Bayesian Hierarchical Modeling of Toronto Airbnb Prices

STA2201H - Final Project

Kristine Villaluna

April 2022

Abstract

The City of Toronto is the most populous city in Canada and is recognized as an international centre of business, finance, arts, and culture. With the large amount of visitors Toronto sees everyday, an increasing number of people have been turning to alternative forms of accommodations when they stop by. This analysis aims to examine the different factors that affect the prices of Airbnbs in Toronto such as the size of the listing, information about the hosts, review scores, and the neighbourhoods or districts they are located in. Using a hierarchical fixed effect model, we model the log price using a room type structure using a combination of these covariates and find that the most important factors that affect listing prices are the number of people it accommodates, the number of bedrooms and bathrooms, and the location of the listing. Across all models tried, it was found that entire homes/apartments had a higher starting price than any of the other room types, with hotel rooms having the lowest intercept.

Introduction

The City of Toronto is the capital city of Ontario and has the highest population of any city in Canada, with a recorded population of 2,794,356 individuals according to the 2021 Census of Population (Statistics Canada, 2022). It is recognized as one of the most multicultural cities in the world and has a vibrant tourism economy. The nature of the city attracts individuals from all over the world to come visit for reasons such as business, finance, arts and culture. Sports fans can head over to the Rogers Center to catch a Blue Jays baseball game, or perhaps cheer on the Maple Leafs (hockey) or the Raptors (basketball) at the Scotiabank Arena. Whatever one's interests are, there is something for everyone in the city.

With the rising prices of hotel rooms, an increased number of people have been turning to other forms of accommodations while on vacation. This combined with other factors such as privacy and in suite kitchens/laundry, Airbnbs have been a great alternative to traditional hotels for lots of travelers. Airbnb's business model focuses on a marketplace platform where hosts and guests exchange housing for money (Airbnb, 2022). When listing their home on Airbnb, hosts have the ability to set additional prices for individual nights, weekly stays, cleaning fees, weekend prices, and additional guests.

The objective of this analysis is to examine the different factors that effect the prices of Airbnbs in Toronto. We will be examining a variety of factors, ranging from information on the listing itself, such as the neighborhood the listing is in, the type of room, the number of bathrooms, etc., to information about the host and the reviews a listing has. This is important as we would like to better understand the role that these factors play in the pricing of these specific listings.

Data

Although Airbnb does not have an official API available to the public, data is available from InsideAirbnb.[1] Inside Airbnb is a mission-driven project with the objective to provide data that quantifies the impact of short-term rentals on the residential market. There are a number of variables available in the data set, and for the purpose of this analysis, we have restricted the data frame to certain variables we would be interested in examining further. We have chosen to remove text variables such as name and description, as well as latitude and longitude coordinates. These data are from their most recent release, compiled on December 05, 2021 and contain 15,261 observations. This data is representative of the population, as it is all of the Airbnb listings in Toronto. Thus, we will proceed to use a random sample of the population in our analysis.

We have a diverse set of variables available to us to use in the analysis. The data contains information on the host, such as how long they have been a host for, their average response time, whether they are a superhost or not, and how many listings they have in total. We have information on the listing such as the neighbourhood it is located in, the type of room it is, and how big it is (number of bathrooms/bathrooms, and how many people it can accommodate). Finally, we have information on whether it has availability and its review scores, both overall and for different subcategories.

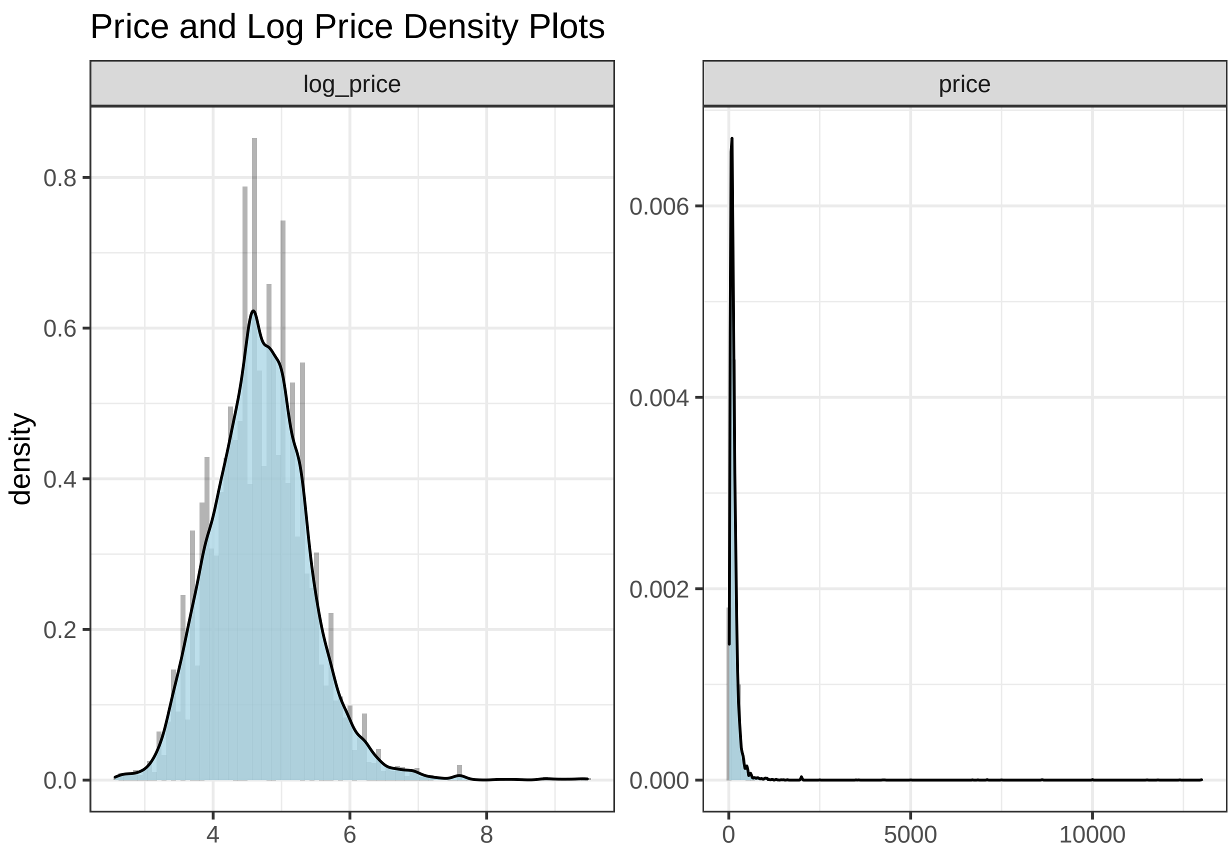
During the data preprocessing, it was found that there was a total of 15,261 observations, but only 15,121 distinct observations (140 duplicate rows). It may be plausible that a host has multiple listings at a given location, however, it is less likely that all the other variables are the same. Thus, these duplicates were removed from the data.

