What Makes The Housing Market “Hot”?

Emily Phillips

Department of Data Science, Bellevue University

DSC 680: Applied Data Science

Dr. Catherine Williams

April 28,2022

Background

The real estate market has had its ups and downs throughout history, but in recent years especially since the slowdown of COVID-19, it looks to be booming and as some sources put it, “hot”! Americans have begun to put more time and money into home-buying and selling, which is raising median housing prices and putting some area markets at a greater demand than others. According to an article titled “Hot housing markets: Where are they - and why are they so hot?” written by Aimee Picchi at CBS News, “At the end of 2020, home prices were about 15% higher than a year earlier, before the pandemic shuttered the U.S. economy, according to NAR data” (Picchi, 2021). Picchi further went on to note that “in about 88% of metropolitan regions, home prices experienced double-digit price growth” (Picchi, 2021). Housing demand increased immensely and the hotness of housing markets could be predominantly tied back to their areas’ affordable housing stock and the liveability of the metropolitan areas. With more people shifting to a work-from-home lifestyle, they started looking for bigger homes at lower prices in warmer and exciting areas, thereby also increasing the hotness of those areas’ housing markets. This idea of housing market hotness is an intriguing one, and it will be thoroughly explored throughout this project.

Business Problem

Understanding and recognizing the popularity of a certain housing market puts one at the advantage of buying (or selling) valuable property. Throughout this project, the main focus was on gathering evidence around the following business questions:

1. Which regions of the United States have the hottest housing markets?
2. Is there a cutoff for the market hotness index where a significant difference between markets becomes apparent?
3. Does the allure of the certain cities (favorable place to live) play into the cities’ assigned market hotness index?
4. How has time (months, years) impacted the hotness of metropolitan areas?
5. Can we predict the market hotness index of cities?

Data Explanation

Multiple data sources were utilized and combined for this project. Two of the datasets originated from Realtor.com: one pertains to monthly data for metro market trends and statistics on active for-sale listings, and the other dataset contains information on the assigned Market Hotness Index for given metro areas which is based on days on market (supply index) and Realtor.com views per property (demand index). Other variables represented in these datasets included median listing prices, median number of days on market, the count of new listings added to the market, etc. and the corresponding variables' percentage change from the previous month and the previous year.

The other dataset was retrieved from a Wikipedia page titled "Metropolitan statistical area", which contained a table containing information on the three-hundred and eighty-four metropolitan statistical areas of the United States. According to this page, "a metropolitan statistical area (MSA) is a geographical region with a relatively high population density at its core and close economic ties throughout the area" (Wikimedia Foundation, 2022). Since the Realtor.com datasets were metropolitan-area focused, the data from Wikipedia was useful for gathering population measures for the represented corresponding areas. The various measures included: the population as of July 1, 2021 as estimated by the United States Census Bureau; the population as of April 1, 2020 as enumerated by the United States Census Bureau, and the percent change in population from April 1, 2020 to July 1, 2021. A part of the focus with the business questions was on how the livability of a certain metropolitan area makes it more suitable as a housing market, and population data was utilized to most closely represent this concept.

The target variable for modeling was Hotness Rank, which is "the specified zip code, county, or metro area's Hotness rank, by Hotness score, compared to all other zip codes, counties and metro areas. A rank value of 1 is considered the hottest" (Conner et al., n.d.).

Methods

Many activities were performed as a part of cleansing and exploring the data. To begin with, all datasets were checked for duplicates, as of which none were found in any of the three sources.

For data cleansing, a large focus was on ensuring that the variables corresponded to their correct data types. Numeric conversion was completed on any numerical measures, and if there was any data in the columns that was not numeric, such as "#NAME?" being found in the Pending Ratio variables, these were replaced with "NaN" to represent missing data. The Realtor.com datasets also contained an attribute that depicted the combined month and year of a given metropolitan area's housing market; this attribute was split into separate year and month columns, in order to gain more value from the dates and utilize them for time-series visualizations.

Once the datasets were properly cleansed, they were merged together to form a cohesive resource on metropolitan areas' housing markets and populations. The dataset was subsetted to focus on the years of 2020 and 2021, in order to tie in with the population measures from the United States Census Bureau. The Realtor.com datasets also contained a column named 'Quality Flag', which "triggered ("1") when data values are outside of their typical range" (Conner et al., n.d.). In essence, this attribute identified outliers within the data, and since outliers are potential dangers to data integrity, the dataset was grouped on the zero values that represented typical data records.

