Sri Deepthi Vootla

**Analysis of the factors influencing the housing prices**

Jayakishan Minnekanti

**1.Abstract:**

The present analysis aims to investigate the key factors that influence housing prices in a specific real estate market. By utilizing a multiple linear regression approach, we sought to identify and quantify the impact of various features on property prices. The dataset used for the analysis contained information on housing attributes such as area, number of bedrooms and bathrooms, presence of guest rooms, basements, and air conditioning, as well as parking facilities and preferred areas. Additionally, the furnishing status of each property was considered.

The findings of the analysis revealed significant factors that play a crucial role in determining housing prices. Notably, the area of the property, the number of bathrooms, and the presence of air conditioning were identified as the most influential factors affecting property prices. Moreover, properties situated on the main road, those with guest rooms, basements, and hot water heating were found to have a positive impact on prices. The furnishing status of a property emerged as an essential determinant, with fully furnished properties commanding higher prices compared to semi-furnished and unfurnished ones. Additionally, the number of stories and parking facilities were also found to influence housing prices significantly.

While the insights from this analysis provide valuable information for property buyers, sellers, and investors, it is essential to acknowledge the study's limitations. The dataset's size and quality, as well as the assumptions made in the regression model, may influence the accuracy and generalizability of the results. Furthermore, the exclusion of certain relevant factors and the encoding of categorical variables may introduce biases. To enhance the analysis's robustness, future research should focus on obtaining larger and more diverse

datasets, considering additional relevant variables, and exploring advanced modeling techniques to capture the complexities of the housing market. Nevertheless, the current study offers valuable insights into the factors influencing housing prices, contributing to informed decision-making in the dynamic real estate landscape.

**2.Introduction:**

The housing market is a critical sector of any economy, as it not only provides shelter to individuals and families but also serves as a significant indicator of economic stability and growth. For both prospective homebuyers and real estate investors, understanding the factors that influence housing prices is of paramount importance. Identifying these factors can help individuals make informed decisions about buying or selling properties and enable investors to strategically allocate their resources for optimal returns.

**The main objectives of this analysis are as follows:**

**Identify Significant Factors:** By performing multiple linear regression, we aim to identify the attributes that have a statistically significant impact on housing prices. Understanding these factors can help individuals and investors prioritize their preferences and make well-informed decisions.

**Quantify the Impact:** Not all housing attributes have the same impact on property prices. Through regression analysis, we can quantify the effect of each significant factor, indicating which features contribute the most to variations in housing prices.

**Provide Insights for Decision-Making:** The insights gained from this analysis will be valuable for prospective homebuyers, sellers, and real estate investors. Armed with this information, they can assess the value of a property and negotiate transactions more effectively.

**Addressing Limitations:** It is essential to acknowledge that no analysis is without limitations. As we delve into the findings, we will also discuss the potential constraints and biases that may affect the accuracy and generalizability of the results. Overall, this report aims to shed light on the factors that influence housing prices in the specified real estate market. Through rigorous analysis and interpretation of the data, we hope to provide valuable insights that contribute to a better understanding of the dynamic housing market and support informed decision-making in the realm of real estate.

**3.Methodology:**

**3.1.Data Collection:**

The dataset used in this analysis was obtained from Kaggle, a well-known platform for data science and machine learning enthusiasts. The dataset was sourced from Kaggle's repository of publicly available datasets and it comprises information related to housing prices in the north-eastern region of USA. The data was collected with the aim of studying the factors that influence housing prices in the area and understanding the dynamics of the real estate market.

**3.2.Data Description:**

The dataset consists of 540 observations, each representing a distinct residential property located within the USA. For each property, a total of 13 attributes were recorded, providing valuable insights into the factors that may influence housing prices in the US.

The variables in the dataset are as follows:

**price:** The selling price of the property in the local currency (USD), serving as the target variable for the analysis.

