EasyChair Preprint

№ 12512

Comparative Analysis of Machine Learning Algorithms for Predicting House Prices

Sachith Yamannage

March 15, 2024

Comparative Analysis of Machine Learning Algorithms for Predicting House Prices

Sachith Nimesh Yamannage

https://orcid.org/0009-0002-9320-1890

Email: sachithnimesh999@gmail.com

Faculty of Science, University of Ruhuna, Matara, Sri Lanka

Abstract

This study conducted a thorough examination of residential property data in the Abbotsford area of M elbourne with the goal of identifying significant trends in housing, regional patterns, and variables aff ecting home values.The dataset contained a number of elements, such as neighborhood information, g eographic coordinates, transaction details, and property attributes.Using a variety of techniques, inclu ding mean imputation, forward filled imputation, machine learning algorithms, and discarding missin g data, the study started with the identification and treatment of missing values.Particularly in the pric e variable, outliers were found, and boxplots and other visualization tools were used for outlier analysi s.For additional analysis, numerical values representing the categorical variables were converted.To in vestigate the distributions of numerical variables and comprehend connections between variables with a focus on correlations with home prices—univariate and bivariate analyses were carried out. Feature engineering, covariance analysis, ANOVA testing, and predictive modeling with regression algorithms like Random Forest, XGBoost, and Support Vector Machine (SVM) were all part of the quantitative analysis process. Metrics like Mean Absolute Error (MAE) were used to assess the performance of the model; the results showed that XGBoost was the most accurate predictor of housing prices. Significant factors influencing home prices were identified by the study, such as building area, property type, number of rooms, and geographic considerations including proximity to important sites. Each component was analyzed in terms of its relative relevance, and the building area and land size. It was noted how the constructed model has limits, such as overfitting and the need for more model refining. The results offer insightful information to scholars, politicians, and real estate professionals who are interested in the dynamics of the housing market.

Keywords: Housing market analysis, Predictive modeling, Melbourne real estate, XGBoost regression, House price prediction

1.Introduction

Background of the study

Nowadays we all know that; the estimated property price of a country is a major factor that use to determine the development of a country. In history, people use manual methods to estimate the property price. But the thing is those manual methods occurs less accuracy. Hence this is a technological era, human use more precious and more productive methods to estimate property price, especially house price.

House price is always not a constant value. There are many factors that can affects to the house price. The location where the house is situated, the facilities at the location such as education availability, health and medical facilities, transport facilities, grocery and shopping facilities, weather of the city, economic and technological conditions of the city are so on. When consider the house, the number of bedrooms, bathrooms, parking area, number of floors and the other comfortable human friendly conditions can affect to the price of a house. Also, the sellers expected profit amount can vary the house price. The main external condition for the house price is government taxes and policies.

By considering the above-mentioned factors in physical attributes, location, and economic factors, we can create a model to predict the future house price using past data. Also, when we consider a situation, a person who do not know the expected value of a house price, he may be face to a trouble when he is finding money or collecting money to buy house. So, in this project we wish to predict the house price using machine learning algorithms.

Machine learning is a useful mechanism that people use nowadays to do their works more one easily. It is based on AI (Artificial Intelligence) which is the important growth of data science. The most people agreed and using definition for machine learning is, the American scientist called Tom Mitchell in 1997 who did a research and more exploratory studies about machine learning, said that the machine learning is the "A computer program is said to learn from experience E with respect to some class of tasks T and performance measure P, if its performance at tasks in T, as measured by P, improves with experience E". We can use any programming language in machine learning; but the most used programming language in machine learning is python. This machine learning algorithms are amazing and clever method to overcome the tasks in computer science field. When we give the data and answers to the machine learning algorithm, it can experience the data and perform a suitable model to predict a new data point.

There are different algorithms (classifications) in Machine Learning process. We can use any of them to predict the house price. Some of them are, Leaner Regression, Decision Tree Regression, K- Means Regression and Random Forest Regression. This machine learning (MI) procedure applies various techniques such as features, labels, reduction techniques and transformation techniques such as attribute combinations, set missing attributes subsequently, looking for new correlations. So, we expect to use machine learning system to do our research for find out most accuracy model which predict better results.

Review of predicting house price using machine learning algorithms.

A scientist called P. Durganjali introduced a model for house price prediction using classification algorithms. she used different classifications algorithms like Leaner Decision Tree, K Means, Linear Regression and Random Forest in that paper to predict house price. (Durganjali, 2019)

Also, a scientist called Sifei Lu, introduced a hybrid regression mechanism to prediction of house price by using limited dataset and features. She examined creative feature engineering technique in that paper. That introduced model has been deployed as the key kernel for Kaggle Challenge "House Price: Advance Regression Techniques." The main purpose of the paper is, predict suitable, reliable, and reasonable price for customers matching to their budgets with priorities. They used Linear Regression, Decision Tree, and Random Forest methods for build a productive and successful prediction model. Also, they used step wise method in Data Collection, Pre-Processing Data, Data Analysis parts for Build the model. Then finalized all model and result into '.txt.' Then find out that the Random forests give a best result when compare with the other methods. After that two finally, found that the Random Forest had the best accuracy in eighty-seven.

In 2019, using 1,970 housing transaction data, a researcher call Koktashev do research on the estimate the house prices in the Krasnoyarsk city. He considers the number of rooms, total area, floor, parking, type of repair, number of balconies, type of bathroom, number of elevators, garbage disposal, year of construction and the accident rate of the house as independent factors of the study. He uses random forest, ridge regression and linear regression to predict house prices. Their study gives that random forest outperforms the other two algorithms, as evaluated by the mean absolute error. (Koktashev, 2019).

The scientist called, Park and Bae developed a house price estimating model with machine learning algorithms, as demonstrated by C4.5, RIPPER, Na¨ıve Bayesian, and AdaBoost. They compare their performance considering the classification precision. Their study aims to help house sellers to make rational decisions in house transactions. The research shows that the RIPPER algorithm, based on precision, consistently outperforms the other models in the productiveness of housing price prediction. (Park, 2015)

Aims

Consider a person in the society. When he/she is looking for a house to living he/she absolutely need a pre-understand about the house price in that location area. If there is a pre-structured model for predict the house price, it will be useful to him/her. On other hand it will also be useful for the house seller. He can earn more profit and sell their property in a productive way. They can get a clear and useful understand about the house price in the market.

Also, we all know that house price is an important indicator that determines the development of a country. So, our aims are also for give a productive model to predict the house price in a country and help government to make decisions on their country.

Objectives

1. What are the key factors that can determine the house price and, in what quarters that can be effect to the house price? And what is the relative importance of each key factor in predicting the house pricing?

2. What is the most suitable and more productive machine learning algorithm to predict house price?

3. Is the developed algorithm dependable in predicting the house price? What are the limitations?

3. Methodology:

Research Approach

We collected dataset through Kaggle website. Our dataset contains the necessary information to investigate Predicting House Price. The proposed overall research approach is the quantitative research approach.

• Our dataset contains more than 6500 data on House Price. And the house prices range from 131000 to 9000000.

• Some of the information included Rooms, Type, Method, Distance, Bathroom, Car, Land size, building area, Year built, council area, Region name, Latitude, Longitude and Property count.

• Further, in our dataset Type, Method, Region Name and Property count are categorical variables. Therefore, we changed them as quantitative variables. s?

Justification

The following four research topics are discussed in this case study. In considering each research question, the following justification for the claim might be given.

1. What are the key factors that can determine the house price and, in what quarters that can be effect to the house price?
   The process of determining prices for houses is complex and influenced by a number 5 of variables. The real estate market is dynamic, and a number of social, cultural, and geographical factors can have an impact on prices. The following are some significant variables that can affect the price of a house, divided into several quarters: Economic Factors, Location, Supply and Demand, Market Conditions, and Global and National Economic Factors and Government policies.
2. What is the relative importance of each key factor in predicting the house pricing?
   The precise location, the state of the economy, and the dynamics of the market can all have an impact on how important certain elements are in determining the price of a house. Assigning a universal weight to each component is difficult because its importance varies over time and between different geographical areas. But I can give you an overall list of some of the factors that are frequently thought to have more influence over the housing market: Location, Economic Factors, Interest rates and Demographic trends.
3. What is the most suitable and more productive machine learning algorithm to predict house price?
   The size of the dataset, the type of data, the existence of non-linear correlations, and the objectives of the prediction task all play a role in selecting the most effective and efficient machine learning method for predicting house values.
4. Is the developed algorithm dependable in predicting the house price?
   What is the limitations A developed algorithm's predictability of housing costs is assessed using a range of measures, and its limitations are considered. Some things to think about are as follows: Model Evaluation Metrics, Cross-Validation, Interpretability and Every machine learning model has limitations. (The following are some typical limitations on house price prediction: Limited data, assumptions, Changing market conditions and Outliers)

Conceptual model

Here is the general conceptual model representing broader categories of predictors under study. The relationships between Price and the key predictors selected from each broader category will be explored during the case study. Cause in our data set there is twenty-one variables, we divided it into four main factors for ease of our creating the conceptual model. Below we mentioned those four main factors and what are the variables that belongs to it. 6

