**ADVANCED MACHINE LEARNING**

**Real Estate Price Prediction Using Machine Learning**

**TABLE OF CONTENTS**

[**SECTION A** 3](#_Toc165580604)

[1.1 INTRODUCTION 3](#_Toc165580605)

[1.2 DATASET 3](#_Toc165580606)

[1.2.1 Data Source 3](#_Toc165580607)

[1.2.2 Data Description 4](#_Toc165580608)

[1.3 LITERATURE REVIEW 5](#_Toc165580609)

[1.4 RESEARCH QUESTIONS 7](#_Toc165580610)

[**SECTION B** 8](#_Toc165580611)

[2.1 DATA ANALYSIS 8](#_Toc165580612)

[2.1.1 Data Cleaning and Processing 8](#_Toc165580613)

[2.1.2 Exploratory Descriptive Analysis (EDA) 10](#_Toc165580614)

[2.1.3 Correlation Between Features 12](#_Toc165580615)

[**SECTION C** 16](#_Toc165580616)

[3.1 TOOLS AND MODELS 16](#_Toc165580617)

[3.2 MACHINE LEARNING MODEL DEPLOYMENT 16](#_Toc165580618)

[3.2.1 Split data into Training and Test Sets 16](#_Toc165580619)

[3.2.2 Feature Engineering 17](#_Toc165580620)

[3.3.3 Unsupervised Machine Learning Model 19](#_Toc165580621)

[3.3.4 Supervised Machine Learning Model 21](#_Toc165580622)

[3.3 MACHINE LEARNING RESULTS SUMMARY 23](#_Toc165580623)

[**SECTION D** 24](#_Toc165580624)

[4.1 LEGAL, SOCIAL, ETHICAL AND/OR PROFESSIONAL ISSUES (LSEPI) 24](#_Toc165580625)

[4.2 CONCLUSION 24](#_Toc165580626)

[REFERENCES 26](#_Toc165580627)

# **SECTION A**

## INTRODUCTION

The complexity and sometimes uncertainty of the real estate market makes it very important for investors, home buyers, agents and brokers to make the best informed decisions when buying or selling a house as this is very crucial to mitigate risks and to make the best deals (Singh et al. 2023). In fact, real estate asset management requires an in-depth understanding and consideration of the prevalent demand and supply dynamics dictated by the fluctuations in the macroeconomic factors of an area (Tekouabou et al. 2023). Possessing the right information ahead of making a decision is therefore the thin line between making the right property choice or not; but having this information is challenging using traditional means of determining the price of a house (Abut et al. 2023).

Conventional means of predicting the price of real estate is often error prone, inaccurate and long-drawn-out because it requires manual application of human experience, understanding of the market forces, and making of subjective analysis (Abut et al. 2023). What if there is an accurate and non-subjective way to predict the price of houses? Advancement in technology through the use of Machine Learning (ML), the availability of large volume real estate data, and the massive computing power available today to process this data has now made it possible to more accurately predict house prices (Abut et al. 2023; Singh et al. 2023).

This study aims at developing a Machine Learning model that predicts the price of houses. Stakeholders will find models like this useful in helping their clients make the best informed decisions and knowing where to invest in a way that suits their business needs.

## 1.2 DATASET

### 1.2.1 Data Source

The dataset used for this study is a data of real estate listings in the United States of America (USA). The data was gotten from, and is hosted on [*Kaggle*](https://www.kaggle.com/datasets/ahmedshahriarsakib/usa-real-estate-dataset/data), a Data Science platform where ML and Data experts find and explore datasets. Originally, the data as hosted on *Kaggle* was gotten from [*Realtor*](http://www.realtor.com), a top real estate listing website in the USA with office based in California (Sakib 2024).

The dataset on the source site is dynamic because new listings are continually added. For the purpose of this work, the data extracted from *Kaggle* as of 18th April 2024 is hosted on google drive so that for the full cycle of this study, the data is static. The *Kaggle* dataset can be found [here](https://www.kaggle.com/datasets/ahmedshahriarsakib/usa-real-estate-dataset/download?datasetVersionNumber=25) and the static dataset hosted on google drive and used throughout this project is found [here](https://drive.usercontent.google.com/download?id=1wBJmx7yGbrjRFdSZOp11PkfPfC3e4NNG&export=download&confirm=%7b%7bVALUE%7d%7d).

### 1.2.2 Data Description

*Throughout this work, Python programming language (and its data science libraries) was used, and the Integrated Development Environment (IDE) used was JupyterLab Notebook. The Notebook that contains* ***all*** *the codes and scripts can be accessed and run* [*here*](https://colab.research.google.com/drive/1oI5eN8E13CeZoGJZPFTsp5UTegtaNxpg?usp=sharing)*.*

The data has 2,226,382 rows and 12 fields. The data has 4 categorical fields and 8 numerical fields.

Field and Attribute Types:

1. *brokered\_by*: This is the privacy-encoded ID of the broker that published the listing. Data type is numerical.
2. *status*: This is the status of the house, and the values can be ‘for\_sale’, ‘ready\_to\_build’, or ‘sold’. Data type is categorical.
3. *price*: This is the amount the house was sold/valued. This feature will be the **target variable** for supervised learning models. Datatype is numerical.
4. *bed*: The number of bedrooms the house has. Datatype is numerical.
5. *bath*: The number of bathrooms the house has. Datatype is numerical.
6. *acre\_lot*: Size of the land on which the house stands in acres. Datatype is numerical.
7. *street*: Privacy-encoded street address where house is located. Data type is numerical.
8. *city*: The name of the city where house is located. Data type is categorical.
9. *state*: The state in USA the house is located. Data type is categorical.
10. *zip\_code*: The zip code of the area where the house is located. Data type is numerical. This will be converted to categorical because since zip code is not numerical.
11. *house\_size*: The building size of the house in square feet. Data type is numerical.
12. *prev\_sold\_date*: The date the house was previously sold. Data type is categorical.

## 1.3 LITERATURE REVIEW

In their study on the use of ML to predict real estate price, Yadav et al. (2023) acknowledged the impact different parameters can have on the price of a property, and were motivated to study the most accurate model that can predict real estate property price. A Bangalore real estate listing data was used and models like linear regression, k-fold cross-validation, Lasso regression, Decision Tree, etc. were used to predict the house price. The research discovered that linear regression was the most effective model as it achieved a score of 85% accuracy. The study acknowledged the limitation of the dataset used, as it had few features and instances and also recommended for future studies the use of ensemble techniques for a more accurate prediction model.

Tekouabou et al. (2023) in their critical review of 72 modern research papers written on the use of Artificial Intelligence (AI) and ML in predicting real estate price aimed at discovering the most popular ML models used, the potentials and impact of ML within the real estate price prediction domain, and the challenges faced in the use of ML for real estate price prediction. The study discovered all studies used supervised machine learning models, and almost invariably used regression method. Due to limitation in the quantity and standard of data used, the studies reviewed relied heavily on simple ML algorithms instead of unsupervised and deep learning algorithms like convolutional neural network (CNN). The study found out that most of the studies agreed that the use of ML in real estate price prediction can actually give better solutions than traditional means as it solves the challenges around the changing and unstable factors that influences the price of real estate. The study advocates for more data to give better predictions, just as Yadav et al. (2023) had recommended.

Real estate data often have a common problem of outliers which can significantly affect the quality of predictions if not properly addressed (Abut et al. 2023). To this end Abut et al. (2023) carried out a research on the use of hybrid methods for real estate price prediction by making use of outlier detection, clustering methods and feature selection. The purpose was to ensure that only representative samples were fed into the models, and only features that are advantageous to the models were considered. The study discovered that this approach significantly improved the accuracy and performance of the ML models in predicting real estate prices. In fact, the hybrid approach used in the study persistently outmatched other methods, as better Mean Absolute Error (MAE), Root Mean Squared Error (RMSE), Mean Absolute Percent Error (MAPE) values were achieved. The results and findings are also similar to that of Singh et al. (2023) in their study aimed at exploring effective ML methods for real estate price prediction.

The number of features and observations in real estate data are mostly few, and for researchers to circumvent this setback, samples are collected over a long period of time which consequently result in limited and sometimes overlooked heterogeneity (Francke and Minne 2024). Furthermore, models of data generated from a particular location are generally inapplicable to other regions, as real estate price heavily relies on econometric components prevalent in each location. In addressing this, Francke and Minne (2024) did a study on the combination of common ML methods and econometrics in predicting real estate price. Econometric models according to the study effectively caters for unobserved heterogeneity that traditional ML methods fail to consider. The use of the combination of ML and econometric models generated an impressive result, with a MAPE of 9% and a 1.00 correlation between the two models. This result is similar to that of the study by Doszyń (2022) where the best prediction performance was achieved by mixed econometric models.

