Quantitative Analysis of Urban Housing Markets

Location: *Southwest Florida*

Authors: Chris Jose Castaneda

Date: 16 May 2025

Table of Contents

1. Problem Statement

Southwest Florida's coastal communities promise a desirable lifestyle marked by scenic beaches, warm climates, and a serene suburban atmosphere. Yet, beneath this attractive facade lies a real estate market characterized by volatility and complexity. Originally setting our sights on major metropolitan areas such as Miami, Chicago, and New York City, we quickly encountered challenges that shifted our analytical focus towards Collier and Lee counties. This pivot provided us with an ideal environment which is both dynamic enough to reflect broader market trends and localized enough to manage complexities effectively.

At the heart of our analysis is the goal of understanding what drives housing prices at a granular, listing-level detail. We aim to pinpoint the key factors influencing price per square foot, whether they're structural attributes such as bedrooms, bathrooms, and square footage, or broader economic indicators like market conditions and mortgage rates. By systematically addressing these considerations, our objective is to provide prospective home buyers with clear, actionable insights, enabling informed decision-making amid rapidly changing market dynamics.

1.1: File Overview

Before you move on, here are all the files that are inside our zip folder. Please read this file overview to understand the purpose of each file

- **app.R**: R Shiny application providing interactive map visualizations, home affordability calculations, and forecasting trends for Southwest Florida housing data. This is for potential clients who are interested in buying a home.
- **april13_listings_raw_listings.csv**: Initial CSV file capturing raw, unprocessed listings data directly after web scraping.
- **Collier_County_Price_Index.csv**: Historical monthly house price index data for Collier County, providing essential economic context and insights into local housing market trends.
- **Data Cleaning.ipynb**: Python notebook performing extensive data preprocessing, including cleaning raw listings, handling missing values, and feature engineering to prepare datasets for analysis.
- **Economic_Parameters.ipynb**: Python notebook retrieving, processing, and visualizing macroeconomic variables such as mortgage rates, Dow Jones, SPY indices, and home price indices for Collier and Lee counties.
- **Lee_County_Price_Index.csv**: Historical monthly house price index data for Lee County, facilitating comparative analysis of housing market conditions within Southwest Florida.
- **Modeling.ipynb**: Python notebook conducting statistical analyses and modeling, including ordinary least squares regression and ARIMA forecasting, aimed at understanding price determinants and market dynamics.

- **MORTGAGE30US.csv**: Historical data on U.S. 30-year fixed mortgage rates, used to assess affordability trends and to model the impact of changing financial conditions on housing market dynamics.
- **SWFL_Data.csv**: Comprehensive dataset of Southwest Florida real estate listings post initial scraping and preliminary data collection.
- **SWFL_Data_Cleaned_Final_Version.csv**: Final cleaned version of the Southwest Florida real estate dataset, including refined features used for in-depth analysis and visualization in the Shiny application.
- **updated_mls_scraper.py**: Python script implementing a robust web scraping pipeline with Selenium to systematically extract detailed property listings from Realtor.com, featuring retry logic, XPath extraction, deduplication, and state management.
- **Zipcode.py**: Python script designed to geocode property addresses into latitude and longitude coordinates using the Geopy library, with a fallback mechanism leveraging U.S. Census data for enhanced accuracy and completeness.

2. Data Collection Methodology

2.1 Web Scraping Pipeline and Evolution

Initially, our data collection strategy targeted prominent real estate platforms such as Zillow to acquire comprehensive listings from major cities like Miami, Chicago, and New York City. However, we quickly faced significant technical challenges. Frequent CAPTCHA security measures on Zillow impeded automated scraping, limiting our access to the extensive datasets necessary for robust analysis. These persistent scraping limitations (See Figure 1) compelled us to reconsider our initial geographic scope.

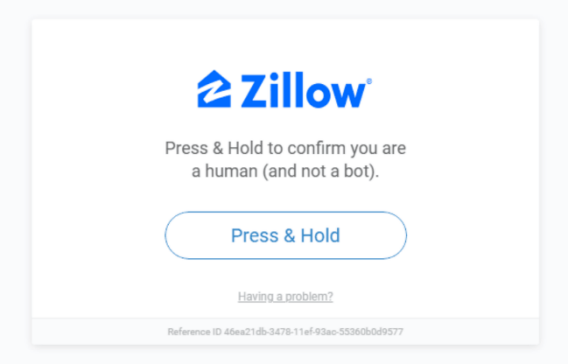


Figure 1- Zillow Web Scraping Captcha Error

To overcome these obstacles, we strategically pivoted toward RealtyOfNaples.com, a platform offering more stable and reliable access for web scraping. Our robust scraping pipeline allowed for more volume and accurate data collection, and comprehensive deduplication procedures to ensure data integrity. Additionally, geocoding was conducted using the Geopy library to ensure precise property location assignments which were aggregated onto the csv file that held the contents of scraped data.

2.2 Public Indices Retrieval

Furthermore, we downloaded more data from the Federal Reserve, specifically targeting Collier and Lee County homes in Southwest Florida, provided a focused and manageable scope, enabling thorough and meaningful analysis. The data contained home price index from both counties. We also downloaded the average 30-year mortgage rate across homes in the U.S as an indicator that could influence price. Lastly,

we achieved stock data looking at the Dow Jones and S&P500 to whether or not periods of economic downturn or prosperity would affect interest rates which in turn would make us wonder if that would lead to a price change.

