Study Problem & Hypothesis:

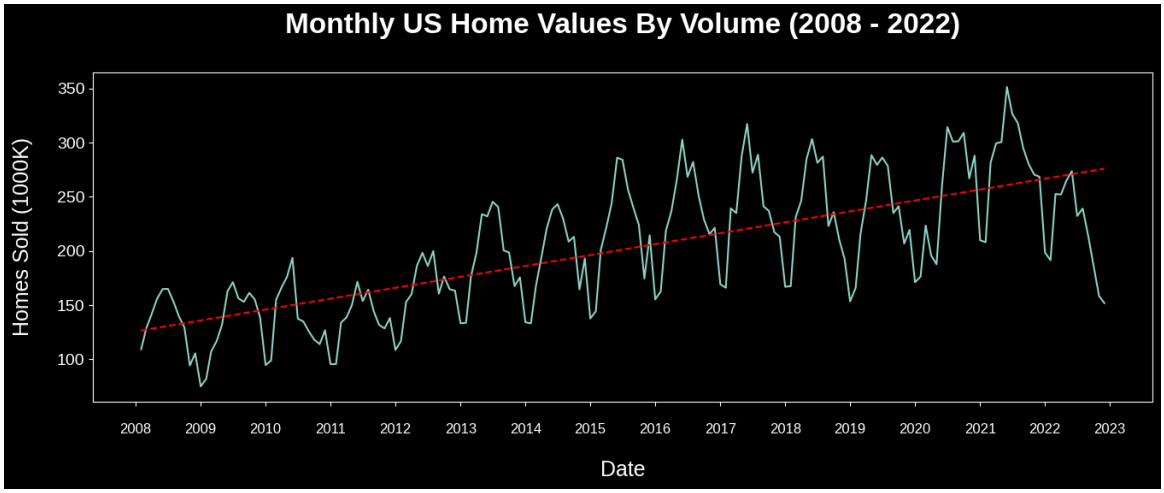
Using data from Zillow Research, can home sales in the United States be effectively forecasted based solely on research data?

> **Null Hypothesis:** An effective predictive time series forecasting model with a mean absolute percentage error of < 20% cannot be generated from the dataset.
>
> **Alternative Hypothesis:** An effective predictive time series forecasting model with a mean absolute percentage error of < 20% can be generated from the dataset.

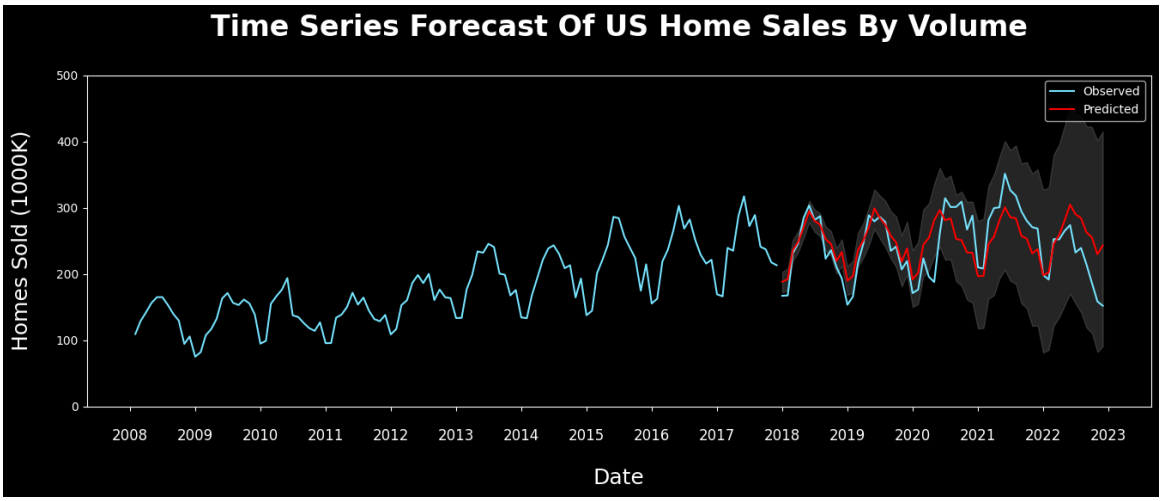
Historical data pertaining to home sales in the United States materializes after a residential property's transfer of ownership via the deed or its formal recording, culminating when financial transactions have been consummated. The development of a predictive model for prospective home sales augments operational efficiency for all stakeholders engaged in the acquisition process, enriching strategic planning capacities across various strata of the supply chain.

Data Analysis Process:

The dataset comprises conidia, cooperative, and single-family residential properties that underwent transactions within the foremost United States Metropolitan Statistical Areas, meticulously assessed at monthly intervals. An initial exploration of the data unveiled a gradual yet consistent seasonal ascent, persisting until the conclusion of 2022, at which point home sales declined precipitously, surpassing the threshold of what was deemed 'anticipated'.



The dataset was bifurcated to establish both a training and testing dataset, with the training dataset encapsulating home sales data from the years 2008 to 2017, while the testing dataset encompassed data from 2018 to 2022. Various forecasting models, including ARIMA and SARIMA, were employed, albeit with limited success. Subsequently, an automated methodology leveraging the Facebook Prophet time series forecasting package was pursued, resulting in enhanced outcomes. Through meticulous hyperparameter tuning, a final optimized model was realized.



The forecasting model achieved a mean absolute percentage error of 12.7%, decisively refuting the null hypothesis in favor of the alternative.