Unlike studies that exclusively make use of tabular data for real estate price prediction, Yousif et al. (2023) conducted a research on the use of both textual and visual features of a USA real estate data. The study used CNN to extract the attributes of the data, and to predict the real estate price using multi-kernel deep learning regression model. The model achieved MAE of 16.60. In a similar research, Chen et al. (2021) compared the performance and accuracy of linear regression with deep learning algorithms and discovered the deep learning models outperformed the linear regression in Mean Squared Error (MSE), RMSE and MAE values.

Key findings from the reviewed studies on real estate price prediction show that linear regression is the most common ML method used. This is so because of limited usable features and instances of real estate data available. While linear regression is effective and outperforms some other types of regression, it is outshined by deep learning models. However, the use of linear regression alongside other approaches like ensemble methods, outlier detection and feature selection can greatly improve prediction performance. This study aims at adopting some of these approaches for better results.

## 1.4 RESEARCH QUESTIONS

Real estate price prediction using ML methods produces better results than traditional methods. However, misleading predictions and unintended errors from the use of ML can have as much devastating effect as an incorrect traditional method. It is therefore imperative we develop models with high levels of accuracy that can be relied on. This importance is highlighted from the reviewed literature. This study aims to answer the following questions:

1. What regression models perform best in the prediction of real estate price?
2. What impact will ensemble and hyperparameter tuning have?
3. What features have the most impact on prediction accuracy?

# **SECTION B**

## 2.1 DATA ANALYSIS

Descriptive statistical analysis was applied on the dataset to expose trends and correlations between features in the data. Data cleaning is the unseen ‘iceberg’ of any data analysis task and it takes up to 80 percent of the time and efforts spent on analysis (O’Brien and Stone 2021). Some data cleaning and processing were therefore performed as this is necessary before descriptive statistics and models are developed.

*Figure 1* below shows a brief view of the dataset showing the 12 features and datatypes described in the [data description section](#_1.2.2_Data_Description).

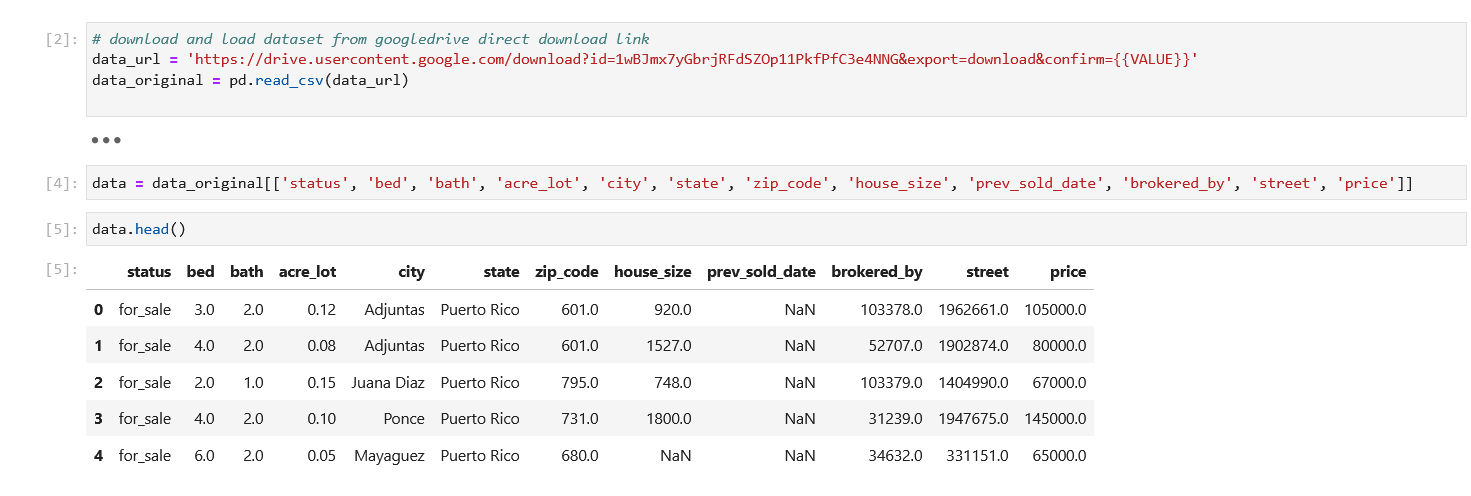


Figure 1: Dataset Preview

### 2.1.1 Data Cleaning and Processing

The first validation was to check for duplicates, as these will cause poor representativeness of data and lead to over-fitted and/or biased outcomes. Fortunately, the data does not contain duplicated rows.

The next check was to confirm if there are missing or null values in the datasets. As *Figure 2* below shows, there are missing values in most of the fields. There is no one-size-fits-all approach in dealing with missing values. We can delete all rows that contain missing values, or replace missing values with a fixed value or mean/mode/median value, or forward-fill with previous values, or ignore the missing values. It depends on the impact the missing value will have on the outcome of the analysis.

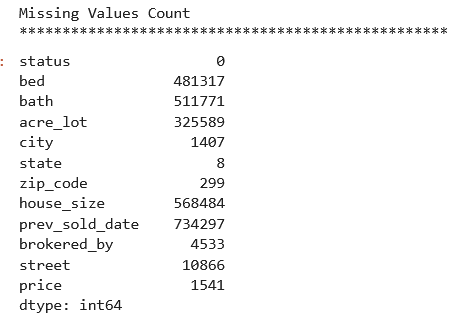


Figure 2: Missing Values in the Dataset

For the purpose of this study, not having missing values in our target variable (‘price’) is important. All rows with missing values in the ‘price’, ‘zip\_code’, ‘state’ and ‘city’ fields were entirely removed. For the ‘prev\_sold\_date’ feature, missing values does not imply the values are missing, but that the houses were not previously sold. So the filled values were converted to ‘Yes’, and missing values were filled with ‘No’. Additionally, a column was created to encode ‘Yes’ values to 1, and ‘No’ values to 0. For other numerical fields, the missing values were replaced with the mode, as using median might not be appropriate if there are outliers in the data.

The next validation was to treat outliers, especially in key features. *Figure 3* below shows a boxplot diagram – with fliers showing the distribution, dispersion and skewness of key features within quartiles. The boxplot diagram shows quite a number of distinctive outliers in the price, bed, bath, acre\_plot and house\_size features outside the whiskers.

