Predicting the Housing Market using Multiple Linear Regression

Garrick Ho

Jun Yan

Department of Statistics

University of Connecticut

December 14, 2023

Abstract

With the housing market fluctuating every year, many real estate owners and businesses face a problem with gaining an accurate representation of the future of the housing market. Lots of people do not know much about the market, so they will go to real estate owners or businesses to learn more about the market. And then see what the market is at and if it is the right time to buy a house. If real estate owners and businesses are able to gain more accuracy in predicting the housing market, they will be able to help more people and possibly gain more business. Not only can this benefit the consumers, the builders are able to move with the market and have a better understanding of it. This paper aims to construct a multiple linear regression model by using machine learning techniques to predict housing costs accurately. This paper ends with an explanation of the limitations of this study and potential future works.

Keywords: housing market prediction; model predictions; multiple linear regression; natural language processing

1 Introduction

The United States has been through three recessions in the last decade. During that decade, house prices have reached the very bottom of the market and also reached the very top of the market. The market has tripled since 1992. After a brief dip in 2002-2003, they bounced back and skyrocketed over 20% from May 2003 to November 2004. [JN05] Because of the constant change, new models for predicting the housing market also need to change so that the model can stay up to date with the changes that occur. This is important because it gives people insights into whether it's a good time to buy or sell a property, helps policymakers make informed decisions, and allows investors to strategize their moves. House price predictions can help the developer determine the selling price of a house and can also help the customer arrange the right time to purchase a house. [Alf+17]

Engineers in the past had to face a significant challenge in forecasting and assessing house prices. Fortunately, with technological advances, machine learning algorithms offer a solution, enabling users to predict house prices by training data sets on various algorithms. The market is currently flooded with software and models designed to enhance the accuracy of house price predictions, minimizing errors to negligible levels. These predictions can be based on the analysis of different aspects of a house, such as kitchen size, bedroom count, bathroom count, and overall land area. These aspects of a house can go on forever. Analysts employ various techniques to compute and analyze data, to ensure that the predictions are aligned closely with previous estimations. [SRR21]

The primary significance of this paper lies in its contribution to the existing body of knowledge on the specified topic. By corroborating the findings of prior research papers, this study aims to reinforce and validate the existing studies. Emphasize the specific insights where this paper aligns with and extends the conclusions from other studies. In the final stages of our analysis, a comparison will be done between the model developed in this paper and the results available on Kaggle. The expected outcome is to try to produce a model that is more efficient than the one provided with the current data set. By benchmarking our model against a Kaggle reference, we seek to evaluate its efficiency and effectiveness in predicting housing prices. This comparative analysis serves as a valuable validation step, allowing us to assess the model's performance. We aim not only to replicate but to surpass the predictive accuracy and performance metrics achieved by the existing Kaggle model.

The structural framework of this paper unfolds in a systematic progression, commencing with the overarching introduction encapsulated in Section 1. Moving seamlessly, Section 2 offers a concise yet insightful introduction to the dataset, intricately delving into the variables at play. This section not only outlines the variables but also expounds on their individual contributions to addressing the core research question. Then, the methodology used in this paper will be explained in Section 3. The applications of the results found will be in Section 4. This section will present the results from the model and have further analysis with tables and figures. More in-depth analysis of how the results can be applied to the real world will be in this section. Finally, the discussion part for this paper will be in Section 5. This final section navigates through the limitations inherent in the current study, providing a transparent acknowledgment of its boundaries and future directions.

2 Data

The data that is being utilized for this study was sourced from Kaggle. It's worth acknowledging that the data set is not expansive; rather, it constitutes a limited sample offering a snapshot of the broader housing market dynamics. This data set is comprised of $n = 545$ observations, each representing a unique house sample, the data set is characterized by thirteen variables. Among these variables, seven fall into the categorical realm, while the remaining six are numerical. This comprehensive data set serves as the foundation for constructing and training a Multiple Linear Regression (MLR) model geared toward predicting housing costs.

The preparatory phase for the data involved a strategic approach, leveraging the use of the pandas library in Python for exploration and visualization. To enhance the compatibility of the data set with the Multiple Linear Regression model, categorical variables represented as strings underwent a transformative process. Specifically, the ’furnishingstatus’ column transformed, breaking it down into three distinct columns: one for furnished, one for semi-furnished, and a final one for unfurnished. This categorical refinement facilitates a more refined representation of furnishing status, laying the groundwork for a more detailed and accurate predictive model.