3. Web Scraping & Data Process

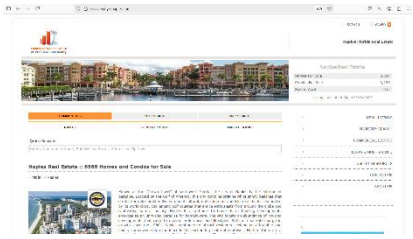


Figure 2 - Web Page of www.RealtyofNaples.com

3.1 Web Scraping

Using RealtyOfNaples.com, we found that it is a platform offering more stable and reliable access for web scraping. Our robust scraping pipeline, developed using Selenium WebDriver with the Firefox Gecko driver, incorporated advanced retry logic to handle network interruptions, precise XPath extraction methods for consistent and accurate data collection. We were able to scrape using the Gecko driver was sourced from the official Mozilla repository (https://github.com/mozilla/geckodriver/releases). The data scraped would be stored in a csv file that was messy. You can find it in the zip folder titled “updated_mls_scraper.py”/

3.2 Extract, Transform, Load (ETL)

After initial web scraping, the extracted data required extensive processing to become analytically useful. Initially stored in a raw, unstructured CSV format ("april13_listings_raw_listings.csv"), the data presented various challenges, including inconsistent formatting, missing values, and extraneous textual information mixed within key numerical fields. To address these issues, we developed a structured ETL pipeline detailed within the Jupyter notebook file titled "Data Cleaning.ipynb." The ETL process involved systematic parsing of prices, removing currency symbols and delimiters, accurately extracting property attributes such as the number of bedrooms, bathrooms, and total square footage, and generating calculated metrics, notably the price per square foot (ppsf). Rigorous data validation steps were implemented to remove outliers which can also be seen by the box plot in Figure 3.

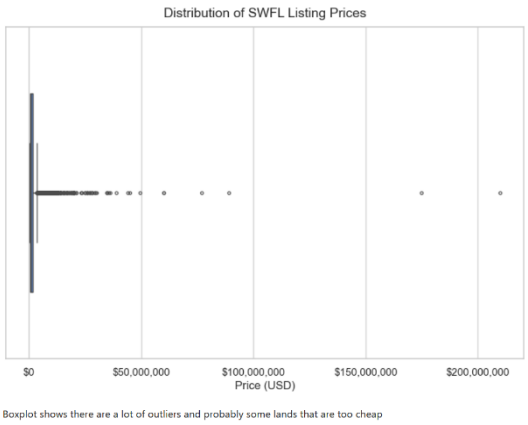


Figure 3- Boxplot Capturing Outliers

One way we eliminated outliers is to see listings that had 0 beds and 0 bathrooms. We made assumptions that these were plots of land as the first 3 listings were plots of land. We also removed outliers that were way above the final quartile such as the most expensive listing that was listed for $210,000,000 since this can skew our findings. It’s important that it’s rare for someone to even attempt a bid, hence it was not a loss. We also removed any listing that was classified as a “Hotel”, “Condo”, “APT” or “#” as these are not homes. Again, we are looking for homes only. We managed to reduce the size of the data by more than half. In addition, we converted community_types to binary (0s and 1s) because there were only two options: Gated or Subdivision/Non-gated. The most difficult process was separating the amenities column

as these were unique amongst a lot of listings but important to look at. We were curious whether the number of amenities, certain zip codes / communities, or even coordinates could affect price

Following initial cleaning, further refinement was performed to ensure geographical accuracy and consistency. The file "Zipcode.py" facilitated geocoding property addresses into precise latitude and longitude coordinates, leveraging the Geopy library complemented by a U.S. Census fallback for enhanced accuracy. After these critical transformations and validations, the finalized dataset was saved as "SWFL_Data_Cleaned_Final_Version.csv," a structured and comprehensive dataset primed for detailed exploratory analyses, statistical modeling, and visualization tasks. This meticulously cleaned dataset serves as the backbone for subsequent modeling phases and the interactive dashboard developed in R Shiny ("app.R") which was our final step. Again, there is more information within the notebooks!

4. Exploratory Data Analysis

4.1 Descriptive Statistics

Our analysis began by assessing key descriptive statistics to understand the distribution and central tendencies within our dataset. We calculated vital metrics, including the count of listings, mean and median price per square foot (ppsf), and the standard deviation which is in the Jupyter notebook labeled “Data Cleaning.ipynb”. These metrics revealed essential insights into market pricing dynamics and variability across listings. Histograms and box plots generated in our notebooks visually confirmed the right-skewed distribution of ppsf, highlighting outliers and areas of high concentration.