• Economic Factor: Price, Property Count
• Location Factor: Suburb, Address, Distance, Postcode, Council Area, Region Name
• Market Demand Factor: Method, Seller G, Date, Year Built
• Conditions Factor: Rooms, Type, Bedroom 2, Bathroom, Car, Land Size, Building Area, Latitude, Longitude

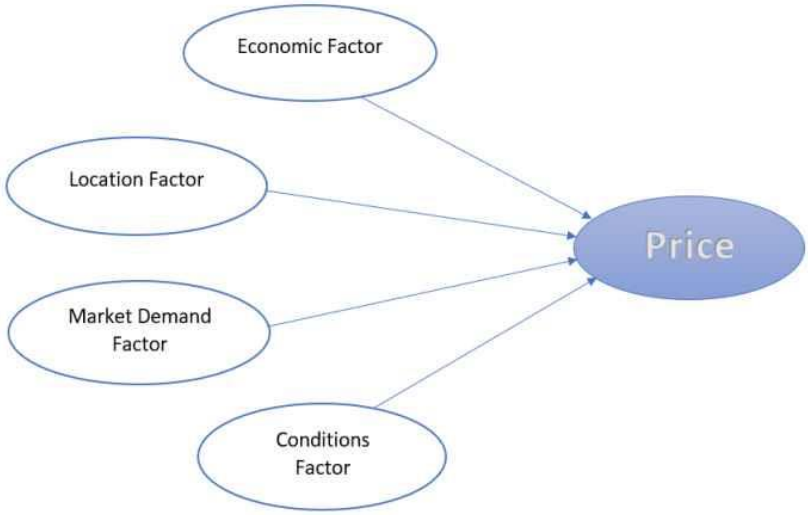


Figure 2.1: General Conceptual Model

Research Design

A branch of artificial intelligence called machine learning (ML) focuses on creating statistical models and algorithms that let computers conduct tasks without the need for explicit programming. The primary goal of ML is to enable computers to learn and improve at a certain task as they are exposed to more data.

Programming has always entailed writing comprehensive instructions that a machine can follow. In Machine Learning instead of being deliberately planned, a system learns from data. The algorithm must first identify patterns, correlations, and trends in the data to produce predictions or choices without having to be specially written for the task at hand.

ML is used in many different fields, such as recommendation systems, fraud detection, natural language processing, picture and speech recognition, and driverless cars, to name just a few. Supervised learning and unsupervised learning are the two primary categories into which machine learning algorithms can be divided.

Machine Learning Algorithms

As we mentioned in introduction part machine learning is a powerful and amazing use of AI (Artificial Intelligence). We can write usual programming codes using any computer language such as R, Python, C+, C++, Java etc. Then we must put our programming code to machine learning algorithm and can get the output by run the code as usually. There are several machine learning algorithms to do this procedure. The classification of machine learning algorithms is visualized below figure. We can use any of those algorithms to do our research.

Note that in our research, we use Python as our programming language, and Random Forest, BGBoost and SVM as the machine learning algorithms.

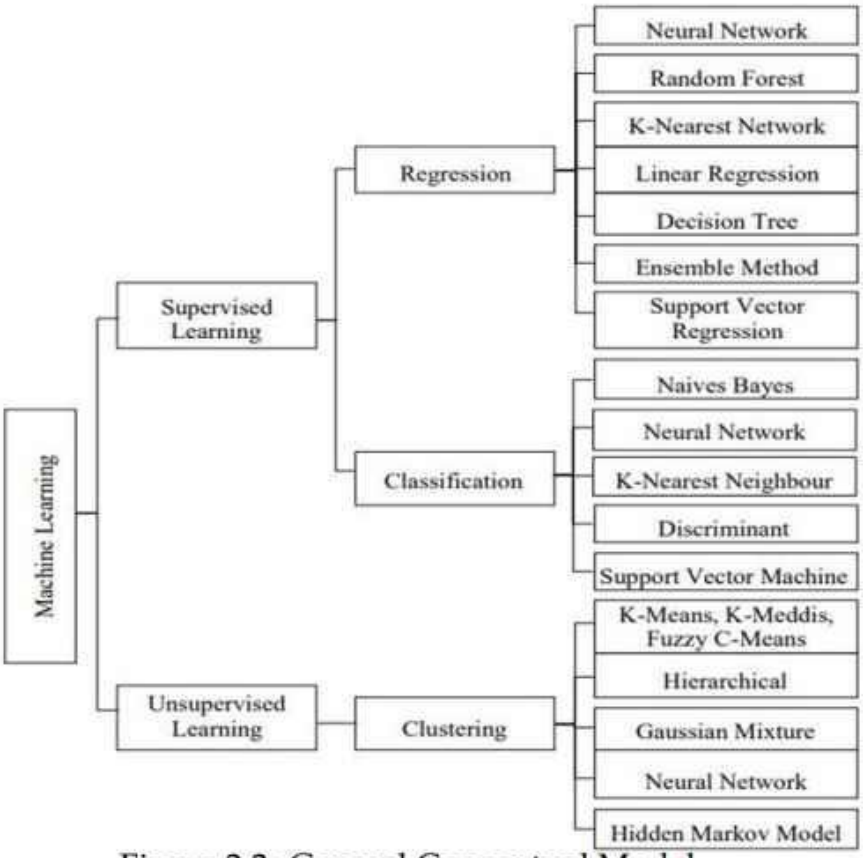


Figure 2.2: General Conceptual Model

Introduction to XGBoost

XGBoost, or eXtreme Gradient Boosting, is a prevailing and widely used machine learning algorithm that excels at predictive modeling tasks. Tianqi Chen developed XGBoost, which has grown up in popularity due to its ability to carry high performance across a wide range of applications, with classification, regression, and ranking.

The algorithm falls under the group of ensemble learning, specifically boosting, in which multiple weak learners (typically decision trees) are joint to create a robust and accurate predictive model. XGBoost expands on outdated gradient boosting by incorporating regularization methods, managing missing values, and using a scalable and parallelizable execution.

Mathematical Foundations

The central awareness behind XGBoost is to iteratively construct a series of weak learners before joining them into a strong learner. The goal is to minimize a loss function, and each weak learner is projected to correct the errors in the present ensemble. Here are the key components, articulated mathematically:

1. **Objective Function**:
   The overall objective function in XGBoost is a sum of a training loss term and regularization terms:

$$\text{Obj}(\theta) = \sum_{i=1}^{n} \text{loss}(y_i, \hat{y}_i) + \sum_{k=1}^{K} \Omega(f_k)$$

where θ represents the model parameters, yˆi is the predicted value for the i-th observation, yi is the true label, K is the number of weak learners, Ω(fk) is the regularization term for the k-th learner, and loss is the specified loss function.

Weak Learner (Tree) Contribution: - The contribution of each weak learner to the overall prediction is given by the sum of its output and a regularization term:

Outputk = X T j=1 wj,k · I(x ∈ Rj,k) + γT

where T is the number of terminal nodes in the tree, wj,k is the weight associated with the j-th leaf in the k-th tree, I(x ∈ Rj,k) is an indicator function denoting whether observation x falls into the j-th leaf of the k-th tree, and γ is the regularization parameter controlling the complexity of the tree. These mathematical formulations prompt how XGBoost optimizes its objective function and combines weak novices to create a powerful predictive model. The algorithm's feat stems from its ability to balance predictive accuracy and regularization, making it flexible and suitable for a wide range of machine learning errands.

Introduction to Random Forest:

Random Forest is a multipurpose and robust ensemble learning algorithm that is usually used in machine learning for classification, regression, and feature selection errands. Random Forest, developed by Leo Breiman and Adele Cutler, is famous for producing accurate and stable predictions transversely a wide variety of datasets. Random Forest paradigms a large number of decision trees during training and proceeds the mode of the classes (classification) or the mean forecast (regression) of the individual trees. The name" Random Forest" comes from the usage of randomness in together the data and feature selection processes.

Mathematical Foundations:

The algorithm is based on the combination of predictions from a collection of decision trees. Let us express key impressions mathematically:

1. Decision Tree Ensemble: - The prediction of the Random Forest ensemble for a given input X is obtained by aggregating predictions from N individual decision trees:

   Yˆ RF(X) = 1 N X N i=1 Yˆ i(X) where Yˆ i(X) is the prediction of the i-th decision tree.

2. Bootstrap Sampling: -
   Each tree in the Random Forest is accomplished on a bootstrap sample, which is a random sample with auxiliary from the original dataset. The probability of an observation being involved in a bootstrap sample is 1 − 1 N, where N is the size of the dataset.
3. Feature Randomness: -
   Each split of a verdict tree considers a random subset of structures. This increases randomness and prevents specific structures from dominating the ensemble. Random Forest has some advantages, including high accuracy, confrontation to overfitting, and the ability to handle large, multidimensional datasets. The combination of several trees and randomness in the training process improves the algorithm's robustness and generalizability. 11 Finally, Random Forest is an influential ensemble learning technique that uses the wisdom of crowds to be combined predictions from various decision trees. Its

adaptability, scalability, and performance make it a widespread choice for a variety of machine learning applications.

