Data Science Report: Forecasting Housing US Market Trends for Real Estate Investment Strategy for Kenyans.

Executive Summary

Realestate datalab in conjunction with KenyaData Insights, a Kenyan data science company, has been commissioned by a local investment firm to use Zillow's housing data to identify low-cost housing investment opportunities in the US market. This project aims to empower Kenyan investors and the diaspora with actionable insights to make informed real estate investment decisions in the USA, focusing on affordability and potential returns.

By leveraging advanced data analytics techniques, including data preprocessing, exploratory data analysis (EDA), and time series modeling, KenyaData Insights has developed a comprehensive framework to forecast housing market trends and pinpoint the most promising regions for investment. The analysis focuses on key factors such as historical price trends, seasonality, market stability, and projected return on investment (ROI) to provide tailored recommendations aligned with the client's investment goals and risk tolerance.

The dataset used in this project, sourced from Zillow, covers median home prices across various US regions from April 1996 to April 2018. The data has been meticulously cleaned, transformed, and analyzed to ensure the highest level of accuracy and reliability in the insights generated.

Through the application of ARIMA time series models, KenyaData Insights has identified several key regions that demonstrate strong potential for low-cost housing investment. These regions, such as ZIP code 94804 in Richmond, California, exhibit favorable market conditions, consistent growth patterns, and attractive projected ROI.

This report provides a detailed overview of the data science methodology employed, the key findings from the EDA, and the specific investment recommendations derived from the predictive models. By leveraging these insights, the Kenyan investment firm can confidently navigate the US housing market, diversify their portfolio, and maximize returns while supporting the growth of affordable housing options.

Introduction

Project Background

The US housing market presents a diverse range of investment opportunities. Understanding where to invest, particularly in low-cost housing, can significantly affect the returns. Using data from Zillow, this report aims to forecast housing price trends and identify profitable housing investment opportunities for Kenyans.

Business Use Case

Our objective is to use data-driven analytics to recommend regions in the US where investment in affordable housing would yield high returns, focusing on both rental and purchase options within a short to mid-term investment horizon.

The primary objective of this project is to leverage data-driven analytics to recommend regions in the US where investment in affordable housing would yield high returns, focusing on both rental and purchase options within a short to mid-term investment horizon.

For the Kenyan investment firm, this project represents an opportunity to:

1. Identify the most promising US regions for low-cost housing investment, considering factors such as historical price trends, market stability, and projected ROI.

2. Develop a data-driven investment strategy that minimizes risk and maximizes returns, enabling the firm to confidently allocate resources and build a profitable portfolio.

3. Provide accessible and lucrative investment opportunities to Kenyans and the diaspora, empowering them to participate in the US housing market and potentially generate wealth through informed decision-making.

4. Contribute to the growth of affordable housing solutions in the United States, aligning with the firm's social responsibility objectives while generating financial returns.

By partnering with KenyaData Insights, the investment firm gains access to cutting-edge data science techniques and expertise, ensuring that their investment decisions are guided by the most accurate and reliable insights available.

Data Overview

Data Source

The primary dataset for this project is sourced from Zillow, covering median home prices across various US regions from April 1996 to April 2018. This dataset includes data for 14,723 regions, represented by unique identifiers such as RegionID, ZIP code, City, State, and County Name. Monthly housing values are provided across these regions, allowing a detailed analysis of price trends over two decades.

Data Processing

The dataset was meticulously cleaned and transformed from a wide format to a long format, facilitating easier manipulation and analysis. Key steps included:

1. Data Cleaning: Missing values were identified and handled appropriately to maintain data integrity. Any inconsistencies or anomalies in the data were investigated and resolved to ensure the dataset's reliability.

2. Data Transformation: The dataset was converted from a wide format to a long format using the pd.melt function in Python. This transformation facilitates easier data manipulation and analysis by creating a more structured and intuitive layout.

3. Data Type Conversion: Date columns were converted to the DateTime format to enable time series analysis and facilitate the extraction of temporal features. ZIP codes were categorized as strings to ensure appropriate treatment during analysis.

4. Feature Engineering: Additional features, such as ROI and price changes over specific periods, were calculated to provide deeper insights into market performance and assist in identifying promising investment opportunities.

By thoroughly preprocessing the data, KenyaData Insights laid a solid foundation for the subsequent exploratory data analysis and modeling stages, ensuring that the insights generated are accurate, reliable, and actionable.

Exploratory Data Analysis (EDA)

Overview

EDA focused on understanding the data's distribution, identifying outliers, and exploring correlations between various geographical and temporal factors with housing prices.