Looking at missing values, the bathrooms variable was completely empty. To overcome this, we extracted the required information from the `bathrooms_text` column. The review score variables had a complete rate of approx. 77%, the bedrooms variable was 93% complete, and then bathrooms and host variables were both approximately 99% complete. Looking closer at the review variables it was found that many of the reviews had a rating of 0, or only had one incomplete review. Regarding the bedrooms and bathrooms, it was not clear if imputation would be possible as there were many different room type and accommodates combinations. Looking at the host related variables, there was such a small number that removing these observations would not drastically alter the composition of the data.

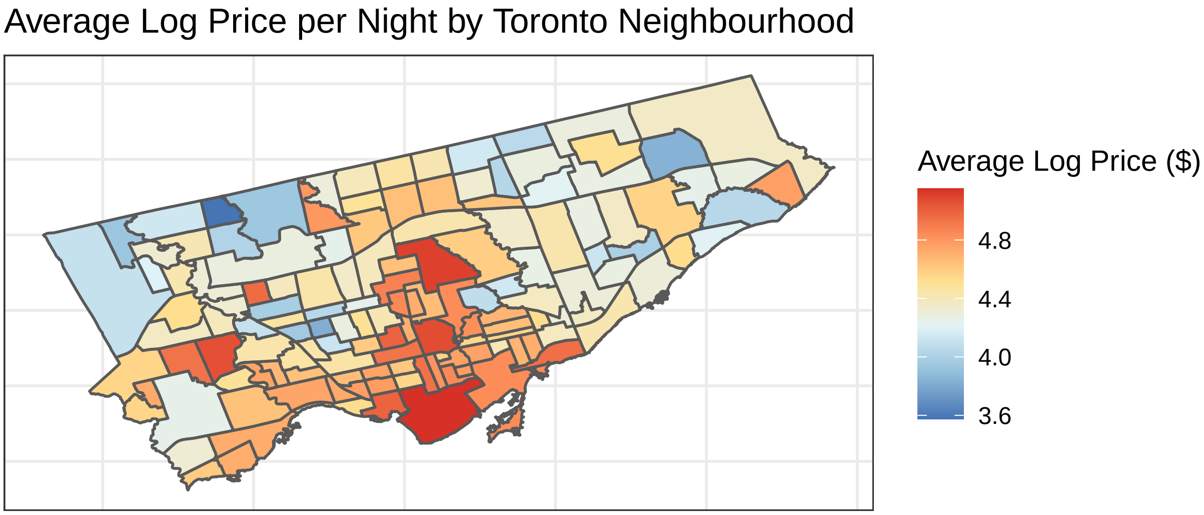
Examining these missing values provided us with lots of information in order to examine where we may want to consider removing variables or observations due to lack of information. Given the size of our dataset, and the fact that these are variables that we expect may be influential on the price of an Airbnb, we proceeded to remove these missing values due to a lack of information. After removing missing values and duplicates, we are left with 10,731 observations.

The main dependent variable of interest to us is the price of the listing. We will apply a log-transformation to obtain our dependent variable of interest, `log(price)` as price is heavily right-skewed as seen below in the top left panel.

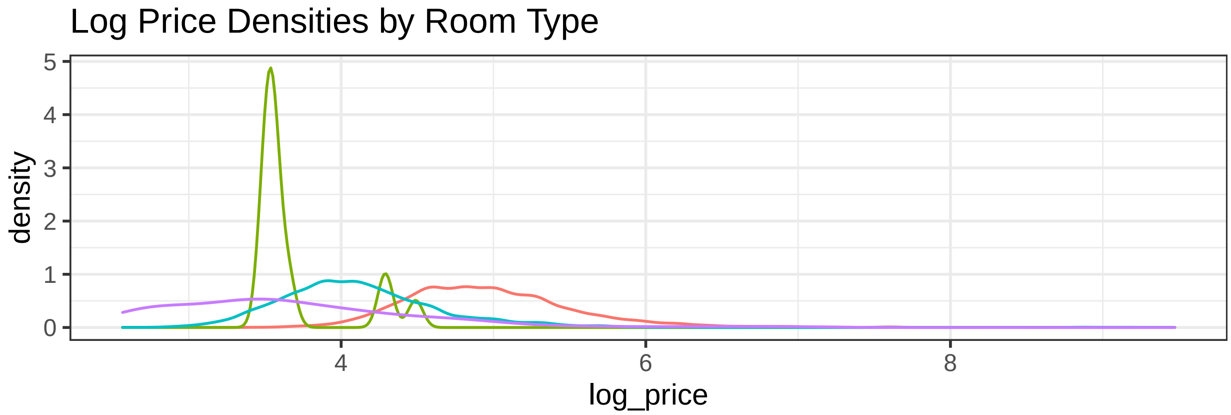
[1] The data dictionary is available here.

