# Real Estate Property Purchasing vs. Renting Valuation

**Team084:** Hong He, Shouyuan Lin, Yuchen Qiu, Zhen Wang, Weili Weng

## 1. Introduction

In today's rapidly changing real estate market, sophisticated analytical tools are increasingly in demand to assist potential buyers, renters, and real estate professionals. Driven by the influx of private equity firms purchasing vast numbers of homes, the market has seen a sharp decline in available properties. To address this gap, we are developing a real estate property valuation system and dashboard. The system aims to provide timely insights into housing trends and opportunities by analyzing the costs of renting and purchasing. Moreover, it will provide guidance on which metro area or zip code offers the most promising return on investment and the decision between purchasing and renting.

Our valuation system is invaluable for homebuyers and renters navigating a competitive market to find housing options that best align with their needs and budgets. Equipped with the analytical capabilities of our system, real estate agents can also offer more compelling guidance to their clients, substantiating their advice with data-driven insights. In achieving our goal, we anticipate a marked improvement in real estate decision-making processes. We will gauge the success and impact of our platform through user satisfaction surveys, monitoring how it facilitates informed choices. A growing user base, drawn by the platform's efficacy and value, will further underscore our success in this endeavor.

## 2. Problem Definition

Our dashboard provides a "rent or buy" recommendation. While easy to use, they overlook essential factors like property type, number of bedrooms, number of bathrooms, location, zip code, interest rate, etc. We chose the most appropriate machine learning model that makes strides in real estate predictions and finally delivers timely insights into housing trends and opportunities.

## 3. Literature Survey

Many studies have utilized machine learning for housing price predictions, considering variables such as property type, location, and ROI. **Mani (2018)**[5] discussed integrating data science with geospatial analysis for housing recommendations. While location plays a key role, concerns arise over the data's timeliness and completeness. **Mora-Garcia (2022)**[12] highlighted risks like data leakage and overfitting as the shift towards machine learning grows. Ullah emphasized the importance of system dynamics models given real estate's non-linear decision-making process. **Zhan et al. (2023)**[1] developed hybrid machine-learning models for housing price forecasting. Their models aim to overcome data limitations and optimize hyperparameters, but real-world applications could face computational challenges. **Bertasso, Pillay, and Boshoff (2015)**[2] delved into homeownership perceptions, demographics, and financial factors affecting housing decisions. Their insights, though valuable, might be limited by geography and outdated data. **Halket and Pignatti Moranodi Custoza (2015)**[3] explored rental scarcity's role in homeownership. Using a novel house-matching algorithm, they identified rental scarcity trends but faced potential data source biases. The **U.S. Congress House Committee (2022)**[4] reported on private equity firms transforming the housing market, further emphasizing the growth in house rentals. **Ullah (2020)**[13] shifted the focus from machine learning, advocating for system dynamics models. These models capture the holistic real estate view, which is especially beneficial in unpredictable markets. To sum up, modern real estate tools, from machine learning to

system dynamics, offer essential insights. Current studies highlight the promise of these tools combined with socio-economic and geospatial analysis. For optimal results, it's essential to understand each method's limitations, update data sources, and regularly validate models against the dynamic real estate market.

## 4. Proposed Method

In the competitive realm of real estate analytics, our approach differentiates itself through a blend of innovative techniques and data-driven robustness:

a. We've curated a unique dataset by combining information from Realtor.com with comprehensive statistics from the US Census and ACS, ensuring accuracy in our housing price predictions.
b. Our innovative method categorizes markets by bedroom counts and zip codes, including only submarkets with at least 10 listings, ensuring a robust analytical base for detailed renting vs. buying cost comparisons.
c. We utilize the Featurewiz package, which employs the SULOV algorithm and Recursive XGBoost for feature selection. This process efficiently narrows down features to identify the most relevant ones to train our models.
d. The technical backbone of our strategy hinges on hybrid models that seamlessly fuse grid search for precise hyperparameter tuning with renowned machine learning algorithms such as KNN, Random Forest, Linear Regressor, BayesianRidge, XGBoost, and Artificial Neural Network (ANN).
e. Our final product will be an intuitive tableau dashboard that demystifies complex data into lucid, understandable visuals, making user navigation seamless. However, risks are associated with our project, such as biases from public data sources, the lack of real-time updates, and potential legal challenges. We will also conduct Beta testing after our tableau dashboard development.

