Prediction of housing prices in Boston

Using Machine learning algorithms

Literature review and Data description

Rabia Danish
Student Number: 501210698
Supervisor's Name: Professor Dr. Ceni Babaoglu
Date of submission: 5th June 2023

Ryerson University

Metropolitan University

Table of Contents

Abstract

In today's ever challenging world of real estate, where housing prices are constantly on the rise, it has become increasingly important for individuals to make well-informed decisions when it comes to property investments. Our research uses available "Housing Prices Dataset" for Boston. By harnessing the power of this extensive dataset, our goal is to unravel the intricate factors that influence housing prices and provide individuals with the knowledge necessary to navigate the market with confidence and foresight. This will allow to determine future prices based on the available features which will enable people to make reliable decisions. Our dataset consists of various features related to the properties, including price, area, number of bedrooms, bathrooms, stories, and other factors that may influence housing prices. The objective of this project is to develop precise prediction models that can assist in estimating housing prices.

We aim in conducting exploratory data analysis to gain valuable insights into the dataset, identifying patterns, and understanding the relationships between the features and the target variable (price). Data preprocessing techniques such as handling missing values, feature selection, and encoding categorical variables are applied to ensure the data is in a suitable format for analysis.

In this study, we will employ several machine learning techniques, including linear regression, decision tree regression, and random forest regression, to develop predictive models for housing price estimation using cross Validation technique. These models will consider input features such as the number of bedrooms, bathrooms, stories, and the area of the house, aiming to capture the underlying linear relationship with the target variable, price. To assess the performance of these models, we will utilize commonly used evaluation metrics such as Mean Absolute Error (MAE), Mean Squared Error (MSE), and Root Mean Squared Error (RMSE) and R2 coefficient.

By comparing the performance of these models using these evaluation metrics, we will gain insights into their accuracy and reliability in predicting housing prices.

Additionally, correlation analysis is conducted to examine the relationships among the different attributes of houses using python. By determining the degree of correlation between variables, we can identify highly correlated attributes and assess their impact on housing prices. This analysis helps address the issue of multicollinearity, where highly correlated attributes can introduce redundancy and distort the accuracy of the prediction models. By identifying the most influential attributes, we can refine the models and focus on the features that have a significant impact on housing prices.

Overall, this data analytics capstone project aims to use the provided dataset to develop accurate prediction models for housing prices by considering different features which may impact those prices, therefore, enabling stakeholders to make informed decisions regarding real estate investments, pricing policies, and market analysis.

Data set:

The Project aims to use data to achieve the goal:

https://www.kaggle.com/datasets/yasserh/housing-prices-dataset

Introduction

In the ever-changing real estate market, accurate prediction of housing prices is essential for making informed decisions regarding property investments.

The literature review section delves into the significance of predicting house prices and its practical implications. As house prices continue to rise, accurate prediction models are essential for various stakeholders such as landowners, valuers, and policymakers. These models enable them to determine property valuations and appropriate sale prices, assisting potential buyers in making informed decisions. While physical conditions, architectural styles, and location are known to impact housing prices, the specific variables influencing prices can vary across different regions. Therefore, an effective prediction model must consider and accommodate the unique variables that drive housing prices in the specific area under consideration.

A number of scholars have conducted extensive research in predicting house prices. This literature review aims to explore existing research articles that focus on housing price prediction and how they can contribute to the research project based on the provided dataset. The articles mentioned provide insights into the application of regression models and machine learning techniques, such as Random Forest and multiple linear regression, for housing price prediction.

Literature Review

The study conducted by Adetunji et al., employed a regression model to analyze the Boston housing dataset with the goal of predicting house prices based on the dataset's features. The proposed approach involved developing a prediction model using the Random Forest algorithm. The UCI Machine Learning Repository's Boston housing dataset, consisting of 506 entries and 14 features [1], was used to evaluate the performance of the proposed model.

The study focused on exploring the effectiveness of the Random Forest machine learning technique for house price prediction. The Random Forest algorithm is a classifier that leverages multiple decision trees trained on different subsets of the dataset to enhance predictive accuracy. By aggregating the outcomes of these trees, the algorithm determines the final prediction based on the majority vote.

With the Random Forest algorithm at the core, the authors proceed to train the prediction model and meticulously assess its performance using well-established metrics like mean absolute error, mean squared error, and R-squared. A comparison between the predicted prices and the actual prices revealed that the model's predictions were within an acceptable range, with an error margin of ±5. This suggests that the model has the potential to provide reliable predictions for house prices.

Overall, this research highlights the successful application of the Random Forest machine learning technique in house price prediction. The findings demonstrate the model's ability to effectively analyze the Boston housing dataset and generate accurate predictions, making it a promising approach for real estate valuation and decision-making processes [1].
Similarly, the study by Truong et al., focused on the comparative analysis of different machine learning models for housing price prediction using the "Housing Price in Beijing" dataset. The study evaluated the performance of Random Forest, XGBoost, LightGBM, Hybrid Regression, and Stacked Generalization Regression models and analyzes their strengths and weaknesses.

The dataset comprises over 300,000 housing transactions with 26 variables as features for predicting the average price per square meter [2]. The evaluation metric used is the Root Mean Squared Logarithmic Error (RMSLE).

Results indicate that Random Forest performs well on the training set due to its suitability for datasets with Boolean features. However, it is prone to overfitting. XGBoost and LightGBM exhibit lower accuracy but do not suffer from overfitting. Hybrid Regression and Stacked Generalization Regression deliver promising results without extensive tuning. Hybrid Regression performs best on the training set, while Stacked Generalization Regression excels on the test set.

