EXAMINING REAL ESTATE INVESTMENTS FOR BRIMAR PROPERTY AGENCIES, USA¶

Group Members:

1. **Maureen Oketch**
2. **Peter Kinyanjui**
3. **Rwenji Murengaita**
4. **Brian Korir**

Business Understanding

- Background
  - Real estate markets undergo gradual shifts over time, with abrupt changes influenced by various factors.
  - Brimar Property Agencies, a newcomer, seeks accurate information on real estate market performance.
  - Project aims to analyze trends in housing prices to provide insights for investor decisions.
- Project Question
  - Explore: Which five zip code areas offer the most favorable investment opportunities for Brimar Investment Agency?

Project Objectives

- To define criteria for evaluating optimal zip codes, considering ROI, regional size ranking, associated risks, and property value volatility.
- To identify patterns, trends, and relationships within the data, including historical price changes, ROI variation, and risk assessment.
- To develop predictive time series model for forecasting ROI.
- To evaluate and select suitable model based on performance and accuracy.
- To provide recommendations for property investments based on key factors.

Data Understanding -Loading the Data/Filtering for Chosen Zip-codes

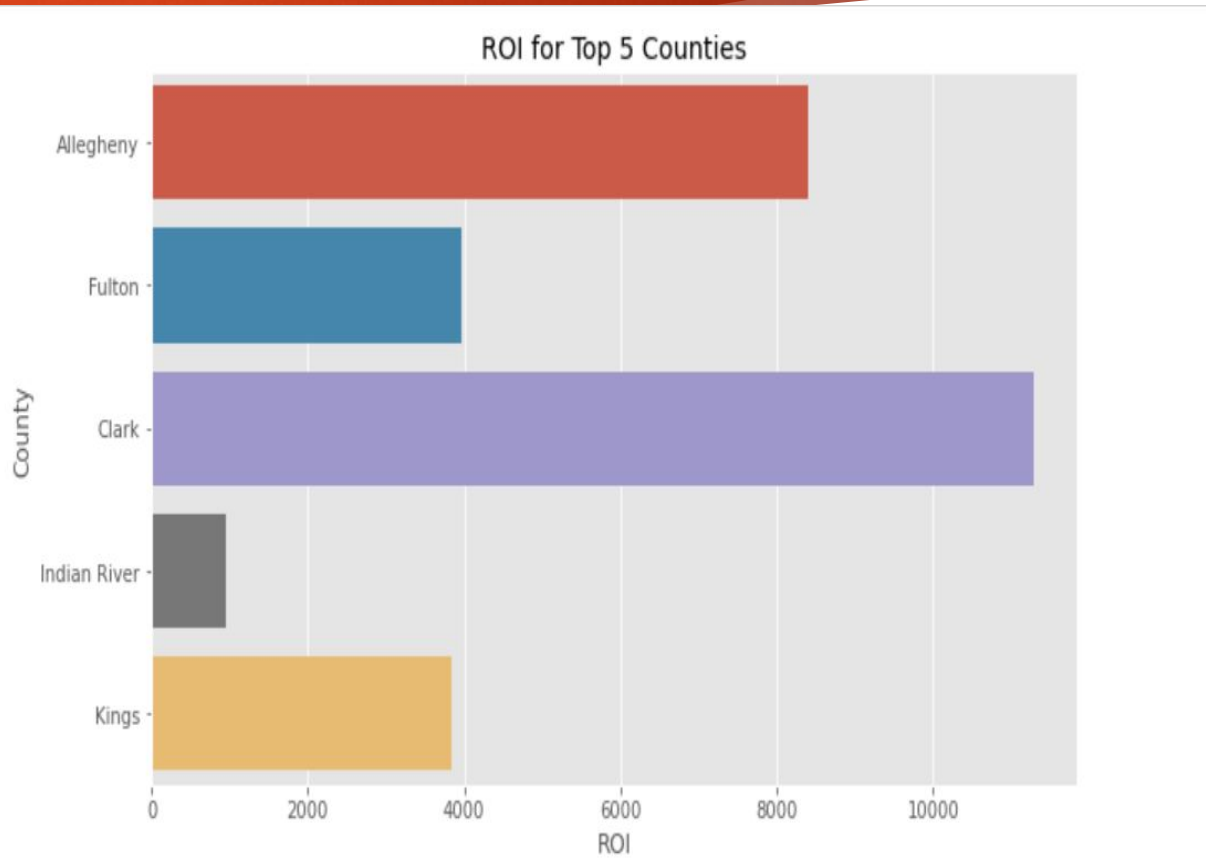
- Data loaded from 'zillow_data.csv' containing median housing sales values for various zip codes from April 1996 to April 2018.
- Initial exploration reveals 14723 rows and 272 columns.
- Metro and price columns have missing values, with other columns having no missing data.
- **Data Summary**
- Data contains geographical information (RegionID, RegionName, City, State, Metro) and housing sales values for each month.
- Various data types include float, integer, and object.
- No duplicates found, but 220 columns have missing values.

