Artificial Intelligence and Machine Learning

Project Report

Semester-IV (Batch-2022)

House Price Prediction

Supervised By:

Mr. Mohd Talib

Submitted By:

Tanishq, 2210992444(G-29)

Vanshika, 2210992505(G-29)

Tanya, 2210992452(G-29)

Tanya, 2210992453(G-29)

Department of Computer Science and Engineering
Chitkara University Institute of Engineering & Technology,
Chitkara University, Punjab

Abstract

Predicting house prices accurately is of paramount importance in the real estate market for both buyers and sellers. In this project, we employ a dataset comprising various features such as area, number of bedrooms, bathrooms, stories, and amenities like guestrooms, basements, and parking availability. Our goal is to develop a robust predictive model using machine learning techniques that can effectively estimate house prices based on these attributes.

We explore the relationships between different features and the target variable, price, utilizing regression analysis. Additionally, we implement feature engineering techniques to enhance model performance and address any missing or irrelevant data.

Through comprehensive experimentation and evaluation, we aim to provide insights into which factors most significantly influence house prices and deliver a reliable prediction tool that can assist real estate stakeholders in making informed decisions. This project contributes to the advancement of predictive modeling in real estate, offering practical applications in pricing strategy development and investment decision-making.

TABLE OF CONTENTS

1. Introduction

The housing market plays a crucial role in the global economy, with house prices significantly impacting individuals and families. Accurately predicting house prices can be a valuable tool for various stakeholders. For potential buyers, a reliable estimate can guide budgeting and inform informed decisions. Sellers can leverage such predictions to set competitive asking prices. Real estate professionals can utilize these models to enhance their market analysis and client guidance.

This project delves into the application of machine learning for house price prediction. We will explore a dataset encompassing a comprehensive range of housing attributes, including physical characteristics like area and number of bedrooms, locational factors like proximity to main roads and preferred areas, and amenities such as basements, air conditioning, and parking availability.

By analysing and harnessing the power of this data, we aim to:

- Identify the key factors that significantly influence house prices.
- Develop a robust machine learning model capable of accurately predicting house prices within the dataset.
- Evaluate the performance of different models to determine the most effective approach for this specific data.

This project contributes to the growing field of data-driven real estate analysis and offers valuable insights for navigating the complexities of the housing market.

2. Problem Statement

The housing market is dynamic and influenced by a multitude of factors. Accurately determining the fair market value of a house is a complex task that traditionally relies on experience, market trends, and subjective evaluations. While these methods provide valuable insights, they can be susceptible to bias and lack a comprehensive understanding of the underlying factors driving house prices.

This project addresses the challenge of developing a more objective and data-driven approach to house price prediction. We aim to leverage the power of machine learning to:

- **Reduce subjectivity:** By analysing a comprehensive dataset of housing characteristics, we can move beyond subjective evaluations and identify the key features that objectively impact house prices.
- **Improve prediction accuracy:** Machine learning models can learn complex relationships between various features and house prices, leading to more accurate predictions compared to traditional methods.
- **Increase transparency:** The interpretability of machine learning models can provide valuable insights into the relative importance of different features, allowing for a better understanding of what drives house prices in a specific market.

By tackling these challenges, this project seeks to develop a reliable machine learning model that can accurately predict house prices within the given dataset. This model can provide valuable information for potential buyers, sellers, and real estate professionals, ultimately contributing to a more informed and efficient housing market.

3. Tools Used

This project will utilize a combination of programming languages, libraries, and potentially cloud platforms to achieve its goals. Here's a breakdown of the anticipated tools:

Development Environment:

- Google Colab: This free Jupyter notebook environment offered by Google provides a platform for code execution and analysis within your web browser, eliminating the need for local installations.

Programming Language:

- Python: Python is a popular choice for data science tasks due to its readability, extensive libraries, and large community support.

Libraries:

- Pandas: This library excels in data manipulation and analysis, allowing us to explore, clean, and prepare the housing dataset.
- NumPy: NumPy provides efficient numerical computing capabilities, crucial for calculations and transformations within the data.
- Scikit-learn (sklearn): As the go-to library for machine learning in Python, scikit-learn offers a wide range of regression algorithms for model building and evaluation.
- Matplotlib/Seaborn**:** Data visualization plays a vital role. These libraries will help us create informative charts and graphs to understand relationships within the data and explore feature importance.

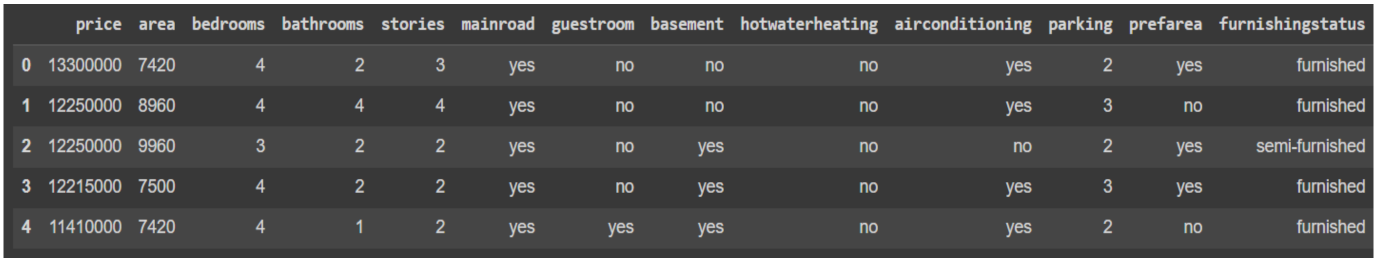
These tools will be the foundation for building and evaluating machine learning models for house price prediction within the project. The specific choice of tools might be subject to change based on the project's specific needs and your comfort level with different technologies.

4. Data Summary

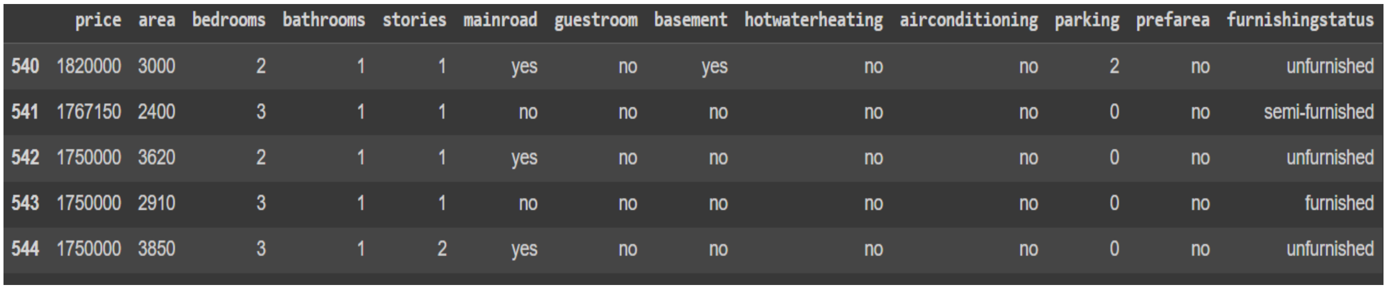
The dataset contains information about various housing features, likely intended for house price prediction. Here’s a breakdown of the columns:

- Price: This numerical data column represents the house price.
- Area: This numerical data column describes the total area of the house.
- Bedrooms: This numerical data column represents number of bedrooms in the house.
- Bathrooms: This numerical data column represents number of bathrooms in the house.
- Stories: This numerical data column represents number of stories in the house.
- Main road: It is categorical data which tells whether the house is located near a main road (yes/no).
- Guestroom: It is categorical data and it tells whether the house has a guest room (yes/no).
- Basement: It is categorical data and it tells if the house has a basement (yes/no).
- Hot-Water Heating: It is categorical data and it tells whether the house has hot water heating (yes/no).
- Air-Conditioning: It is categorical data and it tells whether the house has air conditioning (yes/no).
- Parking: This numerical data column represents number of parking lots.
- Pref Area: It is categorical data and it tells Whether the house is located in a preferred area (yes/no).
- Furnishing Status: It is categorical data and it tells the furnishing status of the house (furnished/semi-furnished/unfurnished).

This is the House Prices dataset. In the below table, it shows the top and bottom 5 rows respectively.



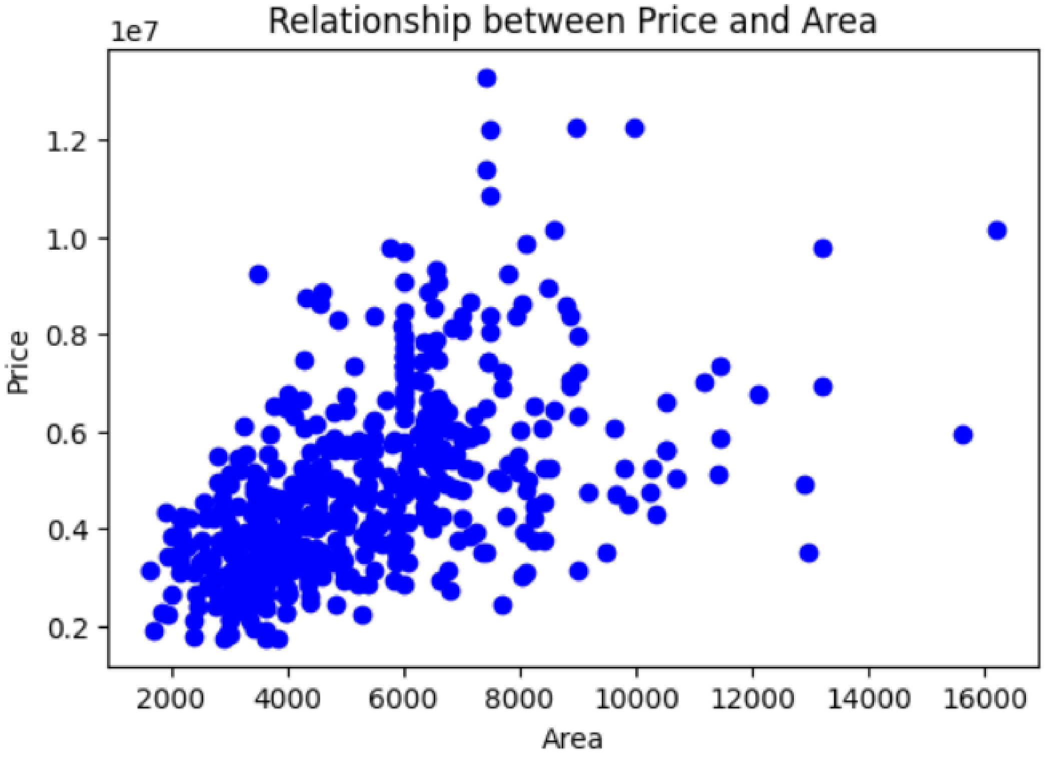
	price	area	bedrooms	bathrooms	stories	mainroad	guestroom	basement	hotwaterheating	airconditioning	parking	prefarea	furnishingstatus
0	13300000	7420	4	2	3	yes	no	no	no	yes	2	yes	furnished
1	12250000	8960	4	4	4	yes	no	no	no	yes	3	no	furnished
2	12250000	9960	3	2	2	yes	no	yes	no	no	2	yes	semi-furnished
3	12215000	7500	4	2	2	yes	no	yes	no	yes	3	yes	furnished
4	11410000	7420	4	1	2	yes	yes	yes	no	yes	2	no	furnished



	price	area	bedrooms	bathrooms	stories	mainroad	guestroom	basement	hotwaterheating	airconditioning	parking	prefarea	furnishingstatus
540	1820000	3000	2	1	1	yes	no	yes	no	no	2	no	unfurnished
541	1767150	2400	3	1	1	no	no	no	no	no	0	no	semi-furnished
542	1750000	3620	2	1	1	yes	no	no	no	no	0	no	unfurnished
543	1750000	2910	3	1	1	no	no	no	no	no	0	no	furnished
544	1750000	3850	3	1	2	yes	no	no	no	no	0	no	unfurnished

