Model of Toronto Airbnb listing prices: Does being instantly bookable impact the prices of accommodations?

Yutong Lu - 1005738356

December 19, 2021

Abstract

Airbnb is a peer-to-peer platform that provides short-term accommodations with various features to worldwide users. We are interested in studying whether one of the listing characteristics, being instantly bookable, may affect the listing prices by attempting to make a causal inference. Data used in this analysis contains the Toronto Airbnb listing information scraped on either November 6th or 7th by Inside Airbnb from the Airbnb website. The information for each listing includes features of the host, the accommodation itself and reviews of the listing. We use the propensity score matching to try to control for the confounding variables and mimic some aspects of causal inference, with a treatment of being instantly bookable. A multiple linear regression model with the treatment variable is built while controlling for the propensity score. Based on the results, being instantly bookable appears to be positively related to the listing prices. However, the treatment does not appear to be significant in the resulting model, and thus we may not infer a relationship between being instantly bookable and the listing prices. To conclude, we argue that since being instantly bookable does not necessarily relate to the price of the listing, the other baseline characteristics of the listings may be more critical in describing and predicting the Airbnb listing prices in Toronto.

Keywords: Sharing Economy, Airbnb, Pricing, Causal Inference, Propensity Score Matching.

Introduction

Nowadays, many companies have adopted business models in sharing economy, which provides a platform for people to collaboratively utilize personal inventories such as vehicles and apartments, with compensation [29]. Airbnb, an online marketplace for short-term rentals and accommodations, is a major peer-to-peer platform with over 5.6 million listings in over 220 countries and regions worldwide as of June 30, 2021 [1].

For each listing on the Airbnb website, users can find many features associated with the accommodation itself and the host of this listing. Among the various features provided, some of them may impact the pricing of the listing. According to Cheng and Jin, Airbnb users value location, amenities, and hosts as the key attributes of the Airbnb experience [6], so we may explore whether accommodations with higher-rated key attributes may have higher prices. Simultaneously on the hosts' side, Gutt and Herrmann reported that hosts who can view their ratings increase their prices compared to hosts who do not achieve the rating visibility [14], indicating that the reviews left by consumers may influence the pricing of Airbnb listings. Thus, many factors related to the user, host, and the website itself may impact the listing price.

Interestingly, if the users enter a check-in date close to the current date of accessing the website, Airbnb will prompt out a tip suggesting the "Instant Book" filter. With the "Instant Book" filter on, the shown listings are bookable without the host's approval, whereas the listings that are not instantly bookable may

take longer to book because the host still needs to review the booking request and may either confirm or disapprove the request. Based on the research by Wang and Nicolau [24], if the Airbnb listing is instantly bookable, it tends to have a lower price because the hosts may use instant booking and lower prices to attract more people to choose this listing.

As a result, the research question is whether being instantly bookable has an effect on the price of the Airbnb listings, and what other listing and host characteristics may impact the price of listings in Toronto. Based on the literature, we hypothesize that being instantly bookable may impact the listing prices and instantly bookable Airbnb listings have lower prices on average.

We are interested in attempting to make a causal inference between instant book and price using the public information of Toronto listings on the Airbnb website. Being instantly bookable may be one of the essential features of the listings when users have an immediate need for accommodation or have a sudden change in the travelling plan. By providing the option of instantly booking the listing, the host may attract more people but take risks such as accepting a visitor that booked an inconvenient date that appears to be available on the calendar. The goal of this research is not only to investigate the relationship between instant book and the listing price but also to study what other factors may be used to model the Airbnb price while controlling for the propensity of each listing to be instantly bookable or not. Furthermore, this study is important because, hopefully, we could provide the users, hosts, and the Airbnb team with a clearer idea about the aspects that might influence the listing prices and help them improve their experience and business on this platform.

In the following sections of this report, we will introduce the dataset in the Data section, including its context, collecting process, numerical and plot summaries of the important variables in the dataset. Then we will introduce our methods for model building and variable selection in the Methods section and present our results in the Results section. Finally, we will discuss the results, limitations and potential future research of our study in the Conclusions and address the ethical considerations in the Appendix. Any relative graphs and tables will be presented in the Appendix section as well.

Data

Data Collection Process

The Toronto Airbnb data is from the Inside Airbnb website, which provides datasets that are sourced from the Airbnb site and all the information is publicly available, including names, listings, and details of the review [8]. This Toronto dataset used in the analysis contains the Airbnb listings that were web scraped on either 6th or 7th November 2021, thus only reflecting the available listings' information during this time frame. Note that this dataset is directly downloaded from the "Get the data" [12] section of the website, and none of the information was web scraped by us directly. Data policies and community guidelines of Inside Airbnb site [7] were closely followed during the usage of this dataset. The data was previously cleansed and aggregated by Inside Airbnb [8].

Based on the data collection process, this dataset has some foreseeable limitations. Firstly, it contains listing information scraped from the Airbnb site on either 6th or 7th November 2021, which is close to the holiday season of the year. As a result, both the price of the listings and the availability of the coming 365 days may be affected, where the prices may be higher than other seasons, and fewer listings may be available in the near future. Due to the time-specific nature of our data, we may not be able to generalize our results to another season in the year. Also, according to the Disclaimers [8] of the website, some spam reviews are allowed by Airbnb, but Inside Airbnb suggests that the spams have little effect on the analysis. However, it is still possible for us to get biased results due to spam reviews.

Data Summary

Data Context

The Toronto Airbnb listing dataset contains observational data with 15155 observations and 74 variables in the original CSV file downloaded from the Inside Airbnb website [12]. Each observation represents a listing on either 6th or 7th November 2021 on the Airbnb website. All the listings are from different neighbourhoods in Toronto, Ontario, Canada. The variables in the dataset include listing ID, URL, other identifiers, web-scrape date and ID, information about the host of the listing, information about the accommodation, and reviews on the listing in different aspects.

Data Cleaning Process

With the original CSV file from Inside Airbnb, we first inspected the dataset and found out that values of the variable "price" had dollar signs in front of the numbers. As a result, we removed the dollar sign in the prices and changed the type of this price variable to numeric such that it correctly defined the price of the listings. Because there were many redundant variables or identifiers in the original dataset, we only selected the variable we needed in the analysis. These variables were room type, beds, price, minimum nights, number of reviews, review scores on rating, review scores on accuracy, review scores on cleanliness, review scores on check-in, review scores on communication, review scores on location, review scores on value, whether this listing is instant bookable, calculated host listings count, reviews per month, and the availability for 365 days in the future. In the dataset with newly selected variables, some observations had missing values for at least one of these variables. As a result, all the observations with missing values were moved and a complete-case dataset was created, with no missing values for any of the variables.

Then, according to our research question, we created a binary variable named "treatment" that can either take a value of one or zero. If an observation was instantly bookable, which was indicated by its value for the variable "instant bookable" being "t", then it was assigned a value of 1 for the treatment variable. On the other hand, if a not instantly bookable listing had a value "f" for the variable "instant bookable", then it was assigned 0 for the treatment variable. Because the treatment variable contained the exact same information as the "instant bookable" variable, the variable "instant bookable" was omitted in the final cleaned dataset.

After all the data cleaning processes above, we obtained a cleaned dataset with 11023 observations, 16 variables, and no missing values.

Important Variables

There are 16 variables in the cleaned dataset, and all of them are important variables that either will be the response variable or the potential predictor variables in the model of Airbnb listing price. Only one of the variables, room type, is a character variable. This variable can take one of the four types of listings, which are "entire home or apartment", "private room", "hotel room", and "shared room".

All the other variables are numerical variables, including price, beds, minimum nights, number of reviews, review scores on rating, review scores on accuracy, review scores on cleanliness, review scores on check-in, review scores on communication, review scores on location, review scores on value, calculated host listings count, reviews per month, availability in future 365 days and treatment. The variable price is the listing prices by the time of getting the data. The variable bed indicates the number of beds in the accommodation. Minimum nights mean the lowest number of nights required to book this listing. There are nine variables about the review of the listing. The number of reviews and the reviews per month are numerical counters that record the total number of reviews of a particular listing and the reviews that the listing gets per month. Accuracy, cleanliness, check-in, communication, location, and value are the star ratings provided by the guests of the accommodation, and rating indicates the overall review score of a listing. The rating scores are between 1.00 to 5.00. Calculated host listings count represents the total number of listings that a specific host has. The variable availability in 365 days indicates the number of days that a particular listing

is available in the following 365 days. Finally, the variable treatment is a binary numerical variable that represents whether a listing is instantly bookable, where an instantly bookable listing will have a value of 1 and 0 otherwise.

