Mapping U.S. Housing Trends with Big Data

Authors: Christian Ahamba, Solange Ruiz, Damian Trent Wilson
Department of Information Systems, California State University Los Angeles
CIS 5200 System Analysis and Design
cahamba@calstatela.edu sruiz85@calstatela.edu dwilso33@calstatela.edu

Abstract: The U.S housing market plays a vital role in the national economy, and its analysis is critical for various stakeholders. This project investigates spatial and temporal trends in housing prices using Redfin market data, spanning over 10 years. We employed big data tools including Hive and Hadoop for querying and aggregation, and Tableau for visualization. Our work identifies state-level price volatility, market shifts over time, and insights for investors, homebuyers, and policy makers. Findings indicate a clear regional disparity in housing market volatility.

1. Introduction

The U.S housing market is a key economic indicator. Understanding its trends supports decision-making for investors, buyers, and policymakers. Our project analyzes temporal and spatial variability in median listing prices across the United States. We used Hive on Hadoop to process over 8.5GB of Redfin housing data, and Tableau for visualization. The goal was to uncover patterns of volatility and regional disparities.

2. Related Work

After overviewing the first study, "**A Dynamic Spatiotemporal and Network ARCH Model with Common Factors**" **(Hou, 2024).** This study introduces a dynamic spatiotemporal volatility model that extends traditional approaches by incorporating spatial, temporal, and spatiotemporal spillover effects, along with volatility-specific observed and latent factors. The model offers more general network interpretation, making it applicable for studying various types of network flow. The use of Bayesian estimation via the Markov Chain Monte Carlo (MCMC) method, the model provides a strong framework for analyzing the spatial, temporal, and spatiotemporal effects of a log-squared outcome variable on its volatility. The model's flexibility is demonstrated through applications to the U.S. housing market and financial stock market networks, highlighting its ability, capturing vary degrees of connection.

The second analyzation, "**Exploring the Spatial Segmentation of Housing Markets from Online Listings**" **(Abella, 2024)**. Abella and colleagues present a methodology based on multipartite networks to detect spatial segmentation in housing markets using data from online listings. By constructing a bipartite network connecting real estate agencies and spatial units, and projecting it into a network of spatial units, the study applies clustering methods to segment markets into spatially coherent regions. The methodology is robust across different clustering algorithms and spatial scales, with case studies in France and Spain, offering insights into market segmentation relevant for urban studies and policymaking.

The third study, "**Spatial Analysis and Modeling of the Housing Value Changes in the U.S. during the COVID-19 Pandemic" (Zhang, 2024).** This research utilizes Geographically Weighted Regression (GWR) to model and analyze the spatial relationships between housing price change rates and various factors during the COVID-19 pandemic. By incorporating data from Zillow Home Index Value (ZHVI), COVID-19 case numbers, and socioeconomic variables, the study examines how different regions in the U.S. experienced housing value changes. The analysis provides insights into the spatial heterogeneity of housing market dynamics during a significant global event.

The analyzation of similar studies offered valuable insights into housing market dynamics through various statistical and econometric models; our project distinguishes itself by several key aspects. As some of those aspects are the use of big data, integration of temporal and spatial visualizations, price volatility, stakeholder insights and scalable and adaptable frameworks.

The use of a Hadoop + Hive setup on a 5-node cluster enables the efficient processing of a substantial 7GB+ dataset, facilitating real-time analysis and scalability that traditional models may lack. By combining time-based and state-level visualizations in Tableau, our project provides an intuitive and interactive platform for stakeholders to explore complex data patterns. Unlike other studies that may focus on average trends or market segmentation, our project places a central emphasis on price volatility—a critical yet often overlooked market indicators offering a contrasting perspective on its implications across different states and time periods.

Our project focuses on delivering practical insights tailored to investors, homebuyers, and policymakers and ensures that the analysis is not only descriptive but also prescriptive, facilitating informed decision-making. The use of HIVE and Tableau allows for scalability and adaptability, making it feasible to extend the analysis to more granular

levels, such as zip-code or neighborhood, as computational resources permit. In summary, our HIVE approach offers a comprehensive and user-centric framework that enhances the understanding of housing market dynamics beyond traditional methods.

