**Factors Influencing Housing Prices in Major U.S. Cities.**

Bernard Owusu Sefah

Data Science Bellevue University

DSC 520: Stats for Data Science

Week 9 Assignment

5/01/2024.

Introduction: I want to research factors influencing housing prices in major U.S. cities. Understanding the key drivers of home values is essential for real estate investors, homebuyers, sellers, and policymakers. It is a data science problem because it analyzes datasets with many variables to uncover patterns and build predictive models.

**Research Questions**

1. What are the most essential features in predicting home values in different cities?
2. How does the influence of factors like square footage, number of bedrooms, etc., vary across regions?
3. Can we build an accurate model to estimate home prices based on property characteristics?
4. How have the key drivers of home prices changed over time?
5. Are there any leading indicators that tend to precede major shifts in home values?
6. Can clustering homes by their features reveal distinct market segments?
7. How do socioeconomic and demographic factors of a neighborhood impact home prices?
8. What role do macroeconomic variables like interest rates and job growth play?

**Approach**: I plan to gather home sales and characteristics data for several major cities. I will merge this with data on economic and demographic factors by neighborhood. I will use exploratory analysis and visualization to identify key relationships and then build predictive models using multiple regression, decision trees, and clustering techniques. I will evaluate model performance and refine the feature set.

This approach should provide a solid understanding of the main variables driving home prices and how well we can predict prices. It will enable the development of valuation models while identifying areas needing further research.

**Data**

1. Zillow Research Housing Data: Includes home values, listings, sales prices, and rental data by ZIP code and city from 2008 to the present (<https://www.zillow.com/research/data/>).
2. American Community Survey (ACS) Census Data: Provides socioeconomic and demographic data by ZIP code, including income, education levels, age, and occupations (<https://www.census.gov/programs-surveys/acs>).
3. Federal Reserve Economic Data (FRED): Macroeconomic indicators like mortgage rates, GDP growth, unemployment rate, etc. (<https://fred.stlouisfed.org/>).
4. Redfin Housing Market Data: Another source of home sales data with fields like sale price, days on the market, price per sq ft, etc. (<https://www.redfin.com/news/data-center/>).

**Required Packages**

* readxl for reading Excel files.
* dplyr for data manipulation.
* ggplot2 for data visualization.
* caret for machine learning modeling.
* cluster for clustering analysis.
* sf for handling geospatial data.
* lubridate for working with dates and times.
* corrplot for visualizing correlations.
* tidyr for data tidying.
* stringr for string manipulation.

**Plots and Tables**

* Scatter plots of home price vs. sq footage, # beds, etc.
* Line plots showing price trends over time.
* Heatmaps of correlations between numeric features.
* Bar charts of average price by categorical variables like ZIP code.
* Maps displaying geographic variation in prices and critical factors.
* Tables summarizing model performance metrics.

**Questions for Future Steps**

* What are the best practices for merging datasets from different sources/granularities?
* How do I handle missing values and outliers?
* What techniques are helpful for feature selection?
* How do I tune the hyperparameters of the ML models?
* What are good ways to visualize geographic data?

**Data Import and Cleaning**

I did a thorough process of importing and cleaning multiple datasets related to housing prices, economic indicators, and demographic data. Here's a summary of the data preparation and cleansing steps:

1. The necessary libraries were loaded, including readr (changed to use to this instead of the readxl), dplyr, ggplot2, tidyr, lubridate, corrplot, ggrepel, and anytime.
2. The datasets were read into R using the read\_csv() function from the readr package. This included housing\_data, income\_data, unrate\_data, demand\_data, city\_data, and aspus\_data.
3. The date columns in unrate\_data, demand\_data, and aspus\_data were verified and adjusted to ensure they were in the correct format using as.Date().
4. The housing\_data was cleaned and prepared by converting the Date column to character, parsing it using dmy() from lubridate, renaming columns for clarity, converting Median\_Sale\_Price to numeric, and selecting relevant columns.
5. The income\_data was prepared for merging by converting the City column to lowercase and converting the Median column to numeric.
6. The city\_data was cleaned and prepared by converting the Date column to the correct format, separating the RegionName column into City and State, and converting the City column to lowercase.
7. All datasets were merged based on the DATE column using left\_join().
8. Missing numeric data were imputed with median values using ifelse() and mutate(across()) from dplyr.