## 5. Experiments and Evaluation

### *5.1 Data Collection*

We used a free web scraping Python library called HomeHarvest to get 437 thousand recent home sales records from the past 270 days and 107 thousands ongoing rental and for sales listings posted within 30 days. We scraped house listings from major cities in the most significant 20 metro areas and saved them as pandas data frames, and then combined them into a 3 separate property dataset: “sold”, ”for sale”, and “for rent”. We focused on residential property types like single-family homes, condos, townhouses, and apartments etc. We combined this information with over two thousands zip code-level data obtained from The American Community Survey (ACS) API, which collects other additional 25 features such as the US population's diverse demographic and housing market details.

### *5.2 Data Cleaning and Exploration*

The scraped dataset demanded extensive cleaning, including handling missing values, removing unreasonable data, categorizing features, and normalizing numerical data. Additionally, irrelevant columns were removed. To understand the data better, we utilized visualization techniques like histogram charts (**Figure 1**) and correlation heatmaps (**Figure 2**. We performed a log transformation on the sold_price variable and removed outliers based on each features’ distribution analysis. The correlation heatmap revealed strong associations of sold_price with various factors like zip_code, related number of bedroom units/percentage, age over 35 and less than 85 percent, household median income, price_per_sqft, sqft, and beds.

Employing the Featurewiz package, we conducted automated feature selection by setting the corr_limit to 0.5, aiming to reduce multicollinearity. The output presents the 10 selected features for the dataset: ['price_per_sqft,' 'sqft,' 'style_encoded,' 'year_built_encoded,'

'ageover35andlessthan85,' 'zipcode_encoded,' 'VacanthousingunitsPercent,' 'Renteroccupied,' 'Vacanthousingunits,' and 'related number of bedroom units'].

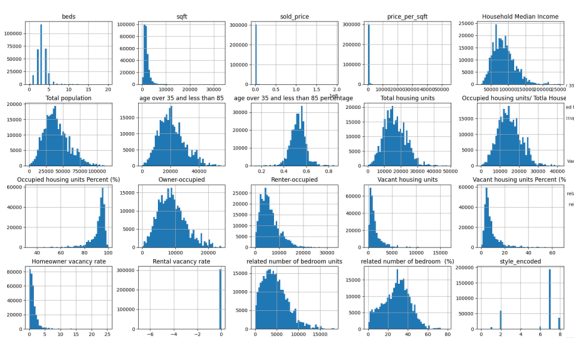

*Figure 1: Features Distribution*

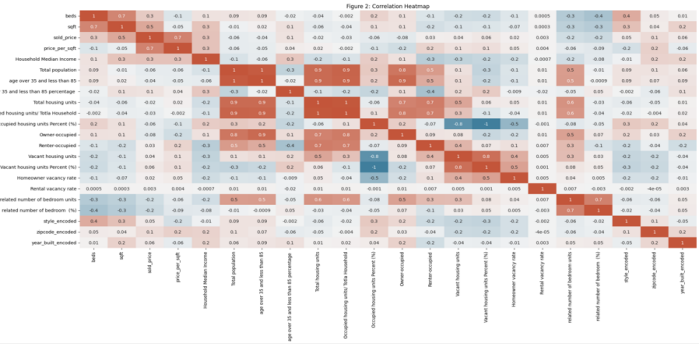

*Figure 2: Correlation Heatmap*

### *5.3 Model Building*

We firstly split the "sold" dataset into 70% for training and 30% for testing models. We developed a suite of esteemed machine learning models (Random Forest, ANN, Linear Regression, Bayesian Ridge, KNN, and XGBoost ) to rigorously explore the hyperparameter space and evaluate comparative model performance effectively.