5. Visual Representations

5.1. Price vs Area



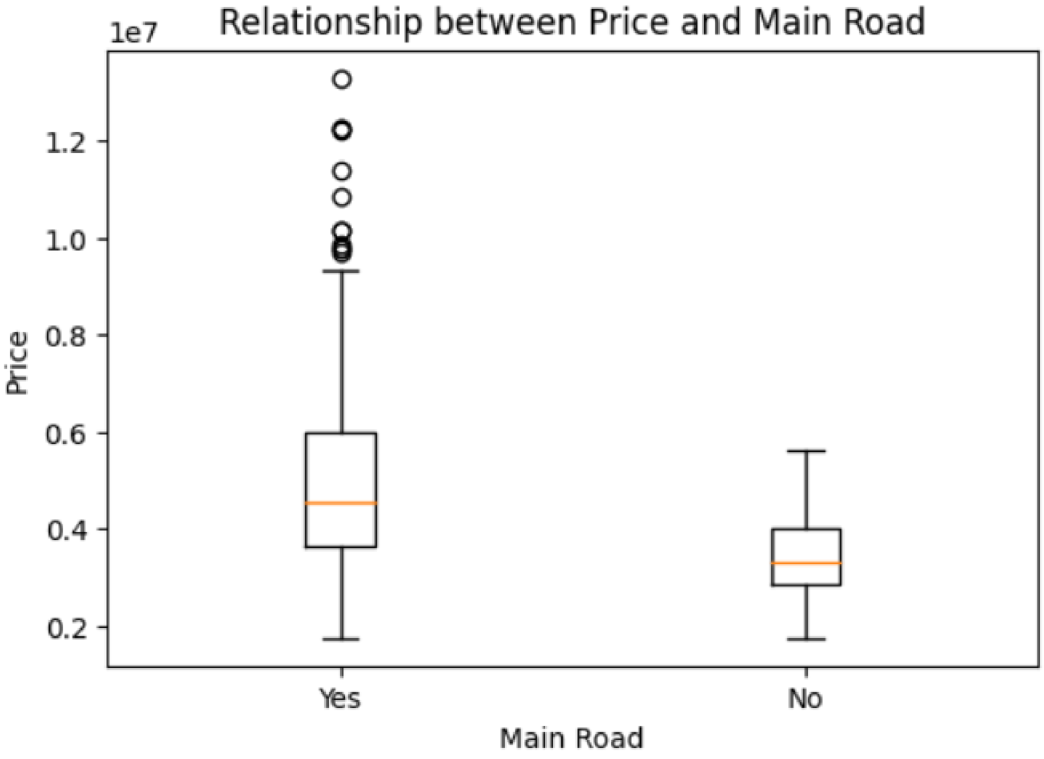
- I chose a scatter plot to visualize the relationship between price and area in the housing dataset because it is a suitable chart type for exploring the relationship between two numerical variables.

- There appears to be a positive correlation between price and area.

- The data points are somewhat spread out, indicating that there is some variation in price for houses of similar area.

- There are a few data points that appear to be outliers. These are houses with either a very high price or a very low price for their area.

5.2. Price vs Main Road



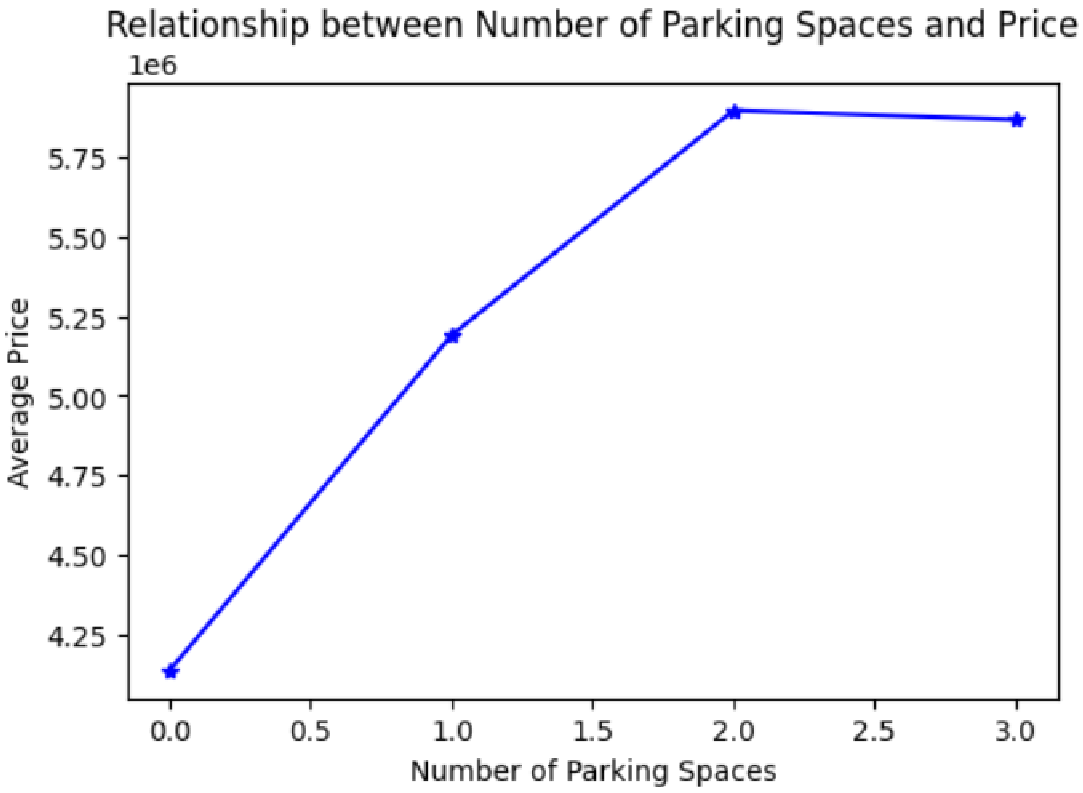
- I chose a boxplot to visualize the relationship between price and main road in the housing dataset because it is a suitable chart type for comparing the distributions of two groups.

- The median price of houses on the main road is higher than the median price of houses not on the main road.

- There are more outliers for houses on the main road.

- Real estate agents and property developers can use the insights to develop more effective pricing strategies such as they can charge a premium for houses that are located on the main road.

5.3. Price vs Parking



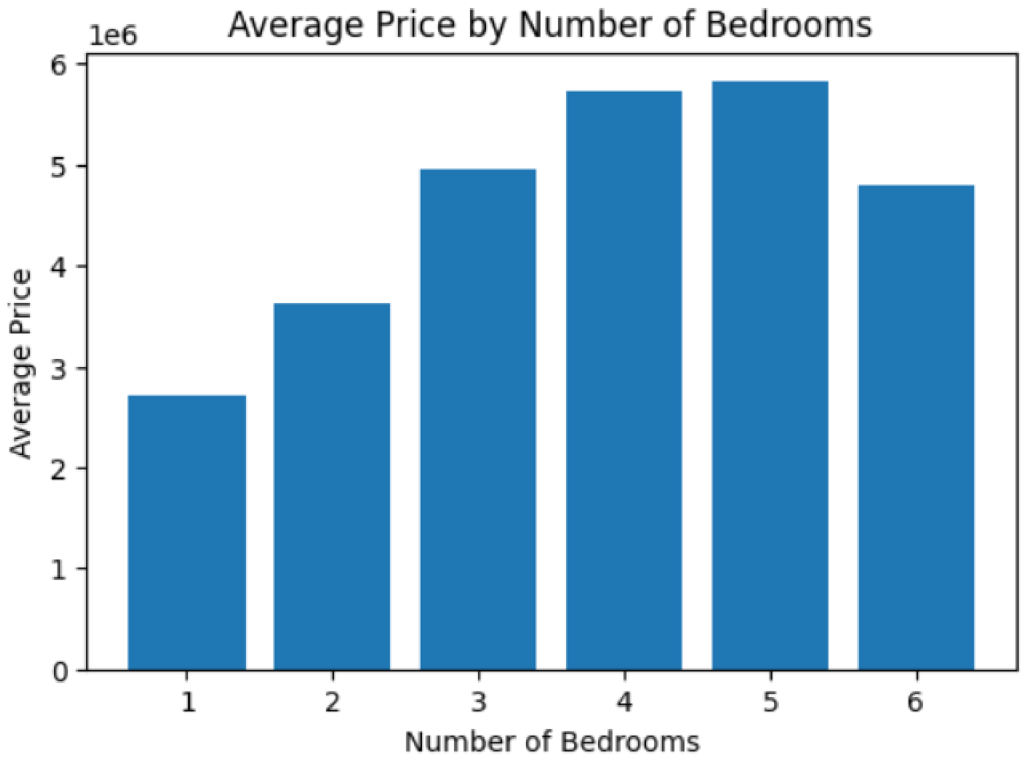
- A line chart to visualize the relationship between the number of parking spaces and the price of a house because it is a suitable chart type for showing trends and patterns in data.

- Line charts are also relatively easy to read and understand.

- There is a positive correlation between the number of parking spaces and the price of a house. This means that, on average, houses with more parking spaces tend to be more expensive than houses with fewer parking spaces.

- Real estate agents and property developers can use the insights to develop more targeted marketing and advertising campaigns.