Data Preparation

- Reshaping data to long format for analysis.
- Slicing data to most recent 10 years for relevance.
- Handling missing values: Filled 'Metro' missing values with 'Missing' and used linear interpolation for 'Price' column.
- Exploratory Data Analysis (EDA)
- Visualize trends in housing prices over time for selected zip codes.
- Identify correlations between different variables such as location, size rank, and housing prices.

Feature Engineering

- Calculate and plot ROI function to compute Return on Investment (ROI) for each county.
- Visualize the top 5 counties with the highest ROI, aiding in effective resource allocation and investment decision-making.

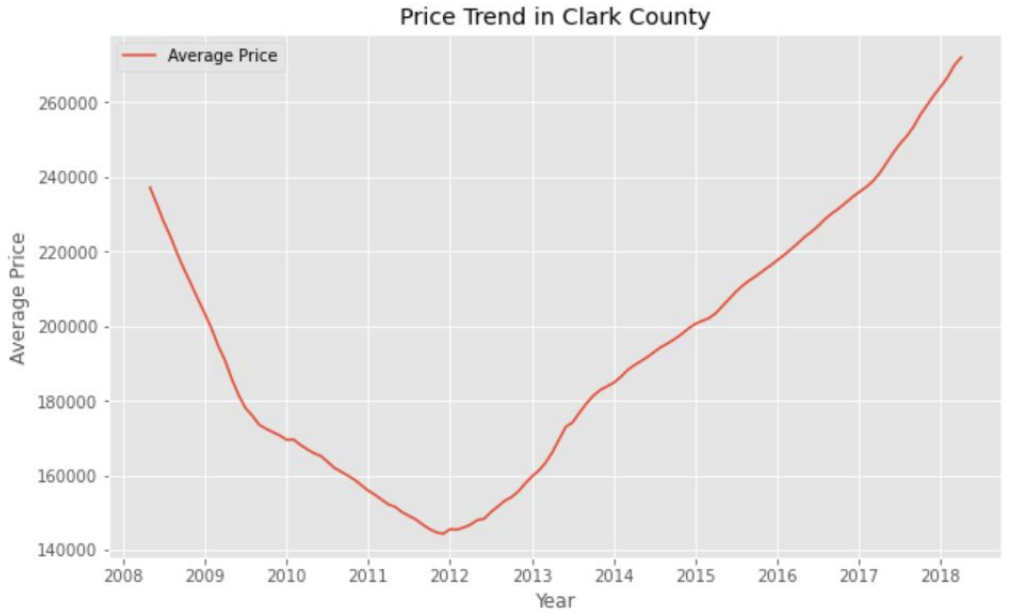


Observations on ROI for Top Counties

- Clark County emerges as the most profitable, with Allegheny, Fulton, Kings, and Indian River following closely.
- Diverse county performance underscores the need for refined investment strategies.
- ROI analysis enables investors to focus on areas with the greatest profit potential.

