Real Estate Investing Potential Report

Gettysburg College - DS 325

Lorenzo Zullo

Abstract

This project aims to provide insights for real estate investors, leveraging machine learning and data analytics to identify high-growth, undervalued markets. Using *Zillow's* ZHVI data and *Realtor* property listings, Lasso, Ridge, and Linear Regression models predict home prices and assess key investment factors. I ranked cities based on growth potential, undervaluation, and market volatility, creating an Investment Score to highlight top opportunities. Advanced visualizations, including heatmaps, choropleth maps, and scatter plots, provide a comprehensive geographic view of investment potential. The final analysis pinpoints high growth rate, low price-to-ZHVI ratio, and low price volatility on a map. This serves as a data-driven guide for investors seeking to maximize returns while minimizing risk in the real estate market.

Introduction

Investing in real estate is one of the most lucrative ways to build wealth, but identifying the best markets to invest in can be incredibly challenging. With housing prices fluctuating, markets becoming increasingly competitive, and investment risks varying across regions, investors often struggle to determine where and when to buy. Traditional methods of real estate analysis like intuition, local knowledge, and outdated data, can lead to missed opportunities or costly mistakes.

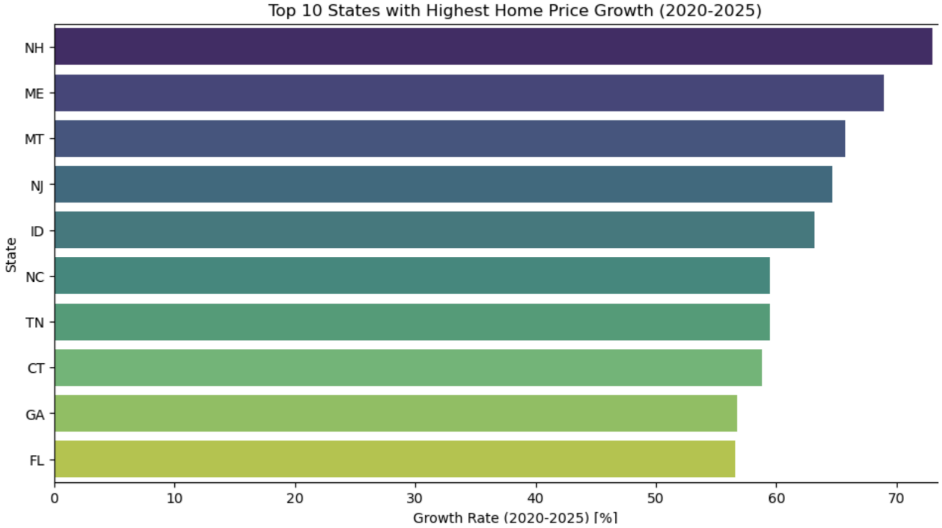
This project tackles that problem head-on by leveraging data-driven insights to pinpoint undervalued markets with high growth potential. By analyzing millions of real estate transactions alongside Zillow’s Home Value Index (ZHVI), we show cities that have the highest investment potential, helping investors make informed decisions. We rank cities based on price-to-value ratios, projected growth, and market stability (volatility), ensuring that investors can maximize return on investment (ROI) while minimizing risk. For real estate professionals, new investors, and those looking to capitalize on market trends, this project offers a data-backed approach to finding the best opportunities in an unpredictable housing market.