**area:** The total built-up area of the property in square feet, indicative of the property's size and livable space.

**bedrooms:** The number of bedrooms in the property, providing an understanding of the property's capacity to accommodate occupants.

**bathrooms:** The number of bathrooms in the property, influencing its functional appeal and convenience.

**stories:** The number of stories in the property, categorizing it into single-story or multi-story.

**mainroad:** A categorical variable indicating whether the property is situated near the main road, which may affect accessibility and convenience.

**guestroom:** A categorical variable indicating whether the property includes a separate guest room, adding to its appeal for visitors.

**basement:** A categorical variable indicating whether the property has a basement, offering additional storage or living space.

**hotwaterheating:** A categorical variable indicating whether the property is equipped with a hot water heating system, contributing to comfort and value.

**airconditioning:** A categorical variable indicating whether the property features an air conditioning system, making it attractive in warmer climates.

**parking:** The number of parking spaces available for the property, a significant factor for buyers with multiple vehicles.

**prefarea:** A categorical variable indicating whether the property is located in a preferred or upscale area, potentially impacting its prestige and value.

**furnishingstatus:** A categorical variable indicating the furnishing status of the property, with categories such as furnished, semi-furnished, and unfurnished.

**3.3.Data Preprocessing:**

We conducted essential data preprocessing steps to ensure the dataset's suitability for analysis.

One of the critical preprocessing techniques employed was one-hot encoding, which effectively transformed categorical variables into a binary format, enabling their integration into the predictive models. The dataset contains several categorical variables, such as 'furnishingstatus,' 'mainroad,' 'guestroom,' 'basement,' 'hotwaterheating,' 'airconditioning,' and 'prefarea,' that play a pivotal role in determining housing prices. However, most machine learning algorithms are designed to handle numerical data, making it necessary to convert these categorical variables into a suitable format.

To overcome the challenge of incorporating categorical variables into the regression models, we employed one-hot encoding. This technique replaces each categorical variable with the binary values for each unique category. As a result, categorical variables with 'n' unique categories are transformed into 'n' binary numbers.

For example, the 'furnishingstatus' variable initially had three categories: 'furnished,' 'semi-furnished,' and 'unfurnished.' After applying one-hot encoding, the categories were replaced with three numerical values: 1, 2 and 3. All the other categorical variables in the dataset having ‘yes’ and ‘no’ were replaced with 1 and 0.

**4.Exploratory Data Analysis:**

Exploratory Data Analysis (EDA) is an essential phase in our study, aimed at gaining insights into the underlying patterns and relationships within the dataset. Through visualizations such as scatter plots for numerical variables and box plots for categorical variables, we aim to uncover trends, identify outliers, and understand the potential factors influencing housing prices.

**4.1 Scatter Plots for Numerical Variables**

To investigate the relationships between the target variable, 'price,' and other numerical predictors, we utilized scatter plots. These visualizations depict how 'price' varies concerning the property's size ('area'), the number of bedrooms ('bedrooms'), bathrooms ('bathrooms'), stories (‘stories’) and parking spaces ('parking').

Scatter plots allow us to observe any potential correlations between the target variable and numerical predictors. For instance, the 'area' vs. 'price' scatter plot can reveal whether larger properties tend to command higher prices. Similarly, we can analyze how the number of bedrooms and bathrooms impacts housing prices, as well as the relationship between parking availability and property values.

A diagram of a plot matrix

Description automatically generated

**Figure-1**

From the figure-1 we can observe that the area is having a positive linear relationship with the price and both bathrooms and stories are also showing a strong correlation with the price whereas, bedrooms and parking are having a weak correlation compared with the other numerical variables.

**4.2 Box Plots for Categorical Variables**

To understand the influence of categorical variables on housing prices, we employed box plots. These visualizations represent the distribution of 'price' across different categories of each categorical variable, such as 'furnishingstatus,' 'mainroad,' 'guestroom,' 'basement,' 'hotwaterheating,' 'airconditioning,' and 'prefarea.'