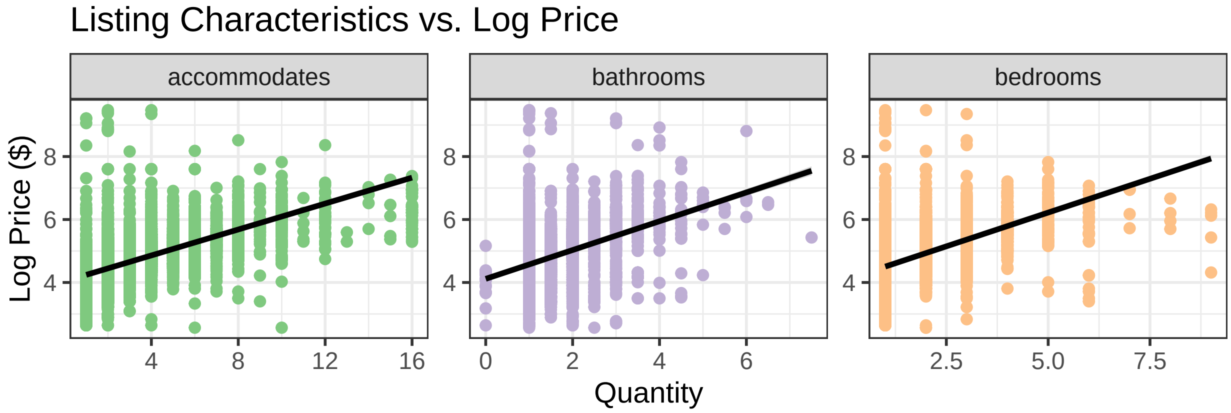

Price and Log Price Density Plots
log_price
price
density
0.0
0.2
0.4
0.6
0.8
4
6
8
0.000
0.002
0.004
0.006
0
5000
10000

Given that we are interested in modeling the log price, we can take a look at the average log price per neighborhood displayed in the map below. Neighborhood geography data is sourced from Open Data Toronto (2022). We see that price varies quite a bit depending on the neighbourhood the listing is located in. The top two most expensive neighbourhoods are the Waterfront Communities-The Island and Bridle Path-Sunnybrook-York Mills, with the least expensive being Black Creek.



After reviewing the remaining covariates, it was found that the log price varies considerably for each given room type as displayed in the top plot below. With this in mind, we will proceed to use this variable to structure our hierarchical model. In the bottom plot, we see that variables related to the size of the listing (accomodates, bathrooms, bedrooms) all have a positive relationship with log price. This is no surprise, as intuitively, the larger the Airbnb is, the more expensive it will be.





Further investigation into the data found that the Waterfront Communities-The Island has the most listings in the data. Looking at the proportions of the different room type it was found that most listings are for entire homes/apartments, with the next most frequent being a private room. Shared rooms and hotels are much less common.

More plots relating to the host and review variables are available in the .Rmd file. Overall, in the EDA it was found that variables related to the listing, the neighborhood, and room type were most influential on the price. Variables relating to the host and review scores did not have as strong of a relationship. Of the review variables, the overall rating, accuracy, cleanliness, and location would be of interest.

Methods

For this analysis, we will hierarchically model log price with fixed effects, using a room type hierarchical structure (within neighborhoods). This allows us varying intercepts for the covariates, but a constant slope. This hierarchical structure has been motivated by the EDA where it was found that price varied vastly across different room types.

Thus, the model specification is as follows:

$$\begin{aligned}
y_i|\alpha_{j[i]} &\sim N(\alpha_{j[i]} + \beta^T x_i, \sigma_y^2) \text{ for } i = 1, 2, ..., n \\
\alpha_j &\sim N(\mu_\alpha, \sigma_\alpha^2), \text{ for } j = 1, \dots, J \\
\mu_\alpha &\sim N(0, 1) \\
\sigma_\alpha^2 &\sim N^+(0, 1) \\
\sigma_y^2 &\sim N^+(0, 1) \\
\beta_k &\sim N(0, 1)
\end{aligned}$$

Where

- i is the observation number
- n is the total number of observations
- j is the room type number
- J is the total number of room types
- k is the number of covariates
- x_i is the vector of covariate data for observation i
- α_j is the room_type level mean log price
- β is a vector of k coefficients corresponding to each of the covariates

As a baseline, a linear (non-hierarchical) model will also be fit in order to compare the models. The simpler model will be specified as follows:

$$\begin{aligned}
y_i &\sim N(\alpha + \beta^T x_i, \sigma_y^2) \text{ for } i = 1, 2, ..., n \\
\alpha &\sim N(0, 1), \\
\sigma_y^2 &\sim N^+(0, 1) \\
\beta_k &\sim N(0, 1)
\end{aligned}$$

To validate our model, we will check model diagnostic plots such as traceplots and pairs plots to ensure the chains have mixed properly and that we have sampled the entire space. Checking convergence diagnostics such as the effective sample sizes (n_eff) and Rhat values evaluate how well the sampler is doing at sampling the posterior distribution of the given model. After doing these checks, we will then assess the model fit by examining some posterior predictive check (PPC) plots and compare the models using leave-one-out cross-validation (LOO-CV). This will help us understand if this statistical model is appropriate for the data or better than other models.