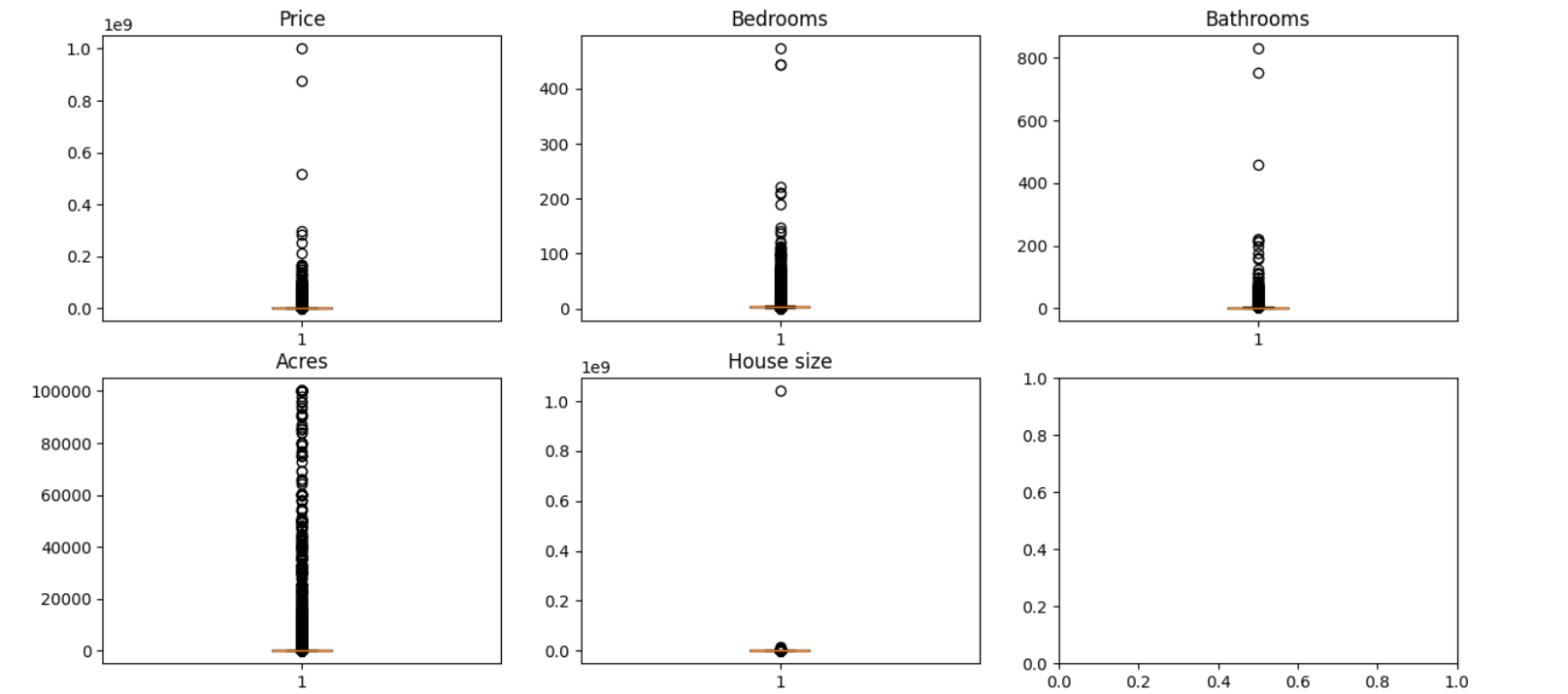


Figure 3: Boxplot Diagram before Removal of Outliers

To avoid false spread of the data which can lead to bias in the data, outliers that are outside 1.5 times the interquartile range below and above the lower and upper quartiles respectively were removed. More specifically for the ‘price’ feature, house prices lower than $100,000 were removed to align data to realistic values. *Figure 4* shows the dispersion of data after the outliers have been removed.

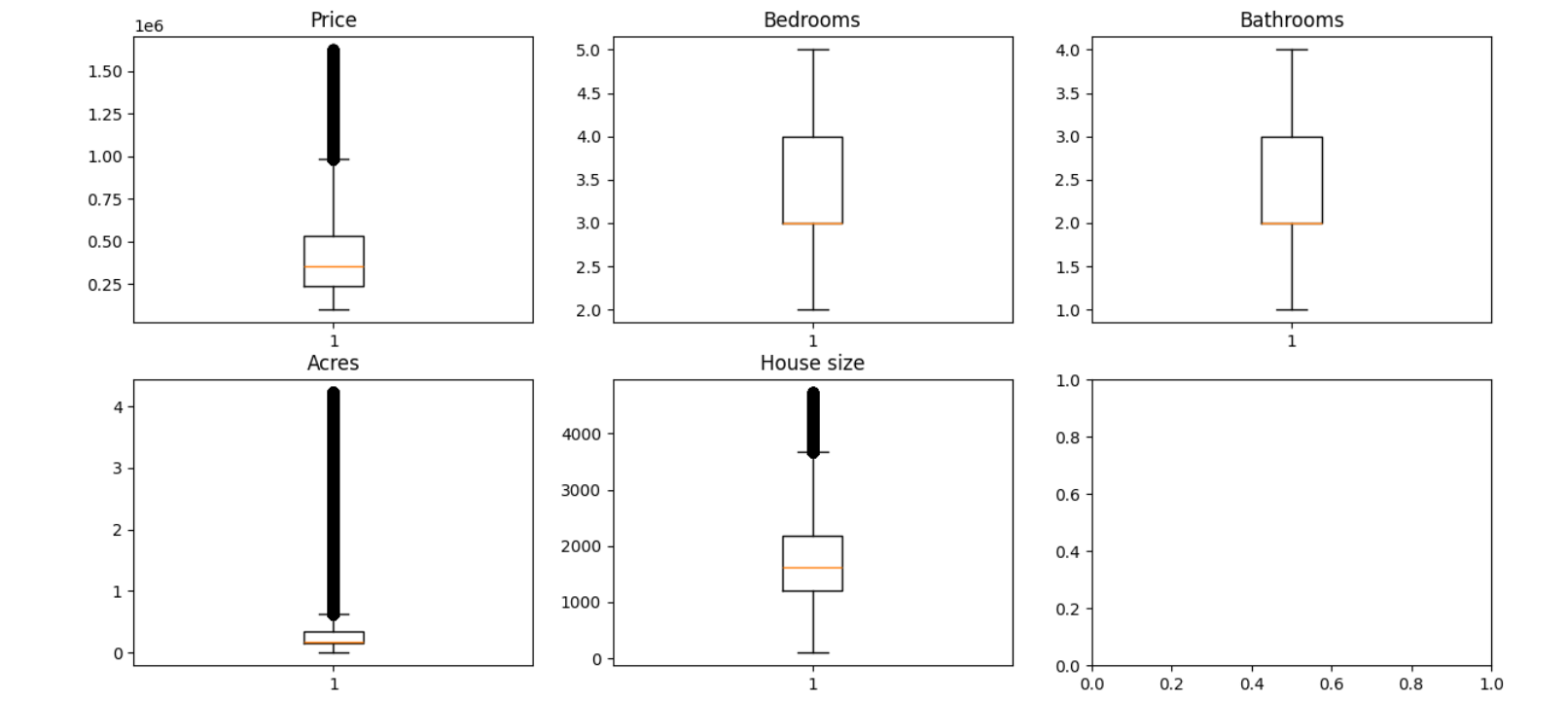


Figure 4: Boxplot Diagram after Removal of Outliers

This study is about real estate price prediction in the USA. All observations in the dataset were checked to ensure only USA states are in the ‘state’ column. Also, the ‘brokered\_by’ and ‘street’ columns were removed from the dataset as they are uninformative, encoded for privacy reasons, and they can cause over-fitting thereby making the model noisy and not being able to generalize to new data. Another processing done in the dataset was the correction of the data type of the ‘zip\_code’ feature. Zip code is a nominal variable and not continuous variable. A zip code of value ‘6000’ is not greater than a zip code of value ‘5321’.

### 2.1.2 Exploratory Descriptive Analysis (EDA)

Having prepared and cleaned the data, we give a brief overview of the structure of the data, and more importantly extract insights out of the data in this section. This helps to see patterns and correlations in the data, and also understand individual features in the data. Necessity for additional cleaning of data other than what had been done can also be revealed with the descriptive statistics.

The data now has 1,492,668 observations and 11 features (*Figure 5*). Also, there are now 5 categorical and 6 numerical variables in the dataset.

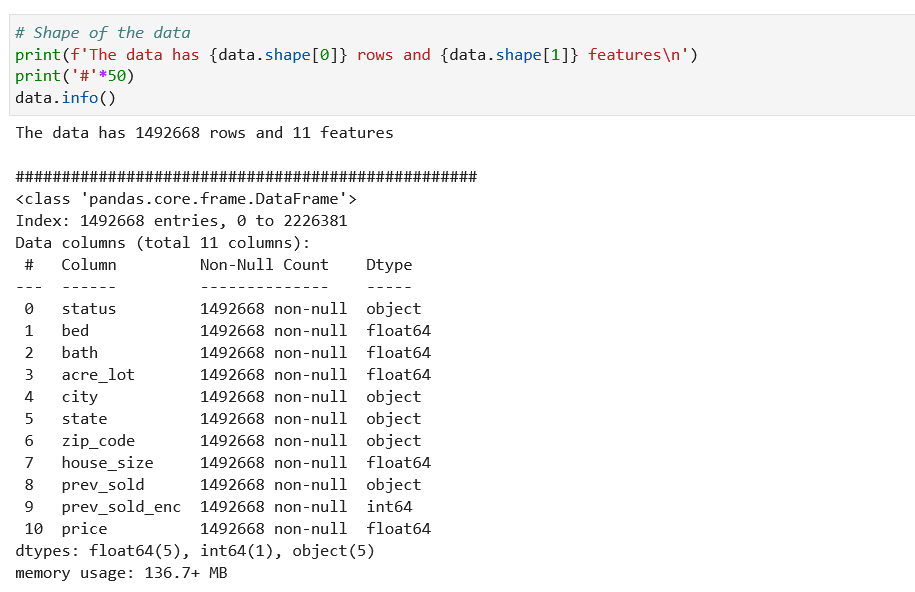


Figure 5: Shape of the data

Describing the data, we can see the summary of both the numerical and categorical attributes (*Figure 6*).