An Introduction to Support Vector Machines (SVMs)

Support Vector Machines (SVM) is a powerful machine learning algorithm that is usually used for classification and regression tasks. SVM was announced by Vladimir Vapnik and his colleagues in the 1960s and has established to be effective in various domains due to its skill to handle both linear and non-linear affairs in data.

Mathematical foundations:

The vital idea behind SVM is to find the ideal hyperplane that separates different classes in the eye space. The algorithm is particularly operative in high-dimensional spaces and is famous for its durability and versatility. In a binary classification setting, given a set of training data with labels yi ∈ {−1, 1} and corresponding feature vectors xi, SVM aims to find a hyperplane defined by:

$$w \cdot x + b = 0$$

Here, w is the weight vector, x is the input vector, and b is the bias term. The distance from any data point to this hyperplane is expressed as:

$$\text{Distance}(x, w, b) = |w \cdot x + b| \, \|w\|$$

The optimal hyperplane maximizes the boundary, which is defined as the distance among the hyperplane and the next-door data point in either class. Mathematically, this profits the formulation of the SVM optimization tricky:

$$\text{Minimize } 1\ 2\ \|w\|\ 2 \text{ subject to } yi\ (w \cdot xi + b) \geq 1.$$

In cases where the data is not linearly divisible, SVM introduces the concept of a" soft margin" by letting some data points to be misclassified. This leads to the formulation of a soft-margin SVM:

The Kernel Trick: SVM can be prolonged to handle nonlinear relationships with the kernel trick. The goal is to indirectly map the input data to a higher-dimensional space wherever a hyperplane can efficiently separate the classes. Mutual kernel functions contain linear, polynomial, and radial basis function (RBF) kernels. Support Vector machines are valuable tools in machine learning because they afford a solid mathematical foundation for cataloguing and regression tasks. Their capability to handle both linear and nonlinear relations, combined with the perception of maximizing margins, makes them useful in a wide variety of real-world applications.

Data Cleaning Process

• Dataset:

The data set used for our study has been taken from the Kaggle platform. It is a snapshot of dataset created by "Tony Pino." It was taken from publicly accessible Domain.com.au data that are published on a weekly basis. Address, Real Estate Type, Suburb, Selling Method, Rooms, Price, Real Estate Agent, Sale Date, and Distance from C.B.D. are all included in the dataset. In this dataset we have 18,397 data with one dependent and twenty independent variables.

• Data Dictionary:

Variable Name	Variable Description	Variable Type	Measurement Units
X	Count	Numerical	No Unit
Suburb	Suburbs in Melbourne (Like Abbotsford, Airport West, Albert Park, etc.)	Categorical	No Unit
Address	Address of the houses in Melbourne.	Categorical	No Unit
Rooms	Number of rooms	Numerical	No. of Rooms
Type	House Types (Like house cottage villa, semi terrace, etc.)	Categorical	No Unit
Price	Price in dollars	Numerical	Dollars
Method	Status of the House sales (Like sold, Not sold, No bid, etc.)	Categorical	No Unit
Seller G	Real estate agent	Categorical	No Unit
Date	Date sold	Numerical	Calendar Date
Distance	Distance from CBD	Numerical	Kilometers
Postcode	Postal code of the house	Numerical	No Unit

Bedroom2	Scraped number of Bedrooms (from different source)	Numerical	No Unit
Bathroom	Number of Bathrooms	Numerical	No Unit
Car	Number of car spots	Numerical	No Unit
Land Size	Land size	Numerical	Squared meter
Building Area	Building Area	Numerical	Squared meter
Year Built	Build year	Numerical	Calendar Year
Council Area	Governing Council for the area	Categorical	No Unit
Latitude	Latitude	Numerical	Degrees
Longitude	Longitude	Numerical	Degrees
Region Name	General Region (West, North West, North, North east …etc)	Categorical	Direction
Property Count	Number of properties that exist in the suburb	Numerical	No Units

• Preparation for analysis

As above mentioned, this dataset includes twenty-two variables with seven categorical variables and fifteen numerical variables. As part of the preliminary data preparation, the variables should be filtered and modified prior to the analysis in light of their observed qualities.

Remove variables that do not provide information to predict house price.

According to our dataset, the following variables x, Suburb, Address, Date, seller G cannot be estimated. Most of these are in categorical and has too many levels. In order to that they were removed before the analysis.

Variable Selection

Sometimes, not all predictors are relevant to explaining the variability in the response variable. Removing non-significant predictors can lead to a simpler and more interpretable model.

Missing Value Handling

There are some missing values in our dataset. Specifically, we observed missing values in the variables such as distance, postcode, Bathroom, car, land size, building.

```
          X         Suburb        Address          Rooms
          0              0              0              0
       Type          Price         Method        SellerG
          0              0              0              0
       Date       Distance       Postcode       Bedroom2
          0              1              1           3469
   Bathroom            Car       Landsize   BuildingArea
       3471           3576           4793          10634
  YearBuilt    CouncilArea       Lattitude      Longtitude
       9438              0           3332           3332
 Regionname Propertycount
          0              1
```

area, Year built, latitude, longitude, and property count. The below figure shows the missing counts of each variable.
The below figure shows the missingness map of the dataset. Missing values were overseen by the following four methods.
1.Mean imputation
2.forward filled imputation
3.Machine Learning method
4. Drop missing value

After that, those missing values had been deducted from the raw data. finally, we select as the suitable method is dropping missing value method since min RMSE value.

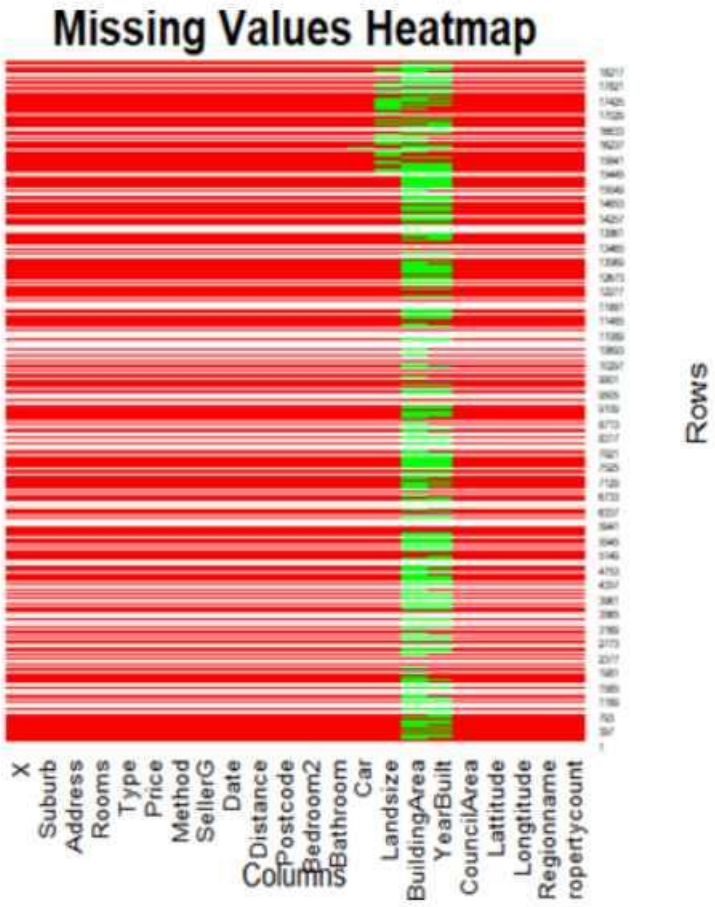


Figure 3.2: missingness map

```
RMSE (Mean Imputation): 428062.78553278913
RMSE (Forward Fill Imputation): 439799.7750078213
RMSE (Machine Learning Model Imputation): 428062.78553278913
RMSE (Drop Missing Values): 375455.4313943043
```

Figure 3.3: Missing Values with Methods

Checking Outliers

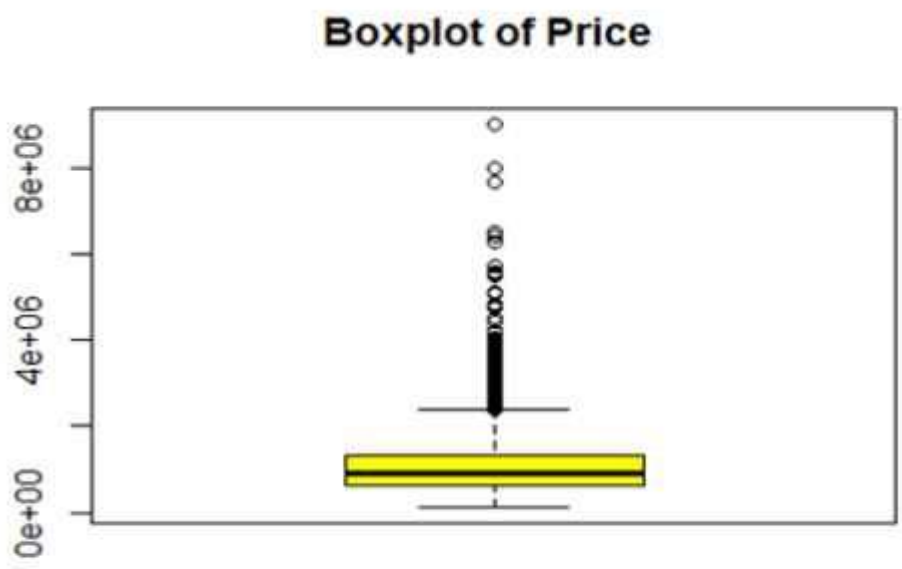


Figure 3.4: Visualization of Outliers

Visualization of Outliers Before the analysis the outliers of all the variables also checked. As a result of that, the price variable had the most outliers. The below boxplot shows the outliers of the price.

