## **BUSINESS UNDERSTANDING**

Overview

The real estate market presents a dynamic and complex investment landscape, where the value of investments can significantly impact the financial well-being of investors. Identifying lucrative investment opportunities requires analyzing vast amounts of historical data to predict future trends, with the inherent volatility and regional diversity of the real estate market. Investors seek to maximize returns while managing the risks associated with market fluctuations and regional economic changes. The primary stakeholder in this project is Prime Nest Investment real-estate company which is focused on identifying the top 5 zip codes for real estate investment and optimizing their portfolios performance by selecting regions with the highest potential for growth and stability.

Challenges

· Historical Data Only: The dataset does not include future or real-time data, limiting the analysis to historical trends.

· Missing Values: Any missing data could affect the analysis's accuracy.

· Lack of External Factors: External economic indicators such as interest rates, employment rates and GDP growth which can significantly influence real estate prices are not included

Proposed Solution

· Implement robust data cleaning methods such as interpolation or imputation techniques to address missing values

· Employ SARIMA (Seasonal Autoregressive Integrated Moving Average) models to capture both seasonal patterns and trends in the time series data.

· Conduct scenario analysis to evaluate the impact of various economic conditions and policy changes on the housing market

· To train a time series model with a r2 score of above 80% and RMSE of less than 2000 compared to the mean house price, indicating a high level of accuracy

Conclusion

PrimeNest Investments will use the project's findings to focus their investment portfolio on zip codes with the best potential. Market analysts, financial advisors, and individual investors can leverage these insights for making informed real estate decisions. Beyond identifying investment opportunities, the project enhances understanding of real estate market dynamics and equips stakeholders with actionable insights to navigate investment complexities, thereby supporting their financial goals.

## **DATA UNDERSTANDING & CLEANING**

The dataset is sourced from Zillow Research a reputable provider of historical real estate market data and is stored in a CSV format. It encompasses historical home values across various zip codes in the United States, spanning from April 1996 to April 2018.

Each row represents a unique zip code and the columns meaning and data types are as follows:

RegionID - A unique identifier for each region or zip code. Integer Datatype

RegionName - The name of the region, which typically corresponds to the zip code. Integer Datatype

City - The city where the region is located. Object Datatype

State - The state where the region is located. Object Datatype

Metro - The metropolitan area that the region is part of. Object Datatype

CountyName - The county where the region is located. Object Datatype

SizeRank - A ranking of the region based on its size or importance, with 1 being the largest or most significant. Integer Datatype

Date - The date of the recorded real estate price. Stored as separate columns as floats and integers Data Types

Price - The real estate price recorded for that region and date. Float Datatype

The integers facilitate mathematical computations*,* floats enable precise numerical analysis and forecasting

## **DATA CLEANING**

The data has 14723 rows and 272 columns before any preparation is done.

The data is in a wide format and needs to be converted to long format using pandas function *pd.melt()*

To check for missing values we’ll use pandas function *data*.*isnull().sum()* to get the sum of missing values column wise

Check for duplicates is also done using pandas function data.*duplicates()* to return a count of all duplicate values in the dataset

Outliers in the dataset were identified using interquartile ranges whereby all values beyond the upper quartile and below the lower quartile are regarded as outliers

All null values in the price column are dropped by row whereas the Metro column is dropped entirely as it has the highest number of missing values and won’t necessarily be used in this analysis

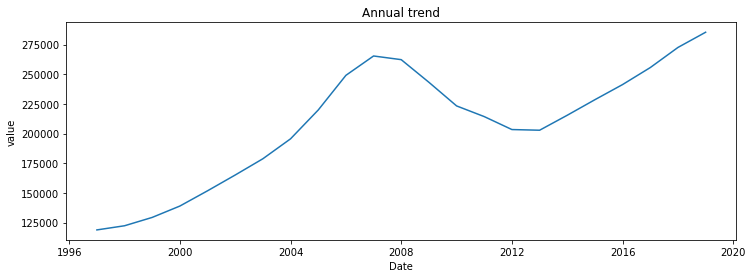
Outliers in the price column are also removed as they will introduce biases to the model

## **EXPLORATORY DATA ANALYSIS**

EDA in time series analysis involves plotting the house prices against the time, which is the index. This is done to see patterns in our dataset over time. Exploring trends for either an upward, downward, stationary or cyclic trend.

