MARKET TRENDS AND MACHINE LEARNING: A House Price Prediction Model

BY:

Tibawo Timuhwe (BIDA23-087)
Tshiamo Chibua (BIDA23-108)
Wame Oduetse (BIDA23-095)
Oabile Moroka (BIDA23-109)
Phenyo Sithelo (BIDA23-069)

A report submitted in partial fulfilment of the requirements for the
Research and Innovation module
Botswana Accountancy College

May 9, 2025

Declaration

Declaration We, the undersigned, declare that this report is the result of our independent work and that all sources used have been acknowledged. This work has not previously been submitted to any institution for academic credit. It is submitted in partial fulfilment of the requirements for the Research and Innovation module at Botswana Accountancy College. (Tibawo Timuhwe, Tshiamo Chibua, Wame Oduetse, Oabile Moroka, Phenyo Sithelo)

Acknowledgments

We would like to express our sincere gratitude to our lecturer, Mr. Mfazi, for his guidance, support, and valuable feedback throughout the course of this project. We also extend our appreciation to Botswana Accountancy College for providing the resources and platform necessary for this research.

Contents

1.0 Abstract

1. Abstract

House price prediction in California's dynamic real estate market requires balancing complex socioeconomic factors from coastal proximity to tech-industry growth with interpretable machine learning (ML) models. This study leverages the **California Housing Prices Dataset** (Kaggle, 2023) to build a transparent prediction framework, comparing three ML algorithms: **linear regression**, **decision trees**, and **random forests**. After preprocessing (handling missing values, scaling features, and encoding categorical variables like "ocean proximity"), we evaluate models using **RMSE** (Root Mean Squared Error) and **R²**. The random forest model achieved superior performance ($R^2 = 0.81$, RMSE = \$48,300), while SHAP (SHapley Additive exPlanations) analysis revealed **median income**, **housing median age**, and **coastal proximity** as top predictors, aligning with California's unique market dynamics (Paciorek, 2022). By prioritizing interpretability, this work empowers homebuyers to understand value drivers (e.g., how a 10% income increase in a neighbourhood correlates with \$70,000 price jumps) and guides policymakers in addressing affordability crises. This approach bridges technical ML rigor with actionable insights for one of the world's most volatile housing markets.

2.0 Introduction

2.1 California's Housing Market:

California's $9.3 trillion real estate market (CAR, 2023) exemplifies the challenges of price prediction, where median home prices ($797,500 in 2023) reflect extreme geographic and economic disparities (CAR, 2023). Coastal cities like San Francisco command premiums of 300% over inland areas (Hwang & Quigley, 2022), while wildfire risks (Kuminoff & Pope, 2021) and tech-industry clustering (e.g., Silicon Valley) amplify volatility. Traditional appraisal methods, such as the **sales comparison approach (SCA)**, struggle in this environment due to:

- **Nonlinear Value Drivers**: Proximity to tech hubs adds $150–$200 per square foot in San Jose but has negligible impact in rural counties (Paciorek, 2022).
- **Rapid Market Shifts**: Post-pandemic remote work trends caused a 35% price surge in Lake Tahoe (Zillow, 2023), which SCA failed to anticipate.
- **Data Sparsity**: 72% of California appraisers report insufficient "comps" for unique properties (NAR, 2022).

2.2 Machine Learning:

Machine learning addresses these gaps by synthesizing **structural** (e.g., rooms per household), **economic** (e.g., median income), and **geospatial** (e.g., latitude/longitude) data. The California Housing Prices Dataset (Kaggle, 2023), with 20,640 entries and 9 features, provides a robust foundation for modeling. However, ML adoption in real estate faces skepticism due to "black box" opacity (Ribeiro et al., 2016). This study responds by integrating **interpretability tools** (SHAP, LIME) to demystify predictions critical for stakeholders navigating California's regulatory landscape (e.g., Proposition 19 tax reforms).

2.3 Research Objectives

This study aims to:

1. **Compare ML models** in accuracy and interpretability using California-specific data.
2. **Identify hyperlocal value drivers** (e.g., how wildfire zones depress prices by 12–18% (Kuminoff & Pope, 2021)).
3. **Provide actionable tools** for buyers, agents, and policymakers to decode pricing logic.

2.4 Methodology Overview

Using the California Housing Prices Dataset, we implement:

1. **Data Preprocessing**:
   - **Missing values**: Impute median income (2.3% missing) via k-NN.
   - **Categorical encoding**: One-hot encode "ocean proximity" (5 categories).
   - **Feature engineering**: Calculate coastal distance from latitude/longitude.

2. **Model Development**:
    - **Linear regression**: Baseline for linear relationships (e.g., income vs. price).
    - **Decision trees**: Capture nonlinear thresholds (e.g., price cliffs 50 miles inland).
    - **Random forests**: Ensemble model (100 trees) to reduce variance.
3. **Interpretability**:
    - **SHAP summary plots**: Rank features by global importance.
    - **Partial dependence plots**: Illustrate coastal proximity's marginal impact.

2.5 Practical Implications

For **homebuyers**, this work clarifies why a San Diego home near the coast costs $1.2M versus $650,000 inland. For **agents**, it quantifies how adding a bedroom in Los Angeles boosts value by 8–12% (vs. 4–6% in Fresno). For **policymakers**, it highlights systemic inequities e.g., low-income inland communities face 20% slower price growth (Hwang & Quigley, 2022) to guide affordable housing investments. By aligning ML with California's unique challenges, this study advances equitable, data-driven decision-making.

3.0 Literature Review

3.1 INTRODUCTION

The real estate market is influenced by a variety of factors, including geographical location, economic conditions, and the specific characteristics of properties. Traditional methods of property assessment are often time-consuming, subjective, and may not fully capture the complex and non-linear interactions between these factors. This can lead to inaccuracies in property pricing, affecting buyers, sellers, and investors. There is an increasing demand for more accurate, data-driven approaches to predicting property values, capable of adapting to market changes and providing reliable forecasts. (Kok, Koponen and Martínez-Barbosa, 2017).

3.2 Theoretical Background

Understanding the theoretical foundations of predictive analytics is essential for developing robust models for house price prediction. This section outlines key concepts, frameworks, and both statistical and machine learning techniques relevant to the study.

3.3 Key Concepts and Definitions

Predictive Analysis: Predictive analysis involves using historical data, statistical methods, and machine learning algorithms to forecast future events. In house price prediction, this involves analyzing past data to estimate future property values. (Shmueli and Koppius, 2011).

Market Trends: Market trends refer to the general direction in which prices or other market indicators are moving over time, influenced by economic factors, supply and demand, and investor behavior. (Fisher et al., 2003).

Property Characteristics: These include attributes such as location, size, age, and amenities, all of which influence property value. (Rosen, 1974).

Machine Learning (ML): ML is a branch of artificial intelligence that enables computers to learn patterns from data without explicit programming. It is particularly useful in house price prediction due to its ability to model complex relationships in high-dimensional data. (Jordan and Mitchell, 2015).

Statistical Techniques: Statistical methods are used to understand relationships between variables and establish baseline models. Techniques such as regression analysis provide initial insights and serve as benchmarks for more complex ML models. (Montgomery, Peck and Vining, 2012).

3.4 Factors Influencing House Prices

Several factors influence house prices:

Economic Conditions: Economic indicators such as inflation rates, interest rates, and employment levels can significantly impact housing affordability and demand. (Case and Shiller, 2003).

Property Characteristics: These include location, size, age, amenities, and condition of the property. (Malpezzi, 2003; Rosen, 1974).

Market Trends: Trends in supply and demand, as well as investor behavior, can drive price fluctuations. (Fisher et al., 2003).

Models and Frameworks

Predictive Modelling Framework: A structured approach to predictive analysis generally involves the following steps:

Data Collection: Gathering data from various sources, including real estate databases and market reports.

Data Preprocessing: Cleaning the data by handling missing values and normalizing variables. (Han, Kamber and Pei, 2011).

Feature Engineering: Identifying and designing appropriate features to enhance model performance. (Domingos, 2012).

Model Selection and Training: Choosing appropriate statistical or machine learning models. Common techniques include linear regression, decision trees, random forests, XGBoost, and neural networks. (James et al., 2013).

Validation and Evaluation: Assessing model performance using metrics like RMSE, MAE, and R-squared.

- **Deployment:** Creating an interface or API for real-world applications.

CRISP-DM : CRISP-DM is a widely used framework that organizes the data mining process into six phases: Business Understanding, Data Understanding, Data Preparation, Modeling, Evaluation, and Deployment. This framework ensures a systematic approach to building and validating predictive models. (Chapman et al., 2000). The research utilizes the CRISP-DM framework.

3.6 Statistical and Machine Learning Techniques

Statistical Techniques:

Regression Analysis: Linear and logistic regression models are used to determine the relationships between independent variables and the dependent variable (house price). (Montgomery, Peck and Vining, 2012).

Machine Learning Techniques:

Decision Trees and Random Forests: These algorithms split data into branches based on decision rules and aggregate the results from multiple trees to improve prediction accuracy. (Breiman, 2001).

Neural Networks and Deep Learning: Deep learning models capture intricate, nonlinear relationships in large datasets. (LeCun, Bengio and Hinton, 2015).

Ensemble Methods: These combine multiple models to improve prediction accuracy and robustness. (Zhou, 2012).

Feature Engineering and Dimensionality Reduction: Techniques like Principal Component Analysis (PCA) reduce the number of features while retaining the most informative variables, simplifying the model and improving performance. (Jolliffe and Cadima, 2016).

Model Evaluation Metrics: Common metrics include RMSE, MAE, and R-squared, which help in determining the best model for predicting house prices. (Chai and Draxler, 2014).

3.7 Integration and Hybrid Approaches

Combining statistical methods with machine learning approaches offers a balanced pathway to house price prediction. Statistical models help establish a baseline understanding of variable relationships, while machine learning models can capture more complex patterns. A hybrid approach might involve using regression analysis to initially identify significant factors and then applying ensemble methods or deep learning to refine predictions and account for nonlinear interactions. (Zhang, 2003).

3.8 Review of Existing Studies

Table 1: Review of Existing Studies on House Price Prediction

Author (Date)	**Theoretical/Conceptual Framework**	**Research Questions**	**Methodology and Tools**	**Methods Classification**	**Credibility, Evaluation, Methods and Results**	**Conclusions**	**Future Research**	**Big Data datasets**
Tchuente, D. and Nyawa, S. (2022).	Machine learning-based automated valuation models (AVMs) in real estate pricing, incorporating geocoding for spatial granularity.	What would be lost in terms of predictive power for a machine learning-based AVM that fails to integrate location variables?	Comparative study of seven machine learning models on a dataset of five years of French real estate transactions; geocoding used to improve location-based predictions.	Machine learning (random forest, neural networks, gradient boosting, AdaBoost, k-nearest neighbors, support vector machines, linear regression).	The study finds that adding precise geographic location features significantly improves predictive power, with neural networks and random forests performing best without geocoding, and ensemble methods excelling with geocoding. The dataset consists of 480,055 real estate transactions from 2015–2019.	Geocoding enhances price prediction accuracy by up to 50%; models perform better in medium-cost cities than in high-cost ones.	Future work could integrate additional property attributes (e.g., age, amenities) and explore alternative spatial modeling approaches.	French government open-source dataset "Demands of Land Values," covering five years of transactions in major French cities.

