First-Time Home Buyers Guide: Predicting How Much a House Costs with Certain Demographics

Brandon T. Shellenberger

Northwest Missouri State University, Maryville MO 64468, USA
S570782@nwmissouri.edu

Abstract. Adding abstract later

Keywords: data analytics · housing prices · first-time home buyers · linear regression · predictive models

1 Introduction

It is always difficult to sell and buy a home, especially when there are many options, from the number of bedrooms/bathrooms to how big the entry property is. It can begin to confuse people who are buying their house for the first time. As an added stress, the Federal Housing Administration (FHA) recommends that first-time buyers must have a 3.5% down payment when receiving a loan through their program. [8] Since the prices of homes have increased, the down payment on the loan has also increased. In 2018, only 12% of US counties were considered unaffordable based on the median price, while in 2024, that number rose to more than 30% of US counties. [5] This report will show the basics of finances when planning to buy a home and is aimed at first-time buyers.

The data being used will mostly come from Zillow's data page. [4] An example spreadsheet shows the median price of homes in different regions from January 2018 to January 2025. This data is readily available to anyone and is updated monthly. Zillow uses the median instead of the average because the median is not easily affected by outliers and is closer to the majority, especially when working with large data. Other data sources will be used from Kaggle.

1.1 Problem to Analyze

I want to use this report as a guide for first-time home buyers when deciding the best time to buy a house. I will be looking at regression models to help predict the sale prices for houses and what to expect for a down payment. Because each region has different prices, I will pick a few of the major areas to perform the analysis.

After collecting the data from sources, the data will be cleaned and will only have valuable information. I will group certain columns for more precise analysis and use these columns in the prediction models. I will also be showing a time-series chart with predictions as the visualization.

A couple of key components will be to try different columns to figure out the best scores. I will be using the number of bedrooms/bathrooms, but other columns might have a greater impact on prices. Another key component will be the predictive aspect. This will be challenging since I will be using linear regression on multiple regions.

2 Methodology

This section will walk through the process of the data being used, cleaning the data for proper analysis, and running the analytics to discover answers. After this section, I will compare the results and talk about the limitations.

2.1 Dataset

There are two different datasets I will be using throughout the pre-processing phase. The first is the data from Zillow. [4] This online company is used by customers to ”shop” for buying/renting a house/apartment, and they have a separate webpage dedicated to storing data for public use. The multiple CSV files are updated monthly. I have downloaded the file with the median monthly price of single-family homes. To go into more detail, here are a few of the attributes I will be using for the analysis:

- SizeRank - This shows how large the city is by population with 0 being the largest. I will use this to help separate the large and small cities of the 940 cities in total.
- RegionName - Name of all the cities including their state.
- All Dates - This is a monthly time series starting in January 2000 and moving up a month for each column.

The second dataset is from the Kaggle page titled ’Housing Price Prediction’. [7] I will be using a package called ’Kagglehub’ to extract the data from the webpage. Each time the program runs, this package will handle searching for the data and moving the data into the correct directory for analysis. This dataset has thirteen columns; some are continuous/discrete numerical values and others are boolean. The columns I will be focused on are:

- price - This column is the price of each home.
- bedrooms - Number of bedrooms
- bathrooms - Number of bathrooms
- basement - This is a boolean value of the house having a basement.

The Zillow dataset will be used to determine future housing prices and the Kaggle dataset will be used for pricing based on certain criteria.

2.2 Data Cleaning

This section is where the data will be curated to fit the analysis process. Curating is the organization and management of datasets to meet the needs of a specific analysis. [3] I used Excel for simple curating and a notebook for more complex cleaning and structuring.

I had to do some work on the Zillow dataset to meet the needs of my analysis. When I loaded the dataset to my root directory, the attributes were the months and the records were each city. As I looked closer at the data in Excel, I noticed that a lot of the cities did not start recording their values until a certain date at random. An example would be that New York would start in January 2000 (the beginning of the dates) and Indianapolis would start in June 2003. After each city's first recorded month, there were no gaps. If I dropped all the cities with null values, there would be about 300 records from the nearly 940 at the beginning. Instead of dropping these records, I decided to cut back on the number of months and drop all dates from January 2000 to January 2016. Afterward, I used Python to drop the columns 'RegionID', 'StateName', 'RegionType', and 'SizeRank' and transposed it to better fit the regression models.

The Kaggle dataset was less curated with no duplicate/null values. The attributes I deleted were 'area', 'mainroad', 'guestroom', 'hotwaterheating', 'parking', and 'prefarea'. Because the models I want to use in the analysis only use numerical values, I had to split some attributes using a One-Hot Encoding method. This is where a categorical column can be split into separate columns, one for each of the categories found in the original column. For example, the attribute 'basement', which determines if there is a basement in the house, is either yes or no as in Table 2. This attribute is split using the Python package Pandas.get_dummies() [1], splitting the column into basement_yes and basement_no and inputting a 1 for yes and 0 for no as in Table 3. Other attributes that were curated this way are the 'airconditioning' and 'furnishingstatus'.