Results

In this section, we will discuss the results of the following models:

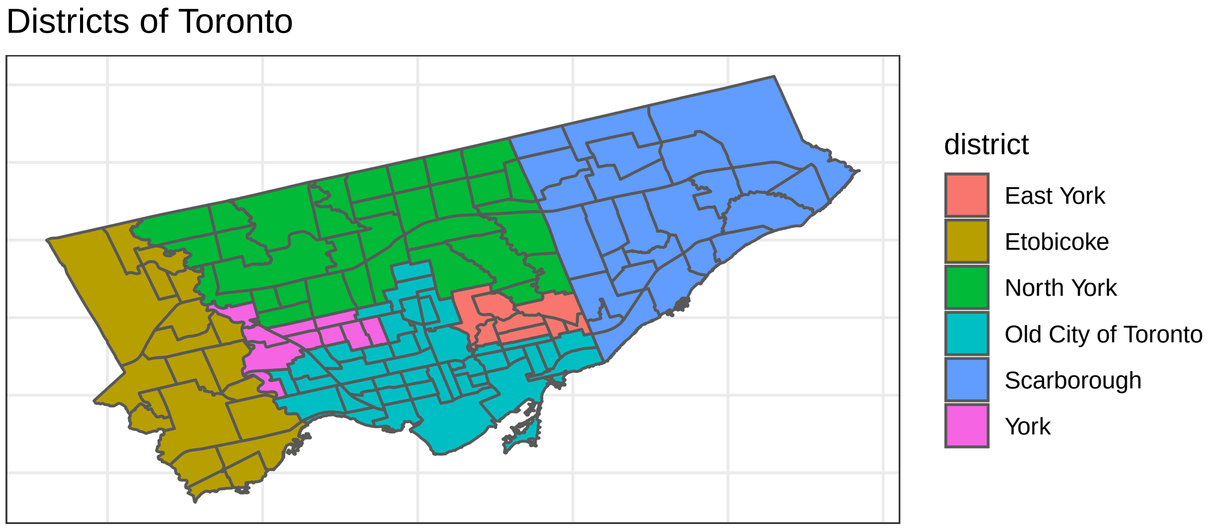
- Model 1: Hierarchical Model with all Covariates (including Neighbourhoods)
- Model 2: Hierarchical Model with all Covariates (including Districts)
- Model 3: Hierarchical Model with Selected Covariates (including Districts)
- Model 4: Linear Model with all Covariates (including Districts)

Model 1 - Neighbourhoods and all Covariates

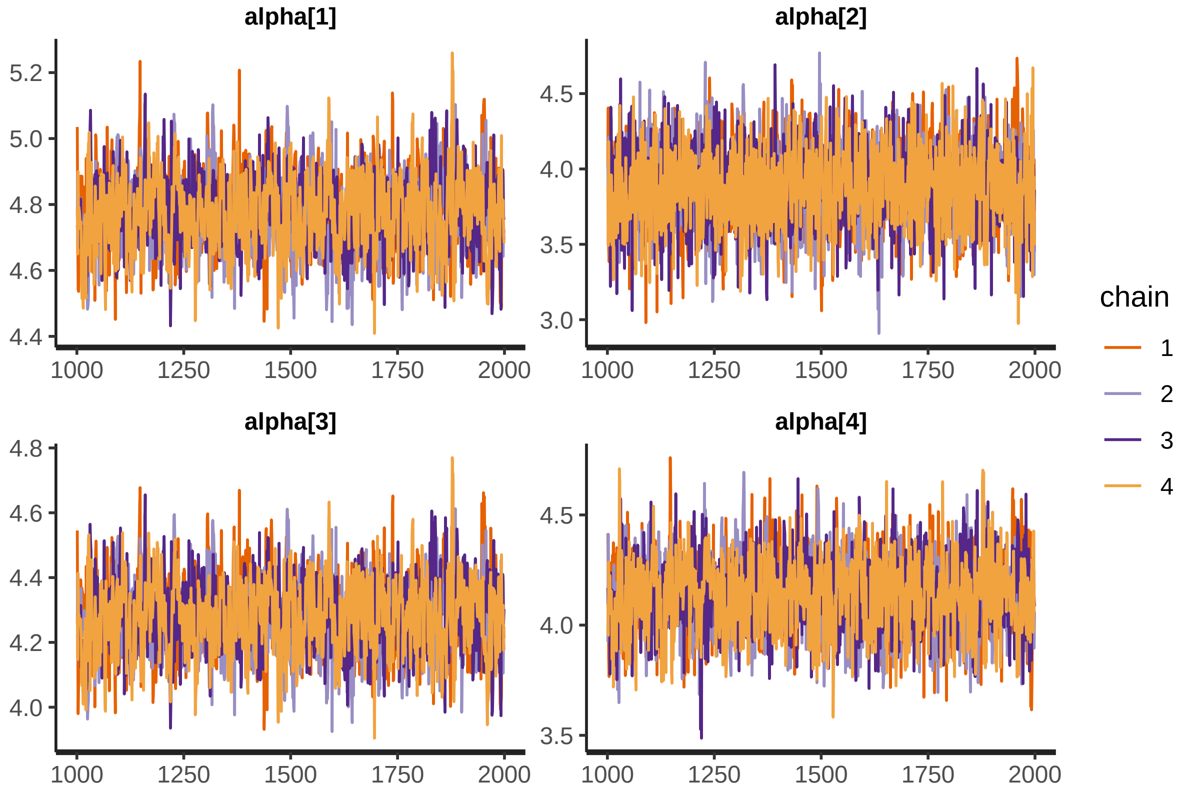
To begin our analysis, we fit a hierarchical fixed effects model modeling log price, using all of the available covariates, including neighbourhoods. Before fitting the model, numeric variables were mean-centered, and categorical variables were encoded as dummy variables. After running this first model, a warning was received that indicated that the Bulk Effective Samples Size (ESS) was too low, which indicated posterior means and medians may be unreliable. Checking the summary, the Rhat values seemed to be ≈ 1, with 1.02 as the highest. Because of this warning, we then considered grouping the neighbourhoods into larger geographical areas, and perhaps limiting the covariates to those which appeared to be of interest in the exploratory data analysis.

