**REAL ESTATE MARKET DYNAMICS THROUGH TIME SERIES MODELING**

**BUSINESS UNDERSTANDING**

**Stakeholder Identification:**

The primary stakeholder in this project is a fictional real-estate investment firm focused on identifying the top 5 zip codes for real estate investment. This firm represents investors aiming to optimize their investment portfolio's performance by selecting regions with the highest potential for growth and stability. Secondary stakeholders include market analysts, financial advisors, and individual investors who rely on precise market forecasts to make informed decisions.

**Project Value:**

This project addresses the real-world problem of optimizing real estate investments by developing a time series model to forecast the future prices of real estate in different zip codes. By leveraging the Zillow Research dataset, the project aims to:

- Provide a data-driven basis for investment decisions.

- Identify zip codes with the highest potential for appreciation, thereby maximizing ROI.

- Offer insights into market trends, enabling stakeholders to manage risks better and align investments with long-term financial goals.

**How Stakeholders Will Use the Project:**

The real-estate investment firm will use the project's findings to strategize their investment portfolio, focusing on zip codes identified as having the best investment potential. Market analysts and financial advisors can leverage the insights to advise their clients on real estate investments. Individual investors, in turn, can use the recommendations to make informed decisions about where to invest their money within the real estate market.

**Metric of Evaluation**

In the analysis of time series, the criterion for assessing the suitability of models will be the Mean Absolute Percentage Error (MAPE). This metric is selected because it calculates errors as a percentage of the actual values, thus giving a weighted measure of the inaccuracies. This approach is advantageous, especially in managing outliers, since errors are normalized by the true values. Unlike the Root Mean Square Error (RMSE), which simply subtracts predicted values from actual ones without accounting for the scale of the data, MAPE ensures that outliers do not disproportionately affect the perception of the model's accuracy.

**Introduction**

In an era marked by significant economic fluctuations and a dynamic real estate market, understanding and predicting market trends has become essential for investors, policymakers, and stakeholders. The ability to forecast real estate prices with accuracy not only offers a competitive edge in the investment landscape but also aids in making informed decisions that can lead to substantial economic benefits. This report delves into an in-depth analysis conducted using time series modeling on a comprehensive dataset provided by Zillow, aiming to uncover the intricacies of real estate market dynamics.

**Background**

The real estate market's complexity is influenced by various factors including economic indicators, interest rates, demographic trends, and even geopolitical events. These elements interplay to shape the market's behavior, making the task of predicting future prices challenging yet crucial. The advent of data science and machine learning offers new avenues to approach this challenge, providing tools and techniques capable of dissecting and understanding these complex patterns.

**Objectives**

The main objective of this project is to utilize time series modeling techniques to forecast real estate prices across various zip codes. By doing so, the project aims to provide the investment firm with valuable insights into which zip codes offer the most promising opportunities for high returns on investment, while also considering factors like risk and investment horizons.

* To develop time series models capable of accurately forecasting real estate prices for different zip codes.
* To provide the investment firm with actionable recommendations regarding the top 5 zip codes for investment, supported by the rationale derived from the time series analysis.
* To consider additional metrics beyond profit margins, including risk assessment, to ensure well-informed and balanced investment decisions. To interpret the model's findings and discuss their implications for real estate investment.

**DATA UNDERSTANDING**

This section outlines the dataset's composition, including the nature of the data, the variables included, and the preliminary observations made upon initial inspection.

**Dataset Description**

The dataset encompasses a wide range of variables that reflect the dynamics of the real estate market across different geographic locations in the United States. Key variables include region identifiers, property types, price indices, and time-stamped values that track price changes over a specified period. The data spans several years, offering a longitudinal view of the market's evolution.

* RegionID-Represents a unique ID for each region.
* RegionName -Represents the name of the region/ also the zipcode.
* City-Represents the city where the region is located.
* State-Represents the state where the region is located.
* Metro-Represents the metropolitan area where the region is located (if applicable).
* CountyName-Represents the name of the county where the region is located.
* SizeRank-Represents the relative size of the region compared to other regions in the dataset.
* 1996 upto 2018-Represents the median home price for the region in months and years.