Table 1: Data Description

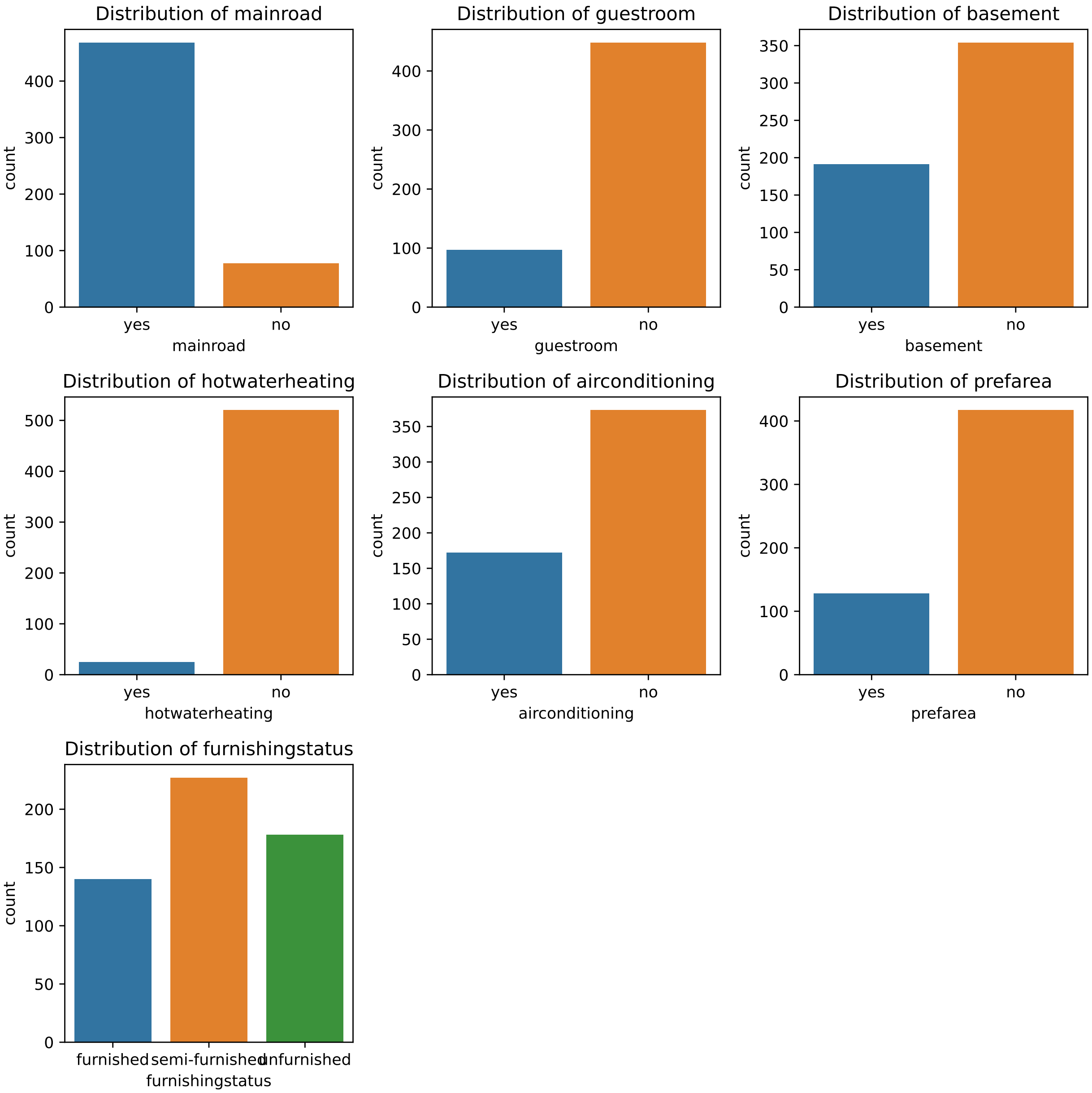
Variable	Data Type	Instance	Purpose
price	Int	13300000	Y
area	Int	7420	X_n
bedrooms	Int	4	X_n
bathrooms	Int	2	X_n
stories	Int	3	X_n
mainroad	Str	yes, no	X_n
guestroom	Str	yes, no	X_n
basement	Str	yes, no	X_n
hotwaterheating	Str	yes, no	X_n
airconditioning	Str	yes, no	X_n
parking	Int	2	X_n
prefarea	Str	yes, no	X_n
furnishingstatus	Str	furnished, semi-furnished, unfurnished	X_n

In Table 1, the ’variable’ column shows the names of all the variables that are in the dataset. The ’Data Type’ column shows the data type of each variable in the dataset. Note that ’Int’ means integer and ’Str’ means string. The ’Instance’ column shows an example of what a data point for that column would look like. The ’Purpose’ column shows the importance of each variable to the machine learning step.

