Math 345 Final Project: An Analysis of Craigslist Housing Listings

Jaylin Lowe and Jenna Korobova

6/8/2020

Introduction

Since its establishment in 1995, Craigslist has been a widely used online resource for individuals to post advertisements in a variety of categories, such as jobs, household articles, services, pets, and housing. In particular, it has become increasingly common for homeowners and landlords to advertise homes and apartments for rent through Craigslist. Moreover, individuals seeking roommates often post advertisements for their rentable spare rooms. We are interested in investigating what characterizes a listing having a higher rent, whether the most influential effects are at the listing or state level, and what interactions exist between the variables.

Methods

Our analysis is based on a dataset obtained from Kaggle. The original dataset consisted of 384,977 observations, where each observation was a privately sold housing option displayed on Craigslist in January of 2020. It contains information about each housing option, including the state, city/region, price, type of building, square footage, number of bedrooms, number of bathrooms, whether cats or dogs were allowed, whether smoking was allowed, if there was wheelchair access, if there was a charging station for electrical vehicles, and if furniture was included. However, we also obtained state level data and added that information to the dataset. Median income, population, and land area were all obtained from the Census Bureau. We also included whether a state had the death penalty or not, the median tax property rate, the political party of the current governor, and the percentage of white and black people in each state. This information was obtained from deathpenalty.org, mortagecalculator.org, ballotpedia.org, and governing.com, respectively. Full citations can be found in Appendix A.

Since the original dataset included apartments, houses, flats, condominiums, and a variety of other housing options, we decided to restrict our analysis to apartments only. Almost all of the original observations (318,032) were apartments, so we chose to focus on those. We also removed all observations with a price higher than 5,000 dollars. There were only a small number of observations that large, and we wanted to focus on the prices of apartments we considered more reasonable. After excluding these observations, we drew a random sample of size 2,000 from our dataset. Since the original dataset was so large, it would have been impractical to create a model from the entire dataset. However, we also later realized that there were unusually small prices in our dataset, ranging from 0 to 5 dollars. We considered these to be unreasonably small prices, likely inaccurate, and subsequently removed them from the dataset. Our dataset also originally included the Washington DC area, but after realizing that the population density for DC was near 10,000 people per square mile while the next highest state was near 1,000 people per square mile, we decided to exclude observations from DC entirely. After making these final changes, our final dataset consisted of 1,976 observations.

Since only about 1% of all observations had a charging station for electric vehicles, we decided to ignore that explanatory variable. We also combined variables tracking whether cats were allowed and whether dogs were allowed into one explanatory variable. This variable has a value of 1 for apartments where both dogs and cats were allowed, and 0 otherwise. After these modifications, the final explanatory variables we considered for modeling were: price, number of bedrooms, number of bathrooms, square footage, if pets were allowed, if smoking was allowed, if there was wheelchair access, if furniture was included, median income of the state (in dollars), political party of the state's governor, population density of the state (in people per square mile), median property tax rate of the state, whether the death penalty was legal in the state, and the percentages of black and white people in the state.

Our strategy for building multilevel models began with exploratory data analysis at each level. We first explored the effects of random intercept by comparing an unconditional means model and random intercept model grouping by state. We then used simulation to test if we needed random slopes for some of our variables. Since testing for random slopes on our variables resulted in overly complex models, we only looked at random slopes for listing level variables that our exploratory data analysis suggested might be significant. These variables were: number of bedrooms, number of bathrooms, and square footage. Our simulation revealed that random slopes for these variables were unnecessary, so we proceeded with investigating the fixed effects for all variables with only a random intercept for state.

Results

EDA

We began our exploratory data analysis by analyzing several of the numeric variables in the dataset—**Table 1** shows the summary statistics for these variables. We first noticed that the distribution of rent price was skewed right, as the mean price was 1138 dollars and the median price was 1005 dollars. This led us to consider the log-transformed version of price, which had a median of 6.92 and a mean of 6.95, confirming that the log-transformation yielded a more symmetric distribution of price. The distributions of bedrooms and bathrooms also appeared to be relatively symmetric, with medians of 2 and 1 respectively. On the other hand, the distributions of square footage (median of 900), median income (median of 59,995), population density (median of 156.24), and tax rate (median of 1.02) seemed to have a slight right skew.

