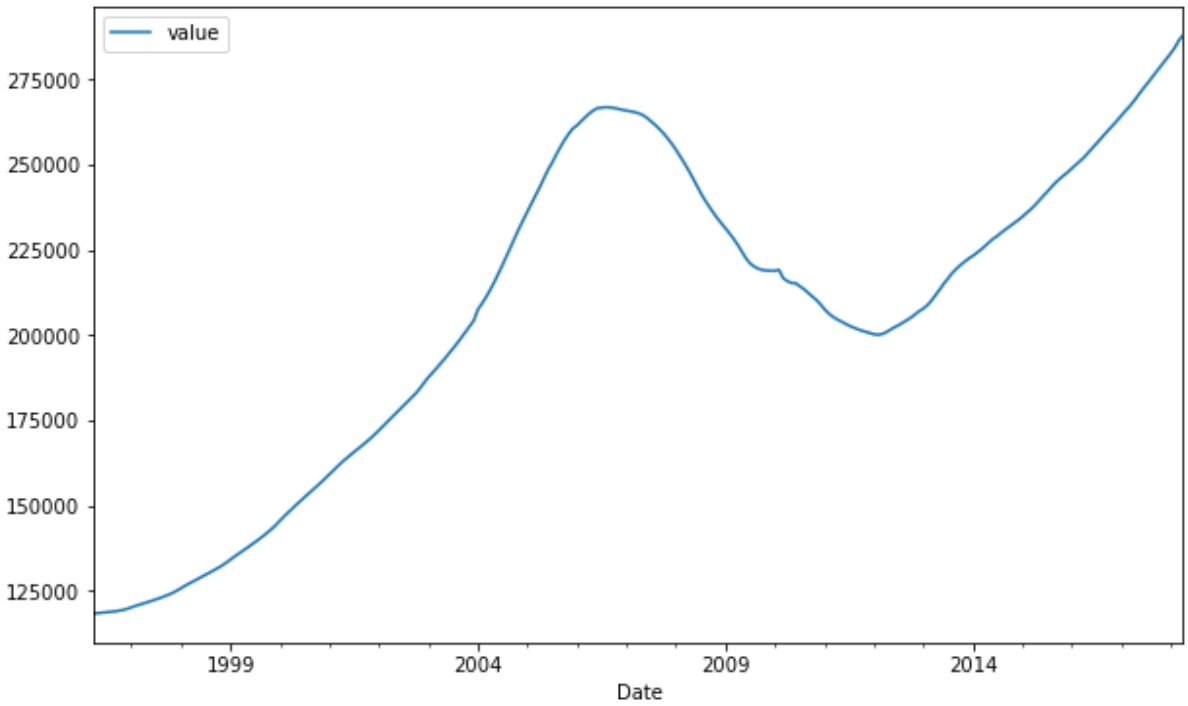

We proceeded to conduct time series analysis.

### 4.1.3 Time Series Analysis

The data did not follow a normal distribution as indicated by the histogram below.




After monthly, quarterly, semi-annual, and annual resampling, the data exhibited an upward trajectory while lacking a distinct seasonality. Below is a plot after monthly resampling.