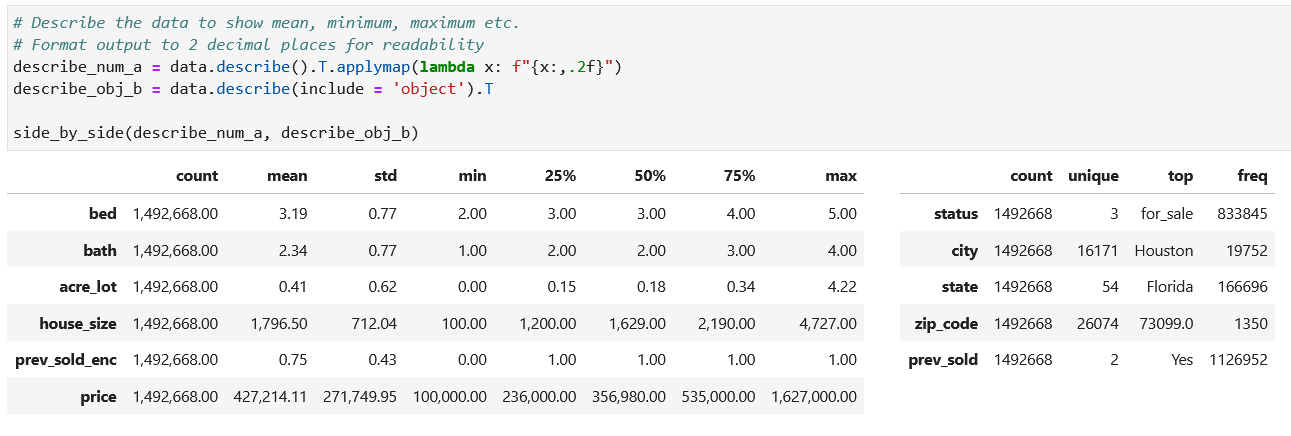


Figure 6: Describing the Data

From *Figure 6*, we see that average bedroom number was 3, and the average number of bathrooms that houses have was 2. Each house had at least 2 bedrooms and 1 bathroom, while the upper quartile of houses had 4 bedrooms and 3 bathrooms. The average price of house was $427,214 with bottom and upper 25% of houses priced at $236,000 and $535,000 respectively. Minimum price of house was $100,000 while the most expensive house was $1,627,000. Most of the houses in the data were previously sold and listed as ‘for\_sale’. The city and state with the highest number of listings was Houston and Florida respectively.

A histogram of all numerical attributes is plotted to show a visual representation of the value range of all instances (*Figure 7*).

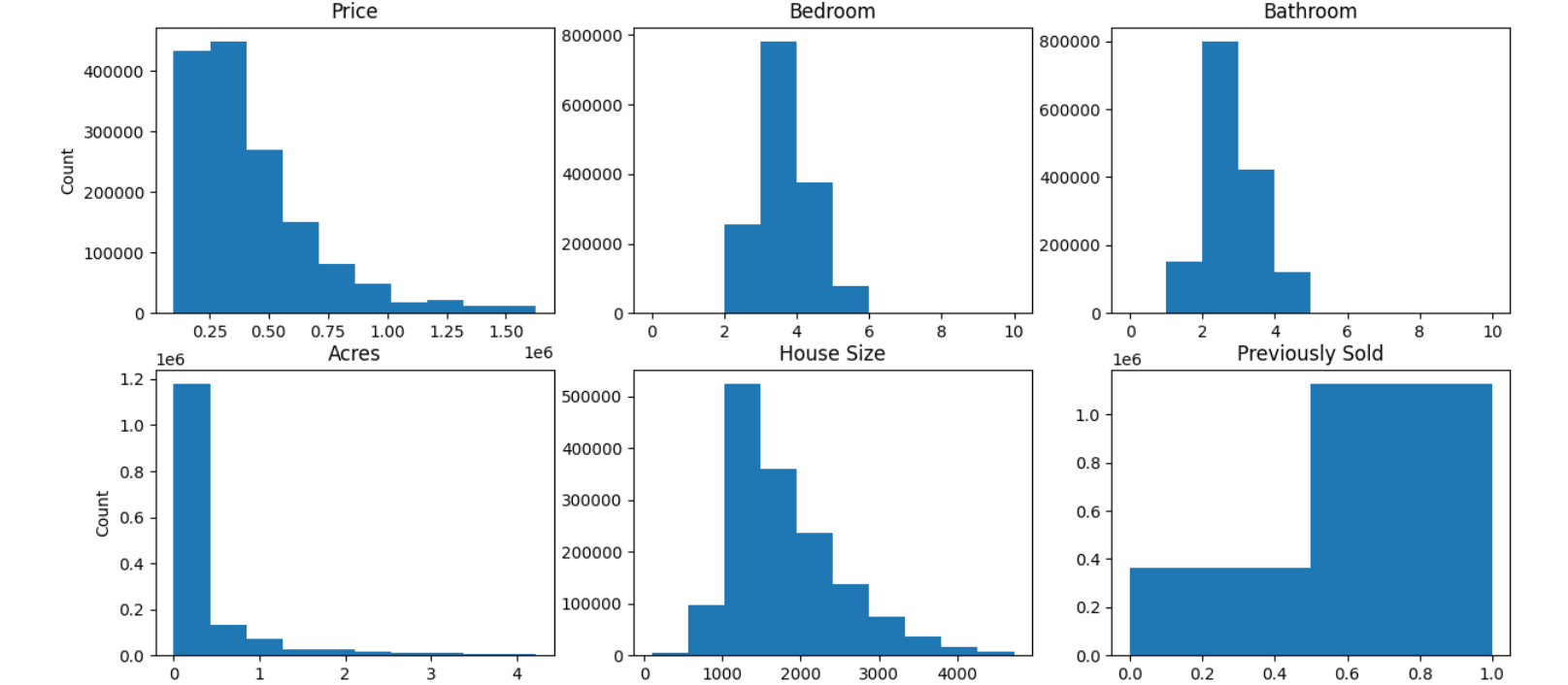


Figure 7: Histogram of Numerical Features

The histogram reveals the price, acres and house size features are positively skewed. This means the median price, land size and house size are lower than the mean and that there are more houses having lower prices, land size and house size. The opposite is for the ‘previously\_sold’ feature which is negatively skewed. The bedroom and bathroom features are to some extent symmetric.

### 2.1.3 Correlation Between Features

To analyse correlations between features, we generate a Pearson correlation coefficient clustermap and ANOVA test between every attribute pair, and also how each attribute correlate with the ‘price’ feature.

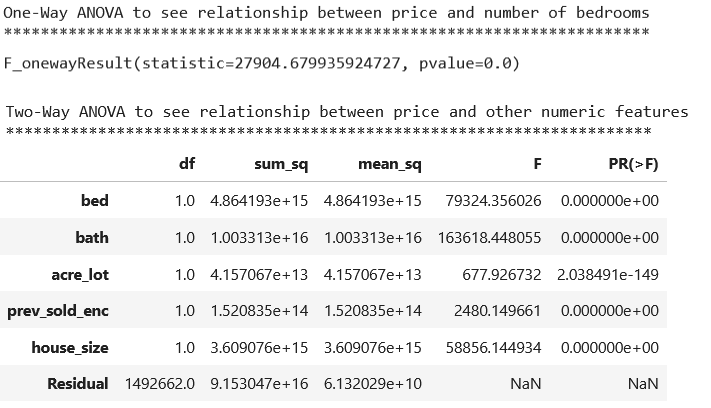


Figure 8: One-Way and Two-Way ANOVA

One-way ANOVA results in *Figure 8* above shows a strong relationship between number of bedrooms and house price (pvalue=0.0). Similarly, the two-way ANOVA results show the independent variables (bed, bath, prev\_sold\_enc and house\_size) have significant effect on the price.