Changing categorical variables to numerical

The variables such as Type and Method both are included in our analysis so that those variables were changed into numeric values. Method variables contains these below categories.

Metadata:

Collaborator: DanB

4.Findings

The" home pricing" dataset give the impression to comprise information on property listings. Each property encompasses a variation of properties. Let us look at the key points: This dataset bargains info on residential properties in Melbourne's Abbotsford region. Each row parallels to a separate assets listing. The dataset includes info such as the property's address, number of rooms, generous of property (for example, a house), price, style of sale, and selling agent. It also postulates the size of the land and the number of parking spaces, along with the property's features, such as the number of bedrooms and bathrooms. The dataset also embraces information nearby the building's dimensions and the year it was built. Several of the homes have substantial historical backgrounds, dating back to the early 1900s. The property's location is revealed, together with its latitude, longitude, postcode, and distance from the city center. The assets are located in the" Northern Metropolitan" area.

Finally, the dataset includes the number of homes and the council area to provide some information about the neighborhood and local region. This dataset looks to be valuable for investigating geographic patterns, property trends, and Melbourne suburb dynamics. It may be useful to researchers interested in housing statistics and trends, data analysts, and real estate professionals. The picture below depicts the first rows of our cleaned dataset.

```
df.head()
```

	Rooms	Type	Price	Method	Distance	Postcode	Bedroom2	Bathroom	Car	Landsize	BuildingArea
0	2	h	1035000.0	S	2.5	3067	2	1	0	156	79.0
1	3	h	1465000.0	SP	2.5	3067	3	2	0	134	150.0
2	4	h	1600000.0	VB	2.5	3067	3	1	2	120	142.0
3	3	h	1876000.0	S	2.5	3067	4	2	0	245	210.0
4	2	h	1636000.0	S	2.5	3067	2	1	2	256	107.0

YearBuilt	CouncilArea	Lattitude	Longtitude	Regionname	Propertycount
1900	Yarra	-37.8079	144.9934	Northern Metropolitan	4019
1900	Yarra	-37.8093	144.9944	Northern Metropolitan	4019
2014	Yarra	-37.8072	144.9941	Northern Metropolitan	4019
1910	Yarra	-37.8024	144.9993	Northern Metropolitan	4019
1890	Yarra	-37.8060	144.9954	Northern Metropolitan	4019

Figure 4.1: Frist 5 Row in data set

Data Summary

To acquire an initial glimpse, we developed a numerical summary of the dataset. Mean, median, standard deviation, minimum, and maximum values for important.

	Price	Rooms	Distance	Postcode	Bedroom2
count	6.830000e+03	6830.000000	6830.000000	6830.000000	6830.000000
mean	1.077604e+06	2.978184	10.148960	3104.262225	2.951391
std	6.733202e+05	0.970479	5.991423	91.208614	0.970789
min	1.310000e+05	1.000000	0.000000	3000.000000	0.000000
25%	6.300000e+05	2.000000	6.100000	3044.000000	2.000000
50%	8.900000e+05	3.000000	9.200000	3083.000000	3.000000
75%	1.334000e+06	4.000000	13.000000	3147.000000	4.000000
max	9.000000e+06	8.000000	47.400000	3977.000000	9.000000

	Bathroom	Car	Landsize	BuildingArea	YearBuilt
count	6830.000000	6830.000000	6830.000000	6830.000000	6830.000000
mean	1.594143	1.606881	487.495461	143.446606	1964.444070
std	0.714366	0.944613	910.805627	89.970692	37.706332
min	1.000000	0.000000	0.000000	0.000000	1196.000000
25%	1.000000	1.000000	167.000000	93.000000	1940.000000
50%	1.000000	2.000000	404.000000	126.000000	1970.000000
75%	2.000000	2.000000	641.000000	173.000000	2000.000000
max	8.000000	10.000000	37000.000000	3112.000000	2018.000000

	Lattitude	Longtitude	Propertycount
count	6830.000000	6830.000000	6830.000000
mean	-37.808012	144.991877	7433.780527
std	0.080042	0.104983	4352.096045
min	-38.164920	144.542370	389.000000
25%	-37.856797	144.925522	4381.250000
50%	-37.802190	144.997000	6567.000000
75%	-37.756900	145.056100	10175.000000
max	-37.408530	145.526350	21650.000000

Figure 4.2: Five number summary

numerical variables are included (e.g., Price, property count, Longitude etc.). Below figure depicts a descriptive analysis of a cleaned dataset

Categorical Variable Analysis

Understanding the distribution of categorical variables is critical in data exploration. Within a dataset, evaluate and provide insights on four categorical variables: Type, council Area, Reigning tome, and ’Method.’ Figure no explains categorical variable value counts independently.

```
Type Distribution:
Type
h    4660
u    1528
t     642
Name: count, dtype: int64

Method Distribution:
Method
S     4373
SP     985
PI     826
VB     602
SA      44
Name: count, dtype: int64

Regionname Distribution:
Regionname
Southern Metropolitan         2312
Northern Metropolitan         2006
Western Metropolitan          1539
Eastern Metropolitan           686
South-Eastern Metropolitan     213
Eastern Victoria                28
Northern Victoria               26
Western Victoria                20
Name: count, dtype: int64
```

```
CouncilArea Distribution:
CouncilArea
Moreland              658
Boroondara            576
Moonee Valley         504
Darebin               433
Glen Eira             426
Maribyrnong           401
Yarra                 339
Port Phillip          336
Stonnington           335
Banyule               279
Melbourne             241
Bayside               223
Hobsons Bay           220
Brimbank              193
Monash                175
Manningham            150
Whitehorse            139
Kingston              111
Hume                   97
Whittlesea             89
Wyndham                47
Knox                   42
Melton                 42
Maroondah              35
Frankston              30
Greater Dandenong      21
Nillumbik              18
Casey                  16
Yarra Ranges           10
Macedon Ranges          5
Cardinia                5
Name: count, dtype: int64
```

Figure 4.3: categorical data summary

Univariate Analysis

violin plots to show the distributions of important numerical data (such as Price). This will help determine whether there is a pattern (normal, skewed, etc.). We used histograms to show the distributions of important numerical data (such as Price). This will help determine whether there is a pattern (normal, skewed, etc.).

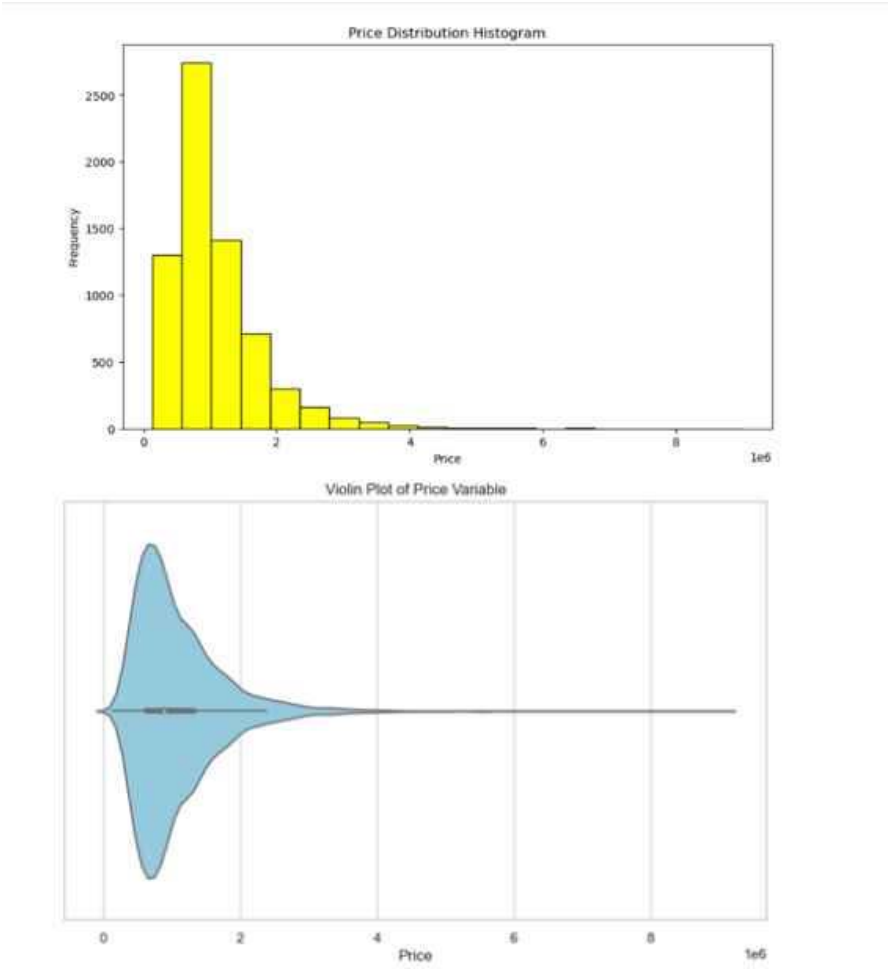


Figure 4.5: Univariate analysis

Plotting another numerical variable using dashboard then can get quick idea about distribution from seeing the dashboard.