4.2 Geospatial Patterns

Geographic visualization significantly enhanced our exploratory analysis, allowing us to clearly depict spatial variations in housing prices across Southwest Florida. We created two distinct interactive maps leveraging latitude and longitude coordinates from our geocoding pipeline. The primary and most informative map utilized clustering techniques to visualize property listings, color-coded by their price per square foot (ppsf). This clustering effectively highlighted areas of higher market activity and pricing

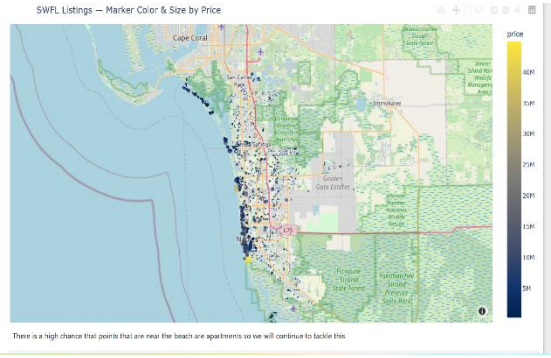


Figure 5 – First Interactive Map

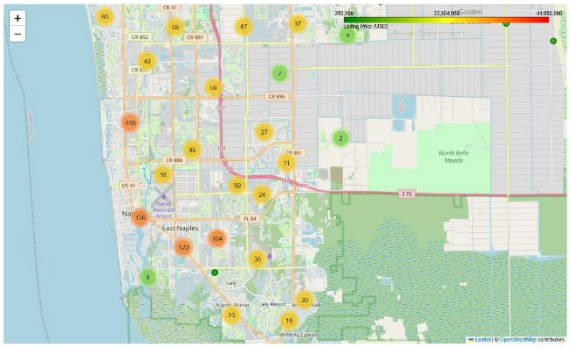


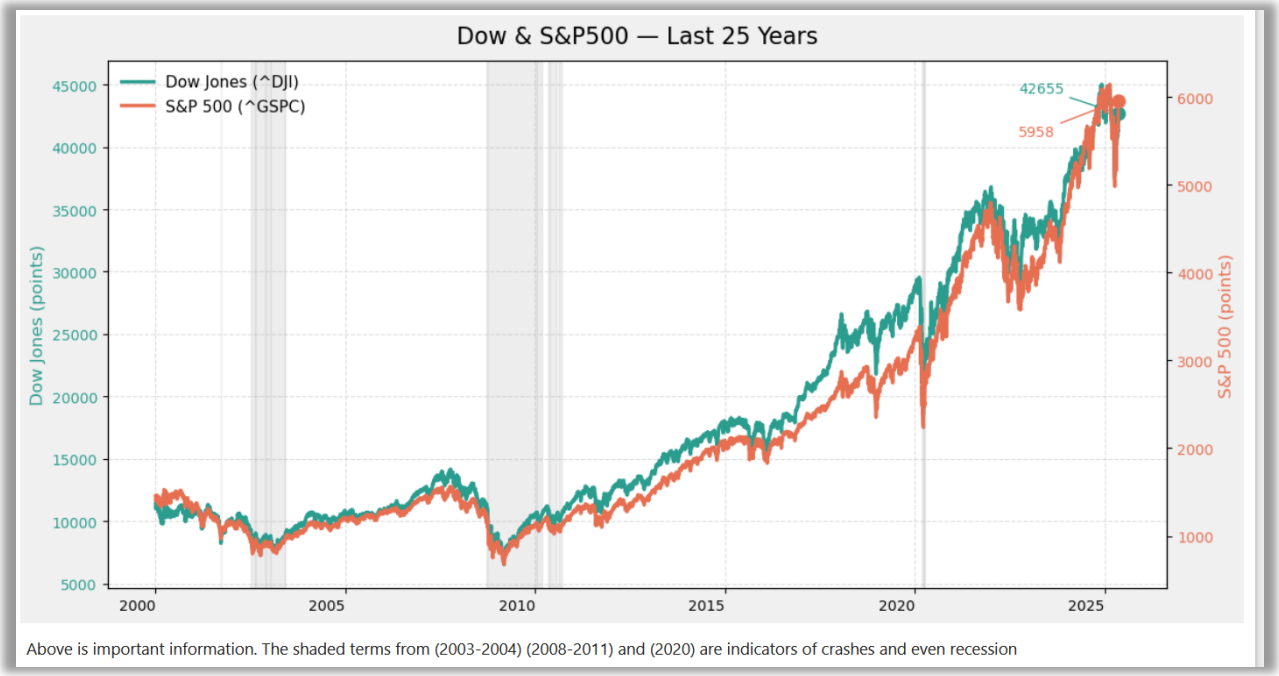
Figure 4 - Interactive Clustered Map Across Southwest Florida

hotspots, particularly evident near coastal and urban centers. The second map complemented this analysis by providing broader visual context of regional distribution. Together, these visual tools helped intuitively illustrate pricing disparities between premium coastal neighborhoods and more affordable inland communities. It could already be seen that

certain zip codes had more expensive homes. Fill free to interact with it on the Jupyter notebook titled "Data Cleaning.ipynb" or the shiny R "App.R" file.