3 Methods

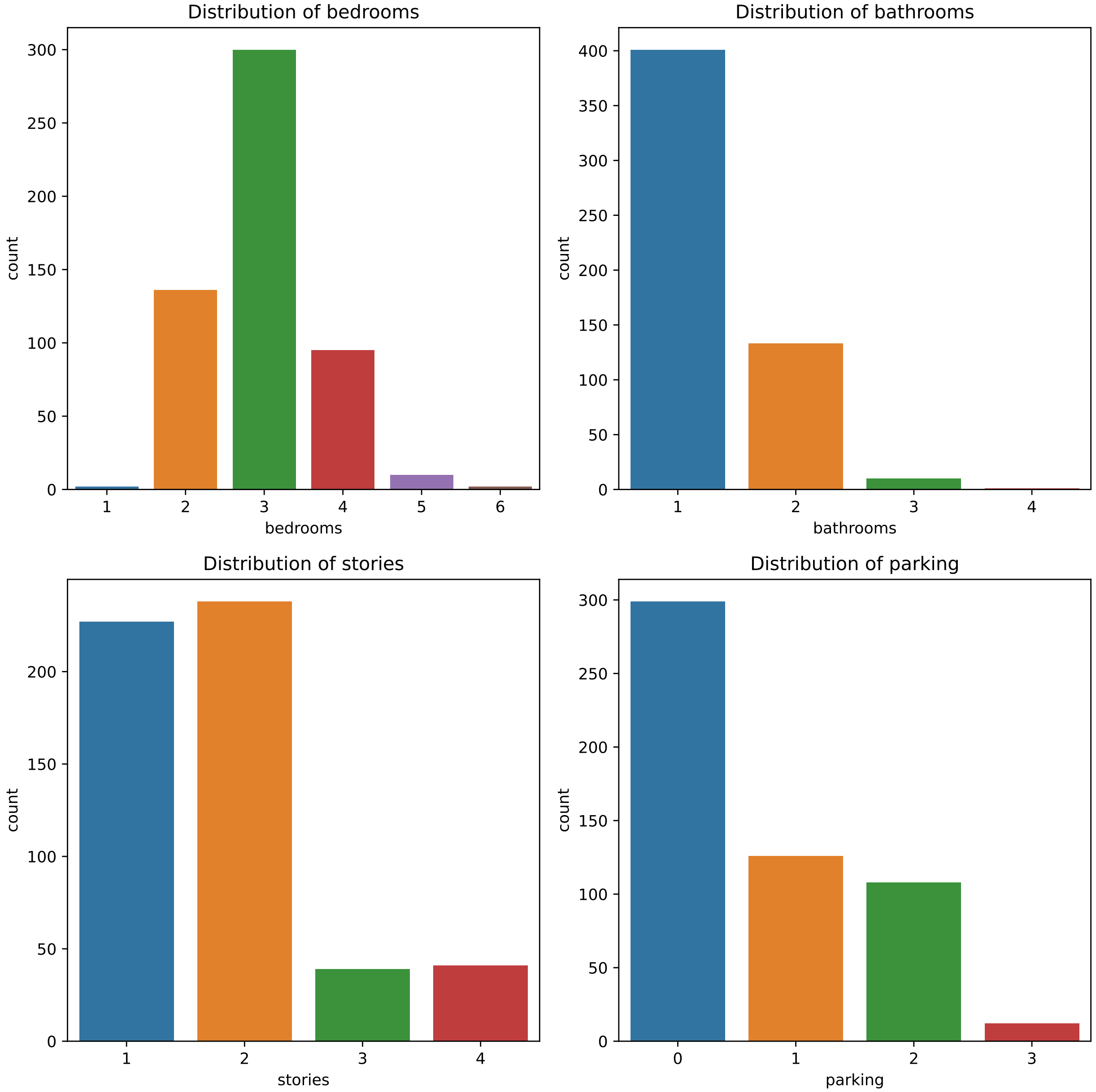
The data was first processed by making all the variables in the data able to be read by the model. The first variable that was altered was the status of the house being furnished. It was converted into dummy variables. A one in the dummy variable would mean that the