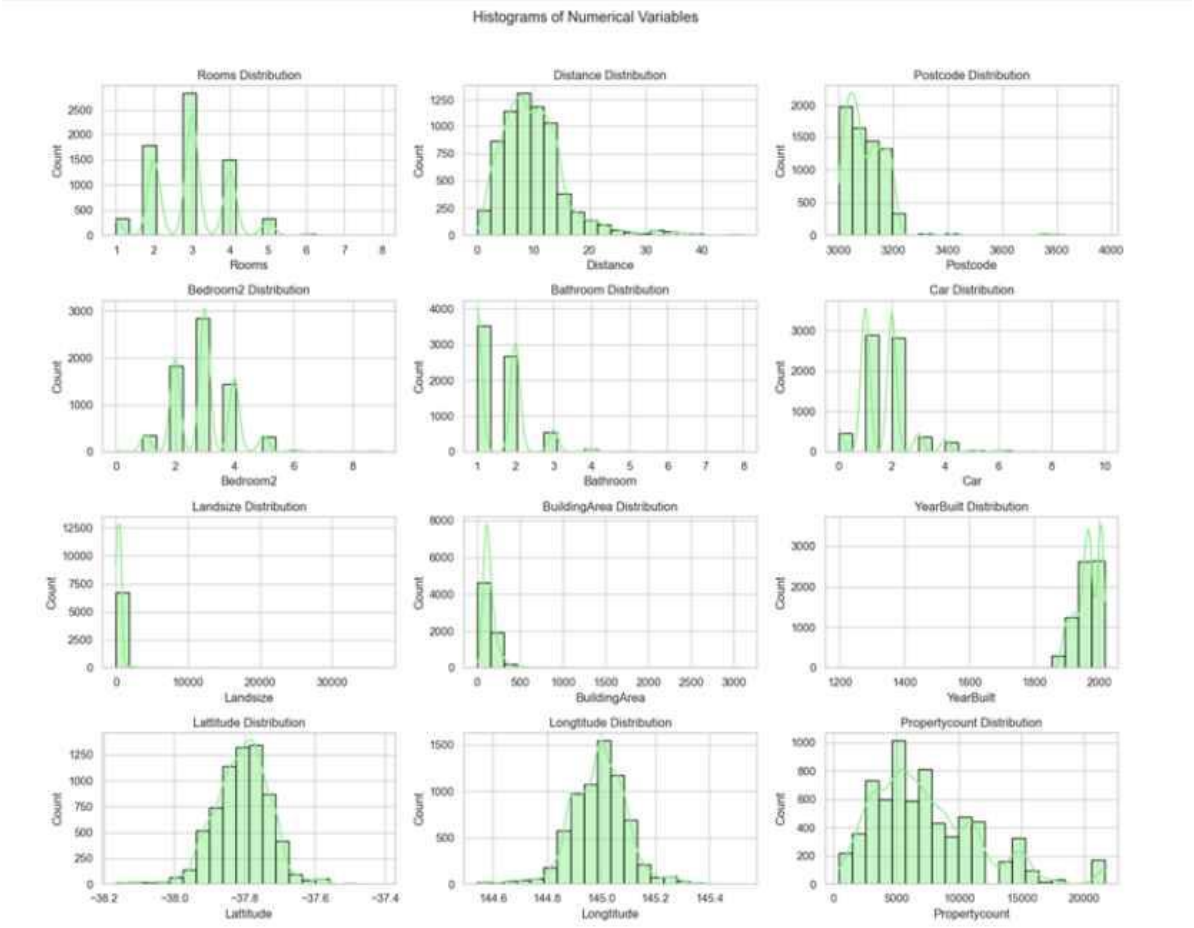


Figure 4.6: Histogram of variables

Bivariate Analysis

Bivariate analysis, which analyzes the relationships between two variables in a dataset, is an essential aspect of data exploration. Bivariate analysis, as opposed to univariate analysis, seeks to discover links, relationships, or patterns that emerge when two variables are studied concurrently. First, examine the correlations between our numerical variables, and then utilize a correlation heatmap to determine which factors are most related to our response variable.