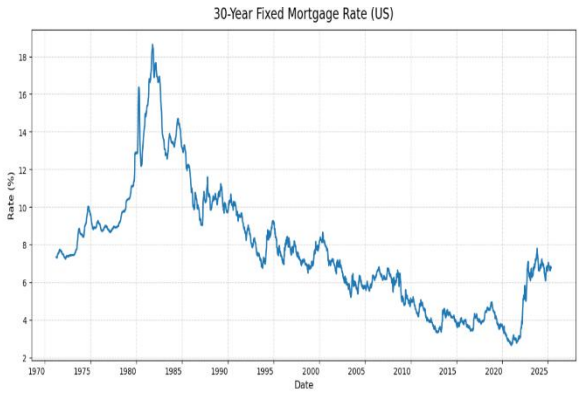
4.3 Temporal Trends

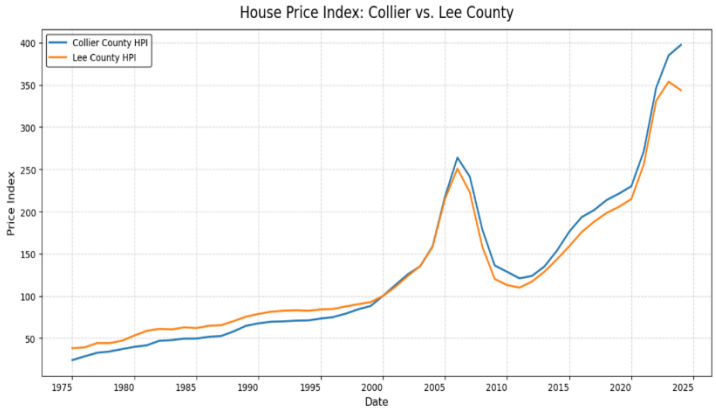
We analyzed temporal trends by visualizing long-term economic indicators alongside local housing market data to capture both macroeconomic influences and regional price dynamics. The historical trends of major financial indices, specifically the Dow Jones and S&P 500, were examined to identify broader market regimes, with particular attention to recessionary periods marked by sharp declines, such as those following the 2001 dot-com crash and 9/11 attacks, the 2008 financial crisis, and the 2020 COVID-19 pandemic. Understanding these periods is crucial, as macroeconomic shifts directly impact housing affordability and consumer confidence.



Furthermore, the local House Price Indices (HPI) for Collier and Lee counties provided a clear visual narrative of regional housing market fluctuations, illustrating significant growth, downturns, and recovery periods, especially around the 2008 recession and post-pandemic surges. Complementing this analysis, the historical visualization of the 30-year fixed mortgage rates revealed interest rate trends that influence purchasing power, mortgage affordability, and ultimately, home buying decisions. Together, these temporal analyses equipped us with a deeper understanding of how broader economic conditions interplay with local housing trends, critical for informed real estate decision-making.

4.4 Market Regime Analysis





We generally see a similar batter be similar
during the home price index. In the 30 year mortgage rates we see them increase leading up to the dot com crash, housing market crash and COVID.

5. Statistical Modeling Approach

5.1 Cross-Sectional Regression Analysis

Our statistical modeling process began with exploratory visualizations such as correlation heatmaps to identify relationships among potential predictors. Numerous linear regression models were evaluated, and through careful iteration, we determined the most robust model incorporating critical predictors and their interactions, as illustrated in the provided regression results.

The final Ordinary Least Squares (OLS) regression equation is expressed as seen in Figure 6:

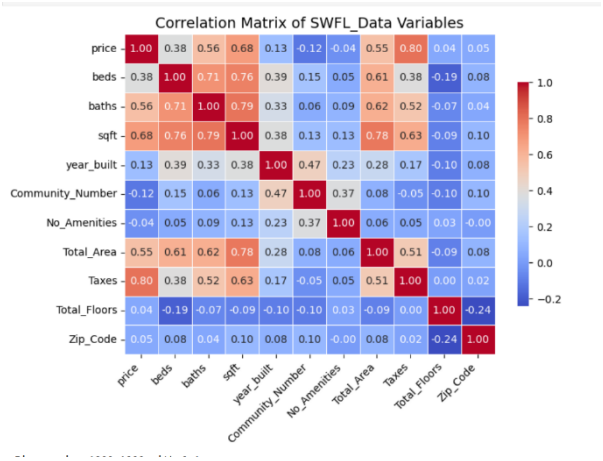


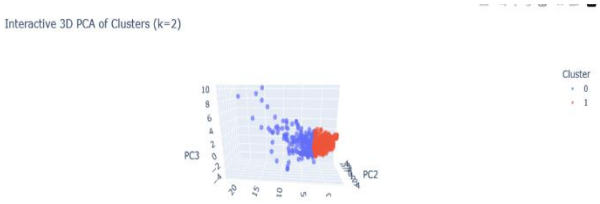
Figure 7 – Heatmap (Refer to the one in "Modeling.ipynb" instead of the one in "Data Cleaning.ipynb"

$$\begin{aligned}\text{price} = {} & \beta_0 \\ & + \beta_1\,(\text{baths}) \\ & + \beta_2\,(\text{sqft}) \\ & + \beta_3\,(\text{No\_Ammenities}) \\ & + \beta_4\,(\text{CommunityTypeCode}) \\ & + \beta_5\,(\text{sqft} \times \text{baths}) \\ & + \beta_6\,(\text{sqft} \times \text{beds}) \\ & + \beta_7\,(\text{beds} \times \text{baths}) \\ & + \beta_8\,(\text{sqft\_wetbar}) \\ & + \beta_9\,(\text{taxes\_zipcode}) \\ & + \varepsilon\end{aligned}$$