Figure 1: Categorical Variable Plots



In figure 1, the plots show the distributions for all the categorical variables. Most of the houses in this data set are located near a main road. Only one-fifth of the houses have a room for guests. For the basement category, two-fifths of them have one. Almost none of the houses have hot water heating. There is air conditioning in about two-fifths of the data. About one-fifth of the houses are located in a preferred area. And lastly, the house can come either furnished, semi-furnished, or unfurnished. Most of the house came semi-furnished, next being unfurnished and then fully furnished.

Figure 2: Numerical Variable Plots


Distribution of bedrooms

Distribution of bathrooms

Distribution of stories

Distribution of parking

In figure 2, four different plots are shown to display the distribution of the numerical variables in the data. The average number of bedrooms in these homes is three, with two being the second most. For the number of bathrooms, the majority of homes have one. About half of the homes in the data are one-story homes while about the other half are two stories. For parking, most of them have no parking spots. A few of them have one or two parking spots.

house was in that current status and a zero would mean that it was not in that status. After that, the next step was to convert the yes and no entries to 1's and 0's. This was done to the variables mainroad, guestroom, basement, hotwaterheating, airconditioning, and prefarea. Before splitting the data set into training and testing groups, the variables were scaled because the values for price and the area of the house were significantly larger than the other values in the data set. Scaling is needed here because we want the values from each variable to contribute equally to the model and avoid the overpowering features with larger values. Finally, the data is split into 80% for training and 20% for testing. The size of the training group was arbitrarily chosen to be greater than the size of the testing set to ensure that the multiple linear regression model is trained on as much data as possible to find and learn meaningful patterns. It is widely accepted for training data to be larger than testing data. An inadequate amount of training data can result in a poor-performing model that underfits and does not generalize to new instances well. A multiple linear regression model was fitted to the training group and then validated on the testing group using the scikit-learn library in Python. Several performance metrics about the predictions were then assessed.

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + ... + \beta_n X_n + \epsilon \tag{1}$$

Equation 1 is the Multiple Linear Regression model being used. Y is the dependent variable, also known as the value that is being predicted. In this case, it would be the cost of a house. X_n is the independent variable, which is the characteristics of the house that are used to predict the house cost. β_n is the average amount by which the dependent variable increases and decreases depending on when the independent variable increases one standard deviation. When all the other independent variables are held constant. β_0 represents the value of the dependent variable when all the independent variables are equal to zero. ϵ is the error term which represents the margin of error within the linear regression model. [UG13]

4 Application

This linear regression model aims to predict the house price with the different characteristics of a house by using the trained data and testing it against the testing data. To see how well the model performed in predicting the house prices, R-squared was used and a visualization was created to help see the results. The figure below represents how the model did with predicting the house prices. As we can see, there is a clear trend in the plot showing the accuracy of the model. One noticeable trend in the plot is that the model's predicted value of the house is undervalued compared to the actual price of the house. This trend with the model can be supported with the R-squared value.

$$\begin{aligned} R^2 &= 1 - \frac{\text{sum squared regression (SSR)}}{\text{total sum of squares (SST)}}, \\ &= 1 - \frac{\sum(y_i - \hat{y}_i)^2}{\sum(y_i - \bar{y})^2} \end{aligned} \tag{2}$$

Equation 2 is the formula for R-squared. R-squared is a measure of the goodness of fit of a model. The sum squared regression is the sum of the residuals squared and the total sum of squares is the sum of the distance the data is away from the mean all squared. It is a percentage that takes the values between 0 and 1. If the percentage is 1, then the variation in the y values is accounted for by the x values. If the percentage is 0, then none of the variations of the y values is accounted for by the x values. [Kas18]

Table 2: Results

Measurement	Outcome
R^2	0.6529
Mean Absolute Error	970043.4039
Root Mean Squared Error	3.0776

Table 2 shows the results of the multiple linear regression model from the paper.