- ***Linear Regression:*** Linear regression predicts prices by analyzing relationships between property features (like size, and location) and prices. It has the benefit of simplicity, interpretability, and fast computation. However, it assumes linear relationships and is sensitive to outliers. After examining the model's output, we concluded that the real estate dataset is inappropriate for the linear regression model due to nonlinearity.
- ***XGBoost:*** XGBoost, which stands for Extreme Gradient Boosting, is an advanced implementation of gradient boosting that is widely used in the field of machine learning. It is known for its performance and speed and is often used in data science competitions and in industry applications. We applied the XGBoost model to our housing dataset. Despite obtaining great numbers in many metrics, such as EVS and RMSE, it performed poorly on the new housing dataset due to overfitting.
- ***Random Forest:*** In the context of real estate price prediction, Random Forest is particularly well-suited due to its ability to handle complex datasets with many variables and non-linear relationships. It is robust to outliers and can model the conditional relationships required for real estate valuation, such as the interactions between location, size, and property condition. The model we built utilizes hyperparameter tuning, which systematically searches through various parameter combinations to find the most effective model settings, thereby improving the model's predictive power.
- ***Bayesian Ridge and K-Nearest Neighbor (KNN):*** Scikit-learn is a popular Python library that implements Bayesian Ridge Regression and K-Nearest Neighbors models. We used Scikit-learn 'BayesianRidge' and 'KNeighborsRegressor' to build these two models, fit them to the training data, and predict the final sale price with features like "zip_code," "sqft," and "list_price." When choosing the k value, a too-large K value may over-smooth and lose essential patterns in the data, so we finally choose k = 12 after comparison.
- ***Artificial Neural Network (ANN):*** When dealing with ANN, we chose a deep-learning Python library called TensorFlow. Prior to model training, we standardized the features using StandardScaler() provided by Scikit-learn. It ensures that all selected features contribute equally to the model, preventing any particular feature from dominating the

learning process. Finally, we built the neural network model, compiled it, and then trained the model.

### *5.4 Model Evaluation and Statistical Tests*

After training the model, we using the testing dataset to calculate the following 6 assessment metrics:

| Models | RMSE | MAE | MAD | MAPE | RMSLE | EVS |
|---|---|---|---|---|---|---|
| Linear Regression | 181,329,225 | 731,546 | 55,607 | 0.013 | 0.017 | 0.832 |
| XGBoost | 9,766 | 5,910 | 3,903 | 1.127 | 0.017 | 0.998 |
| Random Forest | 121,147 | 3061 | 609 | 0.222 | 0.112 | 0.972 |
| BayesianRidge | 107,404 | 35,073 | 20,775 | 0.060 | 0.089 | 0.981 |
| KNN | 184,407 | 30,265 | 12,025 | 0.040 | 0.070 | 0.943 |
| ANN | 96,195 | 28,905 | 21,391 | 0.039 | 0.066 | 0.985 |

Upon reviewing the evaluation metrics, it is apparent that the Random Forest, Bayesian Ridge, XGBoost, and Artificial Neural Network models all demonstrate an Explained Variance Score (EVS) exceeding 0.95, indicating strong predictive capabilities. Focusing on the finer details, the RF model showcases superior performance in terms of Mean Absolute Error (MAE), Mean Absolute Deviation (MAD), Mean Absolute Percentage Error (MAPE), and Root Mean Squared Logarithmic Error (RMSLE). These metrics are critical for our use case because they measure prediction accuracy and the consistency of the model, which are essential factors in real estate price prediction. Given its balanced performance across these key indicators, the Random Forest model was selected for deployment. It has been utilized to project selling prices for properties within our 'for_sale' real estate dataset, ensuring our price predictions are reliable and accurate.

### *5.5 Data Prep for Visualization*

We used our selected machine learning model to predict a purchase price for each currently listed for sale property on realtor.com. Additionally, we incorporated significant socio-economic factors such as interest rates into our analysis. Then, we assumed a 20% mortgage down payment and calculated the monthly mortgage payment plus real estate tax as the monthly buying cost. On the other hand, we scraped the current for-rent property from realtor.com and used the listing price as the monthly renting cost. Based on the comparison between buying and renting monthly costs, we then offered recommendations on whether to buy or rent a property, considering which option would be more cost-effective.