3. Methodology

This section outlines the technical and analytical framework employed in our study. It includes an overview of the dataset used, the computational environment for data processing, and the detailed workflow for preparing, analyzing, and visualizing the data. By leveraging a big data stack built on Hadoop and Hive, paired with visualization tools like Tableau, we were able to manage a multi-gigabyte dataset efficiently and extract meaningful insights about housing market trends across the U.S.

3.1 Dataset Overview

Source: Kaggle (Redfin Housing Market Data)
Format: TSV
Size: ~8.5GB
Time Span: 10 years

File Name	Size
county_market_tracker.csv	363 MB
Neighborhood_market_tracker.csv	4.7 GB
State_market_tracker.csv	20 MB
Us_national_market_tracker.csv	983MB
Zip_code_market_tracker.csv	2.8MB

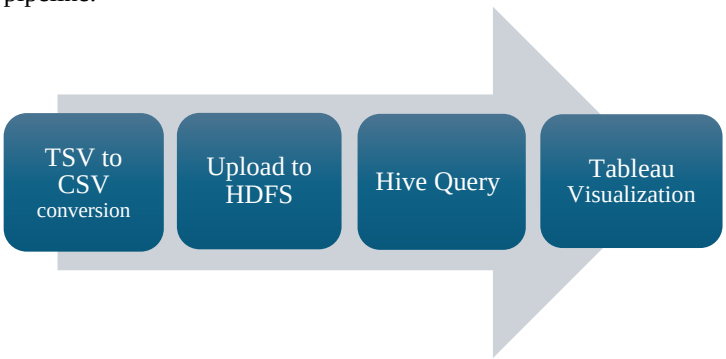
3.2 Cluster Set Up

This section describes the technical infrastructure used to process and analyze the large scale of housing dataset. The project ran on a distributed Hadoop cluster consisting of five nodes, two configured as master nodes to manage the system, and three as worker nodes to perform the data processing tasks. Each node was equipped with 32 GB of RAM and a 2.3 GHz CPU, ensuring sufficient memory and processing power for handling the 8.5 GB dataset. To interact with the dataset stored in the **Hadoop Distributed File System (HDFS)**, the team used Hive, a data warehousing tool that enables SQL-like querying over large datasets. Queries were executed through Beeline Hive via a secure SSH connection to the cluster at **IP address 144.24.13.0.**

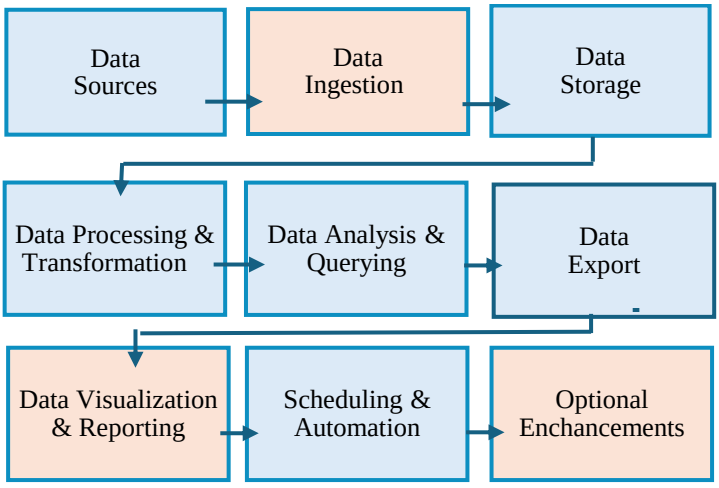
Component	Specification
Hadoop Cluster	5 nodes (2 masters, 3 workers)
Memory	32 GB RAM per node
CPU	2.3 GHz CPU
Query Interface	Hive via Beeline
SSH Access	144.24.13.0

3.3 Data Processing Workflow

The explanation for the **Data Processing Workflow** section is now expanded with clear descriptions for each step in the pipeline.