The calculated R-squared value for the model produced in this paper was 0.6529. A

Figure 3: Linear Regression Plot

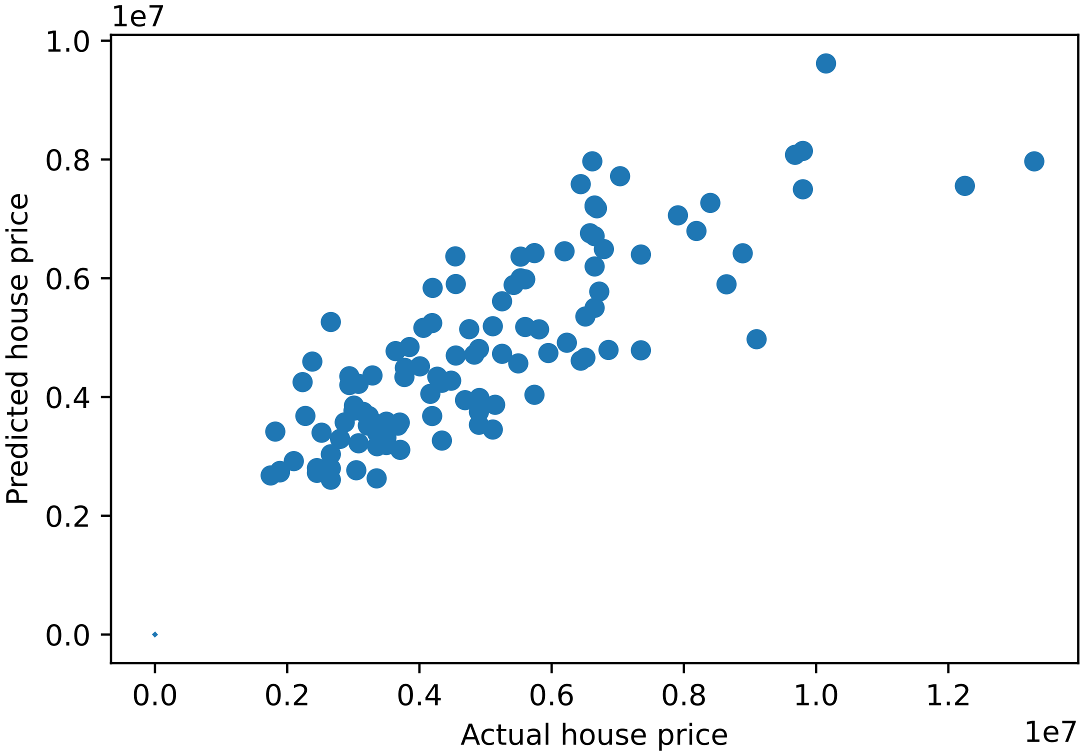
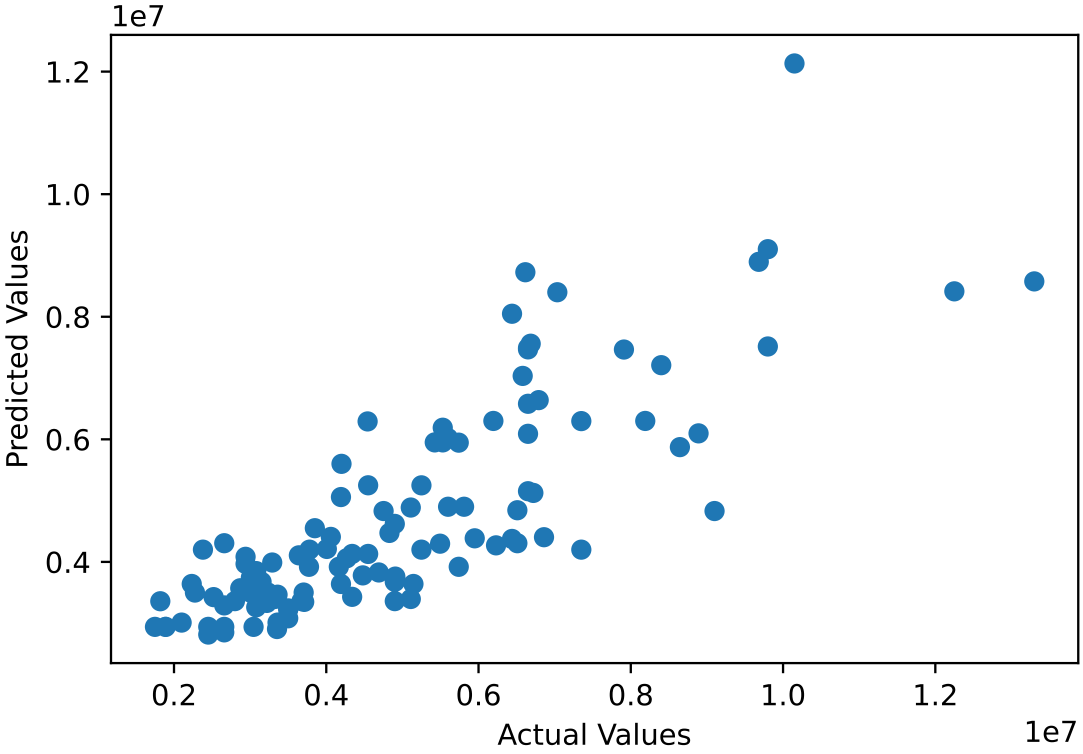