Based on our research question, the response variable in our analysis is listing price, the treatment is whether a listing is instantly bookable or not, and all the other variables are the potential predictors for our model.

Numerical Summaries

Table 1: Summary statistics in the Toronto Airbnb dataset with 11023 observations.

Variable	Mean (standard Deviation)	Median
Beds	1.701 (1.073)	1
Price	133.424 (104.496)	103
Minimum nights	24.875 (38.558)	28
Number of reviews	34.85 (59.01)	12
Review scores on rating	4.72 (0.447)	4.85
Review scores on accuracy	4.776 (0.43)	4.91
Review scores on cleanliness	4.678 (0.499)	4.83
Review scores on check-in	4.834 (0.381)	4.95
Review scores on communication	4.836 (0.398)	4.97
Review scores on location	4.824 (0.332)	4.93
Review scores on value	4.687 (0.461)	4.8
Calculated host listings count	4.691 (10.599)	1
Reviews per month	1.622 (4.345)	0.57
Availability in 365 days	125.879 (130.617)	84

Table 1 is a numerical summary of the mean, median and median of the numerical variables in the dataset, except for the binary numerical variable treatment. From the table, we can see that the average price of the Airbnb listings is 133 CAD, which is higher than its median 103 CAD. As a result, we may infer that there are some listings with high prices that bring the average price of the listings higher.

Table 2: Average listing price, grouped by being instantly bookable or not

Treatment (Instant book)	Number of observations	Average listing price (CAD)
0	8145	135.549
1	2878	127.410

According to Table 2, there are 8145 listings with values of 0 for the variable treatment, indicating that they are not instantly bookable. On the other hand, there are 2878 listings that are instantly bookable with values of 1. The number of non-instantly-bookable listings are more than twice the number of instantly bookable listings, which means that in the propensity score matching in the analysis, we will end up with a much smaller dataset compared to our original dataset with 11023 observations. The average listing price difference between the two groups is 8.139 CAD, where the listings that are not instantly bookable appear to have a higher mean price.

Plot Summaries

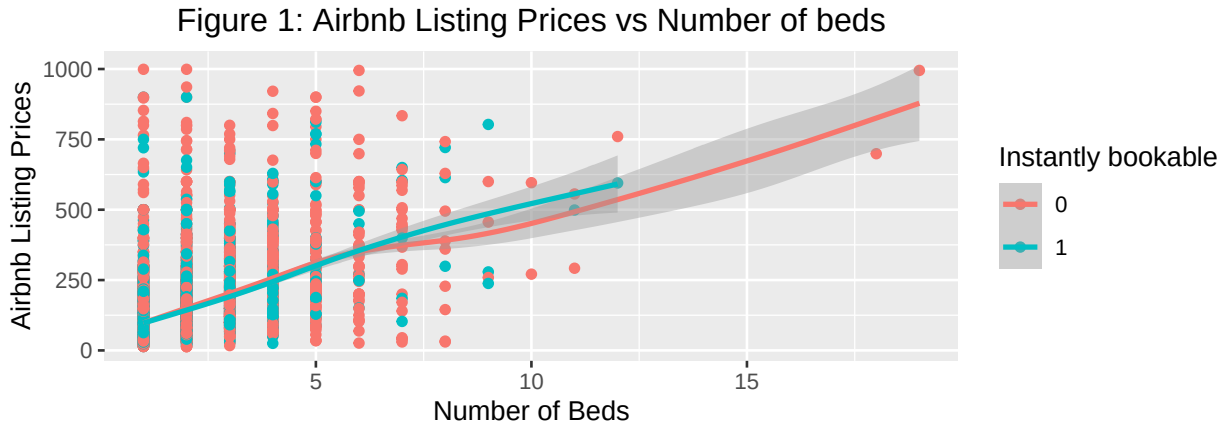

Figure 1 is a scatter plot between the prices of listings and the number of beds, where each observation is colored differently according to whether it is instantly bookable or not. Smooth lines are also fitted to the points. For the entire range of values for the numbers of beds, almost all the listings that have the highest price are not instantly bookable, and there is only one exception when the number of beds is 9. From the plot, it appears that most listings have fewer than 5 beds but there are also two listings with more than 15 beds in the accommodations,and both of them are not instantly bookable.

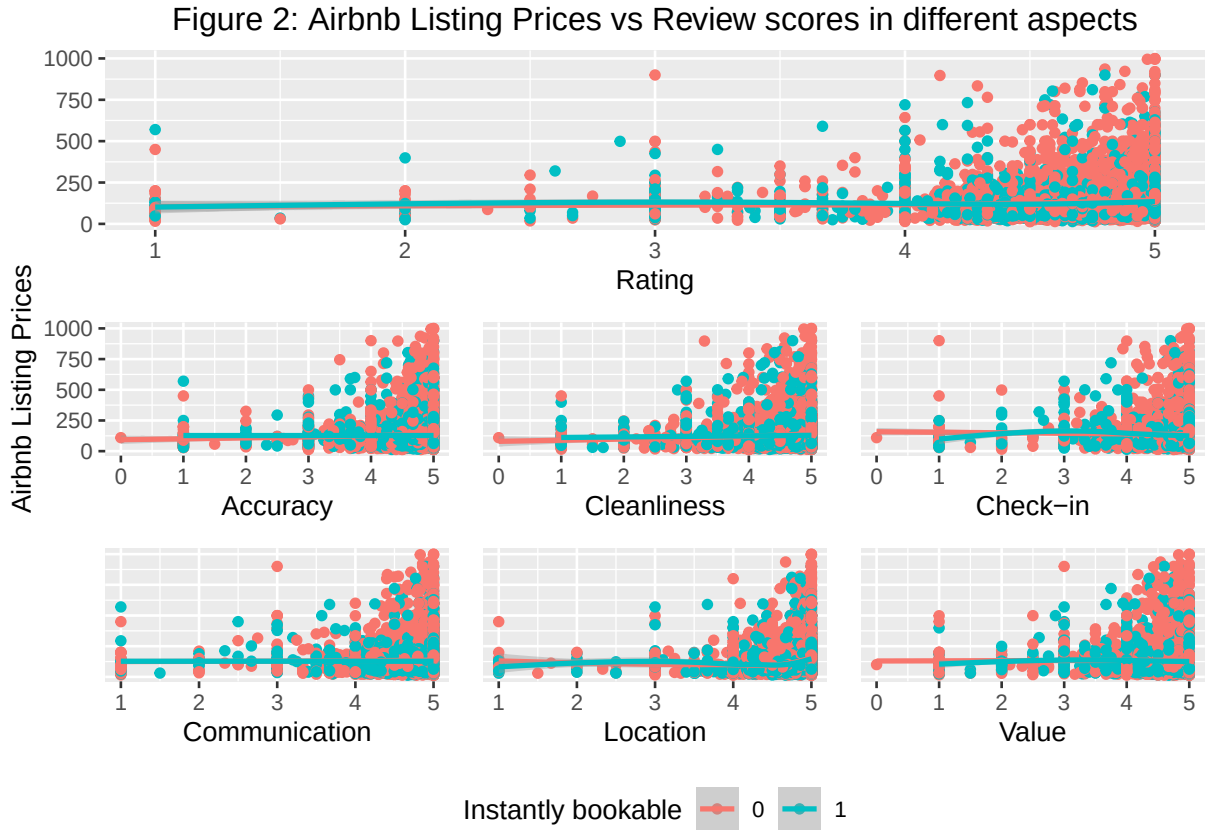

Figure 2 shows the scatter plots of review scores in different aspects, including the overall rating, accuracy, cleanliness, communication, check-in, location and value. Smooth lines are also fitted for each plot. It appears that most listings have review scores higher than 4, with most points accumulating at the right end of the plot. Also, for most aspects of the review, the listings with the highest prices are mainly not instantly bookable.