### *5.6 Visualization and Observations*

We designed an intuitive and interactive dashboard featuring both high-level and detailed-level visualizations. To enhance decision-making, our design includes color coding, filters, navigation, maps, heatmap, bar charts, table, box-and-wikery plots and URL action link. These visual assists users in identifying suitable investment or purchasing opportunities based on their customized preferences. Please find the demo of this dashboard at the following URL: https://www.youtube.com/watch?v=tbYkrp02Btc

#### *5.6.1 High Level*

At a high level, the dashboard enables users to discern which states and cities are more favorable for buying or renting and illustrates how shifts in mortgage rates, from 7.5% to 5% for instance-affect market trends, as shown in ***Figure 3***. Metro Areas in MI,IN,NJ still have opportunities to buy at 7.5% mortgage rate while more metro areas will have buying opportunities if mortgage rate goes down to 5%. Another observation of the headmap shows over 50% of listings originate from five major metro areas. Among these, Miami-Fort Lauderdale

and Houston-The Woodlands-Sugar Land stand out with a notably higher number of listings. Additional observation is that the 3 and 4 beds single family counts for also 40% of listings. In adaptation to the feature of the dashboard, the map offers an intuitive understanding of the geographical layout directly on the dashboard. Interactive features on the map allow users to apply filters to refine their search for the listing information of the right heatmap. Users can also apply filters for specific cities, zip codes, house types and number of bedrooms and bathrooms. The data dynamically adjusts to these selections, and users can then navigate to the next page by clicking the button *"Next Page"* for details.