Variable	Minimum	1st Quartile	Median	Mean	3rd Quartile	Max	SD
Price	25.00	800.00	1005.00	1138.00	1350.00	4845.00	509.53
Log(Price)	3.23	6.69	6. 92	6.95	7.21	8.49	0.42
Bedrooms	0.00	1.00	2.00	1.72	2.00	5.00	0.75
Bathrooms	0.00	1.00	1.00	1.42	2.00	4.50	0.55
Square Footage	25.00	724.80	900.00	913.10	1063.20	5600.00	296.02
Median Income	44097.00	55462.00	59995.00	61738.00	70315.00	83242.00	9275.55
Population Density	1.01	88.05	156.24	186.24	241.38	1018.25	166.54
Tax Rate	0.32	0.77	1.02	1.17	1.61	2.31	0.53

Next we considered boxplots for the distribution of price by each of the level one dummy variables, as displayed in **Figure 1**. We noticed that the distributions of rent price were similar regardless of whether or not pets were allowed, with a median of about 1000 dollars for both. Furthermore, we noted that the median rent price for non-smoking listings was about 400 dollars more than the median rent price for smoking listings, and the range of prices for non-smoking listings was greater than that of smoking listings. We also observed that listings with wheelchair access tended to be more expensive (by about 50 dollars) and had a greater range of prices than listings without wheelchair access. Finally, we noticed that listings that were already furnished had a slightly lower median price but a greater range than unfurnished listings.

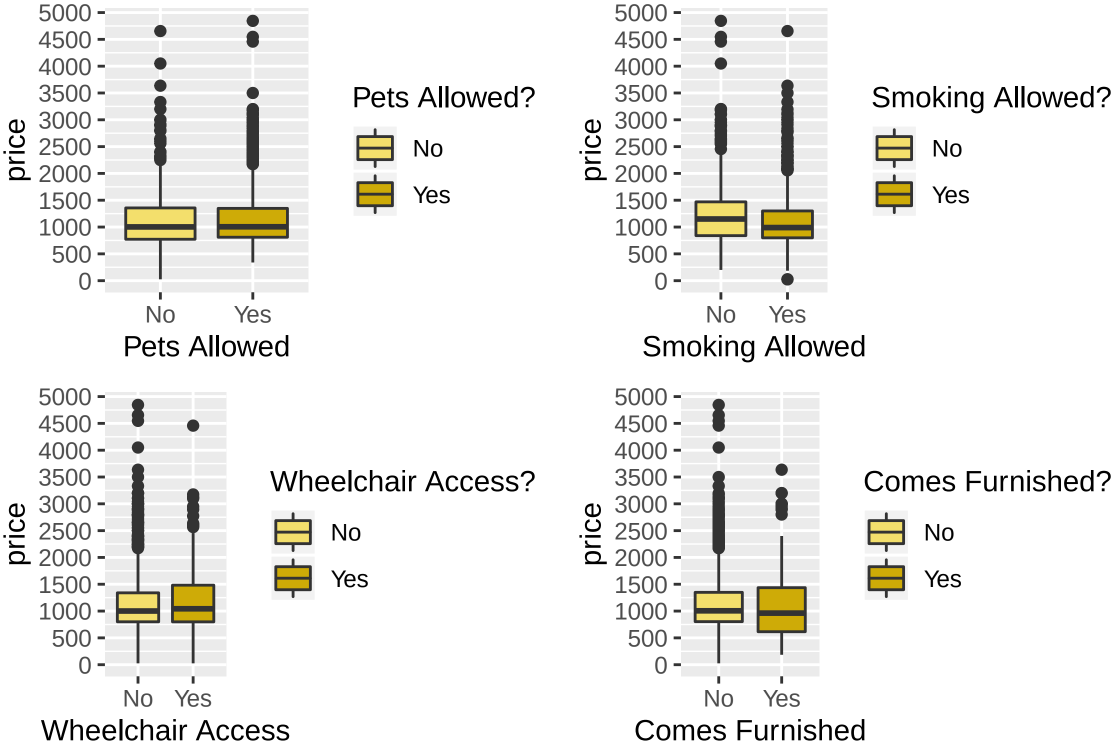

Figure 1: Boxplots for Level 1 Dummy Variables by Rent Price

At the state level, we considered the median price of listings in a given state, as shown in **Figure 2**. The 5 states with the highest priced listings were Hawaii, New Jersey, California, New Hampshire, and Massachusetts (median listing prices ranged between just over 1600 dollars to 1800 dollars), and the states with the lowest priced listings were Mississippi, West Virginia, Oklahoma, Kansas, and Missouri (median listing prices ranged between just over 600 dollars to about 750 dollars). This makes sense in the context of housing trends: Hawaii, California, and states on the East Coast have notoriously high rent prices, while states in the lower Midwest and lower East tend to have more affordable rent.

