Dream House International, Real Estate Time Series Analysis

Group members:

1. Lynn Achieng

2. Josephine Maro

3. Troye Gilbert

4. Edmund Nyaribo

OVERVIEW

This project involves building a time series model using Zillow data from Zillow Research to aid real estate investors in making informed investment decisions. The dataset comprises property information, and the project encompasses data preprocessing, time series transformation, exploratory data analysis, model selection, training, and evaluation. The model's objective is to forecast property price trends, which will be presented to investors through a user-friendly interface. Recommendations on where to invest will be provided based on these predictions and supplemented with additional insights from EDA. The project also includes documentation, deployment, maintenance, and a feedback loop to continuously enhance the model's accuracy and relevance to real estate investment needs.

BUSINESS UNDERSTANDING

Following its success in the real estate business over the years, Dream House International, a real estate agency, has provided housing solutions to many of the people living in the United States. Dream House International now wants to expand its reach further in the existing states but want to focus in the best ones for further investment. Using the Zillow dataset, this project aims to determine the best 5 investment opportunities for Dream House International.

Objectives

1. To identify the top 5 best states for Dream House International to invest in.

2. To identify the top 5 zipcodes with the highest ROI.

3. To forecast future real estate prices for the zip codes over various time horizons.

DATA UNDERSTANDING AND PREPARATION

For data understanding and preparation, begin by thoroughly exploring the Zillow dataset, checking for missing values, and addressing outliers.

The dataset, originally consists of 14723 rows and 272 columns.

- RegionID - This is unique identifier for the Region.
- SizeRank - This is the ranking based on the region size.
- RegionName – It has the code of the Region.
- StateName - State.
- City - This column provides the City Name.
- Metro - Provides the name of the metro city around the Region.
- County Name - County Name.
- Months Column - These Columns contains the value of houses in the Region for every month.

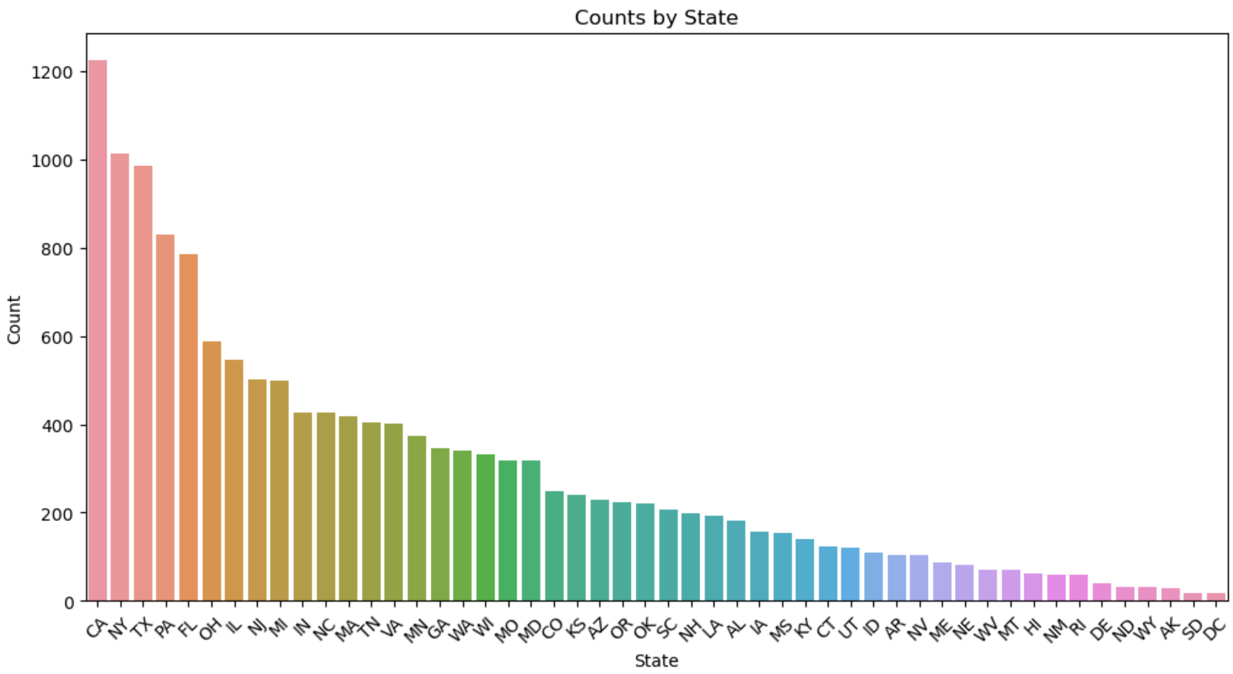
Data Preparation and Analysis

The data set was then prepared for analysis by changing the dataset from wide format to long format, transforms the dataset into a time series format that is best for performing analysis on. After the transformation the dataset will have 3899740 rows and 13 columns. Additional columns were added that help us in understanding the data well.