All this was done to ensure that the final dataset (full\_data\_imputed) is ready for my analysis.

To provide a concise overview of the final dataset without printing the entire data frame, I used glimpse(full\_data\_imputed) to show the dataset since my dataset combined a lot of other datasets.

**What information is not self-evident?**

Some information that may not be immediately apparent from the data includes:

1. The underlying factors driving the variation in housing prices across different cities and regions.
2. The presence of any seasonal patterns or long-term trends in housing prices.
3. The impact of economic indicators, such as unemployment rate or GDP, on housing prices.
4. The relationship between housing prices and demographic factors, such as population growth or income levels.
5. The existence of any spatial dependencies or neighborhood effects on housing prices.

These are the types of insights that require further analysis and exploration of the data to uncover.

**What are different ways you could look at this data?**

There are several ways to approach and analyze this data to answer the research questions:

1. Geographical analysis: Examine housing prices and related factors at different geographic levels, such as by city, region, or zip code, to identify spatial patterns and variations.
2. Time series analysis: Explore the data over time to identify trends, seasonality, and any structural changes in the housing market.
3. Segmentation analysis: Segment the data based on property characteristics, such as number of bedrooms or property type, to understand their influence on housing prices.
4. Economic analysis: Investigate the relationship between housing prices and economic indicators, such as unemployment rate, GDP, or consumer sentiment.
5. Comparative analysis: Compare housing market dynamics across different cities or regions, or before and after significant events like the 2008 financial crisis or the COVID-19 pandemic.

How do you plan to slice and dice the data? To create new summary information and gain deeper insights, the code demonstrates slicing and dicing the data in different ways:

1. Filtering the data by region, city, or time period to focus on specific subsets of interest using the filter() function from dplyr.
2. Aggregating the data at different levels (e.g., yearly, quarterly) to identify broader trends and patterns using functions like group\_by() and summarize() from dplyr.
3. Joining additional datasets, such as demographic or crime data, to explore the impact of neighborhood characteristics on housing prices using left\_join() from dplyr.
4. Calculating summary statistics (e.g., mean, median, growth rates) at different categorical levels to compare and contrast market segments using functions like summarize () and across() from dplyr.

**How could you summarize your data to answer key questions?**

To answer the key research questions, the data can be summarized using various techniques:

1. Descriptive statistics: Calculate measures like mean, median, standard deviation, and range for key variables to understand their central tendencies and variability.
2. Correlation analysis: Compute correlation coefficients between housing prices and various predictors to identify the strength and direction of their relationships.
3. Regression analysis: Fit regression models to quantify the impact of different factors on housing prices and assess their statistical significance.
4. Time series decomposition: Break down the housing price time series into trend, seasonal, and residual components to better understand the underlying patterns.
5. Geographic visualization: Create maps and spatial plots to visualize the geographic distribution and variations in housing prices and related factors.

These summaries can provide insights into the key drivers of housing prices, the relationships between variables, and the overall trends and patterns in the data.

**What types of plots and tables will help you to illustrate the findings to your questions?**

Various plots and tables can be used to effectively illustrate the findings and answer the research questions:

1. Scatter plots: Visualize the relationship between housing prices and continuous variables like square footage or income.
2. Line plots: Show the trend of housing prices over time, both overall and by region or city.
3. Bar plots: Compare average housing prices, sales volumes, or other metrics across different categories like regions or property types.

Do you plan on incorporating any machine learning techniques to answer your research questions? Explain.

Yes, I plan on using machine learning techniques that can be incorporated to answer the research questions and gain deeper insights:

1. Regularized regression methods can be used to handle multicollinearity and identify the most important predictors of housing prices.
2. Clustering algorithms can be applied to identify distinct housing market segments based on property characteristics or neighborhood attributes.
3. Time series forecasting models can be utilized to predict future housing prices based on historical trends and seasonality.

**What questions do you have now that will lead to further analysis or additional steps?**

After the initial analysis, several questions may arise that can guide further exploration and additional steps:

1. Are there any significant differences in the factors influencing housing prices across different cities or regions? How can these differences be further investigated and explained?
2. What are the long-term implications of the identified trends and patterns in housing prices? How can this information be used for forecasting and decision-making?
3. Are there any specific neighborhoods or zip codes that exhibit unique housing market dynamics? How can these localities be further studied to understand the underlying drivers?
4. Can the developed models be improved by incorporating additional data sources or using more advanced machine learning techniques? What are the potential benefits and limitations of such enhancements?