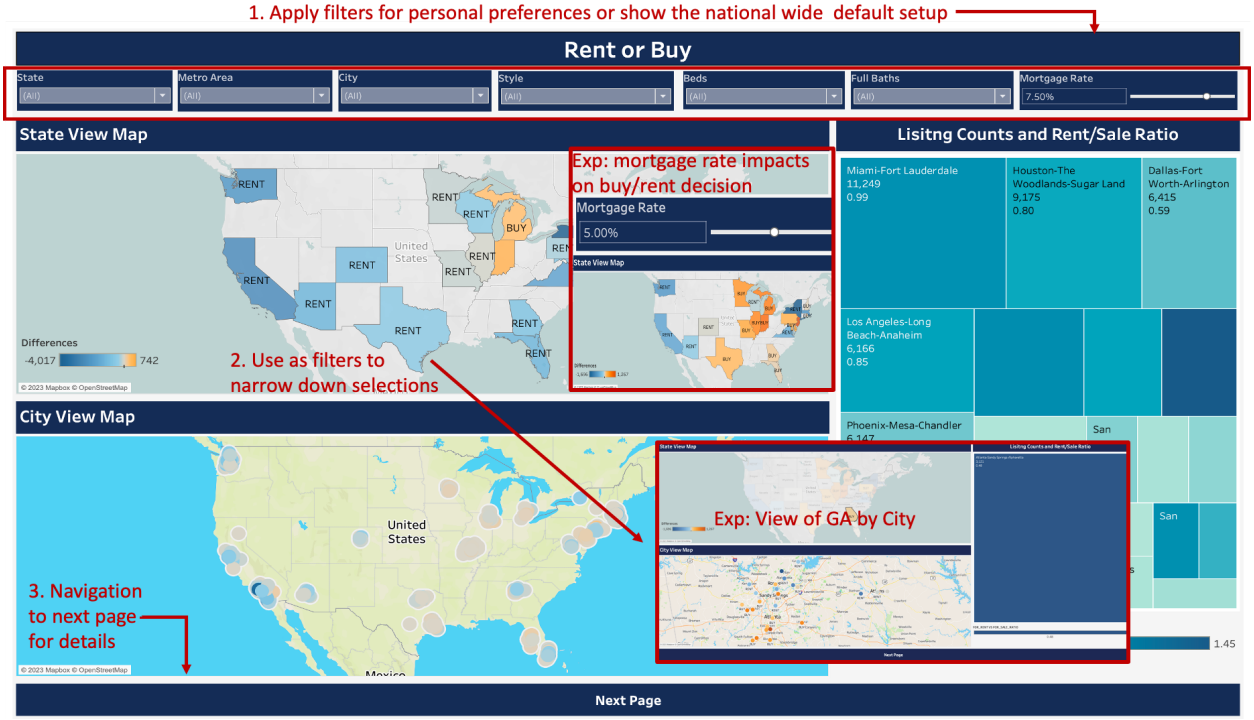


*Figure 3: High-Level Decision Guide for Buying vs. Renting Based on Mortgage Rates*

### *5.6.2 Detailed Level*

As **figure 4** shows, at the top of the dashboard, users can further personalize their view using a variety of filters. For example, a user interested in properties in Atlanta – such as single-family homes or townhouses with 3 or 4 bedrooms – can set these preferences along with a default mortgage rate of 7.5%. Upon adjusting the filters, the dashboard updates to display relevant data. On the left, users will find the current number of available rental properties along with their pricing distribution. Mirroring this, the right side shows analogous data for properties available for sale, filtered and presented according to the same set of criteria. This setup allows users to quickly grasp the total count and pricing trends of houses based on the criteria. Additionally, the dashboard features a central table aggregating and comparing the monthly renting costs versus buying. This table not only presents the raw data and offers a recommendation based on the cost differences between renting and buying, aiding users in making informed decisions. With this comparison, it becomes evident that even the rent is the trend but buying and renting decisions are based on available properties at Zip code level. What 's more, in the middle bottom section of the dashboard, users are presented with a visual map, providing a comprehensive view of selected areas. As users hover over different locations on the map, tooltips appear, offering detailed information about each listing, including a URL link. By selecting the listing, the link directs users to a web browser page for entire property listings, including descriptions, images, and additional relevant information on sites on Realtor.com.