Data cleaning was also done and the findings were as follows:

- Percentage of missing values: 3.94% **-** The missing values were replaced with the word “missing
- There were no duplicate values

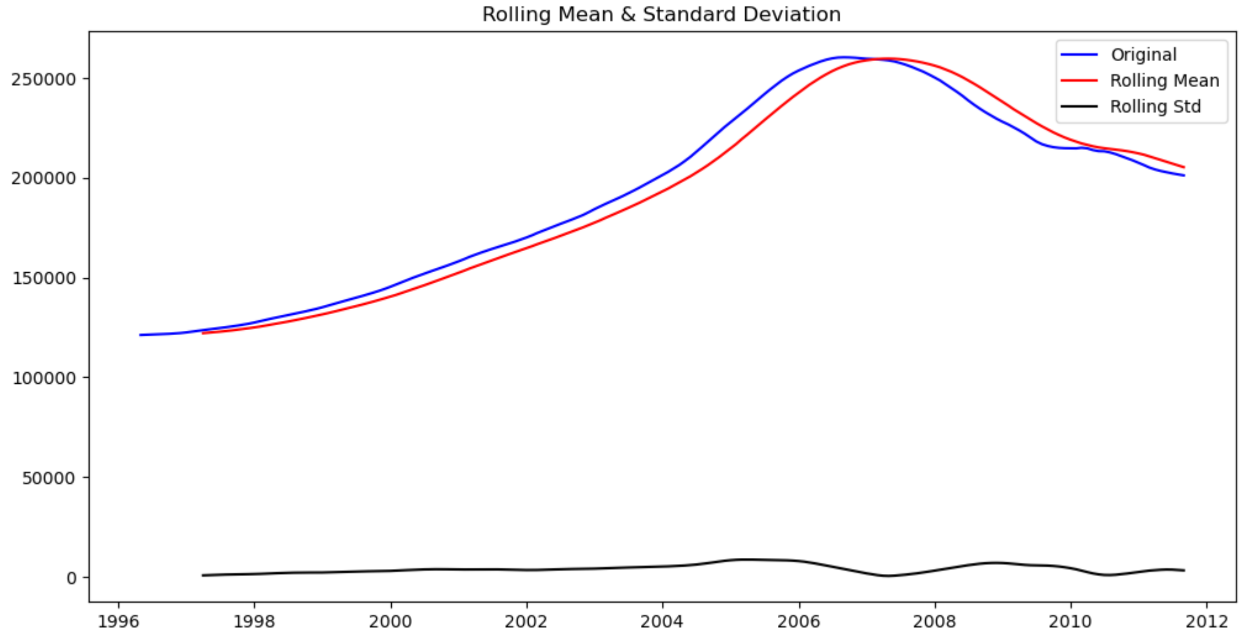
Analysis was done to the data set and we found that California was the state that was most frequently invested in, followed by New York, Texas, Pennsylvania and Florida.



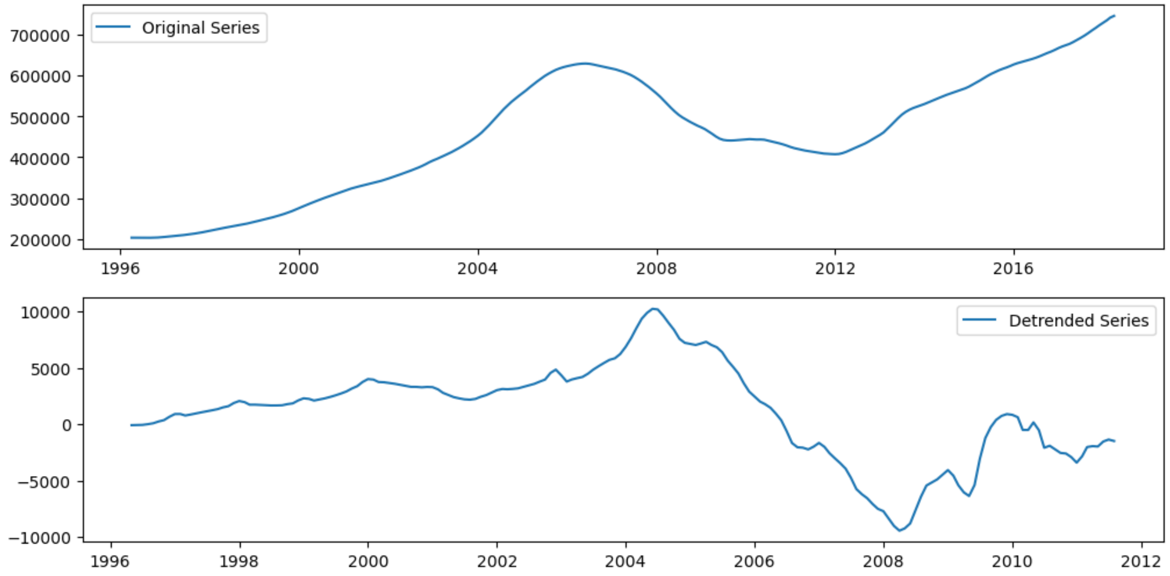
We now focus on these states in order to get the best five zip codes that Dream House International can invest in.

