What affects the Price of Airbnb Listings in NYC?

by Alexey Yushkin and Jack Daoud

Contents

Framing the Problem

Problem Recognition:

Pricing is a complicated but essential business decision. This is true for Airbnb hosts too, which have "limited and inefficient strategies" to "price their spaces" (Li *et al.*, 2015). The problem of price is made even more difficult hosts within New York City, NY where there are about **44,600 listings in NYC** - a lot of competition to consider. This pricing problem can be recognized as:

Which criteria has a statistically significant relationship with the price of an Airbnb listing in NYC?

Review of Previous Findings:

Previous research regarding Airbnb accommodation prices had found that "price is significantly related to the level of the host's accumulated experience and the level of market demand on a specific booking date" (Magno *et al.*, 2018). More specifically, this research paper computed the impact of specific variables on the price of an Airbnb accommodation:

Variable	Description	Impact on Price of Accomodation (%)
Entire home/apartment	Entire home/apartment (dummy) (vs private/shared room)	39.86%
Size	Number of beds	5.11%
Reviews	Number of reviews	-0.26%
Professional	Professional host (host with two or more listings simultaneously) (vs non-professional host)	5.83%
Experience	Number of months since the host joined Airbnb	0.21%
Market demand	Total number of beds in bookable shared accomations available on a specific date	0.02%

Thought Process:

Our analysis aims to infer whether or not certain variables - which are different from variables in the aforementioned literature - are related to the price of an accommodation. These variables are summarized under section 2.3 Variable Selection. The hypothesis for each variable will be as follows:

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

Note: β_i denotes the variable being tested, and in our case, we're assessing approximately 14. More specifically though, we're testing these 14 null hypotheses:

- **H01**: The price of a listing is not related to the **number of reviews** the accommodation received
- **H02**: The price of a listing is not related to the **number of reviews per month** the accommodation received
- **H03**: The price of a listing is not related to the listing's **demand**
- **H04**: The price of a listing is not related to its **distance from subway stations**
- **H05**: The price of a listing is not related to its **distance from bus stops**
- **H06**: The price of a listing is not related to the **number of cultural organizations nearby**
- **H07**: The price of a listing is not related to the **number of shootings nearby**
- **H08**: The price of a listing is not related to the **number of 911 calls nearby**
- **H09**: The price of a listing is not related to its **number of negative reviews**
- **H010**: The price of a listing is not related to its **number of neutral reviews**
- **H011**: The price of a listing is not related to its **number of positive reviews**
- **H012**: The price of a listing is not related to the **number of days since its last review**
- **H013**: No difference exists between the price of an **entire apartment** and that of a **private room**
- **H014**: The price of a listing is not related to the **number of listings a host has**

Solving the Problem

Data Collection:

The data sets we'll be analyzing revolves around Airbnb listings in New York City, NY in the US. The analysis also includes data sets regarding tourist attractions, crime rates, and transportation proximity. The number of observations and variables per data set is outlined below:

Data	Description	Details
Airbnb Listings NYC	Summary information and metrics for listings in New York City.	44,666 observations of 16 variables
NYC Transit Subway - Entrance and Exit Data	This data file provides a variety of information on subway station entrances and exits which includes but is not limited to: Division, Line, Station Name, Longitude and Latitude coordinates of entrances/exits.	1,869 observations of 32 variables
NYC Bus Stop Shelter	This dataset contains the location of Bus Stop Shelters in NYC.	3,428 observations of 18 variables
DCLA Cultural Organizations	Listing of all Cultural Organizations in the Department of Cultural Affairs directory in NYC	2,308 observations of 16 variables
NYPD Shooting Incident Data (Historic)	List of every shooting incident that occurred in NYC going back to 2006 through the end of the previous calendar year.	21,626 obersvations of 19 variables
NYPD Shooting Incident Date (Year to Date)	List of every shooting incident that occurred in NYC during the current calendar year.	1,501 observations of 19 variables
NYPD Calls for Service (911)	Calls for Service to NYPD's 911 system	3,913,878 observations of 19 variables
Airbnb Listing Reviews NYC	Detailed Review Data for listings in New York City	1,003,065 observations of 6 variables

For data sources & snippets, please see the Appendix below.

Data Wrangling:

The data wrangling process involved joining the data sets in 5 sequential steps:

1. Listing data joined with Subway station data
2. Data from step 1 joined with Bus station data
3. Data from step 2 joined with Tourist attraction data
4. Data from step 3 joined with Crime rate data (Shootings / 911 Calls)
5. Data from step 4 joined with Listing reviews data with sentiment analysis

Variable Selection:

The finalized data set produced, after joining all aforementioned data sets & conducting data analysis below, has 19 key variables:

1. `price`: the cost of renting a listing for a day
2. `minimum_nights`: the number of minimum nights permitted for the booking of a listing
3. `number_of_reviews`: the total number of reviews for a listing
4. `reviews_per_month`: the number of reviews per month for a listing

5. `availability_365`: the number of days per year a listing is available for renting
6. `distance_from_subway_station`: the distance in miles between a listing and the nearest subway station
7. `distance_from_bus_stop`: the distance in miles between a listing and the nearest bus stop
8. `n_cult_orgs`: the number of cultural organizations / tourist attractions nearby a listing
9. `shooting_number`: the number of reported shootings nearby a listing
10. `calls_911_number`: the number of calls made to 911 nearby a listing
11. `negative`: the number of negative reviews for a listing
12. `neutral`: the number of neutral reviews for a listing
13. `positive`: the number of positive reviews for a listing
14. `days_last_rev`: the number of days since a listing's last review
15. `room_type`: whether a listing is an Entire Home/Apt, or Private room, or hotel, or shared room
16. `neighborhood_group`: which neighborhood group a listing is located in
17. `calculated_host_listings_count`: the total number of listings on Airbnb per host
18. `long_term_rent`: whether a listing is has a minimum stay of less than 30 days (short-term) or 30 days or more (long-term)
19. `demand`: the total number of days a listing is *not* available within a year

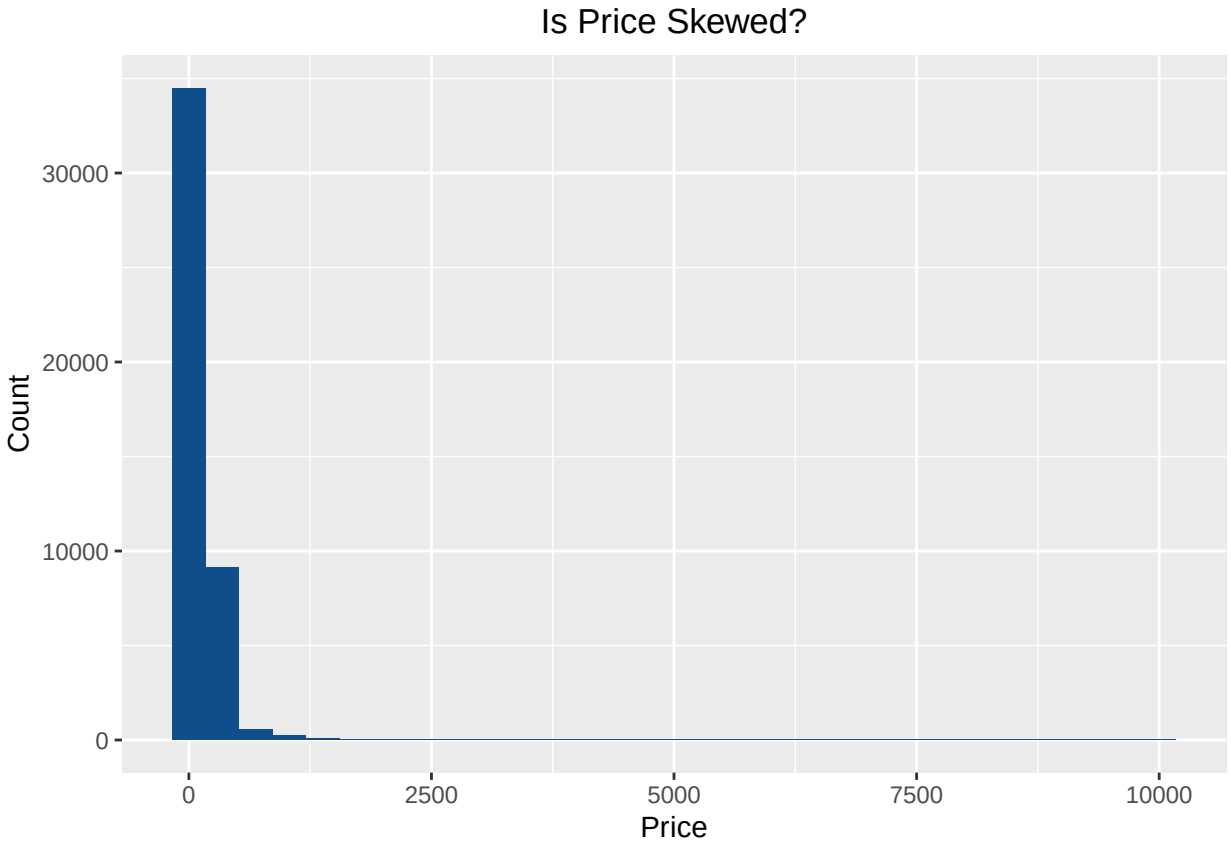
All these variables will be criteria that could answer our business problem:

Which criteria has a statistically significant relationship with the price of an Airbnb listing in NYC?

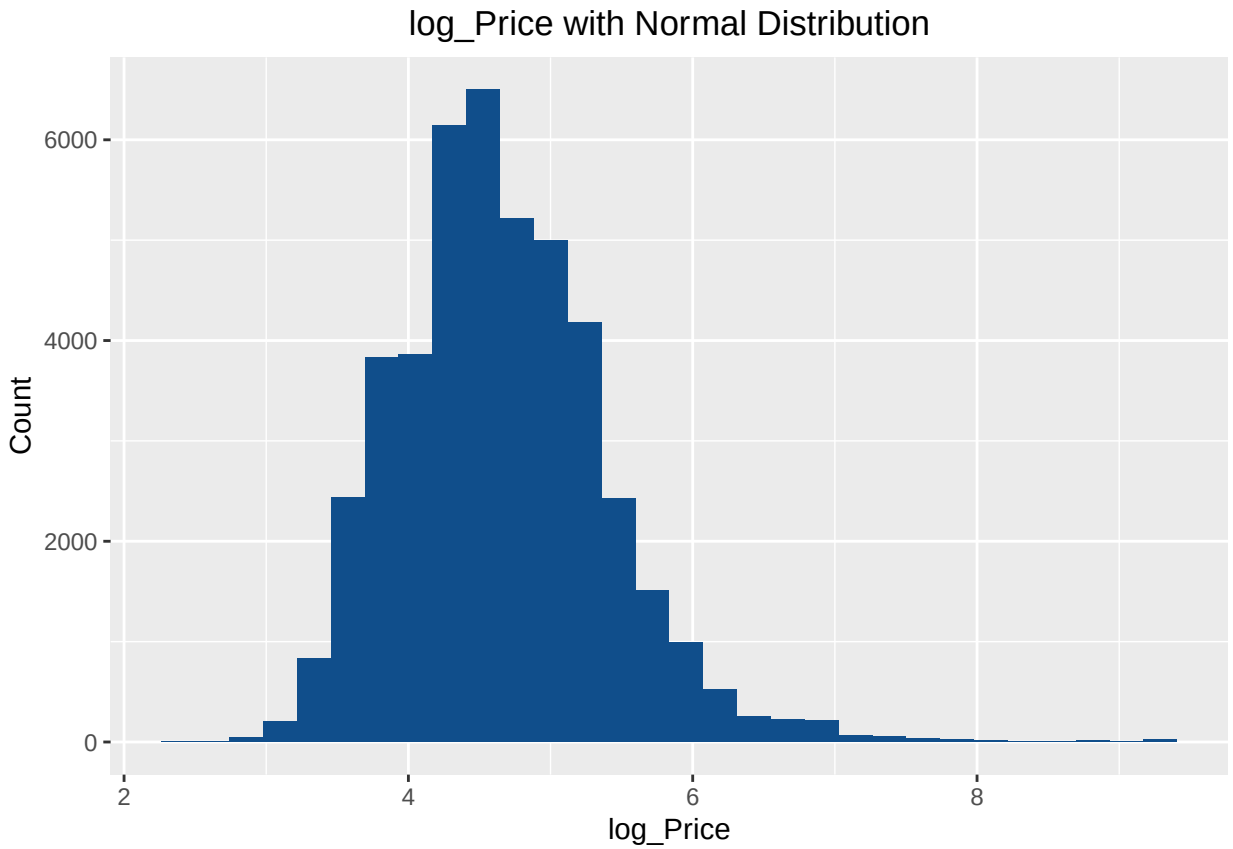
Data Analysis:

Price

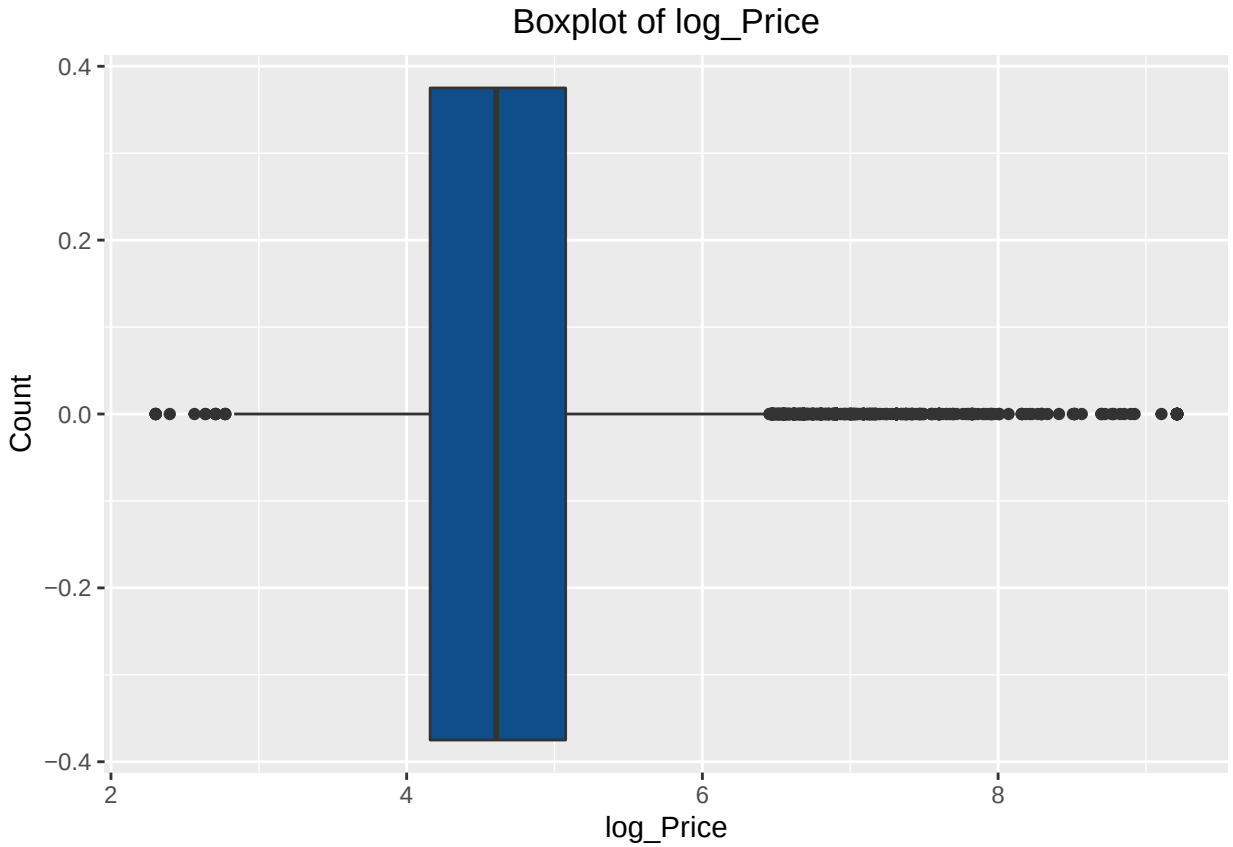
The first step is to assess the main variable of our analysis - the *price* of a listing.