Figure 3 shows a linear regression plot with the predicted house price against the actual house price. Based on this plot that was created, it looks like the model was somewhat accurate in predicting the prices.

R-squared value of 0.6529 indicates the proportion of the variance in the dependent variable that can be explained by the independent variables in a multiple linear regression model. In other terms, about 65.29% of the variability in the observed data can be accounted for by the model. The remaining 34.71% of the variance is unaccounted for and may be attributed to other factors not included in your model or to random variability.

In comparison with the model produced in Kaggle, it is slightly better. The Kaggle model had a R-squared value of 0.5791. A R-squared value of 0.5791 suggests that approximately 57.91% of the variance in the dependent variable can be explained by the independent variables in your regression model. The model that was produced in the paper has a better R-squared value of about 7.38%. This suggests that the regression model outlined in the

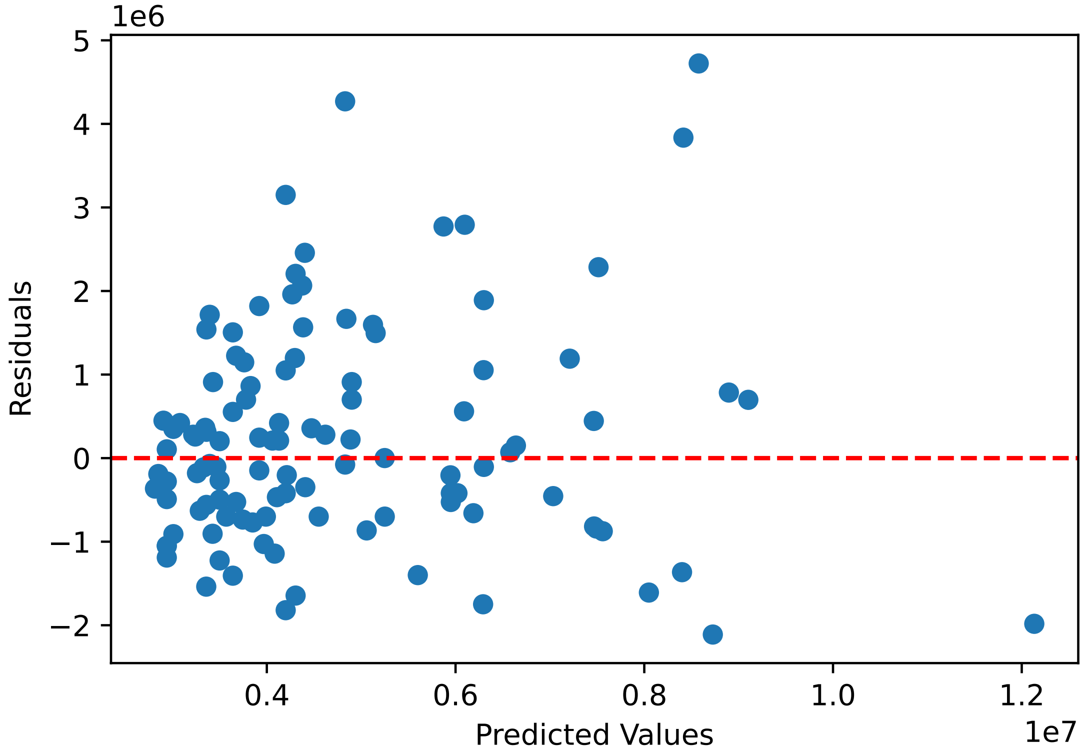
Figure 4: Linearity Check



paper provides a more robust and comprehensive explanation for the variability observed in the dependent variable, thereby enhancing its predictive accuracy and overall effectiveness.