Figure 3: Airbnb Listing Prices vs Room Type

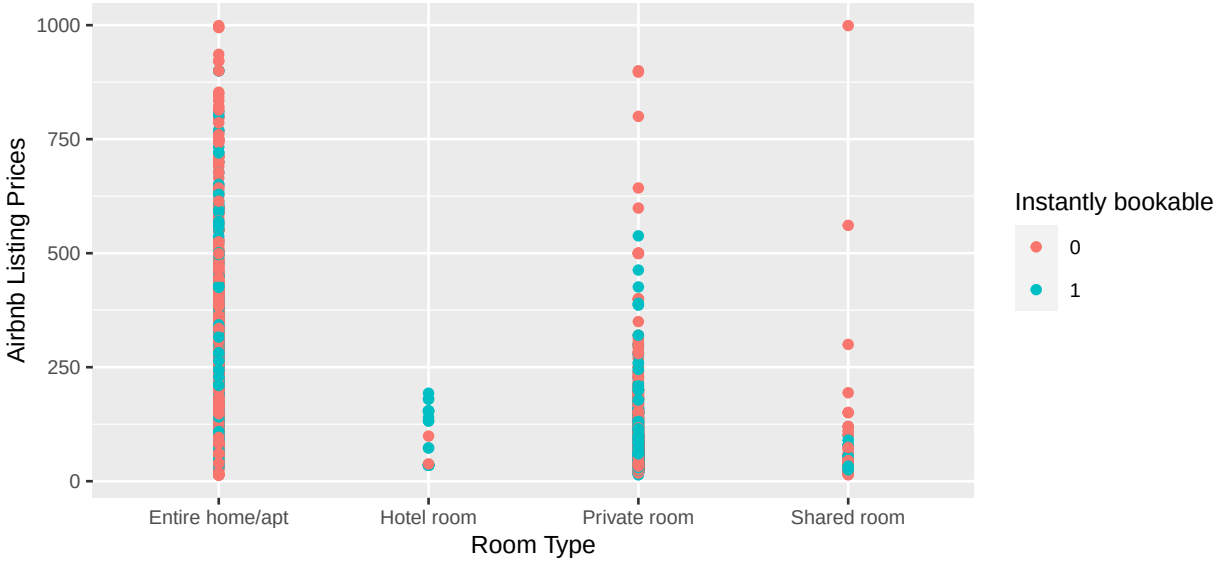

Figure 3 shows the prices of different types of the listing, including entire home or apartment, hotel room, private room and shared room. From the figure, it appears that most hotel rooms are instantly bookable and all of hotel rooms have prices lower than 250 CAD. For the other three types of rooms, the listings with higher prices are mainly not instantly bookable.

Figure 4: Airbnb Listing Prices vs Calculated Host Listings Count

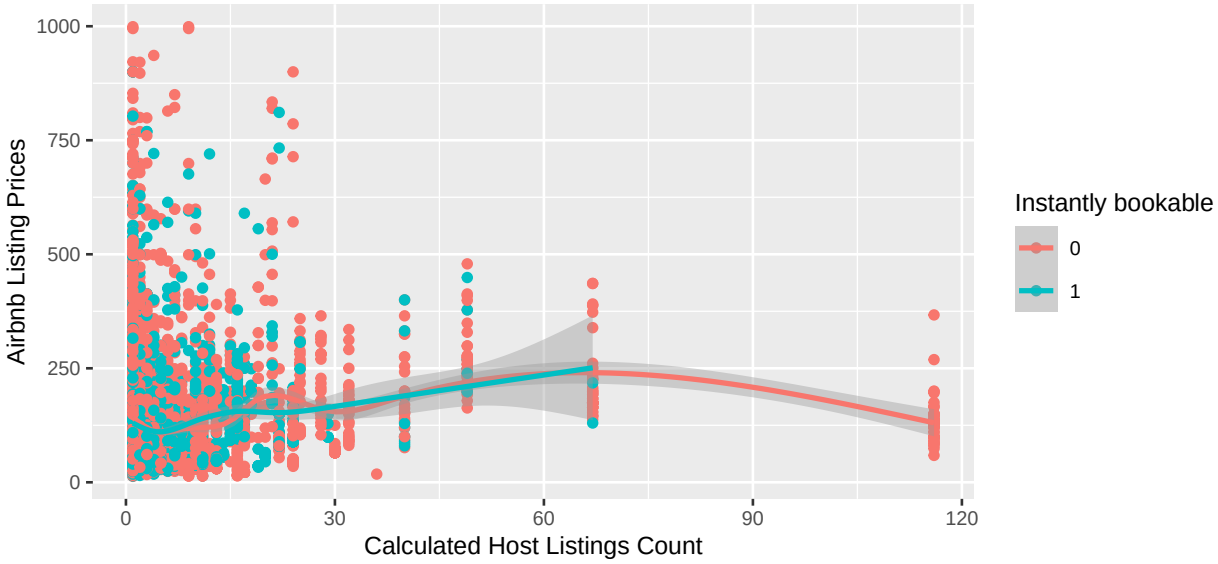

Figure 4 is a scatter plot between the price of Airbnb listings and the calculated host listings count. Two smooth lines are fitted for listings that are either instantly bookable or not instantly bookable. According to the figure, it appears that most observations with hosts that have over 30 listings in total are not instantly bookable, which may infer that more experienced hosts tend to not have their listings instantly bookable. Listings with higher prices appear to accumulate at the left end of the figure, whereasß most listings with higher calculated host listings counts have prices lower than 500 CAD.

In Data section of this report, the tables are created using the `R` package `kableExtra` [30], and the figures are created using `R` packages `tidyverse` [26], `dplyr` [27] and `patchwork` [17].

All analysis for this report was programmed using `R version 4.0.4` [19].

Methods

In this analysis, we will use propensity score matching with logistic regression and then build a multiple linear regression model to describe the Airbnb listing prices in Toronto. For the logistic regression model, we will use a bi-directional stepwise [14] Bayesian Information Criterion (BIC) [20] to select the predictors in our model. On the other hand, we will use adjusted coefficients of determination (R^2_{adj}) [4] and partial F test [28] for model selection for the multiple linear regression model.

Propensity Score Matching

We will first use propensity score matching to try to reduce the effects of confounding variables in our observational data, thus mimicking a randomized experiment to approach a causal inference to some extent [2]. The propensity score of an observation is the probability of this observation being assigned to the treatment group based on its values for some characteristics, regardless of this observation is actually treated or not [2]. Because we want to study whether being instantly bookable or not can impact the listing prices, the treatment in our analysis is being instantly bookable, and listings in the control are not instantly bookable.

Because the propensity score of each observation is a probability between zero to one by definition, we will estimate propensity score using a logistic regression model. The general format of our logistic regression model of propensity score is

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + ... + \beta_m x_m$$

$\beta_1, ..., \beta_m$ represents the coefficients of m predictor variables and β_0 is the intercept with the y-axis. Note that y represents the log odds of being in the treatment group, not the propensity score itself. Thus, for example, β_1 represents the increase in the log odds when x_1 increases by one unit while holding all other predictor values constant. Let p denote the probability of being in the treatment group, or alternatively, the propensity score of the observations, then we have

$$y = log(\frac{p}{1-p})$$

In our analysis, the response variable y in the logistic regression model is the log odds of being instantly bookable. The potential predictors in this model are all the other variables in the dataset except the response variable for the multiple linear regression model, listing prices.