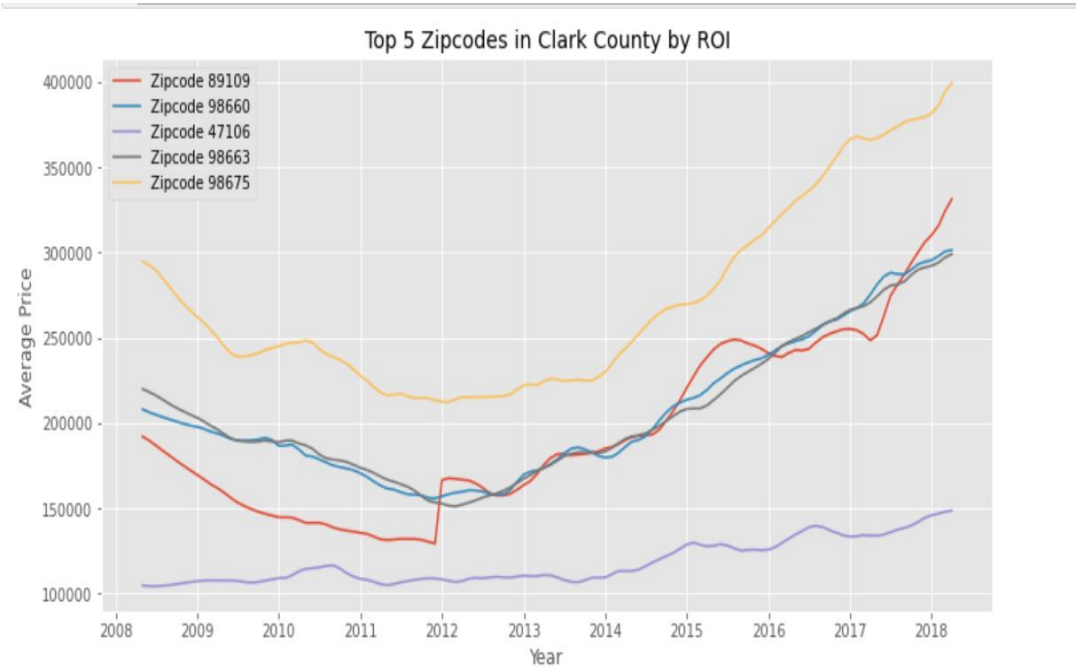
Exploratory Data Analysis for Clark County

- Analyze historical price trends in Clark County using the plot_clark_county_trend function.
- Identify patterns and inflection points to optimize investment timing.