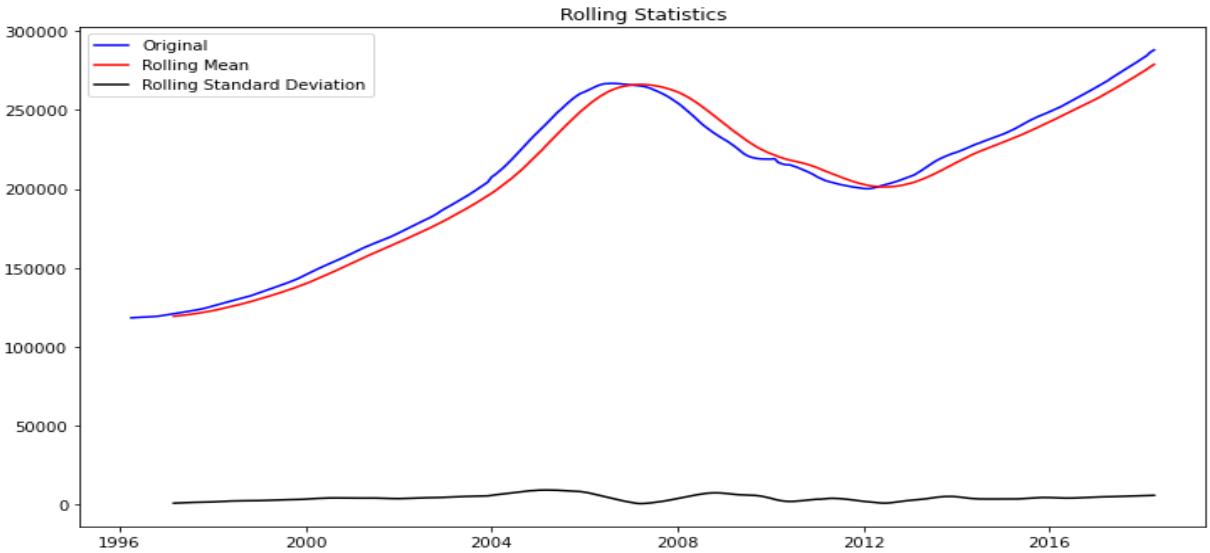



Beyond Notably, there was a conspicuous surge in house prices around 2008. This notable increase can be attributed to the global recession, which profoundly impacted the housing market in the United States during that period. The observed spike in house prices during 2008 aligns with the broader economic downturn, providing context for the temporary deviation from the overall trend in the real estate market.

### 4.1.4 Rolling Statistics

This is performed to check for any stationarity in our data. This will help us identify trends or patterns in the data, and the stationarity check involves examining whether these trends are constant over time.

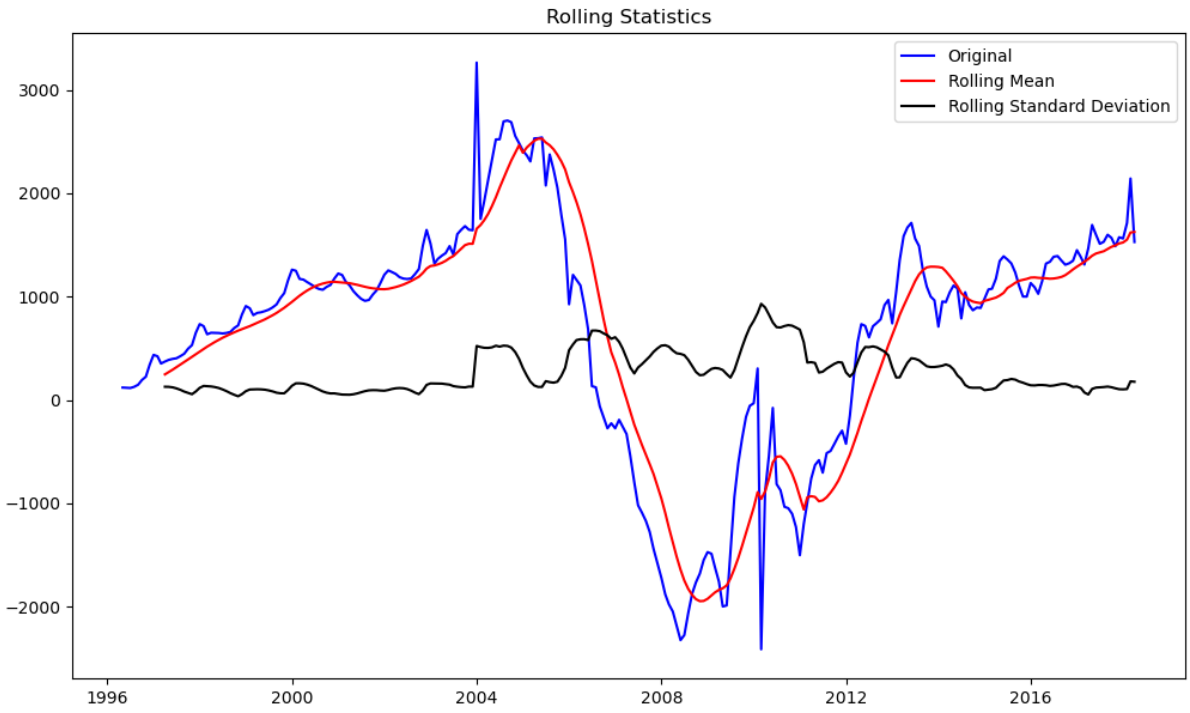
The line graph below shows that our data is not stationary from the rolling mean. To conclude, we performed a Dickey-Fuller Test, confirming our data was not stationary.