3.4 Implementation workflow



4. Results and Visualizations

4.1 National Price Trend (Line Chart)

We created a Hive table called sruiz85_redfin_state. Initially, the state data appeared incorrect due to the use of FIPS codes which is a standardized numeric code assigning to the U.S states. Since Tableau requires full state names for mapping, we updated the schema to use proper state. We then calculated the national average of median listing prices over time, grouping the data by year and export to Tableau. The

results show a steady upward trend, with a notable spike in 2020-2021 across most states Figure 1.

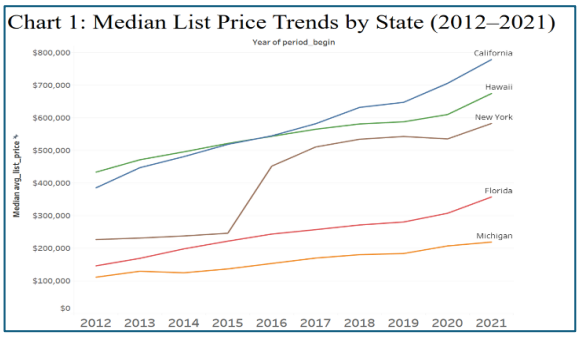


Figure 1. *Statewide Growth Peaks in 2020-2021*

4.2 State Price Map (U.S Map)

This map (Figure 2) shows the typical home prices in each state in 2021. We used color to show which states are more expensive and which are more affordable. Darker colors reflect high prices. States like California and Hawaii are some of the most expensive, while states like Michigan and Mississippi are more affordable, this reflects previous line chart. Clear coastal vs. inland price disparity was observed.

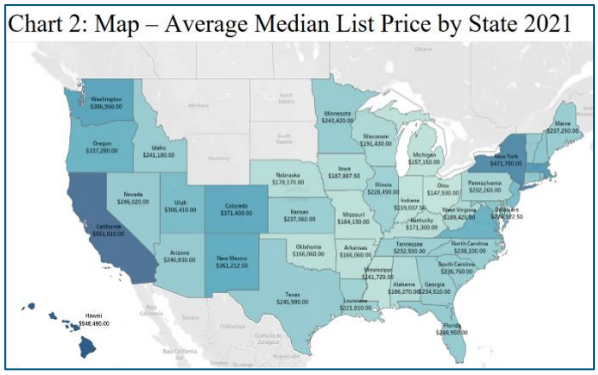


Figure 2 *Coastal vs. Inland price disparity*

4.3 Price Volatile States (Rank Bar Chart)

Using the same data pipeline as previous analyses, we modified the Hive query to compute the standard deviation of listing price per state year combination. This measure captures how much prices fluctuated annually. The resulting data was visualized in Tableau as a heatmap. Where darker red indicates higher volatility and blue indicates stability. Figure 3 highlights that price volatility was most pronounced in Western and Southern state between 2012 and 2021, reflecting regional differences in market stability.

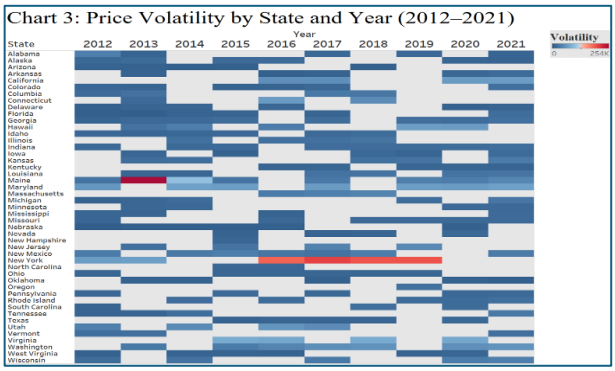


Figure 3. *Annual State Price Volatility*

4.4 Top 10 Volatile States (Bar Chart)

Using Hive, we calculated percentages growth in median listing prices for each state between 2012 and 2021. To do this, we used aggregation with conditional logic specifically MAX () combined with CASE WHEN to isolate prices for the start and end years. The resulting values were used to compute percent growth, which reflects how much prices increased over the 10-year period. The final bar chart below figure 4 visualizes the top 10 states with the highest growth. Idaho, Nevada, and New York left the list with over 175% price increases, meaning home prices in these states nearly tripled. This figure helps highlight which housing markets experienced the most dramatic increases and may be considered the most volatile.