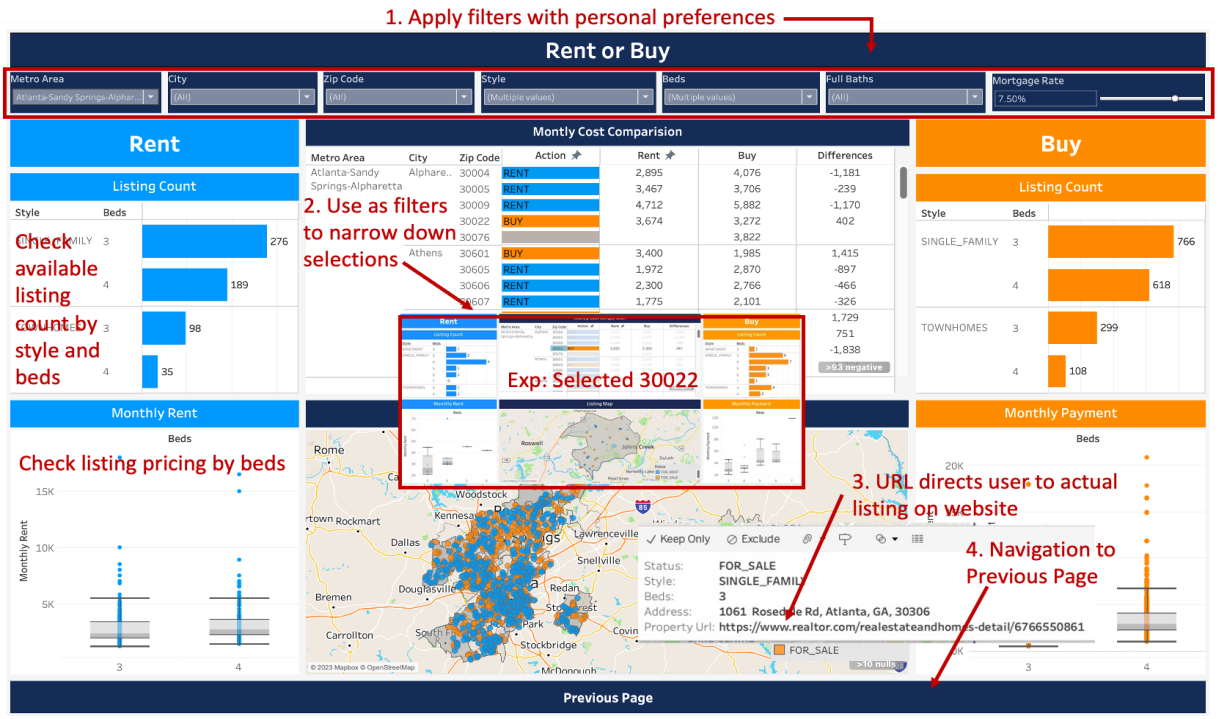


*Figure 4. Detailed-Level Decision Guide for Buying vs. Renting with User-Defined Filters*

## 6. Conclusions and Discussion

Our initial findings indicate a potential paradigm shift in real estate decision-making. We envision our model as a pioneering tool that empowers users to navigate the real estate market with greater clarity and confidence, supported by well-crafted visual aids. Our system offers critical insights into housing trends and opportunities, providing guidance on investment returns and the decision to buy or rent in specific metropolitan areas or zip codes. Key highlights of our approach:

a. Advanced analytics blend of datasets, integrating real estate listings with demographic and macroeconomic building information.
b. Hyperparameter tuning within machine learning algorithms to enhance robustness, addressing market complexities and mitigating data leakage.
c. A user-friendly Tableau dashboard that translates complex data into clear, easily understandable visuals, thereby simplifying user interaction.

However, our approach is not without limitations. Challenges include biases inherent in public data sources and the absence of real-time updates. Looking ahead, we anticipate significant improvements from incorporating a more comprehensive dataset and ongoing model validation, reflecting the dynamic nature of real estate markets. Enhancements will focus on expanding the dataset, integrating real-time data to sharpen accuracy, and iterative beta testing of the dashboard to continuously refine both the user experience and the utility of our tool.

The anticipated success of this project owes much to the collaborative efforts of our team. Throughout the development process, all team members have contributed a similar amount of effort. Active communication via chat apps and regular Zoom meetings enabled us to effectively navigate the challenges posed by a teammate's withdrawal from the class. We adapted our project plan to the actual workload of each task and completed the project within the original timeline.