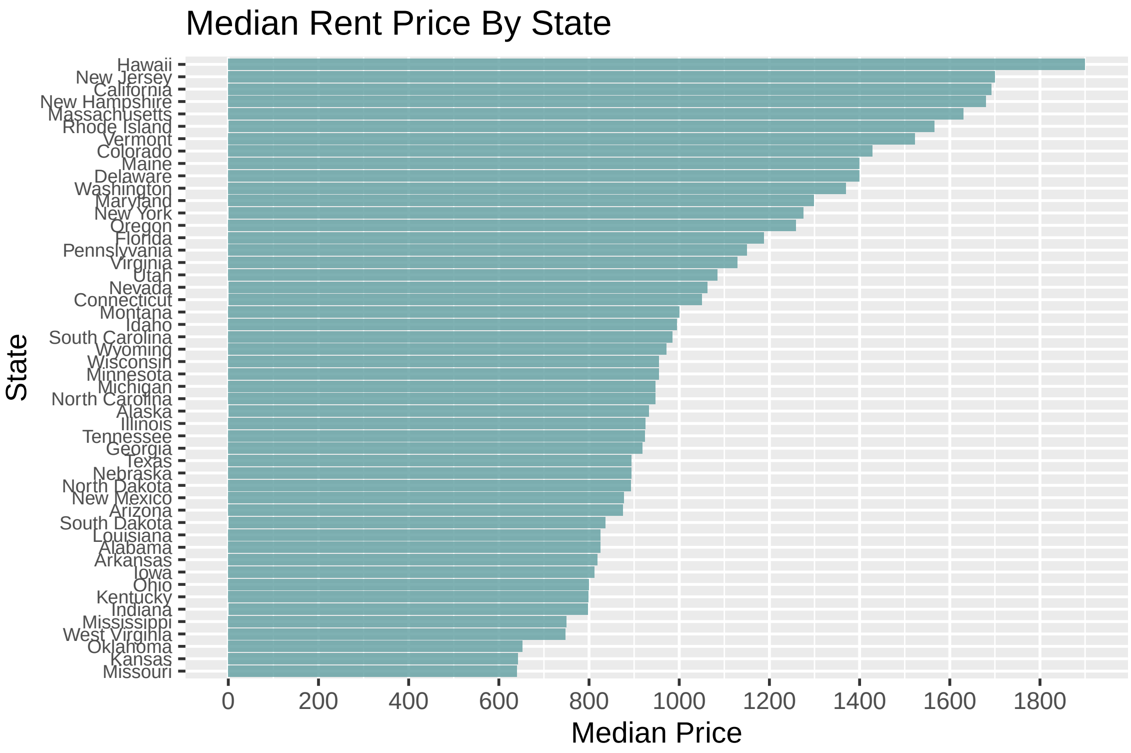

Figure 2: Median Rent Price by State

Model Analysis

Our final model predicted the prices of apartments listed on Craigslist using information about the number of bedrooms, number of bathrooms, square footage, whether pets were allowed, and whether smoking was allowed for each individual listing. It also included information at the state level, including median income, political party of the governor, population density, median property tax rate, percentage of white people and percentage of black people. Our model had a random intercept for the state variable, in order to adjust for the fact that listing prices are likely correlated within each state. 1,047 of the listings were from a state with a Republican governor, while the other 929 had a Democratic governor. 1,442 of them allowed both cats and dogs as pets, and 532 banned cats, dogs, or both. 1,511 of them allowed smoking, while the remaining 465 did not.

The response variable, price, was initially very right skewed, so we log transformed it. We also scaled square feet and median income. Summary statistics for the original distributions of square footage and median income are in the table above, but both price and the log of price are included.

Our final model had the form:

Level 1: $E(\log(\text{Price})_{i,j}) = a_i + \beta_0 \text{scale}(\text{sqfeet})_{i,j} + \beta_1 \text{beds}_{i,j} + \beta_2 \text{baths}_{i,j} + \beta_3 \text{petsAllowed}_{i,j} + \beta_4 \text{smokingAllowed} + \epsilon_{i,j}$

Level 2: $a_i = \alpha_0 + \alpha_1 \text{scale}(\text{MedianIncome})_i + \alpha_2 \text{Governor}_i + \alpha_3 \text{PopDensity}_i + \alpha_4 \text{TaxRate}_i + u_i$

where $\epsilon \sim N(0, \sigma^2)$ and $u_i \sim N(0, {\sigma_b}^2)$

The table below displays the estimates for each of the parameters:

Parameter	Estimate	Standard Error	T Value
α_0	7.152	0.068	104.425
α_1	0.164	0.023	7.099
α_2	-0.113	0.044	-2.571
α_3	0.0002	0.0001	1.882
α_4	-0.098	0.046	-2.113
β_0	0.202	0.012	18.703
β_1	-0.126	0.016	-9.246
β_2	0.071	0.016	4.312
β_3	0.075	0.014	5.231
β_4	-0.033	0.016	-2.107

For each individual listing, the square footage of the rental, along with the number of bedrooms and bathrooms were significant in predicting the price. Specifically, one additional bathroom was associated with an increase in median price by a factor of 1.07. An increase by one standard deviation in the square footage of a listing is associated with an increase in median price by a factor of 1.22. Interestingly, additional bedrooms are associated with lower housing prices. An additional bedroom is associated with a 0.88 multiplicative decrease in median housing price. With 95% confidence, an additional bedroom suggests a decrease in median price by a factor between 0.858 and 0.905. In addition to these three variables, binary variables capturing if pets were allowed and if smoking was allowed were also significant. Apartments that permitted pets were associated with an increase in price by a factor of 1.07 compared to those that did not, but apartments that permitted smoking were associated with a decrease in price by a factor of 0.97 compared to those that did permit smoking.

Four variables consisting of information about the state where the listing was posted were also significant. Of these, higher median income of a state was associated with higher prices, while Republican governors, higher property tax rates, and higher population densities were associated with lower prices. An increase by one standard deviation in median income of a state was associated with an increase in median housing price for listings in that state by a factor of 1.18, while an increase by one percentage point in the median property tax rate is associated with a decrease in median housing price by a factor of 0.91. Listings in states with Republican governors are associated with a drop in price by a factor of 0.89 compared to listings in states with Democratic governors. Finally, an additional person per square foot in the population density of a state is associated with a multiplicative increase by 1.0002 in the median housing price.

Discussion

Our analysis suggests that many aspects of both an apartment itself and attributes of the state it is in are significant predictors of the price of the apartment. Larger apartments with more bathrooms tend to have higher prices than smaller apartments with fewer bathrooms. Surprisingly, apartments with more bedrooms tend to have lower prices. We theorized that this may be true because apartments with more bedrooms might be listed on Craigslist by people looking for roommates. In other words, the apartment listing indicates the total number of bedrooms in the apartment, but the renter would only be using one of those bedrooms. Such listings would likely be cheaper than studio or one bedroom apartments that the renter would fully occupy. Our dataset did not differentiate between entire apartments for rent and people searching for roommates, so we cannot confirm this theory. However, it does suggest an interesting avenue for future research. Apartments that allow pets tend to be more expensive, likely because flexibility may be an indicator of nicer apartment buildings. On the other hand, apartments that allow smoking tend to be cheaper than those that do not. This may be true because higher end apartments tend to ban smoking.

Attributes of the states the listings were posted in are also significant predictors of price. Prices are higher in states with higher median incomes, likely because property values tend to be higher in areas with strong economies. Housing prices are also driven up by a lack of housing, which is more common in areas with higher population density, likely explaining why we found that higher population density is associated with higher prices. Republicans tend to be governors in states with more rural areas, since cities tend to be largely Democratic, which agrees with our assessment that listings in states with Republican governors tend to have lower prices. Higher tax rates were also associated with lower prices, which we found surprising. We expected that higher tax rates would be an indicator of more liberal areas, which would have higher prices. Alternatively, we expected that rent prices may reflect the higher tax rates, since landlords might need to charge higher prices in order to be able to pay the property tax. However, our analysis suggests the opposite. We were not sure what might explain this phenomenon, so more research on property taxes might be illustrative.

The nature of our resources prevented us from being able to consider all of the nearly 400,000 observations in the dataset, thus future work may choose to consider a more comprehensive approach to the data and fit models that take in all of the observations. We hope that our random sample of 2,000 observations was representative, though an analysis that includes all of the observations may nonetheless yield models that differ from our final model. Furthermore, by limiting the scope of our analysis of price, we removed the region variable that the original dataset contained. However, this variable may have been an informative third level to consider in our hierarchical model and thus may be worth exploring in future work. Finally, because we removed Washington D.C. Craigslist listings from our dataset, it might be useful to consider D.C. listings and their characteristics in future work. Rent in Washington D.C. tends to be expensive, and thus those listings may be important to include.