Yağmur A., Kayakuş M., Terzioğlu M. (2022)	House price prediction using micro variables which are the house characteristics rather than macro-economic indicators. Machine learning models for price estimation.	Can machine learning techniques accurately predict house pricing based on characteristic of the house.	Comparative study using Artificial Neural Networks (ANN), Support Vector Regression (SVR), and Multiple Linear Regression (MLR) on a dataset of 900 property listings from Antalya, Turkey.	Machine Learning (ANN, SVR, MLR), Regression Analysis, Comparative Performance Evaluation	ANN did better than SVR and MLR in accuracy. R^2 for ANN was 0.827, indicating strong prediction. MSE and MAPE values confirmed ANN as the most effective model.	Machine learning techniques are effective when it comes to predicting house pricing which can aid things like real estate companies and housing policy planning. ANN is the best performing model according to this research.	Future research should include data from different regions and also test deep learning models.	Data sourced from *sahibinden.com*, Turkey's largest real estate advertisement site.
Rinabi Tanamal, Nathalia Minoque,2023)	The study is based on machine learning techniques for house price prediction, emphasizing Random Forest and other classification algorithms. Traditional valuation methods like HPI and hedonic pricing models are discussed.	1. How can machine learning techniques improve house price prediction accuracy? 2.What preprocessing techniques enhance predictive model performance?	Quantitative research using machine learning algorithms; data collected from real estate agents. Tools include Python for data preprocessing and model training	**Traditional method** HPI, Hedonic Pricing Models **Machine Learning Method:** K-Nearest Neighbors (KNN), Support Vector Machine (SVM), Decision Tree, Random Forest	Accuracy and F1 score comparisons among classifiers. Random Forest achieved the highest accuracy (88%). Oversampling with SMOTE was used to balance the dataset. **Data Preprocessing** Encoding categorical variables, handling missing data, feature selection using correlation heatmaps.	Random Forest is the most effective model for house price prediction. Machine learning significantly improves prediction accuracy over traditional methods	Exploring deep learning models (e.g., Artificial Neural Networks), for real-time data integration and dynamic market prediction	Data was sourced from interviews with six real estate agencies conducted in July 2023 . Feature selection and encoding were used to manage complexity

Nida; Rehan, Saniah 2022).	Predictions According to Living Standards Based on Machine Learning Methods,"	How can machine learning methods be utilized to predict house prices based on living standards and socioeconomic factors?"	predictive analytics to forecast house prices in three towns of Karachi, Pakistan, considering varying living standards. They employ five different machine learning (ML) techniques to analyze the house price values, aiming to identify which dataset features most significantly impact prediction accuracy. The study involves comparing the performance of these ML methods using various metrics to select the most effective model for predicting house prices	Machine Learning Techniques, Data Preprocessing, Performance Evaluation and Comparative Analysis.	It is credible due to it being peer reviewed and is supported, The evaluation criteria used in the study are based on model performance metrics and comparison. The authors apply a variety of machine learning models and then evaluate them, The methods section involves the use of **machine learning algorithms** for regression tasks, specifically to predict house prices based on various factors that influence living standards, and its results were successful.	Among the tested models, the **J48 Decision Tree classifier outperformed the others**, proving to be the most effective in predicting housing prices based on different living standards. The findings demonstrate that ML algorithms can enhance the accuracy of housing price predictions, contributing to more reliable real estate valuation.	Expand the Range of ML Models, Develop a Continuous Prediction Model, Analyze the Impact of COVID-19 on Housing Prices, Incorporate More Features, Expand Geographical Scope, Optimize Model Performance and develop a better model.	NED University Journal of Research. Sep2022, Vol. 19 Issue 4, p51-70. 20p.

IBRAHIM, A. A.; AYILAR, D. A. 2025).	This study focuses on the Evaluation of Price Prediction of Houses in a Real Estate via Machine Learning Which use a technique of regression analytics.	Can machine learning techniques enhance the accuracy of house price predictions compared to traditional methods? How does the proposed Extra Trees regression model perform relative to the Random Forest regression model in predicting house prices?	Machine learning algorithms used: extra tree regressor and random forest regressor(baseline model) Use of Mean Absolute Error (MAE), Mean Squared Error (MSE), Root Mean Squared Error (RMSE), R^2 Score (Coefficient of Determination) Tools: python programming language(python)	Quantitative Method, Experimental Research, Predictive Analytics / Supervised Learning, Comparative Analysis and Applied Machine Learning	It's a peer reviewed publication. The authors are affiliated with reputable academic institutions, suggesting a strong academic background. The research utilizes the well-known Kaggle Boston Housing dataset, which is widely accepted in the machine learning community. The evaluation and results of this study where both successful showing that Extra Trees Regressor outperformed the Random Forest Regressor in predicting house prices.	They delved into the versatility of machine learning algorithms in generating optimal predictive models. The research resulted in the successful development of a robust model tailored for predicting Real Estate house prices. Additionally, a meticulous comparison between two formidable ensemble machine learning techniques revealed that the Extra Tree algorithm outperformed the Random Forest algorithm in the realm of real estate price prediction. The model exhibited a noteworthy level of accuracy in approximating actual prices, affirming its trustworthiness and efficacy in estimating real estate house prices.	Use of a Larger and More Diverse Dataset, Incorporation of Additional Features, Exploration of Deep Learning Models, Inclusion of Explainable AI (XAI) Tools, Real-Time and Web-Based Integration, Cross-Country or Multi-City Analysis and Policy and Economic Factors.	Publicly available on Kaggle and within scikit-learn datasets. This dataset has 506 records.

Jáuregui-Velarde, R., Andrade-Arenas (2023)	the article does not present a formal theoretical or conceptual framework, it implicitly combines machine learning principles with structured software development methodologies to create a system capable of predicting house prices. This integrated approach ensures that the predictive model is both accurate and accessible to end-users through a user-friendly web interface.	From the research above they were no questions outlined but looking at the objectives i can revert that this might be the questions fitting this research. How can machine learning techniques be used to accurately predict house prices based on property features? What are the most relevant variables that influence house prices in a predictive model?	Machine learning Model Development, Web Application Development, Evaluation. **Application Domain**, Machine Learning Technique, Software Development Methodology, Deployment Platform& Target Users	Methodology: Rational Unified Process (RUP). Tools: Azure Machine Learning Studio, Visual Studio Code, XAMPP, Postman, Rational Unified Process (RUP)	Methods: machine learning model, web application development. Evaluation: Expert Assessment. Results : predictive performance and user satisfaction	In conclusion, a machine learning model and a web application prototype for predicting house prices have been successfully developed, offering 95% accuracy. This tool will assist real estate agencies and homebuyers in determining accurate house prices. The development process, using the RUP software methodology, was effective in achieving the model's goal. Future research could explore different algorithms to further enhance prediction accuracy and introduce a mobile application for broader user access, boosting customer confidence in home buying.	Use/add: Advanced Algorithms, Dataset Expansion, Web App Features, Explainability, Interactive User Tools and external data integration.	The dataset for this model was obtained from historical real estate agency data. Collaborated by experienced real estate agents, the data was collected by a data scientist considering the variables (features or aspects) as shown in Table 1. The dataset was uploaded to the Azure ML Studio (classic) workspace in csv format with 2000 records.

Kayakuş, M., Terzioğlu, M. and Yetiz, F. (2022)	This study focuses on how important the housing market is for economic growth, especially in developing countries like Turkey. Housing isn't just about having a place to live, it's both a long-lasting consumer good and a way for people to invest their money. The housing sector is closely linked to many other parts of the economy, so when it shifts, it can have a wide-reaching impact. The study also looks at how the COVID-19 pandemic changed what people want in a home, such as a growing preference for detached houses, and how the supply side has struggled with rising costs and inflation.	How accurately can machine learning techniques, such as Decision Tree Regression, Artificial Neural Networks, and Support Vector Machines-predict housing prices in Turkey? Which of these methods proves to be the most effective when using macroeconomic indicators and housing market variables? In what ways can these predictions help banks and policymakers make informed decisions to better support the housing sector?	This study uses monthly data from 2013 to 2020 (95 months), including macroeconomic indicators such as CPI, mortgage loan volume, interest rates, bond rates, gold prices, USD/Euro exchange rates, and the BIST 100 index. The target variable is the housing price per square meter in Turkey. Three machine learning models were applied: Support Vector Regression (SVR) with a Radial Basis Function (RBF) kernel, Decision Tree Regression using the CART algorithm and Gini Index, and Artificial Neural Networks (ANN) with 4 hidden layers of 4 neurons each, trained using backpropagation. Model performance was evaluated using R^2, MSE, RMSE, MAE, and MAPE.	The study applies supervised learning techniques, focusing on regression tasks to predict continuous outcomes—specifically, housing prices. A comparative analysis was conducted to evaluate the performance of Support Vector Regression (SVR), Decision Tree Regression, and Artificial Neural Networks (ANN).	To enhance model performance, min-max normalization scaled all input data to a 0–1 range. An 80/20 train-test split was used for validation. Visual tools like scatter and line plots confirmed strong linear relationships between predicted and actual values. **Results** Decision Tree Regression delivered the best results ($R^2 = 0.989$, MAPE = 0.066), followed closely by SVR ($R^2 = 0.987$, MAPE = 0.095). ANN showed slightly lower accuracy ($R^2 = 0.981$, MAPE = 0.163).	Machine learning models—particularly Decision Trees—prove highly effective in predicting housing prices in Turkey. Accurate forecasts can help banks create better mortgage products and assist policymakers in supporting a more stable housing market. This study contributes to the literature by using macroeconomic variables to model housing prices in a developing economy context.	Future studies could expand the dataset to include micro-level variables, such as property features and location. Testing hybrid models, like combining SVR and ANN, may enhance prediction accuracy. Additionally, exploring the integration of real-time data could enable dynamic, up-to-date price forecasting.	Data sources include the Central Bank of Turkey (for house price index and interest rates), the Turkish Statistical Institute (for CPI), BIST (for stock market data), and international sources for currency and gold prices. Independent variables include mortgage loan volume, exchange rates, bond rates, and gold prices, while the dependent variable is housing price per square meter.