Box plots provide insights into the central tendencies (median), spread (interquartile range), and potential outliers within each category.

A diagram of a graph

Description automatically generated

**Figure-2**

For example, int the figure-2 we can confirm that the price is increasing when the houses are on the main road compared to the houses that are on the side street or a secondary street. Also the interesting thing that we observed through the box plots is that the furnishing status of a house isn’t showing much impact on the house price whether the house is ‘furnished’, ‘semi-furnished’ or ‘unfurnished’.

**5.Data Mining Techniques:**

Data mining techniques play a crucial role in extracting valuable insights from the data and enhancing the accuracy and robustness of the analysis. By combining traditional statistical methods with data mining techniques, we have gained a deeper understanding of the factors affecting property prices and it will help to make more informed decisions in real estate related tasks.

**5.1.Multi-Collinearity:**

Before proceeding to the actual regression analysis model, we have tested for the multi-collinearity between the variables. If multi-collinearity is present, it can impact the regression model's stability, interpretability, and predictive power. It may lead to inflated standard errors, making it challenging to identify the true significance of predictors.

Detecting multicollinearity can be accomplished using various methods, in this we have chosen Variance Inflation Factor (VIF). Calculating VIF values for each predictor variable and analyzing the obtained results to the other predictor variables can help us understand the correlation among the predictor variables.

**Table-1**

From the VIF values, we can confirm that there is no multi-collinearity among our independent variables. VIF values greater than 5 or 10 generally indicate high multi-collinearity. As the VIF values of our variables are all lesser than 4, so we considered it to be a good sign to proceed to the regression model analysis.

**5.2.Linear Regression:**

multiple linear regression is a statistical technique used to model the relationship between a dependent variable and two or more independent variables (predictors). It extends the concept of linear regression to accommodate multiple predictors and is commonly employed when the outcome variable is influenced by a combination of several factors.

The mathematical representation of multiple linear regression can be expressed as follows:

A black and white text

Description automatically generated

Image1 Source: towardsdatascience

The linear regression analysis was performed to investigate the relationship between various features and property prices. The model was trained on 80% of the data and tested on 20% of the data to evaluate its performance.

**5.3.R-Squared Test:**

Additionally, we have performed the R-squared (R²) test, also known as the coefficient of determination, which is a statistical measure used to assess the goodness of fit of a regression model. It quantifies the proportion of the total variation in the dependent variable that is explained by the independent variables (predictors) in the model. In other words, R² represents the percentage of variability in the response variable that can be accounted for by the predictor variables.

The R-squared value ranges from 0 to 1, where:

* R² = 0 indicates that the model explains none of the variability in the dependent variable. It suggests that the model does not fit the data well.
* R² = 1 indicates that the model explains all the variability in the dependent variable. It suggests that the model perfectly fits the data.

From our analysis the value for the R-squared test we have obtained is 0.7162. This can be interpreted as that our regression model is analyzed to be 71% fitted to our dataset.

**6.Results:**

The linear regression analysis was performed to investigate the relationship between various features and property prices. The analysis revealed that certain features had a substantial impact on property prices.

A graph showing a number of different sizes of bars

Description automatically generated

**Figure-3**

**Impactful Features:**

Notably, the number of bathrooms of the property was found to be a strong predictor of price. Properties with more bathrooms tended to have higher valuations.

The area of a property also had a significant influence on property prices. An increase in the property's area was associated with a corresponding increase in its price. This finding is consistent with the intuition that larger properties generally command higher prices.

Similarly, the number of stories in a property showed a positive correlation with prices, with multi-story buildings fetching higher prices compared to single-story ones.

Additionally, the presence of air conditioning emerged as a crucial factor affecting property prices. Properties equipped with air conditioning systems were associated with higher prices, as they offer added comfort and appeal to potential buyers.

**A pie chart with different colored circles

Description automatically generated**

**Figure-4**

**Positive Impact Features:**

Certain features had a positive impact on property prices. For instance, properties located on the main road were found to command higher prices compared to those on side streets or secondary roads. The accessibility and visibility of main road properties make them more desirable to buyers.

