TIME SERIES ANALYSIS: ZILLOW HOUSING DATA

Contibutors

- Victorine Imbuhila
- Titus Mutuku
- Mary Gaceri
- Sammy Kimani
- Winnie Mauti
- Mwenda James

Business Understanding

Real estate comprises of land, buildings, and physical properties, with applications in residential, commercial, industrial, and agricultural sectors. It plays a pivotal role in the global economy, contributing significantly to the Gross Domestic Product (GDP).

Real Estate in the USA: With a population exceeding 330 million, the US real estate industry holds substantial economic importance, constituting around 6% of the GDP. The sector includes residential and commercial real estate, real estate development, property management, and real estate investment trusts (REITs).

Zamara real estate investment firm is seeking actionable insights to guide their investment decisions. The firm's inquiry goes beyond a simple profit-maximization objective; they recognize the need to balance profit margins with risk mitigation and are open to considering a reasonable time horizon for their investments. This nuanced approach requires a comprehensive analysis that takes into account multiple dimensions of the real estate market.

Challenges in Real Estate Investment: Numerous factors impact the real estate market, such as government policies, demographics, affordability, housing access disparities, location, and economic conditions. Zamara Real Estate Investment Firm seeks to navigate the dynamic landscape of the US real estate market, which contributes significantly to the national GDP.

Problem Statement

Given the substantial economic impact and diversity within the US real estate sector, Zamara faces the challenge of predicting future trends in real estate prices across various zip codes. The primary objective is to develop a robust time series forecasting model that provides accurate and reliable predictions. This involves understanding and leveraging historical real estate data to empower Zamara in making strategic investment decisions. Key considerations include the identification of high-performing zip codes, balancing potential profit margins with associated risks, and establishing a clear time horizon for the forecasts. The ultimate goal is to enhance the firm's ability to strategically allocate resources and optimize returns in the dynamic and competitive US real estate market.

Objectives

Main Objective: Develop a time series model predicting the top five zip codes for real estate investment.

Specific Objectives:

1. Identify and understand seasonal patterns influencing real estate prices in different zip codes.
2. Evaluate which city exhibits the most promising real estate investment opportunities.
3. Forecast property values over the short and long term, aiming to identify the most favourable zip codes for investment.

Metric of Success

The model's success will be measured by achieving a Root Mean Squared Error (RMSE) of less than 5%, coupled with the identification of the zip codes yielding the highest Return on Investment (ROI).

ROI stands for Return on Investment, which is a measure of how much profit or loss a business makes from its investment. A higher ROI means a higher profit and a lower ROI means a lower profit or a loss.

Data Understanding

The dataset is sourced from Zillow Research, a reputable and widely used platform for real estate market data. The dataset is stored in the file *zillow_data.csv* within the project repository. The dataset contains 14723 rows and 272 columns.

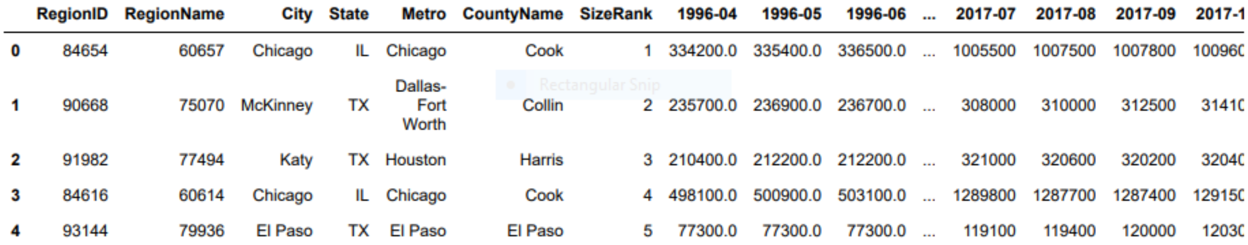
The Zillow dataset provides detailed real estate data, with each row representing a unique zip code. Here's an overview of the dataset structure:

- *RegionID:* A unique identifier for each region.
- *RegionName:* The zip code for the region.
- *City:* The city where the region is located.
- *State:* The state where the region is located.
- *Metro:* The metropolitan area associated with the region.
- *CountyName:* The name of the county where the region is located.
- *SizeRank:* A ranking of the region based on size.
- *Monthly Price Data:* Starting from April 1996 to April 2018, this dataset includes monthly real estate prices for each zip code.

Data Preparation

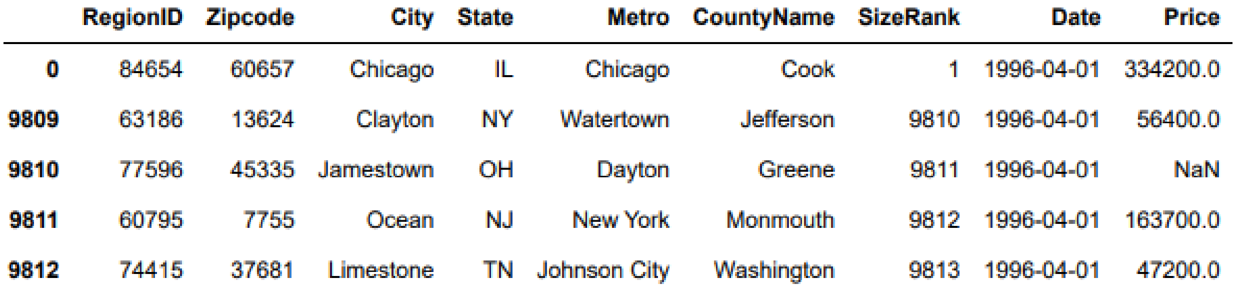
In order to prapare our data for modelling we :

1. Reshape the data from wide to long format. Looking at the format of the data in *zillow_data.csv*, we notice that the actual Time Series values are stored as separate columns as shown below:



	RegionID	RegionName	City	State	Metro	CountyName	SizeRank	1996-04	1996-05	1996-06	...	2017-07	2017-08	2017-09	2017-1
0	84654	60657	Chicago	IL	Chicago	Cook	1	334200.0	335400.0	336500.0	...	1005500	1007500	1007800	100960
1	90668	75070	McKinney	TX	Dallas-Fort Worth	Collin	2	235700.0	236900.0	236700.0	...	308000	310000	312500	31410
2	91982	77494	Katy	TX	Houston	Harris	3	210400.0	212200.0	212200.0	...	321000	320600	320200	32040
3	84616	60614	Chicago	IL	Chicago	Cook	4	498100.0	500900.0	503100.0	...	1289800	1287700	1287400	129150
4	93144	79936	El Paso	TX	El Paso	El Paso	5	77300.0	77300.0	77300.0	...	119100	119400	120000	12030

This is a wide format which makes the dataframe intuitive and easy to read. However, there are problems with this format when it comes to actually learning from the data, because the data only makes sense if you know the name of the column that the data can be found it. In order to pass this data to our model, we'll reshape our dataset to long format. Our dataframe now appears as shown below:

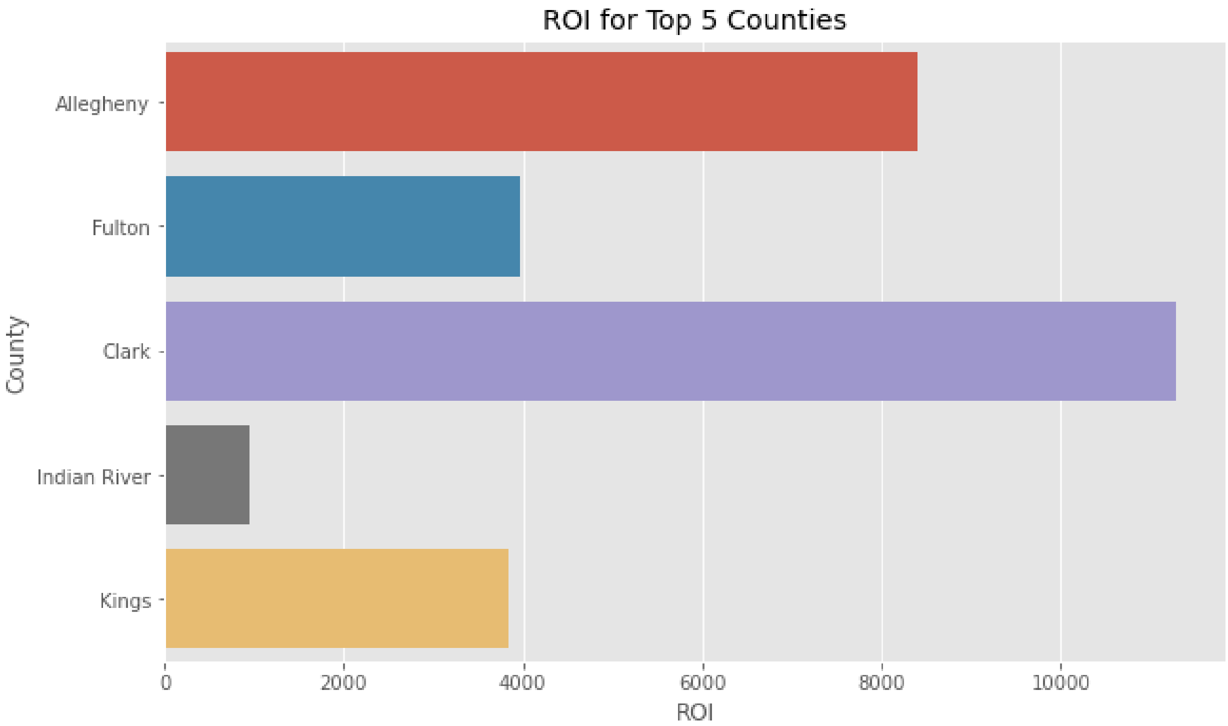


	RegionID	Zipcode	City	State	Metro	CountyName	SizeRank	Date	Price
0	84654	60657	Chicago	IL	Chicago	Cook	1	1996-04-01	334200.0
9809	63186	13624	Clayton	NY	Watertown	Jefferson	9810	1996-04-01	56400.0
9810	77596	45335	Jamestown	OH	Dayton	Greene	9811	1996-04-01	NaN
9811	60795	7755	Ocean	NJ	New York	Monmouth	9812	1996-04-01	163700.0
9812	74415	37681	Limestone	TN	Johnson City	Washington	9813	1996-04-01	47200.0

2. Covert the data types. We convert the *'Date'* column to datetime format to enable easy manipulation of date and time information.

3. Slicing the data. We slice our DataFrame to remain with the most recent data which is relevant for our analysis. This is because it reflects the current trends and market conditions more accurately. By focusing on the last 10 years, we can analyze data that is more representative of current market forces.

4. Perform Feature Engineering. We create a column with the Return On Investment(ROI), we will use this to determine the best county to choose zipcodes from.