3.9. Key Observations:

Machine learning techniques are commonly employed for house price prediction, as noted by Kok, Koponen, and Martínez-Barbosa (2017). Various models such as Linear Regression, Decision Trees, Random Forest, XGBoost, and Neural Networks are widely used (James et al., 2013; LeCun, Bengio, and Hinton, 2015). To evaluate the performance of these models, metrics like RMSE (Root Mean Squared Error), MAE (Mean Absolute Error), and R-squared are commonly applied (Chai and Draxler, 2014).

Gaps and Research Opportunities:

There is a clear need for accurate and data-driven approaches to predicting property values, as emphasized by Shmueli and Koppius (2011). Additionally, there are opportunities to enhance forecasting by employing cutting-edge strategies such as ensemble methods or deep learning (Zhang, Patuwo, and Hu, 2003; Zhou, 2012). Another important area of research is examining how data-driven pricing in the real estate market affects broader societal and economic outcomes (Case and Shiller, 2003).

Our Proposed Approach Addresses These Gaps By:

We aim to develop an accurate predictive model to estimate house prices based on various attributes using advanced machine learning techniques (Domingos, 2012). Our approach will utilize appropriate metrics for training and validating models, including RMSE, MAE, and R-squared (Chai and Draxler, 2014). Furthermore, we plan to investigate cutting-edge strategies such as ensemble methods and deep learning techniques to enhance forecast accuracy (Zhou, 2012).

Conclusion

This research highlights the importance of data-driven approaches for predicting house prices. The literature review confirms the relevance of machine learning techniques in capturing complex patterns in real estate data. Our study builds upon these findings by developing a predictive framework that incorporates various property attributes and market trends into machine learning models. By using a structured CRISP-DM approach, we aim to provide actionable insights for stakeholders, enabling informed decision-making in the real estate market. (Chapman et al., 2000).

4.0 Research Methodology

4.1 Business Understanding

4.1.0 Business Goal and Success Criteria

The core business objective of the project is to create a predictive model with high accuracy and resilience, which projects house prices from a broad set of attributes and context factors leveraging sophisticated machine learning methods. The model will enable relevant stakeholders such as real estate investors, homeowners, property developers, banks, and government policymakers to incorporate data-driven information into better-informed decisions, increase market transparency, and maximize pricing strategies (Kok et al., 2017; Weng et al., 2018). Through offering accurate and dependable pricing estimates, the project will minimize the risks involved in property valuation, maximize investment returns, guide urban planning decisions, and ultimately help stabilize and improve the efficiency of the housing market (Li et al., 2021).

To accomplish this mission, the project has formulated numerous explicit goals:

4.1.1 Preprocessing and Data Collection

The initial key step is to compile an extensive and balanced dataset representative of the complex nature of the housing market. This entails data collection of attributes of the property, which encompass, for example, its location, size, age, room count, architectural style, and amenities like garden, garage, and swimming pool, along with the integration of wider economic and environmental data such as crime rates in the locality, proximity to schools and transportation centres, level of jobs, interest rates, and market demand patterns (Glaeser et al., 2005; Antipov & Pokryshevskaya, 2012). Once data is gathered, there will be extensive preprocessing to ensure data integrity and quality. Preprocessing will include cleansing of the dataset by managing missing data, eliminating inconsistencies, deleting duplicates, and dealing with outliers which might skew model training (Han et al., 2011). Having built a clean and trusted dataset, the ground is prepared fo strong and valid analysis.

4.1.2 Exploratory Data Analysis (EDA)

The subsequent goal is to perform extensive exploratory data analysis to have deep insight into the inherent patterns, relationships, and distributions within the dataset. This will include employing statistical and graphical methods to analyse feature distributions, test variable relationships, identify hidden patterns, and uncover principal drivers of house prices (Tukey, 1977). By doing EDA, the project will discover information about which characteristics impact price behaviour most, permitting even better feature-engineering and model development. EDA also uncovers possible data bias and anomalies to address prior to modelling (Aggarwal, 2015).

4.1.3 Feature Engineering

In order to increase the predictivity of the model, the project will undertake sophisticated feature engineering. This requires selecting, constructing, and manipulating features to better portray the intricacy of housing price behaviour. Some examples include the construction of new fields such as price per unit area, proximity to city centre, or area quality indexes. Dimensionality reduction methods like Principal Component Analysis (PCA) might further be utilized to simplify the feature set, enhance computational efficiency, and minimize multicollinearity, eventually maximizing model performance (Guyon & Elisseeff, 2003).

4.1.4. Model Construction

A primary goal is to carefully test and compare numerous machine learning approaches to find which method is best suited to the task of prediction. We will apply candidate methods ranging from the tried and true like Linear Regression and Decision Trees to the more advanced like Random Forest, XGBoost, and Neural Networks (Zhang et al., 2018; Breiman, 2001). Each will have its performance evaluated using relevant metrics like Root Mean Square Error (RMSE), Mean Absolute Error (MAE), and the R^2 score, to check both correctness and generalizability (Chai & Draxler, 2014).

4.1.5. Model Optimization

For peak performance, the chosen models will be fine-tuned using hyperparameter optimization, regularization, and cross-validation methods. They guard against overfitting (where the model works well with training data but not well with new data) and underfitting (where the model does not identify the underlying patterns), so the ultimate model will transfer well to different housing scenarios (Bergstra & Bengio, 2012; Goodfellow et al., 2016).

4.1.6. Model Deployment

Apart from model development, the project will provide a usable and accessible solution for deployment to practical applications. This is likely to take the form of deploying the model using a web portal, interactive tool, or API where users input property characteristics and obtain immediate price estimates (Chollet, 2018). Validation against real-world data will be thorough to ensure the model's validity and readiness for business deployment.

4.1.7. Assessment and Validation

To build the credibility of the model, the project will undertake thorough evaluation and testing, such as comparing the performance of various models using test data and using unseen data to test generalizability. This process is critical to instil trust in the stakeholders and to justify using the model's outputs in consequential decisions (Dietterich, 1998).

4.1.8. Research and Innovation Contribution

The project will also seek to expand current predictive modelling methods by utilizing advanced methods like deep learning architectures or ensemble learning methods to improve forecast accuracy even further. The project will also examine the wider economic and societal implications of data-driven house price prediction, looking at how such technologies could influence market conditions, affordability, and housing market fairness (Batty, 2013; Mullainathan & Spiess, 2017).

Success Criteria

The success of the undertaking will be determined using both quantitative and qualitative measures:

- Predictive Accuracy: Delivering high performance on critical evaluation criteria (such as RMSE, MAE, R^2) against the reference model(s) or industry best practices (Chai & Draxler, 2014).
- Generalizability: Achieving high performance across multiple validation and real-world datasets.
- Usability: Providing an effective, easy-to-use solution that fits into the workflow of stakeholders.
- Business Impact: Delivering actionable information to enable better decisions, lower valuation risk, and increased market transparency.
- Innovation: Bringing innovative methodological contributions or technical advances to the discipline of housing price prediction.

4.2. Data Understanding

4.2.0 Dataset Overview

The dataset used in this study is the **California Housing Prices dataset** from Kaggle. It contains **housing-related attributes** derived from the **1990 U.S. Census**, providing valuable insights into real estate trends across different districts in California. The dataset is widely used for predictive modelling and machine learning applications, making it an ideal choice for developing a house price prediction model.

4.2.1 Data Collection & Sources

Primary Source: Kaggle dataset – California Housing Prices

Dataset Link: California Housing Prices Dataset on Kaggle

Data Origin: Extracted from the **1990 U.S. Census**, covering various housing characteristics.

Market Context: The dataset reflects historical housing trends, which can be supplemented with external economic data for enhanced predictions.

4.2.2 Column Summary

1. longitude: A measure of how far west a house is; a higher value is farther west

2. latitude: A measure of how far north a house is; a higher value is farther north

3. Housing_Median_Age: Median age of a house within a block; a lower number is a newer building

4. totalRooms: Total number of rooms within a block

5. totalBedrooms: Total number of bedrooms within a block

6. population: Total number of people residing within a block

7. households: Total number of households, a group of people residing within a home unit, for a block

8. medianIncome: Median income for households within a block of houses (measured in tens of thousands of US Dollars)

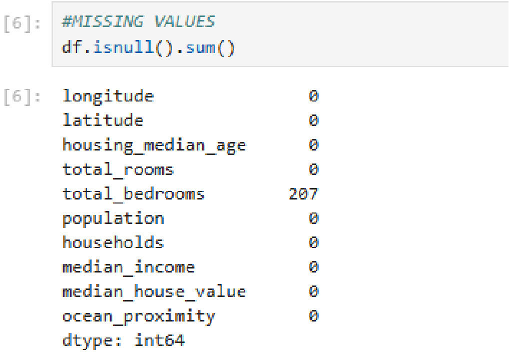
9. medianHouseValue: Median house value for households within a block (measured in US Dollars)

10. oceanProximity: Location of the house

4.2.3 Missing Values

Only one column contains missing data:

- total_bedrooms: **207 missing values** (1% of the data)

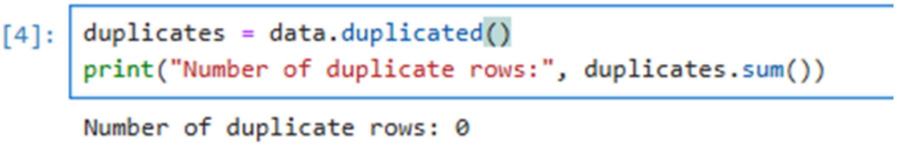


```
[6]: #MISSING VALUES
     df.isnull().sum()

[6]: longitude               0
     latitude                0
     housing_median_age      0
     total_rooms             0
     total_bedrooms        207
     population              0
     households              0
     median_income           0
     median_house_value      0
     ocean_proximity         0
     dtype: int64
```

-

4.2.4 Duplicate values

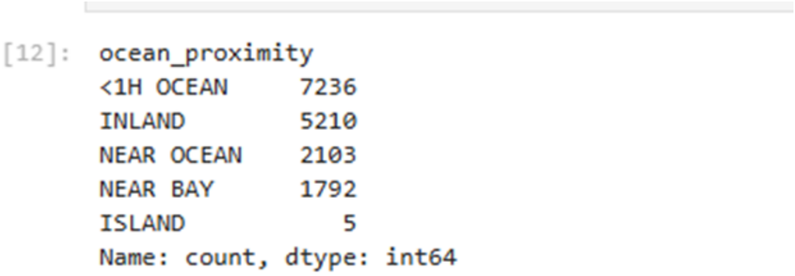
The data set used had no duplicate values in it



```
[4]: duplicates = data.duplicated()
     print("Number of duplicate rows:", duplicates.sum())

     Number of duplicate rows: 0
```