5.4. Average Price vs Number of Bedrooms



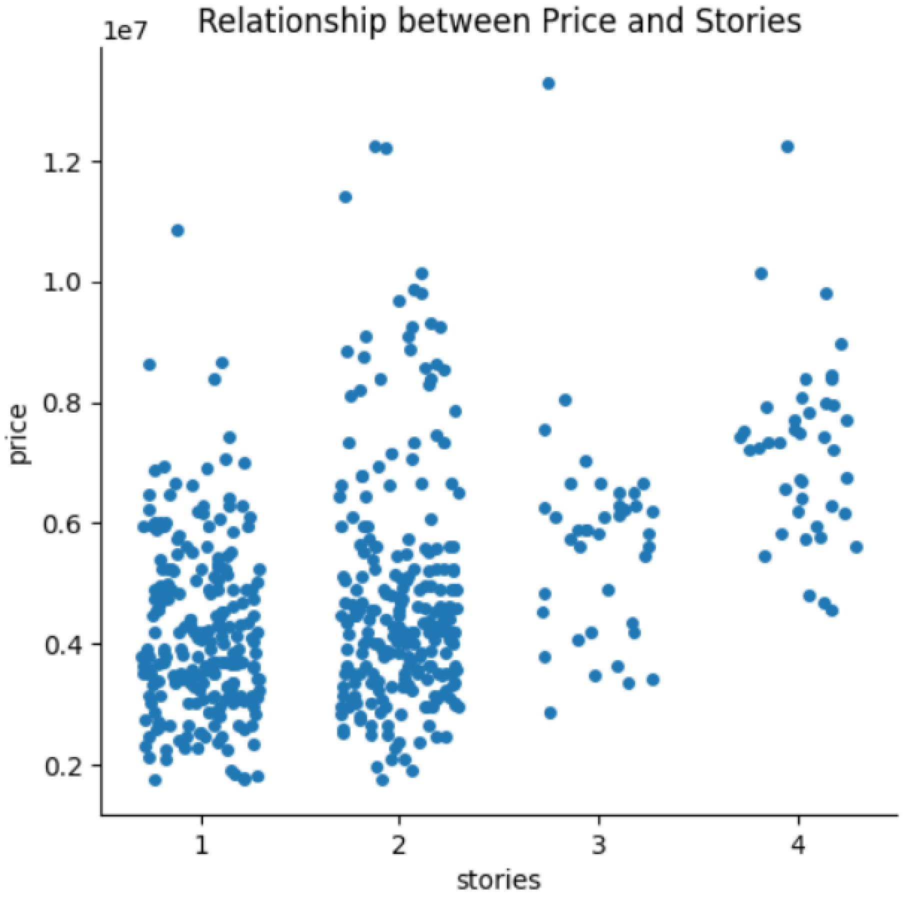
- I chose a bar chart to visualize the average price by number of bedrooms because it is an effective way to compare the average values across different categories.

- Houses with 1 bedroom have the lowest average price.

- Houses with 5 bedrooms have the highest average price among all categories.

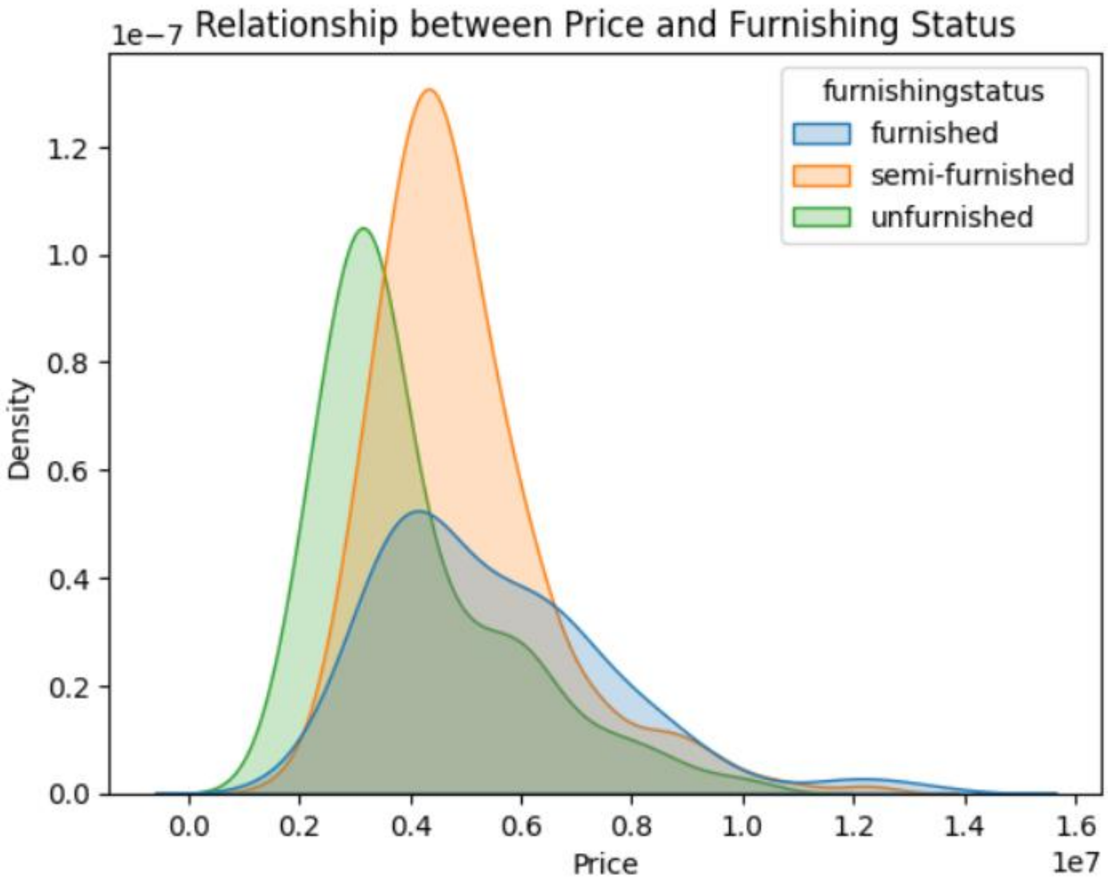
- One of the insights is that houses with 6 bedrooms have a similar average price to houses with 3 bedrooms could be interpreted in a way that there is a limited market for houses with 6 bedrooms, resulting in lower prices for these properties.

5.5. Price vs Stories



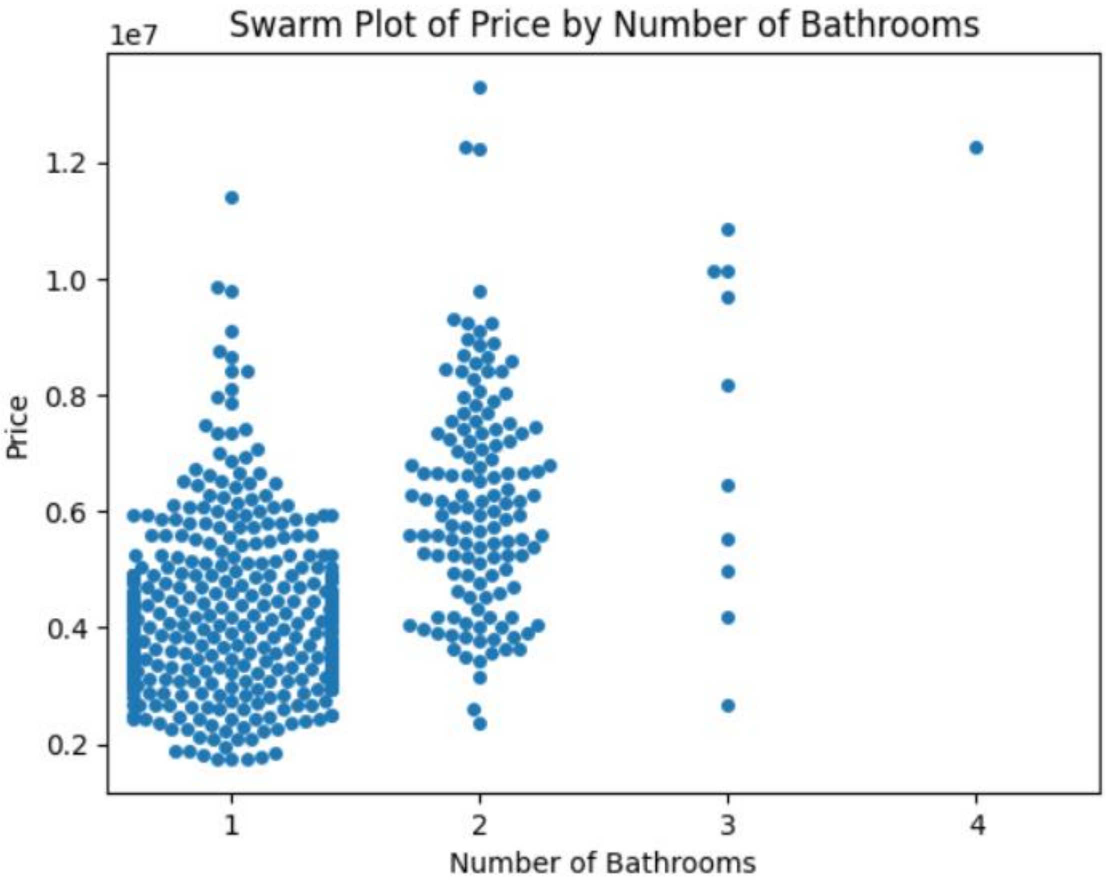
- I picked a cat plot to visualize the relationship between price and stories because it is a good way to see the distribution of the data for each number of stories.
- The median price of houses increases as the number of stories increases.
- There are a few outliers, which are houses with a high price for their number of stories.
- There is a wider distribution of prices for houses with more stories.

5.6. Price vs Furnishing Status



- I picked the KDE (kernel density) plot because it is a good way to visualize the distribution of a variable, especially when there are multiple categories.
- The distribution of prices for fully furnished houses is skewed to the right.
- The distribution of prices for not furnished houses is more evenly spread out which suggests that there is a wider range of prices for not furnished houses, with both expensive and inexpensive options available.