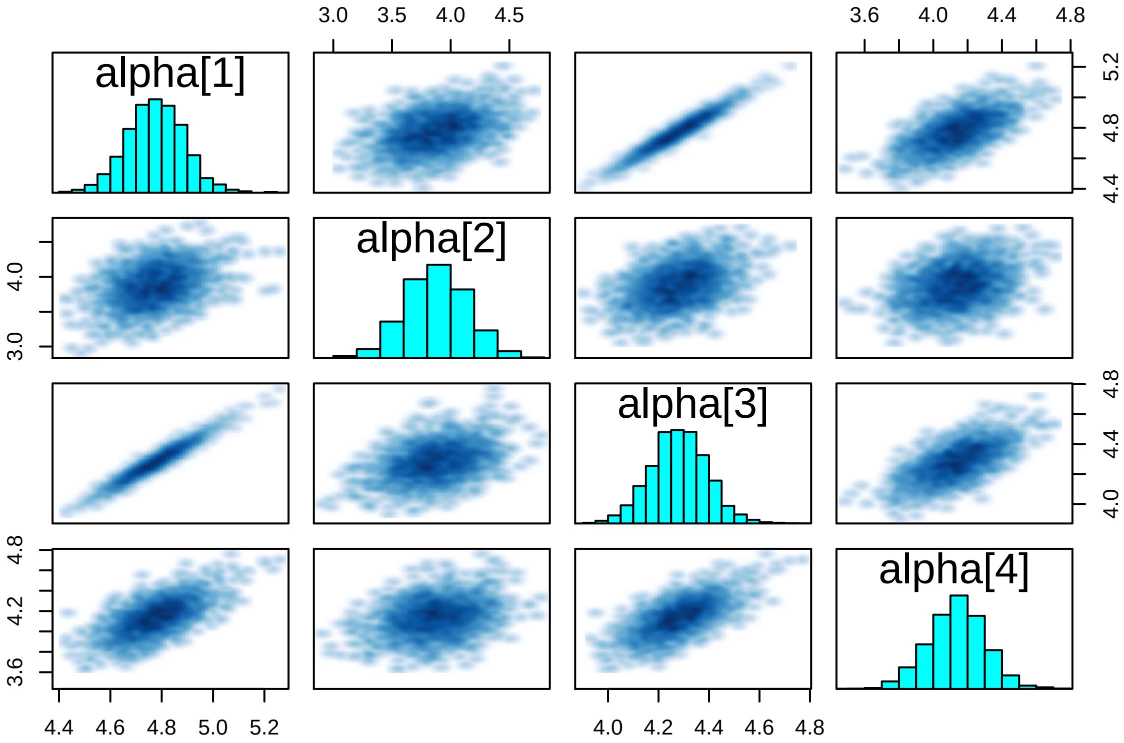
Model 2 - Districts and all Covariates

For the second model, neighbourhoods were grouped into one of the six districts in Toronto, namely, East York, Etobicoke, North York, Old City of Toronto, Scarborough, and York. A visual representation of these districts is displayed in the figure below.



After performing this grouping, the second model was fit with districts instead of neighbourhoods, again hierarchically modeling log price with fixed effects using a room type structure. In this model, all covariates have been included, similar to Model 1. Looking at the Rhat values for this model, the largest Rhat was 1.004751, with the rest of them ≈ 1, indicating good mixing of the chains. Next, we can look at the traceplot and pairs plots for the alphas.