Then, we will perform model selection process using a bi-directional stepwise [14] Bayesian Information Criterion (BIC) procedure [20]. Bayesian Information Criterion (BIC) is an information criterion that strives for a balance between the goodness of fit and the number of predictors [20]. It is a metric based on the maximum likelihood, which is a method for finding the parameter values that can maximize a likelihood function [25]. Assuming that the observed $x_1, ..., x_n$ are realizations of the random variables $X_1, ..., X_n$ that are discrete and independently and identically distributed, we can write the likelihood function as

$$L(\theta) = P(X_1 = x_1, X_2 = x_2, ..., X_n = x_n) = \prod_{i=1}^{n} p(x_i|\theta)$$

where $p(x_i|\theta)$ represents the probability mass function of x_i given the parameter θ. Therefore, there exists a value of the parameter θ that gives the maximum of this likelihood function, representing when the our data is most likely [25]. Alternatively, we can use the natural logarithm of the likelihood function for calculation because logarithm functions are monotonically increasing. The natural log-likelihood function is given by

$$l(\theta) = \ln(L(\theta)) = \sum_{i=1}^{n} \ln(p(x_i|\theta))$$

Bayesian Information Criterion (BIC) has a theoretical expression as

$$BIC = -2\ln(L) + k\ln(n)$$

where L denotes the likelihood function given our data, k denotes the number of parameters to be estimated in this model, and n denotes the number of the observations in the dataset [20]. This expression includes the negative natural logarithm of the likelihood function and a penalty for the number of predictors. We prefer the models with lower BIC, because we want the log-likelihood given our data to be high while not having too many predictors in our model. If we have too many predictors in our model, there is a risk of overfitting the data, which means aligning with the dataset used in our analysis too closely and may have bad performance when using any other datasets or other unobserved data [22]. Because there is a relatively heavy penalty for the number of predictors, BIC prefers simpler models with fewer predictors than some of the other information criteria.

To select predictors in the logistic regression model using BIC, we will perform a bi-directional stepwise [14] procedure using `stepAIC` function in the `MASS` `R` package [23]. We will start with a full model initially, which has a response variable of treatment and all the other variables except for price are the predictor variables. Then in the first step, we will attempt to delete one predictor and see by removing which predictor we would obtain the lowest BIC value. In the following steps, we will use the model with the lowest BIC in the first step and either add or delete a predictor at each step to arrive at a model with an even lower BIC value. The procedure will stop if there is no decrease in the BIC value or there are no more predictors to add or delete. As a result, we will obtain a final model with the lowest BIC value when starting with a full model.

We will fit this logistic regression model to our data, predict each observation's propensity score, and store the propensity score into a column named propensity score in our dataset in an ascending arrangement. Then, we perform single nearest neighbour matching using the function `matching` in the `R` package `arm` [11]. This procedure matches each observation in the treatment group with another observation that is not in the treatment group with the closest propensity score. Because more observations are not treated, or in other words, not instantly bookable, not all the non-instantly bookable observations would have a match with the instantly bookable observations. As a result, we will only keep the observations that have a match in the dataset, and filter out all the unmatched observations.

After performing the propensity score matching, we will have an equal number of observations in both the treatment and control groups, and then we can evaluate the quality of matching by examining the statistics of variables for both groups. We expect minimum differences between statistics of the observations in the treatment group and the control, and any potential problems will be addressed in the limitations.

Multiple Linear Regression

Then, we will build a multiple linear regression model for the Airbnb listing price by controlling for the observations' propensity scores. The general expression of the multiple linear regression is

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + ... + \beta_m x_m + \epsilon$$

where y represents the response variable in the model, β_0 represents the intercept of the y-axis, and $\beta_1, ..., \beta_m$ are the coefficients for $x_1, ..., x_m$ respectively. For example, β_1 represents the increase in the response variable when x_1 increases by one unit while holding all other predictors constant. Note that ϵ is a random error term

that stands for the natural random variation in linear regression. Our analysis requires the variable treatment to be in the multiple linear regression model, and the other potential predictors are selected from the variables that are not included in the logistic regression model for the propensity score. Because the bi-directional stepwise BIC [14][20] procedure does not allow us to force a variable into the model and we could lose the treatment variable in the resulted model of this procedure, we choose not to use bi-directional stepwise BIC [14][20] to select predictors for the multiple linear regression model. Instead, we use a combination of adjusted coefficient of determination [4] and partial F test [28] to select our multiple linear regression model.

We will first calculate the adjusted coefficient of determination R^2_{adj} for the potential linear regression models. The mathematical expression for coefficient of determination R^2 is

$$R^2 = 1 - \frac{RSS}{SST} = 1 - \frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{\sum_{i=1}^{n}(y_i - \bar{y})^2}$$

where RSS is the residual sum of squared that represents the amount of variation not explained by our model and SST represents the total amount of variation [28]. R^2 represents the proportion of the total variation that is explained by our model and thus is between 0 to 1. The higher the R^2, the more variation in the data can be explained by our model [5]. However, as we include more predictors in the model, the value of R^2 tends to always increase even though the added predictor is not necessarily useful [4]. As a result, we cannot use R^2 to compare two models with different number of predictors [4] and we use the adjusted coefficient of determination R^2_{adj} instead, which is given by

$$R^2_{adj} = 1 - \frac{RSS/n - p - 1}{SST/n - 1}$$

The R^2_{adj} only tends to increase by a significant amount when the added predictor really makes the model to explain a lot more variation in the data [4]. We will use `regsubsets` function in the `R` package `leaps` [15] to obtain the model of each possible size that gives the highest R^2_{adj}. We want the variable treatment to be in the model, and thus every model given by this function is the model with the highest R^2_{adj} among all the models of the same size and contains the treatment variable.

We will first fit a full model with all potential predictors and the response variable listing prices and calculate p-values for each estimated coefficient. For each 'best' model given by the `regsubsets` function [15], they have at most the number of predictors as the full model. As a result, we call the models fewer predictors than the full model as the reduced model. Let $\boldsymbol{\beta}$ denotes a vector of all the parameters in the full model, then we have $\boldsymbol{\beta} = (\beta_0, \boldsymbol{\beta}_1, \boldsymbol{\beta}_2)$, where $\boldsymbol{\beta}_1$ are the set of remaining predictors in the reduced model, $\boldsymbol{\beta}_2$ is a vector that denotes the set of predictors removed from the full model, and β_0 is the intercept. The null hypothesis and alternative hypothesis for the partial F test [28] are

$$H_0 : \boldsymbol{\beta}_2 = \mathbf{0}$$

$$H_0 : \boldsymbol{\beta}_2 \neq \mathbf{0}$$

The null hypothesis $H_0 : \boldsymbol{\beta}_2 = \mathbf{0}$ indicates that the coefficients of all the removed variables are zero, and the alternative hypothesis states otherwise. We calculate our F statistic under the null hypothesis using

$$F = \frac{RSS_{(drop)}/k}{RSS_{(full)}/n - p - 1}$$

where RSS is the residual sum of squares that represents the variation of data that is not explained by the model [3]. $RSS_{(drop)} = RSS_{(reduced)} - RSS_{(full)}$ is the residual sums of squares of the reduced value, k is the number of predictors dropped from the full model, n is the number of observations, and p is the number of predictors in the full model. The 1 represents the only intercept in the model. Based on the F statistic, we can then calculate the p-values using $p - value = P(F_{df=(k,n-p-1)} > F)$, where $F_{df=(k,n-p-1)}$ stands for a F distribution that has a degree of freedom of k and $n - p - 1$. The p-value represents the probability of observing a value as extreme or more extreme than our test statistic under the null hypothesis. In this report, we will use a significance level of $\alpha = 0.05$, and compare the p-value with this significance level. If

the p-value is larger than 0.05, then we fail to reject our null hypothesis and thus have evidence to remove these predictors from the full model and use the reduced model instead [28].

Therefore, we will perform the partial F tests using the full model and the 'best' model of each size given by the `regsubsets` function [15], and select the ones that we fail to reject the null hypothesis of partial F tests first. Then, we compare the adjusted coefficient of determination of these models and select the model with the highest R^2_{adj}. After obtaining this final model, we will fit the model to our data and examine the model summary, including the p-value for each estimated coefficient.

The p-value is calculated from a hypothesis test based on the Student's T-test [21], which has test statistic with the expression

$$T = \frac{\hat{\beta}_i - \beta_i}{\sqrt{var[\hat{\beta}_i]}}$$

where β_i denotes the i-th β parameter in our model. The null hypothesis and alternative hypothesis for the T-test are

$$H_0 : \beta_i = 0$$
$$H_a : \beta_i \neq 0$$

The null hypothesis means that this parameter is zero whereas the alternative hypothesis states that this parameter is not zero. We calculate the test statistic under the null hypothesis and calculate the p-value using the formula $p - value = P(|t_{df=n-p-1}| > T)$ where $t_{df=n-(p+1)}$ represents a T distribution with a degree of freedom of the number of observations n minus the number of predictors and intercept $p + 1$. If the p-value is smaller than 0.05, then we will have evidence to reject the null hypothesis and thus we can say that the predictor corresponding to this coefficient is statistically significant in the presence of all other predictors.

