Predicting the Housing Market using Linear Regression

Garrick Ho
Jun Yan

Department of Statistics
University of Connecticut

October 29, 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 robust linear regression model by using machine learning techniques to predict housing costs accurately. By leveraging natural language processing, the linear regression model analyzes home property descriptions and extracts vital insights that significantly contribute to precise pricing predictions. This paper ends with an explanation of the limitations of this study and potential future works that can be done.

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

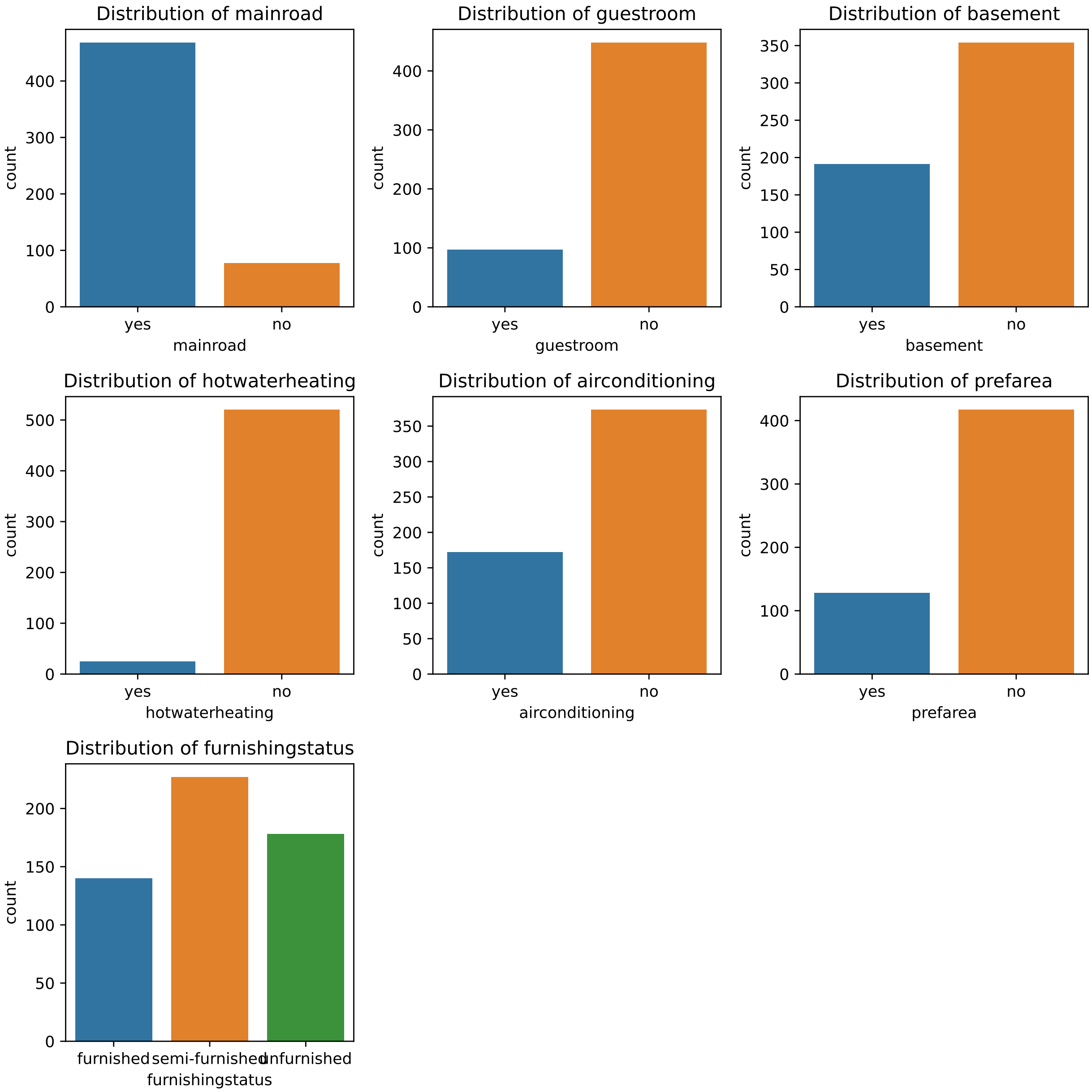
1 Introduction

The rest of the paper is organized in the following way. Beginning with the overall introduction located in Section 1. A brief introduction about the data will be shown in Section 2. Here, the variables will be described and explained how they contribute to the research question. Then, the methodology used in this paper will be explained in Section 3. Next, the simulation will be reported in Section 4. This section will present the results from the model and have further analysis with tables and figures. The applications of the results found will be in Section 5. 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 6. The limitations of the current study and future direction will be discussed here. All the references used in this paper will be included in Section 7.