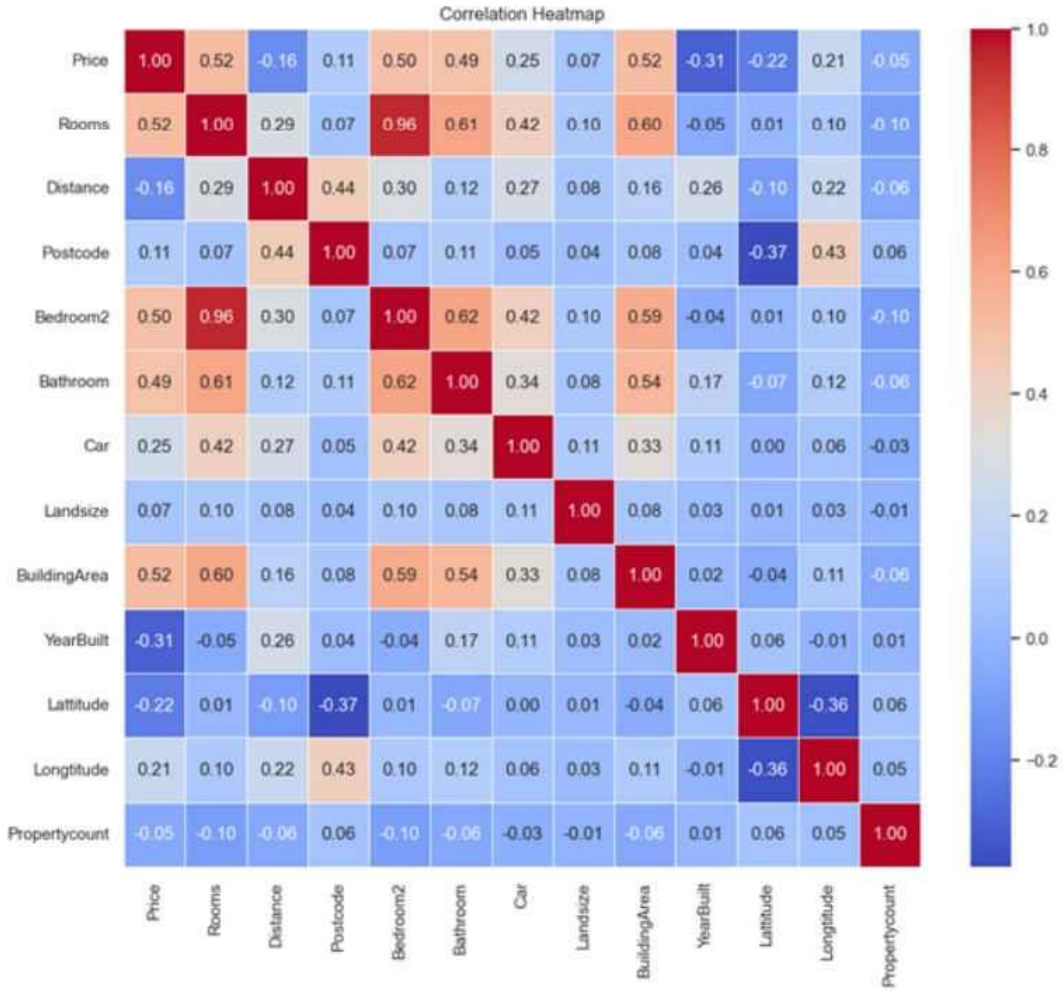


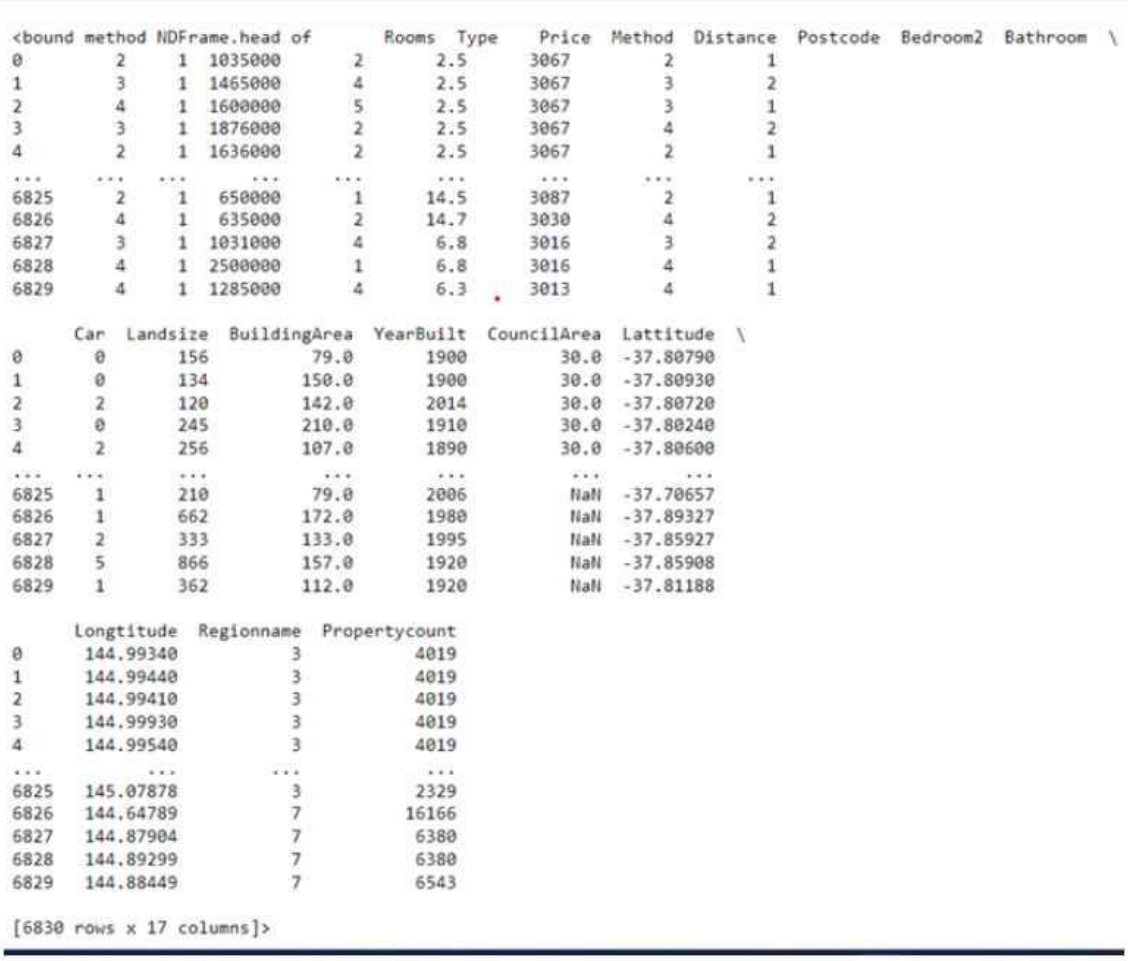
Figure 4.7: Heatmap of variables

The House price dataset exploratory data analysis gave crucial insights that will guide our future quantitative investigation. We discovered significant trends and relationships, such as the strongest correlations between Price, Room, and Build Area. The bathroom and the room are inseparably linked. The analysis provided a foundation for additional inquiry, allowing us to make educated judgments and get a better grasp of the facts in this case study...

Quantitative analysis

– Feature Engineering

After applying one-hot encoding to categorical variables, your dataset now comprises additional binary columns for each category. The unique columns such as 'Rooms,' 'Type,' 'Price,' 'Method,' 'Distance,' 'Postcode,' 'Bedroom2,' 'Bathroom,' 'Car,' 'Landsize,' 'BuildingArea,' 'YearBuilt,' 'CouncilArea,' 'Lattitude,' 'Longtitude,' 'Regionname,' and 'Propertycount' have been retained.



```
<bound method NDFrame.head of       Rooms  Type     Price  Method  Distance  Postcode  Bedroom2  Bathroom  \
0         2     1   1035000       2       2.5      3067         2         1
1         3     1   1465000       4       2.5      3067         3         2
2         4     1   1600000       5       2.5      3067         3         1
3         3     1   1876000       2       2.5      3067         4         2
4         2     1   1636000       2       2.5      3067         2         1
...     ...   ...       ...     ...       ...       ...       ...       ...
6825      2     1    650000       1      14.5      3087         2         1
6826      4     1    635000       2      14.7      3030         4         2
6827      3     1   1031000       4       6.8      3016         3         2
6828      4     1   2500000       1       6.8      3016         4         1
6829      4     1   1285000       4       6.3      3013         4         1

      Car  Landsize  BuildingArea  YearBuilt  CouncilArea  Lattitude  \
0       0       156          79.0       1900         30.0  -37.80790
1       0       134         150.0       1900         30.0  -37.80930
2       2       120         142.0       2014         30.0  -37.80720
3       0       245         210.0       1910         30.0  -37.80240
4       2       256         107.0       1890         30.0  -37.80600
...   ...       ...           ...        ...          ...        ...
6825    1       210          79.0       2006          NaN  -37.70657
6826    1       662         172.0       1980          NaN  -37.89327
6827    2       333         133.0       1995          NaN  -37.85927
6828    5       866         157.0       1920          NaN  -37.85908
6829    1       362         112.0       1920          NaN  -37.81188

      Longtitude  Regionname  Propertycount
0      144.99340           3           4019
1      144.99440           3           4019
2      144.99410           3           4019
3      144.99930           3           4019
4      144.99540           3           4019
...          ...         ...            ...
6825   145.07878           3           2329
6826   144.64789           7          16166
6827   144.87904           7           6380
6828   144.89299           7           6380
6829   144.88449           7           6543

[6830 rows x 17 columns]>
```

Figure 4.8: hot encoded description

The dataset appears to be associated to real estate, with features such as the 33 number of rooms, property type, price, distance to specific locations, number of bedrooms and bathrooms, car spaces, land size, building area, year built, geographical synchronizations (latitude and longitude), council area, region name, and property count in the area.

```
df.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 6830 entries, 0 to 6829
Data columns (total 16 columns):
 #   Column         Non-Null Count  Dtype
---  ------         --------------  -----
 0   Rooms          6830 non-null   int64
 1   Type           6830 non-null   int64
 2   Price          6830 non-null   int64
 3   Method         6830 non-null   int64
 4   Distance       6830 non-null   float64
 5   Postcode       6830 non-null   int64
 6   Bathroom       6830 non-null   int64
 7   Car            6830 non-null   int64
 8   Landsize       6830 non-null   int64
 9   BuildingArea   6830 non-null   float64
 10  YearBuilt      6830 non-null   int64
 11  CouncilArea    6196 non-null   float64
 12  Lattitude      6830 non-null   float64
 13  Longtitude     6830 non-null   float64
 14  Regionname     6830 non-null   int64
 15  Propertycount  6830 non-null   int64
dtypes: float64(5), int64(11)
memory usage: 853.9 KB
```

Figure 4.9: selected variable

Covariance analysis

The correlation matrix offers information about the relationships between the variables in your dataset, particularly the dependent variable, 'Price.' Each cell in the matrix represents the correlation coefficient between two variables, which ranges from -1 to one. A positive value specifies a positive correlation, while a negative value suggests a negative correlation. Here are some key results from the matrix:

Strong positive relationships

The variable 'Rooms' has a strong positive correlation with 'Price' (0.52), representative that the number of rooms has a significant effect on house prices. Similarly, 'Bathroom' and 'BuildingArea' have substantial positive correlations (0.49 and 0.52, respectively) with 'Price.

Strong Negative Relationships

The variable 'Type' has a significant negative correlation (-0.42) with 'Price,' implying that the type of property (e.g., unit, townhouse) is inversely linked to house prices. 'YearBuilt' has a moderately negative correlation (-0.31), inferring that older buildings may have lower prices.

Weak Relationships

Variables like 'Method,' 'Postcode,' 'Landsize,' and 'Car' have weak correlations with 'Price,' inferring that these factors may have a restricted impact on housing prices. The correlation coefficients for these variables are nearly zero.

Geographical Factors

Geographical variables such as 'Distance,' 'Postcode,' 'Lattitude,' 'Longitude,' and 'Regionname,' have fluctuating degrees of correlation with 'Price.' These relationships can be beneficial in determining how location influences house prices.

	Rooms	Type	Price	Method	Distance	Postcode	\
Rooms	1.000000	-0.579983	0.517718	-0.048423	0.289763	0.068674	
Type	-0.579983	1.000000	-0.419183	0.088179	-0.240904	0.008063	
Price	0.517718	-0.419183	1.000000	-0.041462	-0.164975	0.109343	
Method	-0.048423	0.088179	-0.041462	1.000000	-0.058832	-0.031344	
Distance	0.289763	-0.240904	-0.164975	-0.058832	1.000000	0.438274	
Postcode	0.068674	0.008063	0.109343	-0.031344	0.438274	1.000000	
Bathroom	0.613285	-0.273788	0.492481	-0.006614	0.124044	0.112776	
Car	0.420493	-0.281780	0.250916	-0.014382	0.265142	0.049226	
Landsize	0.099031	-0.045246	0.073536	-0.015780	0.082449	0.039998	
BuildingArea	0.603150	-0.388990	0.520492	-0.040639	0.155148	0.081791	
YearBuilt	-0.049272	0.320428	-0.307343	0.024371	0.258462	0.036819	
CouncilArea	-0.161131	0.095130	-0.130870	0.028821	-0.229657	-0.052010	
Lattitude	0.009005	-0.102741	-0.216919	0.000390	-0.101092	-0.374904	
Longtitude	0.096665	0.002320	0.209786	-0.048818	0.215594	0.430579	
Regionname	-0.005093	0.053642	0.090802	0.025594	-0.094625	-0.013522	
Propertycount	-0.100447	0.098420	-0.053336	-0.020742	-0.061433	0.058542	