We also used moving averages and seasonality patterns, to capture temporal trends and eliminated the trends and seasonality using the differencing technique. Additionally, split the data into training and testing sets, reserving the most recent data for validation. This will create a clean, structured dataset ready for time series modeling and forecasting.

Plot showing original series, rolling mean and standard devation:


Rolling Mean & Standard Deviation
Original
Rolling Mean
Rolling Std
250000
200000
150000
100000
50000
0
1996
1998
2000
2002
2004
2006
2008
2010
2012

Original plot and detrended plot:

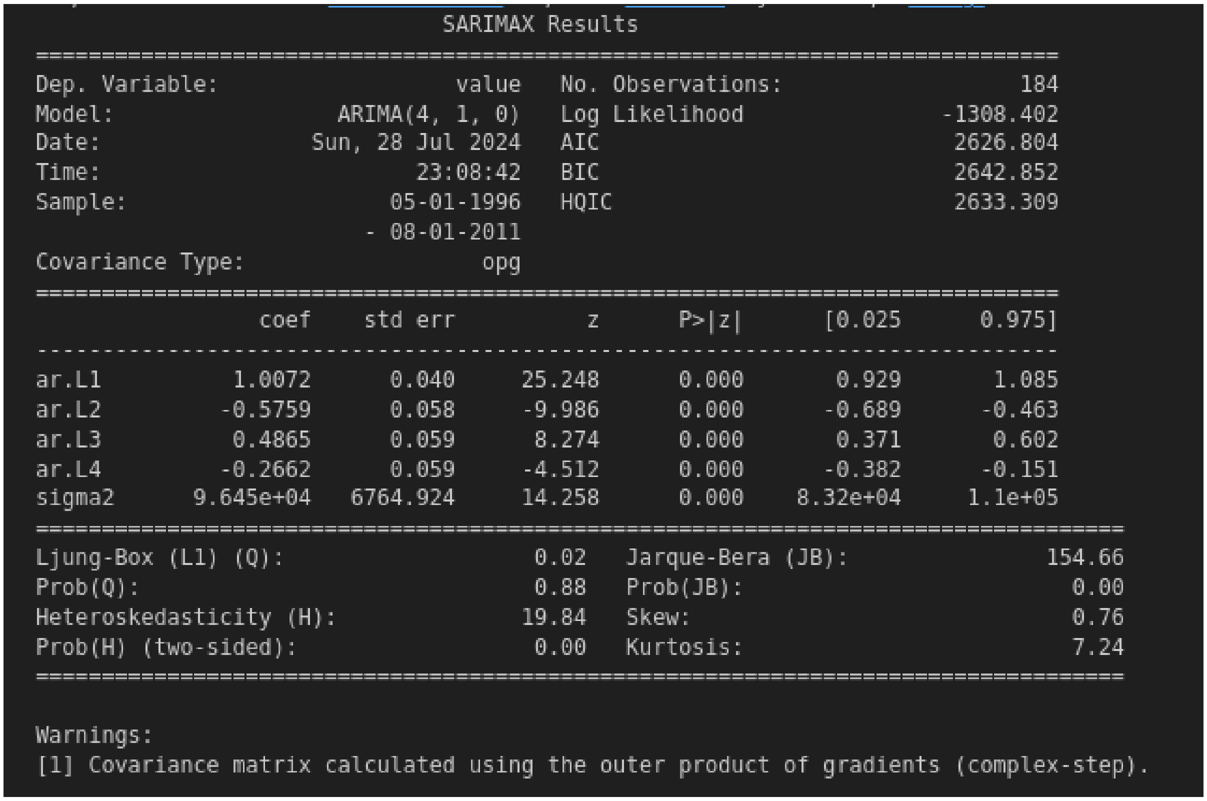


MODELLING

The following models were used during the developing of a machine learning model that can predict what are the top 5 best zip codes in the state of California to invest in. This model will provide predictions and insights into property price trends over time, helping investors identify regions and cities with potential for price appreciation. Ultimately, the project aims to empower investors with data-driven tools that enhance their understanding of real estate market dynamics, enabling them to make more strategic and profitable investment choices.