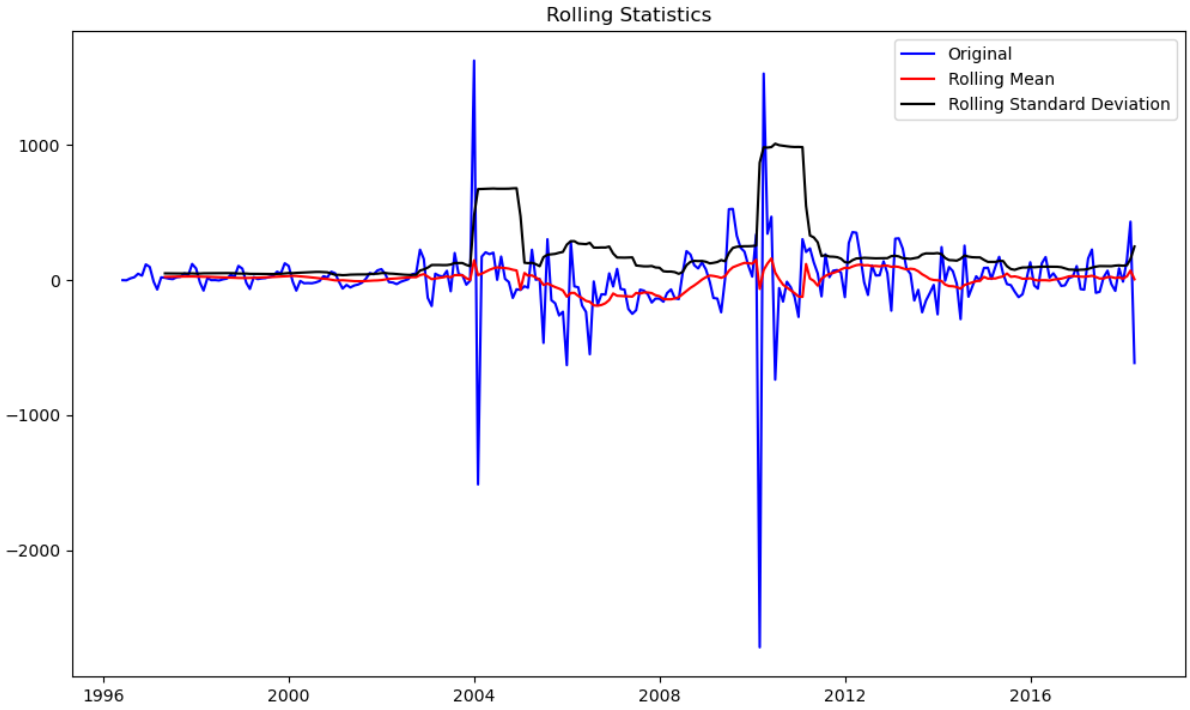
# 5 Modeling

## 5.1 Pre-processing

The first step was to check for stationarity in the data. Below is a plot of the monthly rolling mean and standard deviation, indicating that the data was not stationary. Also, the Dickey-Fuller test returned a p-value of 0.07, which is greater than 0.05, further confirming that the data was not stationary. Below is a graph for the exponential rolling mean and standard deviation.