Stacked Generalization Regression's superior performance can be attributed to K-fold cross-validation and the coupling effect of multiple regressions. Hybrid Regression averages predictions, while Stacked Generalization Regression uses a second stacking level for more accurate predictions based on pre-estimated prices.

Overall the study found that different models had their pros and cons. Random Forest had the lowest error on the training set but was prone to overfitting and had high time complexity. XGBoost and LightGBM had decent accuracy and lower time complexity. Hybrid Regression performed better than the previous methods due to its generalization capability. Stacked Generalization Regression had the most complicated architecture but offered the highest accuracy. Time complexity should be considered, especially since both Hybrid Regression and Stacked Generalization Regression included Random Forest [2].

By utilizing multiple linear regression and Spearman correlation analysis, the study conducted by Zhang et al., aimed to predict housing prices using a multiple linear regression model and the Spearman correlation coefficient to identify influential factors. The analysis was conducted on the Boston housing price dataset. The results showed that the multiple linear regression model was able to effectively predict and analyze housing prices to some extent.

The Spearman correlation coefficient analysis revealed several significant factors influencing housing prices, including the proportion of lower-income groups in the region, the proportion of property land area larger than 25,000 square feet, and the average number of rooms. These factors played a crucial role in shaping housing prices.

The model training process involved optimizing a gradient descent optimizer to calculate the model parameters and minimize the loss function. After training for 100 epochs, the model's prediction results were generally consistent with the real values in the comparison set, indicating its capability to capture the overall trend of housing prices. Subsequently, training for 500 epochs further improved the model's performance.

The validation results confirmed the accuracy and practicality of the empirical model constructed using the Boston housing price dataset. However, it was acknowledged that the prediction accuracy was still limited in certain cases. The study concluded that while the multiple linear regression model was effective to some extent, further research was needed to enhance its universality and explore advanced machine learning methodologies for housing price prediction.

Overall, the results demonstrated the capability of the multiple linear regression model to predict and analyze housing prices. The influential factors identified through the Spearman correlation coefficient analysis provided valuable insights into the housing market. Nonetheless, there is room for improvement to enhance the prediction accuracy and develop more robust algorithms using advanced machine learning techniques [3].

The study by Zhang et al., presented a methodology for predicting housing prices using a multiple linear regression model and the Spearman correlation coefficient. While the results showed some effectiveness in predicting housing prices, there are several limitation to this study

and further research needs to be conducted. Firstly, the study relied solely on a multiple linear regression model, which limited to capture complex nonlinear relationships in the housing market. To enhance the predictive accuracy, the study could benefit from exploring more advanced machine learning algorithms, such as decision trees, random forests, or neural networks. These algorithms have the potential to capture nonlinearities and interactions among variables, leading to improved predictions.

Additionally, the study primarily relied on the Spearman correlation coefficient to identify influential factors. While correlation analysis is valuable, it is important to consider other statistical techniques, such as feature selection algorithms or dimensionality reduction methods, to identify the most relevant and informative features for predicting housing prices. This would help refine the model and potentially improve its predictive performance.

Furthermore, the study could benefit from conducting a more comprehensive evaluation of the predictive model. This could involve using different evaluation metrics, such as mean absolute error or R-squared, to assess the model's performance and compare it with alternative approaches. It would provide a more robust assessment of the model's predictive capabilities.

In conclusion, while the study [3] presented a methodology for predicting housing prices, there are opportunities for improving the predictive model based on the collected dataset including:

1. Feature Selection: Identify the most significant factors influencing housing prices by calculating correlation coefficients or using other feature selection techniques. This will help in selecting the most relevant features for the model.

2. Nonlinear Relationships: Explore the possibility of nonlinear relationships between the independent variables and housing prices. Consider incorporating polynomial features or using nonlinear regression techniques to capture complex patterns in the data.
3. Advanced Regression Techniques: Instead of solely relying on multiple linear regression, consider using decision trees or random forests. These algorithms can capture nonlinear relationships and interactions among variables, potentially improving predictive accuracy.
4. Model Evaluation: Evaluate the model using appropriate metrics such as mean absolute error (MAE) or root mean squared error (RMSE). Compare the performance of different models and select the one that provides the best balance between accuracy and generalizability.

Therefore, by considering the relevant approaches, we can enhance the predictive accuracy and interpretability of the model when analyzing the given housing price dataset.

The study conducted in Taiwan [4], focused on house price prediction and compared three predictive systems: boosting ensemble regression trees (BERT), support vector regression (SVR), and Gaussian process regression (GPR). These systems were optimized using Bayesian optimization (BO) to improve their performance. BERT, implemented with the least squares boosting algorithm, showed the lowest error rate median and low error variability, indicating stability and high accuracy. SVR had the highest error median, suggesting larger deviations in predicted and actual prices, while GPR exhibited significant variations in prediction accuracy. The results demonstrated the superior performance of BERT in terms of stability and accuracy. Overall, BERT was deemed a highly accurate and stable predictive system for house price forecasting, particularly suitable for small data samples. The study contributes valuable insights

for real estate professionals, investors, and analysts in the field of house price prediction [4]. The application of machine learning algorithms for house price prediction in a small town in India emphasized the need for accurate models to aid customers in finding suitable houses [5]. The study incorporates various house attributes, such as the number of bedrooms, age, proximity to amenities, and nearby schools, into the prediction model. Fundamental machine learning algorithms, including decision tree classification, decision tree regression, and multiple linear regression, are implemented using the Scikit-Learn tool. The research develops a model that predicts house availability and estimates prices, comparing decision tree regression and multiple linear regression. The findings show that multiple linear regression performs better for house price prediction. The study highlights the potential of machine learning algorithms for accurate house price prediction and suggests expanding the dataset and exploring advanced techniques for further improvement [5].