5.7. Price vs Number of Bathrooms



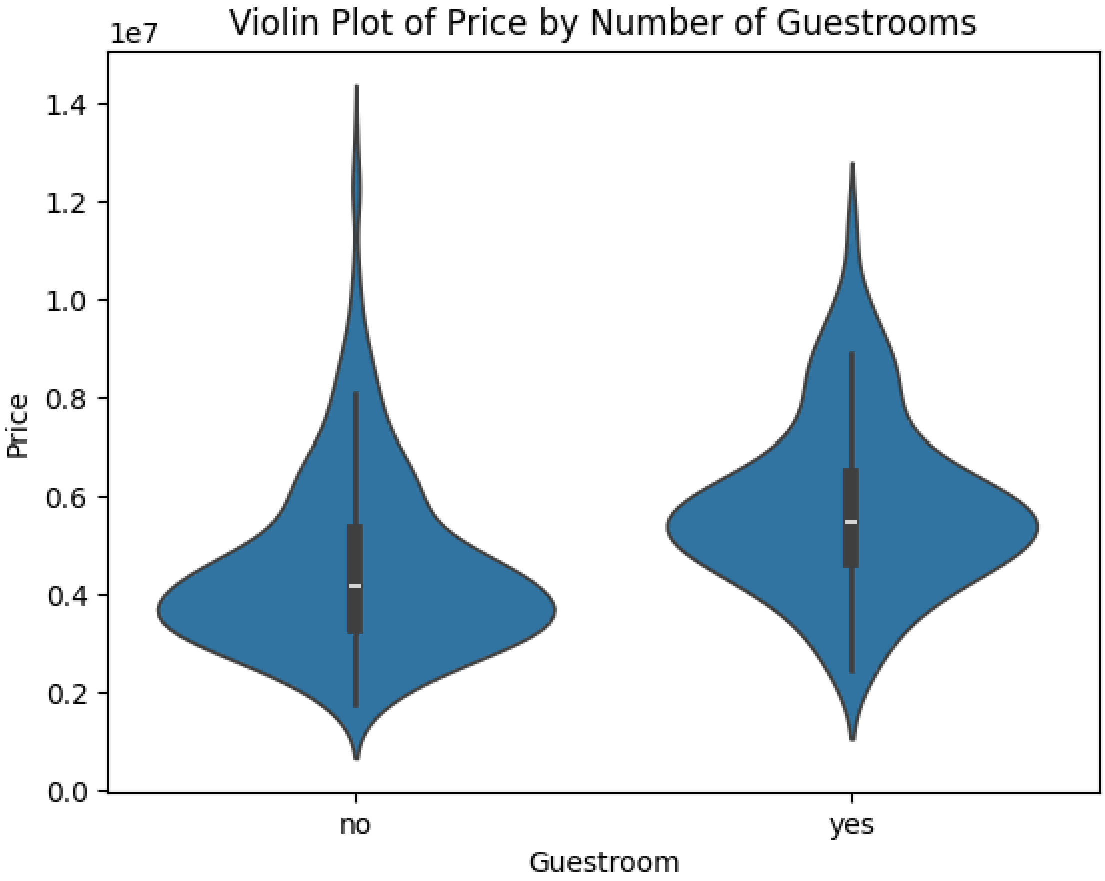
- I picked the swarm plot because it is a good way to visualize the distribution of prices differs for each number of bathrooms.

- There is a general trend of increasing price as the number of bathrooms increases.

- There is a lot of variation in price for houses with the same number of bathrooms. This suggests that there are other factors, besides the number of bathrooms, that affect the price of a house.

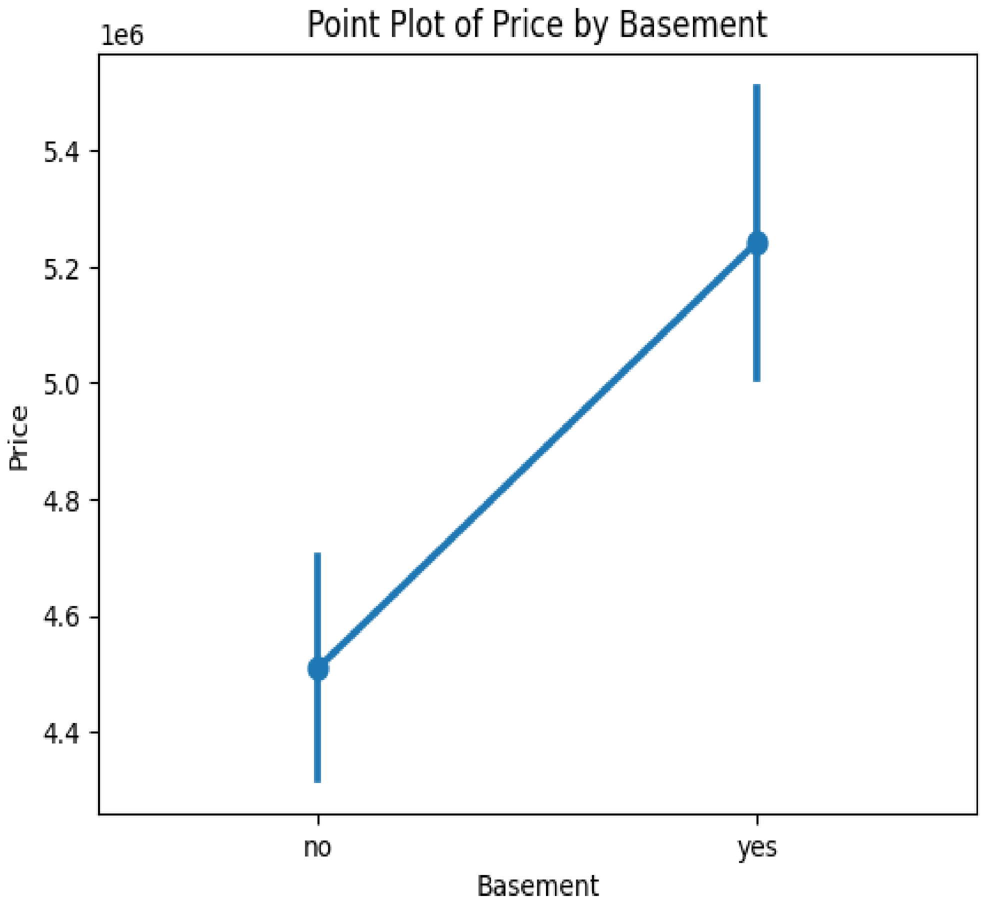
- There are a few houses with a high number of bathrooms and a high price. These houses may be considered luxury properties.

5.8. Price vs Guestroom



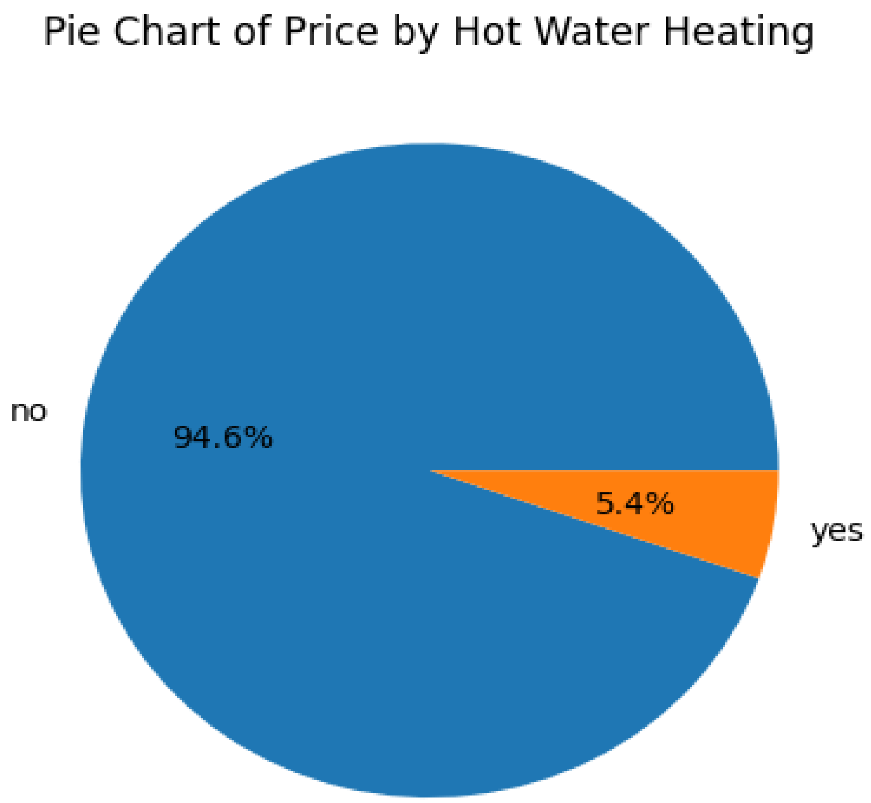
- I picked the violin plot because it is a good way to visualize the distribution of a variable, especially when there are multiple categories
- These insights could be useful for buyers and sellers of houses. Sellers may want to consider the number of guestrooms in their house when setting a price.
- Buyers can use the insights to make informed decisions about how many guestrooms they want in their house.

5.9. Price vs Basement



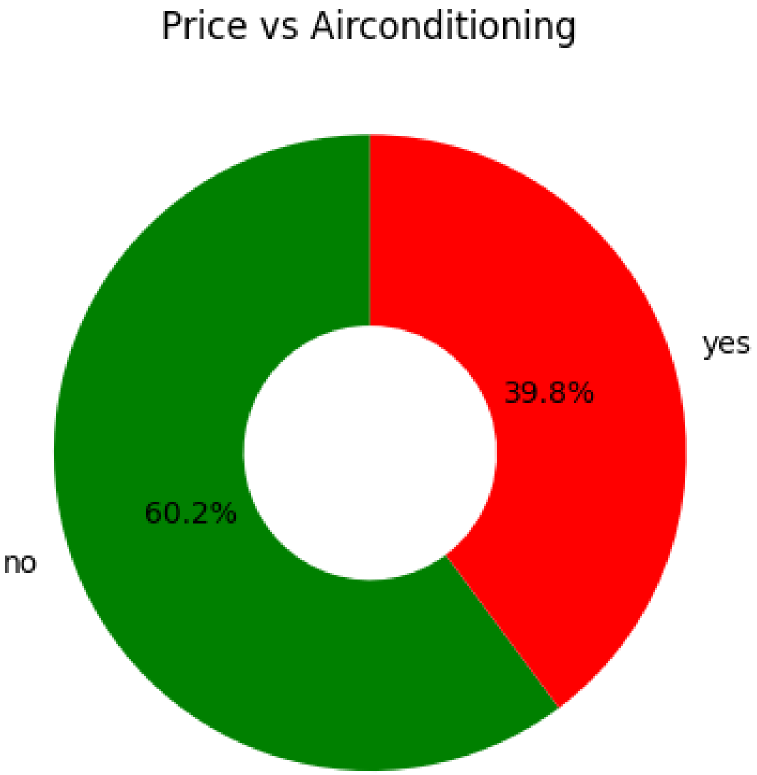
- I picked the point plot because it is a good way to visualize the relationship between two variables, especially when one variable is categorical
- Houses with basements tend to be more expensive than houses without basements. This suggests that buyers are willing to pay more for houses with basements.
- Sellers can use the insights to target the right buyers for their houses.