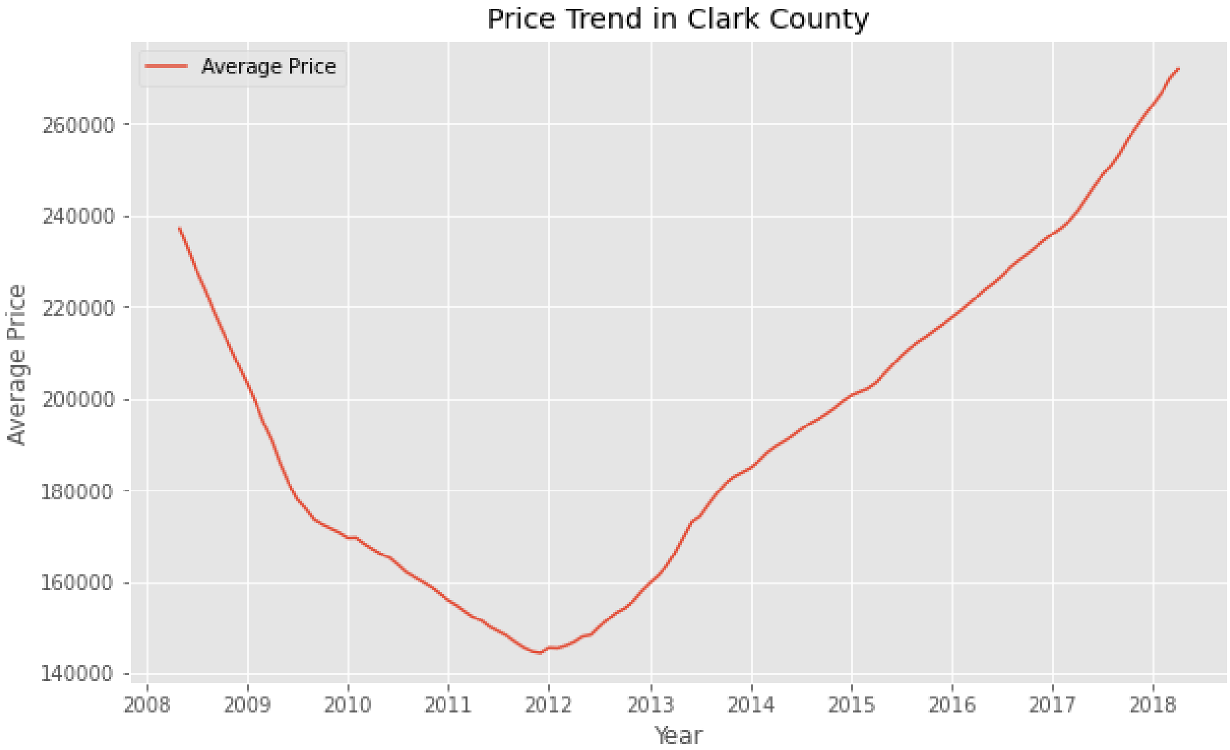
The plot of ROI for top five counties reveals that Clark County has the highest ROI, followed by , Allegheny,Fulton, Kings, and Indian River in descending order. This means that Clark County is the most profitable county to invest in, while Indian River is the least profitable or the most loss-making county.

The plot also shows the difference in ROI among the counties. For example, Clark County's ROI is almost twice as much as Fulton County's ROI, and more than four times as much as Kings County's ROI. This indicates that there is a large variation in the profitability of the counties, and that some counties are much more attractive for investment than others.

5. Check for missing values and handle them. Checking the dataframe info, we find that we have missing values in the *'Metro'* and *'Price'* columns. We fill the 'Metro' column with 'missing' and fill the missing price values using linear interpolation.

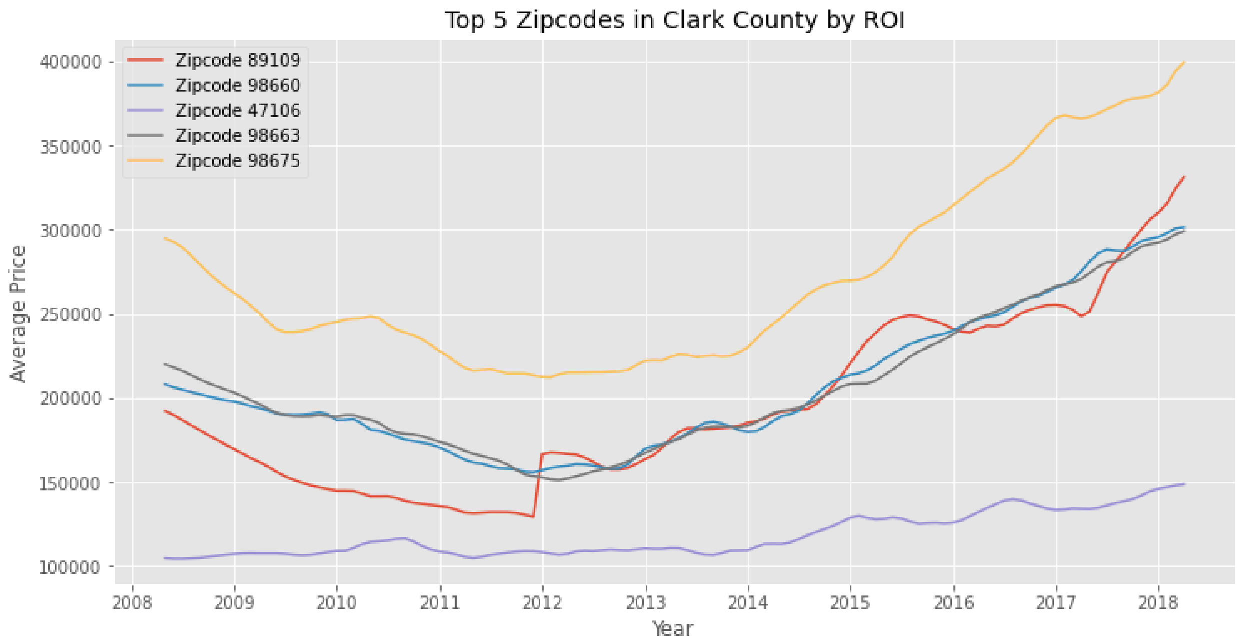
Exploratory Data Analysis

1. **We perform trend analysis** in order to identify and understand the underlying trends of the time series data. We use a line plot shown below:



- The line plot shows the average price trend in Clark County from 2008 to 2018. The trend is negative from 2008 to 2012, and positive from 2012 to 2018. This means that the average price decreased in the first period, and increased in the second period.
- The trend can be described by a piecewise linear function, which is a function that consists of two or more linear segments.
- The average price in Clark County was affected by the global financial crisis of 2008-2009, which caused a severe drop in the housing market.
- The average price reached its lowest point in 2012, which could have been a good opportunity for buyers to enter the market at a bargain price.
- The average price recovered from the crisis and experienced a strong growth from 2012 to 2018, which could have been a result of increased demand, improved economic conditions, and limited supply.
- The average price in 2018 was almost the same as in 2008, which means that the market has returned to its pre-crisis level after a decade of volatility.
- The slope can also be used to compare the different segments of the line plot. For example, we can see that the slope of the second segment is positive and higher than the slope of the first segment, which means that the average price increased faster in the second period than it decreased in the first period. This indicates that the market recovered quickly from the crisis and surpassed its previous level.