Seasonality is also checked since it affects the decisions on the types of modeling to prioritize on. An unexpected trend is also accounted for.

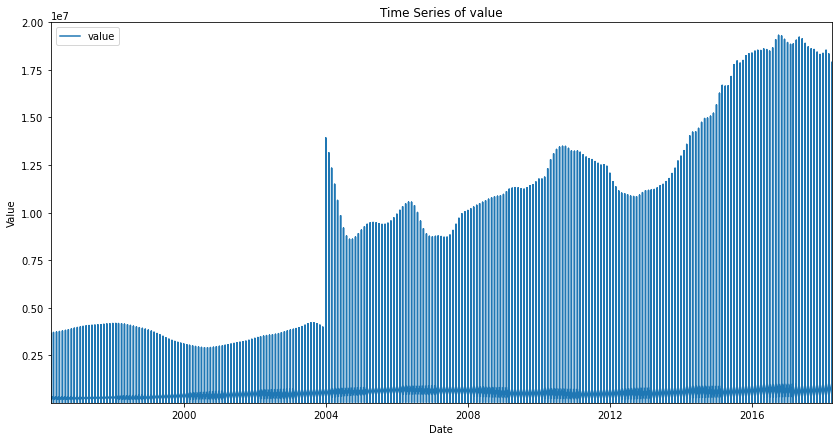
Below is the trend of the entire dataset from 1996 to 2018 against house prices indicated as value.



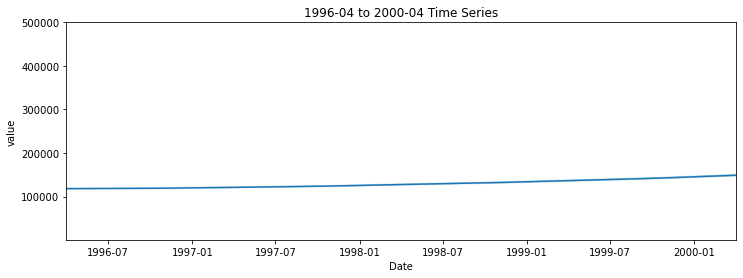
Annual trends show an upwards curve with:

* An upward trend from 2000 to 2008 to about 250000 as price value.
* A slight Downward trend between 2008 and 2012 and eventually an upward trend from 2016 to 2018.

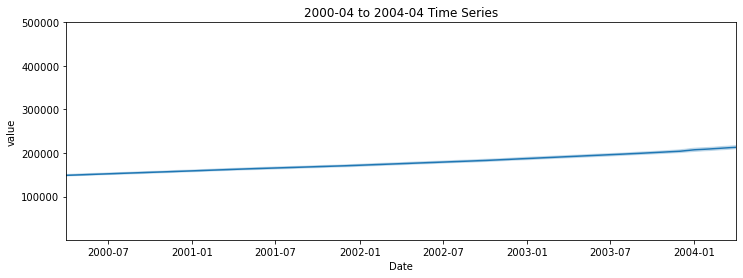
Below is a better view of the dataset fluctuations.



