Predicting the Heat of U.S. Housing Markets

Project Final Report

Peter Scott
Department Name
CU Boulder, College of Engineering
Boulder, CO, USA
pesc3485@colorado.edu

INTRODUCTION

The purpose of this project is to evaluate U.S Housing market database to conduct a correlation analysis and determine if a 'hot market' can be determined based off several attributes such as pending sales, price drops, supply, and other factors. A large portion of the initial steps will be identifying the economic markers that a hot housing market has, so as to better conduct the study. Additionally, the project identifies if certain regions are hotter markets regardless of their supply and demand status, examining how large of a role geographic feature will play into the housing market of a specific area, and whether they are resistant to fluctuations in those economic factors previously listed.

ABSTRACT

I was able to achieve k-means clustering for classification of whether a market is hot, cold, or is a super-hot outlier market. This allowed me to do a heatmap representation of each one of those clusters, showcasing the attributes that go into making a 'hotness index' for this project. After representing that data within a heat map, I displayed a choropleth graph of the US and after aggregating the data and grouping the data by state, I was able to get the mean hotness index for each state and have it displayed on a dashboard. This dashboard meets my goal of having a tool that produces a clear picture for someone unfamiliar with the housing market, because it gives the user a green-good, yellow-neutral, red-worse color guide for quick and clear insight on each specific state's market health.

After the choropleth figure was up and operational, it clearly painted a picture that started to answer my secondary question of whether geographic features play a role in market demand status as I noticed patterns of almost every state touching an ocean was a hotter market. I took this, recompiled my data-list and further broke down the problem by aggregating the data further in 5 regions of the United States and getting the hotness index for each of them and comparing them via bar graph.

I conducted this level of data analysis because it would allow further pattern analysis on geographic location and what impacts they had on the market status. What I ended up concluding was there were far more variables that added complexity to this question, such as climate, accessibility, job market health, cost of living, to name a few. This question ultimately was unable to be questioned from this database simply based off geographic location.

I concluded that this project produced a successful algorithm that could factor in multiple parameters from a housing market and produce a rough index on whether that market is a seller's market, or if it were a "cooler" buyer's market, grouped by city, state, and property type for those targeted searches by users.

KEYWORDS

Housing market, U.S. Housing Data, clustering, predictive modeling, Redfin, Real estate analytics, home buying, predictive modeling.

RELATED WORK

Much of the research that I utilized to validate and support my project focused on machine learning and AI efforts to predict the valuation of properties, specifically forecasting what the houses are worth as a sale price. Although this serves a different purpose than my project, there are still valuable aspects of the study done in 2025 using regression models and ML analytics to predict the housing market – "Assessing AI techniques for Precision in Property Valuation". This specific study clarified that many regressive and ML analytics of the market are often more accurate in predictive metrics than previous hedonic methods from before and the market is filled with tools utilizing these technologies.

Although my project isn't for the valuation of each properties pricing, it is still using algorithms and predictive analysis based off current housing market conditions to project the current state of a market rather than an older method of predictive analysis that is strictly based on historical trends. These algorithms could be turned into a random forest ML model that could predict the market heat indicators that I specified in the project and then extract which important features there are. This could also further grow into a better predictive tool using these ML neural networks so that it could predict a month or two ahead of time to really give a user an incredible tool.

The second main related work that I referenced during this project was "Data-Driven Analysis of the U.S. Housing Market Using Clustering Techniques" – a study conducted in 2019 using k-means clustering to identify distinct housing market types. This was the main origin of my influence to use k-means clustering for distinct market conditions after establishing my hotness index and clustering them for further analysis and data visualization.

