Marissa Walker
Data Scientist
Denton, TX

2024 Median Rental Prices

Exploratory Data Analysis for Machine Learning Honors Project
12 January 2024

Project Overview

Motivation

I am passionate about social justice in many forms, but recently have been especially interested in housing. As a renter, I have experienced firsthand how difficult it can be to find affordable housing in a good condition. The book *Evicted* by Matthew Desmond especially opened my eyes to the horrors of eviction and the lack of tenant protections.

I am curious to look more into housing data, so here I explore a simple notebook I've created to look at median rental price estimates obtained from the United States Department of Housing and Urban Development (HUD). This notebook will outline my main findings from the data set, but further details can be found in the accompanying notebook.

Data Set Description

The data set explored here was downloaded from this HUD website:
https://www.huduser.gov/portal/datasets/50per.html

Specifically, I downloaded the "Data by County" file for 2024. The data consists of 15 columns and 4764 rows, each row representing a different locality.

The columns are:

- Five columns that are codes for the locations (state_code, county_code, county_sub_code, fips2010, and hud_area_code)
- Four columns that include names of the locations with different levels of specificity from town to state (town_name, cntyname, hud_areaname, and state_alpha)
- Five columns for the median rent price estimate for different numbers of bedrooms (0 to 4)
- One column with (I assume) the population from 2020 (pop2020)

A note from the README attached to the data set states:

> There is one record per county or county subdivision (New England town). The rents for all component counties of a metropolitan statistical area ("MSA") or HUD Metropolitan FMR Area (HMFA) are the same, so there will be duplicative rents for each county in a metropolitan area.

Data Exploration

Data Exploration Plan

The initial data exploration plan was:

- Determine basic facts about the data set, such as the number of columns, rows, and null entries
- Visualize the overall rent distributions for different numbers of bedrooms using box plots and histograms
- Explore the relationship between rent and population using scatter plots

Data Cleaning and Feature Engineering

This data set is already very clean, since it is not a raw data set but the result of research by HUD. The only null values are in the town names, which makes sense since the data is grouped by county, and many counties will have multiple towns.

So mostly my data cleaning was just cosmetic:

- I dropped the location code columns that I wasn't interested in.
- I renamed columns to have more human-readable names (e.g. "rent_50_1" renamed to "One Bedroom" for the rental price column for one bedroom units.

However, the note mentioned above (that each county has the same rent duplicated for all rows within the same metropolitan area) does prove to be an important one if we want to look at the relationship between rent and population. In order to account for this, I aggregated by the HUD area name column, using the mean to aggregate one-bedroom rent and adding the populations to create the total population column. (Note: This step could probably have been avoided if I downloaded the "Data by Area" file instead, but it was a good practice exercise anyway!)

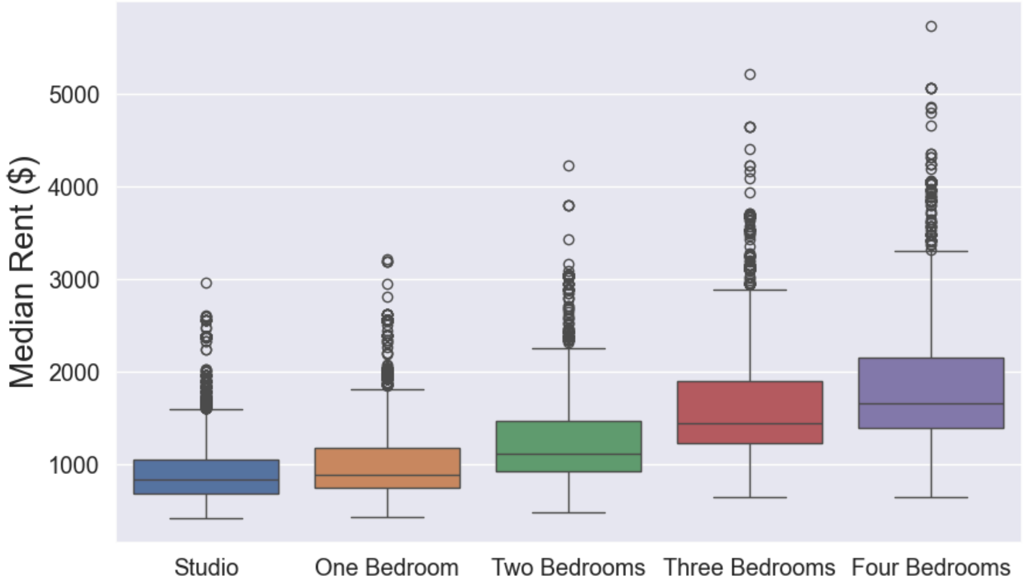
Finally, I also applied a log transform to the rental price data to correct (some) for skew. The skew in the 1-bedroom rent was above 3. After applying the logarithm, it was 1.7. While this is still skewed data, it's less skewed.

Key Findings and Insights

Here I will show a couple of the main plots from the data exploration phase and highlight the key findings from this exploration.