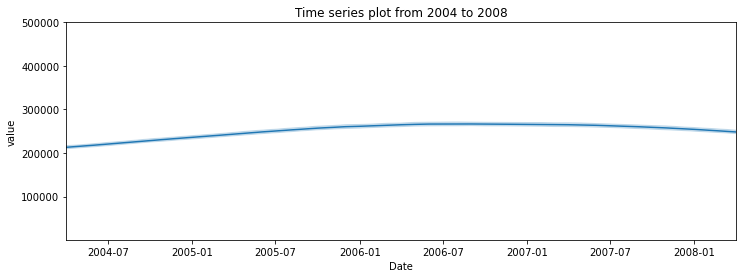
*A closer look at the trends;*

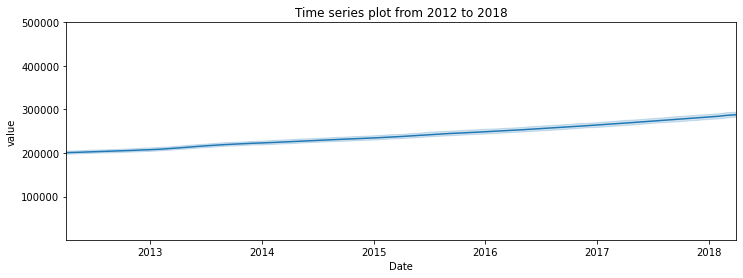


Trend from 1996 to 2000 shows an almost stationary trend.



2000-04 to 2004-04 Time Series plot shows a gradual upwards trend, with a very small slope.

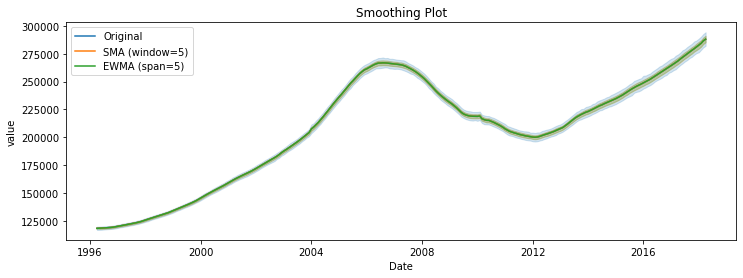
The trends from the year 2004 and 2008 show a slight increase at 2004 from approximately 0.15M to 0.21M in value.



There is an upward trend from 2012 to 2018, reaching its highest peak in 2018.

*Time series smoothing;*

Up close, the data is not very smooth, we will hence use time series smoothing techniques to smooth out our data and narrow down on the best technique to use.

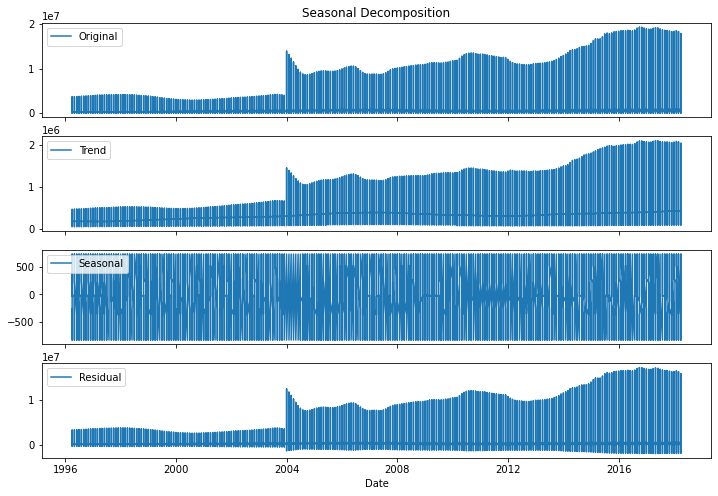


Simple moving average (SMA) and Exponential weighted moving average (EWMN) are used below, and the following are their plots.

Both techniques are effective but EWMA is better since it smoothes out the original time series dataset plot better, and it is more responsive to short term fluctuations as compared to SMA. This is because SMA puts more pressure on old data, while EWMA gives more relevance to recent observations of the price column value.

*Seasonal decomposition:*

Checking for seasonality in our dataset. This affects the type of machine learning algorithms to use.