The final merged dataset contained 4,472 rows and 46 columns, a quite smaller subset from the size of the original housing market datasets. It contained missing data in five of the forty-six columns, with the most missing values being in 532 rows for the 'Price_Increased_Count_Mm' field. The options for filling missing data were weighed; however, with any of the options such as using overall median or mode, it increased the chances of manipulating the data against the concept of metropolitan area independence. Each housing market is unique to location, and by using an overall measure, it meant bringing in information that was not representative of that specific geographical area. Therefore, the missing data was completely removed from the dataset instead. This completed collection was then utilized for data visualization techniques in order to gain more insights into the business questions, which will be explored in the analysis section.

Feature reduction and data scaling was performed prior to modeling. The model was reduced to only include the five most important features in relation to the target variable (hotness rank), with the attributes' Gini Importance values being cut off at 0.008. For model selection, several models which are appropriate for a classification model were used to evaluate performance accuracy. The models evaluated included the following: K-Nearest Neighbor, Naive Bayes, Support Vector Machine and Multinomial Logistic Regression.

Analysis

The main focus of analysis for this project was on yielding insights from various forms of data visualization. The form of visualization that brought the greatest results for understanding was a geospatial chart, which allowed for the depiction of a numerical measure over the map of the metropolitan areas in the dataset. These visualizations can be found in the appendix.

One focus from the business questions was on how the livability of a metropolitan area affects the hotness of its market. With population (census measures) being used as the livability indicator, a geospatial chart was created to depict the median population size of the represented metropolitan areas. The maximum and minimum hotness-ranked areas were also labeled for comparison purposes; the depiction can be found in Appendix A. In 2020, the hottest markets were in Springfield, Ohio and Manchester-Nashua, New Hampshire and had median populations of 136,001 and 422,937 people respectively. For the lowest-ranked market in West Palm Beach, Florida, it had a population of above 6 million people, which is a stark difference from the hotter-ranked areas. The bar chart in Appendix A also gives the same view, but allows for better comparison among the metropolitan areas which fall into a median hotness ranking from 1-5 (top 5) and 296-300 (bottom 5). There is not a cohesive divide between the top five ranked and bottom five ranked housing markets in terms of population, but the top housing markets do start at population sizes much below the larger areas of West Palm Beach and New York-Newark-Jersey City, which have populations in the higher millions (6 mill. and 20 mill. respectively). This difference could signify that the heavily-populated metropolitan areas which may tend

to be more crowded are not going to have as popular of housing markets, since it may lead to a difference in comfortability of living. Cities versus suburbs for home buying and selling is also a popular debate.

Speaking of cities versus suburbs, the regional location of a given metropolitan area could have a significant impact on its hotness rank. In Appendix B, there is a geospatial chart which highlights the median hotness ranking distribution across metropolitan areas. There is a spread for both sides of the ranking scale along the middle of the United States, but it does look like the hotter markets tend to be in California or concentrated in the Midwest region, i.e. Illinois, Indiana, Iowa, Kansas, Michigan, Minnesota, etc. There is also some concentration along the East Coast, in states such as Pennsylvania, New Hampshire, Connecticut and Massachusetts. The location of the least popular markets also became apparent from the plot, with them being primarily in the South. Given California's Mediterranean-like climate and coastal location, there is not much surprise around this finding in terms of hotter-ranked markets, even though California is known for its extensive housing prices as shown in Appendix C's visualization. The demographic of the buyers and sellers there could obviously play a factor. For the Midwest popularity, it is a comfortable and pleasant place to live, and is often known for its hospitality. The median housing prices in these states also err on the lower side ($150k-$250k), which opens the door for people from different financial backgrounds.

Furthermore, in terms of the fourth business question on change over time, Appendix D shows a geospatial chart visualizing median yearly change in hotness rank by metropolitan area. This chart actually gives an opposite view than expected, with the hotter markets as mentioned above decreasing in ranking and the "colder" markets increasing in ranking and becoming more popular from one year to the next. With time comes immense change, and this visualization shows that we can never be too certain of a certain area's market's popularity over time, since there are many factors that come into play with the housing market and a location's livability/trendiness.