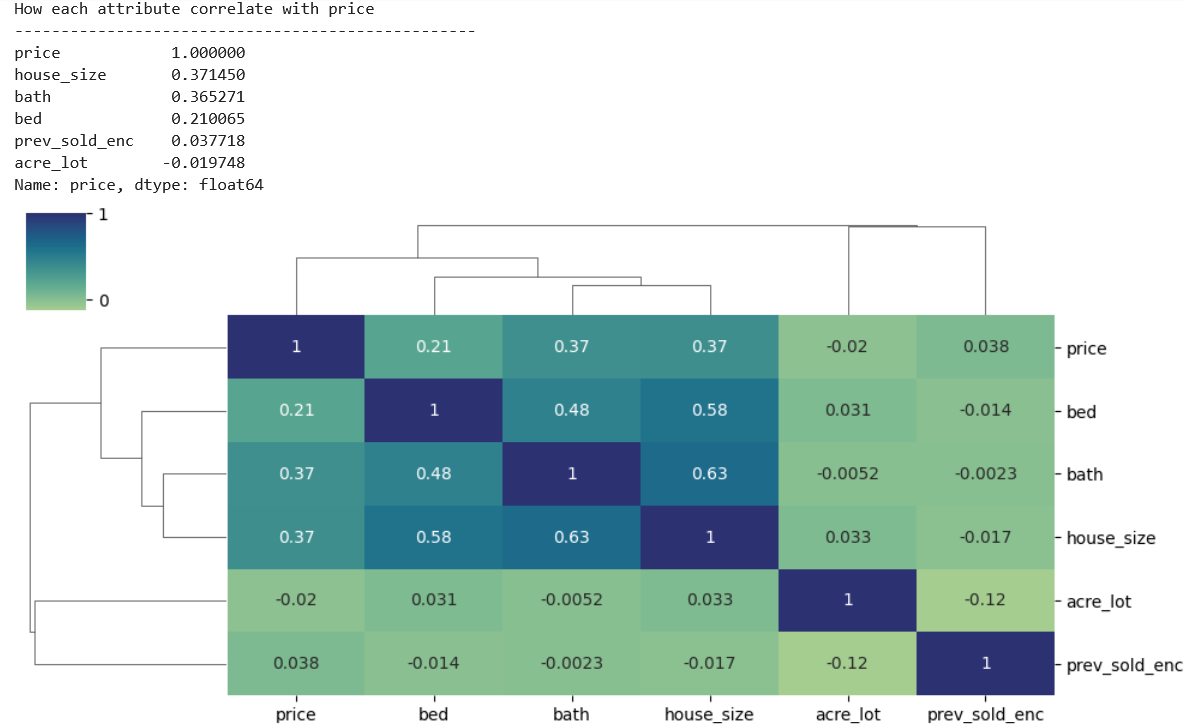


Figure 9: Standard Correlation Coefficient Clustermap

The Pearson correlation map in *Figure 9* above shows that there is a strong positive correlation between number of bathrooms and house size (0.63). The larger the house, the more bathrooms the house will have. Similarly, there’s a strong positive correlation between the house size and the number of bedrooms (0.58), and between the number of bedrooms and the number of bathroom (0.48).

It is observed from the correlation between the price of the house and other features that there is a positive but weak correlation between the price of a house and the house size (0.37), number of bathrooms (0.37) and number of bedrooms (0.21).

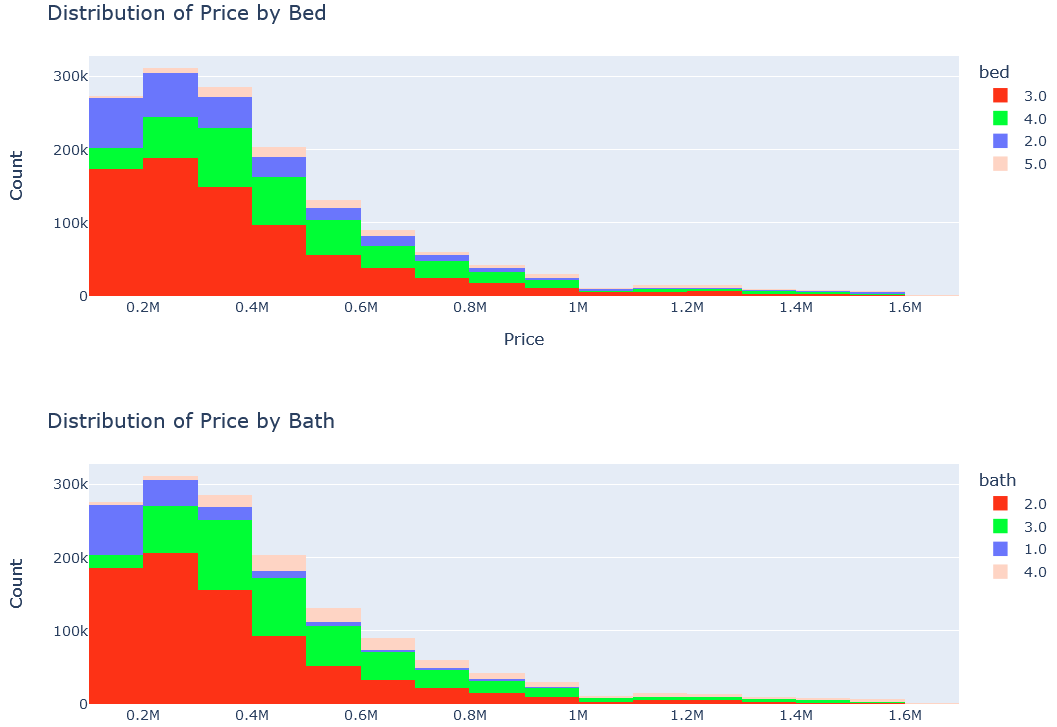


Figure 10: Distribution of House Price by Number of Bedrooms and Bathrooms

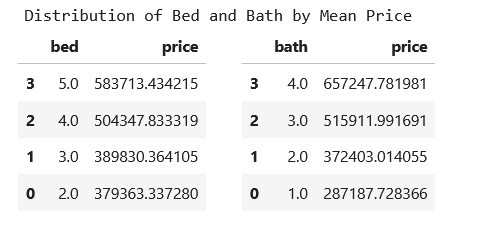


Figure 11: Distribution of Mean House Price by Number of Bedrooms and Bathrooms

*Figures 10 and 11* above shows more intently how the distribution of the house price correlates with the number of bathrooms and number of bedrooms. It shows that the more bathrooms and bedrooms a house has, the higher the price. However, the correlation is weak and this is expected because of the dynamics of real estate. In fact, these dynamics necessitated this study, and in general, the use of ML to predict real estate prices. Location amongst other factors play a role as the price of a 2-bedroom house in a highbrow area might be higher than the price of a 3-bedroom house in a low income area.

Machine Learning models take into consideration these factors and make more accurate prediction than human subjective considerations. This study aims at building ML models that consider all relevant features in the data and their correlations to make accurate predictions.

# **SECTION C**

## 3.1 TOOLS AND MODELS

From the results and recommendations from reviewed literature on real estate price prediction, we see how different models can produce varying levels of accuracy: linear regression yielding better results than other regression models; deep learning and ensemble models’ ability to outperform linear regression; use of hybrid approach to fine-tune regression models, etc. This study in answering stated research questions made use of both supervised and unsupervised ML models, and compared the performance and accuracy across models.

For supervised learning using the ‘price’ feature as label; linear, decision tree, K-Nearest Neighbors (KNN) regression, gradient boosting, Least Absolute Shrinkage and Selection Operator (LASSO), ridge, ridge with cross-validation and elastic net cross-validation regression and ensemble models were used in predicting the price (dependent variable) using the predictor features/independent variables. Regression algorithms’ strength is to understand the relationship between the independent variables and the outcome/dependent variable, and make prediction once relationship has been established. Prediction errors will also be evaluated using performance measures such as MAE, MSE, RMSE and R-Squared.

For unsupervised learning, KNN classification was used to classify house price into price ranges (low, medium, and high price). KNN is predicated on the premise that comparable data points have comparable labels, and then use a selected distance metric to determine the distance between each training sample and the input data point (Srivastava 2019).

The data is split into training and test sets in the ratio 80:20%. Splitting of data helps evaluate the final model and quickly test how well it will generalize to new data.

## 3.2 MACHINE LEARNING MODEL DEPLOYMENT

### 3.2.1 Split data into Training and Test Sets

A test set is created which accounts for 20% of the data (*Figure 12*). Important considerations made when splitting the data to ensure reproducibility:

1. The dataset is randomised and does not contain existing/embedded patterns.
2. The data is not shuffled each time the code is run.
3. Different training and data sets are not regenerated each time the code is run.

The *shuttle* and *random\_state* parameters in the ‘train\_test\_split’ function takes care of these considerations. For the KNN Classification model, the ‘price\_group’ feature was also *stratified* to ensure it is well represented across all classes.



Figure 12: Split of Data into Training and Test Sets

### 3.2.2 Feature Engineering

**Data Preprocessing**

Preprocessing was earlier done in the [data cleaning and processing](#_2.1.1_Data_Cleaning) section before EDA, and some of the reasons for this choice are:

1. Quality insights from EDA is possible when the data is clean.
2. Better understanding of the data to aid choice of model.

**Categorical Features Encoding**

ML algorithms efficiently make use of numerical values and not categorical values. Categorical variables (status, city, state) were therefore encoded so that the ML algorithm can make meaning out of these features. After weighing the capabilities and limitations of different encoding methods, *One-Hot Encoding* was used for ‘status’ and ‘city’ features, while *Target Encoder* was used for the ‘city’ feature because of its high cardinality.

