Impact of Population Shifts on the Housing Market

ShaNiece Twitty and Maria Berg

2025-03-05

Abstract:

This project aims to analyze the growth and decline of the Housing Market in five different cities using population data.We will collect housing data from Redfin, as well as download data from the United States Census Bureau correlated with our chosen five cities from the years 2012 to 2023. With this data, we aim to identify trends and relationships between housing market activity and population changes. We will explore how factors like home sales, pricing patterns, market inventory, and demographic shifts interact with population growth or decline. By analyzing these connections, we hope to gain insights into the broader dynamics influencing the housing market in our selected cities over time. If all is successful, we expect to uncover patterns indicating how demographic changes influence housing affordability and availability.

Introduction:

For our project, we have collected a dataset focused on housing market trends, specifically analyzing the impact of population growth or decline on key real estate metrics. This dataset provides insights into many different housing metrics and population demographics from the years 2012-2023 across various regions over time. We are interested in this data because understanding housing market dynamics in relation to population shifts can help predict market trends, support planning, and guide real estate investments. By analyzing how housing prices and sales activity respond to population changes, we can identify patterns that indicate whether an area is experiencing growth, stability, or decline.

The following is a head of select columns from our data followed by a table with all of our variables and descriptions:

Region	Month.of.Period.End	Median.Sale.Price	PopEst_TOTAL POPULATION
Boston	2012-01	303000	637516
Boston	2012-02	290000	637516
Boston	2012-03	305000	637516
Boston	2012-04	324000	637516
Boston	2012-05	340000	637516
Boston	2012-06	375000	637516

Table 2: Description of Variables

Variable	Description
Region	The name of the metro area where the data was collected.
Median Sale Price	The median sale price of homes in the region for that month.

Variable	Description
Median Sale Price MoM	The month-over-month percentage change in the median sale price.
Median Sale Price YoY	The year-over-year percentage change in the median sale price.
Homes Sold	The total number of homes sold in the region during that month.
Homes Sold MoM	The month-over-month percentage change in the number of homes sold.
Homes Sold YoY	The year-over-year percentage change in the number of homes sold.
New Listings	The number of new homes listed for sale during the month.
New Listings MoM	The month-over-month percentage change in new listings.
New Listings YoY	The year-over-year percentage change in new listings.
Inventory	The total number of active home listings available in the region.
Inventory MoM	The month-over-month percentage change in housing inventory.
Inventory YoY	The year-over-year percentage change in housing inventory.
Days on Market	The median number of days homes stayed on the market before being sold.
Days on Market MoM	The month-over-month change in the median days on the market.
Days on Market YoY	The year-over-year change in the median days on the market.
Average Sale To List	The average percentage of the listing price that homes sold for.
Average Sale To List MoM	The month-over-month percentage change in the sale-to-list price ratio.
Average Sale To List YoY	The year-over-year percentage change in the sale-to-list price ratio.
Month of Period End	The month and year of the data. 'MM-YY' format.
PopEst_TOTAL POPULATION	The estimated total population of the given area.
PopEst_18 YEARS AND OVER	The estimated population aged 18 and older.
PopPct_18 YEARS AND OVER	The percentage of the population that is 18 years or older.
PopEst_TWO OR MORE RACES	The estimated population identifying as two or more races.
PopPct_TWO OR MORE RACES	The percentage of the population identifying as two or more races.
PopEst_HISPANIC OR LATINO (OF ANY RACE)	The estimated population identifying as Hispanic or Latino.
PopPct_HISPANIC OR LATINO (OF ANY RACE)	The percentage of the population identifying as Hispanic or Latino.
PopEst_WHITE ALONE	The estimated population identifying as White alone.

Variable	Description
PopPct_WHITE ALONE	The percentage of the population identifying as White alone.
PopEst_BLACK OR AFRICAN AMERICAN ALONE	The estimated population identifying as Black or African American alone.
PopPct_BLACK OR AFRICAN AMERICAN ALONE	The percentage of the population identifying as Black or African American alone.
PopEst_AMERICAN INDIAN AND ALASKA NATIVE ALONE	The estimated population identifying as American Indian or Alaska Native alone.
PopPct_AMERICAN INDIAN AND ALASKA NATIVE ALONE	The percentage of the population identifying as American Indian or Alaska Native alone.
PopEst_NATIVE HAWAIIAN AND OTHER PACIFIC ISLANDER ALONE	The estimated population identifying as Native Hawaiian or Other Pacific Islander alone.
PopPct_NATIVE HAWAIIAN AND OTHER PACIFIC ISLANDER ALONE	The percentage of the population identifying as Native Hawaiian or Other Pacific Islander alone.
PopEst_SOME OTHER RACE ALONE	The estimated population identifying as some other race alone.
PopPct_SOME OTHER RACE ALONE	The percentage of the population identifying as some other race alone.

Methods:

To collect the data, we accessed publicly available CSV files from the U.S. Census Bureau for demographic information and obtained housing data from the Redfin Data Center. The main challenges included the time-consuming process of downloading individual city and year-specific CSV files, data cleaning (e.g., removing unsupported Unicode characters in R and reformatting dropdown-style data into single columns with corresponding dates), and aligning datasets across different time frames while integrating data from two separate sources. Another challenge we will face is the lack of reliable census data for the year 2020, therefore there will be a "hole" in our trends.

Once cleaned, the data will be analyzed using:

Exploratory Data Analysis (EDA): Summary statistics and visualizations to identify trends.

Correlation Analysis: Investigating the relationship between population growth and housing trends.

Time Series Analysis: Assessing how demographic shifts impact housing prices over time.

Predictive Modeling (if feasible): Using regression analysis to forecast potential future trends.

Expected Outcomes:

If successful, this study will reveal the trends in the housing market based on the population of each city and how much or how little they affect each other. We anticipate finding that shifts in the population and demographics will affect the housing market and/or vise versus. The results could help predict Housing prices or other key housing metrics.As well as provide insides into long term urban development and inform housing policy considerations.

Group Considerations:

Data Collection: (MB - 50%, ST - 50%)

Data Wrangling and Cleaning: (MB - 50%, ST - 50%)

Analysis: (MB - 50%, ST - 50%))

Visualization & Modeling: (MB - 50%, ST - 50%))

Forecasting??? (ST-100%)

GitHub/R Project creation: (MB - 75%, ST - 25%)

Report Writing and Final Presentation: (MB & ST- 100%)