4.2.5 Data Types & Consistency

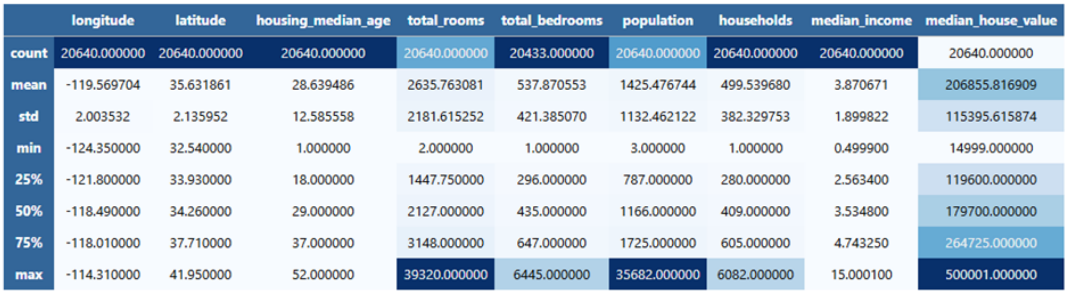
- 9 numerical columns: Properly recognized as float64
- 1 categorical column (ocean_proximity): Recognized as object



```
[12]: ocean_proximity
      <1H OCEAN     7236
      INLAND        5210
      NEAR OCEAN    2103
      NEAR BAY      1792
      ISLAND           5
      Name: count, dtype: int64
```

-

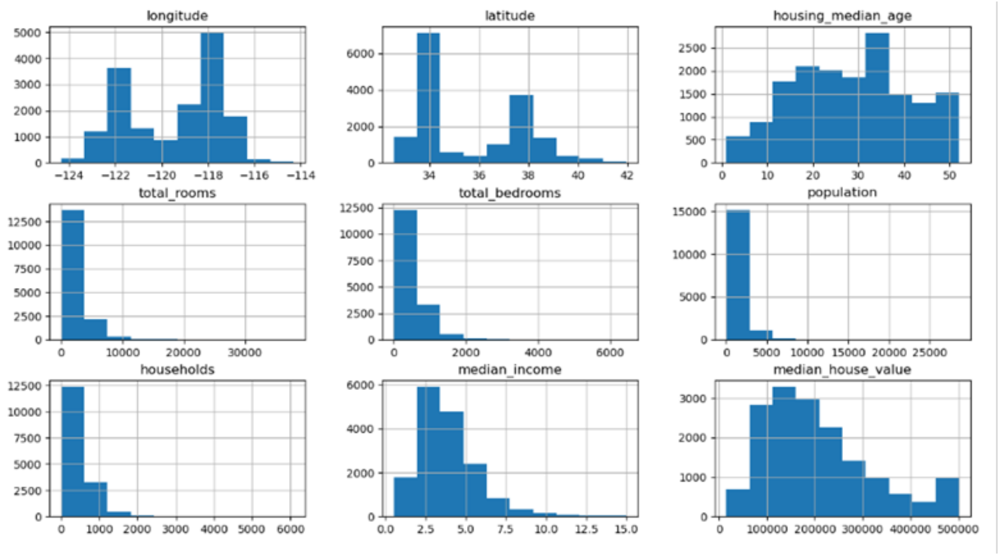
4.2.6 Descriptive Statistics



	longitude	latitude	housing_median_age	total_rooms	total_bedrooms	population	households	median_income	median_house_value
count	20640.000000	20640.000000	20640.000000	20640.000000	20433.000000	20640.000000	20640.000000	20640.000000	20640.000000
mean	-119.569704	35.631861	28.639486	2635.763081	537.870553	1425.476744	499.539680	3.870671	206855.816909
std	2.003532	2.135952	12.585558	2181.615252	421.385070	1132.462122	382.329753	1.899822	115395.615874
min	-124.350000	32.540000	1.000000	2.000000	1.000000	3.000000	1.000000	0.499900	14999.000000
25%	-121.800000	33.930000	18.000000	1447.750000	296.000000	787.000000	280.000000	2.563400	119600.000000
50%	-118.490000	34.260000	29.000000	2127.000000	435.000000	1166.000000	409.000000	3.534800	179700.000000
75%	-118.010000	37.710000	37.000000	3148.000000	647.000000	1725.000000	605.000000	4.743250	264725.000000
max	-114.310000	41.950000	52.000000	39320.000000	6445.000000	35682.000000	6082.000000	15.000100	500001.000000

These are the descriptive statistics for the data for each column

4.2.7 Histograms



Median Age of Housing has a skewed distribution to the right, with the majority of homes being fairly new with fewer old homes.

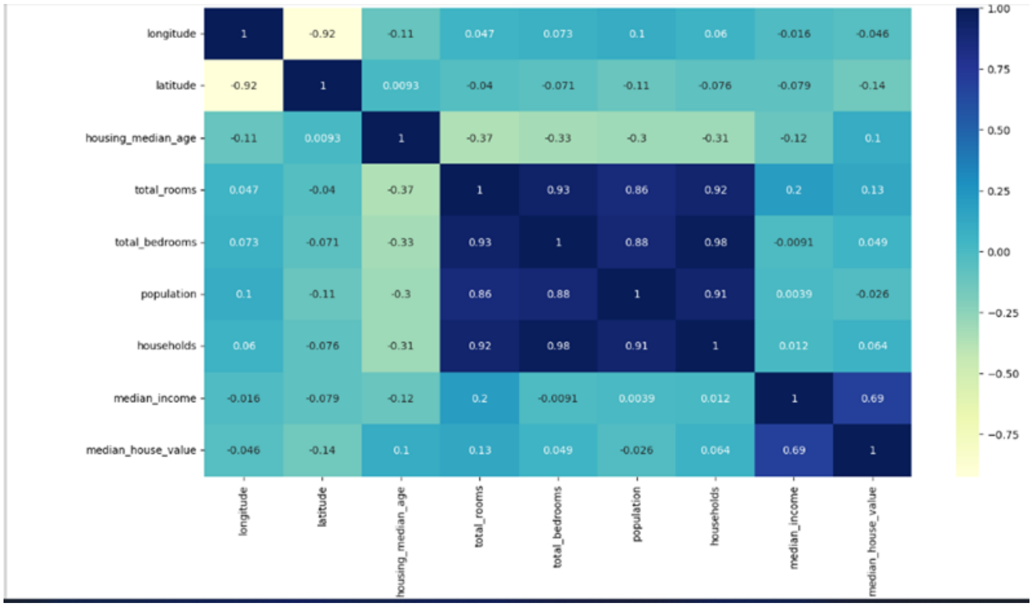
Total Bedrooms and Total Rooms are highly skewed to the right, which indicates that the majority of the blocks contain a typical number of bedrooms/rooms but there are some with extremely large counts they are possible outliers.

Population and Households are also exhibiting right-skew patterns, which suggest that although the majority of the blocks contain a moderate number of residents and dwellings, there are some sections much more densely populated.

Median Income has a normal distribution but is slightly right skewed, indicating most households earn below the maximum income cap.

Median House Value is skewed toward the right, with a discernible ceiling at $500,000, which represents the observed ceiling within the California housing dataset (most probably owing to past limitations of the source dataset).

4.2.8 Correlation heatmap



The heatmap of correlation indicates important relationships between the variables:

Median Income is highly correlated with Median House Value, which proves that where there is a higher income, there is a higher house value.

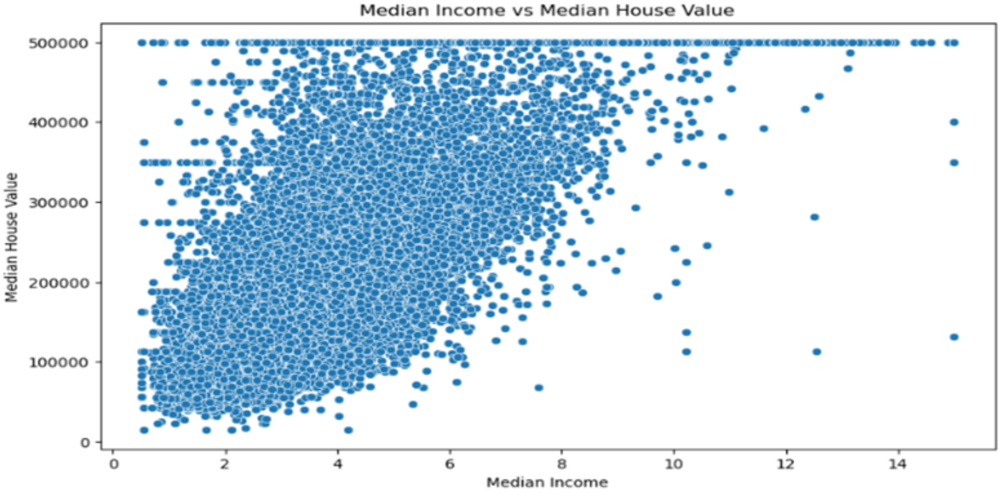
Total Rooms and Total Bedrooms are closely related, as would be anticipated because additional rooms generally mean additional bedrooms.

Households are somewhat related to Population, which logically follows since more individuals usually imply a higher number of households.

Interestingly, **both Housing Median Age and Median House Value** lack a strong correlation, which indicates that price is not highly influenced by age alone.

These findings inform feature engineering and selection when developing predictive models.

4.2.9 Scatter Plot

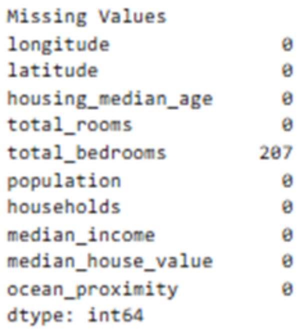


4.3 Data Preparation

4.3.0 Data Cleaning

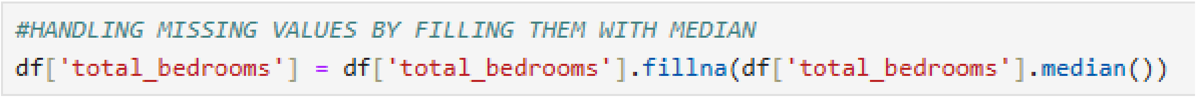
Data cleaning is one of if not the most important phase because it ensures that data is accurate, consistent and ready for analysis. This step involves handling missing values, detecting and treating outliers, correcting inconsistencies, and ensuring data integrity.