*One-Hot Encoding* is useful for categorical variables that are not ordinal, and have smaller cardinality like the ‘status’ and ‘state’ features. New columns that contains binary numbers for each category is created.

*Target Encoding* is useful for categorical features that have high cardinality like the ‘city’ feature. A new column where the values are replaced by a statistic (mean, mode etc.) of the target variable, ‘price’ is created. It is memory efficient, and as our dataset size is not small, the risk of overfitting is considerably reduced.

These two methods ensure we don’t have feature information loss, no ordinal relationship is produced, memory is optimized, cardinality size is catered for, overfitting is controlled, and no feature leakage.

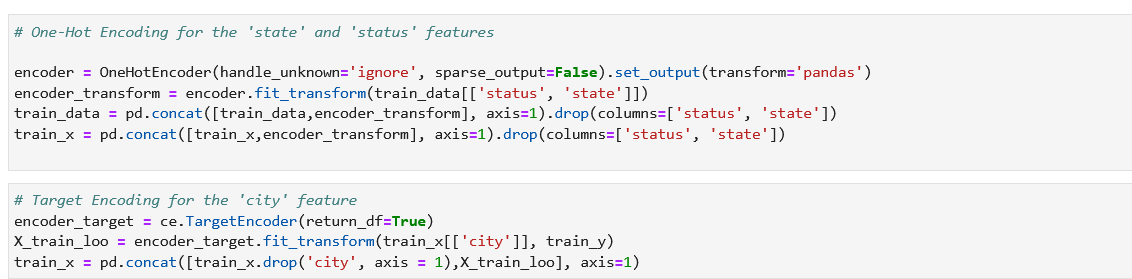


Figure 13: Encoding Categorical Features

**Feature Normalization**

Next step was to normalize the data. Normalization improves ML models by preventing features that have large scales from overly influencing the algorithm learning. For example, the algorithm is prevented from ascribing too much weight to the ‘house\_size’ with value of 2,000 over the ‘bedroom’ size of value 2. For this work, we used Min-Max scaling (Xnormalized = X – Xmin / Xmax – Xmin) and the Z-score normalization (Xstandardized = X-µ / σ) where the features are transformed to a defined range.

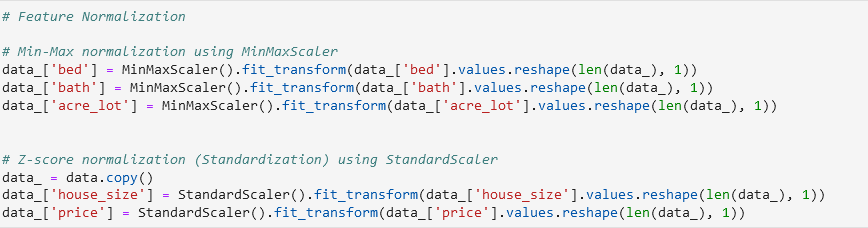


Figure 14: Feature Normalization

### 3.3.3 Unsupervised Machine Learning Model

**K-Nearest Neighbors (KNN) Classification**

Using KNN, the price feature was classified into price groups of low price, medium price and high price (*Figure 15*). Low price is grouped as houses below $500,000, medium below $1,500,000 and high above $1,500,000. Classification prediction accuracy was then tested using different fractions of the test dataset.

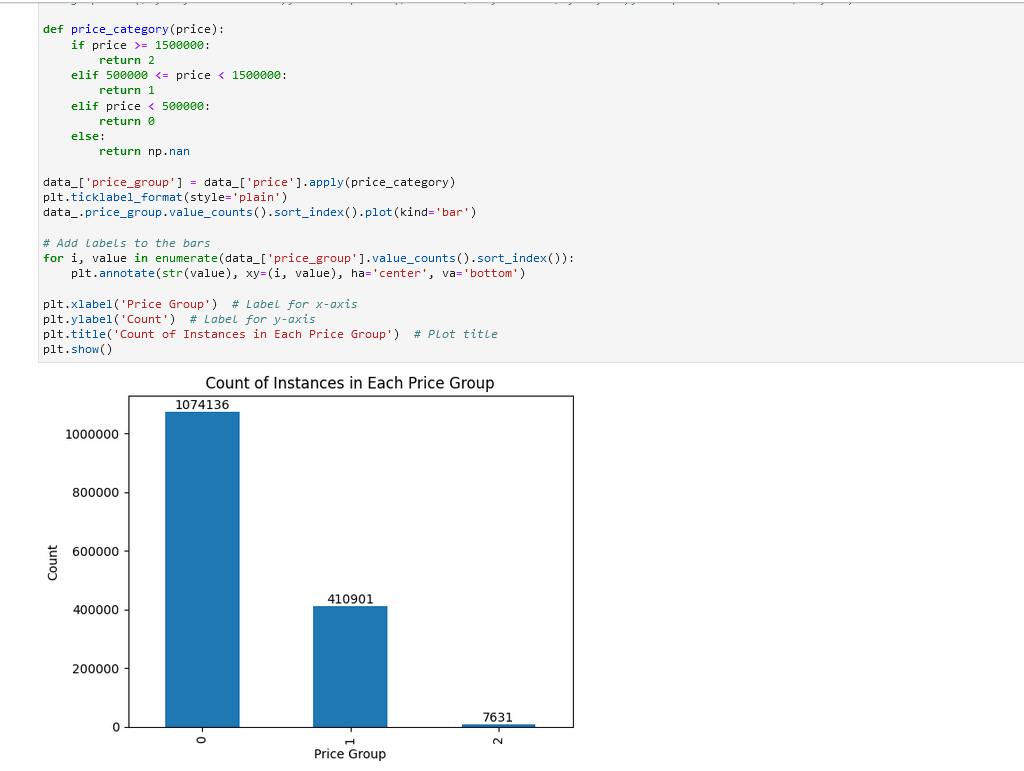


Figure 15: Grouping of Price Feature

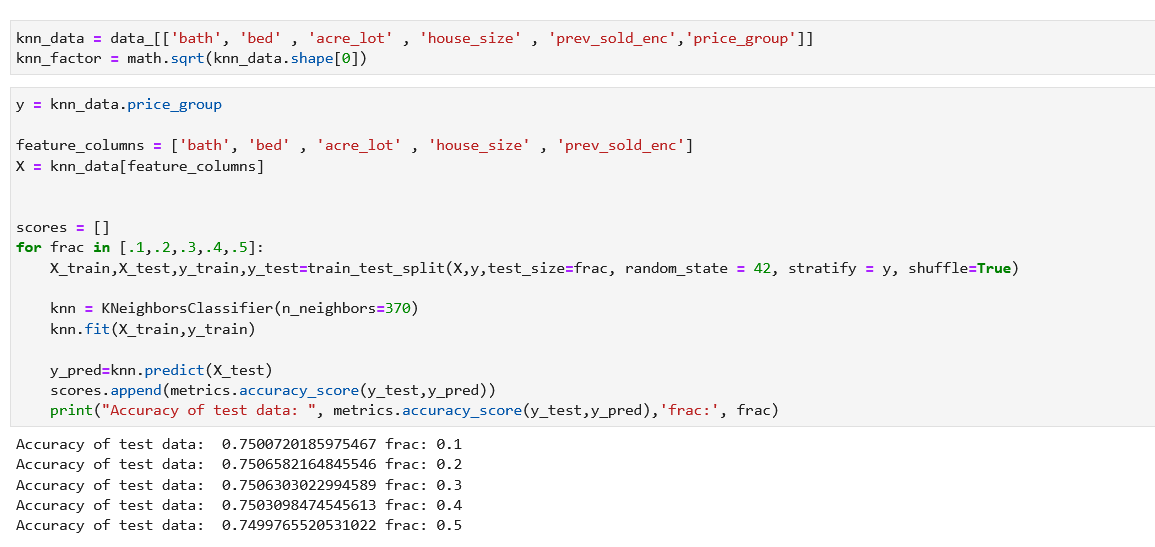


Figure 16: Accuracy of the KNN Classification Prediction

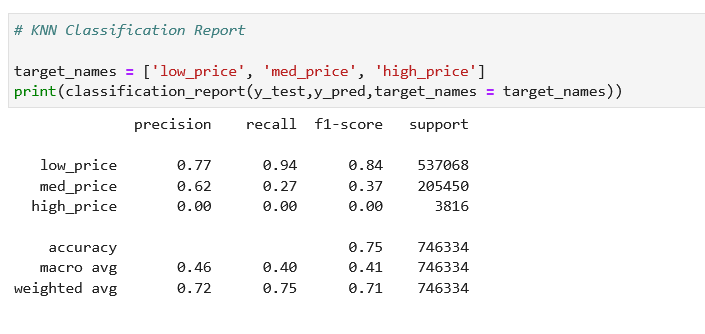


Figure 17: KNN Classification Report

The results (*Figure 16*) show that the accuracy scores of the tested dataset fractions (0.74998 – 0.75007 range) were close with a consistent accuracy of about 75% across different fractions of the test dataset.