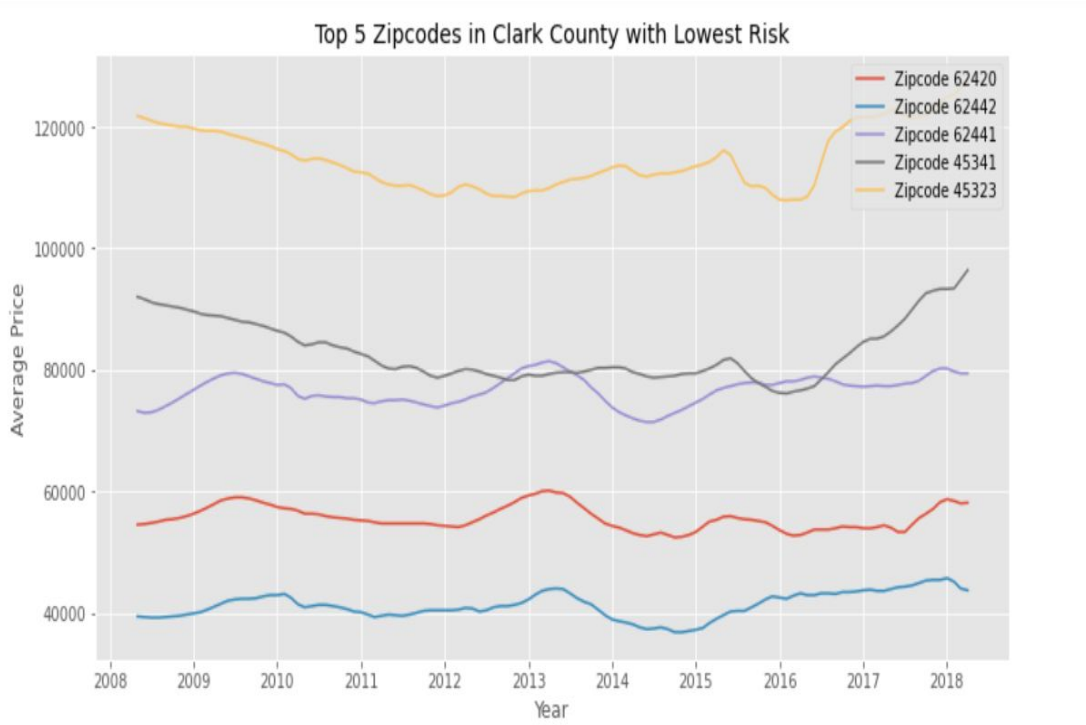
Top 5 Zip Codes within Clark County

- Identify top-performing zip codes using the top_roi_zipcodes_in_clark function.
- Prioritize investments based on ROI, directing capital to high-yield areas.



Lowest Risk Zip Codes in Clark County

- Assess risk factors with the lowest_risk_zipcodes_in_clark function.
- Build a resilient portfolio by balancing ROI opportunities with stability and security.



ROI Analysis for Top Zip Codes

- To delve deeper into the analysis, the top zip codes with the highest ROI are selected for further examination.
- This focused approach allows investors to gain a comprehensive understanding of the market dynamics driving the exceptional performance of these zip codes.
- By studying the underlying factors contributing to their success, investors can replicate strategies and insights to optimize their investment outcomes across other properties and regions.

Checking for Stationarity

- **ADF Test and Differencing**
- Stationarity is a critical assumption in time series analysis.
- The function time_series_indexing applies time series indexing and differencing techniques to assess and enhance the stationarity of the data.
- By ensuring the stationarity of the dataset, investors can apply robust time series models and make more accurate forecasts, enhancing their decision-making processes and reducing investment risks.