The predictive modeling portion of this project fell very short of expected results, with the four models' accuracies, precision scores and recall scores being very low and unacceptable for usage. The

focus of this project became more on analytics than modeling, and the hope is to explore advancements in modeling at a later time.

Assumptions, Limitations & Challenges

The predictive modeling portion of this project posed a great challenge, in terms of yielding even relatively acceptable accuracies, precisions, and recall scores. The target variable had three-hundred classes, and once the dataset was split into training and test sets, there was not enough representation per class to give the models proper training. This lack of representation was a challenge and limitation in being able to reap value from the predictive modeling portion of this project.

In order to bring in data regarding area livability, I utilized the Wikipedia page on metropolitan areas as mentioned above. However, this dataset was quite limited in scope in that it contained actual data for the year of 2020 and then estimated data for the year of 2021. With only two years of representation, the dataset had to be further subsetted to only focus on them, and then population was the only indicator of livability. I made the assumption that this would be enough information to create reasoning on market hotness in regards to the popularity of an area, but it would have been beneficial to find other datasets that focused more on other livability features.

Future Uses & Recommendations

Understanding and recognizing the market hotness of a given area opens doors to other opportunities within that region. For example, from an article titled “Real estate: Why Tampa suddenly has the hottest housing market in the U.S.” from Rachelle Akuffo at Yahoo! Finance, the Florida city was celebrated for outranking other markets “due to its number of potential buyers, scarcity of homes, home sales, and flourishing job market” (Akuffo, 2022). The hotness of a market is useful for home buyers, home sellers, companies and local businesses. More people moving and living also means more people potentially looking for jobs and increasing sales for nearby businesses, since people need to buy products in all simplicity. Market hotness is incredibly useful for many industries, real-estate obviously being one, but utilizing data science and predictive modeling for gaining insights into it can make long strides for the people of the United States.

With a more well-rounded dataset, this data can be used by real-estate agencies, potential home buyers, businesses, etc. to stay ahead of the housing market trends and identify where opportunities are best suited, either for bringing people into a hot housing market or even helping a struggling one.

Ethical Assessment

Unfortunately, the housing market in itself is an incredibly unpredictable entity. In an article titled "Home prices hit another record high in March" written by Anna Bahney on April 20, 2022, the chief economist from the National Association of Realtors (NAR) was quoted, stating "The housing market is starting to feel the impact of sharply rising mortgage rates and higher inflation taking a hit on purchasing power" (Bahney, 2022). The supply and the demand for houses is slowly starting to decrease, which is putting home buyers in a tricky spot for finding affordable houses. Therefore, with the excitement around hot housing markets, it needs to be ensured that the lifestyle of Americans is taken into account from an ethical standpoint. Hopefully, the information is not utilized to scam home buyers or home sellers, but it is a possibility that needs to be envisioned. Homes in popular areas will be promoted for many different reasons, but it needs to be ensured that they are promoted in a manner that does not take advantage of anybody, and that all information is presented truthfully and clearly.