With a regression model that seems better than the one produced in Kaggle, it is imperative to ensure the fulfillment of four fundamental assumptions to guarantee the reliability and validity of statistical inferences. Firstly, the assumption of linearity asserts that the relationship between the dependent and independent variables is adequately represented by a linear function. Homoskedasticity, the second assumption, demands that the variance of the error term in a regression model is constant across all values of the independent variables, emphasizing the importance of a consistent spread of errors. Normality, the third assumption, involves verifying that the residuals follow a normal distribution, ensuring the validity of statistical tests and confidence intervals. The fourth assumption, independence of independent variables, calls for the absence of multicollinearity, emphasizing that predic-

Figure 5: Homoskedasticity Check

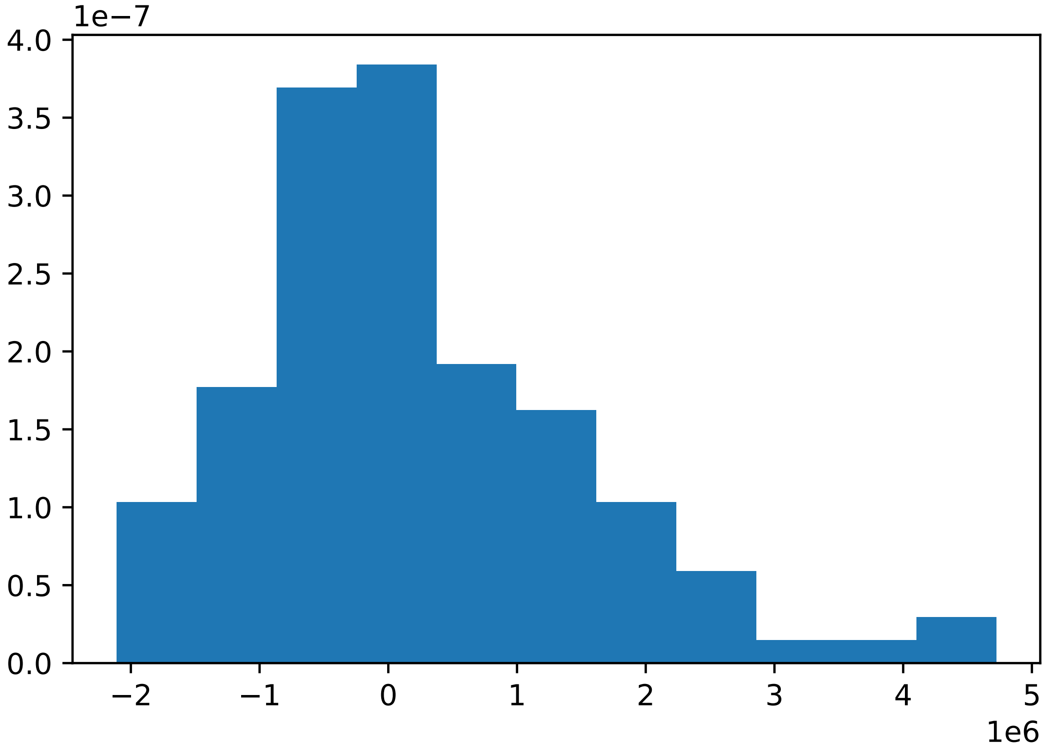


tor variables should not exhibit high correlation, which might lead to unstable coefficient estimates.[Ins23]

Figure 4 shows a scatter plot of predicted values versus actual values. The predicted values are on the horizontal axis and the actual values are on the vertical axis. Each dot on the plot represents a single data point. The data points are tightly clustered around a straight line, which indicates a linear relationship between the two variables. The plot shows a strong positive correlation between the predicted and actual values, meaning that as the predicted values increase, the actual values also tend to increase. This linearity check indicates that the machine learning model can make somewhat accurate predictions.

Figure 5 shows the relationship between the predicted values and the residuals. If the residuals are evenly spread out around the zero line, regardless of the predicted values, then homoskedasticity is present. In this plot, it shows a slight fanning out of the residuals as the

Figure 6: Residuals Distribution



predicted values increase. This suggests that the variance of the error term may be increasing with the predicted values, which would be a violation of the homoskedasticity assumption.

Figure 6 shows the distribution of residuals in the model. The mean of the residuals is close to zero, the distribution is symmetrical around the mean, and the residuals are evenly spread out around the mean. This suggests that the regression model is well-specified and that the assumptions of normality are met. Overall, the residuals distribution plot in the image you provided is a good sign. It suggests that the regression model can make accurate predictions.