	Bathroom	Car	Landsize	BuildingArea	YearBuilt	\
Rooms	0.613285	0.420493	0.099031	0.603150	-0.049272	
Type	-0.273788	-0.281780	-0.045246	-0.388990	0.320428	
Price	0.492481	0.250916	0.073536	0.520492	-0.307343	
Method	-0.006614	-0.014382	-0.015780	-0.040639	0.024371	
Distance	0.124044	0.265142	0.082449	0.155148	0.258462	
Postcode	0.112776	0.049226	0.039998	0.081791	0.036819	
Bathroom	1.000000	0.335331	0.081680	0.539717	0.166412	
Car	0.335331	1.000000	0.113427	0.331702	0.114340	
Landsize	0.081680	0.113427	1.000000	0.082815	0.031474	
BuildingArea	0.539717	0.331702	0.082815	1.000000	0.017940	
YearBuilt	0.166412	0.114340	0.031474	0.017940	1.000000	
CouncilArea	-0.093624	-0.149610	-0.029790	-0.162304	-0.053511	
Lattitude	-0.066983	0.003350	0.013133	-0.042230	0.064333	
Longtitude	0.119573	0.061410	0.026882	0.107153	-0.005064	
Regionname	0.036411	0.005498	-0.009509	0.031412	-0.026114	
Propertycount	-0.062127	-0.033258	-0.014909	-0.063315	0.005116	

Figure 4.10: correlaton matrix

– ANOVA Analysis Report

In our investigation to understand the effect of the categorical predictor 'Method' on house prices, we executed an analysis of variance (ANOVA). The ANOVA results show that 'Method' has a statistically significant effect on house prices (F (4, 6825) = 25.35, p ¡ 0.001). The sum of squares for 'Method' is 4.53e+13, indicating a meaningful change in house prices due to the various methods employed. The p-value of 7.12e-21 is significantly lower than the predictable significance level of 0.05, providing convincing evidence in contrast to the null hypothesis that the means of house prices athwart methods are equal. The residual sum of squares is 3.05e+15, which embodies the inexplicable variation in house prices after accounting for the effect of" Method." The F-statistic comparisons the variance explained by" Method" to the residual variance. The large F-statistic and tremendously low p-value indicate that the categorical predictor 'Method' makes a significant influence on the variability in house prices. Finally, our ANOVA outcomes show that the 'Method' choice is critical in influential house prices, highlighting the importance of taking this factor into account when learning real estate market subtleties. This finding has practical suggestions for housing industry stakeholders, as it provides valuable visions into the factors that influence property estimations.

	sum_sq	df	F	PR(>F)
C(Method)	4.532047e+13	4.0	25.347842	7.116341e-21
Residual	3.050676e+15	6825.0	NaN	NaN

Figure 4.12: summary of ANOVA

Accuracy checking

To appraise the predictive performance of our regression models, we used various metrics, with a main focus on Mean Absolute Error (MAE). MAE measures the average absolute difference between predicted and actual house prices, giving a direct signal of prediction accuracy. The following outcomes were got for each model:

• Random Forest Model: MAE: 166,207.03

• XGBoost Model: MAE: 158,435.11

• Support Vector Machine (SVM) Model: MAE: 453,112.18

The XGBoost model outpaced the other models tested in terms of accuracy, producing the lowest MAE. This means that, on average, the XGBoost model predicted house prices with an absolute difference of about 158,435.11 from the actual prices in the test set. The Random Forest model also accomplished reasonably well, but the SVM model had a large MAE, representing less accurate predictions than the other models. The conclusion to use XGBoost as the best model is supported not only by its superior accuracy, but also by its capability to minimize prediction errors. These findings are useful in determining the most consistent model for predicting house prices based on the given features. In future analyses, we may consider additional metrics like Mean Squared Error (MSE) and R-squared (R^2) to gain an inclusive understanding of model performance. Furthermore, cross-validation and hyperparameter modification could be used to fine-tune the models and improve their predictive skill.

```
print(f"The best model is {best_model[1]} with MAE: {best_model[0]}")
```

```
random forest model MAE is 166207.0322442396
xgboost model MAE is 158435.10539314518
svm  model MAE is 453112.1798872616
The best model is XGBoost with MAE: 158435.10539314518
```

```
[ ]:
```

Figure 4.14: MAE of Models

Model Creation and Prediction

In our search for a precise house price prediction model, we used the powerful XGBoost regression process. The model was trained on a dataset that included a variability of features such as the number of rooms, property type, distance from key locations, and other related attributes. The primary goal was to classify complex patterns in the data to make accurate predictions about house prices. The XGBoost

model was trained on a subset of the dataset, with features (’Rooms’, ’Type’, ’Method’, ’Distance’, ’Postcode’, ’Bathroom’, ’Car’, ’Landsize’, ’BuildingArea’, ’YearBuilt’, ’CouncilArea’, ’Lattitude’, ’Longtitude’, ’Regionname’, and ’Propertycount’) influencing the target variable ’Price.’ The training process elaborate iteratively enlightening the model’s predictive capabilities by optimizing its parameters. With the trained model at our disposal,

we made predictions for new data, shiny a hypothetical scenario with the following feature values: ’Rooms’: 2, ’Type’: 1, ’Method’: 2, ’Distance’: 2.5, ’Postcode’: 3067, ’Bathroom’: 1, ’Car’: 0, ’Landsize’: 156, ’BuildingArea’: 80, ’YearBuilt’: 1900, ’CouncilArea’: 30, ’Lattitude’: +55, ’Longtitude’: 144.9934, ’Regionname’: 3, and ’Propertycount’: 4019

For the given set of features, the model predicts a house price of USD 971,518.69.

This prediction proves the model’s ability to extrapolate from previously learned patterns and make guesses for new, unseen examples.

Figure 4.15: Train model

Limitation of model

– Evaluation and Overfitting Analysis

Our goal of developed a predictive model for house prices led us to use the authoritative XGBoost regression algorithm. To strictly evaluate the model’s performance, we executed a 5-fold cross-validation on the training set with Mean Squared Error (MSE) as the main metric. The average Cross-Validation MSE for the training set is predictable to be USD 99,259,636,005.79. This value is the average squared difference between predicted and actual house prices across all folds. The large degree of the MSE motivates us to study potential overfitting and fine-tune our model for upgraded predictive accuracy. Overfitting, in which a model performs exceptionally well on the training set but struggles with new, previously unseen data, is a concern in predictive modeling. In our case, the large MSE observed during cross-validation indicates that model complexity and hyperparameter tuning should be investigated further. To reduce overfitting, we will investigate the model’s performance on a separate validation or testing set. Discrepancies between training set MSE and validation/test set performance will guide hyperparameter adjustments or the use of regularization techniques. 41 Figure 4.17: House price prediction

– Cross-Validation Evaluation

In assessing the performance of our XGBoost regression model through 5-fold cross-validation, we observe the following negative mean squared error scores:

Fold 1: -8.96e+10

Fold 2: -1.08e+11

Fold 3: -8.38e+10

Fold 4: -8.72e+10

Fold 5: -1.28e+11

These scores, while negative, indicate the average squared difference between predicted and actual values. A higher (less negative) score is preferable, suggesting lower prediction errors. On average, the model displays a considerable ability to make accurate predictions across the folds.

```
]: from sklearn.model_selection import cross_val_score

   cv_scores = cross_val_score(model, X, y, cv=5, scoring='neg_mean_squared_error')

   cv_scores

]: array([-8.95966384e+10, -1.08202523e+11, -8.37988204e+10, -8.71528525e+10,
          -1.27547345e+11])
```

Figure 4.18: cross validation matrix

– Normality of residual

The XGBoost regression model's reliability is assessed by examining residuals, which represent differences between predicted and actual house prices. A symmetric distribution of residuals around zero indicates unbiased predictions, with errors evenly distributed above and below true prices. This symmetry aligns with expectations and enhances confidence in the model's ability to make accurate predictions across diverse scenarios.