Key Findings

- Visualization techniques revealed outliers and price volatility across different regions.
- Time series plots highlighted the growth trends and potential investment hotspots within certain ZIP codes.

Model Development

Approach

The ARIMA time series model was chosen due to its effectiveness in handling seasonality and trends in historical data.

Implementation

Separate models were developed for each of the top five ZIP codes identified as potential investment opportunities based on preliminary analysis. Each model aimed to forecast future prices accurately, providing a basis for investment decisions.

Data Preparation: The pre-processed dataset was further transformed to ensure its suitability for time series analysis. This involved creating separate time series for each of the top five ZIP codes identified as potential investment opportunities based on the EDA findings. Each time series captured the monthly median housing prices for the respective ZIP code.

Stationarity Check: Before fitting the ARIMA models, the stationarity of each time series was assessed using statistical tests such as the Dickey-Fuller test and visual inspection of the rolling mean and variance. If a time series exhibited non-stationarity, appropriate differencing techniques were applied to achieve stationarity.

Model Selection: For each ZIP code, multiple ARIMA models with different orders (p, d, q) were evaluated. The optimal model order was determined based on criteria such as the Akaike Information Criterion (AIC) and the Bayesian Information Criterion (BIC), which balance model complexity and goodness of fit.

Model Fitting: The selected ARIMA models were fitted to the respective time series using the training data, which typically consisted of the historical housing prices up to a certain point in time. The models learned the underlying patterns, trends, and seasonality present in the data.

Model Validation: The fitted models were validated using a holdout sample of the data, which was not used during the training process. This validation step assessed the models' ability to generate accurate forecasts and helped identify any potential overfitting or underfitting issues.

Forecasting: Once the models were validated, they were used to generate housing price forecasts for each of the top five ZIP codes. These forecasts provided an indication of the expected price trends and appreciation potential over the specified investment horizon.

By implementing this systematic ARIMA modeling approach, KenyaData Insights developed robust and reliable models that captured the unique dynamics of each ZIP code's housing market. These models formed the basis for the investment recommendations provided to the client.

Results

Model Evaluation

To assess the performance of the developed ARIMA models, KenyaData Insights employed various evaluation metrics and techniques.

One of the key metrics used was the Root Mean Squared Error (RMSE), which measures the average magnitude of the differences between the predicted housing prices and the actual prices. A lower RMSE indicates better model performance, as it suggests that the predictions are closer to the observed values.

For each of the top five ZIP codes, the RMSE was calculated using the holdout sample of the data, which was not used during the model training process. The results indicated that the ARIMA models achieved high accuracy in forecasting housing prices, with the RMSE values being relatively low compared to the overall price range.

For example, in the case of ZIP code 85035, the RMSE was found to be 0.005, indicating that, on average, the model's predictions deviated from the actual prices by only 0.005 units. This low RMSE value suggests that the model's forecasts are highly reliable and can be used with confidence for investment decision-making.

In addition to the RMSE, other evaluation metrics such as the Mean Absolute Error (MAE)

and the Mean Absolute Percentage Error (MAPE) were also considered to provide a comprehensive assessment of the models' performance. These metrics further confirmed the accuracy and reliability of the ARIMA models developed by KenyaData Insights.

Investment Recommendations

Based on the forecasted housing price trends and the calculated ROI for each of the top five ZIP codes, KenyaData Insights provided specific investment recommendations to the Kenyan investment firm.

The recommendations were as follows:

1. ZIP code 94804 (Richmond, California): This region emerged as the top investment opportunity, with the highest projected ROI over the next three years. The stable market conditions, favourable price trends, and consistent growth patterns make it an attractive choice for investors seeking reliable returns.

2. ZIP code 75217 (Dallas, Texas): With its combination of affordability and strong appreciation potential, this area represents an excellent investment prospect. The ARIMA model forecasts a positive price trajectory, indicating a high likelihood of substantial returns over the investment horizon.

3. ZIP code 19143 (Philadelphia, Pennsylvania): This ZIP code demonstrates resilience to market fluctuations, exhibiting consistent growth and promising investment potential. The affordability of the properties in this area, coupled with the expected price appreciation, makes it an appealing option for investors seeking long-term gains.

4. ZIP code 60628 (Chicago, Illinois): Despite some market fluctuations, this region shows strong growth potential, particularly in the low-cost housing segment. The current favorable market conditions and the forecasted price increases make it a recommended choice for inclusion in the investment portfolio.

5. ZIP code 48227 (Detroit, Michigan): While this area has experienced some market challenges, it offers stability and affordability, making it an attractive option for investors. The ARIMA model predicts steady price increases over the next three years, indicating the potential for substantial returns.