My project builds on these related studies by constructing my composite index for housing market heat, which was based on several market-health factors that were then validated against currently well-known hot markets to determine the accuracy and consistency of this composite score. K-means clustering was then conducted to classify these scores into 3 different categories for further evaluation and data visualization which I evaluated using silhouette and Davies-Bouldin indices. I then packaged all of these scores and clusters up and threw them onto various data visualization figures which allowed for further pattern recognition to answer the second question in determining geographic impact on resistance to market-health conditions and whether or not they stay as "hot-markets". It is a combination of data-driven clustering analysis and gives practical decision support for home buyers in real life applications.

DATASET

Source:
https://www.kaggle.com/datasets/vincentvaseghi/us-cities-housing-market-data/data

Size: 5.9 million data objects, 58 attributes

Description: Composed of a combination of integers, strings, binary and Boolean values, this information is largely numerical and historical evaluation of many different aspects of an areas home market in a specific geographic location like median pricing for list vs. sale, price per square foot, numbers of inventory sold, total inventory, pending sales vs. new listings, price drops, sales above listing price and many other markers that would be sufficient for my data project.

Form of documentation is downloaded to local machine through .tsv file and transformed and compressed into a parquet file for efficient querying and modifications.

Rather than using PostgreSQL I decided to keep the parquet file local due to the scale of the analysis. If I determine to create an application, I will create a live feeding data pipeline into a PostgreSQL database for use by the application.

Quality of the data was quite sufficient for my project analysis, there were same null values in smaller cities or outlier markets, however considering the volume of data included, it was more than quality to produce a highly reproducible and accurate algorithm. No ethical considerations on this data, as it is public data that is aggregated into a composite index for easier user viewing and accessibility to those who aren't as educated on the intricacies of the real estate market.

This is a very justified dataset for the research question because it gave enough attributes for each entry that allowed for a highly in-depth review of all of the factors that are typically used to review a market's health and determining whether it is a seller or buyer's market.

TECHNIQUES APPLIED

The integration of the Redfin dataset was initially done through downloading of a .tsv file, parsing through and querying it utilizing pandas and sqlite. I then found that parquets would be a cleaner and more effective method for this while still utilizing pandas and duckdb to query the database and modify it as needed. I decided to keep all of these within a queryspace.ipynb file because I could rerun queries as needed and it segregated each portion of the code in a very clean fashion.

The initial integration was completed with pandas read_csv command that allowed me to specify to import the data, separated by \t tab, and to look for any na_values and to automatically clean those from the database. I was happy with that imported parquet file and continued on to some simple cleaning, stripping any spaces with underscores and ensuring all column headers were lowercase for consistency, then fixing all date-time formatting issues for the same reason. Upon completion of the cleaning, I compressed it into a parquet file and moved on to the querying stage where I could start to build the hotness index.

This was a pretty complex SQL query, but essentially it takes the price year over year, supply, demand, sale to list time, and inventory year over year, all other attributes were removed for simplifying the model. I then used z-score normalization to standardize each of these housing attributes, which I used a z-score for each housing indicator. I then computed the hotness index which I adjusted based on sensitivity testing during the initial development stages, economic factors that I knew like price and supply/demand having a rather heavy influence on how quick houses sell. This came out to having 0 being a neutral place with the index, positive being warmer, and negatives being cooler markets. I was happy with the scores when I didn't have too many dramatic outliers.

After establishing the composite heat score, I grouped by city, state, and property type, aggregated the mean of those same housing attributes and plugged that into sklearn module to run k-means algorithm and determine how many clusters would be most appropriate for this project – which ultimately ended up being 3 clusters with a silhouette of 0.342 (highest) and Davies-Bouldin= 1.065 (2nd lowest). I reran the k-means with 3 clusters, plugged that into the database assigning each city a cluster and cluster label.

After clustering and binning, I had a new table that gave me each property type for each city and the associated cluster. To validate this, I took roughly 7 cities in the US that are historically hot markets and see where those fell with my 3 clusters, and with the exception of only a couple outliers, they all fell in the "warm" side of the market as I expected them to.