Background

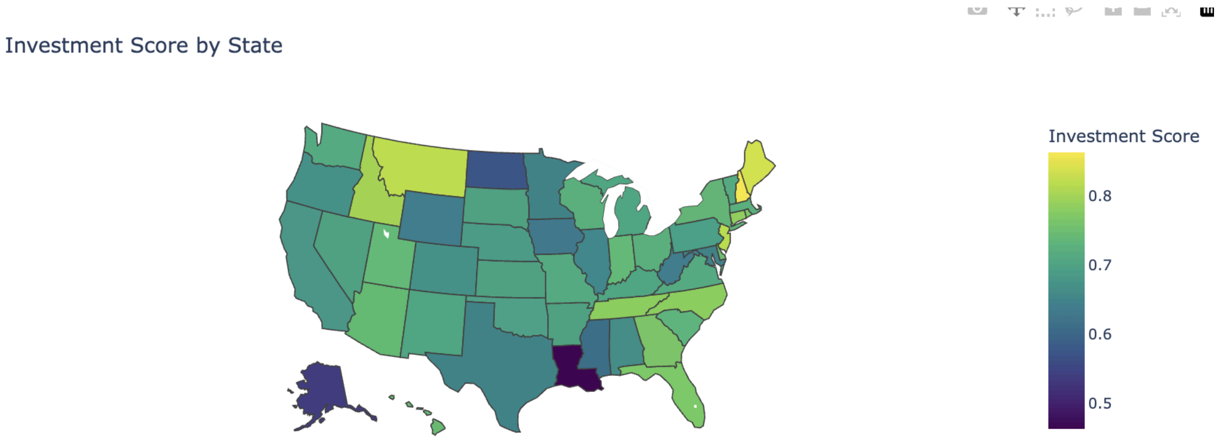
The real estate market has long been an area of interest for economists, investors, and data scientists seeking to understand price fluctuations, investment potential, and risk factors. Machine learning has revolutionized real estate price prediction, allowing investors to better assess property values and market trends. Traditional valuation models often rely on historical trends and economic indicators, but research has shown that machine learning algorithms like random forests and gradient boosting can significantly improve price accuracy (Ho et al., 2021). These models analyze vast datasets, identifying patterns that traditional approaches might miss, making them valuable tools for predicting undervalued or overvalued properties. By leveraging data-driven insights, investors can make smarter purchasing decisions and reduce uncertainty in fluctuating markets.

The integration of big data analytics has further transformed real estate decision-making, providing investors with real-time insights into market conditions. ZHVI is a prime example of how large datasets help track price trends, offering granular data on market appreciation and decline (Gale & Sen Roy, 2023). Studies have shown that leveraging neural network-driven price indices can help pinpoint high-growth areas and optimal investment windows. By combining machine learning, volatility analysis, and big data insights, this project provides a comprehensive framework for identifying the most promising real estate investments.

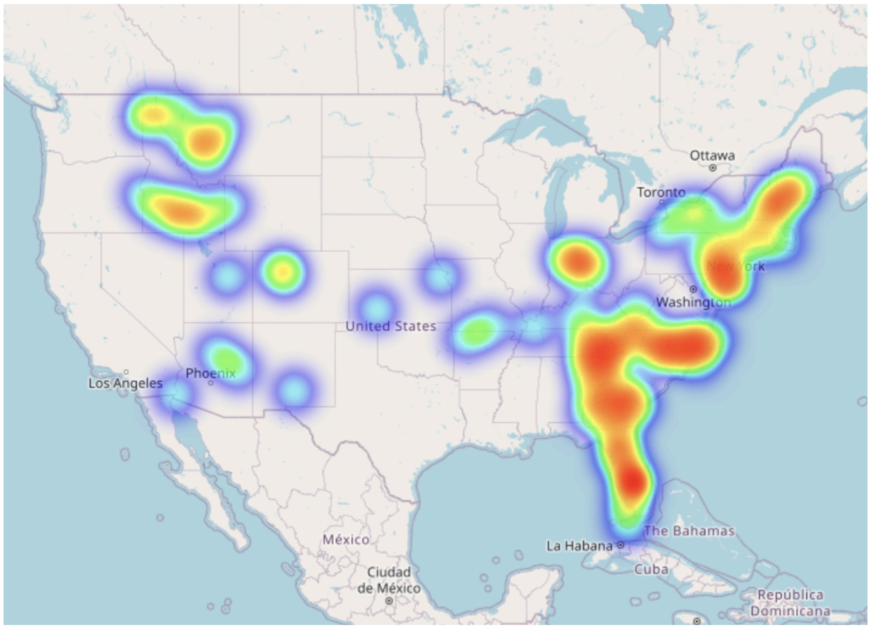
Data



This bar graph displays the top 10 states with the highest home price growth. We can see how New Hampshire has the highest growth rate from 2020 to 2025 with around 72%. Home Price Growth is calculated by taking the difference between the 2020 ZHVI and the 2025 one.



This represents the investment score of each state. We take a weighted average of growth, undervaluation, and volatility to calculate this for each city. From there we average the values for the state and display them using a choropleth map (used: Plotly Technologies Inc., n.d.).