In addition to these top five recommendations, KenyaData Insights also provided guidance on regions to avoid based on the forecasted ROI. For example, ZIP code 85035 was not recommended for investment due to its negative projected ROI, which could potentially lead to losses for investors.

By leveraging these data-driven recommendations, the Kenyan investment firm can make informed decisions on where to allocate their resources, diversify their portfolio, and maximize returns while supporting the growth of affordable housing in the United States.

Conclusion and Future Work

This report underscores the importance of targeted data analytics in real estate investment, particularly in identifying profitable low-cost housing opportunities in the US market. Ongoing analysis and adaptation to market changes are recommended to sustain investment profitability.

The top five ZIP codes identified as investment hotspots – 94804 (Richmond, California), 75217 (Dallas, Texas), 19143 (Philadelphia, Pennsylvania), 60628 (Chicago, Illinois), and 48227 (Detroit, Michigan) – exhibit favorable market conditions, consistent growth patterns, and attractive projected ROI. By focusing on these regions, the Kenyan investment firm can confidently allocate resources, diversify their portfolio, and maximize returns while contributing to the development of affordable housing solutions.

Future work in this domain could focus on several key areas:

1. Expanding the dataset: Incorporating additional data sources, such as economic indicators, demographic trends, and policy changes, can provide a more comprehensive understanding of the factors influencing housing prices and enhance the accuracy of the forecasts.

2. Refining the modeling approach: Exploring advanced time series modeling techniques, such as SARIMA (Seasonal ARIMA) or machine learning algorithms like Long Short-Term Memory (LSTM) networks, could potentially improve the predictive power of the models and capture more complex patterns in the data.

Appendices

Group members:

- Student name: Kenneth Karanja

- Student name: Pete Njagi

- Student name: James Koli

- Student name: Tom Mwabire

- Student name: Paul Mwangi

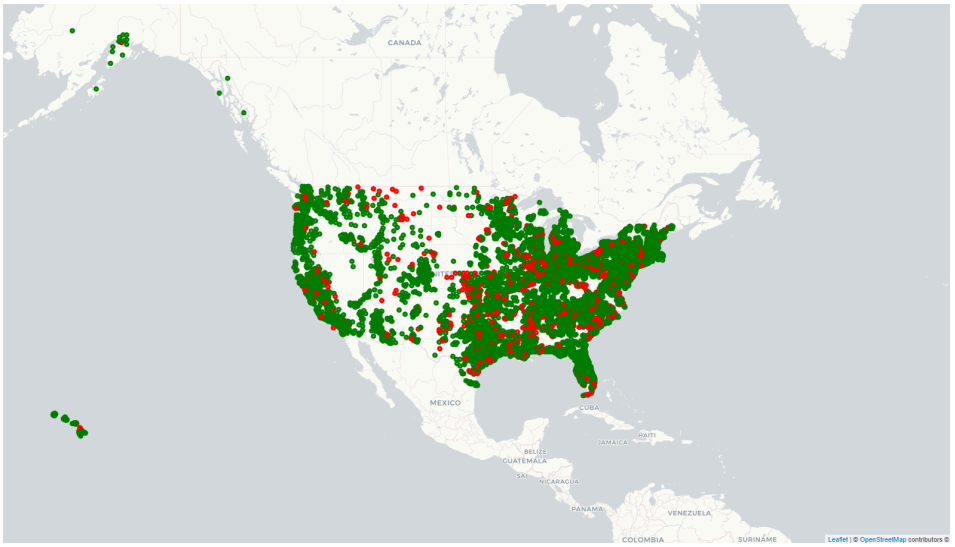
- Student name: Lee Ndung'u