After clustering and binning was completed and validated, I moved on to visualization starting with a heatmap showing each cluster and the associated z-scores of each attribute in the composite score. I

recalculated the clusters but by state for plugging into the choropleth map showing some patterns in hotter markets and their locations, prompting me to compare the scores based on the 4 main regions of the US – West, Southwest, Northeast, Midwest, and Southeast and displaying them as a bar graph.

I felt like this was an appropriate progression in the data mining process, and the analysis that went into this built on itself by further refining the calculated scores and grouping them in different ways to help answer my research questions.

KEY RESULTS

The biggest result of this project is that there is a valid and reproducible algorithm for showing the US market status for specific cities, states, or regions, that can be used to quickly create a composite index and allow a user to see if that market is hot or cold, hot being a seller's market, and cold being a buyer's market. It is a simple application of mathematics and basic economics, however not all home-buyers know where to access this in-depth market information or how to utilize it.

The second aspect of the project results showed that there indeed were patterns in where there were hot markets that could be validated with looking up real estate demand and pricing – which was states that touched the ocean had a much hotter index than those in the middle of the country. This wasn't enough to draw any conclusions about geographic location and market resiliency; however, it definitely had me key in certain areas for further examination. I then tried to determine the market conditions by region which mainly showed that the further northeast and west, specifically pacific northwest had hotter markets. This led me to the conclusion that geographic location was just a small piece of the equation and other factors like job market, cost of living, accessibility, climate and many other factors influenced the housing market rather than the geographic location alone, ultimately determining this database wasn't sufficient enough for that conclusion.

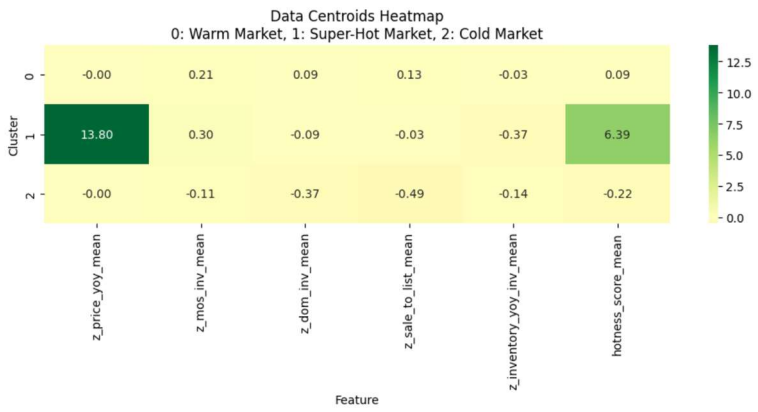
APPLICATIONS

I set out to create an algorithm that could be implemented into a user dashboard that gives someone the opportunity to quickly and accurately gain insights into certain areas of the country they are interested in buying a house. This could be accomplished in list style databases or heatmaps or even choropleths for quick references for a more satisfying visual experience.

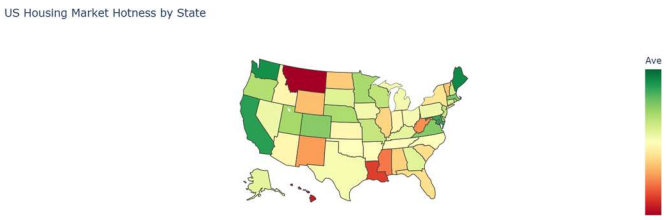
The next step to make this a more complete use experience would be to establish the web application that pulls real time information via the housing market API for updated monthly information and then have a dashboard with various visualizations for the user to define what specific property types they are looking for, or perhaps regions, cities, or states for quick comparisons to narrow their research. This could be paired with other databases and further cost breakdowns for areas they specify to allow the user to get as specific as they need for their uses. d