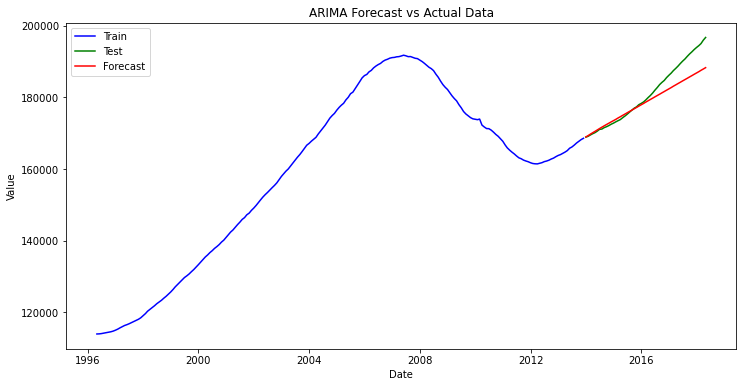


* The first image shows the entire time series dataset plotted.
* The trend plot shows a general upward trend as the years increase with a high slope peak occurring in 2004, and a gradual increase from then with few downward trends.
* There is an almost stationary trend between 1996 and 2004 of the house prices.
* There are some patterns that repeat within a fixed period, indicating seasonality.
* Residuals are seen to increase as the years increase; noise is increasing. They are closer to zero before 2004, and started expanding from then.

## **MODELLING**

The models that will be implemented in this analysis are the ARIMA and SARIMA models.

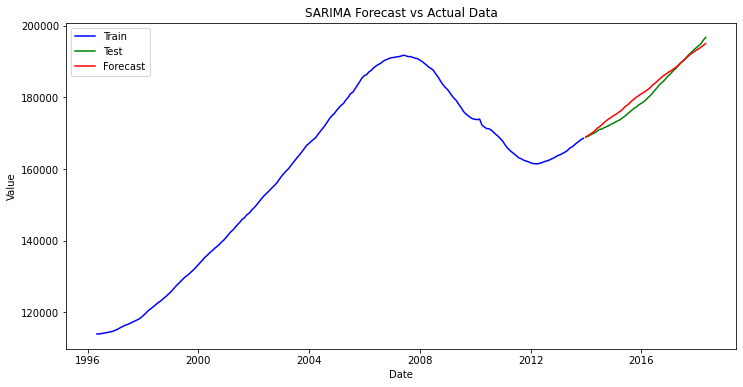
The first iteration of the ARIMA model will use the parameters we identify from our Partial Autocorrelation and Autocorrelation



The ARIMA model with the parameters p=1 , d=2 , q=20 had an r2 score of 81%, but fails to capture the seasonality

For the SARIMA Model parameter tuning was done to find the best values for the autoregressive model, degree of differencing and order of moving-average model.

The best values were 1 for all 3 with a seasonality of 12.



The SARIMA Model has an R2 score of 96% which is relatively good and the forecasts generated will be very good

## **EVALUATION**

For the model evaluation the metrics used are RMSE and R2 score

An r2 score of above 85% with an RMSE below 2000 will depict a high accuracy for the model.

The SARIMA model achieved an R2 score of 96% with an RMSE of 1633 meaning for every forecast made it was off by 1,633 dollars which is relatively low compared to the mean house values.

## **RECOMMENDATION**

\* Regularly update the SARIMA model with new data to maintain the accuracy of forecasts and adapt to changing market conditions.

\* Diversify investments across multiple zip codes to mitigate risk and leverage the varying trends in different regions.

\* Keep track of economic indicators such as interest rates, employment rates, and economic policies that could impact the housing market.

## **CONCLUSION**

The SARIMA model was able to forecast house prices with a relatively low RMSE compared to the mean house price, indicating a high level of accuracy

The time series analysis revealed consistent seasonal patterns in house prices, which were effectively captured by the SARIMA model

## **DEPLOYMENT**

The model was deployed for use to end users using python Framework Fast API.

The get route acts as our home page and prints out a ‘introduction message to the user’

It requires the user to input the date in string format in the order of year, month and day

The post route houses the actual forecasting.

* First storing the input date in variable called target date and converting it to date format
* The last date in the training data is accessed and stored as the current date and converted to date format
* An if statement is presented to ensure the target date is greater than the current date
* Steps are calculated using the target and current date and with this the actual forecasting can happen
* Using the saved sarima model, forecasts are made from the current date to the target date and stored in a variable called forecast
* The forecasts are converted to a list and returned as the output to the user