5.10. Price vs Hot-water Heating



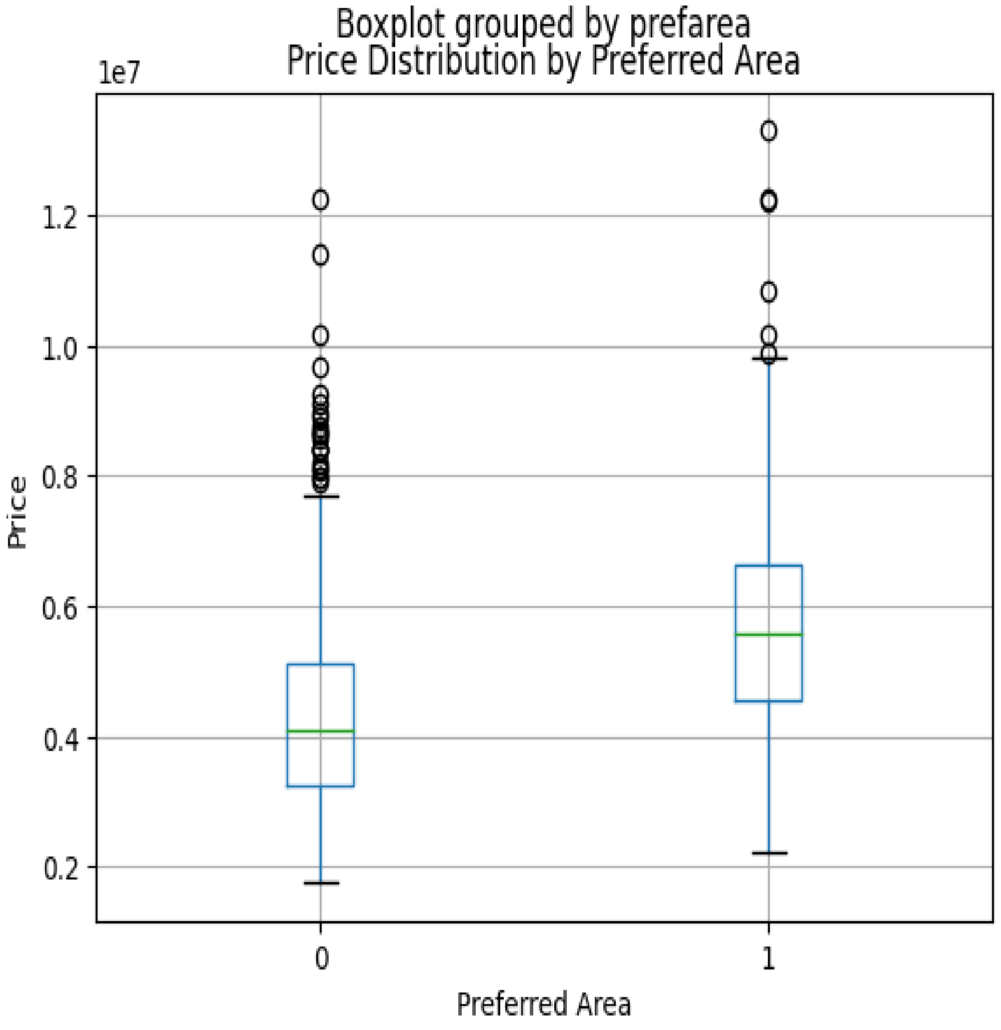
- I picked the pie chart because it is a good way to visualize the proportion of a whole that is made up by each of its parts.
- Houses with hot water heating make up a significant proportion of the total price of houses. This suggests that buyers are willing to pay more for houses with hot water heating.
- Sellers can use the insights to target the right buyers for their houses. For example, sellers of houses with hot water heating may want to target buyers who are looking for energy efficient homes

5.11. Price vs Airconditioning



- I picked the pie-chart(donut chart) because it is a good way to visualize the proportion of different categories within a dataset.
- Houses with central air conditioning are more expensive than houses with no air conditioning.
- The difference in price between houses with and without air conditioning is substantial.

5.12. Price vs Preferred Area

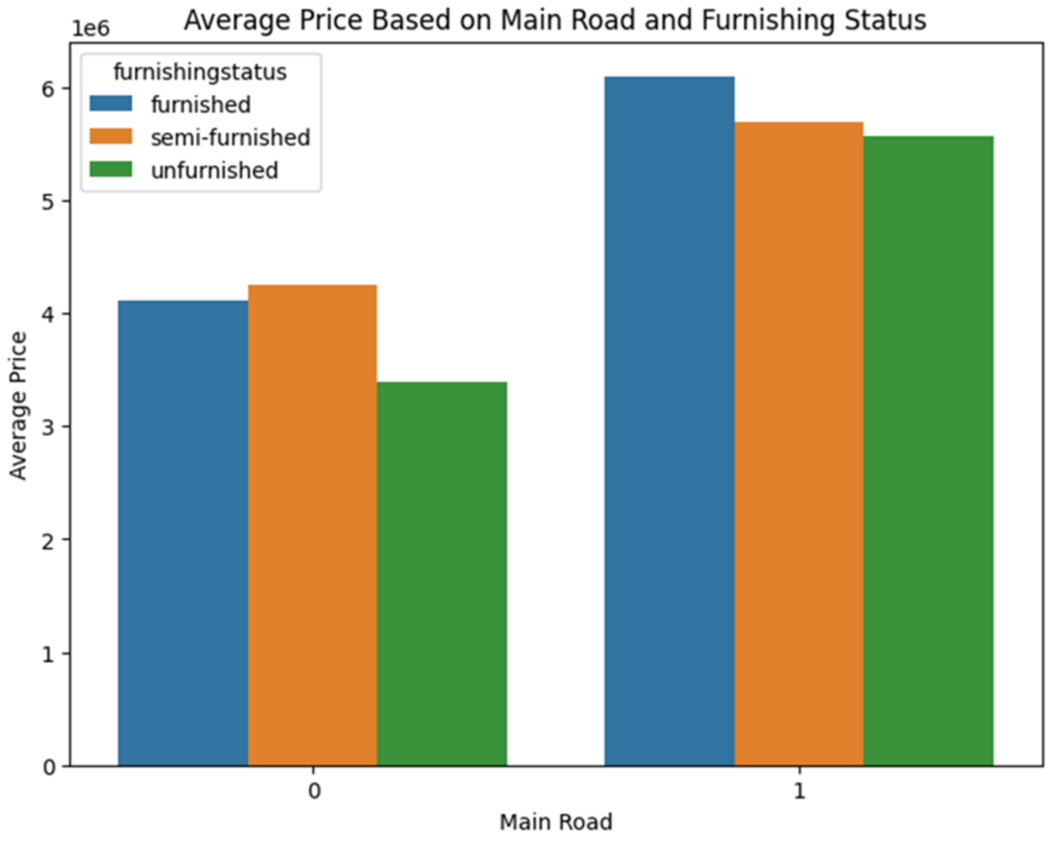


- I picked the boxplot to visualize the relationship between preferred area and price in the housing dataset because a boxplot shows the distribution of the data, including the median, quartiles, and outliers.

- The median price for houses in preferred areas is higher than for houses in non-preferred areas.

- Focusing solely on preferred areas may limit market reach, increase competition, and contribute to social issues, potentially hindering long-term growth and sustainability.

5.13. Average Price Based on Main Road and Furnishing Status

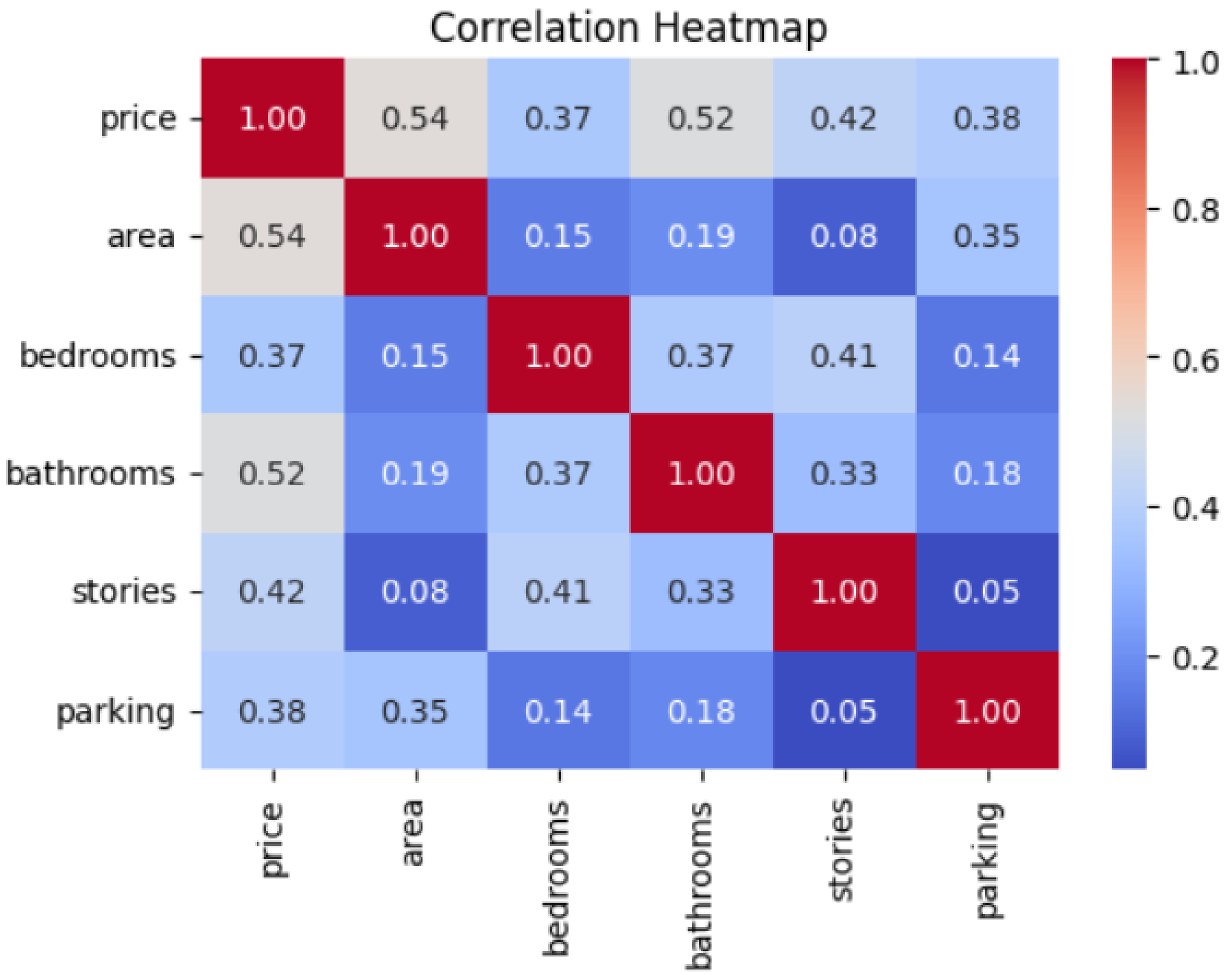


- Grouped bar plots are effective when you want to compare the average prices across multiple categories simultaneously.