Rental Distributions:

The following is a boxplot of the median rental distributions for different numbers of bedrooms across counties in the USA (Note: this plot is before I aggregated by area):

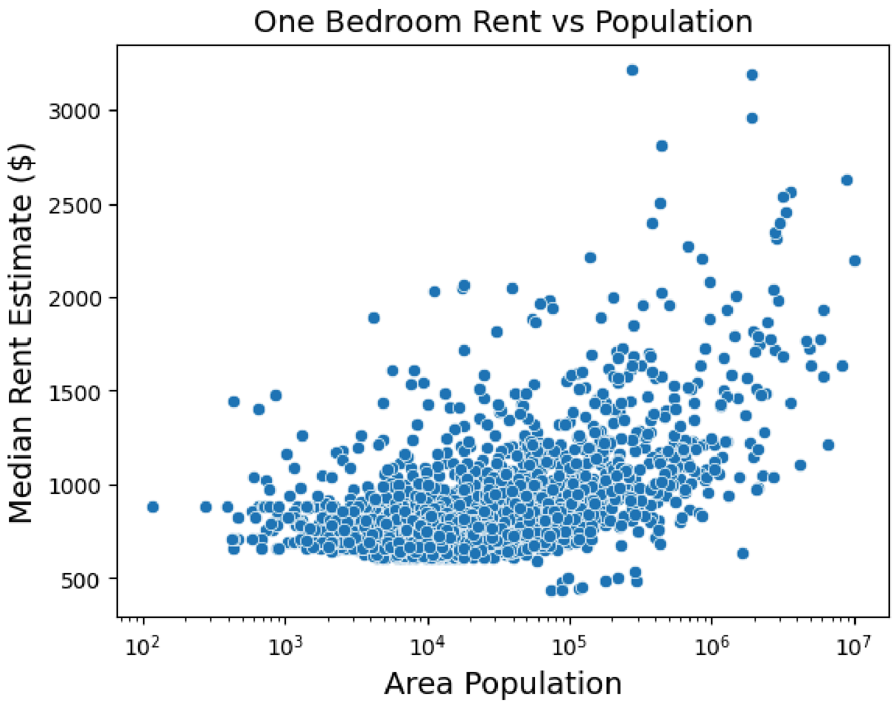


Takeaways:

- There is a clear (and not surprising) trend between the number of bedrooms and rent.
- Rent is too high! Especially among the outlier counties shown by the circles above the boxes in the plot.
- Potential steps for further exploration:
    - Investigate outliers
    - Divide up the data set to explore rental distributions by state or by different sizes of population
    - Download additional data sets from past years to compare

Rent vs Population:

After aggregating the one-bedroom median rent data by HUD area name (as described above), the median rent vs area population plot was created (Note: the x-axis is plotted using a logarithmic scale for better visualization of lower-population areas):



Takeaways:

- There appears to be some relationship between population and median rent, although it is not straightforward. There's a wide distribution of the rent across all populations, but the lowest rent areas seem to tend to be lower population, and highest rent areas correspond to higher population
- At low area populations (below around 60,000), there appears to be a flat bottom on how low the median rent per county can be
- There are a few areas above that population that somehow have lower rent, but in general once you go above 100,000, as the population increases, the minimum median rent by county seems to increase
- Potential steps for further exploration:
  - Investigate the small cluster of points around 100,000 with lower rent
  - Visualize rental prices using mapping (e.g. a choropleth map)
  - Divide the set into different population ranges and compare the distributions of rent (as individual points in the scatter plot make it difficult to determine the numbers)

Hypothesis Testing

Now that I have performed some basic visualizations to gain some understanding of the data, I will dive into forming and testing hypotheses statistically.

Hypotheses

Based on the visualizations and my own curiosity, here are some hypotheses that we could test:

1. Rent for two bedrooms is higher than for a one bedroom
    - Could test any other numbers of bedrooms
    - Very basic and shouldn't be surprising but will be useful practice to test
2. Rent in higher population areas is higher than in lower population areas
    - I would need to think a bit more carefully about how to define this hypothesis - specifically how to group the population ranges
3. Rent in CA is more expensive than in TX
    - Interesting to me since I have lived in CA and TX. California rent is notoriously high so I don't expect to be surprised here, but I'm interested to visualize and compare
4. Rent in 2024 is higher than in 2023
    - This is a hypothesis that would require a further data set to test, but would be fairly straightforward assuming the HUD page kept their data in similar format in previous years.

Hypothesis Testing

I performed the T-Test between the one and two bedroom rental price columns (after transforming them using the natural log function). My code prints out my conclusion based on the p-value:

```
Conclusion: since p_value 6.460569576518981e-222 is less than alpha 0.05 , I reject the null hypothesis that there is no difference between One Bedroom and Two Bedroom rent.
```

This is not at all surprising! But it's a relief to know that my very common sense hypothesis can be easily tested statistically. However, some future work I could do would be to better understand different statistical tests and ways to transform the data, and which to apply for what kinds of data. This was briefly covered in the course but I would like to explore it more.

[ ]:

Summary

Next Steps

I’ve just barely scratched the surface of the insights that can be seen just from this data (let alone compiling data from other sources). I mentioned a few potential next steps in an earlier section, but I’ll repeat some here:

- Investigate outlier counties/areas in the rental distributions, taking into account population
- Compare rental distributions across different regions or population levels, possibly using a choropleth map to visualize the rent
- Download additional data sets from past years to compare
- Test the other hypotheses listed above, and explore different types of significance tests to conduct

Overall Data Quality and Conclusions

This data set was worth exploring, and there is still more to be discovered here! The data set is already clean, which helped immensely. However, I do have questions about the data set and I believe I would benefit from a deeper understanding of its origins and the assumptions that are behind it.

For example, some additional data that would be interesting to explore or know: How were the median rent prices estimated? Is there a data set that divides up the data further, perhaps by type of property (e.g. apartment vs. single-family house)?

Finally, I hope that any next steps I take on this data set (or exploring this with other data) would keep in mind the bigger questions of economic justice. What is the rent-to-income ratio? Who bears the greatest burden of high rent-to-income? How is this affected by systemic issues like racism? Who benefits from this unjust system? Others much more qualified than I am have already explored these questions in depth, but I hope to continue looking into them as I continue my data science journey.