Figure 6 - Final OLS w/ Interaction Terms written in LaTeX

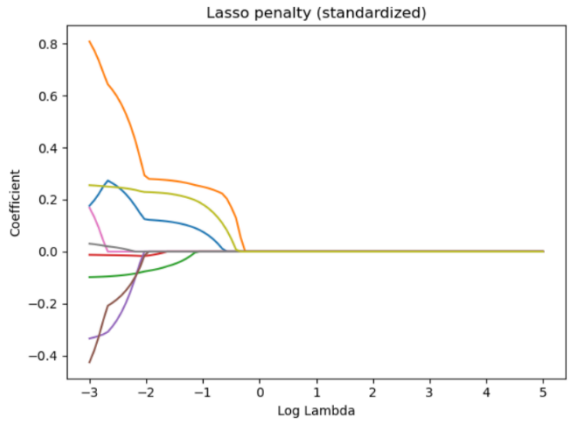
This comprehensive model captured approximately 77.8% of the variation in housing prices, highlighting the significance of interactions between structural attributes like square footage, bedrooms, and bathrooms.

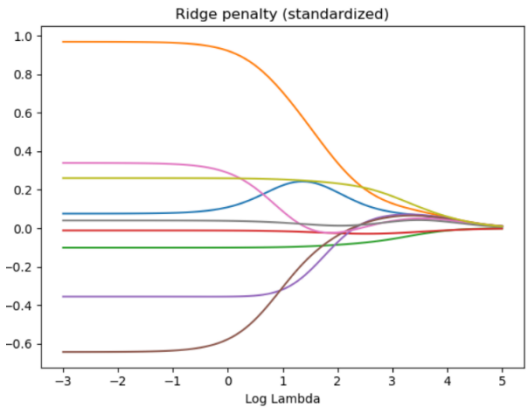
To further validate our findings and explore complex, nonlinear relationships within our data, we implemented advanced machine learning techniques. We began with Ridge and Lasso regression, visualized in coefficient paths plots that clearly illustrate the balance between model complexity and the significance of individual predictors. The Ridge plot indicated gradual shrinkage of coefficients, retaining multiple variables with diminished influence, whereas the Lasso plot highlighted aggressive feature selection, entirely eliminating less significant variables as regularization intensified.



Additionally, we conducted a clustering analysis using Principal Component Analysis (PCA), effectively grouping property listings into distinct segments based on underlying feature relationships. This method offered a clear visual differentiation between groups, suggesting distinct market segments and price-setting mechanisms within our dataset.

To benchmark predictive performance, we compared multiple ensemble and neural network methods,





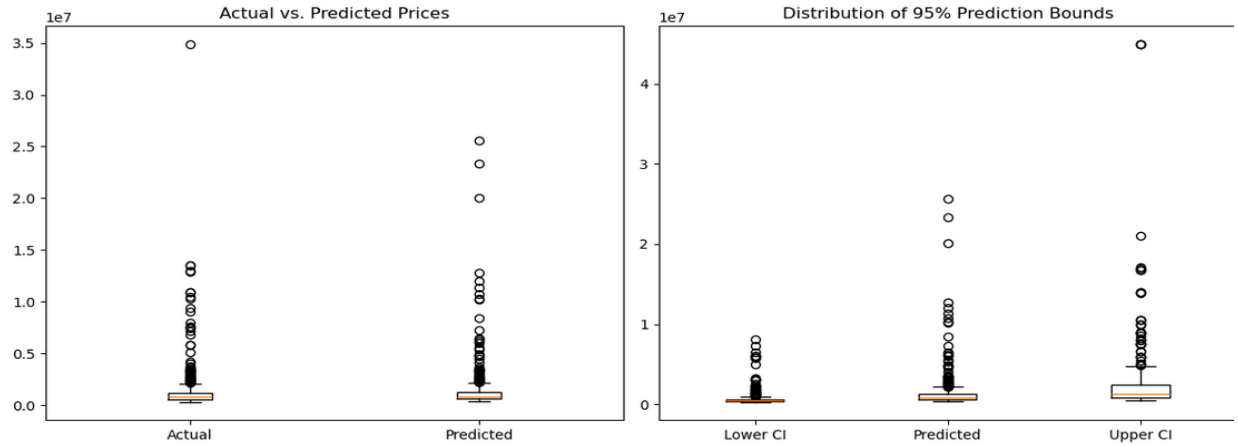
including Random Forest, XGBoost, LightGBM, and a Dense Neural Network (DenseNN). Random Forest and LightGBM notably outperformed others, achieving high R^2 and lower RMSE scores, indicative of strong predictive accuracy and generalization. Finally, predictions generated from these advanced models were supplemented by 95% confidence intervals. The scatter and box plots comparing actual versus predicted prices confirmed the robustness and reliability of our predictions, providing valuable, practical guidance for prospective homebuyers navigating the complexities of the Southwest Florida housing market