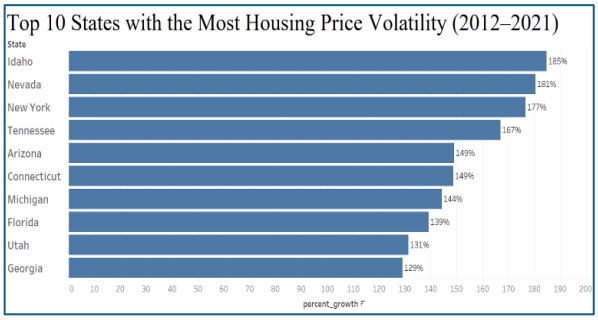


Figure 4. *Top 10 States by Price Growth (2012-2021)*

5. Key Findings

Regions with most stable vs. Most volatile prices:
Most Volatile States: Idaho, Nevada, New York, Tennessee, Arizona.
Volatility: most prominent in Western and Southern states.
Most Stable States: Ohio, Indiana, West Virginia, Iowa, Wisconsin.

Years/Periods with major market shifts:
2013: Sharp volatility spike in Maine (anomaly).

2017: New York shows major price swings.
2020-2021: Nationwide volatility spike and supply-demand imbalance.

Insights for Investors, Buyers and Policy Makers:
Investors: Volatile states may offer higher returns but come with risks.
Buyers: Stable states provide more predictable long-term value, ideal for personal residence.
Policy Makers: Focus on affordability, supply, and zoning in volatile regions.

6. Challenges and Limitations

There were technical and data related challenges that need to be fixed. First Tableau was unable to interpret FIPS codes correctly, for that reason we adjusted the schema to use full state names. Second, missing data for specific states and years reduced the completeness of some visualizations. Hive's lack of support for certain SQL features, such as WITH clause in INSERT statements, forced us to restructure some queries. Finally, visual outputs in Tableau required additional manual formatting to handle null values, percentages, and heatmap readability.

7. Conclusion

Big data tools provide scalable methods to analyze complex housing datasets. Our analysis highlights meaningful trends in pricing and volatility, offering value to stakeholders. Future work may explore machine learning forecasting, as well as extend this foundation through time-series forecasting, enriched feature engineering, and more granular geographic segmentation.

References

[1] Abella, D., Martínez, J. H., et al. (2024). Exploring the spatial segmentation of housing markets from online listings. *arXiv*. https://arxiv.org/abs/2405.08398

[2] Hou, K., Huang H., Liu, Q., & Zhou, T. (2024). Housing Market Forecasting: Large Language Models Beat Traditional Econometric Methods. *arXiv*. https://arxiv.org/abs/2410.16526

[3]J. Woo, *CIS 5200 – Big Data Analytics: Lab materials and course lectures*, unpublished instructional materials, California State University, Los Angeles, 2025.

[4] S. Ruiz, C. Ahamba, and D. Wilson, *Redfin Housing Market Forecasting with Apache Spark* [GitHub Repository]. [Online]. Available: https://github.com/Soligrl/CIS-5200

[5] Thuynyle, *Redfin housing market data* [Data set], Kaggle. [Online]. Available: https://www.kaggle.com/datasets/thuynyle/redfin-housing-market-data

[6] Zhang, X., Liu, Y., Feng, J., & Wang, Y. (2024). Forecasting Housing Prices with Machine Learning: A Survey. *arXiv*. https://arxiv.org/abs/2405.08398