The study conducted by Ho et al., explored the use of machine learning algorithms, including support vector machine (SVM), random forest (RF), and gradient boosting machine (GBM), for predicting property prices in Hong Kong. Using a dataset spanning 18 years and over 40,000 housing transactions, the study compared the performance of these algorithms. The results show that RF and GBM outperform SVM in terms of predictive power, achieving lower mean squared error (MSE), root mean squared error (RMSE), and mean absolute percentage error (MAPE). However, SVM remains valuable for producing accurate predictions within tight time constraints. The study also highlights the influential role of housing attributes such as floor area, property age, and proximity to the Central District on property prices. Overall, the research demonstrates the potential of advanced machine learning techniques in property price prediction, with RF and GBM proving to be effective for accurate estimations. The study suggests careful

feature selection and acknowledges the interpretability challenges associated with machine learning models. Additionally, it emphasizes the importance of considering computation time, recommending SVM for quick forecasts and RF/GBM when high predictive accuracy is desired [6].

Discussion

Our study revolves around the application of machine learning algorithms for predicting house or property prices. Various algorithms, such as Random Forest, XGBoost, LightGBM, Hybrid Regression, Stacked Generalization Regression, Support Vector Regression (SVR), Gaussian Process Regression (GPR), and multiple linear regression, have been explored and compared in different contexts and datasets before. Our study aims to improve predictive accuracy and provide valuable insights into the factors influencing property prices.

Previous studies highlighted the need to consider several critical points. Although, machine learning algorithms show promise in predicting house prices, there are limitations and challenges. The choice of algorithm can impact predictive accuracy, with different algorithms performing better in specific contexts. Overfitting, generalizability, interpretability, and computation time are important factors to consider. Furthermore, the studies often focus on specific datasets or regions, which may limit the generalizability of the findings. There is a need for further research to explore advanced techniques, feature selection methods, ensemble methods, and larger or more diverse datasets.

None of the previous researches had considered the same housing attributes. Each article focuses on different datasets, regions, and machine learning algorithms. While there may be similarities

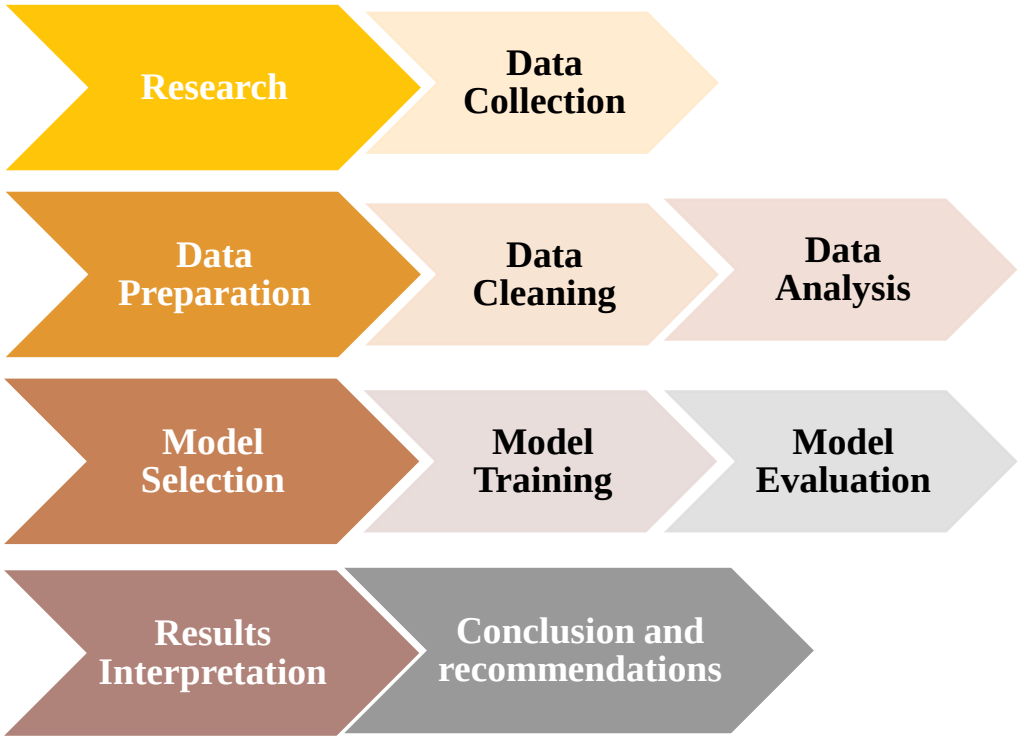
in terms of the use of machine learning for house price prediction, the specific approaches, algorithms employed, and datasets used vary among the studies.

Previous studies also compared and evaluated the performance of different algorithms or techniques on various datasets. Some studies also examined the impact of housing attributes or factors on property prices. Our research provides a foundation for further exploration and improvement in predicting property prices.

Our study builds upon the limitations highlighted in the study by Zhang et al., by incorporating advanced techniques to enhance the predictive model for housing prices. The study suggests feature selection techniques, such as calculating correlation coefficients, to identify the most significant factors influencing housing prices. Additionally, the project recognizes the potential for nonlinear relationships between independent variables and housing prices, proposing the inclusion of polynomial features or nonlinear regression techniques to capture complex patterns. Moreover, the project goes beyond multiple linear regression by considering advanced regression algorithms, including random forest and Decision tree regression. By incorporating these techniques, our research aims to improve the model's predictive accuracy. Furthermore, the project emphasizes the importance of evaluating the model using appropriate metrics such as mean absolute error (MAE) or root mean squared error (RMSE) and comparing the performance of different models. By selecting the model that provides the best balance between accuracy and generalizability, the project enhances the methodology used in [3]. Overall, our aim is to build upon existing research by incorporating these suggestions and improving the predictive accuracy and interpretability of the model for housing prices using advanced techniques, including random forest, linear regression, Decision tree regression and comprehensive model evaluation.