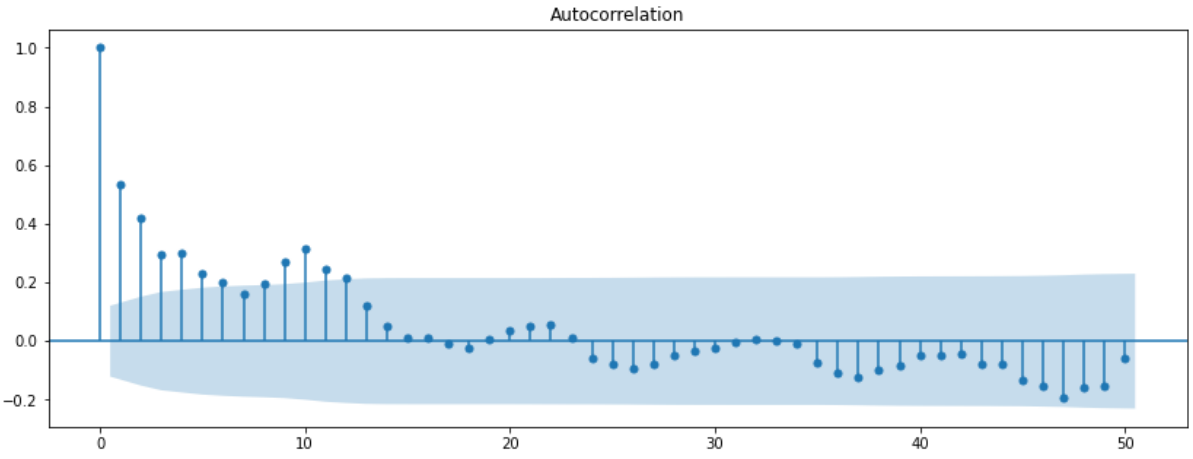

Therefore, we conducted a first-order differencing to remove the non-stationarity in the data. Below is a plot of the monthly rolling mean and standard deviation after the first-order differencing. The Dickey-Fuller returned a p-value of 0.09, further confirming that the data was still non-stationary.

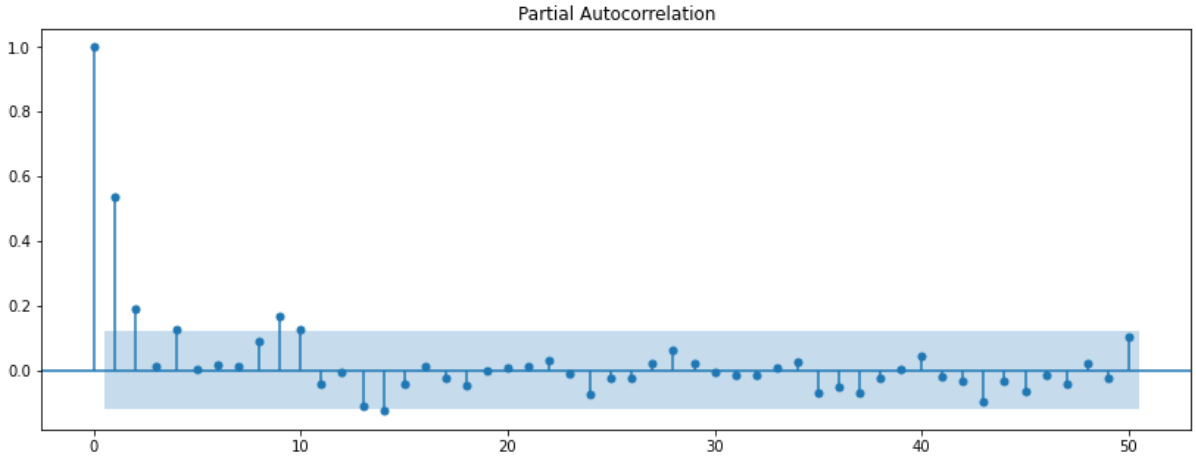



The first-order differencing was not successful in removing the non-stationarity in the data; hence proceeded to second-order differencing, which was successful in removing non-stationarity in the data. The Dickey-Fuller test on the differenced data returned a p-value of 0.00.