As we can see, price is **highly skewed to the right** and therefore **warrants a transformation** if we are to conduct further analysis in the form of linear regression. Furthermore, it seems there are **some listings with a price of zero** that ought to be **excluded**.



Now price represents a more *'normal'* distribution. However, there appears to be **outliers**. We can clearly identify them using a boxplot:



What is the cause of all these outliers in terms of price? It could be a categorical variable, such as the neighborhood group or room type of the listing. It is important to note, although these prices are considered to **statistical outliers**, really expensive listings and really cheap listings have their **business justifications that make them normal**. For example, a listing in a booming location with extremely fancy furniture & many amenities. Therefore, these outliers will be considered throughout the analysis.

Room type & Neighborhood Group

Let's start with looking at the count of each variable

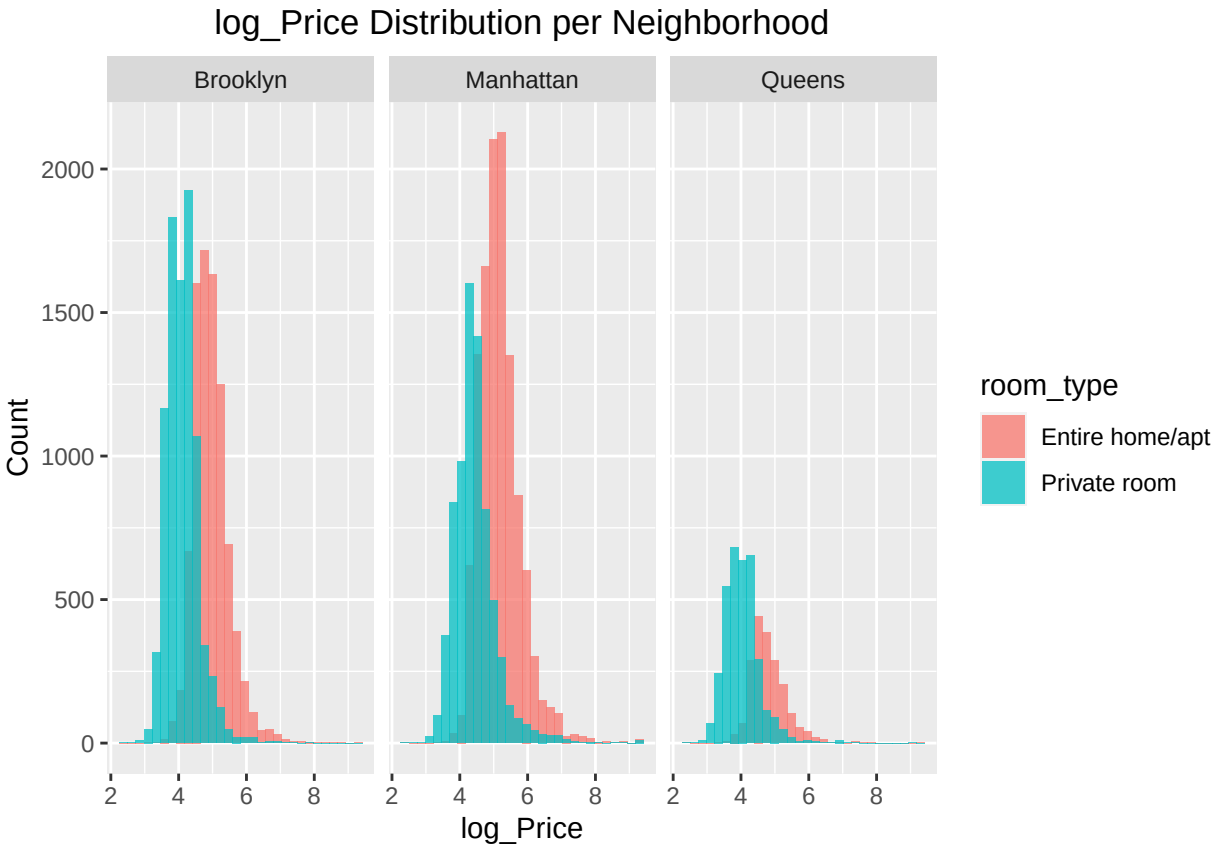
room_type	num_listing
Entire home/apt	22859
Hotel room	368
Private room	20494
Shared room	920

neighbourhood_group	num_listing
Bronx	1158
Brooklyn	17850
Manhattan	19718
Queens	5601
Staten Island	314

Hotel rooms and shared rooms make up approximately **1%** and **2%** of the total listings respectively. Listings

in Staten Island and Bronx make up also **1%** and **2%** of total listings respectively. Such minute proportions are negligible and therefore warrant an exclusion from the analysis.

Now let's look at how each room type relates with price.



Observations of the above visualization can be summarized as follows:

- Private room prices are grouped towards lower levels
- Entire home/apt prices are grouped towards higher levels
- Prices of listings are lower in Queens compared to Brooklyn and Manhattan.

Based on this, there seems to be some relationship between room type, neighborhood group and price. This relationship will be worth analyzing by using a regression model in section 3.

Minimum nights

Any listing with a minimum nights requirement of **over 365 days (greater than 1 year)** will be assumed to be **unreasonable** and therefore **excluded** from the analysis.