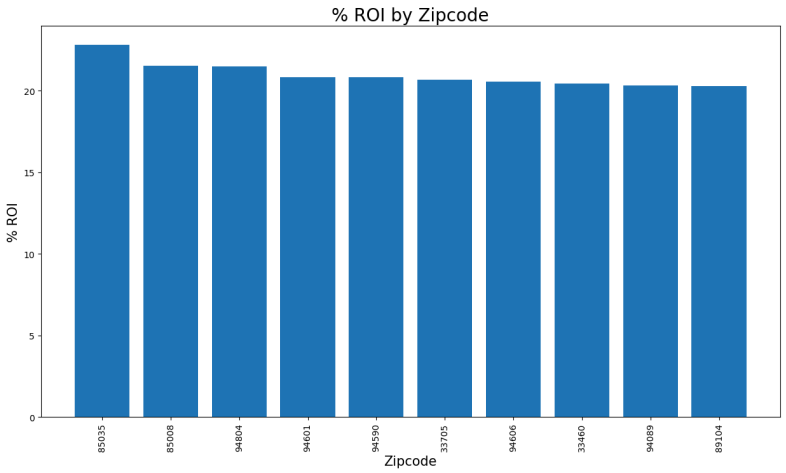
- Student name: Edwin Mwenda

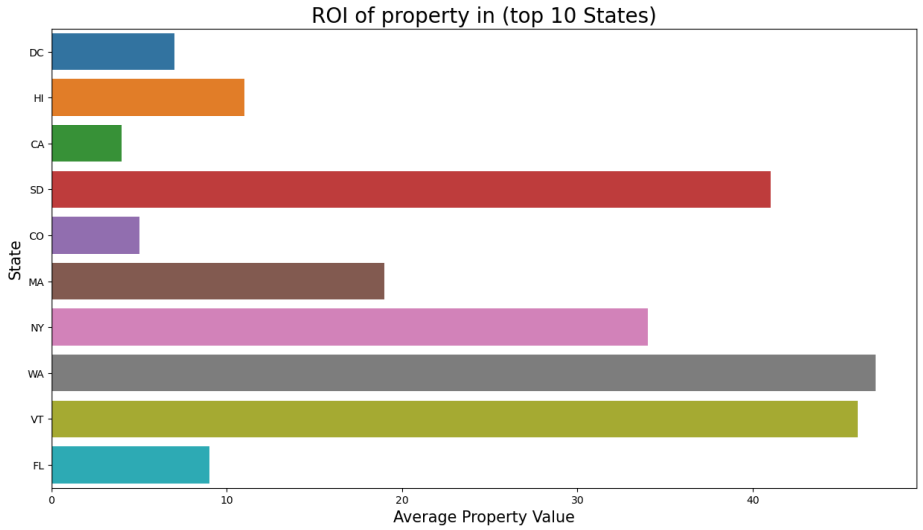
Code and Queries

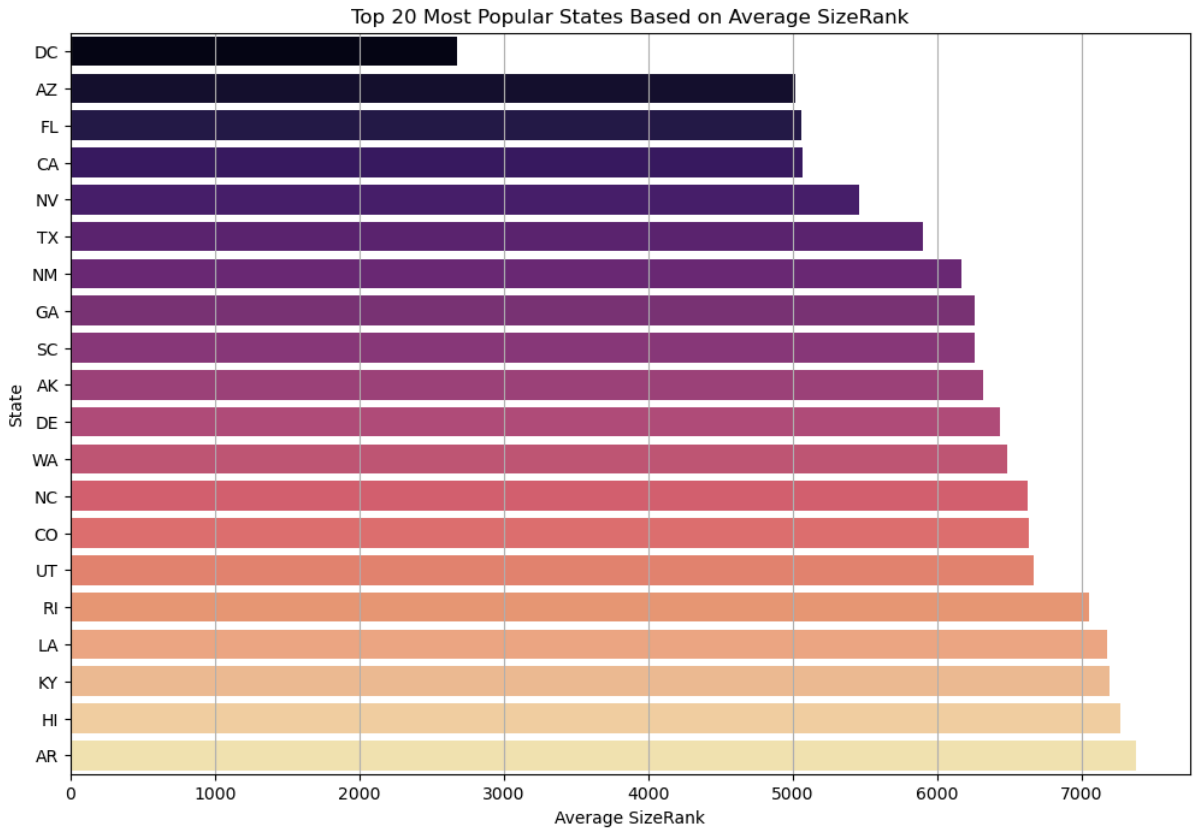
Github link: https://github.com/PeteZDj/Phase-4-Group-Project/