- Houses located on the main road tend to have higher average prices compared to those not on the main road, regardless of furnishing status.

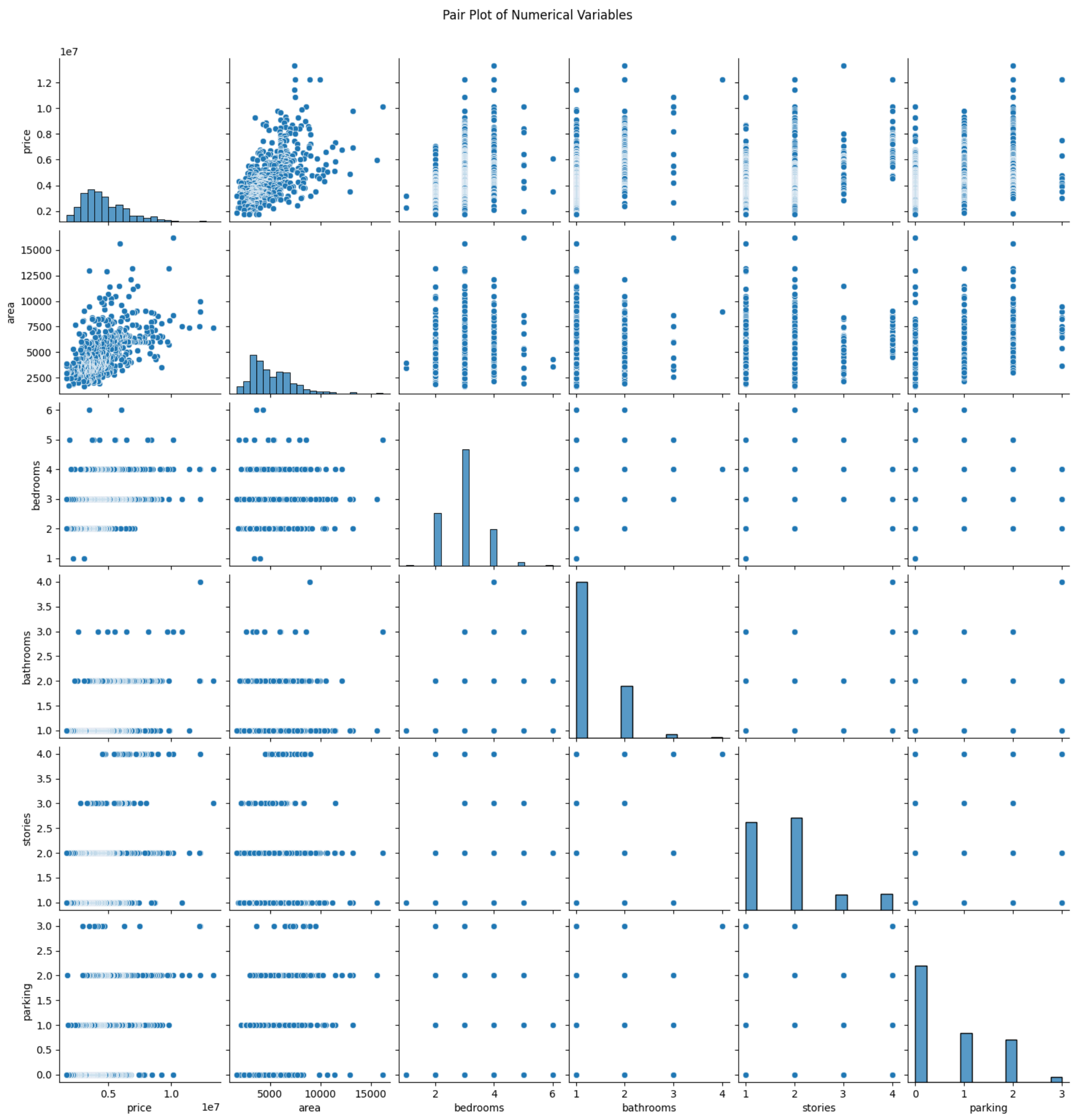
- Investors can prioritize properties that have a higher likelihood of commanding higher prices and generating better returns on investment.

5.14. Correlation Heatmap



- A heatmap effectively displays a large amount of information in a compact and visually appealing manner.
- Price has strong positive correlations with area, bathrooms, and stories.
- Price has a moderate positive correlation with parking
- Overall, heatmaps offer a powerful visualization tool for understanding the relationship between various property attributes and their influence on price, enabling stakeholders to make informed decisions in the real estate market.

5.15. Pair Plot



- A pair plot allows you to visualize all pairwise relationships between features in a single plot.
- There appears to be a positive correlation between price and area.
- Number of stories and parking spaces have a lesser impact on the price compared to area, bedrooms, and bathrooms.

6. Challenges

Feature Selection: It is crucial to identify and prioritize features that have a substantial impact on house prices while eliminating irrelevant or redundant ones. This process involves conducting thorough analyses to understand the correlation between different features and the target variable (house prices). Techniques such as correlation analysis, feature importance ranking, and domain expertise are often employed to guide feature selection. By focusing on the most influential features, models can achieve better predictive accuracy and generalization to unseen data.

Data Quality: The quality of the dataset used for training directly influences the performance and reliability of the predictive model. Ensuring comprehensive coverage of relevant features, accuracy in data collection, and representation of diverse segments of the housing market are essential considerations. Data preprocessing steps, including data cleaning, outlier detection, and imputation of missing values, are employed to enhance data quality and reduce biases that could skew model predictions.

Scaling: Handling large volumes of data efficiently is a common challenge in real estate analytics, particularly with the increasing availability of data from various sources such as property listings, demographic information, and economic indicators. Scaling models to accommodate the size of the dataset and computational resources available requires careful consideration of algorithmic efficiency, distributed computing techniques, and hardware acceleration. Implementing scalable solutions ensures timely analysis and modeling of large datasets without compromising performance.

Overfitting: Overfitting occurs when a model captures noise or random fluctuations in the training data, leading to poor generalization performance on unseen data. Techniques such as regularization, cross-validation, and model selection help prevent overfitting by penalizing overly complex models and evaluating their performance on validation datasets. Balancing model complexity with generalization capability is essential to build robust predictive models that accurately capture underlying patterns in the data while avoiding overfitting.

7. Hypothesis Testing:

Hypothesis testing is a statistical method used to make inferences about a population based on sample data. It involves testing a hypothesis about the characteristics of a population parameter using sample evidence.

Hypothetical Statement - 1:

H1: The area of a house is positively correlated with its price.

Null Hypothesis (H0): There is no significant difference in the average price of houses based on their preferred area.

Alternative Hypothesis (H1): There is a significant difference in the average price of houses based on their preferred area.

Statistical Test: Pearson correlation test

Conclusion: We reject the null hypothesis. There is a significant positive correlation between the area of a house and its price.

Hypothetical Statement - 2:

H1: Houses with more bedrooms tend to have higher prices.

Null Hypothesis (H0): There is no significant association between the number of bedrooms in a house and its price.

Alternative Hypothesis (H1): There is a significant positive association between the number of bedrooms in a house and its price.

Statistical Test: Linear regression analysis

Conclusion: We reject the null hypothesis. There is a significant positive association between the number of bedrooms in a house and its price.

Hypothetical Statement - 3:

H2: There is a significant difference in prices between houses with and without parking.

Null Hypothesis (H0): There is no significant difference in the average price of houses with and without parking.

Alternative Hypothesis (H1): There is a significant difference in the average price of houses with and without parking.

Statistical Test: t-test or ANOVA for comparing means between two groups (with and without parking)

Conclusion: Perform the appropriate statistical test to obtain the conclusion.

8. Feature Engineering

Feature Selection:

The selection of relevant features is essential for capturing the factors that influence house prices. Features such as area, bedrooms, bathrooms, location, stories, parking, and additional amenities are typically considered.

Features are chosen based on their correlation with the target variable (house prices) and domain knowledge of the real estate market.

Handling Missing Values:

Missing values in the dataset need to be addressed to ensure the completeness of the data. Imputation techniques like mean, median, or mode are used to fill missing numerical values such as bedrooms and bathrooms.

For categorical features like air conditioning, missing values can be handled by creating a new category or using advanced imputation methods like KNN imputation.

Encoding Categorical Variables:

Categorical variables are converted into a numerical format suitable for modeling. One-hot encoding is used for categorical variables with multiple levels, while label encoding is applied to ordinal variables with a natural order.

This transformation ensures that categorical information is appropriately represented in the model.

Feature Scaling:

Numerical features are scaled to ensure uniformity in their magnitude. Techniques like StandardScaler or Min-Max Scaler are applied to scale numerical features such as area, bedrooms, and bathrooms to a similar range.

Scaling prevents features with larger magnitudes from dominating the model training process, leading to more stable and accurate predictions.

Feature Transformation:

Feature transformation techniques are used to capture non-linear relationships and improve model performance. Polynomial features are introduced to capture higher-order interactions between variables, such as area squared or bedrooms multiplied by bathrooms.

Log transformation is applied to skewed numerical features to achieve a more Gaussian distribution, enhancing the performance of models that assume normality.

Feature Engineering Techniques:

New features are created to better capture the underlying relationships in the data. Feature interaction involves combining existing features to create new meaningful interactions, such as total rooms calculated as the sum of bedrooms and bathrooms.