**Data Source and Suitability:**

The dataset is sourced from Zillow Research, a reputable provider of historical real estate market data. It encompasses historical home values across various zip codes in the United States, spanning from April 1996 to April 2018. This dataset is particularly suited for the project due to its comprehensive coverage of the real estate market over two decades, offering a rich foundation for analyzing trends, forecasting future real estate prices, and identifying promising investment opportunities.

**Dataset Properties**:

- Size of the Dataset: The dataset consists of 14,723 entries (rows) and 272 columns. Each entry represents a unique zip code.

- Features: Besides geographical identifiers like RegionID, RegionName (zip code), City, State, Metro, and CountyName, the dataset features monthly home values from April 1996 (1996-04) to April 2018 (2018-04).

- Descriptive Statistics: The dataset exhibits a wide range of home values, reflecting the diversity of the U.S. real estate market. The presence of various geographical identifiers supports granular analysis by location.

**Relevance for the Project:**

The temporal nature of the dataset, with over two decades of monthly home values, makes it highly relevant for time series analysis. This analysis can uncover trends, seasonality, and growth patterns, crucial for predicting future values and identifying the best zip codes for investment. The inclusion of geographical identifiers allows for localized analysis, aligning with the project's goal to recommend specific zip codes for investment.

**Feature Justification:**

- Temporal Features: Monthly home value columns are directly relevant for time series forecasting, enabling the identification of long-term trends and cyclical patterns.

- Geographical Features: Identifiers like RegionName, City, and State are essential for localizing the analysis to specific markets, a key requirement for providing targeted investment recommendations.

**Preliminary Observations**

Initial analysis of the dataset reveals several interesting trends:

* Price indices fluctuate significantly over time, indicating periods of both growth and decline.
* Geographic disparities in price changes suggest the influence of local market conditions and policies.
* Certain property types exhibit distinct price trends, hinting at varying demand and investment potential.

**DATA CLEANING**

Data cleaning was done in order to address issues related to the quality of the dataset - to ensure that the data is accurate, consistent, and free from errors. Here are some data cleaning methods engaged in:

Checking for validity of data: The RegionName contains zipcode data. It was renamed to Zipcode.

Checking for missing values: -The missing values in the date columns were filled through interpolation and the missing values in the metro column will be replaced with 'missing'.

There were no duplicates in the data. The zipcode column underwent inspection, revealing a mix of entries with either four or five digits. To standardize the format, a right justification method was employed to prepend zeros to entries with only four digits.

In feature engineering, two new columns were introduced, Return on Investment (ROI) and Coefficient of Variation (CV).

We later converted the data into time series by melting the data from wide view to long view.

**METHODOLOGY**

This section details the methodological approach taken to analyze the dataset, including data preprocessing steps, exploratory data analysis (EDA), and the selection and application of time series modeling techniques.

**DATA PREPROCESSING**

Data preprocessing involved cleaning the dataset, handling missing values, and restructuring the data to suit time series analysis requirements. This process ensured the accuracy and reliability of the subsequent analysis.

Seasonality, alongside trend, was present and it impacts forecasting accuracy in time series analysis. Removing artifacts like interpolated data ensures accurate modeling. There was dentification of genuine seasonal patterns in the data, removal of artifacts such as interpolated data that may distort analysis and clear visualization of seasonality, trend, and residual patterns.

Stationarity was present, where statistical properties remain constant over time, is crucial. The Dickey-Fuller test validates stationarity, guiding further processing steps. There was confirmation of non-stationarity for the p-value > 0.05 and test statistic > critical values, proving necessity for further processing steps to achieve stationarity, such as differencing.

Detrending was done removing underlying trends to analyze fluctuations effectively. Differencing was used to achieve this, isolating the stationary component for analysis. Elimination of trends and seasonality through appropriate differencing.