Based on our research question, we want to see if there is a relationship between being instantly bookable and listing prices. Therefore, if in the final model the p-value for the coefficient of the treatment variable is higher than 0.05, then we fail to reject the null hypothesis that the coefficient for the treatment predictor is zero, and thus there is no relationship between the listing price and the treatment in the presence of all other predictors. On the other hand, if the p-value of the coefficient for the treatment variable is smaller than 0.05, then we may reject the null hypothesis and have evidence that there is a statistically significant relationship between the treatment, or being instantly bookable, and the response variable price in the presence of other predictors. Simultaneously, because we have controlled the propensity score of being in the treatment group, we may also infer that being instantly bookable may have a causal effect on the listing price. In this way, we may answer our research question about the relationship between instant book and prices, as well as the other listing or host characteristics that could impact the listing prices.

In this Methods section, model construction and variable selection use functions from `R` packages `arm` [11], `leaps` [15], `MASS` [23], `tidyverse` [26] and `dplyr` [27].

Results

Table 3: Logistic regression model for treatment

	Estimated coefficient	Standard error	p-value
Intercept	0.5428270	0.2168749	0.0123163
Room type: Hotel room	3.1454015	0.6154323	0.0000003
Room type: Private room	0.3105064	0.0466302	0.0000000
Room type: Shared room	0.0080935	0.2816070	0.9770717
Minimum nights	-0.0029319	0.0008296	0.0004089
Review scores in rating	-0.3570966	0.0453373	0.0000000
Reviews per month	0.0343187	0.0054988	0.0000000

We started with a logistic model with the response variable treatment and all the other variables in the dataset except for price and performed a bi-directional stepwise [14] Bayesian Information Criterion (BIC) [20] procedure. The resulted model summary is shown in Table 3, which contains an intercept, and predictor variables room type, minimum nights, review scores in rating and reviews per month. The BIC value of this model is 1.2498218×10^4, which is 74.385142 lower than the initial full model.

The mathematical model of this logistic regression is given by

$$\begin{aligned}\hat{y} = 0.5428270 + 3.1454015x_{\text{room type: hotel room}} + 0.3105064x_{\text{room type: private room}} \\ +0.0080935x_{\text{room type: shared room}} - 0.0029319x_{\text{minimum nights}} \\ -0.3570966x_{\text{Review scores in rating}} + 0.0343187x_{\text{Reviews per month}}\end{aligned}$$

Note that the variable room type is a character variable with values "entire home or apartment", "hotel room", "private room" and "shared room". Based on alphabetical order, the baseline category is "entire home or apartment", and thus, it is not displayed in the model summary. Thus, when we interpret the coefficients, we would make a comparison with the baseline category. For example, the estimated coefficient of room type hotel room is 3.1454015, which means that for an Airbnb listing that is a hotel room, the expected log odds of being in the treatment group, or being instantly bookable, is 3.1454015 higher than the expected log odds of a listing that is an entire room or apartment, when holding all the other variables constant. Also, notice that with a significance level of 0.05, the only non-significant p-value is for the coefficient of the shared room category of the variable room type. Because all the other room type categories have significant p-values, we are comfortable with keeping the variable room type in the model.

The estimated intercept of this model is 0.5428270, but it does not make sense to interpret this intercept because it would infer that there exists a minimum night requirement of zero, which is not possible considering that the Airbnb bookings are at least for one night. The variables minimum nights, review scores in rating and reviews per month are all numerical variables. Thus, for example, we can interpret the estimated coefficient of review scores in rating as for every one-unit increase in the review scores in rating, the expected log odds of an entire home or apartment for being in the treatment group of instant book will decrease by 0.3570966, when holding all other predictors constant.

Table 4: Summary statistics in the treatment and control groups, each with 2878 observations.

Variable	Mean (Standard Deviation) of the treatment group	Mean (Standard Deviation) of the control
Beds	1.635 (1.01)	1.654 (1.02)
Price	127.41 (97.879)	127.97 (102.942)
Minimum nights	21.956 (33.815)	22.418 (32.552)
Number of reviews	37.286 (64.575)	36.427 (61.886)
Review scores on rating	4.659 (0.527)	4.641 (0.577)
Review scores on accuracy	4.726 (0.504)	4.715 (0.536)
Review scores on cleanliness	4.622 (0.581)	4.608 (0.597)
Review scores on check-in	4.786 (0.462)	4.796 (0.477)
Review scores on communication	4.788 (0.482)	4.787 (0.512)
Review scores on location	4.789 (0.407)	4.786 (0.397)
Review scores on value	4.633 (0.54)	4.62 (0.576)
Calculated host listings count	4.535 (6.577)	4.596 (10.43)
Reviews per month	2.124 (4.812)	2.092 (6.503)
Availability in 365 days	124.494 (132.083)	131.553 (132.562)

With the resulted logistic regression model, we predicted the propensity score of each observation in the dataset and matched each instantly bookable listing with another non-instantly-bookable listing with the

closest propensity score of being in the treatment group. After matching, both the treatment group and the control have 2878 listings, and the summary statistics of their variables are shown in Table 4.

In Table 4, no large discrepancies can be spotted in the mean and standard deviation of the treatment and control group. Among all the variables, only the average availability in 365 days appears to be slightly more different in the two groups, with 124.494 in the treatment group and 131.553 in the control. However, the standard deviation of this variable in both groups also appear to be quite large, with 132.083 in the treatment group and 132.562 in the control group. The difference in the two means of the availability are 7.059, which is within one standard deviation of both groups. Therefore, the mean and standard statistics may be considered as similar between the treatment and control groups, and we may infer that the quality of matching is satisfactory.

Table 5: Highest adjusted R squared for models of each size

Number of predictors	Adjusted R squared
1	-0.0001660
2	0.2855060
3	0.2943663
4	0.3076886
5	0.3135286
6	0.3176182
7	0.3204503
8	0.3222319
9	0.3229720
10	0.3228694
11	0.3228882

Note:
Note that the highest adjusted R squared is highlighted in red.

With matched observations in the dataset, Table 5 shows the the multiple linear regression models with the highest R^2_{adj} of each possible model size. The specific predictors in the models with size 2 to 11 are shown in Appendix Table 8, and the model with size 1 only has the predictor treatment. Notice that when the linear regression model only has the variable treatment, the adjusted R squared drops below zero at -0.0001660, and the addition of even one predictor increases the adjusted R squared to 0.2855060. This infers that the treatment variable alone may not explain the listing prices. The highest adjusted coefficient of determination is 0.3229720 for the model with 9 predictors, which is highlighted in red.

According to the partial F test between this model with 9 predictors and the full model, the p-value is 0.5249, which is greater than our significance level of 0.05. As a result, we do not have evidence against the null hypothesis and can opt for this smaller model with 9 predictors instead. This multiple linear regression model has a response variable of price, and the predictor variables are treatment, beds, review scores in accuracy, review scores in cleanliness, review scores in check-in, review scores in location, review scores in value, calculated host listings count and availability in future 365 days.

Table 6: Multiple linear regression model for Airbnb listing prices

	Estimated Coefficient	Standard Error	p-Value
Intercept	-58.0005847	14.1128515	0.0000402
Treatment	0.1807846	2.1805258	0.9339270
Beds	51.7783270	1.0765752	0.0000000
Review scores in accuracy	11.5156938	4.2674275	0.0069854
Review scores in cleanliness	15.7855351	2.9795896	0.0000001
Review scores in check-in	-34.3684775	3.7015647	0.0000000
Review scores in location	45.0332253	3.7821306	0.0000000
Review scores in value	-18.8442494	3.9458581	0.0000018
Calculated host listings count	0.6904710	0.1279189	0.0000001
Availability in the future 365 days	0.0496467	0.0083689	0.0000000

Note:
The p-value for the coefficient corresponding to the treatment variable is hightlighted in red.