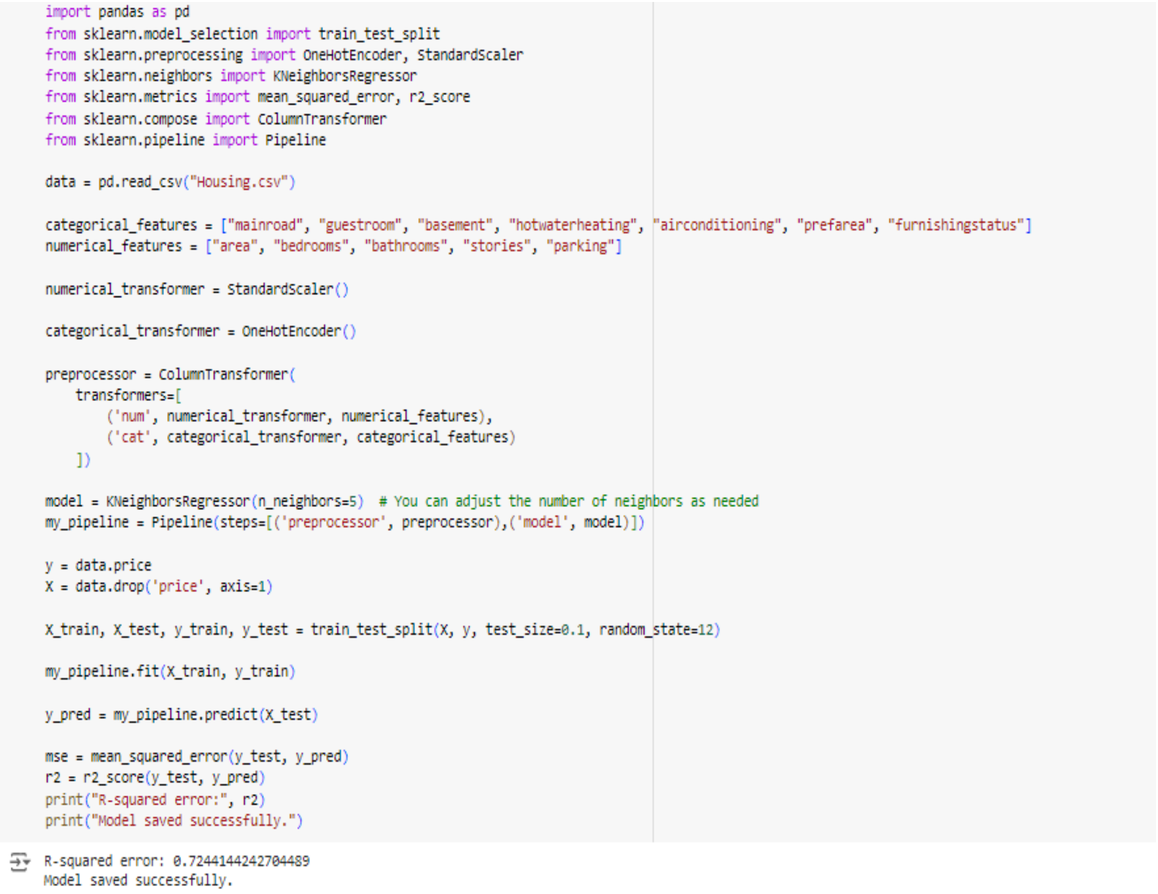
Feature extraction techniques are used to extract relevant information from existing features. For example, creating a binary variable for the presence of a basement.

Domain-specific features are engineered based on domain knowledge, such as proximity to landmarks or amenities, which can provide valuable insights into the factors driving house prices in specific markets.

Effective feature engineering enhances the predictive performance of models and uncovers valuable insights into the factors influencing house prices. By carefully engineering and selecting features, accurate and interpretable models can be built for house price prediction tasks, facilitating informed decision-making in real estate markets.

9. ML Model Implementation

ML Model 1:



```
import pandas as pd
from sklearn.model_selection import train_test_split
from sklearn.preprocessing import OneHotEncoder, StandardScaler
from sklearn.neighbors import KNeighborsRegressor
from sklearn.metrics import mean_squared_error, r2_score
from sklearn.compose import ColumnTransformer
from sklearn.pipeline import Pipeline

data = pd.read_csv("Housing.csv")

categorical_features = ["mainroad", "guestroom", "basement", "hotwaterheating", "airconditioning", "prefarea", "furnishingstatus"]
numerical_features = ["area", "bedrooms", "bathrooms", "stories", "parking"]

numerical_transformer = StandardScaler()

categorical_transformer = OneHotEncoder()

preprocessor = ColumnTransformer(
    transformers=[
        ('num', numerical_transformer, numerical_features),
        ('cat', categorical_transformer, categorical_features)
    ])

model = KNeighborsRegressor(n_neighbors=5)  # You can adjust the number of neighbors as needed
my_pipeline = Pipeline(steps=[('preprocessor', preprocessor),('model', model)])

y = data.price
X = data.drop('price', axis=1)

X_train, X_test, y_train, y_test = train_test_split(X, y, test_size=0.1, random_state=12)

my_pipeline.fit(X_train, y_train)

y_pred = my_pipeline.predict(X_test)

mse = mean_squared_error(y_test, y_pred)
r2 = r2_score(y_test, y_pred)
print("R-squared error:", r2)
print("Model saved successfully.")
```

```
R-squared error: 0.7244144242704489
Model saved successfully.
```

Model Explanation:

Data Preprocessing: The data is pre-processed using a Column Transformer that applies Standard Scaler to numerical features and One Hot Encoder to categorical features. This ensures that all features are on the same scale and categorical variables are appropriately encoded for the model.

Model Building: The pre-processed data is then passed through a K Neighbors Regressor model. KNN Regressor predicts the target value of an instance by averaging the target values of its k nearest neighbours. Here, n_neighbors is set to 5, meaning it considers 5 nearest neighbors to make predictions.

Pipeline Construction: The Pipeline combines the preprocessing steps and the model into a single object, making it easier to use and deploy.

Performance Evaluation:

For evaluating the performance of the model, two metrics are used:

Mean Squared Error (MSE): It measures the average of the squares of errors, which is the average squared difference between the predicted values and the actual values. Lower values indicate better fit.

R-squared (R2) Score: It measures the proportion of the variance in the dependent variable that is predictable from the independent variables. Higher values (closer to 1) indicate a better fit.

Evaluation Metrics Consideration:

Business Impact: The choice of evaluation metrics should align with the business objectives. In real estate pricing prediction, accuracy in predicting prices is crucial for customer satisfaction and optimizing revenue. Therefore, metrics like MSE and R-squared are suitable as they directly measure the accuracy of predictions.

ML Model Selection:

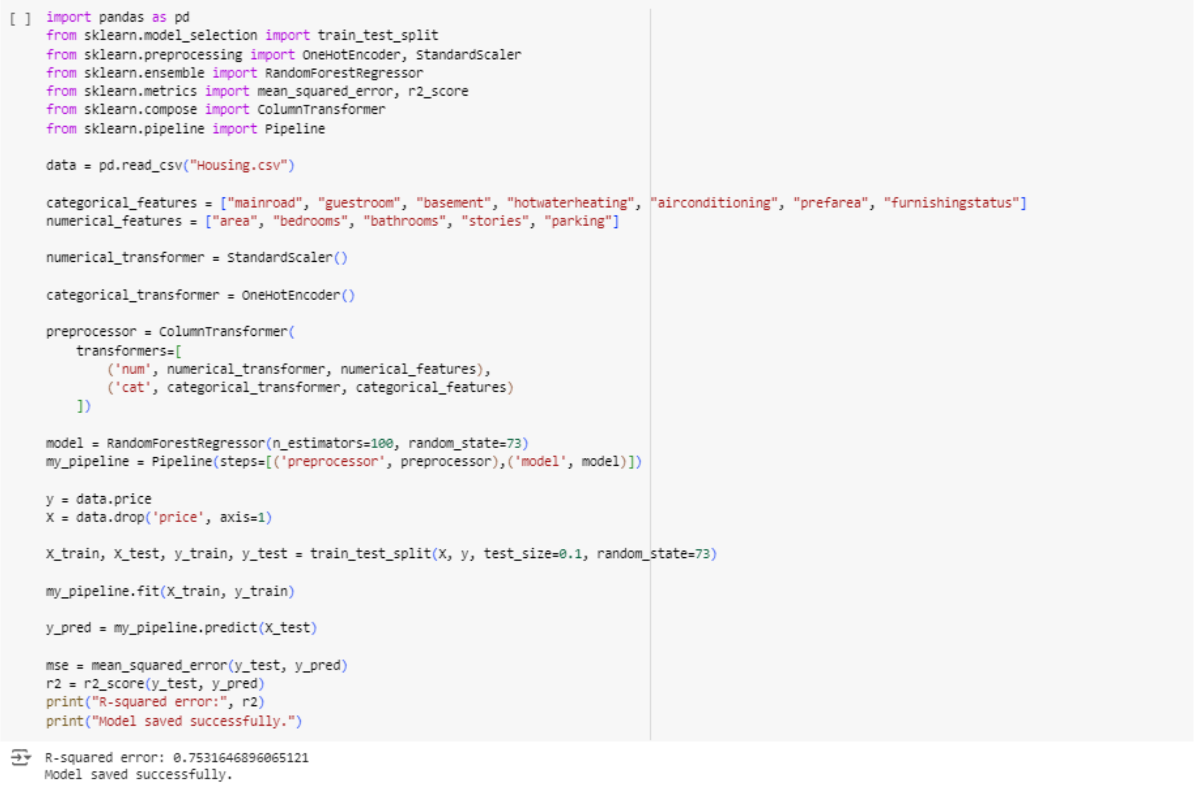
The KNN Regressor is chosen as the final prediction model. KNN is a simple yet effective model for regression tasks, especially when there's no assumption of linearity in the data. It's also non-parametric, meaning it does not assume any underlying data distribution. KNN can capture complex patterns in the data and is easy to understand and implement.

Feature Importance:

KNN does not inherently provide feature importance like some other models such as decision trees or linear models. However, feature importance can still be inferred by observing the weights assigned during the distance calculation. Features with higher weights contribute more to the prediction. For, a more detailed analysis of feature importance, you might consider using techniques like permutation importance or SHAP (SHapley Additive exPlanations) values. These methods provide insights into how much each feature contributes to the model's predictions.

Overall, the KNN Regressor provides a simple yet effective approach for predicting housing prices based on the given features, and its performance can be further analysed and optimized using various evaluation metrics and model explainability tools.

ML Model 2:



```
[ ] import pandas as pd
    from sklearn.model_selection import train_test_split
    from sklearn.preprocessing import OneHotEncoder, StandardScaler
    from sklearn.ensemble import RandomForestRegressor
    from sklearn.metrics import mean_squared_error, r2_score
    from sklearn.compose import ColumnTransformer
    from sklearn.pipeline import Pipeline

    data = pd.read_csv("Housing.csv")

    categorical_features = ["mainroad", "guestroom", "basement", "hotwaterheating", "airconditioning", "prefarea", "furnishingstatus"]
    numerical_features = ["area", "bedrooms", "bathrooms", "stories", "parking"]

    numerical_transformer = StandardScaler()

    categorical_transformer = OneHotEncoder()

    preprocessor = ColumnTransformer(
        transformers=[
            ('num', numerical_transformer, numerical_features),
            ('cat', categorical_transformer, categorical_features)
        ])

    model = RandomForestRegressor(n_estimators=100, random_state=73)
    my_pipeline = Pipeline(steps=[('preprocessor', preprocessor),('model', model)])

    y = data.price
    X = data.drop('price', axis=1)

    X_train, X_test, y_train, y_test = train_test_split(X, y, test_size=0.1, random_state=73)

    my_pipeline.fit(X_train, y_train)

    y_pred = my_pipeline.predict(X_test)

    mse = mean_squared_error(y_test, y_pred)
    r2 = r2_score(y_test, y_pred)
    print("R-squared error:", r2)
    print("Model saved successfully.")
```

```
R-squared error: 0.7531646896065121
Model saved successfully.
```

Model Explanation:

Data Preprocessing: Similar to the previous code, the data is pre-processed using a Column Transformer. Numerical features are standardized using Standard Scaler, while categorical features are one-hot encoded using One Hot Encoder.

