Emily Phillips
Dr. Catherine Williams
DSC 680 – Applied Data Science
May 1, 2022

What Makes the Housing Market "Hot"? – 10 Questions & Answers

1. What other factors pertaining to livability do you think play a role in market hotness?

I would have liked to explore other factors such as tourism numbers, number of people working in the metropolitan area, number of people moving to the area, etc. Population counts are a good measure for the number of people currently living there and the percentage change from 2020 to 2021 helped to understand trends in where people are now and where they are going, but there is more to "livability" than just number of people living somewhere. It would have also been nice to find survey data on metropolitan areas, to better understand how the people of the United States view the areas and where they think the popularity lies.

2. Are there any other visualizations that you thought about using to aid the investigation of the business problem?

My focus for visualizations was in using geospatial charts, bar charts and line charts. However, I also considered using bubble charts, since they allow for a depiction of three numerical measures in respect to each other. The dataset for this project contained mostly numerical measures, so this type of chart would also be useful for investigating the business questions. I also used one scatter plot in the project code, but more of them could have been utilized for understanding relationships between two numerical measures.

3. Do you think any data integrity was lost from the sub-setting of the datasets?

Yes. The Realtor.com datasets originally contained data from years 2016-2022, but with sub-setting, the final dataset only focused on years 2020 and 2021. If all the yearly data had been

included, it would have been useful for really understanding the trends over time in regard to the metropolitan housing market.

4. Has the evolution of the project made you change your mind on approaching the modeling problem from a classification or a regression standpoint?

The modeling portion of this project really threw me for a loop! I think if I were to come back to this project, that I would home in on what I want my target variable to be and ensure that it accurately represents my data and the business problem at hand. Therefore, it could end up being a classification problem or a regression problem. The target variable chosen for this project had three-hundred target classes, which posed a problem for classification accuracy and precision.

5. Did any of your results differ from the facts specified in the referenced sources?

The article about the city of Tampa is from 2022, and my dataset ranges from the years 2020-2021, but from the visualizations for the data, Florida was on the cooler side of the hotness rank scale for these given years. The new year of 2022 looks to be bringing in great fortune and luck for Tampa which is cool to see from the data in comparison to the article!

Also, in reference to the article about rising housing prices, the data results displayed this fact as well, with the median yearly change in listing prices increasing for all, but one) metropolitan areas that had rankings from 1-5 and 296-300 (top 5 and bottom 5 respectively).

6. What did you learn from the data cleansing portion of the project since it was extensive?

Merging datasets can be extremely useful for combining related useful data, but it can also pose risks in losing quality data.

Emily Phillips
Dr. Catherine Williams
DSC 680 – Applied Data Science
May 1, 2022

7. Is there anything that you would do differently that was not specified in the corresponding paper?

The paper covers mostly all these details! My big takeaways though are:

a. Find other area livability datasets
b. Try other types of joins to not lose as much yearly data (or find livability data that extends beyond two years)
8. Why did you focus on metropolitan areas over other geographic types?

My initial thought was to make the geographical focus of the project on U.S. cities. However, the Realtor.com market hotness dataset only had historical data at the metro, county and zip code levels, and the county and zip code levels seemed too granular. Therefore, to utilize both the inventory and market hotness datasets together, I focused on metropolitan areas in order to find overlapping information that could ultimately be merged.

9. Was there an advantage of using the historical data from Realtor.com over just the current snapshot dataset?

With trends being such a critical part of the housing market, historical data was the best route for being able to identify and analyze these trends over time.

10. How did you manage data outliers or skewness?

As mentioned in the paper, the Realtor.com datasets contained a column named ‘Quality Flag’ that flagged any unusual (outside-of-range) records with a “1”. I took this as being my outlier identifier, and removed any of these flagged records from the final dataset. When I tried to investigate outliers and data skewness in the Python code, it was a bit difficult for me to identify

true outliers since my business knowledge on the housing market is not proficient. Therefore, I figured I would trust the experts and take their quality identifiers as the source of truth for data outliers and skewness.