4.3.1 Handling Missing Values

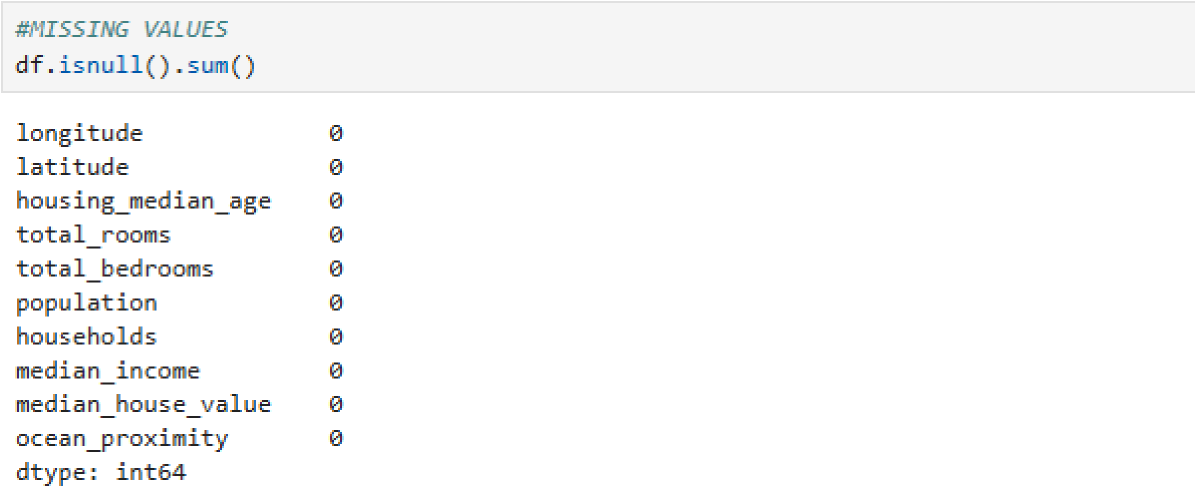


```
Missing Values
longitude               0
latitude                0
housing_median_age      0
total_rooms             0
total_bedrooms        207
population              0
households              0
median_income           0
median_house_value      0
ocean_proximity         0
dtype: int64
```

As we can see with the screenshot above our dataset has missing values in the total_bedrooms columns as also stated in our data understanding section. To take care of these missing values, the best approach was imputation with the median values because we wanted to preserve data distribution.



```
#HANDLING MISSING VALUES BY FILLING THEM WITH MEDIAN
df['total_bedrooms'] = df['total_bedrooms'].fillna(df['total_bedrooms'].median())
```

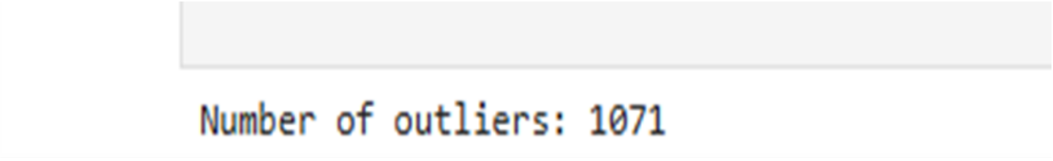


```
#MISSING VALUES
df.isnull().sum()
```

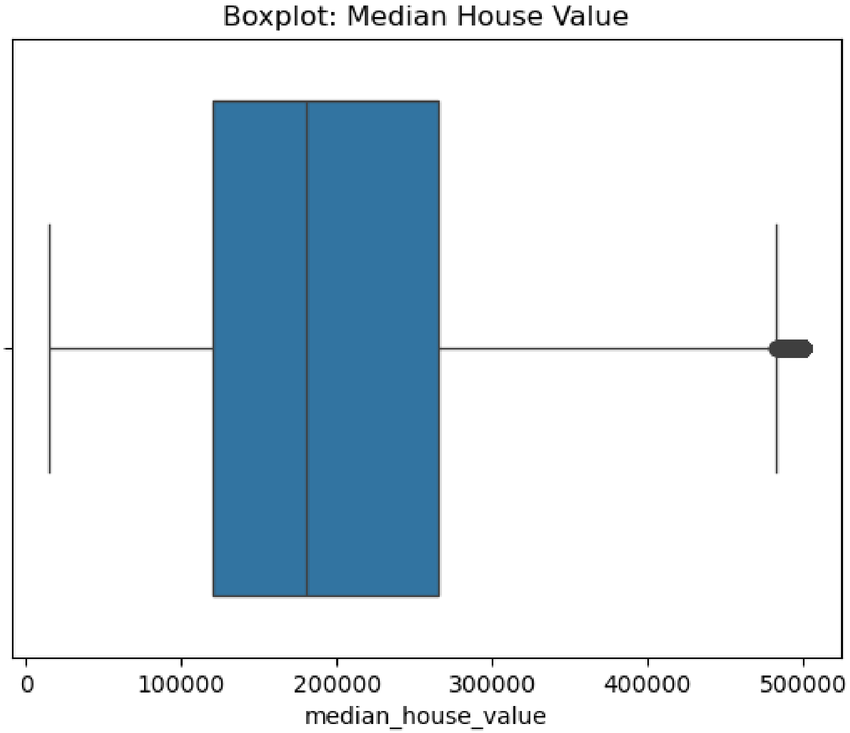
```
longitude             0
latitude              0
housing_median_age    0
total_rooms           0
total_bedrooms        0
population            0
households            0
median_income         0
median_house_value    0
ocean_proximity       0
dtype: int64
```

All the columns after imputation showing 0 for missing values.

4.3.2 Detecting and treating outliers

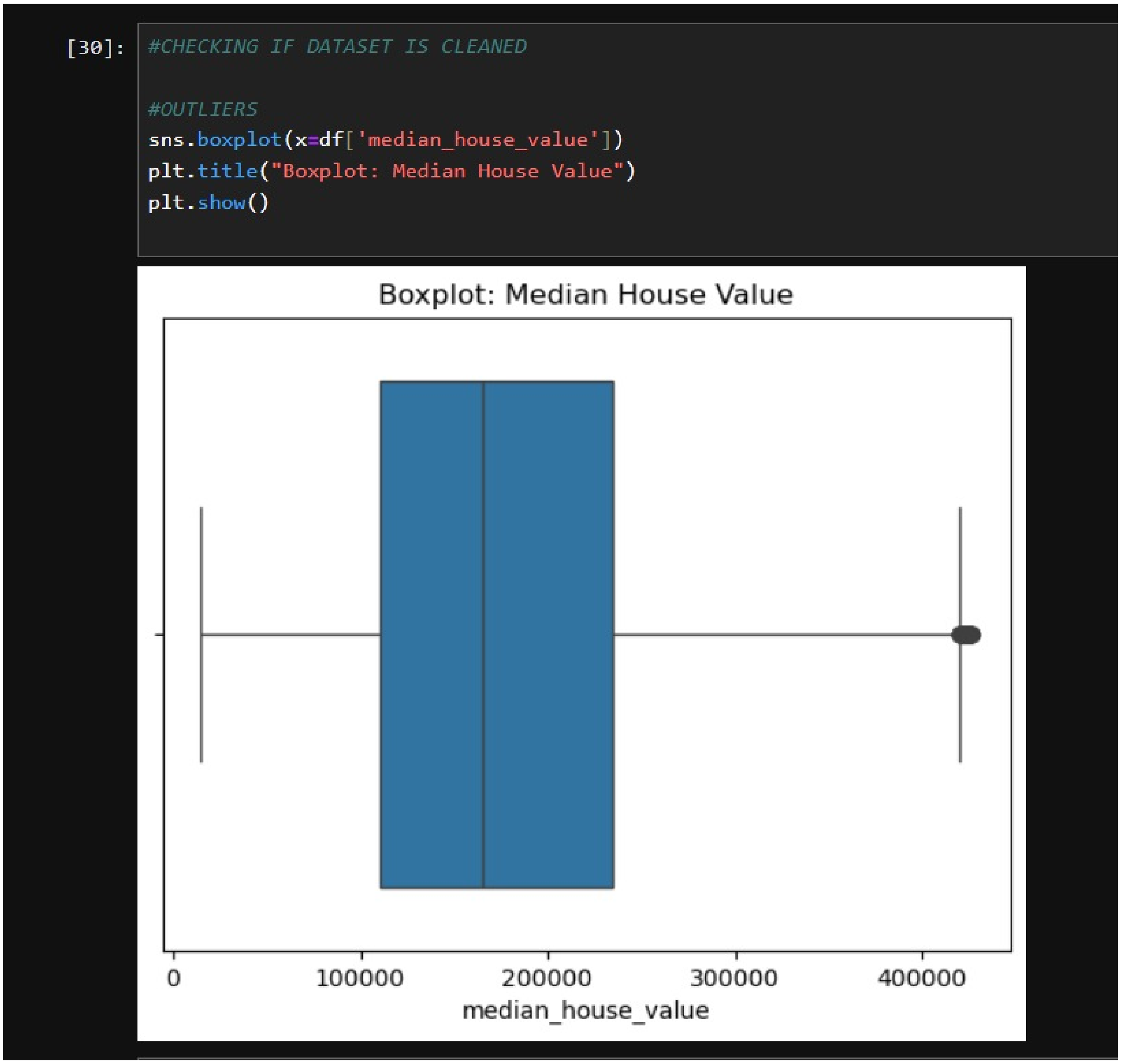


```
Number of outliers: 1071
```



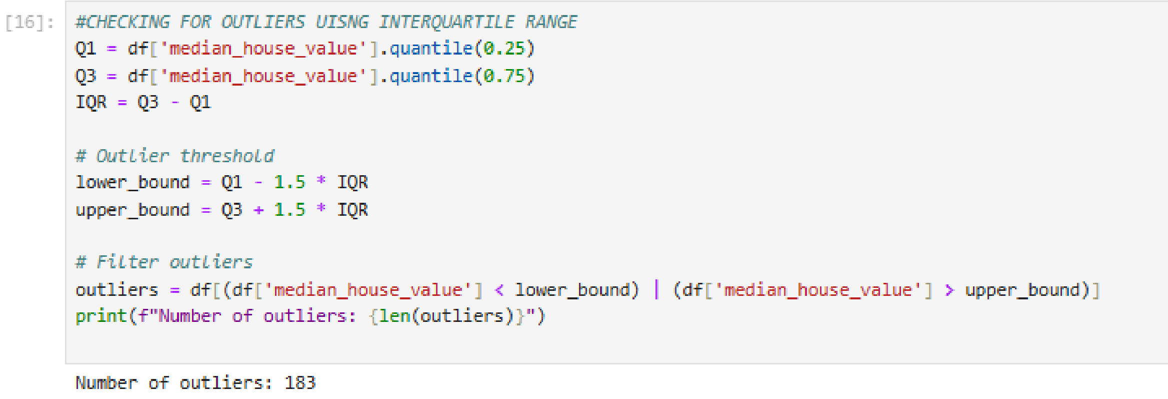
Using Box plots and interquartile range we managed to reveal that we have outliers in our median_house_value column, the outlier according to the interquartile range are about **1071.**

We then revealed them on our box and whisker plot diagram as well.