Table 3 shows the correlation between each independent variable and the other independent variables in the model. A VIF value greater than five indicates that there is a high degree of correlation between the variable and the other variables in the model. This can lead to problems with multicollinearity, which can affect the accuracy of the model.

Table 3: VIF Results

Variable	VIF
Area	1.325
Bedrooms	1.3695
Bathrooms	1.2867
Stories	1.4781
Main road	1.1727
Guestroom	1.2128
Basement	1.3231
Hot water heating	1.042
Air conditioning	1.2118
Parking	1.2128
Prefarea	1.1492
Status furnished	8.5758
Status semi-furnished	12.4043
Status unfurnished	8.8262

5 Discussion and Conclusion

In summary, this paper makes a substantial contribution to the existing body of knowledge within the specified research domain by reinforcing and validating findings from prior research. The study emphasizes specific aspects where it aligns with or extends conclusions drawn in earlier works. In the final stages of analysis, a rigorous comparison with a Kaggle reference is undertaken, to surpass the predictive accuracy of the existing model. This comparative analysis serves as a crucial validation step, providing insights into potential areas for improvement and efficiency enhancement. The paper aspires to advance predictive modeling in housing price prediction beyond current benchmarks, contributing to the evolving landscape of statistical methodologies in the field.

One limitation of this current study is that this data set is very small and has very limited variables. The next step for this study is to gain access to a data set from the actual house market and produce a model with it. Another factor that can limit this study is that nowadays, there are infinite numbers of factors that can affect the cost of a house. So creating a model to include all these new factors is practically impossible.

Over the past decade, the United States has experienced three recessions, highlighting the dynamic nature of its economic landscape. Beyond the traditional focus on housing prices, the imperative now extends to constructing a more robust predictive model with the precision to forecast overall market trends. This expanded model not only discerns fluctuations in the real estate market but also holds the potential to anticipate broader economic shifts. The ultimate goal is to transcend the confines of predicting housing prices alone and, instead, develop a tool that can provide insights into the timing and indicators of future recessions. In this way, the enhanced model becomes a strategic asset for decision-makers navigating the intricacies of economic cycles.

The logical progression for advancing this study involves a strategic expansion of the data set, with a particular focus on capturing a more extensive and diverse range of housing market dynamics. The envisioned next step entails acquiring a larger data set, extending beyond the current scope, and delving into various communities across the United States. Recognizing the inherent socioeconomic variations among different communities, particularly disparities in income brackets, opens up an opportunity for analysis and model refinement.

By systematically gathering data from distinct communities, each characterized by its unique economic landscape, we can aim to create a more comprehensive and representative data set. This approach can enable new researchers to examine how local economic factors and community-specific variables contribute to variations in housing prices. The inherent diversity in income levels and economic conditions across communities provides a rich terrain for exploration. This multifaceted analysis can uncover insights into the nuanced relationships between socioeconomic factors, geographical location, and housing market dynamics. It may unveil community-specific variables that exert a significant influence on housing prices.

References

[Alf+17] Adyan N Alfiyantin et al. “Modeling House Price Prediction using Regression Analysis and Particle Swarm Optimization”. In: *International Journal of Advanced Computer Science and Applications* 8.10 (2017).

[Ins23] CFA Institute. *Basics of Multiple Regression and Underlying Assumptions.* Accessed: 2023-12-03. 2023. URL: `https://www.cfainstitute.org/en/membership/professional-development/refresher-readings/multiple-regression`.

[JN05] Dag H. Jacobsen and Bjørn E. Naug. “What Drives House Prices?” In: *Economic Bulletin* (2005).

[Kas18] Eiiti Kasuya. “On the use of r and r squared in correlation and regression”. In: *Ecological Models and Data in R* 34.1 (2018), pp. 235–236.

[SRR21] Akhilendra P Singh, Kartikey Rastogi, and Shashank Rajpoot. “House Price Prediction Using Machine Learning”. In: *2021 3rd International Conference on Advances in Computing, Communication Control and Networking (ICAC3N).* 2021, pp. 203–206. DOI: `10.1109/ICAC3N53548.2021.9725552`.

[UG13] Gulden K Uyanik and Nese Guler. “A Study on Multiple Linear Regression Analysis”. In: *Procedia - Social and Behavioral Sciences* 106 (2013). 4th International Conference on New Horizons in Education, pp. 234–240. ISSN: 1877-0428. DOI: `https://doi.org/10.1016/j.sbspro.2013.12.027`.