The mathematical model of this multiple linear regression model is given by

$$\begin{aligned}\hat{y} = -58.0005847 &+ 0.1807846x_{\text{treatment: being instantly bookable}} + 51.7783270x_{\text{beds}} \\ &+11.5156938x_{\text{Review scores in accuracy}} + 15.7855351x_{\text{Review scores in cleanliness}} \\ &-34.3684775x_{\text{Review scores in check-in}} + 45.0332253x_{\text{Review scores in location}} \\ &-18.8442494x_{\text{Review scores in value}} + 0.6904710x_{\text{Calculated host listings count}} \\ &+0.0496467x_{\text{Availability in the future 365 days}}\end{aligned}$$

Table 6 shows the model summary of this chosen multiple linear regression model. For example, the estimated coefficient of review scores in location is 45.0332253, and we can interpret this as for every one unit increase in the review scores in location for a non-instantly bookable listing, the expected listing price will increase by approximately 45.03 CAD, holding all other predictors constant. Note that the estimated coefficient for treatment is 0.1807846, and given the fact that it is a binary categorical variable, we can interpret that the average price of instantly bookable listings is about 0.18 CAD higher than listings that are not instantly bookable when all the other variables are constant. However, it appears that with a significance level of 0.05, the only insignificant predictor in this model is the treatment variable with a p-value of 0.9339270, which is highlighted in red in Table 6. As a result, we may not be able to say that the variable treatment is significant and there appears to be no relationship between being instantly bookable and the listing prices.

Figure 5: Added−variable plots for the multiple linear regression model

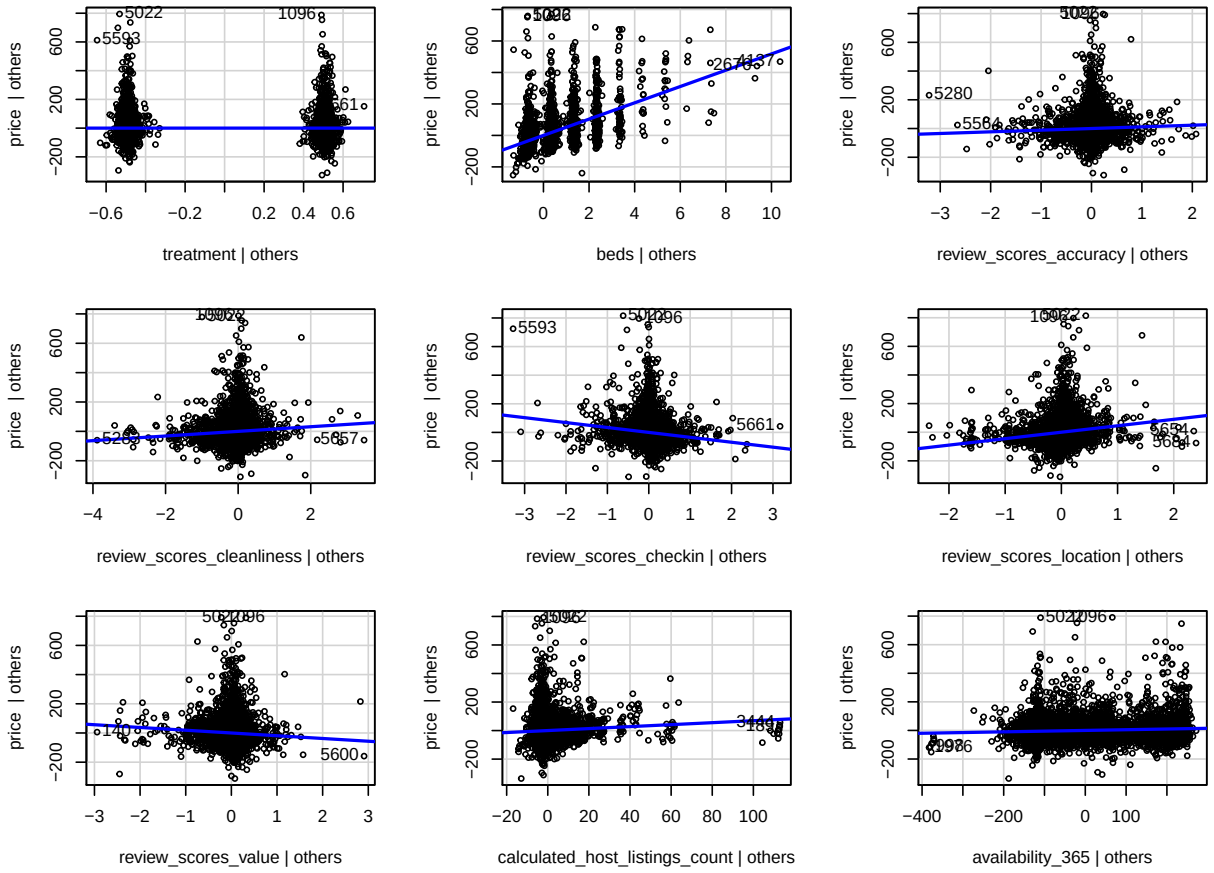

Figure 5 shows the added-variable plots [9][10] of the multiple linear regression model, which demonstrates the correlation between each predictor and the response variable when conditioning on all the other predictor variables [10]. Note that the slope of the blue regression line fitted in each plot has a slope that is equal to the estimated coefficient of each specific predictor in the multiple linear regression model [10]. In Figure 5, it appears that the blue regression fit of the variable treatment is approximately a horizontal line at zero, indicating that there may be no relationship between the treatment of being instantly bookable and the listing prices. This figure is appropriate because it is consistent with the insignificant p-value for the T-test of the coefficient associated with the treatment variable. We hypothesized that being instantly bookable may affect listings prices, and instantly bookable listings have lower prices on average than listings that are not instantly bookable. According to our results, we have evidence to reject our hypothesis.

All analysis for this report was programmed using `R version 4.0.4` [19].

Conclusions

Characteristics of both host and accommodation information for listings may impact the Airbnb listing pricing in Toronto. In this analysis, our research question is whether being instantly bookable may affect the price of Airbnb listings in Toronto and what other listing and host characteristics may relate to the listing prices. We were interested in getting close to a causal inference and we hypothesized that the instant book feature might impact listing price, and instantly bookable listings tend to have lower prices on average.

Data for this analysis was scraped on November 6th or 7th from the Airbnb website by Inside Airbnb [12]. To control for confounding variables and mimic some aspects of a randomized control experiment [2], we

used propensity score matching, and the treatment is being instantly bookable. A logistic regression model was selected using bi-directional stepwise [14] Bayesian Information Criterion procedure [20]. The response variable of this logistic regression model was treatment, and the potential predictors may be all variables in the dataset other than price. We obtained a logistic regression model with four predictors and fit our data to this model to measure the propensity of each observation in the dataset to be in the treatment group. We matched the observations based on their propensity scores using the single nearest neighbour approach [11]. We evaluated the quality of matching by comparing the statistics of the treatment and control groups. After obtaining the matched data, we selected the multiple linear regression model using a combination of adjusted coefficient of determination [4] and partial F test [28] while keeping the variable treatment in the model.

In the results, we obtained a multiple linear regression model with nine predictors in total, and none of the predictors were in the preceding logistic regression model. The variable treatment had a positive estimated coefficient, meaning that based on our results, instantly bookable listings could have higher prices on average. However, in this multiple linear regression model, only the treatment variable appeared insignificant using a significance level of 0.05. As a result, we had evidence against our hypothesis that a relationship exists between being instantly bookable and listing prices. This result made sense according to the analysis by Wang and Nicolau [24], where they reported p-values higher than 0.05 for the significant differences in three out of four quantiles of the variable "instant bookable". In terms of the listing characteristics, it appeared that the number of beds and the availability in the future 365 days were positively related to the price of the listing. The reviews left by past visitors may affect the listing price too, where higher review scores in accuracy, cleanliness and location may result in higher listing prices, but a higher-rated check-in experience and value of the listing may lead to lower listing prices. Also, experienced hosts with more listings in total tend to mark their listings at a higher price.