Table 1: One-Hot Encoding Example

Table 2: Original Table

basement
yes
no
yes
no
no

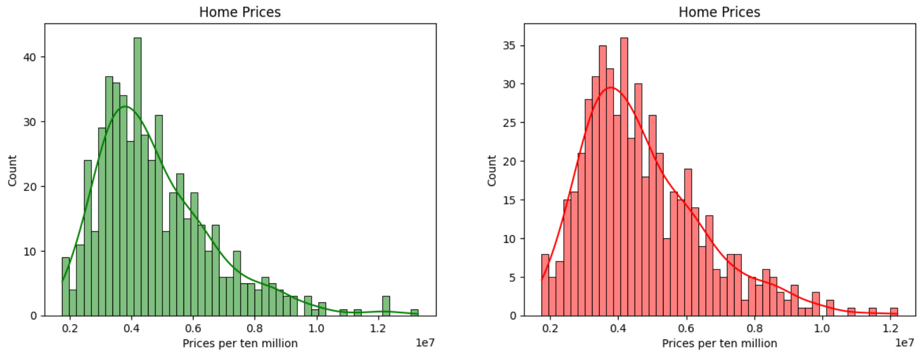
Table 3: After One-Hot Encoding

basement_yes	basement_no
1	0
0	1
1	0
0	1
0	1

After cleaning both datasets, I went back and looked for outliers in the Kaggle dataset. Graphing the price frequency, shown in figure 1a, some prices are larger and might be considered outliers. To determine if these prices were outliers, two methods were used: Inner Quartile Range (IQR) and Z-scores. The IQR is the range above the smallest and below the largest quartiles, Q1 and Q4, respectively, also considered as the middle half of the data [6]. To be an outlier with the IQR method, the values will be smaller or larger than 1.5 times the IQR. After running this analysis, there were 15 outliers and it was determined that there were too many to be deleted from the dataset. The Z-score outliers are determined by the formula below, where x is the data point, μ is the mean of the dataset, and σ is the standard deviation of the dataset.

$$z = \frac{(x - \mu)}{\sigma} \tag{1}$$

Since this is a small dataset, if the Z-score for a data point is above 4 or below -4, it is considered an outlier. [9] With this process, there were 3 outliers. I deleted the outliers found by the Z-score and have shown the price frequencies in Figure 1b.



(a) Before removal of outliers

(b) After removal of outliers

Fig. 1: With and without outliers

2.3 Analysis

The methods being used in the analysis consist of a spread of regression models: multiple linear regression, polynomial regression, decision tree regressor, and random forest regressor. The scores used during the analysis will be the Root Mean Squared Error (RSME) and the Coefficient of Determination R^2. Each model will first train 80% of the data and test the remaining 20%. When running the models, there will be scores for training and testing.

Multiple Linear Regression The multiple linear regression will fit the best linear equation to the dataset. Because there is more than one feature being used, this model will have numerous coefficients, one for each of the features.

Linear regressions are best for linearly scaled data sets. Looking at Table 4, the model does not help predict prices with a high average and low correlation.