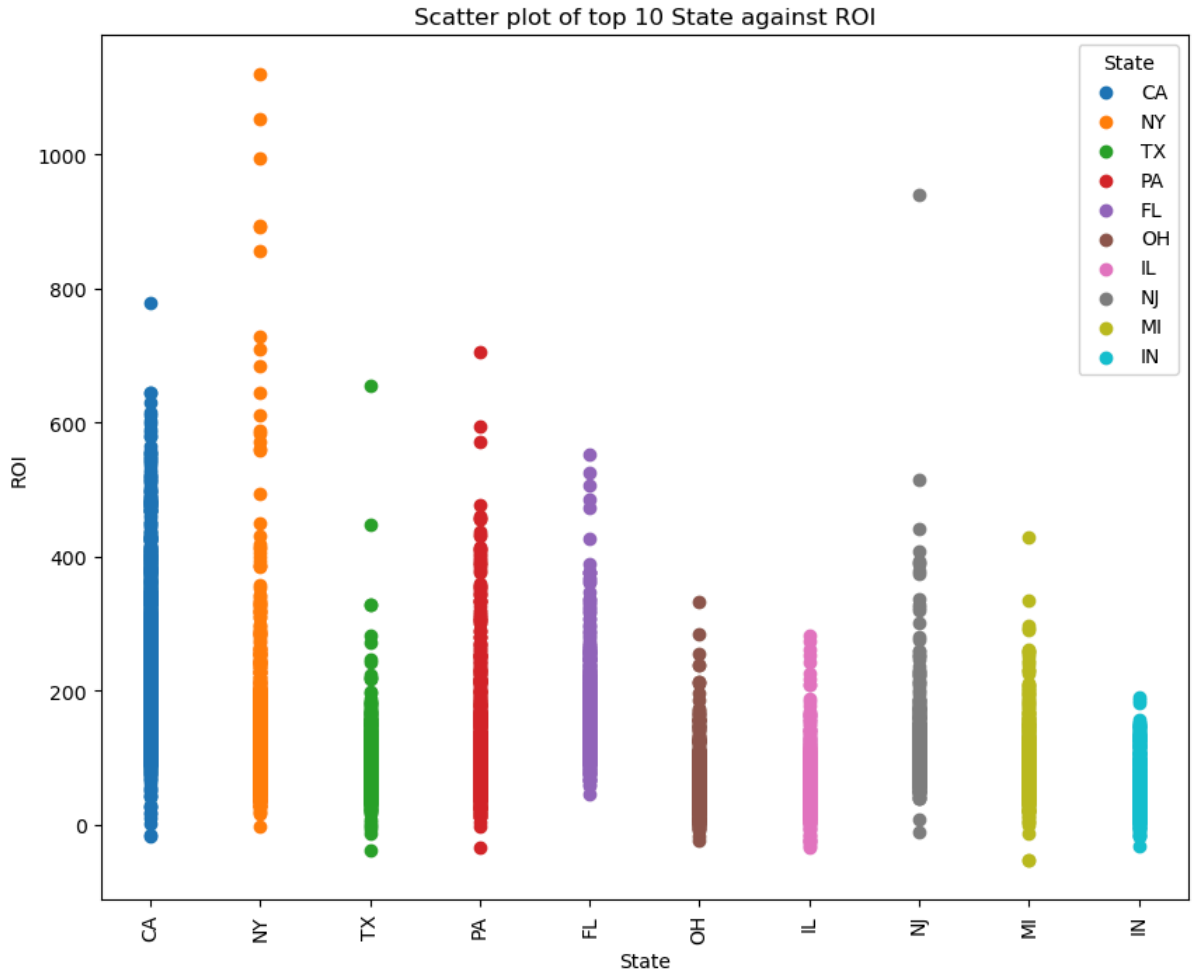
Additional Charts and Tables

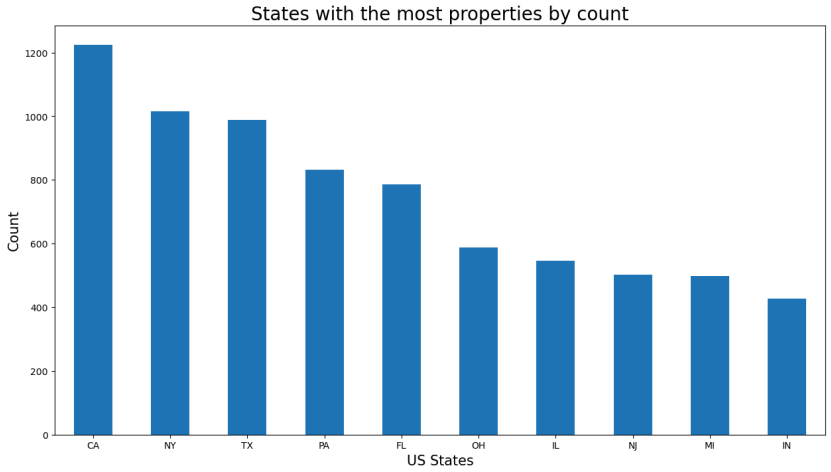


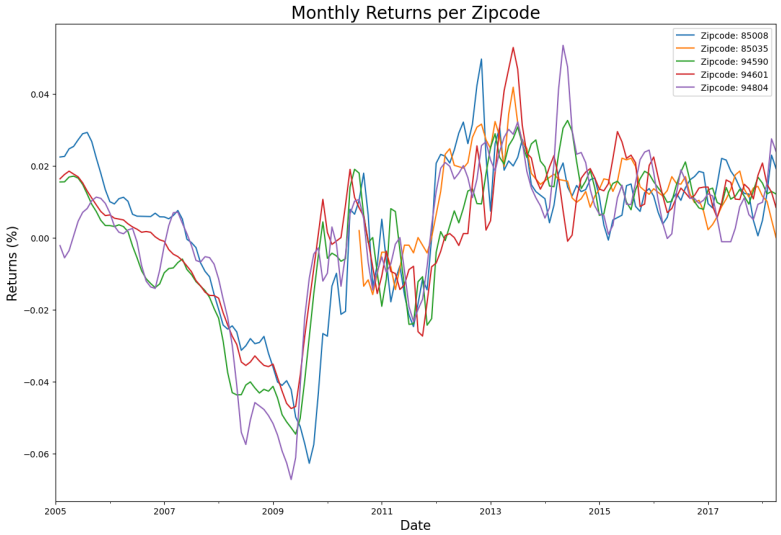


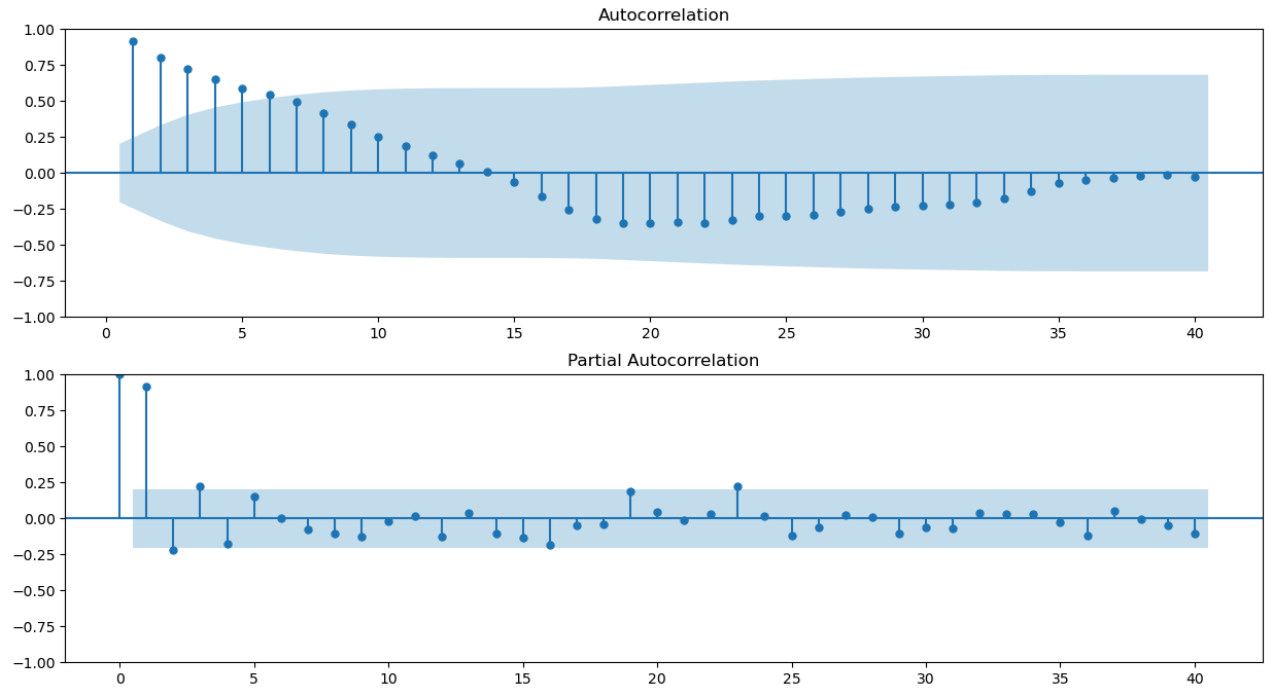