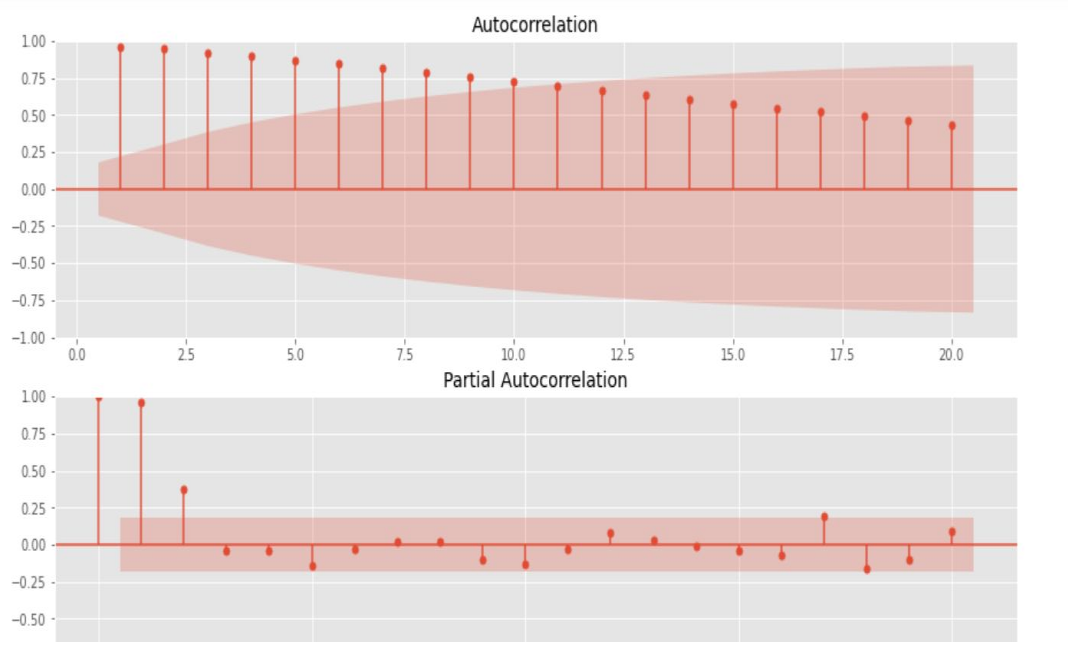
Stationarity Testing Results

- The stationarity of the combined Clark County zip code data is confirmed through statistical tests, such as the Augmented Dickey-Fuller (ADF) test.
- With a significant p-value (p-value < 0.05), the data exhibit stationarity, providing a solid foundation for reliable modeling and forecasting.

```
ADFuller test p-value for combined Clark County zip codes:
p-value: 0.0
Reject the null hypothesis. Data is stationary.
```

Autocorrelation and Partial Autocorrelation (ACF & PACF) Analysis

- Autocorrelation and partial autocorrelation analyses offer valuable insights into the underlying patterns and dependencies within the time series data.
- By examining these plots, investors can identify potential trends, seasonality, and lag effects that influence the data's behavior.



Observations from ACF & PACF Analysis

- The ACF and PACF plots reveal intriguing patterns and correlations within the time series data:
- **Autocorrelation Insights:** The Autocorrelation graph displays non-random data patterns, indicating the presence of trends or seasonality.
- **Partial Autocorrelation Patterns:** The Partial Autocorrelation graph highlights specific lag effects, such as strong correlations at lag 1, providing valuable insights into the data's underlying structure.

Finding Best ARIMA Model

Use of auto_arima:

- Employed auto_arima to determine the optimal p, d, q values for the ARIMA model.
- Parameters:
  - **modeling_data**: Time series data.
  - **trace=True**: Enable tracing of the fitting process.
  - **error_action='ignore'**: Ignore errors to prevent interruption.
  - **suppress_warnings=True**: Suppress warnings for cleaner output.
  - **stepwise=True**: Conduct stepwise search for optimal parameters.

- **Best ARIMA Model:**
  - ARIMA(0,2,2)(0,0,0)[0]
  - Significance:
    - AIC=-493.430
    - Indicates a relatively good fit for the model.
  - Total fit time: 1.063 seconds