2. **We examine the top five zip codes in Clark County** by Return On Investment(ROI) as shown below:



Observation:

The results show the average price trends for properties in five different zip codes in Clark County from 2008 to 2018, ranked by ROI (Return on Investment).

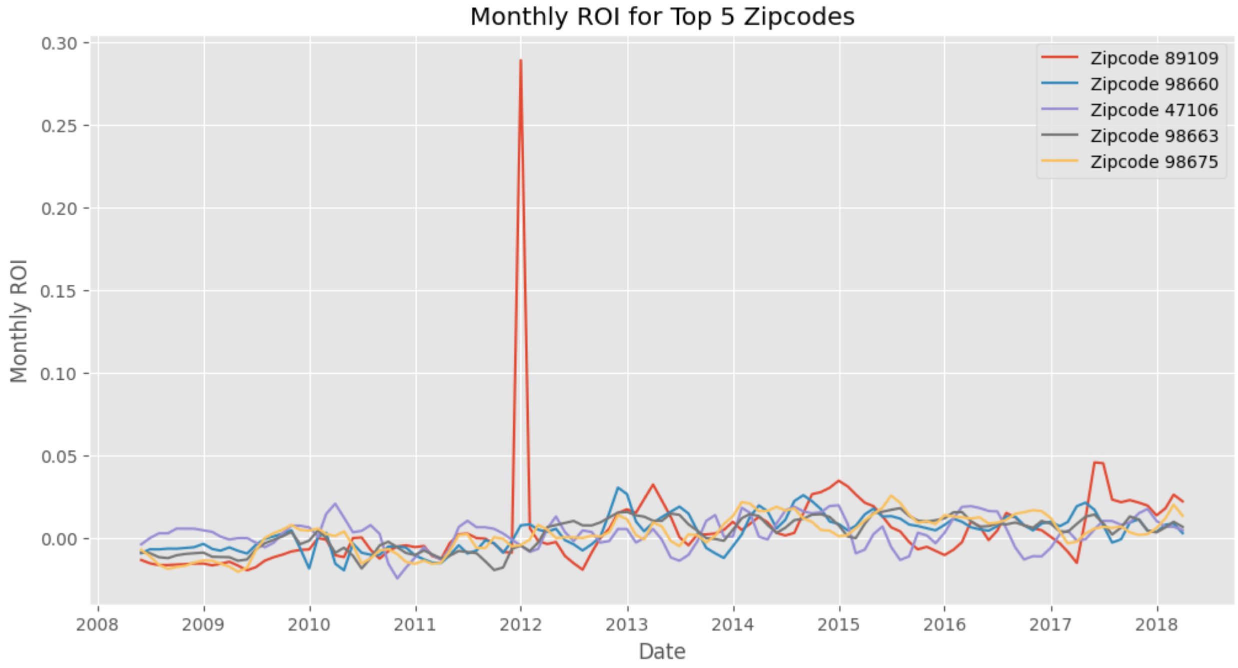
Interpretation:

- *Zip code 89109 (Las Vegas) has the highest ROI of 72.51%*. It also shows a significant increase in average property prices over the years, especially from 2016 onwards. This suggests that Las Vegas is a very attractive and lucrative market for real estate investment, as the property value has grown rapidly and substantially.
- *Zip code 98660 (Vancouver) has an ROI of 44.90%.* The average property prices have been steadily increasing since around 2012. This indicates that Vancouver is a stable and profitable market for real estate investment, as the property value has increased consistently and moderately.
- *Zip code 47106 (Borden) with an ROI of 41.97%* shows a consistent increase in property prices but at a slower pace compared to Las Vegas and Vancouver. This implies that Borden is a reliable and decent market for real estate investment, as the property value has increased gradually and slightly.
- *Zip code 98663 (Vancouver) has an ROI of 35.97%.* The trend is similar to that of Borden but with a slightly lower rate of increase. This means that Vancouver is a dependable and fair market for real estate investment, as the property value has increased similarly and marginally.
- *Zip code 98675 (Yacolt) has the lowest ROI among these five at 35.57%*, and its price trend is relatively stable with a moderate increase. This reveals that Yacolt is a secure and modest market for real estate investment, as the property value has increased steadily and mildly.

3. **We also examine the stationarity of the time series**. In time series modeling, it's commonly assumed that the data exhibits stationarity. This means that the statistical

properties such as mean, variance, and autocorrelation of the series remain constant over time. Stationary data simplifies the modeling process significantly.
To verify if the time series data is stationary, techniques like the Dickey-Fuller test can be employed along with examining the rolling mean.
If the data turns out to be non-stationary, a method known as differencing can be applied. Differencing helps in transforming the data into a stationary form, which is more conducive for time series modeling.

Monthly ROI is the ROI calculated for each month, based on the difference between the property value at the beginning and the end of the month, divided by the initial value, and the graph is plotted as shown below:



Observation:
The results show the monthly ROI for the top 5 zip codes in Clark County from 2008 to 2018.
Interpretation:

- The graph shows that the monthly ROI for all five zip codes fluctuated over time, staying mostly below 0.10. This means that the property value did not change much from month to month, and the profit or loss was relatively small.
- There is a noticeable spike in monthly ROI for all zip codes around 2013, reaching nearly 0.30 for zip code 98675. This means that the property value increased significantly in that year, and the profit was very high.
- The spike in monthly ROI could be due to various factors**, such as increased demand, improved economic conditions, limited supply, or other market forces.
- After 2013, the monthly ROI declined and stabilized for all zip codes, with minor fluctuations. This means that the property value did not change much after the spike, and the profit or loss was moderate.