The presence of guest rooms, basements, and hot water heating systems also showed a positive correlation with property prices. Properties offering these amenities tend to attract higher valuations, likely due to the added convenience and utility they offer.

**7.Conclusion & Limitations:**

The results of the linear regression analysis provide valuable insights into the factors influencing property prices. The area, number of bathrooms, stories, and air conditioning were identified as impactful features. Additionally, properties located on the main road and offering guest rooms, basements, and hot water heating systems tended to have higher prices. The furnishing status of a property also played a role in determining its price.

These findings can be valuable for property sellers, buyers, and real estate professionals in understanding the key factors that drive property valuations and making informed decisions in the real estate market.

**Limitations:**

**Data Quality and Size:**

A small dataset may not fully capture the variability and complexity of the real-world relationships between predictors and property prices. A larger and more diverse dataset with a wider range of property types, locations, and market conditions could lead to a more comprehensive understanding of the factors influencing property prices. Such a dataset would provide a more robust foundation for building the regression model, improving its accuracy, and enabling better generalization of the findings to a broader population of properties.

**Assumptions of Linear Regression Model:**

Linear regression assumes a linear relationship between the predictors and the target variable (property prices in this case). This assumption implies that the change in the dependent variable is directly proportional to the change in each predictor, and the relationship is constant across all levels of the predictors. However, in real-world scenarios, the relationships between variables may not always be strictly linear. Therefore, it is essential to assess the validity of this assumption through techniques like residual analysis, which evaluates the linearity of the model's residuals. If the assumption is violated, alternative modeling techniques such as polynomial regression, spline regression, or non-linear regression might be considered to better capture the true relationships and improve the model's performance in predicting property prices.

**Lacking Variables:**

The dataset's limitation in including certain variables, such as location-specific factors, economic indicators, and buyer preferences, can be critical in accurately predicting property prices. These variables lacking from the data might have significant influence on property valuations, and their exclusion can result in biased or incomplete regression results. For instance, properties located in desirable neighborhoods or close to amenities might command higher prices, irrespective of other property features. Similarly, macroeconomic indicators like interest rates, inflation, or housing market trends can impact property prices. The absence of these factors in the dataset may limit the model's ability to explain the full variability in property prices. To address this limitation, additional data sources should be considered to incorporate these relevant variables and ensure a more comprehensive analysis of property price determinants.

**8.Future Research:**

Future research in the area of property price analysis and prediction can focus on several aspects to improve the understanding and accuracy of the models. Some potential areas of future research include:

**Larger and Diverse Datasets:** Acquiring larger and more diverse datasets from different regions and property markets can enhance the generalizability of the models. A more extensive dataset will enable researchers to capture a broader range of property types, locations, and market conditions, leading to more robust and reliable predictions.

**Incorporating Location-specific Factors:** Including location-specific variables in the analysis can significantly improve the predictive power of the models. Factors such as neighborhood desirability, proximity to amenities, transportation, and schools can play a crucial role in determining property prices. Geospatial data and location-based analytics can be leveraged to incorporate these variables.

**Inclusion of External Factors:** Considering external economic indicators, real estate market trends, and demographic data can provide valuable context for understanding property price variations. Economic factors, such as interest rates, inflation, and GDP growth, can significantly impact property demand and prices.

**Advanced Modeling Techniques:** Exploring advanced regression techniques can improve the stability and interpretability of the models. Additionally, machine learning algorithms like Random Forests, Gradient Boosting, or Neural Networks can be applied to capture non-linear relationships and improve predictive performance.

**Feature Engineering:** Investigating new feature engineering techniques to create additional meaningful predictors can enhance model performance. Feature selection methods can also be employed to identify the most relevant features for property price prediction.

**9.References:**

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

Image1: <https://towardsdatascience.com/multiple-linear-regression-for-manufacturing-analysis-c057d4af718b>