In determining which class accounts for the 75% accuracy, the classification report (*Figure 17*) further explains the precision, recall, f1-score, support, accuracy and the macro and weighted average metrics of the prediction across the price classification. In general, prediction for low price performed best as the ratio of correctly predicted observations to positive observations has a precision score of 77%. The medium price has a prediction score of 62% and high price prediction performed worst with 0% prediction. The recall (ratio of correctly predicted to all observations) score for low price was 94% while for medium price and high price was 27% and 0% respectively. The f1-score (mean of the prediction and recall scores) for low price prediction was 84%, while medium price and high price f1-scores were 37% and 0% respectively. In general, the prediction of the low price class performed best amongst the other classes.

Finally, the confusion matrix (*Figure 18*) shows that for the low price class, 67.67% was correctly predicted. 20.20% of medium price class was false negative and failed to predict as medium, while 4.29% of the low class was false positive and incorrectly predicted as medium. This means the medium price class would have performed better.

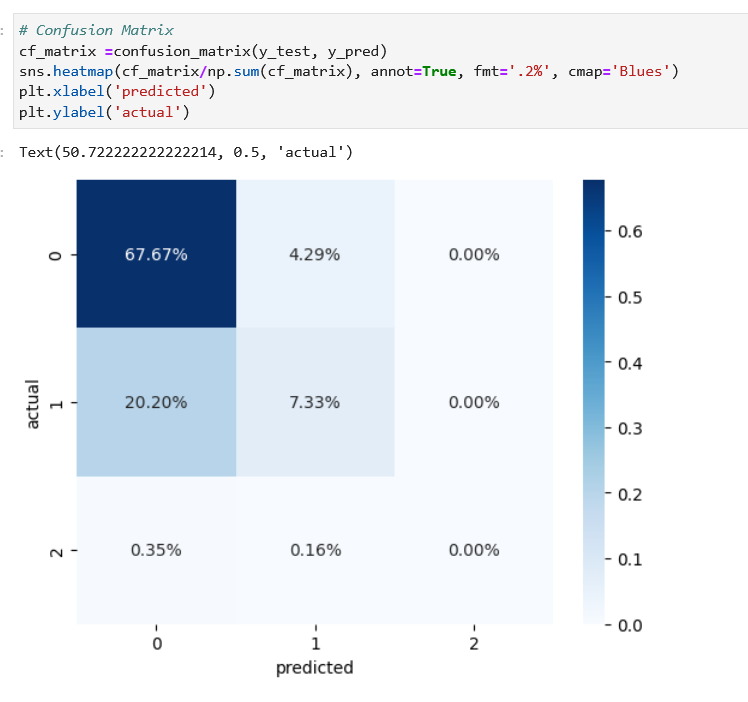


Figure 18: Confusion Matrix (Prediction vs Actual)

### 3.3.4 Supervised Machine Learning Model

For the supervised learning models, eight (8) prediction algorithms were used:

1. Regression Models: Linear, Decision Tree, KNN Regression, LASSO and Ridge
2. Ensemble Model: Gradient Boosting
3. Hyperparameter Tuning/ Cross Validation: Ridge with Cross-Validation and Elastic Net Cross-Validation.

The MAE, MSE, RMSE and R-Squared performance scores were then compared to determine which model performed best.

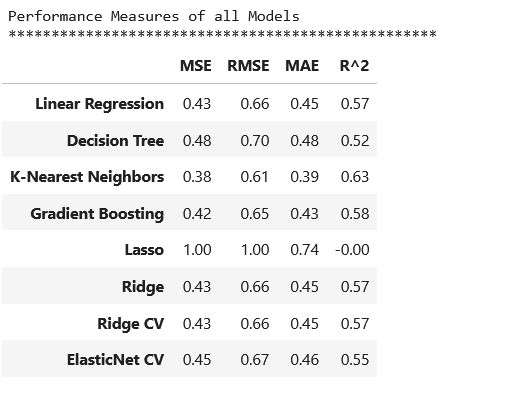


Figure 19: Performance Measures of all Models

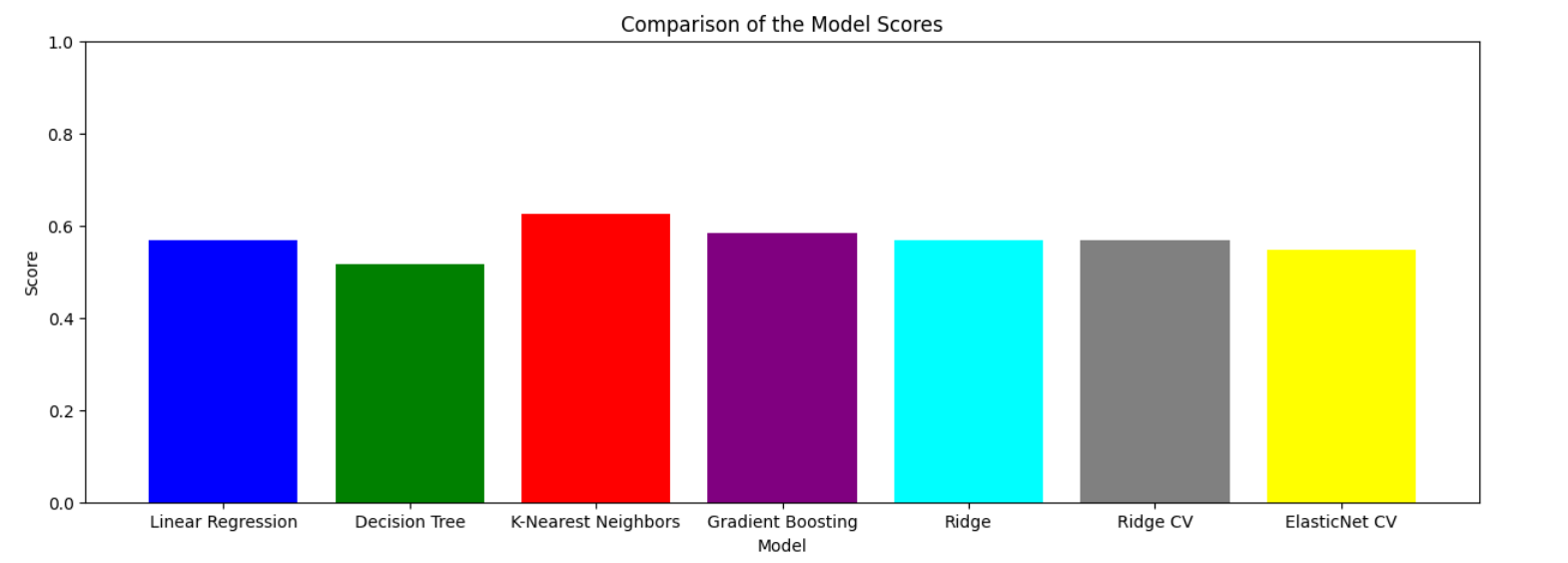


Figure 20: Chart of R-Squared Performance Score of all Models

From the performance scores in *Figures 19 and 20*, we see that the KNN regression model performed best amongst other supervised learning models. KNN had the lowest MSE score of 0.38, which means it had best performance in reducing squared errors. The LASSO model which performs variable regularization and selection to improve prediction accuracy had the worst MSE score of 1. The ensemble model, gradient boosting performed slightly better than KNN with MSE score of 0.42 while decision tree performed worst after LASSO with MSE score of 0.48. The RSME and MAE scores of these models have same performance trend as the MSE, with KNN performing best, LASSO performing worst and other models performing slightly better than the KNN.

The R-Squared (coefficient of determination) measures what variance of the dependent feature is predictable from the independent features. The KNN had the highest R-Squared score of 0.63 which implies that compared to the other models, it clarifies more variance in the dependent feature. The LASSO had the worst R-Squared score of -0.00 which means it could not explain the variance in the dependent feature.