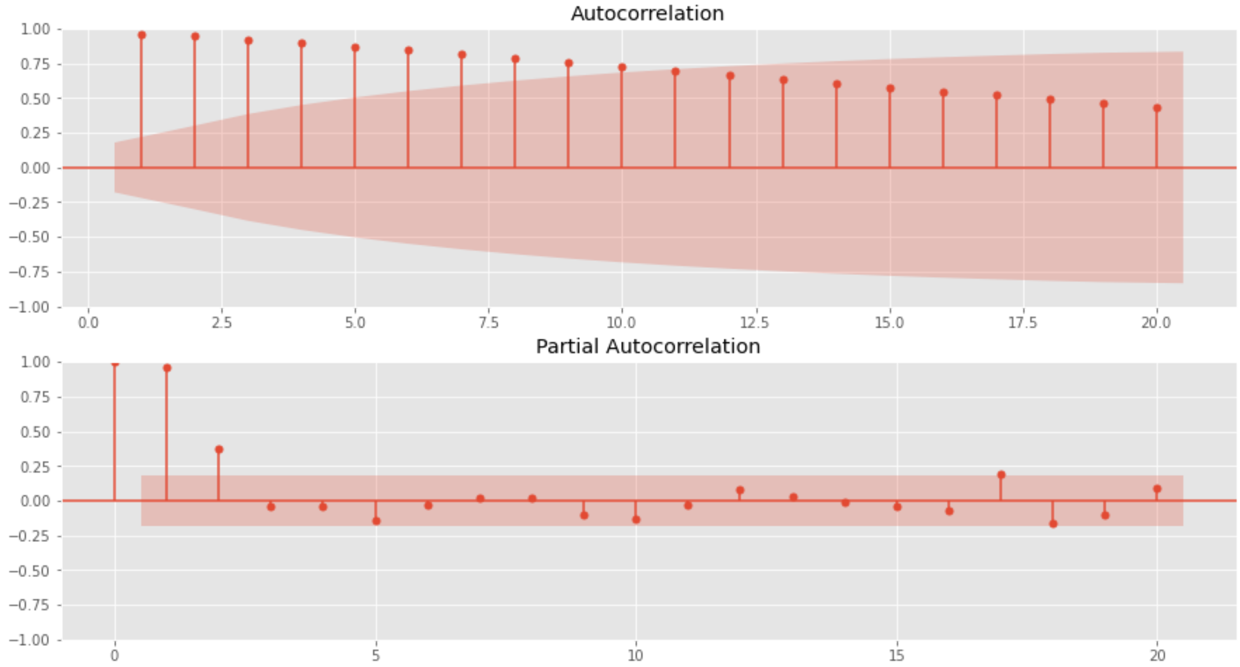
Interpreting ADFuller Results:

The ADF test has a null hypothesis that the time series has a unit root, which means it is non-stationary. The alternative hypothesis is that the time series does not have a unit root, which means it is stationary. The test statistic is compared to critical values to determine whether to reject or fail to reject the null hypothesis. A low p-value (usually below 0.05) indicates strong evidence against the null hypothesis, and a high p-value (usually above 0.05) indicates weak evidence against the null hypothesis.

The ADFuller test p-value for combined Clark County zip codes gives us a p-value of 0.0.The results show that the p-values for combined Clark County zip codes are very low, which means that we can reject the null hypothesis and conclude that the time series for these zip codes are stationary.

Modeling

We plot the ACF and PACF plots to check for seasonality and trends in the time series data. The plots are as shown:



Observation:

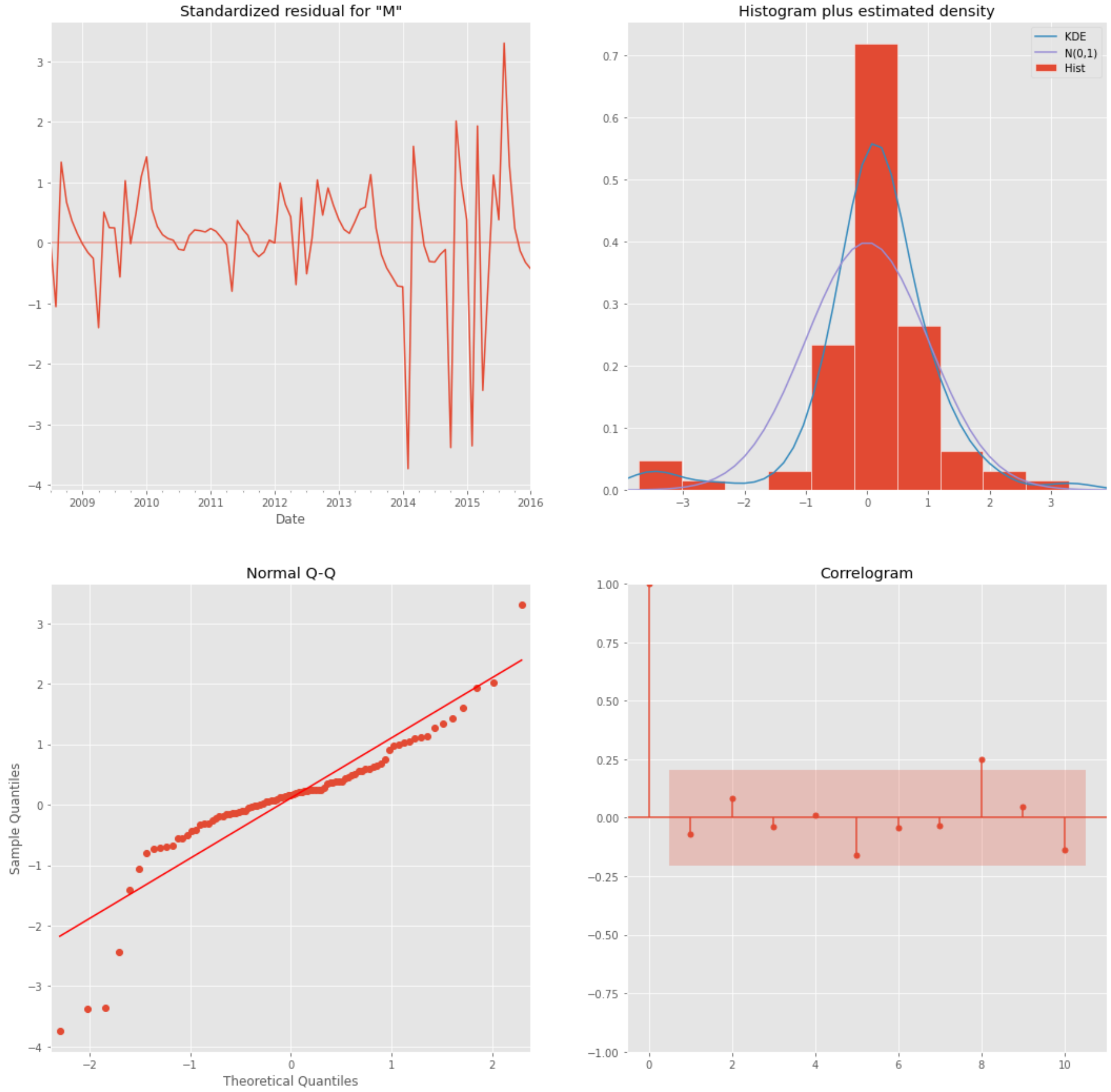
In the Autocorrelation graph, there is a strong positive correlation at lag 0, which is expected as any data point has a perfect correlation with itself. The correlation then decreases as the lag increases but remains positive. This pattern suggests that the data is not random and there is some trend or seasonality.

In the Partial Autocorrelation graph, there is a significant spike at lag 1 indicating that there is a strong correlation when accounting for the influence of other lags, and then it stabilizes around zero.

The Autocorrelation plot indicates seasonality as there are regular peaks at consistent intervals.

1. ARIMA Model

On stationary data, the best model to use is ARIMA.

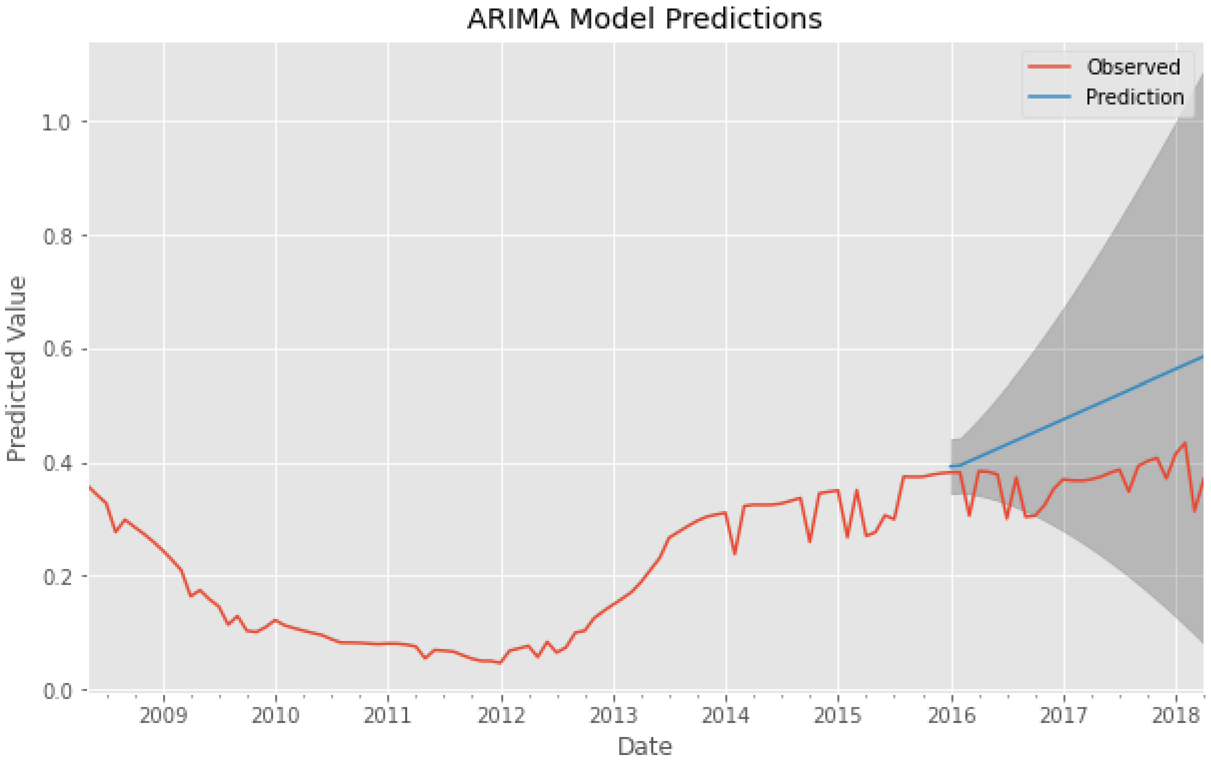


Observations:

- The Standardized Residual for “M” plot indicates a significant spike around 2015, suggesting an outlier or a specific event that caused a large residual at that time. This affects the accuracy of the model.
- The Histogram plus estimated density plot shows the distribution of residuals. It appears to be approximately normally distributed, as indicated by the overlay of the normal distribution curve. This is a desirable property for our model, as it implies that the residuals are random and have no systematic patterns.