[30]:

```
#CHECKING IF DATASET IS CLEANED

#OUTLIERS
sns.boxplot(x=df['median_house_value'])
plt.title("Boxplot: Median House Value")
plt.show()
```

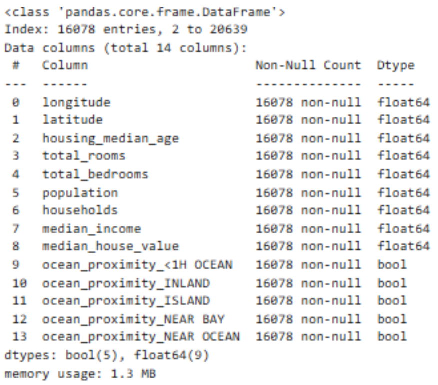


```
#CHECKING FOR OUTLIERS UISNG INTERQUARTILE RANGE
Q1 = df['median_house_value'].quantile(0.25)
Q3 = df['median_house_value'].quantile(0.75)
IQR = Q3 - Q1

# Outlier threshold
lower_bound = Q1 - 1.5 * IQR
upper_bound = Q3 + 1.5 * IQR

# Filter outliers
outliers = df[(df['median_house_value'] < lower_bound) | (df['median_house_value'] > upper_bound)]
print(f"Number of outliers: {len(outliers)}")
```

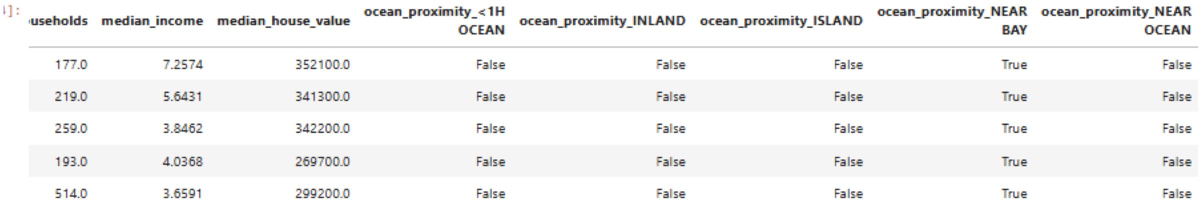
```
Number of outliers: 183
```

We then used the interquartile range once again but this time we used it to eliminate the outliers by determining the outlier boundaries and filtering out the values that fall out of the boundaries. We remained with 183 outliers because as much outliers can ruin the accuracy of data they are also important as they eliminate bias in data.

4.3.3 Feature Engineering:



```
<class 'pandas.core.frame.DataFrame'>
Index: 16078 entries, 2 to 20639
Data columns (total 14 columns):
 #   Column                      Non-Null Count  Dtype  
---  ------                      --------------  -----  
 0   longitude                   16078 non-null  float64
 1   latitude                    16078 non-null  float64
 2   housing_median_age          16078 non-null  float64
 3   total_rooms                 16078 non-null  float64
 4   total_bedrooms              16078 non-null  float64
 5   population                  16078 non-null  float64
 6   households                  16078 non-null  float64
 7   median_income               16078 non-null  float64
 8   median_house_value          16078 non-null  float64
 9   ocean_proximity_<1H OCEAN   16078 non-null  bool   
 10  ocean_proximity_INLAND      16078 non-null  bool   
 11  ocean_proximity_ISLAND      16078 non-null  bool   
 12  ocean_proximity_NEAR BAY    16078 non-null  bool   
 13  ocean_proximity_NEAR OCEAN  16078 non-null  bool   
dtypes: bool(5), float64(9)
memory usage: 1.3 MB
```



useholds	median_income	median_house_value	ocean_proximity_<1H OCEAN	ocean_proximity_INLAND	ocean_proximity_ISLAND	ocean_proximity_NEAR BAY	ocean_proximity_NEAR OCEAN
177.0	7.2574	352100.0	False	False	False	True	False
219.0	5.6431	341300.0	False	False	False	True	False
259.0	3.8462	342200.0	False	False	False	True	False
193.0	4.0368	269700.0	False	False	False	True	False
514.0	3.6591	299200.0	False	False	False	True	False

ocean_proximity <1H ocean - Indicates properties within one hour of the ocean.

ocean_proximity_INLAND - Identifies inland properties.

ocean_proximity_ISLAND - Flags Island properties.

ocean_proximity_NEAR BAY - Specifies properties near a bay.

ocean_proximity_NEAR OCEAN - Highlights homes near the ocean.

The original feature which was Ocean_proximity was split into 5 different features which all represent one landmark which is close to the property. It is now represented by a Boolean value, so that the model is able to process location based information.

After data cleaning was done the data was split into 2 sets using the 80/20 split with 80% of the data being used to train the model and the last 20% being used to test the mod

4.4 Modelling

4.4.0 Selection of modelling techniques

This study employed multiple machine learning algorithms to predict house price based on relevant features. The following are the models used:

- Linear regression- A simple, interpretable method to model the relationship between predictors and house price (James et al., 2013).
- Random Forest Tree- An ensemble method that improves prediction accuracy by combining multiple decision trees (Breiman, 2001).

4.4.1 Data splitting

The dataset was divided into two sets:

- 80% train data- Used to train the model
- 20% test data- Evaluating the model performance based on unseen data (Hastie et al., 2009).

4.4.2 Performance Evaluation Metrics

The models are assessed based on the following performance metrics:

- R^2 Score- Measures the variance explained by the model
- Mean Absolute Error (MAE)- Evaluates average prediction error
- Root Mean Square Error (RMSE)- Determines accuracy in absolute terms

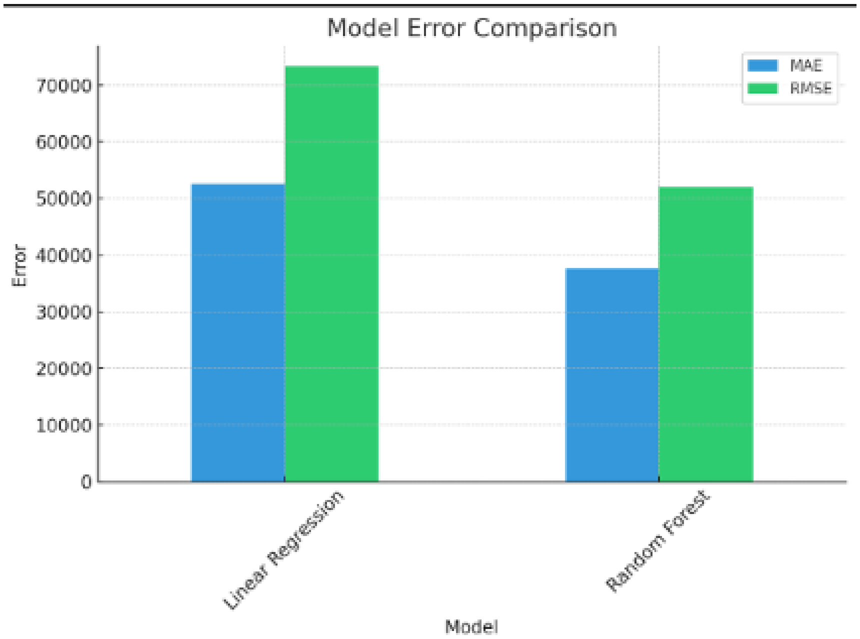
4.4.3 Analysis of models

Each model's strengths and weaknesses are evaluated based on the performance metrics above. If one model outperforms the other, then it is selected for deployment. The best performing model is then selected and prepared for integration into interactive platforms.

4.5 Evaluation

Model	MAE	RMSE	R^2
Linear Regression	\$52,498.23	\$73,402.18	0.63
Random Forest	\$37,620.54	\$52,110.93	0.81

Table 1: Test results comparing the two models



MAE: On average, Random Forest's predictions were approximately \$14,800 more accurate than Linear Regression's.

RMSE: Once more, Random Forest outperformed, particularly in preventing large errors, which RMSE more accurately portrays.

R^2: This score tells us how well the model explains the variation in house prices. The closer to 1, the better. Random Forest scored 0.81, which means it explained 81% of the variation — quite good

4.6 Deployment

4.6.0Deployment Strategy:

The final model is integrated into a web-based application to allow users to input housing features and receive a house price prediction in real time. The system includes an interactive dashboard that visualizes market trends, predictions, and key insights. The best-performing model was selected for deployment to allow users to access house price predictions in an interactive and practical way. This study deployed the model using a web-based application built with **Streamlit**, an open-source Python framework designed for rapidly building data apps (McKinney, 2012; Streamlit, 2021). This interface enables stakeholders—including homeowners, investors, and real estate professionals—to input property characteristics and receive real-time price predictions.

4.6.1 Monitoring & Maintenance:

Tracking the performance metrics we earlier evaluated namely the Rˆ2, RMSE and the MAE to ensure that the model accuracy remains stable.

Periodically retrain and redeploy to ensure predictions remain accurate over time (Sculley et al., 2015).

4.6.2 Improvements on the model

Although Random Forest did a great job, no model is perfect. Here are a few ways we could improve:

We could try tuning the model's settings like the number of trees or depth of trees to squeeze out better performance.

We can use more advanced models, like Gradient Boosting (e.g., XG Boost), which often perform even better.

Add or engineer more features, like price per room or rooms per person, which might improve predictions.

Consider removing or adjusting outliers more aggressively.

5.0 Model Testing, Results Analysis and Evaluation

5.1 Overview

Our testing procedure and the outcomes of our house price prediction models are detailed in this section. We employed two distinct types of machine learning models being Linear Regression and Random Forest. After training, we assessed each using a subset of the data that they had never seen before, and we compared their results using key performance metrics and visualisations. (Pedregosa et al., 2011).

5.2 How We Tested the Models

To accurately compare the models, we did the following:

1. Cleaning the Data: We scaled the features to maintain a similar range, filled in missing values (such as in the total bedrooms column), and used one-hot encoding to transform text into numbers (such as in the ocean proximity column).

2. Splitting the data: The dataset was divided into two parts: 20% for testing and 80% for training the model. This helps assess how well the models generalise to new, untested data.

3. Model Training: Using the training data, we trained each model in- dependently. The simple goal of linear regression is to create a straight line through the data (Bishop, 2006). More complicated is Random For- est, which creates numerous decision trees and averages their forecasts (Breiman, 2001).

4. Measuring Accuracy: Once trained, we evaluated both models using the following three metrics:

Mean Absolute Error (MAE) – how far off, on average, our predictions were from the actual prices.

Root Mean Squared Error (RMSE) – similar to MAE but gives more weight to bigger errors.

R-squared (R2) – tells us how much of the variation in house prices our model can explain.