Table 4: Multiple Linear Regression

Type	RSME	R^2
Train	1136595	0.5922
Test	1227494	0.5246

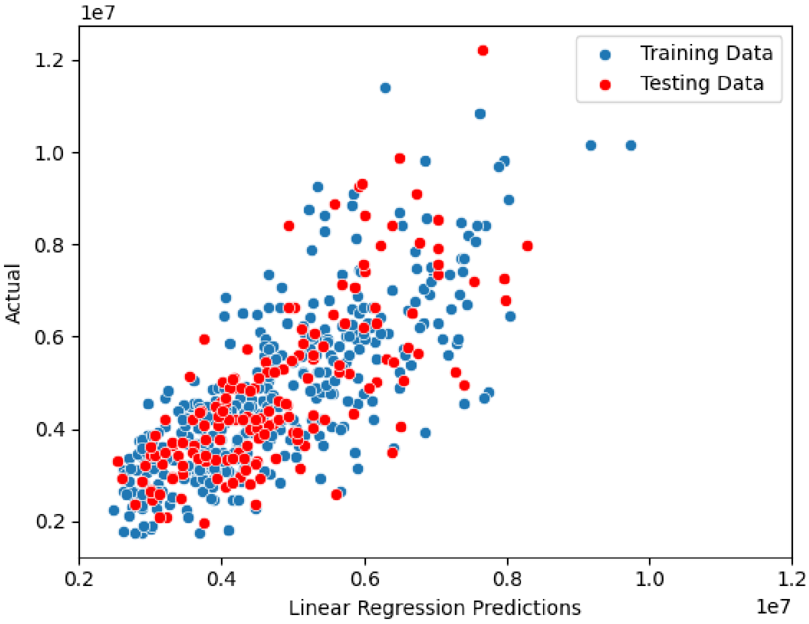


Fig. 2: Predicted vs. Actual

Polynomial Regression This model is similar to linear regression, but instead of a linear relationship, the relationships could be of a higher degree of freedom. In this analysis, the polynomial regression model used degrees of 2, 3, and 4 where the scores are in Table 5. The degree is the highest exponent found in the equation and will have degree-1 direction changes to account for possible fluctuations in the data. Looking at the scores, R^2 from the training sets increases as the degree increases. This cannot be said for the testing sets where it decreases at degree 3, then begins to increase. Anything higher than a degree of 4, the models will overfit the data. Figure 3 shows the predicted vs. actual with a degree of 4.

Table 5: Polynomial Regression

Description	Type	RSME	R^2
Degree 2	Train	1067841	0.6401
Degree 2	Test	1194405	0.5498
Degree 3	Train	964012	0.7067
Degree 3	Test	1309440	0.4600
Degree 4	Train	952656	0.7135
Degree 4	Test	1172812	0.5660

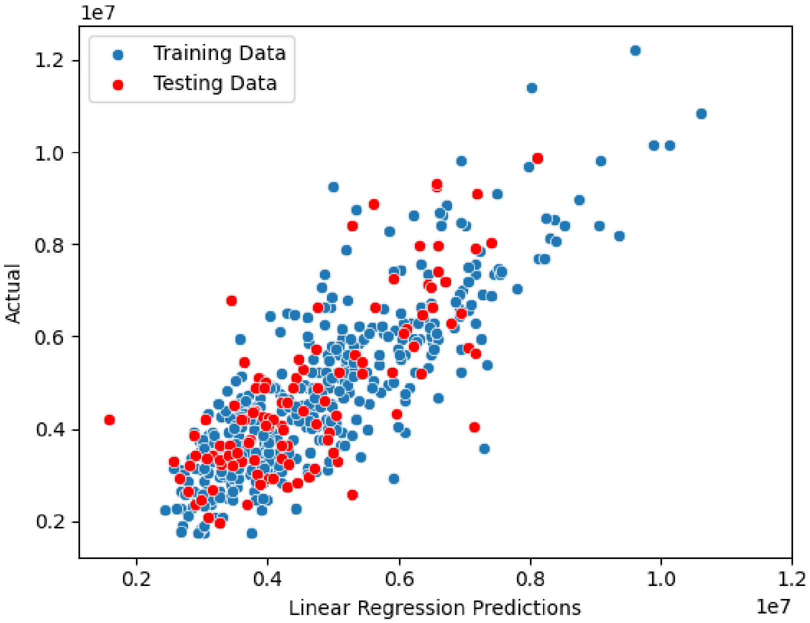


Fig. 3: Predicted vs. Actual

Decision Tree Regressor This is a model that will make a decision based on statistical probabilities. This model looks like a tree, splitting into leaf nodes based on this probability. The model has an attribute for max depth which is the maximum number of rows/branches the model can go. In Table 6, models higher than a maximum depth of 5 start to overfit the data. Figure 4 shows the predicted and actual values with a maximum depth of 5 branches.

Table 6: Decision Tree Regressor

Maximum Depth	Type	RSME	R^2
3 branches	Train	1222539	0.5282
3 branches	Test	1256334	0.5020
4 branches	Train	1117445	0.6058
4 branches	Test	1239294	0.5154
5 branches	Train	1013031	0.6761
5 branches	Test	1216762	0.5328