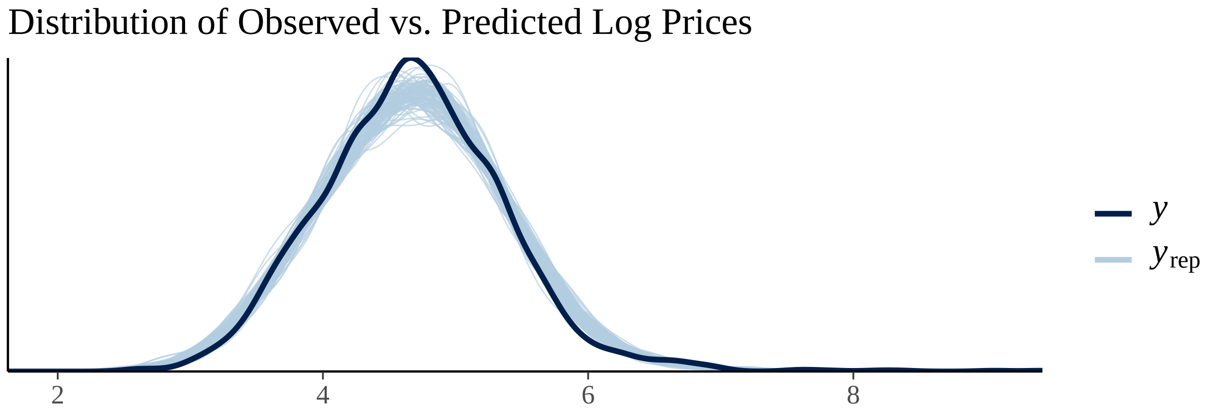


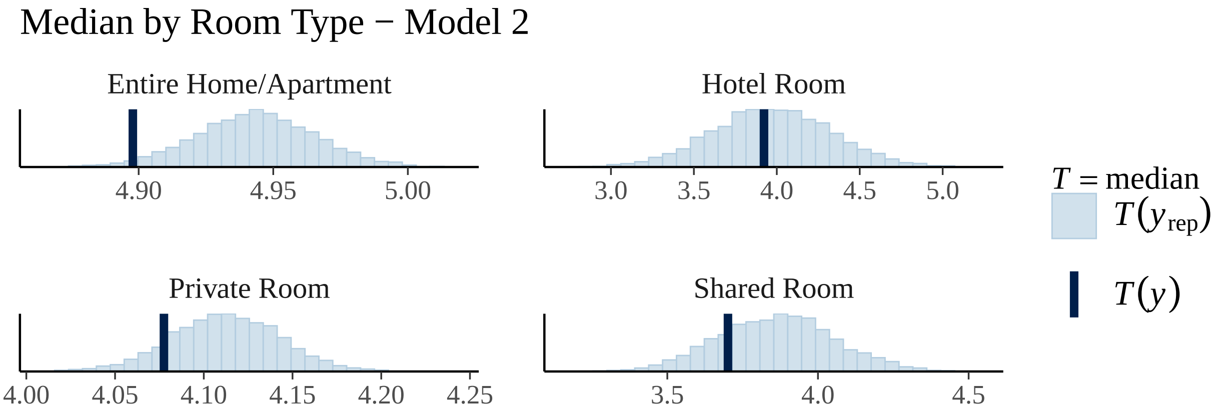


The traceplot indicated that the chains mixed well. Looking at the pairs plot, we see some ellipse shapes which raises some concern as this can indicate that we are not sampling the entire space. After checking all of convergence diagnostics, we can try and assess model fit.

Next, we can look at the overall distributions of the replicated datasets versus the data. The following figure contains a plot of the distribution of our data (y) against 100 different datasets drawn from the posterior predictive distribution on the top. We see that the observed and the predicted log prices seem to follow the same normal distribution which is good.

We can also look at some test statistics that are of interest. In the bottom panel, we see the distribution of the median (log) prices across the replicated data sets in comparison to the median in the data. We see that for entire homes/apartments, the predicted median log price is too high. It is slightly less high for private and shared rooms, and looks to be about average for the hotel rooms.