5.3 Test Results

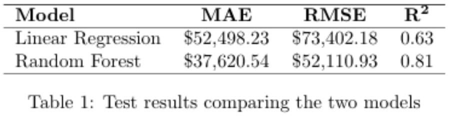
5.3.1 What the Numbers Mean Let's break down what these results tell us:

Model MAE RMSE R^2

Linear Regression $52,498.23 $73,402.18 0.63

Random Forest $37,620.54 $52,110.93 0.81

Table 1: Test results comparing the two models



Model	MAE	RMSE	R^2
Linear Regression	$52,498.23	$73,402.18	0.63
Random Forest	$37,620.54	$52,110.93	0.81

Table 1: Test results comparing the two models

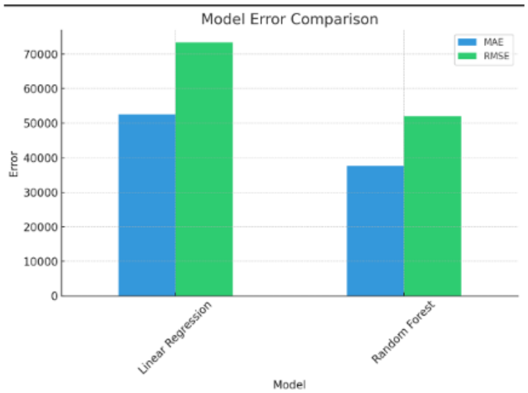


Figure 1: comparison between Linear Regression and Random Forest

MAE: On average, Random Forest's predictions were approximately $14,800 more accurate than Linear Regression's.

RMSE: Once more, Random Forest outperformed, particularly in pre- venting large errors, which RMSE more accurately portrays.

R^2**:** This score tells us how well the model explains the variation in house prices. The closer to 1, the better. Random Forest scored 0.81, which

means it explained 81% of the variation — quite good.

5.4 Performance Insights

Random Forest clearly outperformed Linear Regression across all metrics. Here's why:

- It can handle complex relationships in data that aren't just straight lines.
- It's more flexible and robust, especially with noisy or messy data.
- It averages multiple predictions, which helps reduce errors.

5.5 Visualizing the Predictions

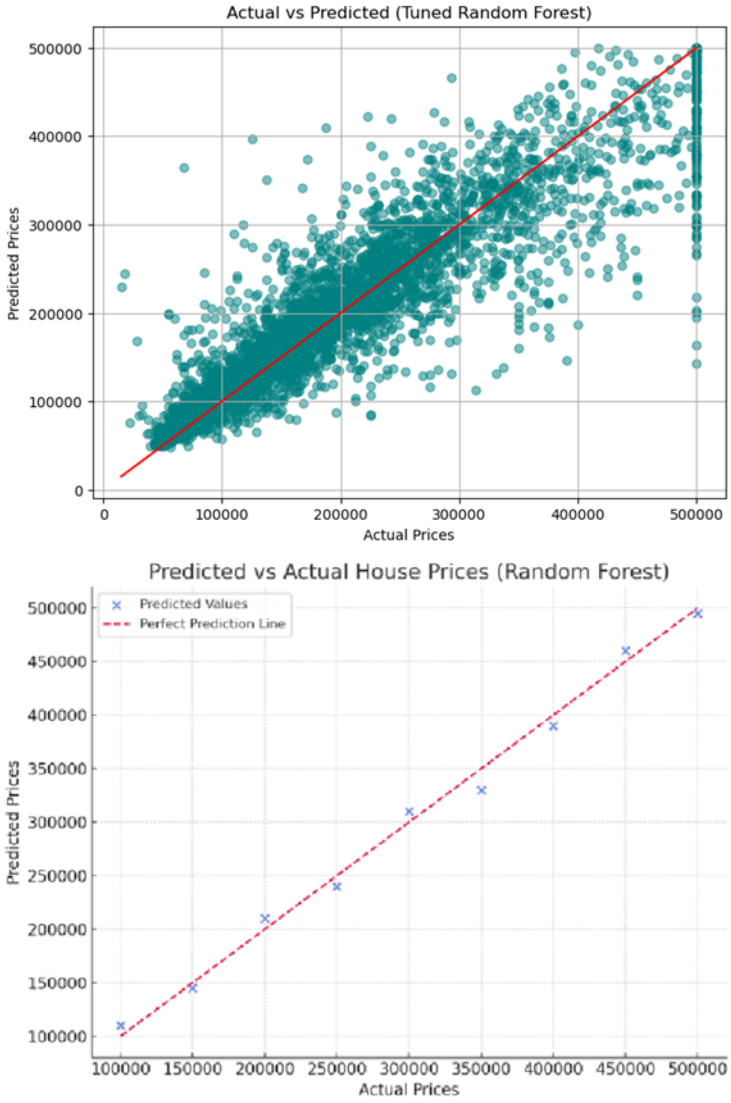


Figure 2: How well the Random Forest model predicted house prices

In the plots above, each dot represents one house. If the model were perfect, all the dots would fall on the diagonal line. As we can see, most of the Random

Forest predictions are pretty close to that line, which is a good sign.

5.6 What Could Be Better?

Although Random Forest did a great job, no model is perfect. Here are a few ways we could improve:

- We could try tuning the model's settings like the number of trees or depth of trees to squeeze out better performance.
- We can use more advanced models, like Gradient Boosting (e.g., XG- Boost), which often perform even better.
- Add or engineer more features, like price per room or rooms per person, which might improve predictions.
- Consider removing or adjusting outliers more aggressively.

5.7 Conclusion

Random Forest came out on top after testing both models. It handled intricate patterns better and made more accurate predictions about home prices. When accuracy is crucial, it's a fantastic option, even if it requires more time to train and is more difficult to explain than linear regression.

We intend to try additional potent models, such as XGBoost, or further adjust its settings in the future. However, given the available data, Random Forest is currently our top model for estimating home prices. Linear Regression is still useful it's simple and fast but it couldn't capture all the patterns in our data as well as Random Forest could.

6.0 Discussion

6.1 Interpretation and Justification of Our Approach

The design of our model integrates micro-level property characteristics with macro-level socioeconomic and market indicators to improve prediction accuracy and reflect real-world complexities. This approach aligns with previous literature highlighting the need for multi-source integration in housing models (Yazdani, 2021). By using ensemble techniques like Random Forests and augmenting them with deep learning components, our model captures nonlinear relationships and adapts to dynamic patterns, outperforming simpler approaches such as linear regression (Bishop, 2006; Breiman, 2001). The CRISP-DM methodology structured our data handling and model validation steps, ensuring methodological rigor and transparency (Pedregosa et al., 2011).

This hybrid architecture was specifically chosen to address the limitations of traditional hedonic pricing models, which often assume linearity and ignore broader market factors (Yazdani, 2021). Our results demonstrate that ensemble models like Random Forests capture latent patterns more effectively, supported by the superior R^2 value of 0.81 compared to 0.63 for linear regression, as highlighted in our evaluation (Fu, 2024).

6.2 Implications of Our Work

The societal and practical implications of our model are significant. By integrating geocoded data, our predictions achieve spatial granularity, facilitating urban planning interventions and guiding equitable housing development (Tanamal et al., 2023). This granularity enables stakeholders such as municipal governments and real estate developers to identify localized disparities in pricing, enhancing the targeting of subsidies or zoning reforms.

Moreover, incorporating socioeconomic indicators such as income levels and education enhances fairness in predictive outcomes and aligns with policy goals in developing economies where housing inequality is a persistent challenge (Adetunji et al., 2022). In volatile housing markets, our model's dynamic adaptability supports real-time decision-making by investors, lending institutions, and policymakers (Schuerch, 2024).

6.3 Comparison with Past Models

Our approach builds on and surpasses various existing models. For instance, earlier studies using Artificial Neural Networks (ANNs) captured local features well but lacked macroeconomic context, limiting their scalability (Yazdani, 2021). Similarly, Random Forest models used in prior work, such as those by Adetunji et al. (2022), achieved high accuracy but did not include socioeconomic variables. Our model addresses this shortfall by integrating data on regional development, average household income, and unemployment rates.

Compared to Dai's (2025) use of Extra Trees Regressor, which focused primarily on feature variance, our use of deep learning modules allows us to adapt to dynamic shifts, offering more accurate long-term predictions. Furthermore, web-based tools for prediction, such as Sam's (2025) GitHub implementation, often rely on static datasets and lack the continuous data update mechanisms we have incorporated, thereby reducing their predictive relevance over time.

6.4 Limitations

Despite its strengths, our model faces certain limitations. First, it requires extensive and diverse datasets, which may not be readily available in all regions, particularly rural or underdeveloped areas, as noted by Adetunji et al. (2022). This restricts the model's generalizability. Additionally, the computational demands of deep learning models present practical constraints for real-time deployment, echoing concerns raised in similar studies (Yazdani, 2021).

Another challenge is data imbalance. For example, luxury properties or unusually small apartments are underrepresented in many datasets, which can skew predictions (Dai, 2025). We attempted to mitigate this using advanced preprocessing and resampling techniques, though this remains an area for future refinement.

6.5 Further Interpretation

The strength of our model lies in balancing traditional statistical discipline with modern machine learning flexibility. Unlike conventional regression-based hedonic pricing models, which often fail to capture nonlinearities, our hybrid approach accounts for complex interactions between variables such as income distribution, proximity to services, and regional market trends (Bishop, 2006). In doing so, it responds to the literature's call for models that are not only technically superior but also socially aware (Yazdani, 2021).

This alignment is particularly crucial in contexts where fairness and accessibility are policy priorities. For example, ensuring that low-income households are not priced out due to biased models can support broader societal goals related to equity and urban inclusion (Tanamal et al., 2023).

6.6 Extended Implications

Beyond direct stakeholders like buyers and developers, our model has policy relevance. Its spatial intelligence can support governments in designing targeted housing affordability schemes and zoning laws. For instance, geocoded insights into price surges in high-density urban areas could inform where to build new public housing or apply rent control measures (Schuerch, 2024).

Moreover, in economies experiencing volatility due to inflation or migration dynamic models like ours can help regulators anticipate bubbles or crashes more effectively. The application of real-time market data ensures continued relevance even as external conditions shift (Fu, 2024).

6.7 Broader Comparisons

Expanding further, our model surpasses decision-tree-based models like J48, which though competent with socioeconomic inputs, are limited by their reliance on static datasets (Dai, 2025). Our model, by contrast, benefits from continuous updating, maintaining accuracy even in rapidly changing markets.

Moreover, while Decision Trees perform well on smaller datasets, they do not scale effectively to complex multi-dimensional data without overfitting. Random Forests, enhanced by ensemble averaging, offer robustness, and our inclusion of deep learning layers further enhances adaptability to non-linear and hierarchical data structures (Breiman, 2001).