**Reference**


1. Zhan, C., Liu, Y., Wu, Z., Zhao, M., & Chowc, W. S. (2023). A Hybrid Machine Learning Framework for Forecasting House Price. *Expert Systems with Applications*, 233, 120981. ISSN 0957-4174. https://www.researchgate.net/publication/372346729_A_hybrid_machine_learning_framework_for_forecasting_house_price
2. Bertasso, C., Pillay, D., & Boshoff, D.G.B. (2015). The Rent vs Buy Decision of Residential Property. *The 3rd Virtual Multidisciplinary Conference*. DOI: 10.18638/quaesti.2015.3.1.244. https://www.researchgate.net/publication/314611349_The_Rent_vs_Buy_Decision_of_Residential_Property
3. Halket, J. & Pignatti Moranodi Custoza, M. (2015). Homeownership and the Scarcity of Rentals. *Journal of Monetary Economics*, 76,107-123. https://econpapers.repec.org/artCoefficient of Determinationicle/eeemoneco/v_3a76_3ay_3a2015_3ai_3ac_3ap_3a107-123.htm
4. United States. Congress. House. Committee on Financial Services. Subcommittee on Oversight and Investigations. (2022). Where have all the houses gone? Private equity, single-family rentals, and America's neighborhoods: Virtual hearing before the Subcommittee on Oversight and Investigations of the Committee on Financial Services, U.S. House of Representatives, One Hundred Seventeenth Congress, second session, June 28, 2022. *U.S. Government Publishing Office*. GOVDOC: Y 4.F 49/20:117-90. CHRG-117hhrg48334.pdf (govinfo.gov)
5. Mani, A. (2018). House Hunting — the Data Scientist Way. *GeoAI*. https://medium.com/geoai/house-hunting-the-data-scientist-way-b32d93f5a42f
6. Choy, L. H. T., & Ho, W. K. O. (2023). The Use of Machine Learning in Real Estate Research. Journal of Land Innovations – Data and Machine Learning, 12(4), 740. https://doi.org/10.3390/land12040740
7. Baldominos, A., Blanco, I., Moreno, A. J., Iturrate, R., Bernardez, O., & Afonso, C. (2018). Identifying Real Estate Opportunities Using Machine Learning. *Journal of Applied Sciences*. https://arxiv.org/pdf/1809.04933.pdf
8. Probasco, J. (2023). Renting vs. Buying a Home: Which is better for you? *Fortune*. https://fortune.com/recommends/mortgages/renting-vs-buying-a-home/
9. Domes, S. (2021). Using Machine Learning to Predict Your Rent. *Medium*. https://scottdomes.medium.com/using-machine-learning-to-predict-your-rent-91c783fdf6d6
10. Cox, A. & Followill, R (2018). Rent or Buy: A 30-Year Perspective. *Financial Planning Association May 2018*. https://www.financialplanningassociation.org/article/journal/MAY18-rent-or-buy-30-year-perspective
11. Gollapudi, S. & Panigrahi, D. (2019). *Online Algorithms for Rent-or-Buy*. *In Proceedings of the 36th International Conference on Machine Learning*, 97:2424–2432. https://proceedings.mlr.press/v97/gollapudi19a/gollapudi19a.pdf
12. Mora-Garcia, R.T., Cespedes-Lopez, M.F. & Perez-Sanchez, V.R., (2022). Housing Price Prediction Using Machine Learning Algorithms in COVID-19 Times. *Land*, 11(11), p.2100. https://www.mdpi.com/2073-445X/11/11/2100
13. Ullah, F. and Sepasgozar, S.M., (2020). Key factors influencing purchase or rent decisions in smart real estate investments: A system dynamics approach using online forum thread data. *Sustainability*, 12(11), p.4382. https://www.mdpi.com/2071-1050/12/11/4382

14. Beracha, E. and Johnson, K.H., (2012). Lessons from over 30 years of buy versus rent decisions: Is the American dream always wise?. *Real Estate Economics*, 40(2), pp.217-247. https://ideas.repec.org/a/bla/reesec/v40y2012i2p217-247.html
15. Rampini, L., & Re Cecconi, F. (2022). Artificial Intelligence Algorithms to Predict Italian Real Estate Market Prices. *Journal of Property Investment & Finance*. https://www.emerald.com/insight/content/doi/10.1108/JPIF-08-2021-0073/full/html
16. Lorenz, F., Willwersch, J., Cajias, M., & Fuerst, F. (2022). Interpretable Machine Learning for Real Estate Market Analysis. *Real Estate Economics, 2022 The Authors*. Real Estate Economics published by Wiley Periodicals LLC on behalf of American Real Estate and Urban Economics Association. DOI: 10.1111/1540-6229.12397. https://onlinelibrary.wiley.com/doi/pdf/10.1111/1540-6229.12397
17. Conway, J. E. (2018). Artificial Intelligence and Machine Learning: Current Applications in Real Estate. *Massachusetts Institute of Technology*. https://dspace.mit.edu/handle/1721.1/120609
18. Ajala, E. (2018). The Economics Of Web Scraping Report. https://www.imperva.com/blog/the-economics-of-web-scraping-report/
19. Truong, Q., Nguyen, M., Dang, H., & Mei, B. (2020). Housing Price Prediction via Improved Machine Learning Techniques. *Procedia Computer Science*, 174, 433-442. https://goo.by/YZYvMm
20. Chan S. P. (2001) "Case Study 2: Housing Decisions: Renting versus Owning". *Auburn University*. https://www.eng.auburn.edu/~park/documents/case_study/case2.pdf