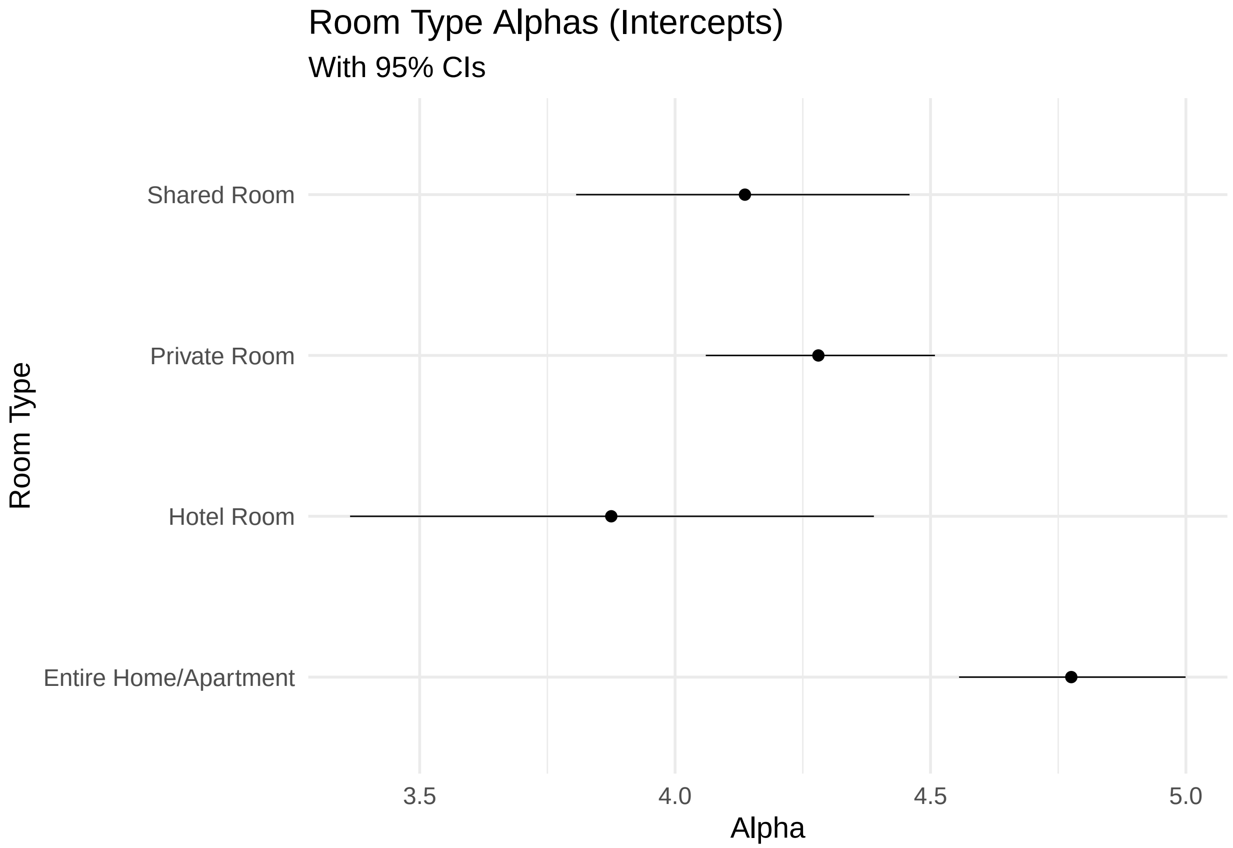




Finally, we can take a look at the mean room type alphas along with their 95% credible intervals as seen below. Note that to save space, the beta plot has been omitted, but the results will be discussed.

We see that unsurprisingly, entire homes/apartments have the largest intercept, followed by private rooms. We see that shared rooms and hotels have a much wider credible interval, indicating more uncertainty in those lower prices.

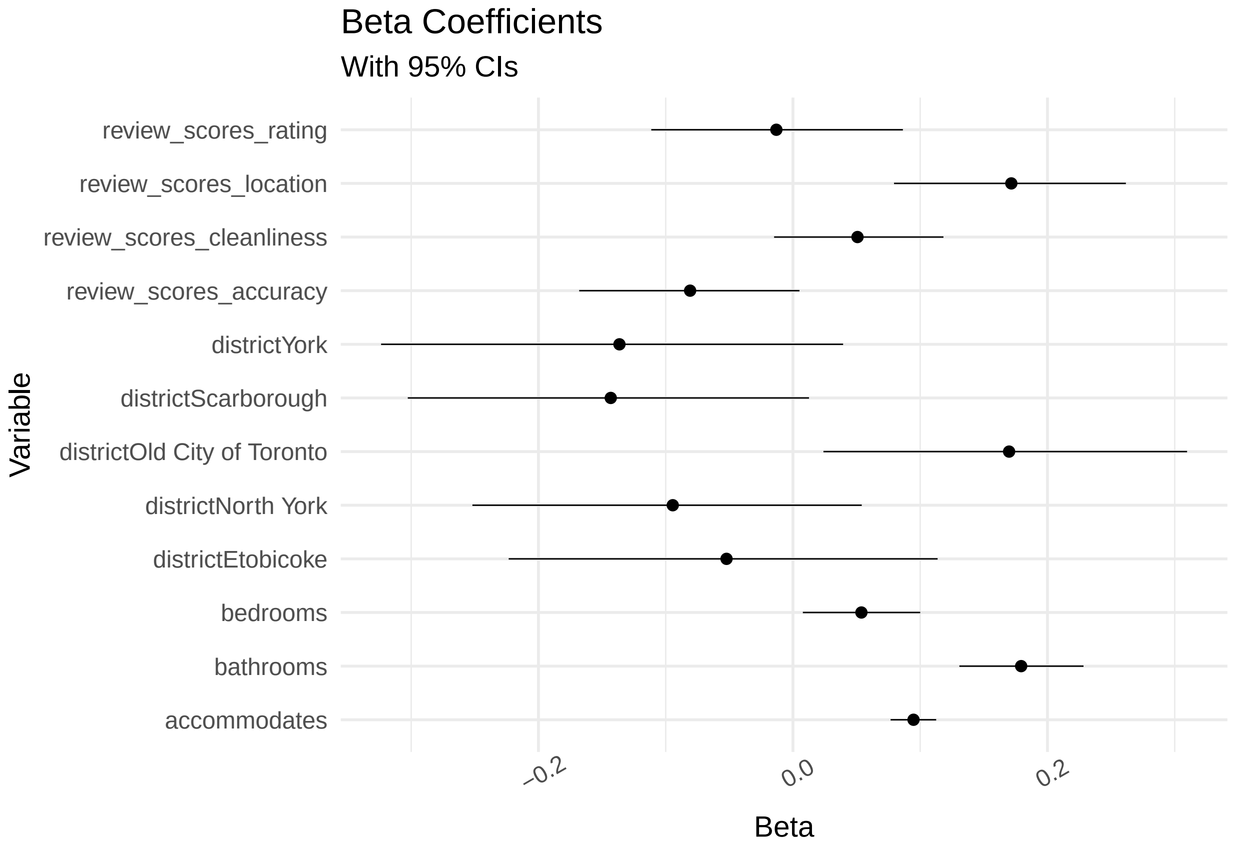
Looking at the beta coefficients, we see that the covariate with the largest coefficient is review scores location. Out of all of the covariates, the ones that do not contain include zero in their 95% CI are the following: review scores for value, overall rating, location, and communication, host response time within an hour, district of Old City of Toronto, and size variables like accommodates and number of beds/bathrooms.



