Peer Review

EAINT MAY, NORA PHELAN, ZACK TAYLOR, ASH TODD, CMPU 250, Vassar College

1 INTRODUCTION

In the introduction, the motivation and research questions are clearly stated and the sources are relevant and cited. The goal of the paper is also clearly stated. The motivation could be clearer. It is not entirely obvious why comparing Airbnb and Zillow data matters or how this comparison connects to the broader issues of bias in AI-driven pricing tools. Additionally, there is no clear hypothesis and there is no explanation present for the research question "How does short-term rental activity influence AI-driven New York Metro Area price estimates?" as there is no discussion of how short-term rentals might influence Zestimate data. On top of this, the background given for the Zillow algorithm is quite sparse. It would be helpful to know more about how Zestimate was created and used, including how it's trained and what inputs are used. It is also unclear how the comparison of the Zestimate data and Airbnb data can show bias. A brief explanation could make the paper's focus more understandable. The rest of the paper could also be easier to follow with the inclusion of some key terms and their definitions (AVMs, Disparate Impact Rule, socioeconomic disparity, etc.).

2 DATA

The data section appropriately cites all data sources, including Zillow's ZORI, Inside Airbnb, and NYU Furman Center's CoreData, with clear references and dates. The timeline for data collection is also laid out: Zillow's data spans from March 2024 to February 2025, Airbnb's from March 2025, and the demographic data is sourced from a May 2023 update. While this provides a strong foundation for transparency, the criteria for choosing the ten NYC neighborhoods are only briefly mentioned. Expanding on how these neighborhoods were "strategically selected" would strengthen the transparency and clarity.

Data cleaning is generally well explained: column names were standardized, missing values were removed, and specific filters were applied to Airbnb and demographic datasets. However, the section would benefit from more detail on what specific columns or values were considered "unnecessary" and how the decision to remove them was made. Additionally, the process of standardizing columns is mentioned, but not explained. The acknowledgment of limitations of Zillow and Airbnb data is helpful and adds to the transparency of the data analysis process. The cases and relevant variables are not described in this section. It would be helpful to describe them to give more background information about what each case represents and what key variables are being analyzed to provide more context for understanding how the data ties back to answering the research questions.

3 METHODS

In the methods section, the visualizations used correspond to the stated research questions. But these visualizations can be tweaked to improve readability. Keeping neighborhood order consistent in the first two bar graphs would allow for easier comparison between AirBnB and Zillow prices. Additionally, the heat maps are hard to read, as the Black population by percentage is out of order in figure 2.1, and there are big jumps in that percentage that don't translate well into the heat maps. Instead of heat maps, scatter plots with each dot color coded by neighborhood would better show the differences in the data. These would have the same x-axis, and the value represented by the color gradient would be on the y-axis. Figures are also not given figure numbers, so those labels should be added. Lastly, we are unsure if the color palette used for the heatmaps is colorblind friendly, so that would be something to double check.

4 RESULTS

The data visualizations currently in the methods section should be placed into the results section of this report. These plots are the physical representation of the resulting data analysis meant to answer the research questions. In addition to this, the results section should contain a written description of the visualizations breaking down the information presented by the plots.

When answering the research questions there should be more specific references to the findings. In question three, the report says "race is a stronger predictor of price disparity than income", however there is no reference to a specific percentage or datapoint to explain why this is supported by the findings. There should also be clarification for what data represents actual home values. In question two the report says "zestimate values often deviate from actual market activity…" It would be helpful to explain how it deviates and which direction it deviates. Additional clarity could also be provided by explaining what data shows "actual market activity". In question four, the report says "a strong correlation between Airbnb listing prices and Zillow's rent estimates". It would be helpful to show the supporting evidence or explain the correlation between the listings.

5 DISCUSSION

In the discussion section, the answers to research questions are summarized and supported. Also, the limitations of the analysis, generalizability of the findings, and ethicality of the work are all addressed adequately. However, after the sentence describing "Through investigating the alignment of Zillow home value estimates, Airbnb listing prices, and neighborhood demographic data for nine NYC neighborhoods," it would be good to describe some of the results from this investigation again for coherency.

6 ANALYSIS CODE

The analysis code looks correct and is readable and easy to follow with sections following the format of the writeup. In terms of style, the code is mostly clean and organized into distinct blocks for each figure. The code is commented well. The markdown descriptions include all the details from the writeup and this helps ground the reader in the research questions being answered. Regarding accessibility, the plots are well-labeled, include color bars, and use legible font sizes.

7 GENERAL

Overall, the preliminary analysis and corresponding code is written clearly and easy to follow. There is one consistent typo in the document, where the number of neighborhoods is stated to be either nine or ten at different points. There are also some formatting inconsistencies between sections. The font used should be kept the same, and section and subsection heading should be different sizes. Additionally, parts of the preliminary analysis template were left in where they should have been removed.

The project repo is well organized, but missing a pdf copy of the preliminary analysis.

Going forward, focus on simplifying visualizations and keeping your content coherent and consistent. There are a few points within the report where the descriptions of findings become repetitive and some small bits of background/context are missing. Other than that the report is easy to follow and comes to useful conclusions.

8 FINAL CONSIDERATIONS

Overall, the report is coherent and easy to follow. It flows well and the motivations, findings, and subsequent conclusions are clear. One additional thought to consider: what solutions would you propose to solve the issues discovered in your analysis?