- The base model is simple and serves as a benchmark.
- The ARIMA model involves identifying the order of differencing, autoregression, and moving average components to handle non-seasonal data.

Sample of ARIMA results:



```
                               SARIMAX Results
==============================================================================
Dep. Variable:                  value   No. Observations:                  184
Model:                 ARIMA(4, 1, 0)   Log Likelihood               -1308.402
Date:                Sun, 28 Jul 2024   AIC                           2626.804
Time:                        23:08:42   BIC                           2642.852
Sample:                    05-01-1996   HQIC                          2633.309
                         - 08-01-2011
Covariance Type:                  opg
==============================================================================
                 coef    std err          z      P>|z|      [0.025      0.975]
------------------------------------------------------------------------------
ar.L1          1.0072      0.040     25.248      0.000       0.929       1.085
ar.L2         -0.5759      0.058     -9.986      0.000      -0.689      -0.463
ar.L3          0.4865      0.059      8.274      0.000       0.371       0.602
ar.L4         -0.2662      0.059     -4.512      0.000      -0.382      -0.151
sigma2      9.645e+04   6764.924     14.258      0.000    8.32e+04     1.1e+05
===================================================================================
Ljung-Box (L1) (Q):                   0.02   Jarque-Bera (JB):               154.66
Prob(Q):                              0.88   Prob(JB):                         0.00
Heteroskedasticity (H):              19.84   Skew:                             0.76
Prob(H) (two-sided):                  0.00   Kurtosis:                         7.24
===================================================================================

Warnings:
[1] Covariance matrix calculated using the outer product of gradients (complex-step).
```

- The SARIMA model extends ARIMA by incorporating seasonality, improving forecast accuracy for seasonal time series data.

The first models, for the base, ARIMA and SARIMA, we used the dataset based on the carlifornia state in getting the performance of the models. The California dataset will serve as a bench mark for the subsequent models that will be built using the states. Then we will ascertain the performance of the models by comparing them with the base model.

EVALUATION

In this project, two evaluation metrics were used to assess the performance of different models in predicting the top 5 best zip codes to invest in.

- MSE (Mean Squared Error): Measures the average of the squares of the errors.
- RMSE (Root Mean Squared Error): Square root of MSE, providing error magnitude in the same units as the original data.
- Compare multiple ARIMA models using AIC (Akaike Information Criterion) and select the model with the lowest AIC.

When doing evaluation of the models, we got the RMSE value of the SARIMA model to be lower than that of the ARIMA model, which is expected. This is to show that the SARIMA Model outperforms the ARIMA model and does better in predicting future possible outcomes than the ARIMA model.

CONCLUSION

Based on the evaluation metrics (MSE and RMSE), the SARIMA outperformed other models with the lowest RMSE. This model was selected as the final model for predicting best zip codes to invest in. The following are the results of the metrics used during the modeling:

- Base Model: Serves as a baseline, expected to have higher MSE and RMSE compared to ARIMA and SARIMA.
- ARIMA Model: Expected to perform better than the base model by capturing non-seasonal patterns.
- SARIMA Model: Expected to perform the best if there is seasonality in the data, by capturing both non-seasonal and seasonal patterns.

Based on the evaluation metrics (MSE and RMSE), the SARIMA outperformed other models with the lowest RMSE. This model was selected as the final model for predicting best zip codes to invest in.The following are the results of the metrics used during he modeling:

The following states were selected as the best 5 to invest in with their corresponding zip codes

- California, Stinson Beach
- New York, Wainscott
- Texas, Luling
- Pennsylvania, Porter
- Florida, Rotonda West

Future Work

In future, we could extend our research to explore other cities and enhance our model by gaining a deeper insight into the data and trends. This would involve continuously refining and optimizing our predictive model.

With additional time, which is a resource-intensive commitment, we could analyze return data for each ZIP code individually. This approach would allow us to select ZIP codes based on predicted returns and minimized losses, rather than relying solely on historical data.

Moreover, we could explore alternative time series forecasting methods to evaluate their effectiveness.

Lastly, incorporating more up-to-date data would ensure that our predictions are more accurate and reflective of the current real estate landscape.

For More Information

Please review my full analysis in my Jupyter Notebook and my Presentation