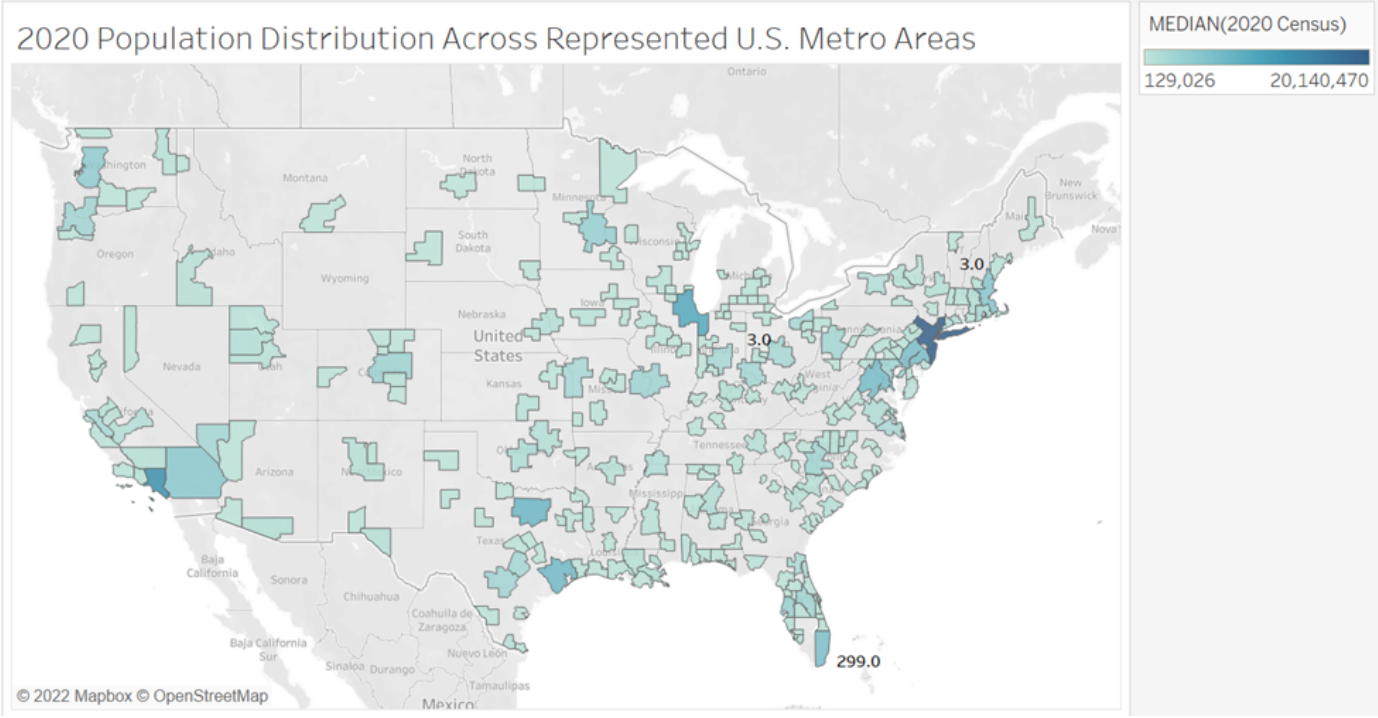
Conclusion

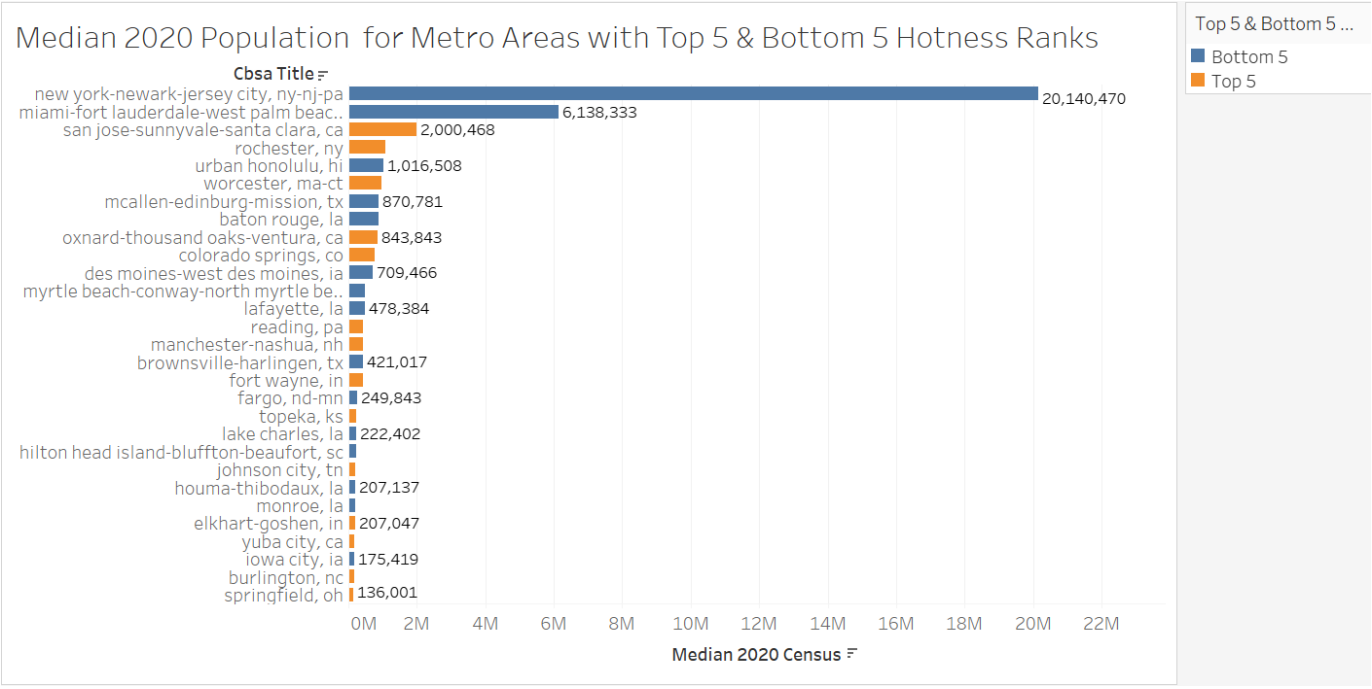
Given the high variability of the housing market, it is incredibly advantageous to have a reliable eye on it, and especially to know where it is thriving and increasing in hotness. Through the usage of data analytics, many individuals can stay informed on the factors which play into market hotness such as location, population and time as explored above. By bringing everything under one sphere of investigation, it makes it easier to consolidate patterns and trends and truly identify where people can get the best value for their homes. The choice of a home is one that should not be made lightly, and the hotness ranking of a given area can be of much importance in ensuring that people get the best possible living situations for their needs and wants.