Our study builds upon the models and offers stakeholders in the real estate market, such as professionals, investors, and analysts, more reliable tools for making informed decisions.

Tentative Methodology for the project



By following this methodology, data scientists and analysts can systematically collect, clean, analyze, and train models on the available dataset. This process helps in gaining a comprehensive understanding of the data, developing accurate predictive models, and evaluating their performance for making informed decisions.

Data Preparation

The steps ensure the data is in a suitable format for analysis and model training, improving the accuracy and reliability of the results.

Data Collection

The dataset utilized in this project was a sourced from Kaggle Inc [7]. The dataset contains information about different houses in Boston. The dataset consists of 545 samples and 12 feature variables.

Data Dictionary

The dataset can be described briefly as follows.

Table 1. Data dictionary

Attributes	Explanation	Variable type
Price	Price of the Houses	Numeric/ Continuous
Area	Area of a House	Numeric/ Continuous
Bedrooms	Number of House Bedrooms	Categorical/nominal
Bathrooms	Number of Bathrooms	Categorical/nominal
Stories	Number of House Stories	Categorical/nominal
Mainroad	Weather connected to Main Road	Categorical/nominal
Guestroom	Weather the house has a guest room	Categorical/nominal
Basement	Weather the house has a basement	Categorical/nominal
Hotwaterheating	Weather the house has a hot water heater	Categorical/nominal
Airconditioning	Whether the house has an air conditioning system.	Categorical/nominal
Parking	The number of parking spaces available for the house.	Categorical/nominal

Prefarea	Whether the house is located in a preferred area or not.	Categorical/nominal
Furnishingstatus	The furnishing status of the house.	Categorical/nominal

```
RangeIndex: 545 entries, 0 to 544
Data columns (total 13 columns):
 #   Column            Non-Null Count  Dtype
---  ------            --------------  -----
 0   price             545 non-null    int64
 1   area              545 non-null    int64
 2   bedrooms          545 non-null    int64
 3   bathrooms         545 non-null    int64
 4   stories           545 non-null    int64
 5   mainroad          545 non-null    object
 6   guestroom         545 non-null    object
 7   basement          545 non-null    object
 8   hotwaterheating   545 non-null    object
 9   airconditioning   545 non-null    object
 10  parking           545 non-null    int64
 11  prefarea          545 non-null    object
 12  furnishingstatus  545 non-null    object
dtypes: int64(6), object(7)
```

Fig 1. Data types in python

Missing Values

After reading the "Hosuing.csv" dataset into a Pandas dataframe. We looked into datatypes, missing values and duplicate rows.

```
data.isnull().sum()
```

```
price             0
area              0
bedrooms          0
bathrooms         0
stories           0
mainroad          0
guestroom         0
basement          0
hotwaterheating   0
airconditioning   0
parking           0
prefarea          0
furnishingstatus  0
dtype: int64
```

```
data.duplicated().value_counts()
```

```
False    545
dtype: int64
```

Fig 2. Missing values and Duplicates

Our analysis of the dataset revealed that there are no instances of missing values or duplicates. This indicates that the dataset is complete and unique. However, we can see that there are categorical variables in the dataset and we need to convert categorical variables into numerical representations before training the model. To handle categorical variables, we can use one-hot encoding or label encoding. One-hot encoding creates binary columns for each category, while label encoding assigns a unique numerical value to each category.

Outlier's detection for numerical attributes

Table 2. Outlier counts and percentages for the attributes

Attributes	# of Outliers	% of Outliers
Price	6 outliers	1.10%
Area	7 outliers	1.28%
Bedrooms	2 outliers	0.37%
Bathrooms	11 outliers	2.02%
Stories	0 outliers	0%
Parking	0 outliers	0%

The analysis aimed to determine the impact of outliers on housing price prediction. Outliers are observed in the price, area, bedrooms, and bathrooms columns. The count of outliers for each attribute is relatively low, ranging from 2 to 11 instances. No outliers are found in the stories and parking columns. The question at hand was whether removing outliers would enhance the accuracy of the linear regression model.

The results indicate that removing outliers could be beneficial for certain attributes, such as price, area, bedrooms, and bathrooms. Outliers have the potential to distort predictions and introduce bias into the model. By eliminating outliers, the model's accuracy may be improved. However, careful consideration is necessary when deciding to remove outliers, taking into account domain knowledge and specific requirements. Since the dataset had a relatively low count of outliers, their removal may not significantly impact the overall dataset. Further analysis is recommended to assess the validity of outliers and determine if they represent genuine extreme values or data errors. Documentation of the outlier identification criteria and an evaluation of the potential impact on the data's distribution and characteristics are essential aspects to consider.

Statistical Summary for Numerical attributes

Using the describe() function we generate a table that shows the summary statistics (count, mean, standard deviation, minimum, quartiles, and maximum) for numerical attribute of the dataset.

Table 3. Statistical Summary

	price	area	bedrooms	bathrooms	stories	parking
count	5.450000e+02	545.000000	545.000000	545.000000	545.000000	545.000000
mean	4.766729e+06	5150.541284	2.965138	1.286239	1.805505	0.693578
std	1.870440e+06	2170.141023	0.738064	0.502470	0.867492	0.861586
min	1.750000e+06	1650.000000	1.000000	1.000000	1.000000	0.000000
25%	3.430000e+06	3600.000000	2.000000	1.000000	1.000000	0.000000
50%	4.340000e+06	4600.000000	3.000000	1.000000	2.000000	0.000000
75%	5.740000e+06	6360.000000	3.000000	2.000000	2.000000	1.000000
max	1.330000e+07	16200.000000	6.000000	4.000000	4.000000	3.000000

During our analysis, we explored important aspects of the housing dataset to gain a better understanding of the data. Here are the key findings:

Price: The average house price is around 4.77 million, ranging from 1.75 million to 13.3 million. Most houses have prices below 5.74 million.