2 Data

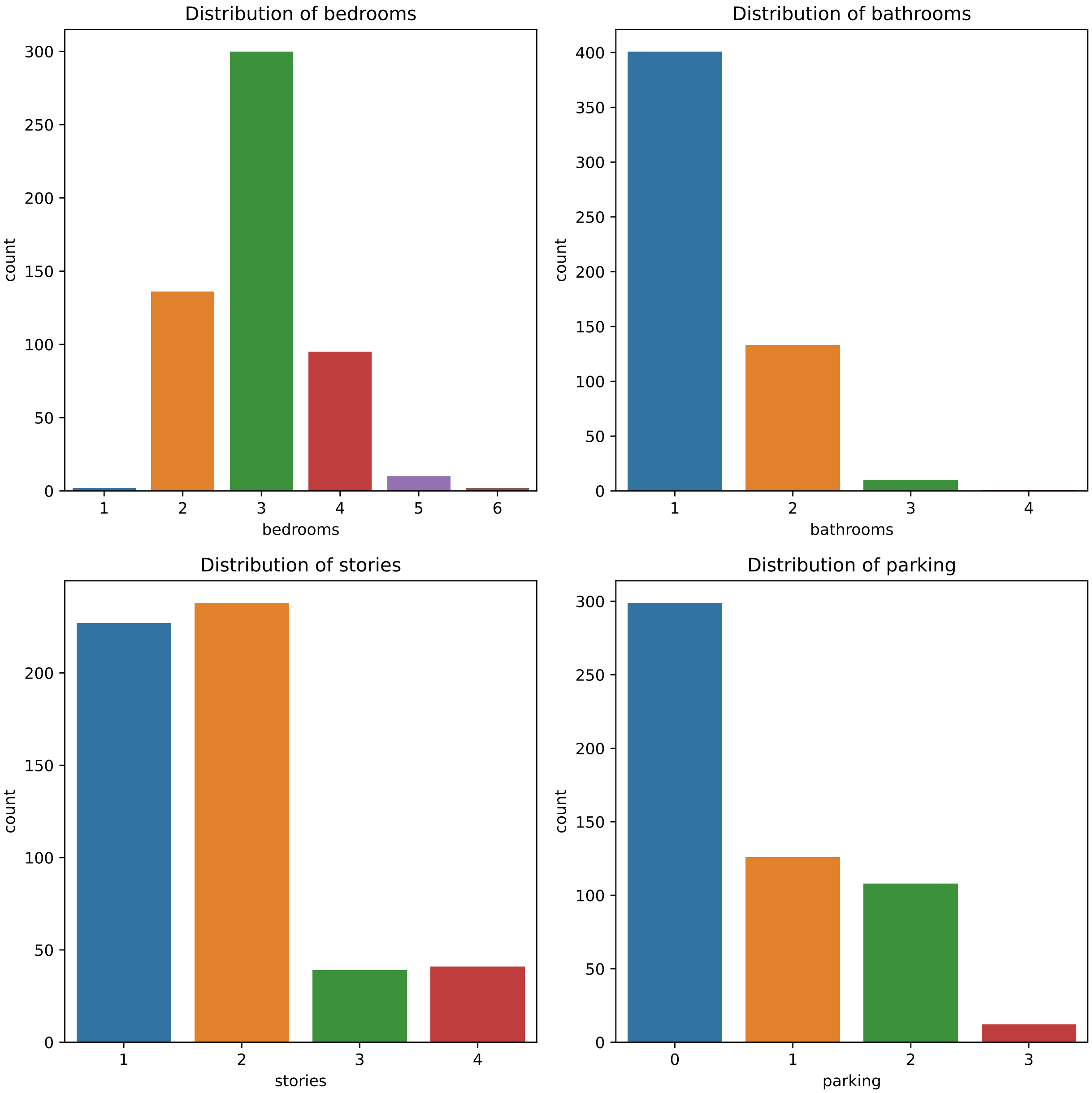
The data that will be used for this study was sourced from Kaggle. It is a data set that contains $n = 545$ observations (different house samples) of thirteen variables, seven of which are categorical variables and six are numerical variables. The data set contains features and corresponding labels for training and testing a Multiple Linear Regression (MLR) model to predict the house cost. The data was explored and visualized using the pandas library in Python. For the columns that were strings, they have been transformed into an integer data type so that they can be read in the model. More specifically, the 'furnishingstatus' column was broken down into three different columns. One for furnished, one for semi-furnished, and lastly one for unfurnished.

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



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.

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

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

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.

4 Simulation

5 Application

6 Discussion and Conclusion

7 References

[2] [1]

References

[1] Adyan N Alfiyantin, Ruth E Febrita, Hilman Taufiq, and Wayan F Mahmundy. Modeling house price prediction using regression analysis and particle swarm optimization. *International Journal of Advanced Computer Science and Applications*, 8(10), 2017.

[2] 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)*, pages 203–206, 2021. doi: 10.1109/ICAC3N53548.2021.9725552.