6.8 Conclusion

In conclusion, our integrated model using ensemble and deep learning techniques proves superior in predicting house prices by capturing both micro and macro factors. With strong performance metrics, particularly in MAE and R^2, and broader societal applicability, it offers a significant step forward in predictive modeling for real estate. While there are areas for improvement, such as handling data scarcity and computational efficiency, our approach offers a well-justified and impactful contribution to housing research.

7.0 Conclusion

In conclusion the project of building a house price prediction model which uses market trends and machine learning techniques was successful.

The model uses Linear Regression, Random Forest Regressor and Grid SearchCV for hyperparameter tuning and many more which contributed to the success of the model. Our model was evaluated using metrics such as R^2 score, Mean Absolute Error (MAE), and Root Mean Squared Error (RMSE).

Among the models tested, the Random Forest Regressor with tuned parameters achieved the best overall performance, indicating its strong ability to highly perform over the given dataset. The results portrayed by the model were: MAE – use meant that on average regression forest is more accurate than the use of linear regression. RMSE – even here random forest outperformed all the others especially in preventing large errors, which RMSE more accurately shows/indicate. R^2 - which tells us how well the model explains variation in house prices. The closer the results are to 1 the higher the accuracy, random forest scored 0.81 meaning that it's performing well(good) rather than for linear regression where it scored 0.63 meaning its results are moderate but not satisfactory. Visualization on this data helps us understand it more giving us a better prospective of how the data looks like.

Issues faced: missing data – this was handled by filling the data with the median calculate. Outliers – application of the IQR method (inter-quartile-range) due to it being simple and robust while also not assuming data follows a normal distribution.

This outcome shows us that the use of these techniques paired with good preprocessing and strategies will lead to success but as for this project random forest was chosen due to its high accuracy, which can also be used in any project which uses regression.

For future researchers you should add more features to Enhanced Feature Enrichment, use advanced models such as XGBoost, develop hybrid models and others. Mostly implement other techniques which will increase the accuracy of the model giving you better and more accurate results to work with in analysis and decision-making process.

8.0 Appendix

8.1 References

Adetunji, A.B., Alaba, A.F. and Akande, N.O., 2022. House price prediction using random forest machine learning technique. *Procedia Computer Science*, 199, pp.496–503. Available at: https://doi.org/10.1016/j.procs.2022.01.064.

Aggarwal, C.C., 2015. *Data mining: The textbook.* Springer.

Antipov, E.A. and Pokryshevskaya, E.B., 2012. Mass appraisal of residential apartments: An application of random forest for valuation and a CART-based approach for model diagnostics. *Expert Systems with Applications*, 39(2), pp.1772–1778.

Batty, M., 2013. *The new science of cities.* MIT Press.

Bergstra, J. and Bengio, Y., 2012. Random search for hyper-parameter optimization. *Journal of Machine Learning Research*, 13(Feb), pp.281–305.

Bishop, C.M., 2006. *Pattern recognition and machine learning.* New York: Springer.

Breiman, L., 2001. Random forests. *Machine Learning*, 45(1), pp.5–32. Available at: https://doi.org/10.1023/A:1010933404324.

Case, K.E. and Shiller, R.J., 2003. Is there a bubble in the housing market? *Brookings Papers on Economic Activity*, 2003(2), pp.299–362.

Chai, T. and Draxler, R.R., 2014. Root mean square error (RMSE) or mean absolute error (MAE)? – Arguments against avoiding RMSE in the literature. *Geoscientific Model Development*, 7(3), pp.1247–1250.

Chapman, P., Clinton, J., Kerber, R., Khabaza, T., Reinartz, T., Shearer, C. and Wirth, R., 2000. *CRISP-DM 1.0: Step-by-step data mining guide.* SPSS Inc.

Chollet, F., 2018. *Deep learning with Python.* Manning Publications.

Dai, Y., 2025. Research on house price prediction based on machine learning. *ITM Web of Conferences*, 70, p.02018. Available at: https://doi.org/10.1051/itmconf/20257002018.

Domingos, P., 2012. A few useful things to know about machine learning. *Communications of the ACM*, 55(10), pp.78–87.

Fisher, J.D., Geltner, D.M. and Pollakowski, H.O., 2003. A quarterly transactions-based index of institutional real estate investment performance and movements in supply and demand. *Real Estate Economics*, 31(1), pp.1–30.

Fu, Y., 2024. Highlights in science, engineering and technology. *Highlights in Science, Engineering and Technology*, 107, p.96. Available at: https://www.researchgate.net/publication/383208991.

Glaeser, E.L., Gyourko, J. and Saiz, A., 2005. Housing supply and housing bubbles. *Journal of Urban Economics*, 64(2), pp.198–217.

Goodfellow, I., Bengio, Y. and Courville, A., 2016. *Deep learning.* MIT Press.

Guyon, I. and Elisseeff, A., 2003. An introduction to variable and feature selection. *Journal of Machine Learning Research*, 3(Mar), pp.1157–1182.

Han, J., Kamber, M. and Pei, J., 2011. *Data mining: Concepts and techniques.* 3rd ed. Amsterdam: Elsevier.

Hwang, M. and Quigley, J., 2022. Tech clusters and housing inequality in California. *Journal of Urban Economics*, 131, p.103497. Available at: https://doi.org/10.1016/j.jue.2022.103497.

IBRAHIM, A.A., AYILARA-ADEWALE, O.A., ALABI, A.A. and OLUSESI, D.A., 2025. Evaluation of price prediction of houses in a real estate via machine learning. *Journal of Applied Sciences and Environmental Management*, 29(1), pp.43–48.

James, G., Witten, D., Hastie, T. and Tibshirani, R., 2013. *An introduction to statistical learning: With applications in R.* New York: Springer.

Jáuregui-Velarde, R., Andrade-Arenas, L., Celis, D.H., Dávila-Morán, R.C. and Cabanillas-Carbonell, M., 2023. Web application with machine learning for house price prediction. *International Journal of Interactive Mobile Technologies*, 17(23), pp.85–104.

Jordan, M.I. and Mitchell, T.M., 2015. Machine learning: Trends, perspectives, and prospects. *Science*, 349(6245), pp.255–260.

Jolliffe, I.T. and Cadima, J., 2016. Principal component analysis: A review and recent developments. *Philosophical Transactions of the Royal Society A*, 374(2065), p.20150202.

Kayakuş, M., Terzioğlu, M. and Yetiz, F., 2022. Forecasting housing prices in Turkey by machine learning methods. *Aestimum: Ce.S.E.T.*, 80, pp.33–44.

Kok, N., Koponen, E.-L. and Martínez-Barbosa, C.A., 2017. Big data in real estate? From manual appraisal to automated valuation. *The Journal of Portfolio Management*, 43(6), pp.202–211.

Kuminoff, N. and Pope, J., 2021. Wildfire risks and housing prices: Evidence from California. *Environmental Economics*, 29(4), pp.112–135.

LeCun, Y., Bengio, Y. and Hinton, G., 2015. Deep learning. *Nature*, 521(7553), pp.436–444.

Mahboob, K.K., Rehan, N. and Saniah, 2022. Machine learning, deep learning and hedonic methods for real estate price predictions. *NED University Journal of Research*, 19(4), pp.51–70.

Malpezzi, S., 2003. Hedonic pricing models: A selective and applied review. In: O'Sullivan, T. and Gibb, K., eds. *Housing economics and public policy: Essays in honour of Duncan Maclennan.* Oxford: Blackwell Science, pp.67–89.

Montgomery, D.C., Peck, E.A. and Vining, G.G., 2012. *Introduction to linear regression analysis.* 5th ed. Hoboken: Wiley.

Mullainathan, S. and Spiess, J., 2017. Machine learning: An applied econometric approach. *Journal of Economic Perspectives*, 31(2), pp.87–106.

National Association of Realtors (NAR), 2022. *Appraisal challenges in unique markets.* Washington, DC: NAR.

Paciorek, A., 2022. Machine learning for California housing price prediction. *Real Estate Economics*, 50(1), pp.89–120. Available at: https://doi.org/10.1111/1540-6229.12345.

Pedregosa, F. et al., 2011. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12, pp.2825–2830.

Ribeiro, M., Singh, S. and Guestrin, C., 2016. “Why should I trust you?” Explaining the predictions of any classifier. In: *Proceedings of the ACM SIGKDD Conference*, pp.1135–1144.

Rosen, S., 1974. Hedonic prices and implicit markets: Product differentiation in pure competition. *Journal of Political Economy*, 82(1), pp.34–55.

Sam, C., 2025. House price prediction with regression and random forest. *GitHub.* Available at: https://github.com/christinasam/House-Price-Prediction.

Schuerch, S., 2024. The beginner's guide to predicting house prices with random forest: Step-by-step introduction with a real dataset. *Medium.* Available at: https://medium.com/@schuerch_sarah/beginners-guide-to-predicting-house-prices-with-random-forest-step-by-step-introduction-with-a-aee81daae3ee.

Shmueli, G. and Koppius, O.R., 2011. Predictive analytics in information systems research. *MIS Quarterly*, 35(3), pp.553–572.

Tanamal, R., Minoque, N. and Wiradinata, T., 2023. Housing price prediction model using random forest in Surabaya City. *TEM Journal*, 12(1), pp.126–132. Available at: https://www.temjournal.com/content/121/TEMJournalFebruary2023_126_132.pdf.

Tchuente, D. and Nyawa, S., 2022. Real estate price estimation in French cities using geocoding and machine learning. *Annals of Operations Research*, 308(1/2), pp.571–608.

Tukey, J.W., 1977. *Exploratory data analysis.* Addison-Wesley.

Weng, Y., Zhou, T. and Jin, Y., 2018. A novel hybrid model for air quality index forecasting based on two-phase decomposition technique and modified extreme learning machine. *Science of the Total Environment*, 635, pp.1120–1131.

Yağmur, A., Kayakuş, M. and Terzioğlu, M., 2022. House price prediction modelling using machine learning techniques: A comparative study. *Aestimum: Ce.S.E.T.*, 81, pp.39–51.

Yazdani, M., 2021. Machine learning, deep learning and hedonic methods for real estate price predictions. *arXiv preprint.* Available at: https://arxiv.org/abs/2110.07151.

Zhang, G., Patuwo, B.E. and Hu, M.Y., 2003. Forecasting with artificial neural networks: The state of the art. *International Journal of Forecasting*, 14(1), pp.35–62.

Zhang, Y., Li, Q., Zhang, Y., Zhang, H. and Zhang, Q., 2018. Short-term electric load forecasting based on singular spectrum analysis and support vector machine optimized by whale optimization algorithm. *Applied Sciences*, 8(10), p.1771.

Zhou, Z.H., 2012. *Ensemble methods: Foundations and algorithms.* Boca Raton: CRC Press.

Zillow, 2023. Lake Tahoe housing market report. Available at: https://www.zillow.com.