Model	Normalized Accuracy	R2	RSME
Random Forest	*96.5%*	*0.743*	*1.21E6*
XGBoost	*96.9%*	*0.642*	*1.43E6*
LGBM	*96.1%*	*0.684*	*1.34E6*
DenseNN	*95.3%*	*0.533*	*1.63E6*

The table summarizes the performance comparison across four advanced machine learning models used to predict housing prices in Southwest Florida. Each model—Random Forest, XGBoost, Light Gradient Boosting Machine (LGBM), and Dense Neural Network (DenseNN)—is evaluated on three key metrics: Normalized Accuracy, R^2 (coefficient of determination), and Root Mean Squared Error (RMSE). Normalized Accuracy reflects the relative predictive accuracy across models, with values consistently

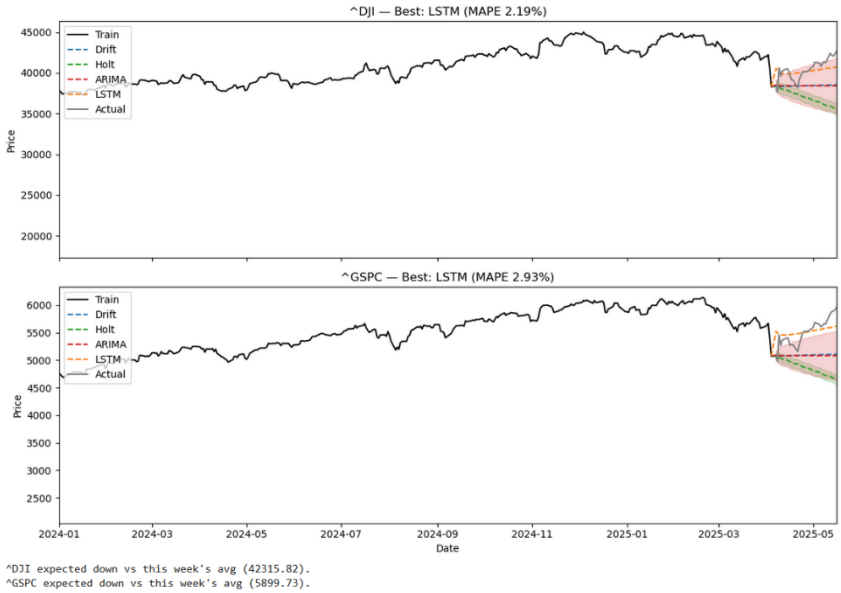
high (above 95%), demonstrating each model's effectiveness in general predictions. XGBoost achieves the highest normalized accuracy at 96.9%. The R^2 value measures how well each model explains the variability in housing prices. Random Forest notably outperforms with an R^2 of 0.743, indicating it explains approximately 74.3% of price variance, suggesting strong predictive capability compared to others. The RMSE metric quantifies the average prediction error in absolute terms, where lower values indicate more precise predictions. Random Forest again leads with the lowest RMSE (1.21E6), confirming its superior predictive accuracy relative to the other models tested.

Overall, this comparison highlights the Random Forest model as particularly robust for price prediction in this context, providing reliable insights for decision-making in the Southwest Florida housing market.



Finally, the boxplots comparing actual versus predicted prices and the associated 95% prediction intervals illustrate strong alignment between model predictions and real-world data. Notably, the visualizations indicate consistency in predicted values, though some upper-bound predictions capture potential outliers, reinforcing the model's practical reliability for pricing insights.

5.2 Time-Series Forecasting



In our time-series forecasting analysis, we employed multiple forecasting methods to predict future price dynamics for Collier County's housing market, as illustrated in the first visual. Methods used included Simple Exponential Smoothing (SES), ARIMA, Vector Autoregression (VAR), and various neural network approaches such as LSTM, RNN, and NAR. The forecasts highlighted varying degrees of uncertainty, depicted by shaded

confidence intervals. ARIMA notably demonstrated stable, conservative predictions, while methods like LSTM and NAR showed higher variability, reflecting their sensitivity to recent trends.

The second visual further reinforced our time-series analysis by applying forecasting models, including Drift, Holt's Linear Trend, ARIMA, and LSTM, to broader financial market indices such as the Dow Jones Industrial Average (DJIA) and the S&P 500 (GSPC). Here, LSTM emerged as the best-performing model, achieving the lowest Mean Absolute Percentage Error (MAPE)—2.19% for DJIA and 2.93% for the S&P 500. This emphasized the model's capability to adaptively capture nuanced financial trends, crucial for correlating financial market behavior with real estate prices.