Area: The average house size is approximately 5150 square units, ranging from 1650 to 16200 square units. The majority of houses have an area below 6360 square units.

Bedrooms: On average, houses have about 3 bedrooms, ranging from 1 to 6. Most houses have 3 or fewer bedrooms.

Bathrooms: Houses typically have around 1.3 bathrooms, ranging from 1 to 4. The majority of houses have 2 or fewer bathrooms.

Stories: Most houses have 2 or fewer stories, with an average of 1.81 stories and a range of 1 to 4.

Parking: On average, houses have approximately 0.69 parking spaces, ranging from 0 to 3. The majority of houses have 1 or fewer parking spaces.

These findings provide valuable insights into the dataset, giving us a clearer picture of the distribution and characteristics of the variables. This knowledge can be used to further analyze and develop models for predicting housing prices based on these features.

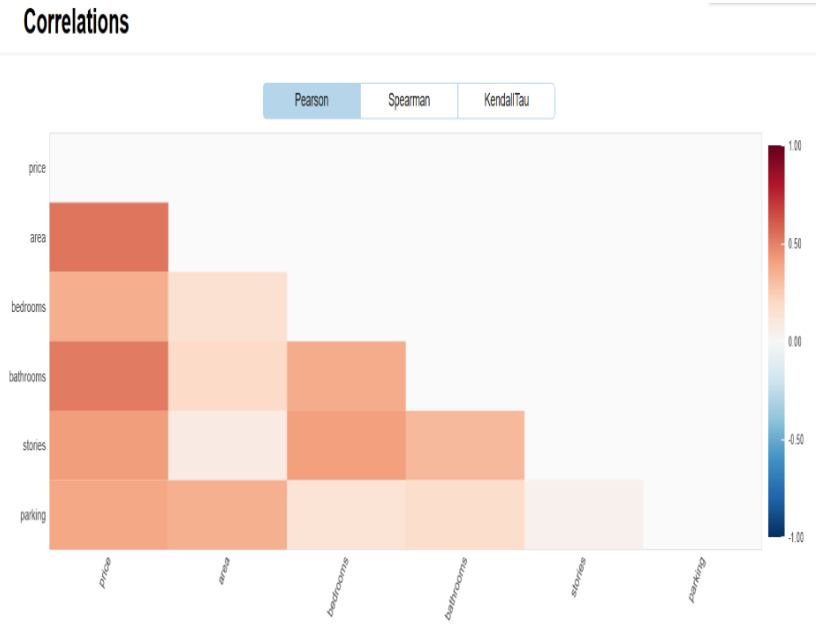
Exploratory Data Analysis (EDA)

Python was chosen as the programming language for this project due to its versatility and the availability of numerous packages for basic statistical analysis and building complex models. The following packages were loaded for data cleaning, preparation, building, and plotting the dataset: Numpy, Pandas, Sklearn, Scipy, Seaborn, dataprep, and Mathplotlib. We downloaded the dataprep library for the summary of the exploratory data analysis.

Dataset Statistics

Number of Variables	13
Number of Rows	545
Missing Cells	0
Missing Cells (%)	0.0%
Duplicate Rows	0
Duplicate Rows (%)	0.0%
Total Size in Memory	251.7 KB
Average Row Size in Memory	473.0 B
Variable Types	Numerical: 2 Categorical: 11