VISUALIZATION OF RESULTS

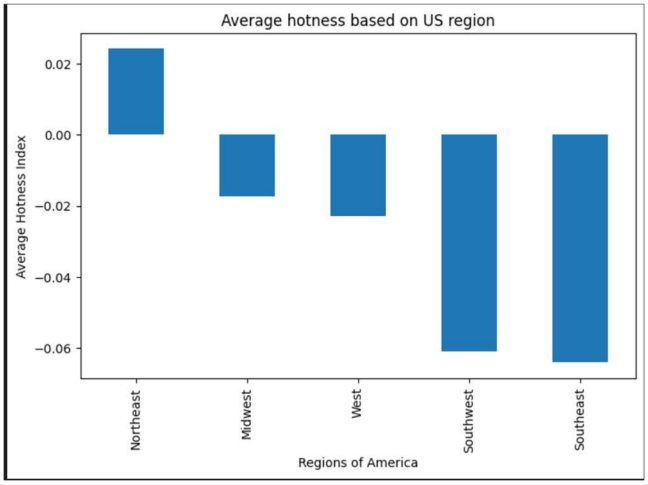
Below is the heatmap that is broken into 3 clusters (temperatures of the market) with the associated attributes for each cluster and their score. 0 is the set neutral zone, positive (green) representing warmer markets, and then cluster 1 is the super-hot outlier cluster.



Choropleth of the US state average market temps by cluster. With a color key on the side, this creates a really nice visual for users to determine which states have a hotter real estate market based on how deep green the states are shown.



Finally, is the bar graph that was used to compare the different regions of the US. This again uses the same hotness index matrix, 0 being neutral and shows the aggregate of each region to determine if that region as a whole is warmer or cooler.



EVALUATION OF METHODS

Metrics – Clustering will be done using Davies-Bouldin Index or silhouette scoring for checking the accuracy of my clustering. For evaluation of performance, I will check prediction accuracy and then utilizing my validation measures of known hotspots to evaluate the predicted outcomes.

Validation – Data sample will be compared to a known hotspot like Miami or Asheville to confirm accuracy of metrics and see if predictions are accurate for these areas. The other validation that will be done is Temporal Validation to see if the predicted hotspots last over time or if there is a pattern to the duration of their popularity.

Visualization – I will validate spatial consistency of my data through heatmaps or choropleth map for some of the models as those can be most effective when talking about US economics as a whole and can perhaps provide a nice tool to link in with an application as to have a real-time platform of the hotspots that are happening throughout the United States.

Update: I have defined and computed a “hotness score” metric that will be used for comparison and charting. For validating this score, I will be using temporal consistency and known hot markets to ensure my formula is producing appropriate numbers and scaling.

My next step in the evaluation process is to add visual and cluster performance metrics such as a silhouette score.

TOOLS

Languages – Python: including modules like NumPy, Pandas.

Database – Parquet Files local. I determined that I will utilize PostgreSQL database only if I decide to make a dashboard application that needs a live feed data and be accessed by many different users.

Visualization – Matplotlib, Plotly, Seaborn.

Additional Applications – VSCode coding IDE, Jupyter notebooks for analysis

REFERENCES

[1] Vaseghi, V. (n.d.). *US Cities Housing Market Data* [Data set]. Kaggle. Retrieved October 12, 2025, from https://www.kaggle.com/datasets/vincentvaseghi/us-cities-housing-market-data/data

[2] Ali, W., Samarasinghe, D. A. S., Feng, Z., & Rotimi, J. O. B. (2025). *Assessing AI techniques for precision in property valuation: A systematic review of the four valuation methods*. In Proceedings of the 23rd CIB World Building Congress (Vol. 1, Article 322). Purdue University. https://doi.org/10.7771/3067-4883.1790

[3] Chen, L., Zhang, Y., & Patel, R. (2019, December). *Data-driven analysis of the U.S. housing market using clustering techniques.* In 2019 IEEE International Conference on Big Data (Big Data) (pp. 4567–4574). IEEE. https://doi.org/10.1109/BigData47090.2019.9006132