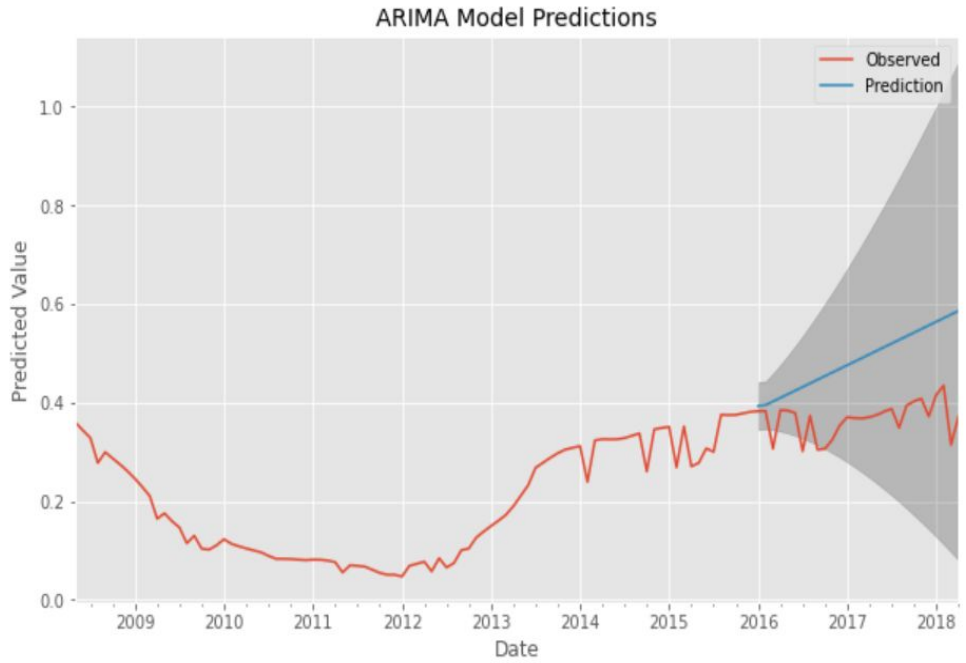
```
Performing stepwise search to minimize aic
 ARIMA(2,2,2)(0,0,0)[0]             : AIC=-490.837, Time=0.18 sec
 ARIMA(0,2,0)(0,0,0)[0]             : AIC=-350.072, Time=0.05 sec
 ARIMA(1,2,0)(0,0,0)[0]             : AIC=-425.100, Time=0.02 sec
 ARIMA(0,2,1)(0,0,0)[0]             : AIC=-465.261, Time=0.06 sec
 ARIMA(1,2,2)(0,0,0)[0]             : AIC=-492.637, Time=0.12 sec
 ARIMA(0,2,2)(0,0,0)[0]             : AIC=-493.430, Time=0.13 sec
 ARIMA(0,2,3)(0,0,0)[0]             : AIC=-492.926, Time=0.10 sec
 ARIMA(1,2,1)(0,0,0)[0]             : AIC=-490.578, Time=0.08 sec
 ARIMA(1,2,3)(0,0,0)[0]             : AIC=-490.931, Time=0.15 sec
 ARIMA(0,2,2)(0,0,0)[0] intercept   : AIC=-491.824, Time=0.16 sec

Best model:  ARIMA(0,2,2)(0,0,0)[0]
Total fit time: 1.063 seconds
```

ARIMA Model Development

- **Modeling Approach:**
  - Split data into training and testing sets.
  - Train ARIMA model on training data.
- **Model Summary:**
  - Coefficients:
    - ma.L1: -1.3634
    - ma.L2: 0.4635
    - sigma2: 0.0006
  - Observations:
    - No. Observations: 93
    - Log Likelihood: 206.128
    - AIC: -406.257
    - BIC: -398.724
- **Diagnostic Plots:**
  - Analyzed diagnostic plots for model evaluation.
    - Standardized Residual for “M” Plot
    - Histogram Plus Estimated Density Plot
    - Normal Q-Q Plot
    - Correlogram or ACF Plot