The heatmap visualizes the top 100 investment score properties, with hotspots concentrated in Florida and the East Coast. These regions showed high growth potential and relatively undervalued properties, making them prime targets for real estate investment. I used *Gopesh* to understand how to use Geopy for geocoding, *Stack Exchange* to fix timeout error, and *Real Python* to bring it all together in the folium map (2022, 2015, 2025).

Methods

We used two primary datasets: the USA Real Estate Dataset (Sakib, 2023) and Zillow's Home Value Index (Zillow Research, 2025). Preprocessing included handling missing values, feature engineering (e.g., price-to-ZHVI ratio and growth rate), encoding categorical variables, and removing outliers to ensure data quality. Features such as bed, bath, house_size, acre_lot, ZHVI_latest, and growth_2020_2025 were selected based on correlation analysis.

To predict home prices, we implemented Linear, Lasso, and Ridge Regressions, applying standardization and hyperparameter tuning (via cross-validation) for optimal performance. The best model was chosen based on R^2, Mean Absolute Error (MAE), and Root Mean Squared Error (RMSE). Additionally, we developed an Investment Score to rank cities by growth potential, affordability, and market stability, helping investors pinpoint high-opportunity markets. The results were visualized using choropleth maps, scatter plots, and bar charts to provide a data-driven real estate investment framework.

Results

Ridge Regression performed best, achieving an R^2 score of 0.57775. However, the high MAE ($116K) and RMSE ($172K) suggest that the models struggled with price variability, particularly for luxury homes and volatile markets. The models performed well in predicting median-priced homes however, with bathroom count and ZHVI being strong indicators. However, outliers, high-end properties, and volatile markets led to large prediction errors, suggesting that additional location-based features or more advanced modeling techniques (e.g., ensemble methods) may improve performance. Beyond price prediction, we ranked top investment markets based on growth potential, undervaluation, and stability, visualizing key insights through choropleth maps, scatter plots, and bar charts to guide investors toward high-opportunity markets while flagging riskier regions.

Conclusions

This project set out to solve a critical challenge in real estate investing, identifying the best markets for property investment based on growth potential, undervaluation, and market stability. By leveraging machine learning models and big data analytics, we can predict home prices and rank markets using an Investment Score. The results highlight high-growth, undervalued markets, helping investors make data-driven decisions rather than relying on intuition or outdated methods.

Key takeaways include the importance of historical price trends (ZHVI) and market growth rates in predicting investment potential. While the Ridge Regression performed best, outliers and volatile markets posed challenges, suggesting that more advanced models like Random Forest, could improve predictions. Our heatmap visualization pinpointed investment hotspots, particularly in Florida, the East Coast, and parts of the Southwest, reinforcing the value of spatial analysis in real estate investing.

One challenge was handling luxury home price variations, which skewed predictions. Future research could explore property-type segmentation (e.g., single-family vs. multifamily homes) and incorporate macroeconomic indicators (e.g., mortgage rates, job growth). Additionally, integrating geospatial data beyond city/state level—such as neighborhood-specific insights—could further refine investment predictions. Ultimately, this project provides a strong foundation for data-driven real estate investing, offering investors the tools to make informed, strategic decisions in an evolving market.

References

Gale, H., & Sen Roy, S. (2023). *Optimization of United States residential real estate investment through geospatial analysis and market timing.* Applied Spatial Analysis and Policy.

Gopesh. (2022, March 9). *Geocoding with Python using Nominatim.* Medium. https://gopesh3652.medium.com/geocoding-with-python-using-nominatim-a-beginners-guide-220b250ca48d

Ho, W. O., Tang, B. S., & Wong, S. W. (2021). *Predicting property prices with machine learning algorithms.* Journal of Property Research.

Plotly Technologies Inc. (n.d.). *Choropleth maps in Python using Plotly.* https://plotly.com/python/choropleth-maps/

Real Python. (2025). *Creating web maps from data with Python and Folium.* https://realpython.com/python-folium-web-maps-from-data/

Sakib, A. S. (2023). *USA real estate dataset.* Kaggle. https://www.kaggle.com/datasets/ahmedshahriarsakib/usa-real-estate-dataset

Stack Exchange User. (2015, October 29). *Avoid time-out error Nominatim Geopy OpenStreetMap.* GIS Stack Exchange. https://gis.stackexchange.com/questions/173569/avoid-time-out-error-nominatim-geopy-openstreetmap

Zillow Research. (n.d.). *Zillow housing data and real estate market trends.* https://www.zillow.com/research/data/