## 5.2 ARIMA Models

The next step was to plot the autocorrelation and partial autocorrelation plots.






From the ACF and PACF plots, the baseline model was an ARIMA(1,0,1).

We conducted a 75% split for a better model evaluation to have training and testing sets. The baseline model had an RMSE of 121.94 and an MAE of 85.94, thus requiring further improvement. The next models were ARIMA(1,0,3) and ARIMA(3,0,1). We created a function running through different combinations of p, d, and q to determine the best model returning the lowest RMSE and MAE. The ARIMA(7,1,8) was the best model, with a RMSE and MAE of 118.11 and 86.26, respectively.

Below is a table indicating the RMSE and MAE of the different ARIMA models.

| | ARIMA(1,0,1) | ARIMA (1,0,3) | ARIMA (3,0,1) | ARIMA (7,1,8) |
|---|---|---|---|---|
| RMSE | 121.94 | 122.23 | 122.24 | 118.11 |
| MAE | 85.97 | 86.27 | 86.18 | 86.26 |

The ARIMA models rely on stationarity, homoscedasticity, and normality in the data, and our data did not follow a normal distribution. Also, the model returned high AIC and BIC scores therefore we sought to examine whether the Prophet would perform better.

### 5.3 Prophet

Prophet is a time series forecasting library from Meta. We first had to reset the index in our data to build a Prophet model. We then fitted a Prophet model with default hyperparameters. However, the RMSE and MAE were 142.35 and 95.67, greater than any of the ARIMA models.

Thus, we undertook a manual hyperparameter tuning. The parameters chosen were a changepoint_prior_scale of 0.1 and the multiplicative seasonality_mode. The model had an RMSE and MAE of 122.51 and 87.13, respectively. Also, we performed a hyperparameter tuning with optuna, which indicated that the best prophet model had an RMSE of 121.87 and an MAE of 86.22.

## 6 Conclusions

From the above project, we met our objectives as follows:

1. The best zipcodes to invest based on ROI were:
    i. 11211 – Brooklyn, New York

ii. 11222 – Brooklyn, New York
iii. 11216 – Brooklyn, New York
iv. 7302 – Jersey City, New Jersey
v. 11215 – Brooklyn, New York

2. The best cities to invest in were:
    i. New York
    ii. New Jersey
    iii. Winscott
    iv. Amagansett
    v. Hartsel
3. The best states to invest were:
    i. New York
    ii. New Jersey
    iii. Colorado
    iv. California
    v. Washington DC

We also noted that the real estate prices had an upward trend, meaning that the value increased over time, although there was no clear way to determine the best time to enter the market as the data was non-stationary.

Further, the ARIMA(7,1,8) model was the best predictive model to forecast future real estate values.

# 7 Recommendations

From our findings, it is advisable to invest in Real Estate; the data showed an upward trend, indicating appreciating values over the years.

To the real estate investors, we recommend investing in the following states: New York, New Jersey, Colorado, California, and Washington DC; from the analysis, these states showed promising Returns on Investment. The best zip codes were found within the states mentioned; these are 11211 - Brooklyn, New York; 11222 - Brooklyn, New York; 11216 - Brooklyn, New York; 7302 - Jersey City, New Jersey; and 11215 - Brooklyn, New York.

As a way to mitigate risk, we recommend using the model created to forecast future real estate values.

# 8 Next Steps

To collect more data on Real Estate Values - more data will better inform the model and lead to better predictive results. Continuous model training to improve accuracy