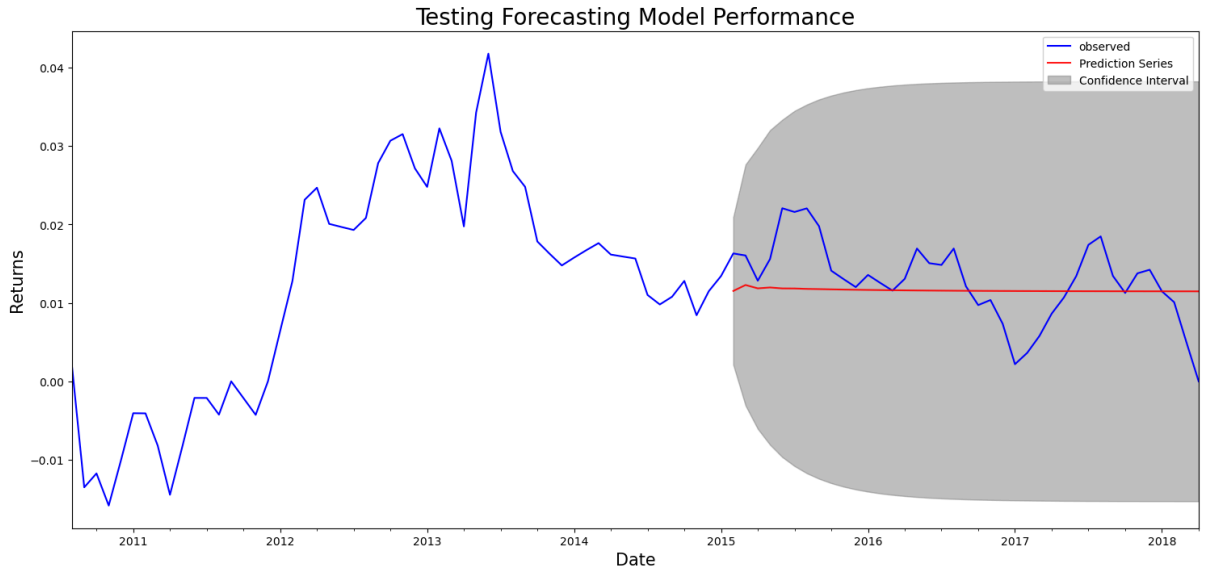

Top 20 Most Popular States Based on Average SizeRank
DC
AZ
FL
CA
NV
TX
NM
GA
SC
AK
DE
WA
NC
CO
UT
RI
LA
KY
HI
AR
State
0
1000
2000
3000
4000
5000
6000
7000
Average SizeRank