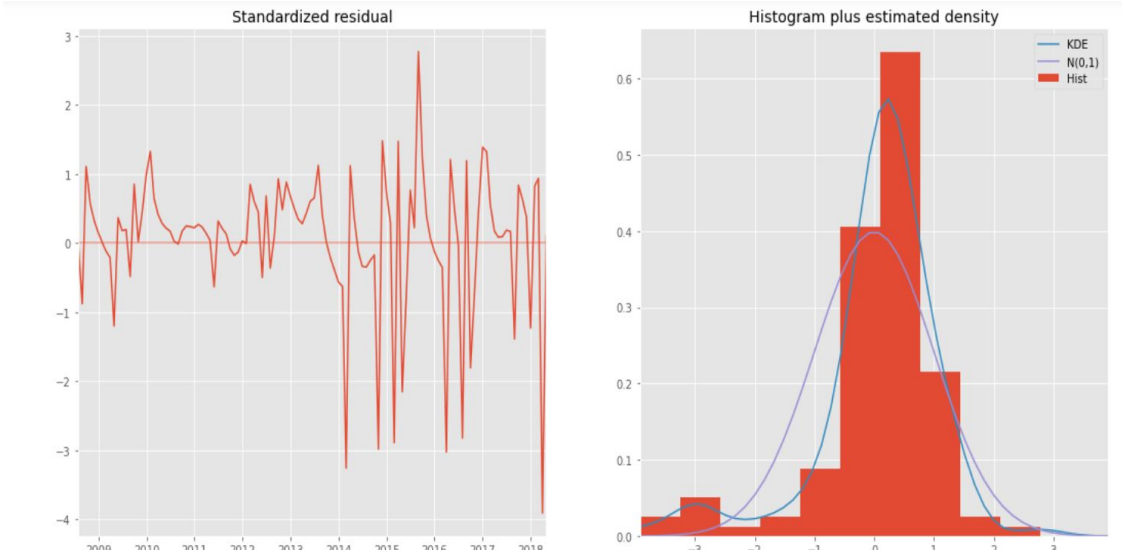
ARIMA Model Testing

- **Testing Performance:**
  - Generated predictions for the test period.
  - Calculated confidence intervals for predictions.
- **Performance Evaluation:**
  - Root Mean Squared Error (RMSE):
    - Value: 0.13316429334548938
    - Indicates the average deviation of predictions from actual values.
- **Predictions Visualization:**
  - Plotted observed data and predicted values.
  - Highlighted confidence intervals for predictions.



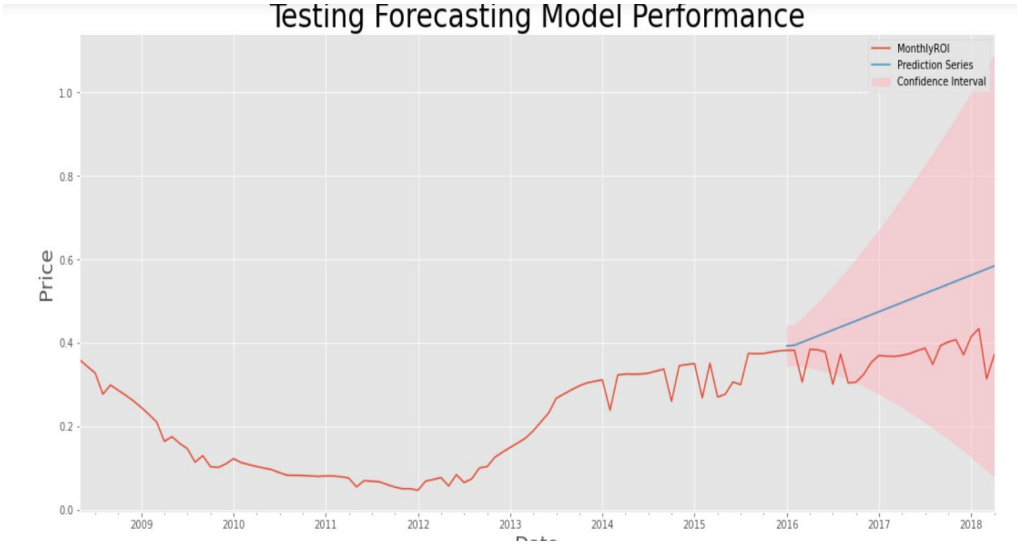
SARIMA Model Development

- **Seasonal ARIMA (SARIMA):**
  - Considered SARIMA due to observed seasonality in the data.
  - Parameters:
    - **seasonal_order=(0, 0, 0, 0)**: No seasonal component.
- **SARIMA Model Fit:**
  - Fitted SARIMA model to training data.
  - Model Summary:
    - Coefficients and statistics.
- **Diagnostic Plot:**
  - Visualized diagnostic plots for SARIMA model evaluation.



SARIMA Model Testing

- **Performance Evaluation:**
  - Generated predictions for the test period.
  - Calculated confidence intervals for predictions.
- **Performance Comparison:**
  - Compared RMSE of SARIMA model with ARIMA model.
  - Highlighted improvement or deterioration.
- **Predictions Visualization:**
  - Plotted real vs. predicted values with confidence intervals.



Prophet Model Development and Evaluation

- **Introduction to Prophet:**
  - Suitable for data with strong seasonal patterns.
  - Handles seasonality effectively.
- **Model Training:**
  - Trained Prophet model on training data.
  - Utilized default hyperparameters for initial training.
- **Model Testing:**
  - Generated predictions for the test period.
  - Calculated RMSE to assess performance.
- **Performance Analysis:**
  - Examined RMSE value to gauge model accuracy.
  - Compared Prophet model performance with previous models.
- **Hyperparameter Tuning:**
  - Explored manual hyperparameter tuning.
  - Tested various combinations of hyperparameters.
- **Optimization Outcome:**
  - Determined best hyperparameters for the Prophet model.
  - Highlighted potential improvement in model performance.