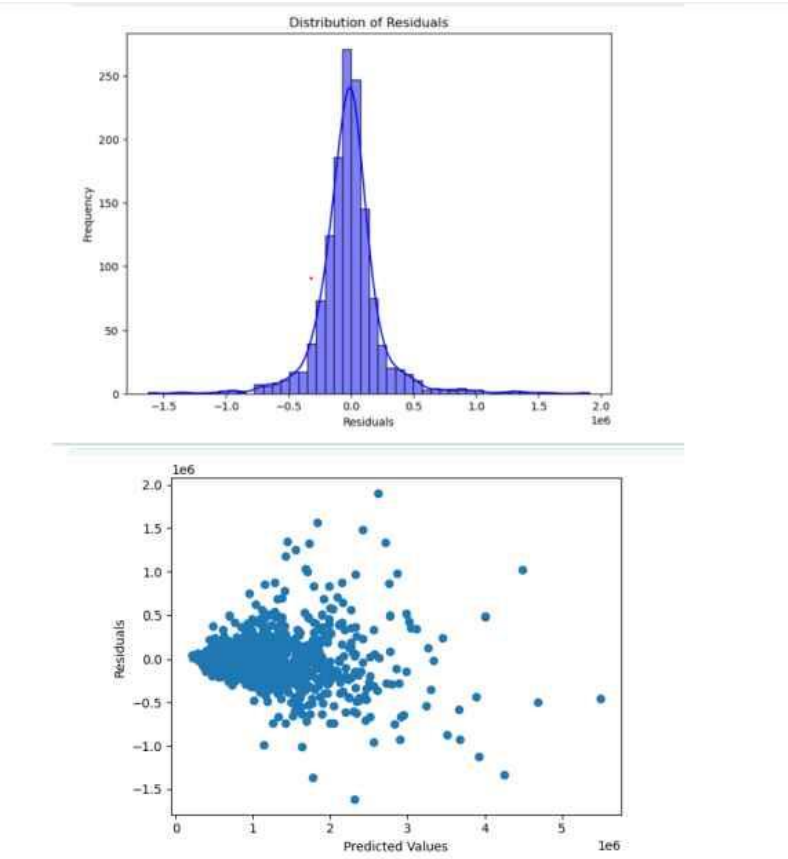


Figure 4.19: residual distribution

– Feature Importance

The analysis of feature importance from our XGBoost regression model sheds light on the factors that have the greatest effect on house price prediction. Particularly," BuildingArea" and" Landsize" have the maximum importance scores, at 676.0 and 624.0, singly. These findings direct that the size and dimensions of properties play a key role in determining their market value. Furthermore, geographical coordinates, denoted by" Lattitude" and" Longtitude," have a significant impact, emphasizing the location's influence on property prices. Temporal and structural considerations: Among other distinguished contributors," YearBuilt" and" Distance" have importance scores of 486.0 and 497.0, respectively.

A property's construction year, as well as its proximity to key amenities or city centers, clearly influence its price. Understanding these temporal and spatial dimensions lets us to appreciate the subtle collaboration of historical context, accessibility, and assets values. Method and Property Characteristics: Features with importance scores of 218.0, 198.0, and 100.0 offer additional intuitions into the factors that influence house prices. Price predictions are seriously influenced by the sales method used, the total number of properties in each area, and the type of property. Identifying these factors provides an all-inclusive understanding of the various elements that shape the real estate landscape, emphasizing the multifaceted nature of property valuation.

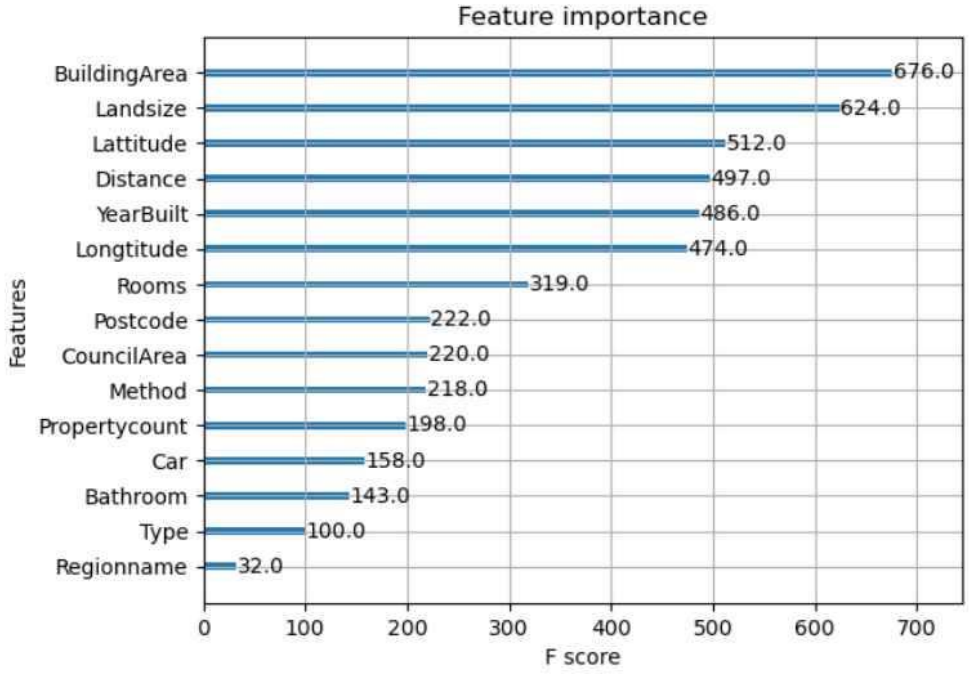


Figure 4.20: Feature important plot

5.Conclsion

Each of the study goals' findings will be discussed.

1. Determinants of House Price and independent variables Relative Importance
   The analysis revealed the key factors influencing house prices. The number of rooms ('Rooms'), property type ('Type'), building area ('BuildingArea'), and bathroom count ('Bathroom') all have strong correlations with house prices. Geographical factors such as distance to specific locations, postcode, latitude, and longitude all play a key role. The relative importance of each factor, as determined by the feature importance analysis, provides additional insight. Notably, 'BuildingArea' and 'Landsize' emerge as the most crucial factors in determining market value. Geographic coordinates, known as 'Lattitude' and 'Longitude,' play a key role in determining property prices. Temporal factors like 'YearBuilt' and 'Distance' to key amenities play significant roles. The quarterly analysis could provide a more nuanced understanding of how these factors change over time and across regions. This temporal perspective may reveal trends or patterns that would not be apparent in an aggregate analysis. For example, certain quarters or seasons may experience increased demand, which affects prices. This thorough investigation would add depth to the predictive model and improve forty-seven its accuracy.
2. Selection of the Most Suitable Machine Learning Algorithm.
   An evaluation of machine learning algorithms, with Random Forest, XGBoost, and Support Vector Machine (SVM), exposed that XGBoost was supplementary accurate in predicting house prices. The result to use XGBoost as the favored algorithm is well-founded, given its capability to capture complex patterns in data and decrease prediction errors. Boost's capability to manage both numerical and categorical features, combined with its feature selection ability, makes it an exceptional prime for real estate price forecast. The algorithm's performance is dependable with its reputation as an authoritative and versatile tool in various data science applications.
3. Reliability and Limitations of the Developed Algorithm.
   The reliability of the developed algorithm is assessed by evaluating its predictive performance using metrics such as mean absolute error (MAE). The results show that the XGBoost model performs well, with a low MAE, indicating that predictions are accurate on average. However, the limitation is the potential for overfitting during the 5-fold cross-validation on the training set. Overfitting, in which the model performs exceptionally well on the training set but struggles with new, unknown data, is a widespread problem. The large Cross-Validation MSE on the training set necessitates further investigation into model complexity and hyperparameter tuning to improve predictive accuracy. The plan to assess the model on a separate validation or test set and adjust as needed is a prudent step toward addressing this limitation. The normality of residuals analysis gives confidence that the model can make unbiased predictions. The symmetric distribution of residuals around zero indicates that the model predicts house prices accurately.

6. References

[1] Author, A. A. (Year, Month Day of publication). Title of the article. Title of the Journal, Volume (Issue), Page range. DOI or URL

[2] Geron, A. (2019). Firsthand Machine Learning with Scikit-Learn, Keras, and ´ TensorFlow. O'Reilly Media.

[3] Marcelino, P. (2017, February 28). Comprehensive Data Exploration with Python. Kaggle. https://www.kaggle.com/pmarcelino/comprehensive-data-exploration-withpython

[4] https://www.geeksforgeeks.org/house-price-prediction-using-machine-learningin-python/

[5] Durganjali, P., Pujitha, M. (2019). House Resale Price Prediction Using Classification Algorithms. 2019 International Conference on Smart Structures and Systems (ICSSS), 1-4.

[6] Lu, Sifei Li, Zengxiang Qin, Zheng Yang, Xulei Goh, Rick. (2017). A hybrid regression technique for house prices prediction. 319-323. 10.1109/IEEM.2017.82899048.Appendix

7. Appendix

- Metadata - https://github.com/sachithnimesh/Predict-House-price-choosing-suitable-ML-model

X	Suburb	Address	Rooms	Type	Price	Method	SellerG
1	Abbotsford	25 Bloomburg St	2	h	1035000	S	Biggin
2	Abbotsford	5 Charles St	3	h	1465000	SP	Biggin
3	Abbotsford	55a Park St	4	h	1600000	VB	Nelson
4	Abbotsford	124 Yarra St	3	h	1876000	S	Nelson
5	Abbotsford	98 Charles St	2	h	1636000	S	Nelson

Table 1: Real Estate Data

Date	Distance	Postcode	Bedroom2	Bathroom	Car	Landsize	BuildingArea	YearBuil
4/2/2016	2.5	3067	2	1	0	156	79	1900
4/3/2017	2.5	3067	3	2	0	134	150	1900
4/6/2016	2.5	3067	3	1	2	120	142	2014
7/5/2016	2.5	3067	4	2	0	245	210	1910
8/10/2016	2.5	3067	2	1	2	256	107	1890

Table 2: Property Data

Latitude	Longitude	Region Name	Property Count
-37.8079	144.9934	Northern Metropolitan	4019
-37.8093	144.9944	Northern Metropolitan	4019
-37.8072	144.9941	Northern Metropolitan	4019
-37.8024	144.9993	Northern Metropolitan	4019
-37.806	144.9954	Northern Metropolitan	4019

Table 3: Geographical Data