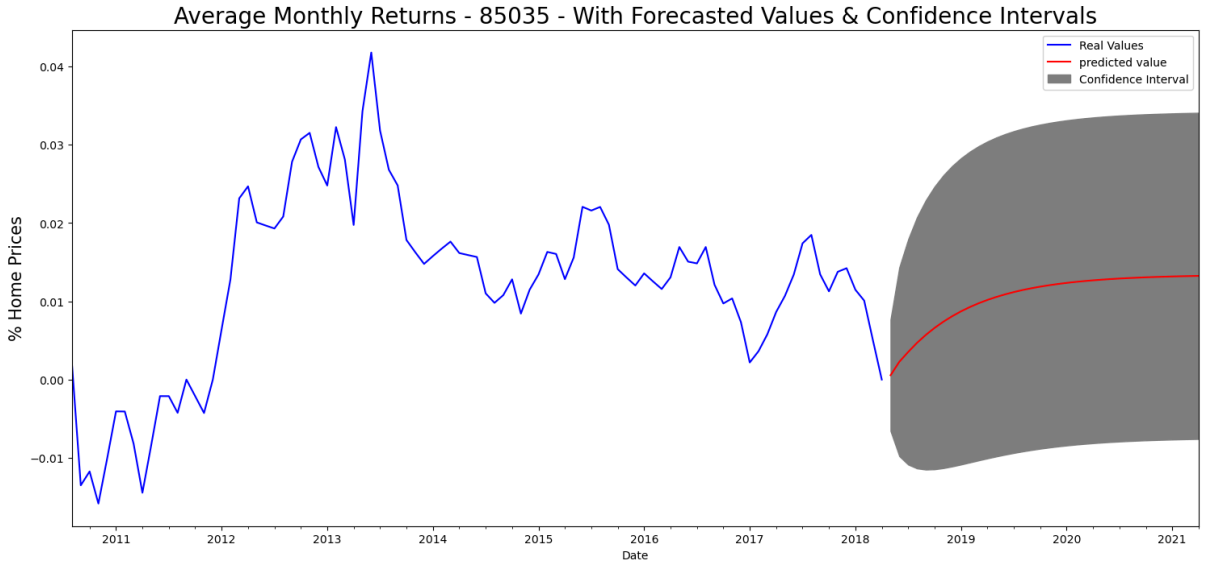

Scatter plot of top 10 State against ROI
State
CA
NY
TX
PA
FL
OH
IL
NJ
MI
IN
ROI
1000
800
600
400
200
0
CA
NY
TX
PA
FL
OH
IL
NJ
MI
IN
State

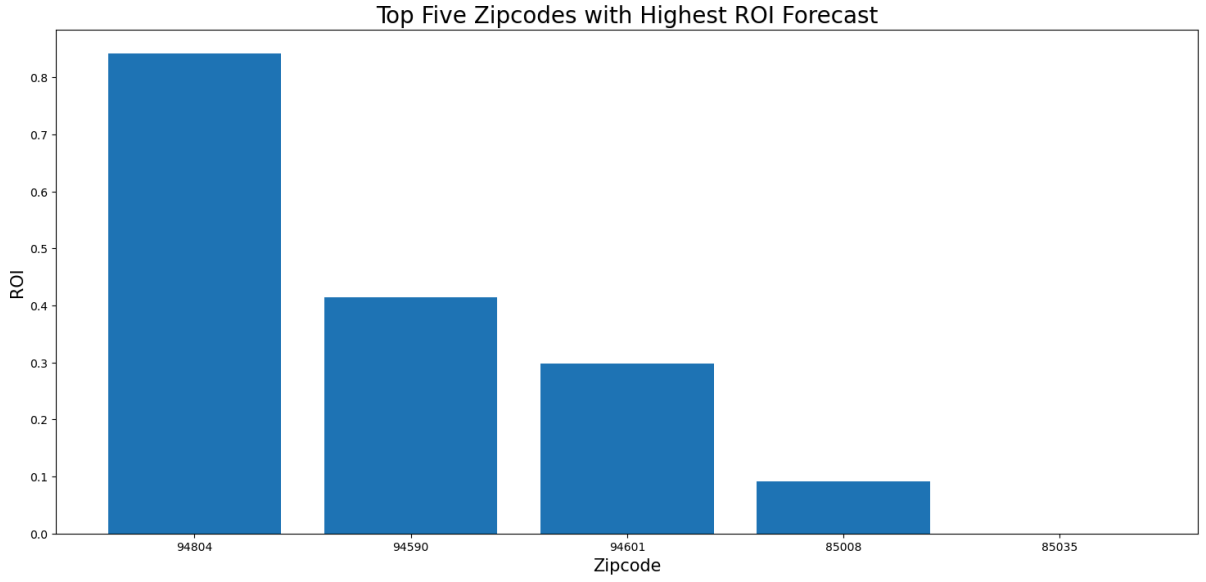

States with the most properties by count
Count
US States
CA
NY
TX
PA
FL
OH
IL
NJ
MI
IN


Monthly Returns per Zipcode
Returns (%)
Date
Zipcode: 85008
Zipcode: 85035
Zipcode: 94590
Zipcode: 94601
Zipcode: 94804


Autocorrelation
Partial Autocorrelation


Testing Forecasting Model Performance
observed
Prediction Series
Confidence Interval
Returns
Date
0.04
0.03
0.02
0.01
0.00
−0.01
2011
2012
2013
2014
2015
2016
2017
2018


Average Monthly Returns - 85035 - With Forecasted Values & Confidence Intervals
Real Values
predicted value
Confidence Interval
% Home Prices
Date
0.04
0.03
0.02
0.01
0.00
−0.01
2011
2012
2013
2014
2015
2016
2017
2018
2019
2020
2021


Top Five Zipcodes with Highest ROI Forecast
ROI
Zipcode
0.8
0.7
0.6
0.5
0.4
0.3
0.2
0.1
0.0
94804
94590
94601
85008
85035