- In the Normal Q-Q plot, most points closely follow the theoretical line, indicating that residuals are normally distributed with some deviations. The deviations at the ends of the plot suggest that there are some extreme values in our data that may not fit well with the normal distribution assumption.
- The Correlogram or ACF plot shows that there is no significant autocorrelation in our residuals, as most are within the confidence interval. This means that our model has captured most of the information in the data and there is no remaining structure in the residuals.

Testing model performance

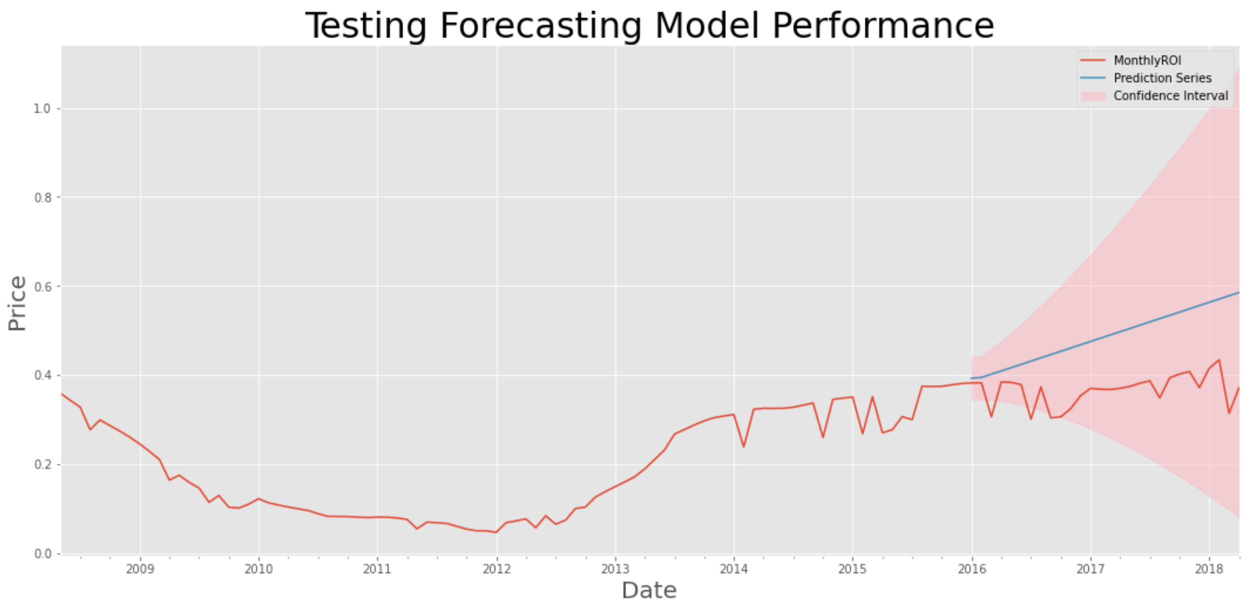


Observation:

- The observed data fluctuates between 0.0 to around 0.6 predicted value. There is a noticeable dip around 2012 before it starts increasing again.
- The prediction begins from around 2016 and shows an upward trend, indicating that the predicted value is expected to increase over time.
- The shaded grey area surrounding the Prediction line indicates confidence intervals or error margins associated with these predictions.

2. *SARIMA Model*

From the ACF and PACF models, we saw that the data exhibits seasonality. We use a SARIMA model since it is equipped to capture and model seasonal components, providing a more accurate representation of the underlying patterns.



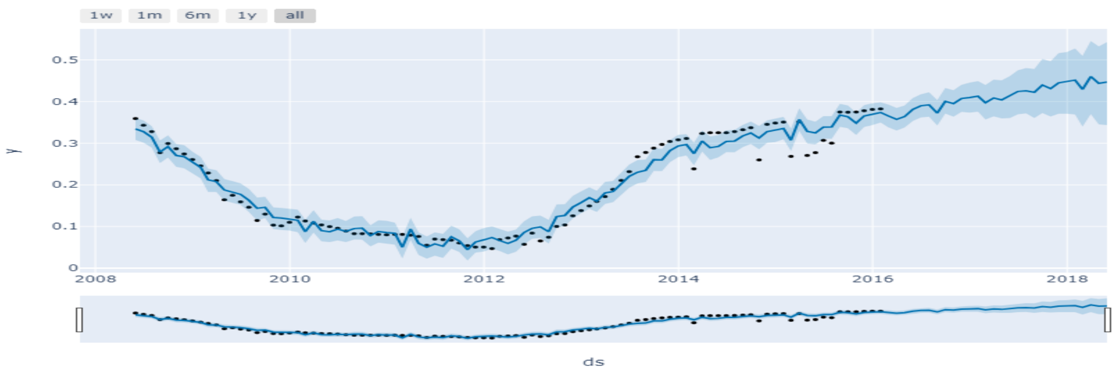
The SARIMA model gives us an RMSE value of 0.1305 which is better than the ARIMA model suggests that, on average, the model's predictions deviate by approximately 0.1305 percentage points from the actual monthly ROI values. We want to compare this value with the prophet model.