Autocorrelation Function (ACF) and Partial Autocorrelation Function (PACF) plots reveal the correlation structure of the detrended series, guiding modeling decisions. There is autocorrelation in the time series at several lags. Therefore, the time series is non-random. There are also significant partial correlations which further continues to support that the series is not random

**EXPLORATORY DATA ANALYSIS (EDA)**

EDA was conducted to uncover underlying patterns in the data, with a focus on identifying trends, seasonality, and outliers. Visualizations such as time series plots, autocorrelation plots, and decomposition charts played a crucial role in this exploratory phase. House prices have been trending upwards from 1996-2008 until the house market crash where the house prices drastically went down and stabilized around 2012.

**MODELING**

**Time Series Modeling**

Several time series models were evaluated for their suitability in forecasting real estate prices. The analysis explored models including SARIMA (Seasonal ARIMA), and Prophet models, assessing their performance based on criteria such as the Akaike Information Criterion (AIC), Bayesian Information Criterion (BIC), and cross-validation scores.

**Analysis and Findings**

The core of the report presents the analysis conducted using the selected time series models, discussing the models' performance, the accuracy of the forecasts, and any challenges encountered during the modeling process. Insights drawn from the models are discussed in depth, highlighting their implications for understanding and predicting real estate market trends.

**Model Comparison and Selection**

A comparative analysis of the models' performance revealed key strengths and limitations, guiding the selection of the most appropriate model for forecasting. Factors considered in this comparison included predictive accuracy, computational efficiency, and the models' ability to capture the dataset's underlying patterns.

**Forecasting Results**

The selected model's forecasting results are presented, offering a forward-looking perspective on expected real estate price trends. The forecasts are discussed in the context of current market conditions, economic indicators, and potential future developments that could influence the real estate market

**FbProphet Model**

There was a difference in the forecast between our model and the prophet model. The prophet model has a higher MAPE than our dynamic model. Confidently go with the dynamic forecast. In future analysis, analyze the different zip codes to be able to answer the investors question, as with this specific zip code used, highly discourage the investor to invest there at the moment

**CONCLUSION**

**Conclusion:**

The Zillow dataset is highly suitable for addressing the real-world problem of identifying the best zip codes for real estate investment. Its comprehensive coverage of historical home values across a wide range of geographical locations, coupled with its temporal depth, provides a robust foundation for conducting time series analysis. By carefully addressing the dataset's limitations, this project can leverage the data to generate valuable insights for the stakeholder, guiding investment decisions in the dynamic real estate market.

\* There is a difference in the forecast between our model and the prophet model

\* The prophet model has a higher MAPE than our dynamic model.

\* In future analysis, we would need to do analysis of the different zip codes to be able to answer the investors question, as with this specific zip code used, we would highly discourage the investor to invest there at the moment

\* We would need to split our data into train and test set and redo the analysis with this data.

**Limitations and Implications:**

- Historical Data Only: The dataset does not include future or real-time data, limiting the analysis to historical trends and necessitating assumptions about future market behavior.

- 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. This absence might limit the comprehensiveness of the forecast.

- Missing Values: Any missing data, especially in the early years, could affect the analysis's accuracy. Careful preprocessing and imputation strategies will be required to mitigate this.

**Next Steps**

Real-Time Monitoring and Updating: Establish a framework for real-time monitoring of relevant market indicators and updating the forecasting models accordingly. This will enable the investment firm to adapt quickly to changing market conditions and make timely investment decisions.

Incorporate External Factors: Integrate external economic indicators such as interest rates, employment rates, and GDP growth into the analysis to provide a more comprehensive understanding of the real estate market dynamics. This could involve obtaining relevant datasets and merging them with the existing Zillow dataset for enhanced forecasting accuracy.

Validation and Sensitivity Analysis: Conduct thorough validation and sensitivity analyses to assess the robustness of the model outputs. This involves testing the models against different scenarios, adjusting key parameters, and examining the stability of results to ensure the reliability of investment recommendations.