# 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. [4] 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. [2] 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. [1] 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 way for first-time home buyers when 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. [1] 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.
- StateName - 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 Kaggle page titled 'Housing Price Prediction'. [3] 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 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

Making the data fittable

### 2.3 Analysis

Analyzing data based on the essential question

## 3 Results

Comparing Results

## 4 Limitations

Problems/Holdbacks

## 5 Conclusion

Summary of the project

## References

1. (Feb 2025), https://www.zillow.com/research/data/
2. 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)
3. KUMARdatalab, H.: Housing price prediction (Jul 2023), https://www.kaggle.com/datasets/harishkumardatalab/housing-price-prediction
4. Lee, D., Tracy, J.: Are first-time home buyers facing desperate times? Liberty Street Economics (2025)