Correlations



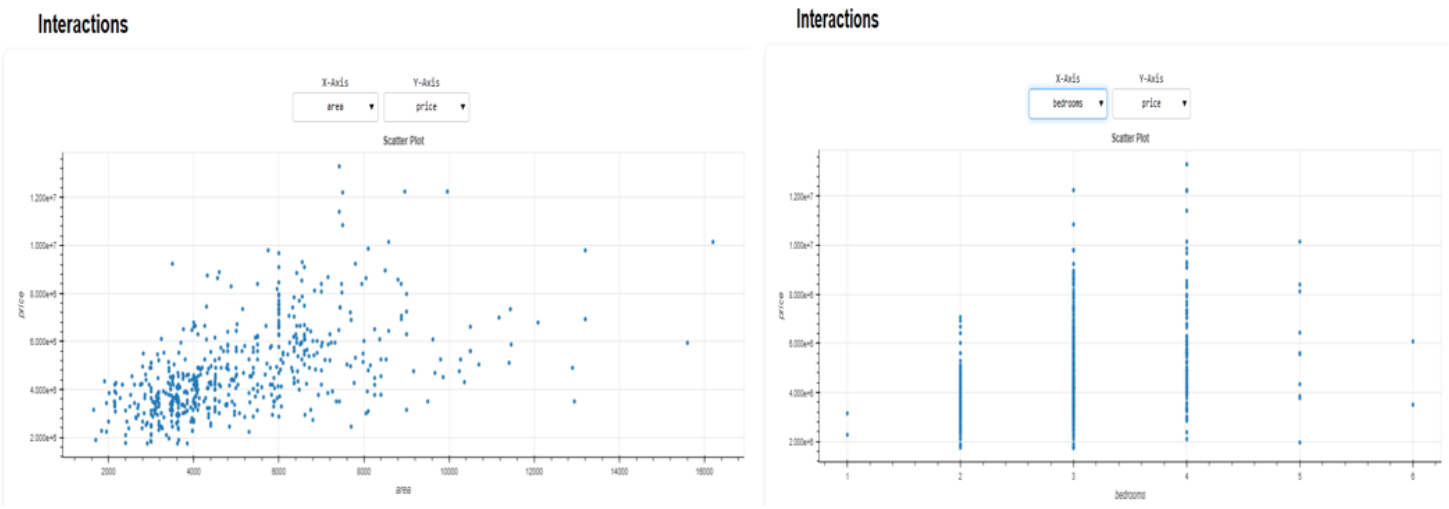


Fig 3. Exploratory data analysis

The Analysis of Main Factors Affecting Housing Price

Housing price is affected by multiple factors and features of a specific house. According to the previous research, some analysts have proposed several variables that significantly influence the overall housing price. House factors can be divided into several types. The most influential type is residential factors, including residence, usability, and number of rooms. When people consider purchasing a house for living purposes, the factors above are the main determinants for the living quality. Buyers with family members would typically attach more importance to the essential feature of the house, like the living area and number of rooms, which have a significant