3. *PROPHET Model*

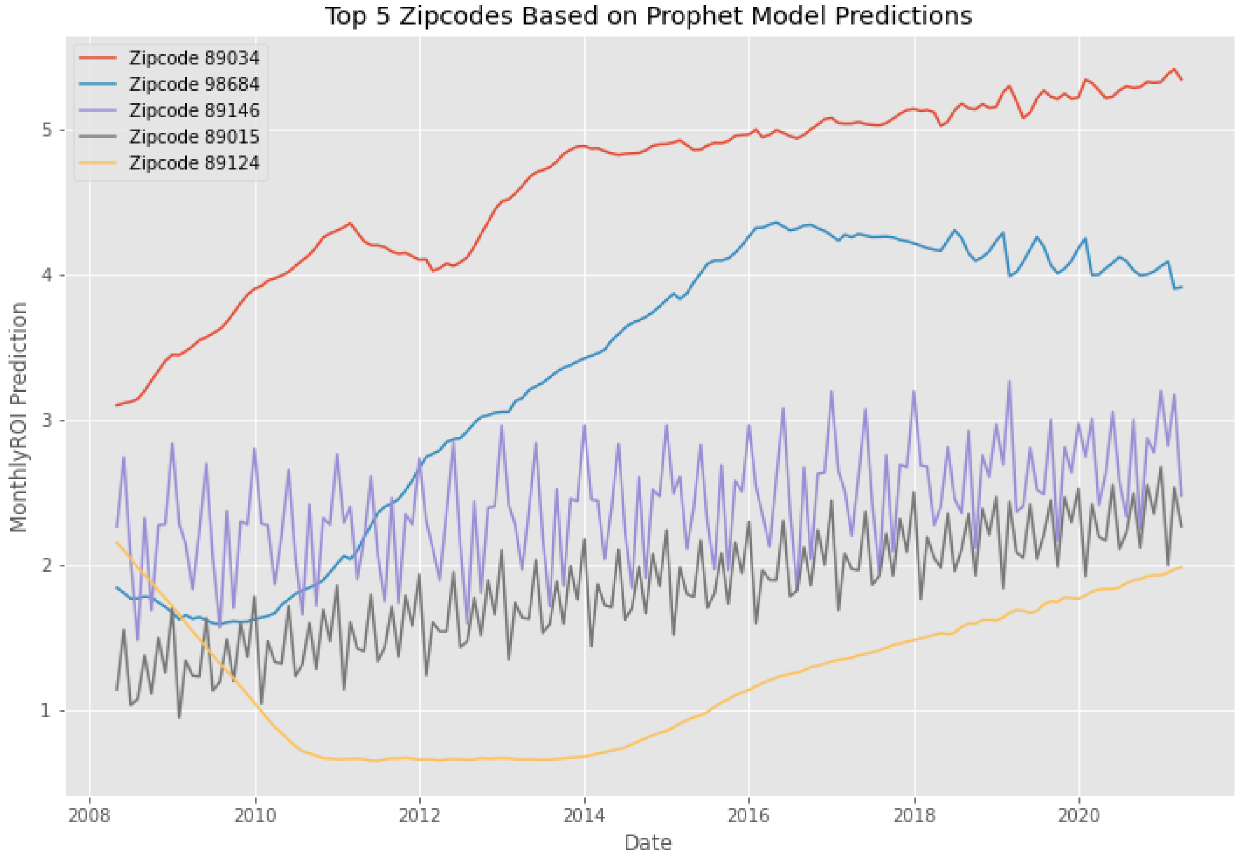
We choose Prophet model since it is designed to handle seasonality, especially in datasets with strong seasonal patterns. The RMSE value of 0.13245648205825172 is lower than that of the SARIMA model, therefore we tune it in order to improve it. The tuned model gives us an RMSE value of 0.05782154676800822. This is the better model since it has the best RMSE out of all the other models.

Forecasting

We forecast future values using the tuned prophet model.



The graph shows a time series data forecasted into the future from 2008 to 2018.
The plot shows a dip in values around 2012 and a steady increase from there onwards.



Zipcode 89034 shows a consistent increase in Monthly ROI Prediction over time. Zipcode 96864 has fluctuating predictions but shows an overall increase. Zipcode 89146 also increases but at a slower rate compared to others. Zipcode 89015 has highly fluctuating predictions with no clear trend of increase or decrease. Zipcode 89124 shows an initial increase until around 2014 and then remains relatively stable 1.

Based on the graph, it appears that Zipcode 89034 has consistently had the highest predicted ROI over the years.

Conclusion

The study identified the most promising zip codes and counties for real estate investment. The data indicated a positive trend in real estate value over time, but no clear seasonal pattern was observed.

We concluded that the best zip codes to invest in are 89034, 98684, 89146, 89015, and 89124. The best counties to invest in are Clark County, Allegheny, Fulton County, Kings County, and Indian River County.

The data showed an upward trend in real estate value over time, but there was no clear seasonality pattern to determine the best time period to invest in real estate. A predictive time series model was created to help predict future real estate values.

Recommendations

1. Optimal Zip Codes for Investment: 89034 (Mesquite, Nevada), 98684 (Vancouver, Washington), 89146 (Las Vegas, Nevada), 89015 (Henderson, Nevada), 89124 (Las Vegas, Nevada). These zip codes showcase the highest Return on Investment (ROI) and are recommended for investors seeking promising opportunities.

2. Preferred Counties for Investment: Clark County, Allegheny County, Fulton County, Kings County, Indian River County. Investing in real estate within these counties is recommended, offering diverse opportunities and potential for robust returns.

3. Strategic Timing. While no clear seasonal patterns were identified, the overall upward trend suggests that the real estate market is favorable for investment. Investors are advised to consider the long-term growth potential rather than specific timing considerations.

Next Steps:

1. Implementation of Predictive Model. Integrate the predictive time series model into investment strategies, using it as a valuable tool for making informed decisions and optimizing portfolio performance.

2. Detailed Due Diligence. Conduct a comprehensive due diligence process, including property inspections, market analysis, and local economic factors, to further refine investment decisions and mitigate risks.

3. Diversification Strategies. Explore diversification strategies within the recommended zip codes and counties, spreading investments across different property types and neighborhoods to enhance portfolio resilience.

4.Continuous Monitoring. Stay abreast of market trends, economic indicators, and any emerging patterns to adapt investment strategies accordingly. Regularly update the predictive model with new data for improved forecasting accuracy.

By following these recommendations and next steps, investors can position themselves strategically in the real estate market, capitalize on identified opportunities, and navigate the dynamic landscape with confidence.