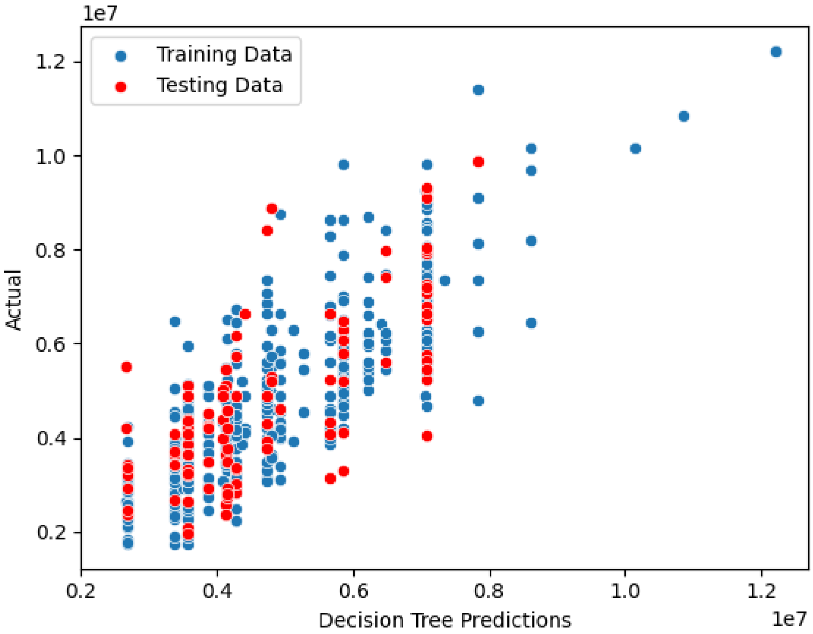


Fig. 4: Predicted vs. Actual

Random Forest Regressor This model uses multiple decision trees and averages the results to help control overfitting of the data. There are two main attributes with this model, maximum depth (same as decision tree models) and estimators which determine the number of trees in the forest [2]. The difference in scores did not change much when the number of estimators changed. Knowing this, Table 7 will show the scores while keeping the estimators at 100 and changing the maximum depth. Figure 5

Table 7: Random Forest Regressor (est = 100)

Maximum Depth	Type	RSME	R^2
3 branches	Train	1139001	0.5902
3 branches	Test	1216965	0.5327
4 branches	Train	1027976	0.6664
4 branches	Test	1172175	0.5664
5 branches	Train	912367	0.7372
5 branches	Test	1159471	0.5758

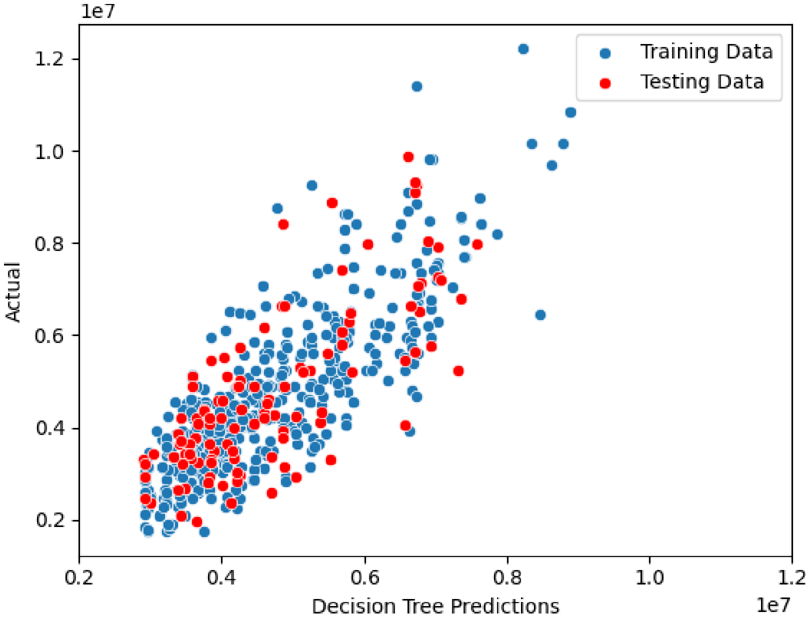


Fig. 5: Predicted vs. Actual

In the next section, we will compare the top results from each model.

3 Results

Comparing Results

4 Limitations

Problems/Holdbacks

5 Conclusion

Summary of the project

References

1. pandas.get_dummies &x2014; pandas 2.2.3 documentation — pandas.pydata.org. https://pandas.pydata.org/docs/reference/api/pandas.get$_dummies.html, [Accessed31-03-2025$]
2. RandomForestRegressor — scikit-learn.org. https://scikit-learn.org/stable/modules/generated/sklearn.ensemble.RandomForestRegressor.html, [Accessed 18-04-2025]
3. What Is Data Curation? Why Is It Important? — Alation — alation.com. https://www.alation.com/blog/what-is-data-curation/, [Accessed 30-03-2025]
4. (Feb 2025), https://www.zillow.com/research/data/
5. Farha, C., McCoy, J., Rodziewicz, D.: First-time homeownership became less affordable across most of the united states in recent years. Economic Bulletin pp. 1–4 (2025)
6. Frost, J.: Interquartile Range (IQR): How to Find and Use It — statisticsbyjim.com. https://statisticsbyjim.com/basics/interquartile-range/, [Accessed 06-04-2025]
7. KUMARdatalab, H.: Housing price prediction (Jul 2023), https://www.kaggle.com/datasets/harishkumardatalab/housing-price-prediction
8. Lee, D., Tracy, J.: Are first-time home buyers facing desperate times? Liberty Street Economics (2025)
9. Shiffler, R.E.: Maximum z scores and outliers. The American Statistician **42**(1), 79–80 (1988)