References

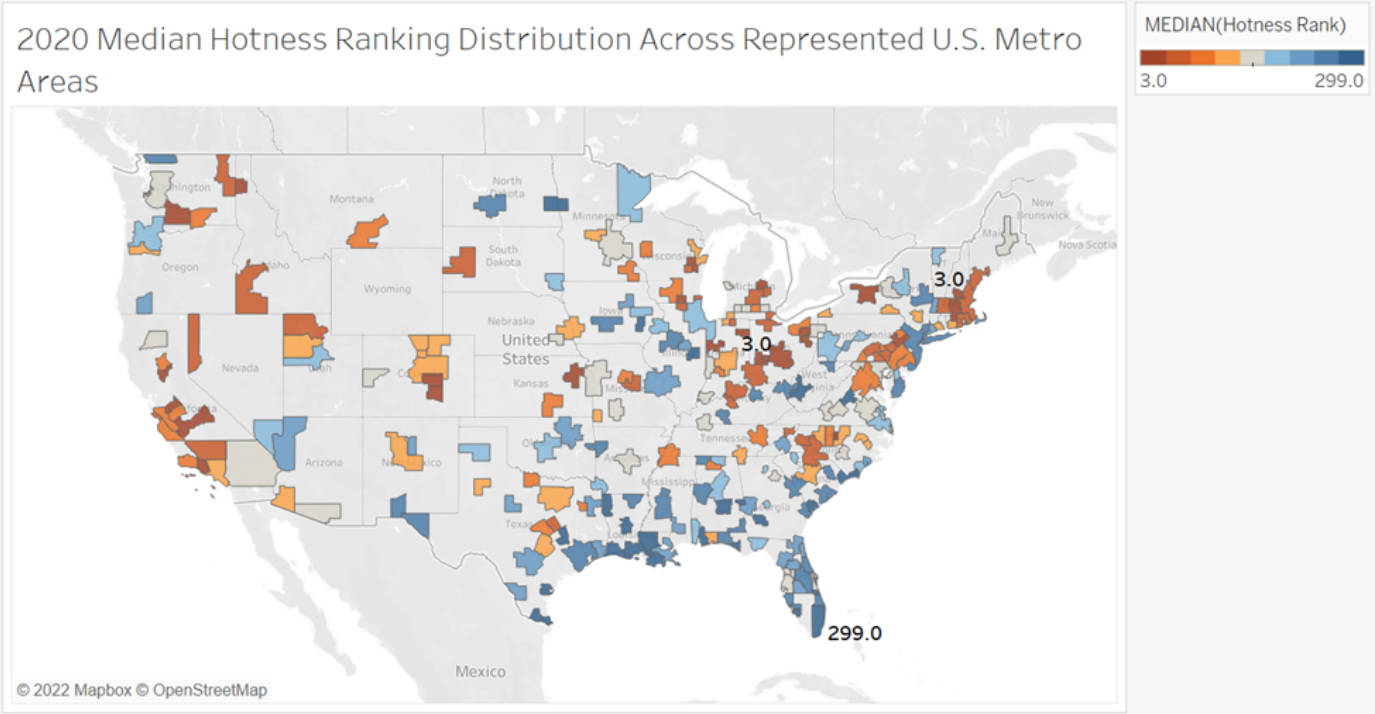
1. Picchi, A. (2021, April 20). *Hot Housing Markets: Where Are They - and why are they so hot?* CBS News. Retrieved May 1, 2022, from https://www.cbsnews.com/news/housing-market-hot-why
2. Wikimedia Foundation. (2022, April 28). *Metropolitan statistical area*. Wikipedia. Retrieved May 1, 2022, from https://en.wikipedia.org/wiki/Metropolitan_statistical_area
3. Conner, R., Puckett, L., & Murphy, N. (n.d.). *Real Estate Data*. Realtor.com Economic Research. Retrieved May 1, 2022, from https://www.realtor.com/research/data/
4. Akuffo, R. (2022, April 26). *Real estate: Why Tampa suddenly has the hottest housing market in the U.S.* Yahoo! Finance. Retrieved May 1, 2022, from https://finance.yahoo.com/news/real-estate-tampa-hottest-housing-market-us-190202227.html
5. Bahney, A. (2022, April 20). *Home prices hit another record high in March*. CNN. Retrieved May 1, 2022, from https://www.cnn.com/2022/04/20/homes/us-existing-home-sales-march/index.html