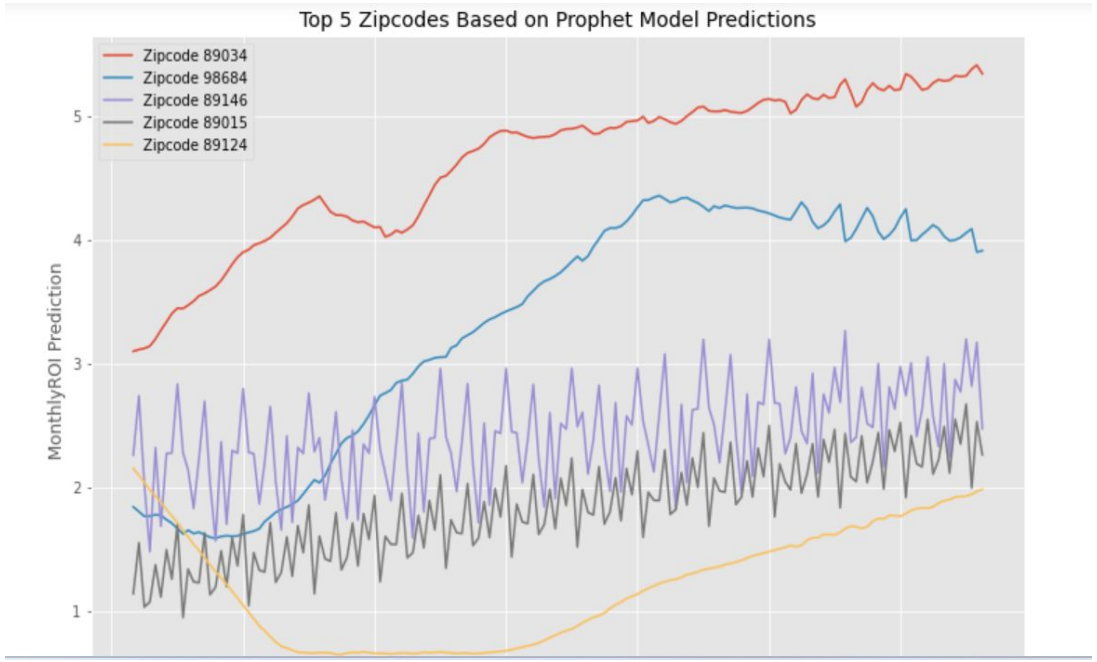
Forecasting Results

- Table displaying forecasted values for selected time periods.
- Visual representation of the forecast using plot_plotly.

- the tangible results of our forecasting efforts using the tuned Prophet model. The table showcases the forecasted values, including the predicted values (yhat), lower and upper bounds of the prediction intervals (yhat_lower and yhat_upper), along with the corresponding dates (ds).
- Additionally, the visual representation provides a clear illustration of the forecasted trend over the specified time horizon, aiding in the interpretation of future real estate value trends.

Zip Code Analysis

- Summary of the forecasting process for different zip codes.
- Observations on the predicted ROI trends for various zip codes.



Recommendations and Conclusion

- Summary of the recommendations and next steps outlined.
- Encouragement for investors to leverage the insights for strategic positioning in the real estate market.
- Optimal Zip Codes for Investment: List of recommended zip codes.
- Preferred Counties for Investment: List of recommended counties.

Next Steps

- Implementation of Predictive Model: Integration of predictive model into investment strategies.
- Detailed Due Diligence: Importance of thorough due diligence process.
- Diversification Strategies: Suggestions for portfolio diversification.
- Continuous Monitoring: Emphasize the need for ongoing market monitoring and adaptation.


Thank You!