Lastly, the forecast table summarizes our immediate-term predictions for key housing market indicators. For the following week, the Collier County Price Index is forecasted at $404,960, the Lee County Price Index at $343,490, and the 30-year Mortgage Interest Rate at 6.76%. These figures provide actionable near-term insights, enabling stakeholders and prospective homebuyers to anticipate short-term market movements accurately and strategize their investment or purchasing decisions accordingly

Data	Forecast in a Week
Collier County Price Index	$404,960
Lee County Price Index	$343,490
Mortgage Interest Rate	6.76%

6. Shiny App Overview

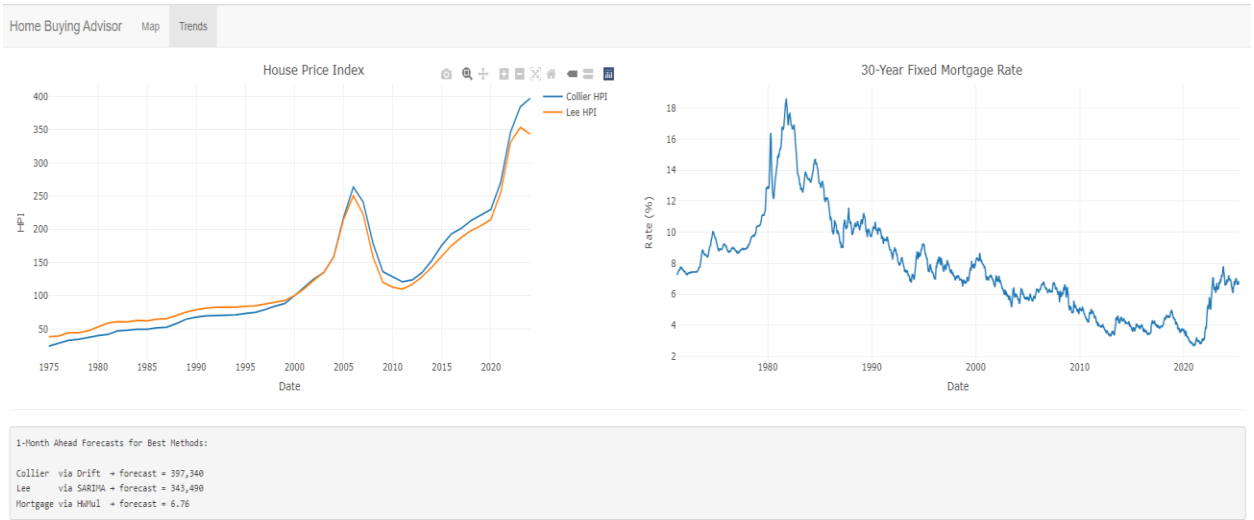


Figure 8 - Shiny App Trend Page

We developed an interactive dashboard titled "Home Buying Advisor" using R Shiny to translate our comprehensive data analysis into accessible insights for prospective homebuyers. The app features two primary tabs designed for user-friendly interaction. The "Trends" tab visualizes historical trends of local housing markets through interactive plots for the Collier and Lee County House Price Indices (HPI) and

30-year fixed mortgage rates, along with clearly displayed short-term forecasts, empowering users to interpret historical data and anticipate immediate market movements effectively.

The "Map" tab offers an interactive geospatial interface, enabling users to filter property listings by county, price range, bedroom and bathroom count, salary, down payment, and mortgage rates. The map visually clusters listings by their respective price per square foot, allowing users to easily identify affordable or premium regions. Integrated affordability calculations further enhance the tool by instantly providing tailored insights on the maximum affordable home price based on user-provided financial inputs. Collectively, these features transform complex analytical outputs into practical, actionable guidance for making informed real estate decisions.

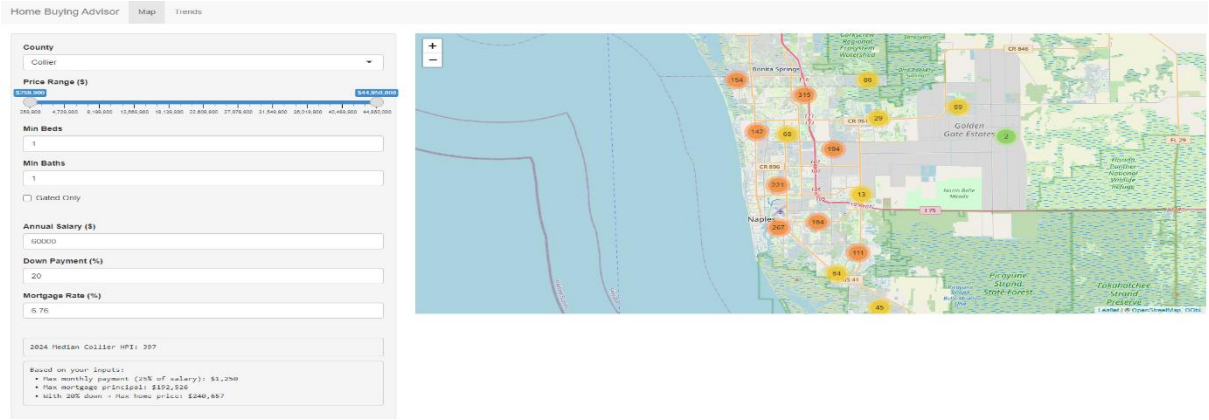


Figure 9 - Shiny App Map & Input Page

7. Evaluation of Methodology

7.1 Strengths

From this project, we can confidently display an end-to-end case study regarding the housing market to potential clients. The methodology employed in our analysis demonstrates several key strengths. First, our pipeline ensures comprehensive end-to-end reproducibility—from initial data collection via robust web scraping techniques to final deployment through the interactive Shiny dashboard. Such reproducibility is critical, not only for validation and reliability but also for facilitating future updates and iterations of our models. Furthermore, integrating climate risk considerations like historical hurricane occurrences and flood risks significantly enhances our understanding of environmental impacts on property valuation. This integration ensures our models reflect real-world vulnerabilities, making the outcomes highly relevant for stakeholders. Finally, our modular codebase provides distinct advantages by enabling parallel development, seamless integration, and straightforward maintenance. Distinct separation of concerns (data cleaning, geocoding, statistical modeling, and visualization) allows continuous refinement without compromising overall system stability.