Appendix A - Metropolitan Area 2020 Population Estimates

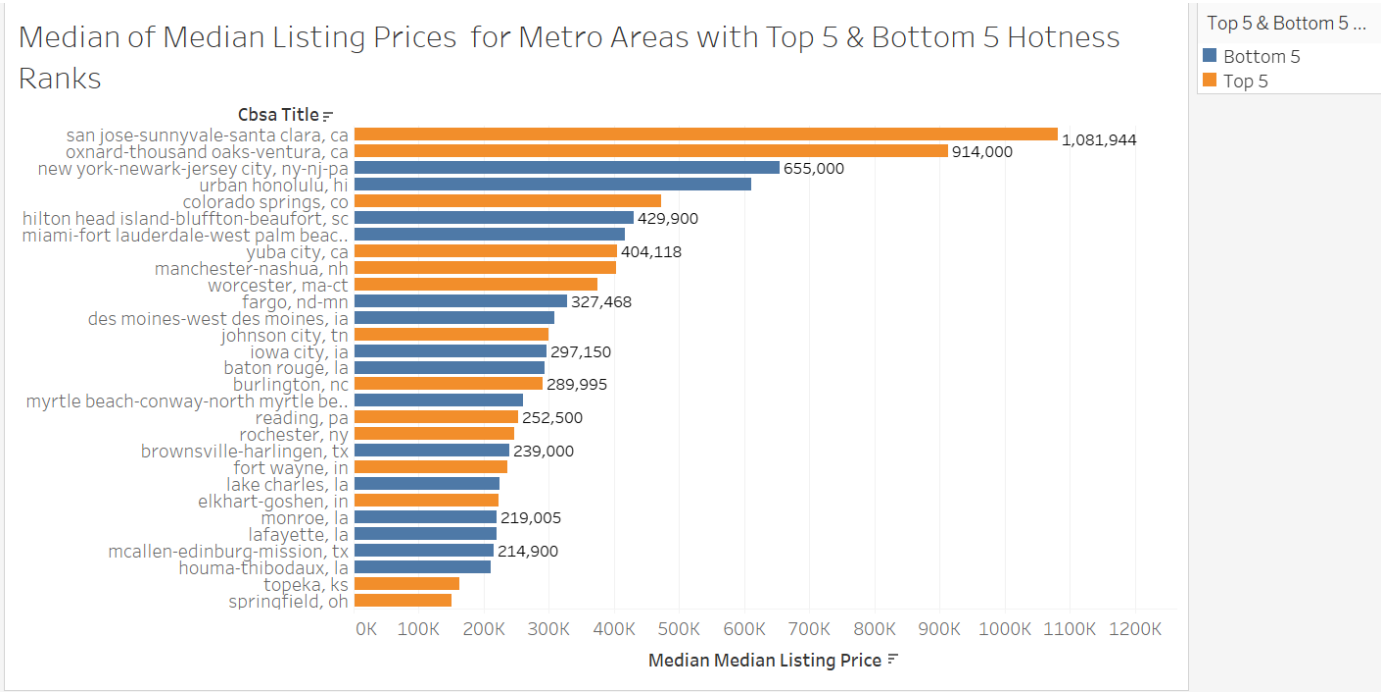




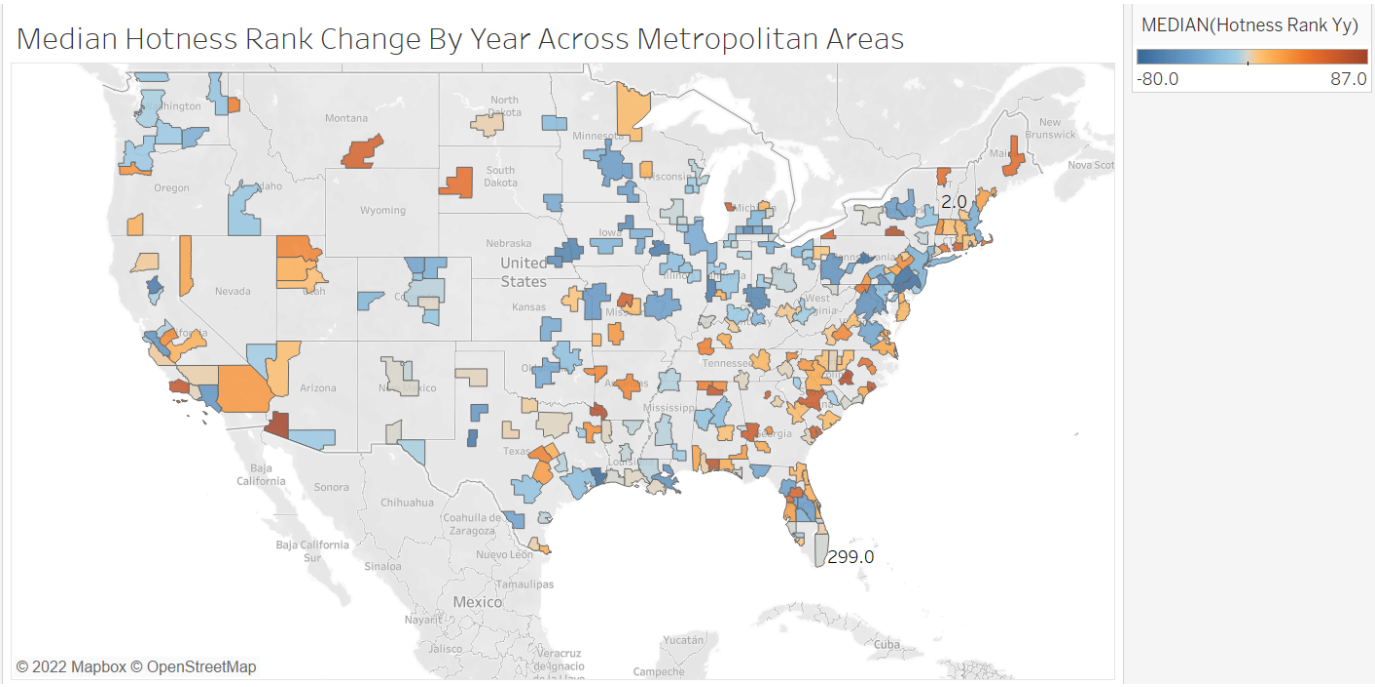
Appendix B - Median Hotness Ranking Distribution Across Metropolitan Areas



Appendix C - Median of Median Listing Prices Among Top-Ranked and Bottom-Ranked Metropolitan Areas



Appendix D - Median Yearly Change in Hotness Rank by Metropolitan Area for Years 2020 & 2021



Appendix E: Implementation Plan

Step	Who	What
1.1	Stakeholders	Identify the business problem and the problem's impact on the organization
2.1	Managers & Product Owners	Develop a vision for the project and OKRs to set goals for the work
3.1	Data Scientists	Gather and prepare the housing market data
3.2	Data Scientists	Brainstorm ideas for visualizations
3.3	Data Scientists	Create data visualizations and workshop to align with business problem
3.4	Data Scientists & Stakeholders	Provide overview of visualizations and mark where changes/updates need to be made
4	Data Scientists & Stakeholders	Deploy visualization dashboard