Furthermore, there seems to be a legitimate distinction between listings in terms of minimum number of days for length of stay. A listing with a minimum number of nights of **less than 30 days** will be denoted as **Short-term**, whereas listings with minimum number of nights of **greater than or equal to 30 days and less than 1 year** will be denoted as **Long-term**.

```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```

Approximately **15%** of listings are for **long-term** stays whereas **85%** are for **short-term**. Furthermore, we'll assume that long-term pricing strategies for listings are static and predictable, whereas short-term pricing strategies are more dynamic and less predictable. Based on this, *our analysis will focus on the short-term listings and exclude long-term ones.*

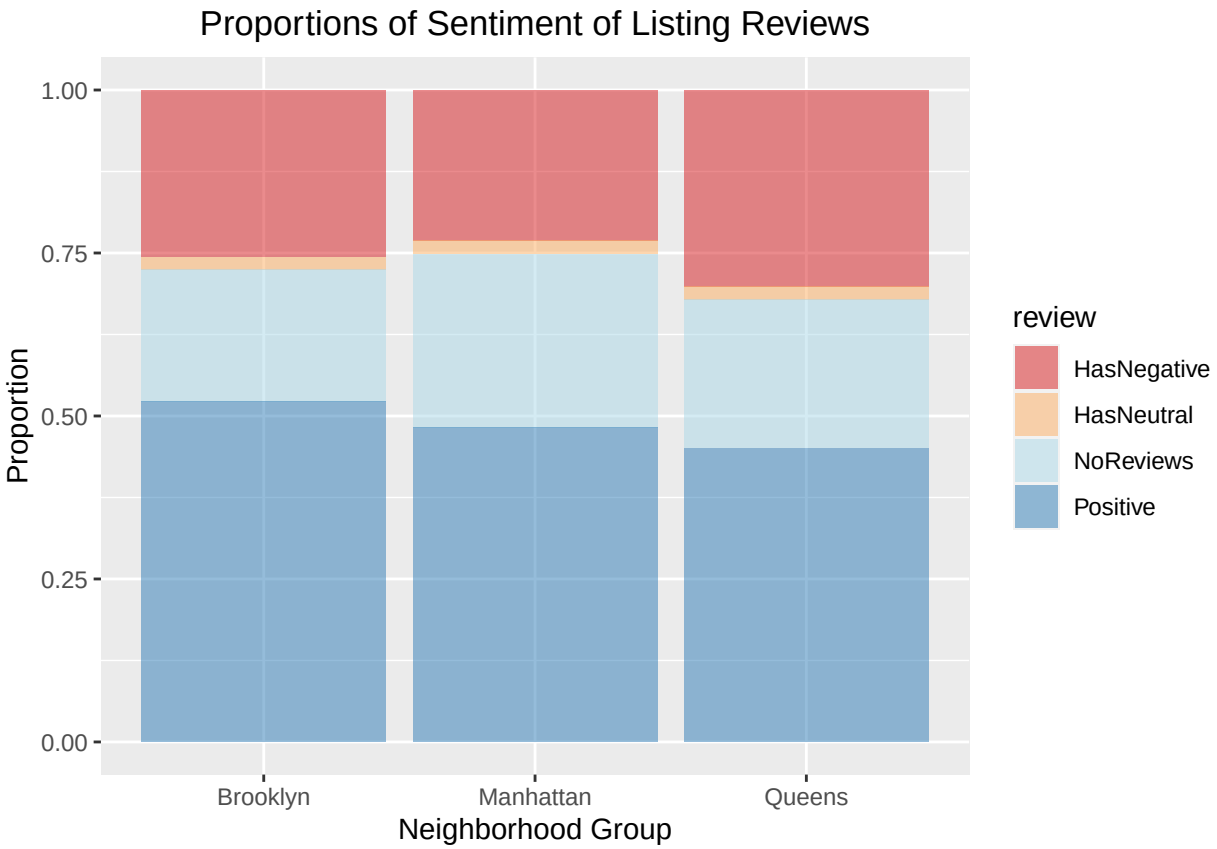
Sentiment Analysis

When looking at reviews as a customer, one usually looks out for negative reviews. Therefore, we'll assume that *negative reviews have more weight* in comparison to neutral or positive reviews.

Based off that assumption, the categorical variable **review** was created with the following conditional:

- If listing has 1 or more negative reviews, then "review" = HasNegative
- If listing has 0 negative but 1 or more neutral reviews, then "review" = HasNeutral
- If listing has 0 negative and 0 neutral but 1 or more positive reviews, then "review" = Positive
- If listing has no reviews, then "review" = NoReviews

A more appropriate approach might be to assign the category of HasNegative based on a proportion of the negative reviews at a certain threshold. For e.g., if a listing has 10% or more negative reviews from its total reviews, then it should be labeled as HasNegative. However, demarcating such a threshold will have to be assessed (this will be a step we'll conduct post the initial check-in)



The proportions are strikingly close across neighborhoods. We can see that approximately **50%** of listings have **purely positive reviews** across neighborhoods, with *Queens having the least purely positive reviews.* About **25%** of listings **don't have reviews at all**, most of which are in Manhattan. This hints at short stays that don't necessarily warrant reviews. Lastly, almost **25%** of listings **have negative reviews** across each neighborhood, with Queens having the most. What is negatively affecting reviews in Queens? Is it the hosts? The neighborhoods? The renters?

What about review sentimentality and price?

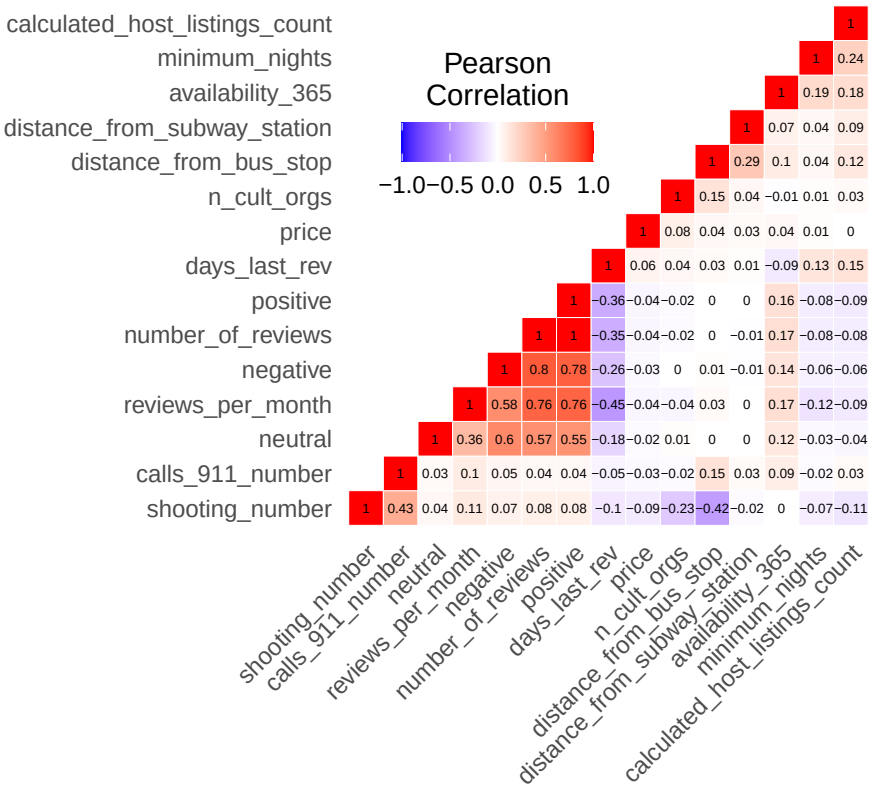
The distributions of review sentimentality and price are strikingly similar, hinting at the fact that either

a) there might not be a relationship between review sentimentality and price, or
b) if there's a relationship, it will be a weak one

This will have to be assessed in section 3 by running a Multiple Linear Regression model using either dummy variables or the quantities of each review sentiment.

Correlation Matrix

Plotting a correlation matrix can help guide which variable we ought to assess with regards to modeling in section 3.



We can see that review sentimentality, number of reviews, and reviews per month are *highly positively (+0.5) correlated* with one another. The same 3 variables are highly negatively correlated with days since last review. Interestingly, number of shootings and distance from bus stops is also highly negatively correlated. Unfortunately, we can see that there is not much correlation between price and all these numerical variables. This could likely be due to the fact that not all variables have linear relationships. This also suggests that perhaps categorical variables in the form of dummy variables would be better explanatory variables for price.

Thought Process Adjustments:

We came to realize that a few of our variables have skewed distributions which warrant log transformations:

1. `demand`
2. `number_of_reviews`
3. `reviews_per_month`
4. `calls_911_number`
5. `negative`
6. `neutral`
7. `positive`
8. `negative_prop` (proportion)
9. `neutral_prop`
10. `positive_prop`
11. `days_last_review`

Other than those transformations, there will be no major adjustments other than the aspect in which we will assess the relationships between all the aforementioned variables and price. When looking at each variable, we'll have to split the modeling based on specific categories such as listing type (`Entire home/apt`), rent duration (`long_term_rent`), and `neighborhood group`. Furthermore, when assessing review sentimentality, we will attempt a variety of approaches by looking at sentimentality as a number, factor, and proportion.

Modelling

Model Benchmark

First, let's see how much of price is explained by the entire list of 15 aforementioned variables. This is with the addition of the sentiment_review_proportions, making it a total of 18 variables:

```
## Adjusted R squared: 0.4386143
```

So the list of variables we're assessing can explain up to 44% of the price of an Airbnb listing. But which variables truly play a role? i.e. which variables can be neglected?

term	estimate	std.error	statistic	p.value
(Intercept)	4.7808817	0.0437466	109.2857357	0.0000000
demand_log	-0.0498883	0.0017848	-27.9515296	0.0000000
number_of_reviews_log	-0.0868043	0.0215327	-4.0312720	0.0000556
reviews_per_month_log	0.0420945	0.0048527	8.6744646	0.0000000
distance_from_subway_station	-0.0000038	0.0000108	-0.3520402	0.7248105
distance_from_bus_stop	-0.0000380	0.0000045	-8.3936421	0.0000000
n_cult_orgs	0.0235400	0.0011659	20.1896864	0.0000000
shooting_number_log	-0.0974091	0.0025786	-37.7766255	0.0000000
calls_911_number	-0.0000012	0.0000003	-4.1764549	0.0000297
negative_log	-0.0769306	0.0113801	-6.7601006	0.0000000
neutral_log	-0.0173320	0.0305951	-0.5664961	0.5710605
positive_log	0.1020215	0.0246896	4.1321644	0.0000360
days_last_rev_log	0.0547798	0.0032531	16.8393654	0.0000000
entire_home_apt	0.7632362	0.0058255	131.0171179	0.0000000
distance_from_transportation	0.0000462	0.0000132	3.5098986	0.0004489
negative_prop_log	0.0037003	0.0012308	3.0065336	0.0026445
neutral_prop_log	-0.0001440	0.0025802	-0.0558244	0.9554821
positive_prop_log	-0.0063749	0.0029218	-2.1818203	0.0291300
calculated_host_listings_count	-0.0042807	0.0001847	-23.1763924	0.0000000

From the above table, we can see that some variables can be excluded from further modeling because they are statistically insignificant, i.e. they have a p-value greater than 0.05:

- `distance_from_subway_station`
- `neutral_log`
- `neutral_prop_log`

For the rest of the variables, we can reject $H0_i$ at a confidence level of 95% (significance level 0.05). But first we have to run a few more models to be sure.

Furthermore, we can see that **some variables have a inverse relationship with price**, i.e. when this variable increases, price is likely to decrease. These are the variables along with a rational explanations for the inverse relationship:

- `demand_log`: this makes sense based off the law of demand which states price and demand are inversely related, so that if price decreases, demand will increase.
- `number_of_reviews_log`: hosts with a large number of reviews possibly set low prices to attract more customers.
- `distance_from_bus_stop`: if a listing is further away from bus stops, it will cost less because of a low transportation score.
- `shooting_number_log`: a listings with a higher recorded number of shootings in the area will have a lower price due to a decreased sense of safety.
- `calls_911_number`: same as the variable above.

- `negative_log`: the more negative reviews a listing has, the lower the price will be. This is so hosts can attract price-sensitive customers that're willing to overlook a bad reputation.
- `calculated_host_listings_count`: this is likely due to the fact that hosts with a higher number of listings normally set lower prices to attract more customers.

On the other hand, **some variables have a positive relationship with price**, i.e. when this variable increases, price is likely to increase with it:

- `reviews_per_month_log`: this actually raises another question, why is it that an increase in the number of reviews per month increases price, whereas an increase in the total number of reviews decreases price?
- `n_cult_orgs`: the more tourist attractions nearby a listing, the higher the price.
- `positive_log`: the more positive reviews a listing has, the higher the price.
- `days_last_rev_log`: the higher the number of days since the last review for a listing, the higher the price.
- `entire_home_apt`: the type of listing has the most significant impact on price (this reiterates the aforementioned previous findings).
- `distance_from_transportation`: the coefficient for this variable is non-intuitive, because it states that the higher the distance from transportation, the higher the price. Perhaps the explanation is that we already used `distance_from_bus_stop` and `distance_from_subway_station` so `distance_from_transportation` brings an insignificant correction.

Now let's assess the relationship of statistically significant variables with price grouped under these categories:

1. `long_term_rent`
2. `neighborhood group`
3. `review`

Model by Neighborhood

```
## Adjusted R squared:
```

```
## $Brooklyn
## [1] 0.4781637
##
## $Manhattan
## [1] 0.3468879
##
## $Queens
## [1] 0.4410756
```

Interestingly, the **adjusted r squared** changes across neighborhoods. For Queens, it's identical to the first overarching model run at **44%**. For Brooklyn and Manhattan, it increased to **48%** and decreased to **35%** respectively. This means our model is best at explaining the price of an Airbnb listing in Brooklyn. It also means there are lurking variables that one ought to consider when it comes to pricing Airbnb listings in Manhattan.

The variables being assessed for a relationship with price behave differently depending on the neighborhood of the listing. These are the statistically significant variables per neighborhood:

Brooklyn	Manhattan	Queens
demand_log	demand_log	demand_log
distance_from_bus_stop	number_of_reviews_log	reviews_per_month_log
n_cult_orgs	reviews_per_month_log	n_cult_orgs
shooting_number_log	n_cult_orgs	calls_911_number
calls_911_number	shooting_number_log	negative_log
negative_log	negative_log	days_last_rev_log

Brooklyn	Manhattan	Queens
days_last_rev_log	positive_log	calculated_host_listings_count
distance_from_transportation	days_last_rev_log	entire_home_apt
calculated_host_listings_count	distance_from_transportation	
entire_home_apt	calculated_host_listings_count	
	entire_home_apt	

The variables that are statistically significant across all neighborhoods are:

- `demand_log`
- `n_cult_orgs`
- `negative_log`
- `days_last_rev_log`
- `calculated_host_listings_count`
- `entire_home_apt`

Model by Rent Duration

```
## Adjusted R squared:
```

```
## $`Short-term`
## [1] 0.4424669
##
## $`Long-term`
## [1] 0.4858908
```

The **adjusted r squared** doesn't differ significantly across durations of rent. For Short-term, it's identical to the first overarching model run at **44%**. For Long-term, it increased to **49%**. This means our model is best at explaining the price of an Airbnb listing that is long-term rather than short-term. This makes sense intuitively, because long-term pricing is more static and easier to predict, whereas short-term pricing is more dynamic and more difficult to predict.

Just as the variables behaved differently when being assessed by neighborhood, the same applies when being assessed by rent duration. These are the statistically significant variables per rent duration:

Long-term	Short-term
demand_log	demand_log
number_of_reviews_log	reviews_per_month_log
reviews_per_month_log	distance_from_bus_stop
n_cult_orgs	n_cult_orgs
shooting_number_log	shooting_number_log
calls_911_number	calls_911_number
negative_log	negative_log
positive_log	days_last_rev_log
days_last_rev_log	entire_home_apt
entire_home_apt	distance_from_transportation
calculated_host_listings_count	negative_prop_log
	calculated_host_listings_count

The variables that are statistically significant across both rent durations are:

- `demand_log`
- `reviews_per_month_log`
- `n_cult_orgs`

- shooting_number_log
- calls_911_number
- negative_log
- days_last_rev_log
- calculated_host_listings_count
- entire_home_apt

Model by Review Sentiment

```
## Adjusted R squared:

## $Positive
## [1] 0.4307711
##
## $HasNeutral
## [1] 0.4663408
##
## $HasNegative
## [1] 0.453184
##
## $NoReviews
## [1] 0.3908127
```

The **adjusted r squared** doesn't differ significantly across types of reviews. Listings with no reviews can explain approximately 39% of price, whereas listings that have Negative or Neutral reviews can explain 45% and 47% respectively. Lastly, listings with purely positive reviews can explain about 43% of price. WHAT CAN WE SEE?

These are the statistically significant variables per review:

Positive	HasNeutral	HasNegative	NoReviews
demand_log	demand_log	demand_log	demand_log
distance_from_bus_stop	distance_from_bus_stop	distance_from_bus_stop	distance_from_bus_stop
n_cult_orgs	n_cult_orgs	n_cult_orgs	n_cult_orgs
shooting_number_log	shooting_number_log	shooting_number_log	shooting_number_log
positive_log	positive_log	positive_log	
		negative_log	
days_last_rev_log	days_last_rev_log	days_last_rev_log	
		calls_911_number	calls_911_number
entire_home_apt	entire_home_apt	entire_home_apt	entire_home_apt
distance_from_transportation		distance_from_transportation	distance_from_transportation
calculated_host_listings_count	calculated_host_listings_count	calculated_host_listings_count	calculated_host_listings_count

The variables that are statistically significant across types of reviews are:

- demand_log
- distance_from_bus_stop
- n_cult_orgs
- shooting_number_log
- calls_911_number
- entire_home_apt
- calculated_host_listings_count

Model Comparison

The three models we looked at are:

1. Model Set 1 - 3 Models Categorized by Neighborhood Group
2. Model Set 2 - 2 Models Categorized by Rent Duration
3. Model Set 3 - 4 Models Categorized by Review Sentiment

The variables that are statistically significant within each set of models are:

Model Set 1	Model Set 2	Model Set 3
demand_log	demand_log	demand_log
		distance_from_bus_stop
	reviews_per_month_log	
n_cult_orgs	n_cult_orgs	n_cult_orgs
	shooting_number_log	shooting_number_log
negative_log	negative_log	
days_last_rev_log	days_last_rev_log	
	calls_911_number	calls_911_number
entire_home_apt	entire_home_apt	entire_home_apt
distance_from_transportation		distance_from_transportation
calculated_host_listings_count	calculated_host_listings_count	calculated_host_listings_count

The variables that are statistically significant across all models or at least 2 models are:

All Model Sets	Two Model Sets
demand_log	demand_log
n_cult_orgs	n_cult_orgs
calculated_host_listings_count	calculated_host_listings_count
entire_home_apt	entire_home_apt
	distance_from_transportation
	shooting_number_log
	negative_log
	days_last_rev_log
	calls_911_number

Model Conclusion

Now we build our final model with all the variables that are statistically significant in at least two of our previous models. How much of price can be explained?

Note: `distance_from_transportation` and `calls_911_number` were removed due to being statistically insignificant after our initial run of the model

```
## Adjusted R squared: 0.4168248
```

Our final model with 7 variables has 42% explanatory power whereas our initial model had 44%. That's approximately 2% decrease when compared to the initial model which tested 18 variables. Barely any change in explanatory power but less than half the number of variables were used. Therefore, although the explanatory power decreases slightly, the simplicity of the model increases substantially (reduced from 18 variables to 7).

One important note is that one of our variables `entire_home_apt` was already inferred to have a strong relationship with price in our Section 1.2 Review of Previous Findings. We chose to include this variable in

our analysis because it is extremely significant in such a way that if we didn't use the variable, we'd have to split the data set and analyze each one separately (doubling the number of models).

Moreover, our final model that has an explanatory power of 42% explains more when it comes to price than the model found in the review of previous findings with an explanatory power of 27% (Magno *et al.*, 2018).

Communication

So what factors should we consider when attempting to solve the pricing problem of an Airbnb listing? More specifically...

Which criteria has a statistically significant relationship with the price of an Airbnb listing in NYC?

Let's revise the aforementioned list of 14 null hypotheses (Section 1.3 - Though Process). The null hypotheses crossed out have been rejected due to statistical significance (i.e. a p-value of less than 0.1):

- **H01**: The price of a listing is not related to the **number of reviews** the accommodation received
- **H02**: The price of a listing is not related to the **number of reviews per month** the accommodation received
- **H03**: The price of a listing is not related to the listing's **demand**
- **H04**: The price of a listing is not related to its **distance from subway stations**
- **H05**: The price of a listing is not related to its **distance from bus stops**
- **H06**: The price of a listing is not related to the **number of cultural organizations nearby**
- **H07**: The price of a listing is not related to the **number of shootings nearby**
- **H08**: The price of a listing is not related to the **number of 911 calls nearby**
- **H09**: The price of a listing is not related to its **number of negative reviews**
- **H010**: The price of a listing is not related to its **number of neutral reviews**
- **H011**: The price of a listing is not related to its **number of positive reviews**
- **H012**: The price of a listing is not related to the **number of days since its last review**
- **H013**: No difference exists between the price of an **entire apartment** and that of a **private room**
- **H014**: The price of a listing is not related to the **number of listings a host has**

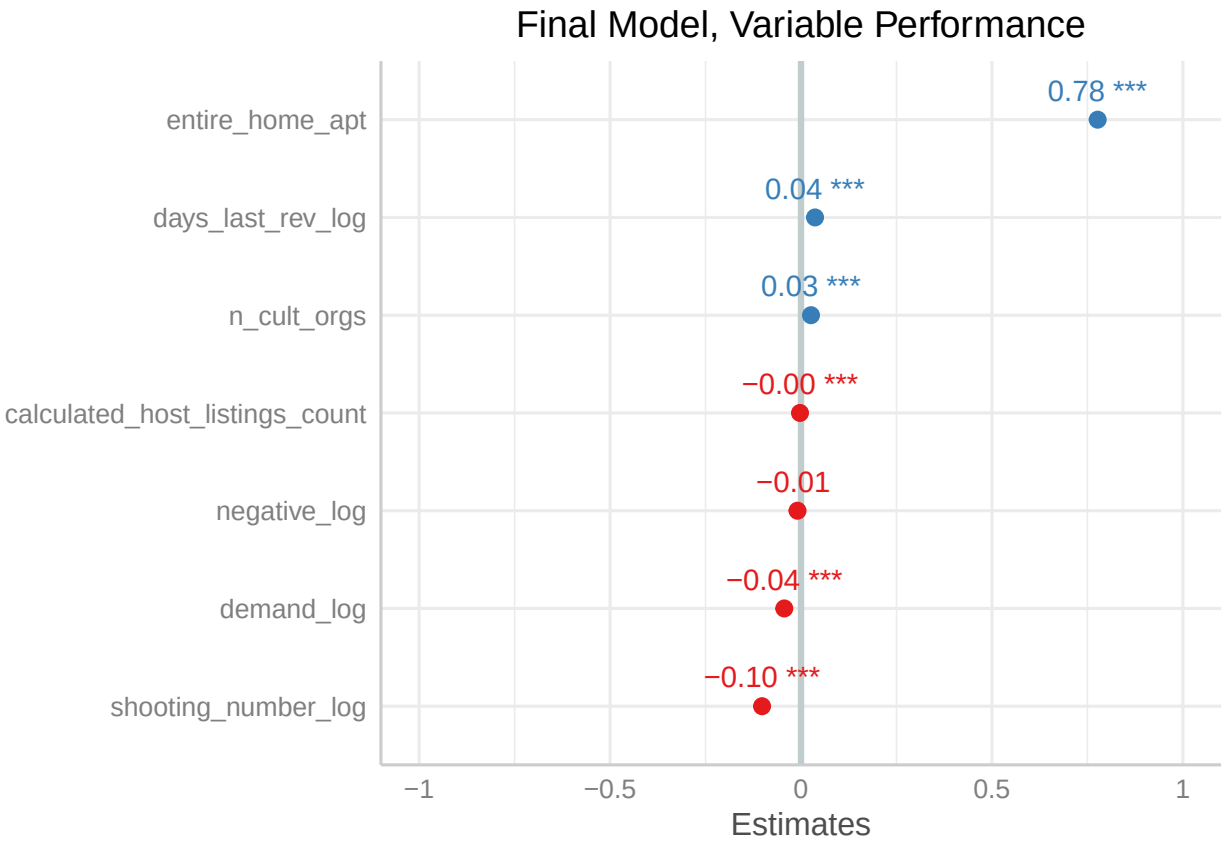
Now we're left with these 7 alternative hypotheses as answers to our question and possible solutions to the pricing problem:

- **HA1**: The price of a listing *is* related to the listing's **demand**
- **HA2**: The price of a listing *is* related to the **number of shootings nearby**
- **HA3**: The price of a listing *is* related to its **number of negative reviews**
- **HA4**: The price of a listing *is* related to its **number of cultural organizations nearby**
- **HA5**: The price of a listing *is* related to its **number of days since its last review**
- **HA6**: The price of a listing *is* related to the **number of listings a host has**
- **HA7**: A difference *exists* between the price of an **entire apartment** and that of a **private room**

So these are the criteria which have a statistically significant relationship with the price of an Airbnb listing in New York. But what about the impact of each criteria on price?

	variable_description	percent_of_impact	pvalue
entire_home_apt	Entire home/apartment (dummy) (vs private/shared room)	77.68	0.000
shooting_number_log	Number of shootings nearby the listing (logarithm transformed)	-10.20	0.000
demand_log	Number of days the listing is not available in a year (logarithm transformed)	-4.38	0.000
days_last_rev_log	Number of days past since a listing received a review (logarithm transformed)	3.67	0.000

	variable_description	percent_of_impact	pctvalue
n_cult_orgs	Number of tourist attractions nearby the listing	2.62	0.000
negative_log	Number of negative reviews a listing has (logarithm transformed)	-0.92	0.063
calculated_host_listings_count	Number of total listings currently on Airbnb by a single host	-0.27	0.000



Here we can see that our findings both reiterate previous findings as well as add onto them...

Both in our analysis and previous findings, the variable `entire_home_apt` has the highest impact on the price of a listing.

The adding onto previous findings comes along with the assessment of the unique variables of:

1. Number of shootings (`shooting_number_log`): high negative impact
2. Demand (`demand_log`): medium negative impact
3. Days since last review (`days_last_rev_log`): medium positive impact
4. Number of cultural organizations nearby (`n_cult_orgs`): medium positive impact
5. Number of listings per host (`calculated_host_listings_count`): low negative impact
6. Has negative reviews (`negative_log`): low negative impact

In regards to #2, we used a different approach in computing the demand of a listing that resulted with a higher explanatory power in comparison to that of previous findings in Section 1.2.

Furthermore, in regards to #3, the idea that less reviews (or at least an increased elapsed time between reviews) positively impacts price also reiterates previous findings:

"In addition, the findings show that the number of reviews received on Airbnb by an accommodation is negatively correlated to its price. This result is consistent with the evidence provided by Gibbs et al. (2018),

who suggested that the lower the price, the higher the number of bookings and, in turn, the higher the number of reviews." (Magno *et al.*, 2018)

Three Major Insights

In conclusion, the **three major insights** we can derive from our analysis (that was not found in previous research) are:

1. Number of shootings in a neighborhood, which we have chosen as a criminal ratio, has a significant negative impact on price.
2. Number of tourist attractions nearby the listing have a significant positive impact on the price.
3. Existence of negative reviews has a significant negative impact on the price.

Appendix

Data Sources

Data	Link
Airbnb Listings NYC	http://data.insideairbnb.com/united-states/ny/new-york-city/2020-10-05/visualisations/listings.csv
NYC Transit Subway - Entrance and Exit Data	https://data.ny.gov/Transportation/NYC-Transit-Subway-Entrance-And-Exit-Data/i9wp-a4ja
NYC Bus Stop Shelter	https://data.cityofnewyork.us/Transportation/Bus-Stop-Shelters/qafz-7myz
DCLA Cultural Organizations	https://data.cityofnewyork.us/Recreation/DCLA-Cultural-Organizations/u35m-9t32
NYPD Shooting Incident Data (Historic)	https://data.cityofnewyork.us/Public-Safety/NYPD-Shooting-Incident-Data-Historic-/833y-fsy8
NYPD Shooting Incident Date (Year to Date)	https://data.cityofnewyork.us/Public-Safety/NYPD-Shooting-Incident-Data-Year-To-Date-/5ucz-vwe8
NYPD Calls for Service (911)	https://data.cityofnewyork.us/Public-Safety/NYPD-Calls-for-Service/n2zq-pubd
Airbnb Listing Reviews NYC	http://data.insideairbnb.com/united-states/ny/new-york-city/2020-10-05/data/reviews.csv.gz

Snippets of Original Data

01) Airbnb Listings in NYC:

```
## # A tibble: 5 x 18
##      id name  host_id host_name neighbourhood_g~ neighbourhood latitude
##   <dbl> <chr>   <dbl> <chr>     <chr>            <chr>            <dbl>
## 1  2595 Skyl~    2845 Jennifer  Manhattan        Midtown           40.8
## 2  3831 Whol~    4869 LisaRoxa~ Brooklyn         Clinton Hill      40.7
## 3  5121 Blis~    7356 Garon     Brooklyn         Bedford-Stuy~     40.7
## 4  5136 Spac~    7378 Rebecca   Brooklyn         Sunset Park       40.7
## 5  5178 Larg~    8967 Shunichi  Manhattan        Hell's Kitch~     40.8
## # ... with 11 more variables: longitude <dbl>, room_type <chr>, price <dbl>,
## #   minimum_nights <dbl>, number_of_reviews <dbl>, last_review <date>,
## #   reviews_per_month <dbl>, calculated_host_listings_count <dbl>,
## #   availability_365 <dbl>, distance_from_subway_station <int>, station <chr>
```

02) NYC Transit Subway:

```
##   Division     Line Station.Name Station.Latitude Station.Longitude Route1
## 1      BMT 4 Avenue      25th St         40.66040         -73.99809      R
## 2      BMT 4 Avenue      25th St         40.66040         -73.99809      R
## 3      BMT 4 Avenue      36th St         40.65514         -74.00355      N
## 4      BMT 4 Avenue      36th St         40.65514         -74.00355      N
## 5      BMT 4 Avenue      36th St         40.65514         -74.00355      N
##   Route2 Route3 Route4 Route5 Route6 Route7 Route8 Route9 Route10 Route11
## 1                                           NA     NA      NA      NA
## 2                                           NA     NA      NA      NA
## 3      R                                    NA     NA      NA      NA
## 4      R                                    NA     NA      NA      NA
## 5      R                                    NA     NA      NA      NA
##   Entrance.Type Entry Exit.Only Vending Staffing Staff.Hours   ADA ADA.Notes
## 1         Stair   YES               YES     NONE             FALSE
## 2         Stair   YES               YES     FULL             FALSE
## 3         Stair   YES               YES     FULL             FALSE
## 4         Stair   YES               YES     FULL             FALSE
## 5         Stair   YES               YES     FULL             FALSE
##   Free.Crossover North.South.Street East.West.Street Corner Entrance.Latitude
## 1          FALSE            4th Ave          25th St     SW          40.66049
## 2          FALSE            4th Ave          25th St     SE          40.66032
## 3           TRUE            4th Ave          36th St     NW          40.65468
## 4           TRUE            4th Ave          36th St     NE          40.65436
## 5           TRUE            4th Ave          36th St     NW          40.65449
##   Entrance.Longitude        Station.Location       Entrance.Location
## 1         -73.99822 (40.660397, -73.998091)  (40.660489, -73.99822)
## 2         -73.99795 (40.660397, -73.998091) (40.660323, -73.997952)
## 3         -74.00431 (40.655144, -74.003549) (40.654676, -74.004306)
## 4         -74.00411 (40.655144, -74.003549) (40.654365, -74.004113)
## 5         -74.00450 (40.655144, -74.003549)  (40.65449, -74.004499)
```

03) NYC Bus Stops:

```
##   CounDist BoroCD AssemDist                                    the_geom
## 1       34    301        53  POINT (-73.94783099999995 40.70681200000007)
## 2       34    301        50  POINT (-73.94516199999998 40.71909700000003)
## 3       34    301        53 POINT (-73.94578299999995 40.702976000000035)
## 4       34    301        50  POINT (-73.94083599999993 40.72019500000005)
## 5       35    302        50  POINT (-73.96979499999998 40.69344000000007)
##   CongDist StSenDist SHELTER_ID      LOCATION AT_BETWEEN LONGITUDE LATITUDE
## 1        7        18     BR0003   MONTROSE AV LORIMER ST -73.94783 40.70681
## 2       12        18     BR0014     GRAHAM AV HERBERT ST -73.94516 40.71910
## 3        7        18     BR0026      BROADWAY LEONARD ST -73.94578 40.70298
## 4       12        18     BR0028  KINGSLAND AV HERBERT ST -73.94084 40.72019
## 5        8        25     BR0040 VANDERBILT AV  MYRTLE AV -73.96980 40.69344
##   AssetID BoroCode BoroName            Street SegmentID PhysicalID NODEID
## 1       1        3 Brooklyn   MONTROSE AVENUE     31244      91546      0
## 2       2        3 Brooklyn     GRAHAM AVENUE     35598      45754      0
## 3       3        3 Brooklyn          BROADWAY     31350      43972  19920
## 4       4        3 Brooklyn  KINGSLAND AVENUE     65844      48893  40551
## 5       5        3 Brooklyn VANDERBILT AVENUE     30196      60512      0
```

04) DCLA Cultural Organizations:

```
##           Organization.Name                    Address     City State Postcode
## 1 122 Community Center Inc.           150 First Avenue New York    NY    10009
```

```
## 2      13 Playwrights, Inc. 195 Willoughby Avenue, #402 Brooklyn    NY    11205
## 3                 1687, Inc.                 PO Box 1000 New York    NY    10014
## 4          18 Mai Committee 832 Franklin Avenue, PMB337 Brooklyn    NY    11225
## 5  20/20 Vision for Schools        8225 5th Avenue #323 Brooklyn    NY    11209
##     Main.Phone..                   Discipline             Council.District
## 1 (917) 864-5050                            Manhattan Council District #2
## 2 (917) 886-6545                      Theater Brooklyn Council District #39
## 3 (212) 252-3499 Multi-Discipline, Performing Manhattan Council District #3
## 4 (718) 270-6935 Multi-Discipline, Performing Brooklyn Council District #33
## 5 (347) 921-4426                  Visual Arts Brooklyn Council District #43
##                Community.Board   Borough Latitude Longitude Census.Tract
## 1                             Manhattan 40.72826 -73.98479           34
## 2  Brooklyn Community Board #6  Brooklyn 40.69205 -73.96418          193
## 3 Manhattan Community Board #2 Manhattan       NA        NA           NA
## 4  Brooklyn Community Board #6  Brooklyn 40.66946 -73.95842          213
## 5 Brooklyn Community Board #10  Brooklyn 40.62408 -74.02484          142
##       BIN        BBL                 NTA
## 1 1005894 1004370001        East Village
## 2 3054896 3019050080        Clinton Hill
## 3      NA         NA
## 4 3029691 3011870049 Crown Heights South
## 5 3152153 3060090001           Bay Ridge
```

05) NYPD Shooting Incidents:

```
## # A tibble: 5 x 19
##   INCIDENT_KEY OCCUR_DATE OCCUR_TIME BORO  PRECINCT JURISDICTION_CO~
##          <dbl> <chr>      <time>     <chr>    <dbl>            <dbl>
## 1    216972672 08/24/2020 15:11      BRONX       44                0
## 2    217953750 09/16/2020 16:50      MANH~       33                0
## 3    217540562 09/07/2020 02:49      BROO~       71                0
## 4    217773056 09/12/2020 00:40      BROO~       75                0
## 5    218560117 09/30/2020 21:07      BROO~       75                2
## # ... with 13 more variables: LOCATION_DESC <chr>,
## #   STATISTICAL_MURDER_FLAG <lgl>, PERP_AGE_GROUP <chr>, PERP_SEX <chr>,
## #   PERP_RACE <chr>, VIC_AGE_GROUP <chr>, VIC_SEX <chr>, VIC_RACE <chr>,
## #   X_COORD_CD <dbl>, Y_COORD_CD <dbl>, Latitude <dbl>, Longitude <dbl>, `New
## #   Georeferenced Column` <chr>
```

06) NYPD Calls for Service:

```
## # A tibble: 5 x 19
##   OBJECTID CAD_EVNT_ID CREATE_DATE INCIDENT_DATE INCIDENT_TIME NYPD_PCT_CD
##      <dbl>       <dbl> <chr>       <chr>         <chr>                <dbl>
## 1        1    65201967 01/13/2020  01/13/2020    12/30/1899              42
## 2        2    65211387 01/13/2020  01/13/2020    12/30/1899              42
## 3        3    65216018 01/13/2020  01/13/2020    12/30/1899              43
## 4        4    65209020 01/13/2020  01/13/2020    12/30/1899              44
## 5        5    65215853 01/13/2020  01/13/2020    12/30/1899              47
## # ... with 13 more variables: BORO_NM <chr>, PATRL_BORO_NM <chr>,
## #   GEO_CD_X <dbl>, GEO_CD_Y <dbl>, RADIO_CODE <dbl>, TYP_DESC <chr>,
## #   CIP_JOBS <chr>, ADD_TS <lgl>, DISP_TS <lgl>, ARRIVD_TS <lgl>,
## #   CLOSNG_TS <lgl>, Latitude <dbl>, Longitude <dbl>
```

07) Airbnb Listing Reviews:

```
## # A tibble: 5 x 19
```

```
##   OBJECTID CAD_EVNT_ID CREATE_DATE INCIDENT_DATE INCIDENT_TIME NYPD_PCT_CD
##      <dbl>       <dbl> <chr>       <chr>         <chr>                <dbl>
## 1        1    65201967 01/13/2020  01/13/2020    12/30/1899              42
## 2        2    65211387 01/13/2020  01/13/2020    12/30/1899              42
## 3        3    65216018 01/13/2020  01/13/2020    12/30/1899              43
## 4        4    65209020 01/13/2020  01/13/2020    12/30/1899              44
## 5        5    65215853 01/13/2020  01/13/2020    12/30/1899              47
## # ... with 13 more variables: BORO_NM <chr>, PATRL_BORO_NM <chr>,
## #   GEO_CD_X <dbl>, GEO_CD_Y <dbl>, RADIO_CODE <dbl>, TYP_DESC <chr>,
## #   CIP_JOBS <chr>, ADD_TS <lgl>, DISP_TS <lgl>, ARRIVD_TS <lgl>,
## #   CLOSNG_TS <lgl>, Latitude <dbl>, Longitude <dbl>
```

Sentiment Analysis

Please check "./Sentiment_Analysis" for the R scripts used to conduct our sentiment analysis on Airbnb listing reviews. The order of scripts and their respective functions is as follows:

1. reviews_files.R - used to create a new folder with txt files of reviews with names [review_id].txt (didn't include them, there are 1M+ of them)
2. files_processing_bing.R
3. file_processing_nrc.R
4. files_processing_afinn.R

All these scripts make up the sentiment analysis. The accuracy comparing with Monkeylearn (an online software that automatically conducts sentiment analysis) was 0.903, 0.907, 0.907. For the full analysis we used file_processing_nrc.R because it gave better results in terms of prediction of negative reviews.

5. checking_model.R - used to validate algorithms comparing with Monkeylearn prediction and threshold adjustments.

References

Magno, F., Cassia, F., Ugolini, M. M. (2018). "Accomodation prices on Airbnb: effects of host experience and market demand". TQM Journal. Volume 30 Number 5. Pages 608-620.

Li, J., Moreno, A. and Zhang, D.J. (2015), "Agent behavior in the sharing economy: evidence from Airbnb",Working paper, [1298] Ross School of Business, University of Michigan, Ann Arbor, MI.