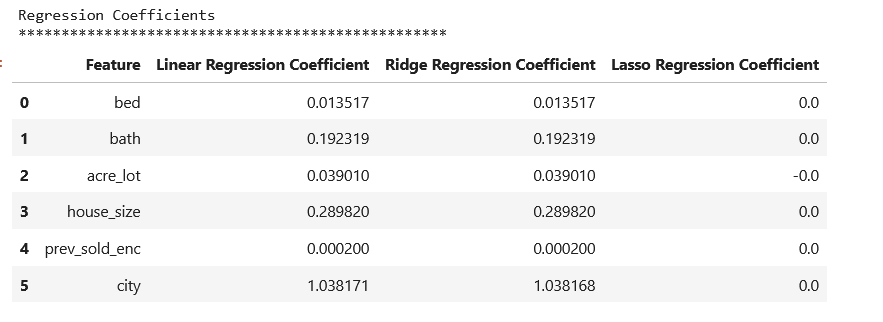


Figure 21: Regression Coefficients

In estimating the impact of each feature on the prediction of real estate price, the linear and ridge regression coefficient in *Figure 21* above shows that the city feature has the most significant impact on price (Coefficient value = 1.038) followed sequentially by house, bath, acre\_lot, bed and prev\_sold features. The city a property is located therefore has the most impact on price.

## 3.3 MACHINE LEARNING RESULTS SUMMARY

Of all the models, KNN Classification performed best, followed by KNN Regression and gradient boosting. This aligns with the study by Tekouabou et al. (2023) that use of simple regression models do not perform as much as unsupervised and deep learning models. Contrary to most results in reviewed literature however, linear regression did not perform better than other regression models; as KNN regression performed best, followed by gradient boosting ensemble model. This study however agrees with Yadav et al. (2023) that ensemble models could in fact improve linear regression. KNN Classification performed best in predicting low price class and poorly in high price. City was the most impacting feature on the price of real estate. The prediction accuracy of the best model (KNN at 75%) is however not as good as expected and from results of reviewed literature.

# **SECTION D**

## 4.1 LEGAL, SOCIAL, ETHICAL AND/OR PROFESSIONAL ISSUES (LSEPI)

The use of real estate price prediction models in actual scenarios poses challenge in terms of stakeholders’ visibility to the interpretation of the inner workings of these models, and how they arrive at their decisions (Fátima et al. 2024). This results in stakeholders’ reluctance in the use of these models because in the event of a wrong prediction, neither the model scientist nor the broker can give explanations of failure cause or improvement points. CNN for instance has been discovered to often produce surprising errors in its classification (Afifi and Brown 2019; Lang et al. 2022), and interestingly, the use of AI by law is only advisory and not responsible for effects of failures (Bartlett 2023). In the event of a financial loss caused by a wrong prediction, who is to be held responsible?

This is an ongoing research across different studies, where researchers are beginning to study the explainability of AI and ML algorithms, and the ‘black-box’ paradox. In fact, explainability in deep learning and complex algorithms is more difficult as not much success has been achieved in explaining causal relationships behind the correlations these algorithms discover (Lang et al. 2022; Longo et al. 2020). Moreover, giving sufficient explanation on how models arrived at their decision is legally binding, as this is needed for a user to give consent to its use. If not, what if models maliciously developed to give some investors undue advantage were used? This is outlined in the European Union General Data Protection Regulation (GDPR) articles 13, 14 and 15 (Information Commissioner's Office 2023).

This study agrees with previous studies that deep learning models can perform better than simple models, but what trade-off is acceptable? Better model or explainable model? Better model or stakeholder’s acceptance? These are some of the legal, ethical and professional issues with the use of ML in real estate price prediction.

## 4.2 CONCLUSION

This study discovered that the KNN Regression and Classification models developed can produce up to 75% accuracy in predicting real estate price. However, this is lower than scores from other studies. This means if this model were to be adapted in real-world, wrong decisions will be made 25% of the time. In a sector that involves huge financial investment, this might not be practical.

As this study found, the city a house is located plays the most impactful role in determining the price. Also, prediction of low priced houses was more accurate than medium or high priced houses.

Limitations this study faced is majorly on the quality of the dataset, and this limitation was faced by most reviewed literature. A substantial number of observations which would have been beneficial in training the model for more accurate results were unusable. However, the combination of data improvement methods really helped improve the results. This include outlier detection, variable encoding, normalization and standardization methods. Encoding of the ‘city’ feature posed some challenge, as it was difficult deciding best encoding technique due to its high cardinality. Relying on the ‘state’ feature is not sufficient as a substitute, as a state can have cities with very different economic levels.

This study agrees with other studies on need of quality data, more relevant features like spatial and unstructured data (maps photographs, economic data etc.), and would recommend future studies to combine different datasets so that features can be added, and a rich dataset can be used.

# REFERENCES

Abut, F., Arlı, H.Ş., Akay, M.F. and Adıgüzel, Y. (2023) A New Hybrid Approach for Real Estate Price Prediction Using Outlier Detection, Feature Selection, and Clustering Techniques. *2023 8th International Conference on Computer Science and Engineering (UBMK)* 23,.

Afifi, M. and Brown, M.E. (2019) What Else Can Fool Deep Learning? Addressing Color Constancy Errors on Deep Neural Network Performance. *IEEE/CVF International Conference on Computer Vision (ICCV), Seoul, Korea (South)* 243–252.

Bartlett, B. (2023) The possibility of AI-induced medical manslaughter: Unexplainable decisions, epistemic vices, and a new dimension of moral luck. *Medical Law International* 23 (3), SAGE Publishing.

Chen, Y., Xue, R. and Zhang, Y. (2021) *House price prediction based on machine learning and deep learning methods*. 699–702https://ieeexplore.ieee.org/document/9587907 Accessed 19 April 2024.

Doszyń, M. (2022) Econometric Models of Real Estate Prices with Prior Information. Mixed Estimation. *Real Estate Management and Valuation* 30 (3), 61–72.

Fátima, T.N., Manuela, A. and Miguel, de C.N. (2024) The Impacts of Open Data and eXplainable AI on Real Estate Price Predictions in Smart Cities. *Applied Sciences* 14 (5), 2209.

Francke, M. and Minne, A. van de (2024) Combining machine learning and econometrics: Application to commercial real estate prices. *Real estate economics* Wiley-Blackwell.

Information Commissioner's Office (2023) *Legal framework*. https://ico.org.uk/for-organisations/uk-gdpr-guidance-and-resources/artificial-intelligence/explaining-decisions-made-with-artificial-intelligence/part-1-the-basics-of-explaining-ai/legal-framework/ Accessed 4 March 2024.

Lang, M., Bernier, A. and Knoppers, B.M. (2022) Artificial Intelligence in Cardiovascular Imaging: ‘Unexplainable’ Legal and Ethical Challenges?. *Canadian Journal of Cardiology* 38 (2), 225–233.

Longo, L., Goebel, R., Lecue, F., Kieseberg, P. and Holzinger, A. (2020) Explainable Artificial Intelligence: Concepts, Applications, Research Challenges and Visions. *Lecture Notes in Computer Science* 12279, 1–16.

O’Brien, A. and Stone, D. (2021) A Case Study in Managing the Analytics ‘Iceberg’: Data Cleaning and Management using Alteryx. *Journal of Emerging Technologies in Accounting* 18 (2), 221–245.

Sakib, A.S. (2024) *USA Real Estate Dataset*. https://www.kaggle.com/datasets/ahmedshahriarsakib/usa-real-estate-dataset/data Accessed 18 April 2024.

Singh, B., Singh, J., Kumar, A. and Gupta, S. (2023) Investigating the Effectiveness of Machine Learning Techniques for Real Estate Price Prediction. *2023 4th International Conference on Smart Electronics and Communication (ICOSEC)* .

Srivastava, T. (2019) *Introduction to KNN, K-Nearest Neighbors : Simplified*. https://www.analyticsvidhya.com/blog/2018/03/introduction-k-neighbours-algorithm-clustering/ Accessed 21 April 2024.

Tekouabou, S.C.K., Gherghina, Ş.C., Kameni, E.D., Filali, Y. and Gartoumi, K.I. (2023) AI-Based on Machine Learning Methods for Urban Real Estate Prediction: A Systematic Survey. *Archives of Computational Methods in Engineering* 31, Springer Science+Business Media.

Yadav, S.K., Dhanda, N., Sahai, A., Verma, R.U. and Pandey, S. (2023) Real Estate Price Prediction Using Machine Learning. *Lecture notes in electrical engineering* 1046, 95–111.

Yousif, A., Baraheem, S., Vaddi, S.S., Patel, V.S., Shen, J. and Nguyen, T.V. (2023) Real estate pricing prediction via textual and visual features. *Machine vision and applications* 34 (6), Springer Science+Business Media.