Weaknesses

There are some limitations to the study. As mentioned in the data collection section of this report, the listing information was scraped from the Airbnb website on either November 6th or 7th, 2021 [12], and we treated the two days as the same period. However, there may be differences in the website listings on these two days, such as deleted or added listings, price changes and review score changes. We may be missing some updates for the listings scraped on November 6th as it does not appear in the original dataset that the listing information scraped on the 6th was re-scraped on the 7th. As a result, this may lead to potential biases in our resulted models. Also, because the data was scraped at a date close to the holiday season, it is possible that our results may only generalize to this specific time frame and cannot describe the listing characteristics for other seasons in the year.

Another potential problem mentioned in the data collection is that there may be spam reviews for the listings on the website [8]. There is no indicator for potential spam reviews in the scraped data, so we could not filter out the spam reviews for the listings. As a result, the review scores may appear to be higher or lower due to the spam reviews, leading to potentially biased results. Also, we created a complete-case dataset with no missing values when cleaning the data, but we could not demonstrate that the missing values in the original scraped data were random. Therefore, we may have eliminated some new listings with not enough reviews to have a displayed rating or hosts with specific characteristics that may affect whether they set their listings to instantly bookable or not. Again, this may lead to biased results in our models.

There may be weaknesses in our methodologies used as well. Firstly, the propensity score matching methods is still based on observational data instead of randomized control trials [2]. We may not have access to all the confounding variables [16], and thus our results may still be biased. Therefore, we cannot make a true causal statement even though our results show that being instantly bookable or not does not affect the Airbnb listing prices. Another limitation in our method is that we relied heavily on logistic and linear regression assumptions when building the models in this analysis. Therefore, our model may perform better for some ranges of predictor value than others, and the results may be biased.

Next Steps

In future studies, we may change the model selection method for the logistic regression model of the treatment variable. We used Bayesian Information Criterion (BIC) in our analysis, which has a relatively heavier penalty for model complexity [20] than other information criteria such as the Akaike information criterion (AIC) [20]. This led us to have a relatively simpler logistic regression model with fewer predictors. As a result, we could use AIC to choose predictors in the logistic regression model such that we may end up with a model with more predictors, which in turn will leave us with fewer predictors for the final linear regression model.

The next steps may also include investigating prices during another season that is not close to the holiday. Alternatively, time series analysis [18] may be performed on the listings and observe the changes in their prices and other baseline characteristics over a period of time. It is also interesting to investigate the changes in the Airbnb listings brought by the COVID-19 global pandemic.

Discussion

To conclude, whether being instantly bookable does not appear to causally relate to Airbnb listing prices based on our results. As a result, the filter for instantly bookable listings may not be suggesting the users with urgent needs for accommodations with listings of a higher price. However, because instantly bookable listings may be booked without host approval, there may be problems when the listing is not actually available, but the host does not label it as unavailable on the website, resulting in cancellations and refunds that could have been prevented. Also, the process of a visitor requesting and host approval may provide both parties with a chance to be informed of any potential issues and consider carefully before making a decision, thus preventing potential inconvenience in the future. On the bright side, instant booking shortens the confirmation process, thus bringing both the visitors and hosts a more efficient experience in the peer-to-peer market. As a result, since the instant book does not necessarily relate to listing prices, visitors who do not require immediate accommodations may consider other aspects of the listings and the hosts, such as review scores, future availability and host experience, to filter and select the best listings for them. The hosts may label the listings as instantly bookable to attract more visitors while keeping a close record of future availability. Eventually, whether being instantly bookable or not, customers, hosts and the Airbnb website hopefully may work together to create a peer-to-peer platform that supports user-host communication that is convenient and efficient at the same time.

Bibliography

1. *About us.* (June 30, 2021). Airbnb. https://news.airbnb.com/about-us/. (Last Accessed: December 10, 2021)

2. Austin, P.C. (2011). An Introduction to Propensity Score Methods for Reducing the Effects of Confounding in Observational Studies. *Multivariate Behavioral Research, 46*(3), 399–424. https://doi.org/10.1080/00273171.2011.568786.

3. Barone, A. (2021, December 1). *Residual Sum of Squares (RSS).* Investopedia. https://www.investopedia.com/terms/r/residual-sum-of-squares.asp. (Last Accessed: December 10, 2021)

4. Bhalla, D. (n.d.). *Difference between adjusted R-squared and R-squared.* ListenData. https://www.listendata.com/2014/08/adjusted-r-squared.html. (Last Accessed: December 10, 2021)

5. Bloomenthal, A. (2021, October 10). *Coefficient of Determination.* Investopedia. https://www.investopedia.com/terms/c/coefficient-of-determination.asp. (Last Accessed: December 10, 2021)

6. Cheng, M., & Jin, X. (2019). What do Airbnb users care about? An analysis of online review comments. *International Journal of Hospitality Management, 76*, 58–70. https://doi.org/10.1016/j.ijhm.2018.04.004.

7. *Data Policies.* (n.d.). Inside Airbnb. http://insideairbnb.com/data-policies.html. (Last Accessed: December 10, 2021)

8. *Disclaimers.* (n.d.). Inside Airbnb. http://insideairbnb.com/about.html#disclaimers. (Last Accessed: December 10, 2021)

9. Fox, J., & Weisberg, S. (2019). *An {R} Companion to Applied Regression, Third Edition.* Thousand Oaks CA: Sage. https://socialsciences.mcmaster.ca/jfox/Books/Companion/. (Last Accessed: December 15, 2021)

10. Gallup, J.L. (2019). Added-variable plots with confidence intervals. *The Stata Journal, 19*(3), 598–614. https://doi.org/10.1177/1536867X19874236.

11. Gelman, A., & Su, Y.S. (2021). *arm: Data Analysis Using Regression and Multilevel/Hierarchical Models.* R package version 1.12-2. https://CRAN.R-project.org/package=arm. (Last Accessed: December 15, 2021)

12. *Get the data.* (n.d.). Inside Airbnb. http://insideairbnb.com/get-the-data.html. (Last Accessed: November 30, 2021)

13. Gutt, D., & Herrmann, P. (2015). Sharing Means Caring? Hosts' Price Reaction to Rating Visibility. *ECIS 2015 Research-in-Progress Papers.* Paper 54. ISBN 978-3-00-050284-2. https://aisel.aisnet.org/ecis2015_rip/54.

14. Hayes, A. (2021, November 29). *Stepwise regression.* Investopedia. https://www.investopedia.com/terms/s/stepwise-regression.asp. (Last Accessed: December 10, 2021)

15. Lumley, T., & Miller, A. (2020). *leaps: Regression Subset Selection.* R package version 3.1. https://CRAN.R-project.org/package=leaps. (Last Accessed: December 14, 2021)

16. Nuttall, G.A., & Houle, T.T. (2008). Liars, damn liars, and propensity scores. *Anesthesiology (Philadelphia), 108*(1), 3–4. https://doi.org/10.1097/01.anes.0000296718.35703.20.

17. Pedersen, T.L. (2020). *patchwork: The Composer of Plots.* https://patchwork.data-imaginist.com, https://github.com/thomasp85/patchwork. (Last Accessed: December 12, 2021)

18. Rao, S.S. (2021, January 21). *A course in time series analysis.* Texas A&M University Statistics. https://web.stat.tamu.edu/~suhasini/teaching673/time_series.pdf. (Last Accessed: December 12, 2021)

19. R Core Team (2021). *R: A language and environment for statistical computing.* R Foundation for Statistical Computing, Vienna, Austria. http://www.R-project.org/. (Last Accessed: December 10, 2021)

20. Rossi, R., Murari, A., Gaudio, P., & Gelfusa, M. (2020) Upgrading Model Selection Criteria with Goodness of Fit Tests for Practical Applications. *Entropy, 22*(4), 447. https://doi.org/10.3390/e22040447.