7.2 Limitations & Biases

Despite these strengths, our analysis is not without limitations. One notable issue is potential selection bias resulting from scraping listings exclusively from RealtyOfNaples.com, which may exclude off-market transactions or listings only available through other platforms. Consequently, our dataset might not fully represent the broader housing market. Additionally we made a lot of assumptions when making our time series models stationary over time. In reality, markets experience structural breaks and regime shifts, potentially reducing the accuracy of long-term forecasts. Moreover, reliance on FEMA flood zone shapefiles introduces the risk of outdated or incomplete spatial data, potentially overlooking recent geographical changes or updates affecting flood risk designations.

7.3 Mitigation Strategies

We have proactively developed strategies to mitigate these limitations. Regular maintenance and updates to our scraping pipeline ensure continued access to accurate, current data while broadening our data sources to minimize selection bias. Furthermore, exploring regime-switching models, such as Markov-switching ARIMA, could better accommodate structural breaks and improve forecasting accuracy during volatile periods. Implementing annual updates from FEMA and other authoritative sources will enhance spatial data accuracy, ensuring environmental and risk assessments remain relevant. Collectively, these mitigation approaches are designed to maintain methodological rigor and enhance long-term robustness.

8. Further Implications & Extensions

Several promising directions could extend and enrich our analysis further. A critical area for expansion is modeling insurance costs based on flood zones, significantly enhancing our affordability assessments by factoring in hidden long-term costs associated with homeownership. Including insurance premium predictions would provide families with comprehensive insights into total cost of ownership beyond the initial purchase price. This is crucial as Florida is known as the lightning state and has a lot of fierce hurricanes. If hurricanes happen this causes damage and could drive prices down but also could drive prices up because of the flood insurance.

Additionally, incorporating walkability indices and local amenities such as proximity to schools, parks, transit networks, and essential services could substantially improve our predictive models by capturing quality-of-life factors that significantly influence home valuation. Detailed policy scenario analyses, particularly under varied climate change projections and shifts in flood risk profiles, could further inform buyers and policymakers about potential future housing market impacts and resilience strategies.

We also propose directly engaging families currently searching for homes or who have recently purchased properties in our target markets to gather qualitative data. Such consultations would reveal nuanced buyer preferences, financing hurdles, and experiential insights often absent from purely quantitative datasets. Alongside this, a deeper dive into credit scores and mortgage rates would refine our affordability models, enabling more precise recommendations tailored to varying financial profiles.

Examining actual property valuations, including monitoring online real estate fractional marketplaces, could validate and refine our methodologies based on real-time market responses. This approach would ensure our models remain responsive and reflective of evolving market dynamics.

Further data collection initiatives, particularly targeting comprehensive historical tax assessment data and detailed property price histories, could enable the deployment of sophisticated forecasting methods such as Gradient Boosting and Seasonal ARIMA (SARIMA). Expanding the temporal granularity to examine monthly median price fluctuations, differences by day of the week, impacts of holidays, and correlations with interest rate changes would significantly enhance forecasting accuracy and buyer advisory capabilities

9. Conclusion

Our analysis has provided a detailed, data-driven exploration of the housing market dynamics in Southwest Florida, leveraging robust statistical modeling and advanced forecasting techniques. By integrating diverse datasets, including economic indicators, climate risk profiles, and comprehensive property characteristics, we created actionable insights for prospective homebuyers and stakeholders alike. The modular, reproducible pipeline developed through this project not only ensures methodological rigor but also establishes a solid foundation for future expansions and refinements.

Despite the inherent challenges and biases identified, our proactive mitigation strategies and forward-looking extensions position our analysis to evolve continually. With future enhancements incorporating additional qualitative insights, policy scenarios, and granular forecasting methods, our Home Buying The shiny Advisor application can effectively guide homebuyers through complex market conditions, empowering informed and resilient decision-making.

References

"Naples Florida Real Estate : 6567 Naples Homes & Condos for Sale." *Realty Of Naples*, www.realtyofnaples.com/. Accessed 16 May 2025.

"All-Transactions House Price Index for Lee County, FL." *FRED*, 25 Mar. 2025, fred.stlouisfed.org/series/ATNHPIUS12071A.

"30-Year Fixed Rate Mortgage Average in the United States." *FRED*, 8 May 2025, fred.stlouisfed.org/series/MORTGAGE30US.

"All-Transactions House Price Index for Collier County, FL." *FRED*, 25 Mar. 2025, fred.stlouisfed.org/series/ATNHPIUS12021A.

"Yahoo Finance - Stock Market Live, Quotes, Business & Finance News." *Yahoo! Finance*, Yahoo!, finance.yahoo.com/. Accessed 16 May 2025.