These questions can lead to further analysis, refinement of the models, incorporation of new data sources, and targeted dissemination of the findings to relevant audiences.

**Introduction**

This section continues the exploration of factors influencing housing prices in major U.S. cities by using various datasets and analytical techniques. Understanding the key drivers of home values is essential for real estate investors, homebuyers, sellers, and policymakers. The analysis involves merging multiple datasets, performing exploratory data analysis, and building predictive models to gain insights into the housing market dynamics.

**The Problem Statement Addressed**

The primary research questions addressed in this analysis were:

1. What are the most essential features in predicting home values in different cities?
2. How does the influence of factors like square footage and the number of bedrooms vary across regions?
3. Can we build an accurate model to estimate home prices based on property characteristics?
4. How have the key drivers of home prices changed over time?
5. Are there any leading indicators that tend to precede major shifts in home values?
6. Can clustering homes by their features reveal distinct market segments?
7. How do socioeconomic and demographic factors of a neighborhood impact home prices?
8. What role do macroeconomic variables like interest rates and job growth play?

**How I Addressed This Problem Statement**

**Dataset Sources:**

* Zillow Research Housing Data
* American Community Survey (ACS) Census Data
* Federal Reserve Economic Data (FRED)
* Redfin Housing Market Data

**Data Preparation:**

* Data cleaning involved ensuring date consistency, converting numeric fields, and merging datasets based on relevant features.
* Padding techniques were used to handle datasets with varying lengths.
* Missing numeric data were imputed with median values.

**Data Cleaning Changes:**

Initially, the data cleaning focused on ensuring date consistency and merging datasets based on date columns. Due to the diverse nature of the data sources and their differing granularities, the approach was modified to combine datasets based on relevant features rather than dates. Padding techniques were used to ensure that datasets with varying lengths could be combined without losing information.

**Exploratory Data Analysis (EDA):**

* Summary statistics and data structures were examined.
* Correlation matrices were plotted to understand relationships between variables.
* Scatter plots were used to visualize relationships between median sale prices and income levels.

**Predictive Modeling:**

* A linear regression model was built using predictors such as median and mean income, unemployment rate, economic indicators, and Zillow Home Value Index.
* The model's performance was evaluated using RMSE (Root Mean Squared Error).

**Clustering Analysis:**

* K-means clustering was applied to identify distinct segments in the housing market based on the available features.

**Analysis**

**Correlation Matrix:** The correlation matrix heatmap displayed the relationships between variables, with positive correlations shown in blue and negative correlations in red. This helped identify the strength and direction of relationships between housing prices and various predictors.

**Scatter Plots:** Scatter plots of median income and mean income against median sale prices showed the general trends and relationships between these variables. The linear regression lines provided a visual representation of these trends.

**Bar Plots:** Bar plots were used to compare median sale prices, median income, and homes sold across different regions. These plots highlighted the regional differences in housing market dynamics.

**Time Series Plots:** Time series plots of homes sold and median income over time revealed trends and seasonal patterns in the housing market. These plots helped visualize changes in the market over different periods.

**Clustering Analysis:** The clustering analysis revealed distinct segments in the housing market based on median income and sale prices. The clusters provided insights into different groupings of housing markets with similar characteristics.

**Implications**

The analysis provides valuable insights for various stakeholders:

* **Real Estate Investors:** Can identify high-value markets and make informed investment decisions based on income levels and economic indicators.
* **Homebuyers and Sellers:** Can understand the key drivers of home values and make informed buying or selling decisions.
* **Policymakers:** Can develop strategies to stabilize housing prices and address regional disparities in the housing market.

**Limitations**

The analysis has several limitations:

* **Data Granularity:** City-level data may not capture neighborhood-level variations.
* **Temporal Coverage:** The limited time period may not reflect long-term trends.
* **Model Complexity:** The linear regression model may not capture complex nonlinear relationships.
* **External Factors:** The analysis did not account for external factors such as policy changes or natural disasters.

**Concluding Remarks**

This analysis provides a comprehensive understanding of the factors influencing housing prices in major U.S. cities. By leveraging various datasets and employing multiple analytical techniques, the study identified key predictors of housing prices and provided actionable insights for stakeholders. Future work could include incorporating additional data sources, applying more advanced machine learning techniques, and conducting a more granular analysis at the neighborhood level.