Study Findings:

As previously stated, the refined model displayed a mean absolute percentage error of 12.7% when applied to the test dataset spanning from 2018 to 2022. This achievement comfortably surpassed the predefined threshold of 20% or less for a mean absolute percentage error, thus affirming the model's effectiveness and warranting the rejection of the null hypothesis. The model's forecasting precision exceeded my initial expectations, particularly considering the presence of exogenous variables.

Study Limitations:

The primary limitation for this analysis was obtaining home sales data from a CSV file. Preparation of data took longer than expected, having to understand the structure first, then determining the best methods to transform the data into a usable format, to analyze and support time

series forecasting. For a one-time analysis, this might be acceptable, but if the objective is to develop a time series forecasting model to rely on for business decisions, it is not going to be feasible for data analyst nor management.

The second limitation of using a downable csv file is the potential for low quality and quantity data. In these analyses we wanted to build an efficient and reliable forecasting model based on US single-family homes. Although the data did contain single-family based homes, it also included condominiums and co-operative homes with no distinction as to how many sold homes belonged to each category monthly. There is a big difference between single-family homes and condominiums in which one is a permanent residential home while the other tends to be utilized as either an Airbnb or as a vacationing seasonal home. If management wants to make a business decision based on specific variables, it wouldn't make sense to include additional variables that could affect the accuracy of a prediction model or worse, unable to confidently accept or reject a null hypothesis. Another minor limitation is the inability to create automation when it comes to data transformation and calculations because the file can always come from a different source or be in a different format.

In reference to the techniques implemented on the result of having a limited dataset, it led to having a reduction in model complexity to where hyperparameter tunning needed to be extensive to be able to handle the significant drop of home sales towards the end of year 2022. Worst of all, based on how the dataset was divided between training and testing, the exogenous factors in the data ended up being the years at the end of the testing dataset. For the model to be properly trained, it needs enough data to not only know how to handle trending and seasonal patterns, but also patterns derived from events in which nobody can foresee such as COVID-19.

Recommended Action:

The visual representations of the data indicate a consistent upward trajectory until the year 2016, characterized by seasonal troughs that consistently remained above the previous year's low point. However, post-2016, while the upward trend persisted, the high and low points began to exhibit fluctuations that fell outside the confines of predicted confidence intervals. Notably, in the summer of 2021, a peak in home sales was observed, followed by a precipitous decline that exceeded the anticipated lower bound of the confidence interval. This abrupt disruption in both the seasonal and upward trends, which had its origins in 2008, raises significant concerns within the real-estate market.

To comprehensively address this issue, a rigorous causal analysis is imperative to unravel the underlying reasons for the heightened volatility in home sales. This multifaceted inquiry encompasses an examination of historical data predating 2008, aiming to ascertain if the observed pattern is an integral part of a broader historical context or a recent phenomenon. Additionally, it entails an in-depth analysis of buyer data to assess whether a reduction in the number of homes being listed for sale plays a role. Furthermore, a thorough evaluation of mortgage loans and interest rates may offer valuable insights.

Given the extensive scope of this project, a prudent approach involves dividing it into a series of smaller, well-defined analyses. These smaller scopes can be pursued concurrently by different teams, ultimately converging into a comprehensive causal analysis.

Expected Benefits:

The anticipated advantages of this forecasting model revolve around the optimization of planning for various stakeholders directly engaged in the real estate sector. Specifically, these benefits encompass:

Homebuyers: Empowering them to strategically time their property acquisitions, ensuring optimal market conditions for their investments.

Investors: Facilitating well-informed decisions regarding the timing and location of real estate investments, leveraging accurate forecasts to maximize returns.

Home Price Predictions: Adapting the model to forecast home prices, guiding investment choices by signaling expected price increases or decreases in specific areas.

Lenders and Financial Institutions: Assessing lending risk based on forecasts, enabling timely adjustments to lending standards in response to potential market fluctuations.

Government Agencies and Policymakers: Utilizing forecasts to shape housing-related policies, such as infrastructure development initiatives and affordable housing programs, in alignment with market dynamics.

Homebuilders: Optimizing inventory management by aligning production with forecasted demand, mitigating the risk of overbuilding or underbuilding in fluctuating markets.

Economic Impact Assessment: Providing an accurate gauge of the real estate sector's economic influence, encompassing its contribution to GDP and job creation, thereby enhancing economic planning.

These benefits collectively underscore the model's potential to drive efficiency, informed decision-making, and strategic planning across the spectrum of real estate-related activities.

Sources:

Housing data - Zillow Research. (2023, April 25). Zillow. https://www.zillow.com/research/data/

Ma, B. (2021c, December 13). Time Series Modeling with ARIMA to Predict Future House Price. Medium. https://towardsdatascience.com/time-series-modeling-with-arima-to-predict-future-house-price-9b180c3bbd2f

Maverickss. (2023). Time Series Forecasting using ARIMA/SARIMA/SARIMAX. Kaggle. https://www.kaggle.com/code/maverickss26/time-series-forecasting-using-arima-sarima-sarimax

Li, S. (2018, November 30). An End-to-End Project on Time Series Analysis and Forecasting with Python. Medium. https://towardsdatascience.com/an-end-to-end-project-on-time-series-analysis-and-forecasting-with-python-4835e6bf050b

Cphalpert. (n.d.) census-regions/us census bureau regions and divisions.csv GitHub. https://github.com/cphalpert/censusregions/blob/master/us%20census%20bureau%regions%20and%20divisions.csv