impact on the overall living quality and experience in the house. Besides, the intangible features, like the view of residence and usability, also have a rather considerable influence on the housing price, through affecting buyers' experience on the house and willingness to pay.

On the other hand, floor factors, like the number of stories, have also impacted the housing price significantly. Typically, household prefers the house with the number of the stories most suitable for their daily convenience. A family with children and elders tends to prefer a house with

multifloor construction, which offers different family members separate living areas with appropriate privacy while living together.

Univariate Analysis:

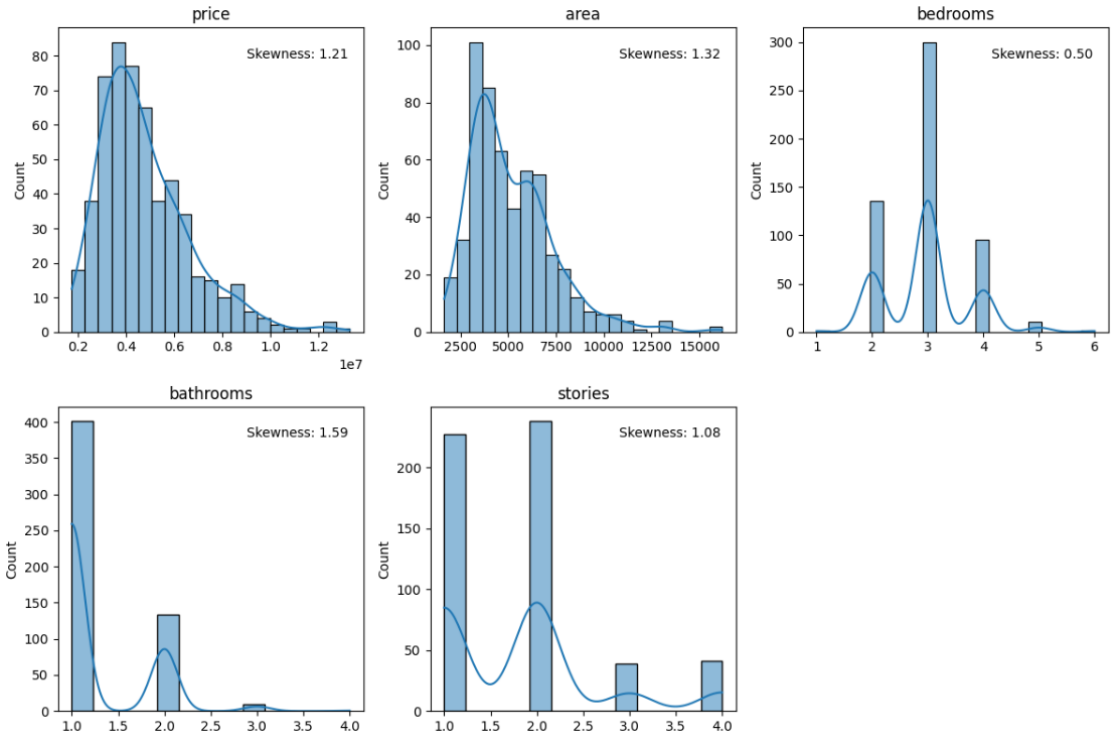
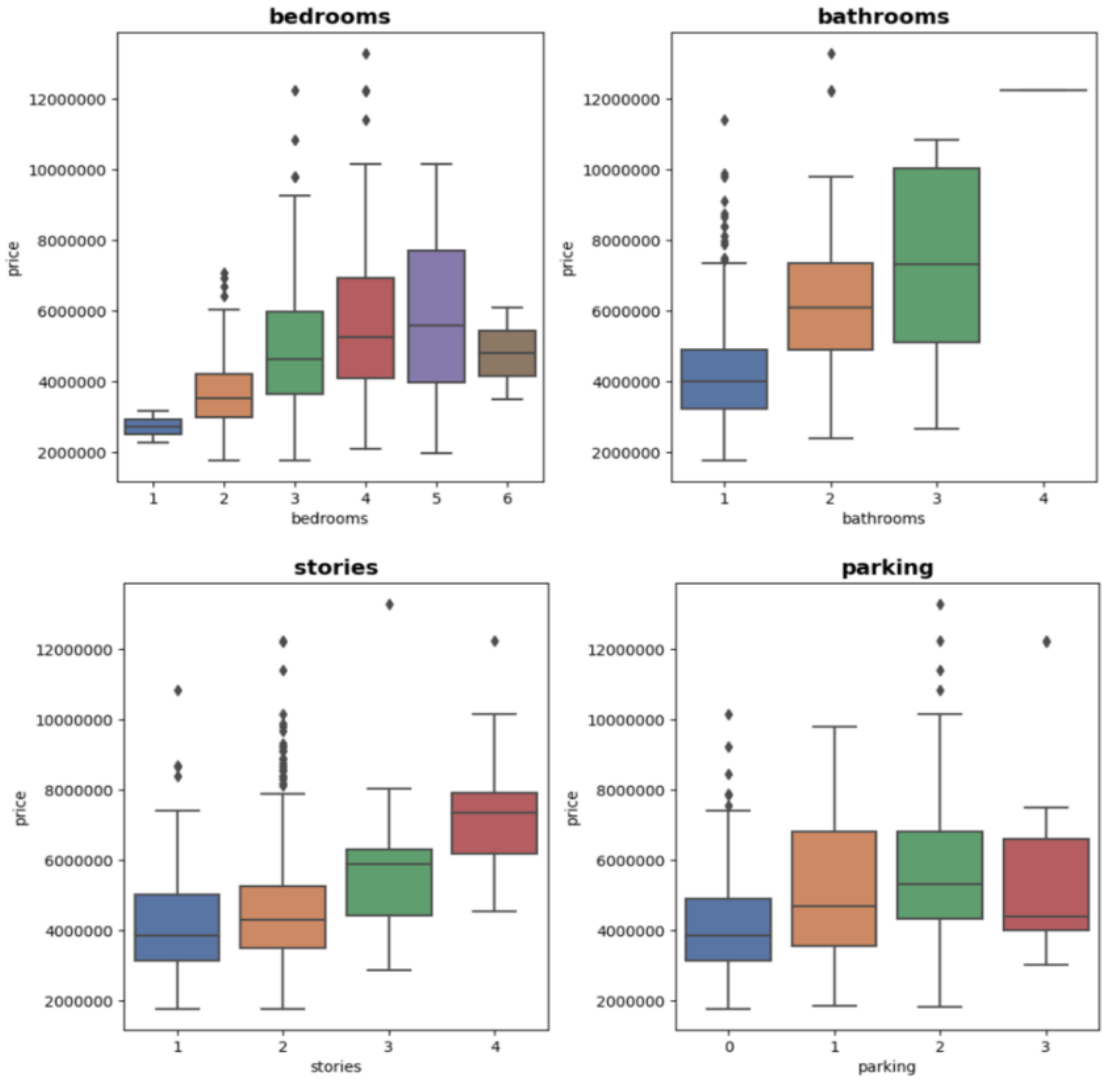