Model 3 - Districts with Selected Covariates

Another model we can investigate is a similar model to Model 2, but with less covariates. In this model, we will try to reduce the covariates to contain only those that were identified in our exploratory analysis as important, namely, district, number of bedrooms, number of bathroom, how many people it accommodates, and some review score variables (overall, accuracy, cleanliness, and location). No variables pertaining to the host will be included as they did not appear to have a relationship with our variable of interest, log price. After running the model, the Rhat values looked good (maximum value of 1.002776), and the chains looked like they mixed well in the traceplot. Looking at the pairs plot, some of the ellipse-like shapes have improved, however, it is still present between alpha one and three. The PPC plots for the model looked similar to the ones from the previous model, with the predicted median log price too high for entire homes/apartments.

As we have fewer covariates in this model, we can further examine the mean beta coefficients with their respective 95% credible intervals. We see that the covariates with the highest beta coefficients are the number of bathrooms, the district of Old City of Toronto, and the location review score. The covariates with the smallest beta coefficients are the districts of York and Scarborough. This is unsurprising as Old City of Toronto contains the downtown core, whereas York and Scarborough are farther in the suburbs. Looking at the 95% CIs, they are most wide for the district variables, and most narrow for the listing size variables like accomodates and number of beds/bathrooms. We see that only the covariates accommodates, number of bedrooms and bathrooms, district of Old City of Toronto, and location are significant in this model.



Model 4 - Baseline Model (Linear Fit, all Covariates)

Finally, as a baseline model, we fit a simple linear model in STAN using all the covariates to try and model log price. Checking the model diagnostics, the chains appeared to have mixed well and all Rhat values were ≈ 1. We can now proceed to compare the models.

Compare Models

In order to compare the models, we will be examining the LOO ELPD. When comparing two fitted models, we can estimate the difference in their expected predictive accuracy by the difference in LOO ELPD. The first step of this is to get the point-wise log likelihood estimates from each model and then get estimates for the ELPD. For this analysis, we will not include Model 1 in the comparisons.

Model 2 vs. Model 3

	elpd_diff	se_diff	elpd_loo	se_elpd_loo	p_loo	se_p_loo	looic	se_looic
model1	0.00000	0.000000	-1403.767	86.26939	43.11327	6.677096	2807.534	172.5388
model2	-15.38791	7.978829	-1419.155	86.18915	28.76178	4.800808	2838.310	172.3783

Here we see that Model 2 had the higher ELPD, and thus is a better model fit.

Model 2 vs. Model 4

	elpd_diff	se_diff	elpd_loo	se_elpd_loo	p_loo	se_p_loo	looic	se_looic
model1	0.00000	0.000000	-1403.767	86.26939	43.11327	6.677096	2807.534	172.5388
model2	-15.38791	7.978829	-1419.155	86.18915	28.76178	4.800808	2838.310	172.3783

When looking at Model 2 versus Model 4, again, Model 2 had the higher ELPD, and thus is the better model.

Model 3 vs. Model 4

	elpd_diff	se_diff	elpd_loo	se_elpd_loo	p_loo	se_p_loo	looic	se_looic
model1	0	0	-1419.155	86.18915	28.76178	4.800808	2838.31	172.3783
model2	0	0	-1419.155	86.18915	28.76178	4.800808	2838.31	172.3783

Interestingly, between Model 3 and Model 4, there is no difference. Therefore, we are indifferent between the two models.

Discussion

Intuitively, we expected to see that the neighborhood the AirBnb is in, the type of room, and the variables regarding the listing itself would be the most influential on price. For example, an entire house downtown or near the water with 3+ beds will be more expensive than a shared room farther from the core. The results from the various models examined in this analysis agreed well with this hypothesis.

We first fit a hierarchical fixed effects model modeling log price, using all of the available covariates, including neighbourhoods. Including neighbourhoods in the model resulted in a model which had a low bulk effective sample size. To combat this, neighbourhoods were grouped into districts and the model was re-run (Model 2). The second model had good diagnostics, and the significant coefficients were review scores for value, overall rating, location, and communication, host response time within an hour, district of Old City of Toronto, and size variables like accommodates and number of beds/bathrooms. Motivated by the EDA, a third model was fit using a smaller subset of variables. This model also had good diagnostics, and the significant coefficients were accommodates, number of bedrooms and bathrooms, district of Old City of Toronto, and review scores for location. For both models, the intercept for entire house/apartment was highest, with hotel room lowest. Finally, a fourth model was fit using a simple linear fit for comparison purposes. After comparing the models using LOO ELPD, it was found that the optimal model was Model 2.

Future Work

As with any analysis, it is important to consider how the work could be extended upon. For future work, and interested individual could try a second model specification - a hierarchical model with mixed effects to allow for varying slopes. In this analysis, we focused on a hierarchical model with fixed effects only. It may be interesting to consider whether an additional bedroom/bathroom is worth more in different room types, for example, in a hotel room versus a private room. After constructing this model we then could compare the models using LOO-CV as we did in this analysis.

References

Airbnb. "What Is Airbnb and How Does It Work?", https://www.airbnb.ca/help/article/2503/what-is-airbnb-and-how-does-it-work?" Accessed April 15, 2022.

City of Toronto, Open Data Portal. 2022. Neighbourhoods. https://open.toronto.ca/dataset/neighbourhoods/

Government of Canada, Statistics Canada (February 9, 2022). "Census Profile, 2021 Census of Population". www12.statcan.gc.ca. Retrieved April 02, 2022.