21. Student. (1908). The probable error of a mean. *Biometrika*, 1–25.

22. Twin, A. (2021, October 22). *Overfitting.* Investopedia. https://www.investopedia.com/terms/o/overfitting.asp#:~:text=Overfitting%20is%20a%20modeling%20error,limited%20set%20of%20data%20points.&text=Thus%2C%20attempting%20to%20make%20the,and%20reduce%20its%20predictive%20power. (Last Accessed: December 10, 2021)

23. Venables, W.N. & Ripley, B.D. (2002). *Modern Applied Statistics with S. Fourth Edition.* Springer, New York. ISBN 0-387-95457-0.

24. Wang, D., & Nicolau, J.L. (2017). Price determinants of sharing economy based accommodation rental: A study of listings from 33 cities on Airbnb.com. *International Journal of Hospitality Management, 62*, 120–131. https://doi.org/10.1016/j.ijhm.2016.12.007.

25. Weisstein, E.W. (n.d.). *Maximum Likelihood.* Wolfram MathWorld. https://mathworld.wolfram.com/Maximum26.html. (Last Accessed: December 7, 2021)

26. Wickham et al., (2019). Welcome to the tidyverse. *Journal of Open Source Software*, *4*(43), 1686. https://doi.org/10.21105/joss.01686. (Last Accessed: December 11, 2021)

27. Wickham, H., François, R., Henry, L. & Müller, K. (2021). *dplyr: A Grammar of Data Manipulation.* https://dplyr.tidyverse.org, https://github.com/tidyverse/dplyr. (Last Accessed: December 11, 2021)

28. Zach. (2020, December 6). *What is a partial F-test?* Statology. https://www.statology.org/partial-f-test/. (Last Accessed: December 11, 2021)

29. Zervas, G., Proserpio, D., & Byers, J.W. (2017). The Rise of the Sharing Economy: Estimating the Impact of Airbnb on the Hotel Industry. *Journal of Marketing Research, 54*(5), 687–705. https://doi.org/10.1509/jmr.15.0204.

30. Zhu, H. (2021). *kableExtra: construct complex table with 'kable' and pipe syntax.* R package version 1.3.4. https://CRAN.R-project.org/package=kableExtra. (Last Accessed: December 11, 2021)

Appendix

A1: Ethics Statement

All the analyses in this study were performed with ethical considerations. Firstly, the data policies of our data source website Inside Airbnb were closely followed, where we did not scrape data from the site by ourselves, and we did not republic the data. All the data, including host and listing information, were publicly available on the Airbnb website. For privacy concerns, Airbnb anonymized the location of listings, and thus we did not have the exact location of each listing [8], and no location information was used in this analysis. The identifiers of the hosts were removed from the original dataset downloaded, including but not limited to host name, description, verification, and neighbourhood. For the reviews of listings, no exact comments or any reviewers' information was included in the dataset. We only obtained the number of reviews and the star reviews that are publicly displayed for each listing. Also, publication bias was explicitly avoided by presenting the insignificant treatment predictor in the final model. For reproducibility, all the decisions made in this analysis were documented. All resources and information external to the course STA304 of the University of Toronto were properly cited in the Bibliography section.

A2: Materials

Here are the first 6 observations of the cleaned dataset with 11023 observations, 16 variables and no missing values.

Table 7: The first 6 observations in the cleaned dataset

	1	2	3	4	6	7
Room type	Entire home/apt	Private room	Private room	Entire home/apt	Entire home/apt	Entire home/apt
Beds	7	1	1	1	2	1
Price	469	93	72	45	100	70
Minimum nights	28	180	28	365	30	28
Number of reviews	7	169	217	26	113	85
review scores in rating	5.00	4.84	4.75	4.92	4.63	4.71
Review scores in accuracy	5.00	4.81	4.73	5.00	4.64	4.88
Review scores in cleanliness	5.00	4.89	4.82	4.82	4.66	4.72
Review scores in check-in	5.00	4.87	4.90	5.00	4.95	4.88
Review scores in communication	5.00	4.90	4.93	5.00	4.96	4.94
Review scores in location	5.00	4.92	4.34	4.82	4.56	4.60
Review scores in value	5.00	4.83	4.73	4.82	4.68	4.80
Instantly bookable	f	t	t	f	f	f
Calculated host listings count	1	2	2	1	4	2
Reviews per month	0.09	1.51	1.74	0.21	0.90	1.51
Availability in 365 days	0	365	365	251	282	310

Note:
The column named 5 is missing because the fifth observation is removed from the original dataset due to its missing review scores in accuracy, cleanliness, check-in, communication, location and value.

Supplementary Plots and Tables

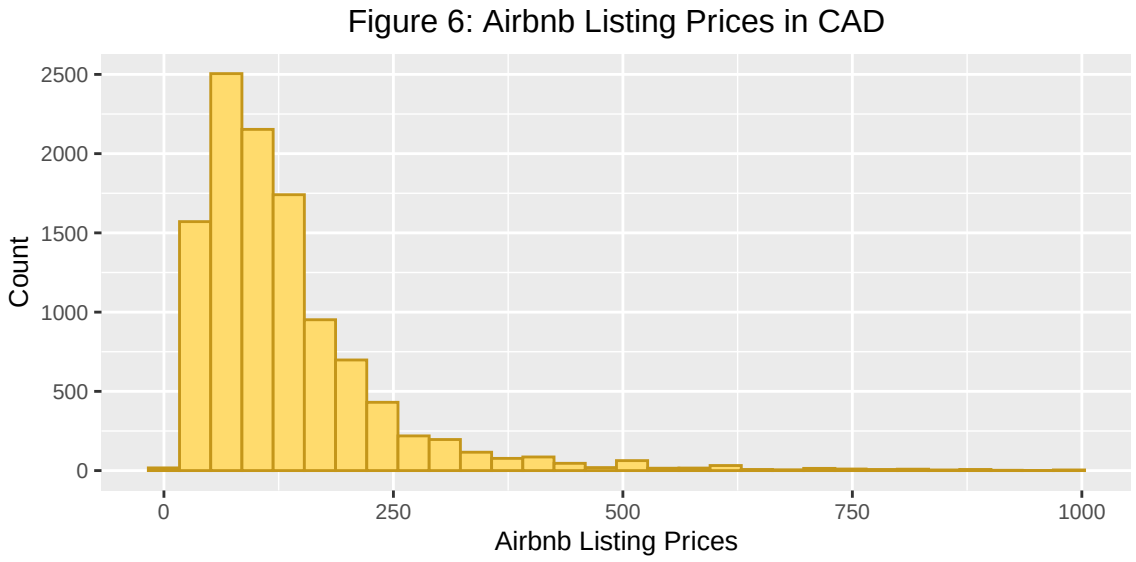

Figure 6 is a histogram for the response variable price in the unmatched dataset with 11023 observations. It appears that price is a right-skewed, unimodal distribution with higher priced listings on the right end of the plot. This also corresponds to the higher mean of price 133.42 CAD when compared to the median of price 103 CAD, where the mean is brought up by the listings with prices that are more extreme.

Table 8: Predictors in models with the highest adjusted R squared of each size

	2	3	4	5	6	7	8	9	10	11
Intercept	1	1	1	1	1	1	1	1	1	1
Treatment	1	1	1	1	1	1	1	1	1	1
Beds	1	1	1	1	1	1	1	1	1	1
Number of reviews	0	0	0	0	0	0	0	0	1	1
Review scores: Accuracy	0	0	0	0	0	0	0	1	1	1
Review scores: Cleanliness	0	0	0	0	0	1	1	1	1	1
Review scores: Check-in	0	0	1	1	1	1	1	1	1	1
Review scores: Communication	0	0	0	0	0	0	0	0	0	1
Review scores: Location	0	1	1	1	1	1	1	1	1	1
Review scores: Value	0	0	0	0	0	0	1	1	1	1
Calculated host listings count	0	0	0	0	1	1	1	1	1	1
Availability in 365 days	0	0	0	1	1	1	1	1	1	1

Note:
This table corresponds to Table 5 in the Results section, where the model with size 9 has the highest adjusted R squared.

Table 8 shows the predictors in the models with the highest R^2_{adj} of each size, which is supplementary to Table 5 in the Results section. Each column represents a model with the size specified by the column name. 1 indicates that this predictor is in the model, and 0 otherwise. Note that the model with 9 predictors has the highest R^2_{adj} among them and thus is chosen as the final model of listing prices.