Fig 4. Univariate Analysis

We have used the seaborn library to plot histograms with kernel density estimation. It also calculates the skewness of each numerical attribute using the skew() function from pandas, and displays the skewness value on the plot. The skewness value indicates the degree of asymmetry in the distribution of the attribute. Positive skewness indicates a right-skewed distribution, negative skewness indicates a left-skewed distribution, and a skewness close to zero indicates a roughly symmetric distribution.

Bivariate Analysis



Toronto Metropolitan University

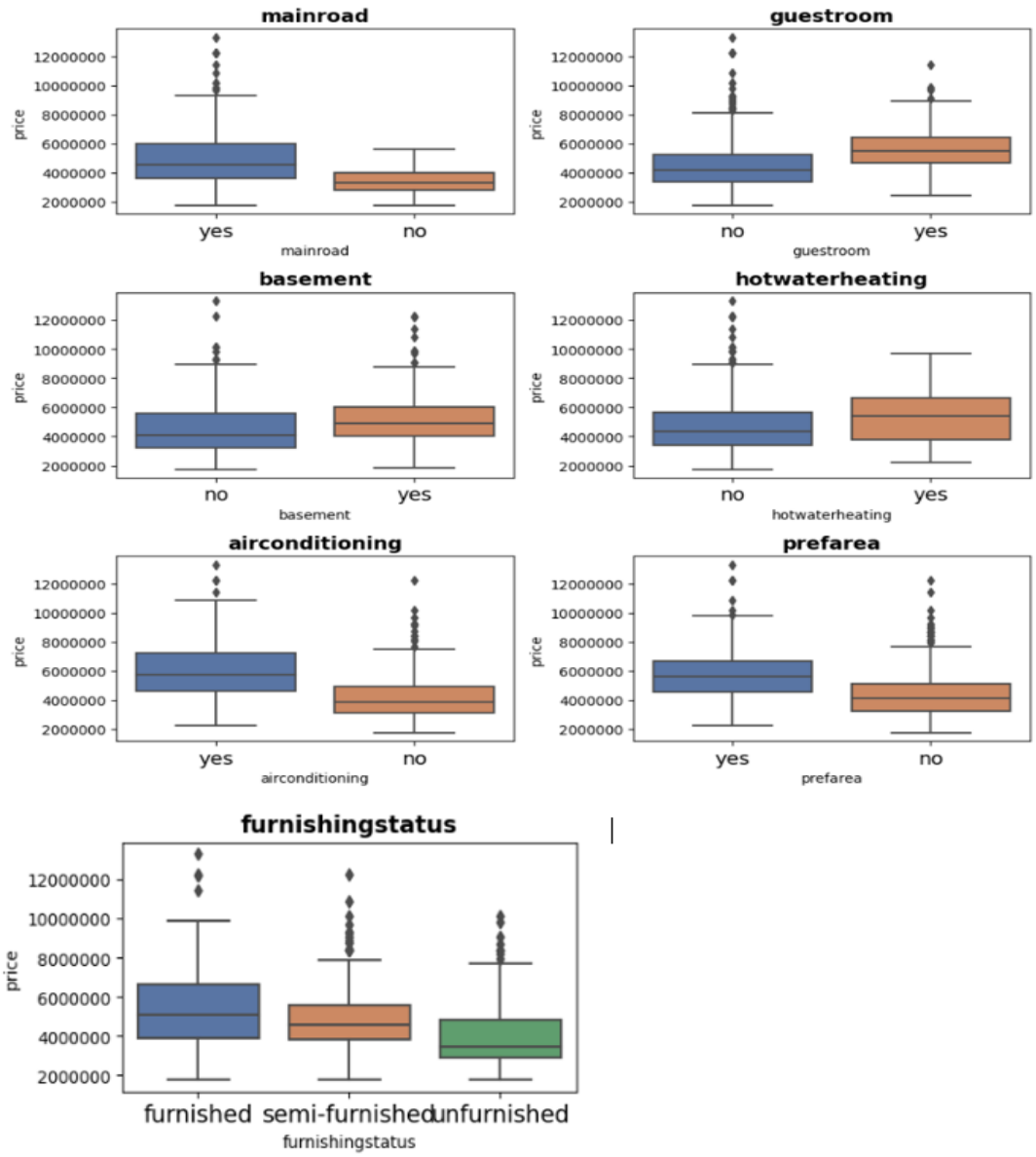


Fig 5. Bivariate Analysis

The box plot shows a visual representation of the central tendencies, range, median, outliers and variabilities of the attributes. Understanding these characteristics helps us in analyzing the relationship between the attributes and the target variable (Price).

Correlation Analysis

The correlation analysis reveals the following key relationships in the dataset:

Price is positively correlated with area (0.54), bedrooms (0.37), bathrooms (0.52), stories (0.42),

and parking (0.38). This suggests that larger houses with more bedrooms, bathrooms, stories, and parking spaces tend to have higher prices.

Area shows a positive correlation with bedrooms (0.15) and bathrooms (0.19), indicating that larger houses tend to have a slightly higher number of bedrooms and bathrooms.

Bedrooms exhibit a positive correlation with bathrooms (0.37) and stories (0.41), suggesting that houses with more bedrooms tend to have more bathrooms and stories.

Bathrooms show a positive correlation with stories (0.33), indicating that houses with more bathrooms tend to have more stories.

The correlation between stories and parking is weak (0.046), suggesting a minimal relationship between the number of stories and parking spaces.

These correlation coefficients provide valuable insights into the relationships between variables and can help understand the factors influencing housing prices. It indicates that house size, number of bedrooms and bathrooms, and availability of parking spaces are important factors to consider when predicting housing prices.

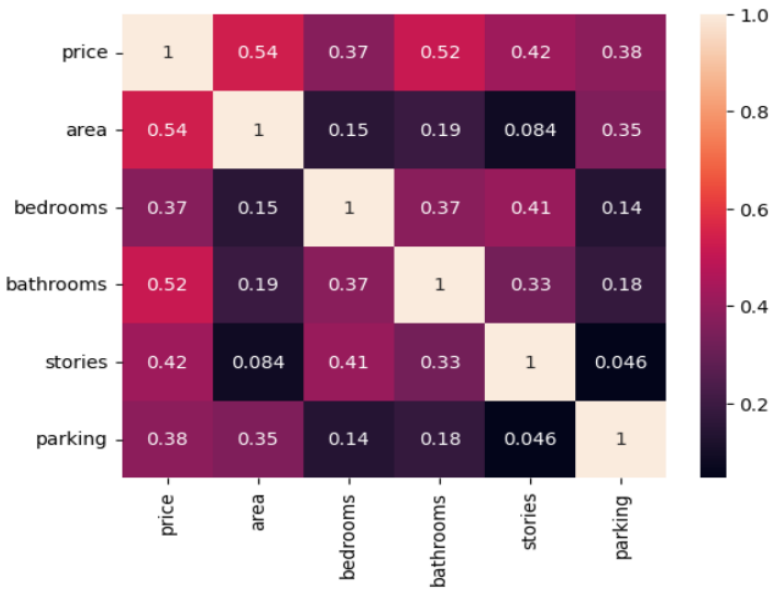


Fig 6. Correlation Analysis

References

1. Adetunji, A. B., Akande, O. N., Ajala, F. A., Oyewo, O., Akande, Y. F., & Oluwadara, G. (2022). House price prediction using Random Forest Machine Learning Technique. *Procedia Computer Science*, *199*, 806–813. https://doi.org/10.1016/j.procs.2022.01.100

2. Truong, Q., Nguyen, M., Dang, H., & Mei, B. (2020). Housing price prediction via Improved Machine Learning Techniques. *Procedia Computer Science*, *174*, 433–442. https://doi.org/10.1016/j.procs.2020.06.111

3. Zhang, Q. (2021). Housing price prediction based on multiple linear regression. *Scientific Programming*, *2021*, 1–9. https://doi.org/10.1155/2021/7678931

4. Lahmiri, S., Bekiros, S., & Avdoulas, C. (2023). A comparative assessment of machine learning methods for predicting housing prices using Bayesian optimization. *Decision Analytics Journal*, *6*, 100166. https://doi.org/10.1016/j.dajour.2023.100166

5. Thamarai, M., & Malarvizhi, S. P. (2020). House price prediction modeling using machine learning. *International Journal of Information Engineering and Electronic Business*, *12*(2), 15–20. https://doi.org/10.5815/ijieeb.2020.02.03

6. Ho, W. K. O., Tang, B.-S., & Wong, S. W. (2020). Predicting property prices with machine learning algorithms. *Journal of Property Research*, *38*(1), 48–70. https://doi.org/10.1080/09599916.2020.1832558

7. Dataset https://www.kaggle.com/datasets/yasserh/housing-prices-dataset

GitHub Link

https://github.com/rabiadanish/CIND-820-Project