Model Building: The pre-processed data is then passed through a Random Forest Regressor. A Random Forest is an ensemble learning method that constructs multiple decision trees during training and outputs the mean prediction of the individual trees for regression tasks.

Pipeline Construction: Again, a Pipeline is constructed to combine the preprocessing steps and the model into a single object.

Performance Evaluation:

For evaluating the performance of the model, the same metrics are used:

Mean Squared Error (MSE): It measures the average of the squares of errors between predicted values and actual values. Lower values indicate better fit.

R-squared (R2) Score: It measures the proportion of the variance in the dependent variable that is predictable from the independent variables. Higher values (closer to 1) indicate a better fit.

Evaluation Metrics Consideration:

The choice of evaluation metrics aligns with the business objectives as discussed previously. MSE and R-squared are appropriate for measuring the accuracy of price predictions in real estate, ensuring customer satisfaction and revenue optimization.

ML Model Selection:

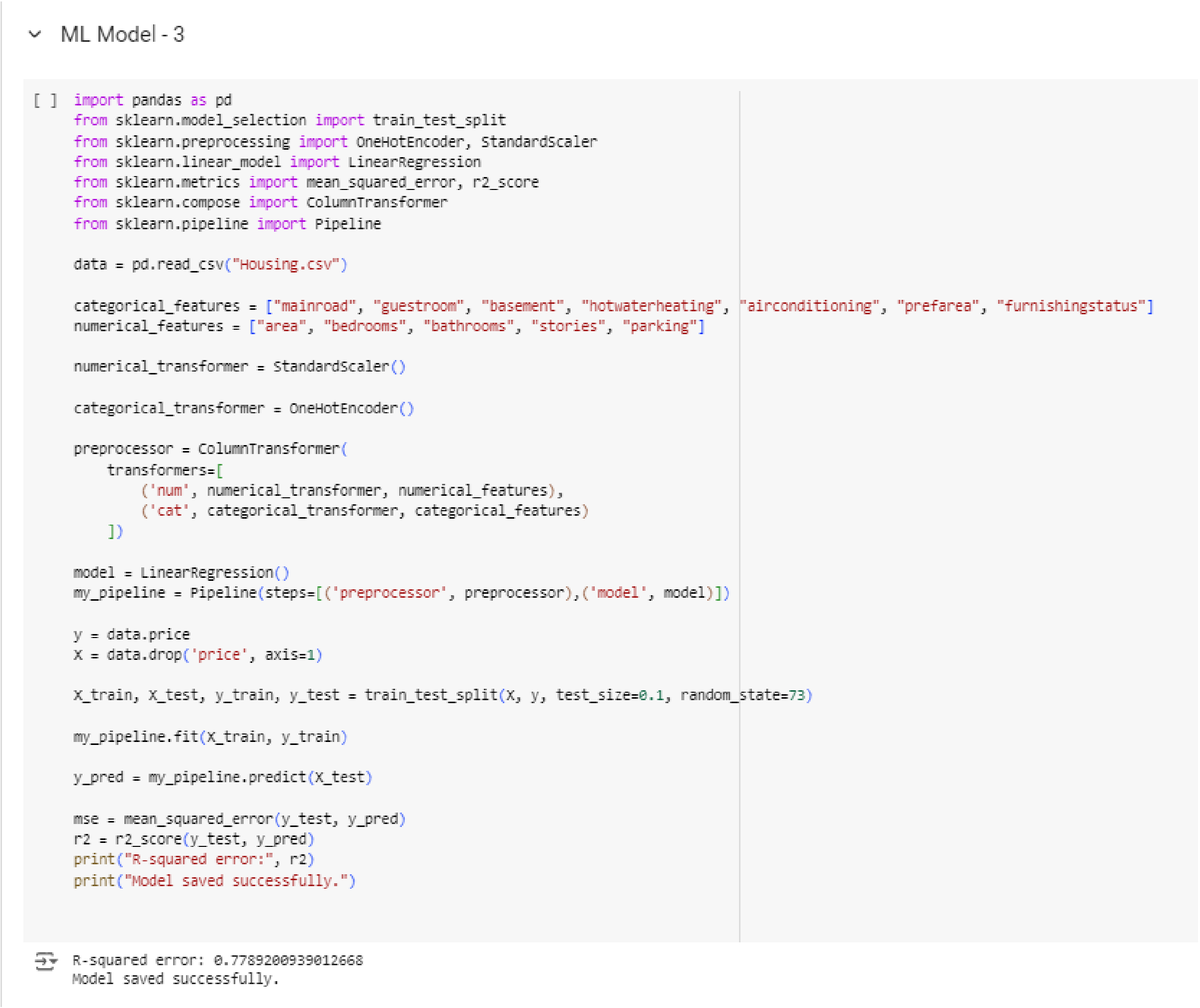
The Random Forest Regressor is chosen as the final prediction model. Random Forests are powerful and versatile algorithms suitable for regression tasks, known for their robustness, handling of non-linear relationships, and ability to capture complex patterns in data. With a higher number of trees (n_estimators), the model tends to generalize better.

Feature Importance:

Random Forests inherently provide feature importance based on how much each feature contributes to reducing the impurity (e.g., mean squared error) across all decision trees in the forest. Features with higher importance values are more influential in making predictions.

Overall, the Random Forest Regressor offers a strong approach for predicting housing prices based on the given features, with potential for further analysis and optimization using various evaluation metrics and model explainability techniques.

ML Model 3:



```
[ ] import pandas as pd
    from sklearn.model_selection import train_test_split
    from sklearn.preprocessing import OneHotEncoder, StandardScaler
    from sklearn.linear_model import LinearRegression
    from sklearn.metrics import mean_squared_error, r2_score
    from sklearn.compose import ColumnTransformer
    from sklearn.pipeline import Pipeline

    data = pd.read_csv("Housing.csv")

    categorical_features = ["mainroad", "guestroom", "basement", "hotwaterheating", "airconditioning", "prefarea", "furnishingstatus"]
    numerical_features = ["area", "bedrooms", "bathrooms", "stories", "parking"]

    numerical_transformer = StandardScaler()

    categorical_transformer = OneHotEncoder()

    preprocessor = ColumnTransformer(
        transformers=[
            ('num', numerical_transformer, numerical_features),
            ('cat', categorical_transformer, categorical_features)
        ])

    model = LinearRegression()
    my_pipeline = Pipeline(steps=[('preprocessor', preprocessor),('model', model)])

    y = data.price
    X = data.drop('price', axis=1)

    X_train, X_test, y_train, y_test = train_test_split(X, y, test_size=0.1, random_state=73)

    my_pipeline.fit(X_train, y_train)

    y_pred = my_pipeline.predict(X_test)

    mse = mean_squared_error(y_test, y_pred)
    r2 = r2_score(y_test, y_pred)
    print("R-squared error:", r2)
    print("Model saved successfully.")
```

```
R-squared error: 0.7789200939012668
Model saved successfully.
```

Model Explanation:

Data Preprocessing: The data preprocessing steps remain consistent with the previous examples. Numerical features are standardized using Standard Scaler, while categorical features are one-hot encoded using One Hot Encoder.

Model Building: The pre-processed data is fed into a Linear Regression model. Linear Regression is a simple yet powerful algorithm that models the relationship between the independent variables and the target variable by fitting a linear equation to observed data.

Pipeline Construction: Similar to before, a Pipeline is constructed to encapsulate the preprocessing steps and the model, facilitating a streamlined workflow.

Performance Evaluation:

The performance of the model is evaluated using two metrics:

Mean Squared Error (MSE): It measures the average of the squares of errors between the predicted values and the actual values. Lower values indicate better fit.

R-squared (R2) Score: It quantifies the proportion of the variance in the dependent variable that is predictable from the independent variables. Higher values (closer to 1) suggest a better fit of the model.

Evaluation Metrics Consideration:

These evaluation metrics are selected based on their relevance to the business objectives. MSE and R-squared provide insights into the accuracy and explanatory power of the model, which are essential for predicting housing prices accurately.

ML Model Selection:

Linear Regression is chosen as the final prediction model. It offers simplicity, interpretability, and ease of implementation. Despite its simplicity, Linear Regression can provide robust results when the relationship between features and target variable is approximately linear.

Feature Importance:

Linear Regression provides feature coefficients as a measure of feature importance. Features with higher coefficients have a stronger linear relationship with the target variable and contribute more to the predictions.

Overall, Linear Regression serves as a reliable model for predicting housing prices based on the provided features, offering straightforward interpretability and decent performance, which can be further improved and analysed using additional techniques and evaluation metrics.

10. Conclusion

Dynamic Market Changes: The real estate market is inherently dynamic, subject to various economic, social, and environmental factors that continuously influence housing prices. These changes can include shifts in interest rates, changes in government policies related to housing, fluctuations in demand and supply dynamics, and broader economic trends such as inflation or recessions. Predicting house prices accurately in such a dynamic environment requires models that can adapt to changing market conditions and incorporate up-to-date information to make reliable predictions.

Data Complexity: Real estate datasets often encompass a wide range of property features, each with its own significance in determining house prices. These features can vary significantly across different regions and property types, adding to the complexity of building a comprehensive predictive model. Managing this complexity requires robust feature engineering techniques, careful selection of relevant variables, and the use of advanced modeling approaches to capture the nuanced relationships between features and house prices.

Bias and Ethics: Biases present in real estate data, such as selection bias in property listings or historical disparities in housing policies, can introduce distortions that impact the accuracy and fairness of predictive models. Addressing biases and ethical concerns involves implementing measures for data transparency, fairness testing, and accountability in model development processes. It also requires ongoing monitoring and evaluation to detect and mitigate any unintended consequences of model predictions on marginalized communities or vulnerable populations.

In conclusion, building a reliable house price prediction model entails navigating the complexities of a dynamic real estate market, managing the intricacies of diverse data sources and features, and addressing biases and ethical considerations to ensure fair and accurate predictions. By incorporating advanced modeling techniques, robust data governance practices, and a commitment to ethical principles, stakeholders can develop predictive models that provide valuable insights into housing market trends while upholding standards of fairness and integrity.

.