Housing Prices in the Northeast US: The Impact of Cities

By Gavin Benedict

Data Source

Summary: Utilizing a open-source file from Kaggle that includes the price of houses through most of the United States of America in the year 2023. The data set aligns with the project aim of determining whether proximity to cities increases the price of housing.

Data Collection: The original data comes from realtor.com, an aggregate site that tracks housing sales. The original data has information from each of the states, but most of the data focuses on the Northeastern US, so I cleaned the data to better focus on this region.

Why This Dataset: This dataset gives an excellent view of the current state of the housing market broken down into zip codes, and includes helpful information on the houses such as the number of rooms and the amount of land attached to the house.

Data Profiing

Data Cleaning

Below are the steps I took to clean the data:

1. After loading the data I determined that there were several states with only a few data points, (The Virgin Islands, Georgia, Puerto Rico, Virginia, South Carolina, Tennessee, Wyoming, and West Virginia) so I set out to first change the data types for the status, city, and state columns to a string, which allowed me to drop the above states.
2. Next I noticed that each borough of New York city was listed as the city, eg. Bronx, to ensure consistency, I renamed the boroughs to be listed as New York City in the city column.
3. Next, I dropped houses where there were either 0 or NaN beds, since a house without bedrooms is likely not a house and is instead a parcel of land for sale.
4. Dropped houses with a price above $1 million, since I am more interested in single family and starter homes. This eliminated a lot of the houses in Massachussettes, strangely enough.
5. Dropped houses with the 'ready to build' status, before dropping the status column to eliminate extra empty space.
6. Dropped houses with more than 6 bedrooms, as these are likely either mansions or apartment complexes, neither of which this analysis was interested in.
7. Finally, I imputed the average value for each column in place of NaN.

Data Description

Column Name	Description	Qualitative / Quantitative	Discrete / Continuous	Nominal / Ordinal / Binary
bed	The # of bedrooms	Quantitative	Discrete	
bath	The # of bathrooms	Quantitative	Discrete	
acre_lot	The amount of land attached to the house in acres	Quantitative	Continuous	
city	The name of the city where the house resides	Qualitative		Nominal
state	The name of the state the house resides in	Qualitative		Nominal
zip_code	The zip code for the house	Qualitative		Nominal
house_size	The square feet of space within the house	Quantitative	Continuous	
price	How many dollars the house sold for	Quantitative	Continuous	

Data Limitations and Ethics

Limitations

The largest limitation for the data is that it only has information for 2023 and largely focuses on the Northeastern USA. This allows us to have a clear snapshot of the current housing costs within this area, but does not indicate where this stands historically for the region.

To aid with this I also included a data set from Nasdaq that focuses on Zillow house prices in the US from 2006 to 2023. This gives a better idea of the trend in housing costs for the last twenty years.

Ethics

This data does not feature any PI, meaning there are no ethical concerns.

Project Goals and Questions

Goals:

- Determine what attribute has the largest impact on house prices.
- Use housing prices to categorize the cost of living between states.

Key Questions:

- What impacts housing prices the most between the number of bathrooms and bedrooms, the square footage of the house, the amount of acreage, or proximity to cities?
- What characteristics of a house have the least impact on price?
- Have housing prices changed in the last decade?
- Which state is the most-friendly to new families in terms of housing prices?