AirBNB Pricing

C. Hill

12/13/2019

EXECUTIVE SUMMARY

We currently predict nightly rates for a new AirBNB property in NYC based on the average of the nightly rates for NYC. The goal of this project is to improve the RMSE of the current model.

Results: The RMSE for the validation set is **282.41** compared to 290.25, an improvement of 3%.

Model Summary: The model incorporates the following effects on AirBNB prices to make predictions: "Neighborhood Group" + "Minimum Nights" + "Room Type"

Though we expected these factors to have a larger impact on final price, it appears that property quality and size, factors not reflected in the data, have a larger impact on the final price of an AirBNB.

Recommendations: Incorporate the following information into the data: 1) A new metric such as a star rating system, in order to differentiate the quality of these properties 2) Information about size and number of beds.

This data could be added by adding questions when users list the property or by incorporating data such as the Zestimate and number of bedrooms from Zillow matching on address.

Introduction

This project uses data from the Kaggle database (https://www.kaggle.com/dgomonov/new-york-city-airbnb-open-data#AB_NYC_2019.csv) and includes AirBNB listings and locations for 2019 in NYC. Each row represents one property. We split this data into a training, test, and validation set to develop the algorithm.

Overview

Data Description: The data shows a list of AirBNB property listings in NYC for 2019 and includes 16 variables:

1) *id*: The id of the property
2) *name*: The name on the property listing
3) *host_id*: The id of the property host assigned by AirBNB
4) *host_name*: The name of the property host
5) *neighbourhood_group*: The area the property is located such as Manhattan or Queens
6) *neighbourhood*: A more localized description of the area
7) *latitude*: Latitude
8) *longitude*: Longitude
9) *room_type*: The type of room offered including three options: Private room, Entire home/apt, or Shared room
10) *Price*: The nightly rate of the listing
11) *Minimum_Nights*: The minimum number of nights allowed to be booked
12) *number_of_reviews*: The number of reviews on the property
13) *last_review*: The date of the last review
14) *reviews_per_month*: The average reviews per month
15) *calculated_host_listings_count*: The number of listings the host has
16) *availability_365*: The number of days per year the listing is available

For this project, we split 20% of the data off for a validation set. We then split the remaining data into a test and training set, with 20% in the test set and 80% in the training set. The training set has 31,288 lines and is summarized in tables 1:3.

Table 1: Train Set Summary

	id	name	host_id	host_name	neighbourhood_group
1	2539	Clean & quiet apt home by the park	2787	John	Brooklyn
3	3647	THE VILLAGE OF HARLEM....NEW YORK !	4632	Elisabeth	Manhattan
5	5022	Entire Apt: Spacious Studio/Loft by central park	7192	Laura	Manhattan
6	5099	Large Cozy 1 BR Apartment In Midtown East	7322	Chris	Manhattan
11	5295	Beautiful 1br on Upper West Side	7702	Lena	Manhattan

Table 2: Train Set Summary Cont.

	neighbourhood	latitude	longitude	room_type	price	minimum_nights	number_of_reviews
1	Kensington	40.64749	-73.97237	Private room	149	1	9
3	Harlem	40.80902	-73.94190	Private room	150	3	0
5	East Harlem	40.79851	-73.94399	Entire home/apt	80	10	9
6	Murray Hill	40.74767	-73.97500	Entire home/apt	200	3	74
11	Upper West Side	40.80316	-73.96545	Entire home/apt	135	5	53

Table 3: Train Set Summary Cont.

	last_review	reviews_per_month	calculated_host_listings_count	availability_365
1	10/19/2018	0.21	6	365
3		NA	1	365
5	11/19/2018	0.10	1	0
6	6/22/2019	0.59	1	129
11	6/22/2019	0.43	1	6

The data lists properties in 5 neighborhood groups with Manhattan and Brooklyn having the highest average property values. Half of properties fall between $50 to $200 and 93% of properties are priced below $300 per night with a long tail of high end properties at much higher rates (See Figures 1-2).

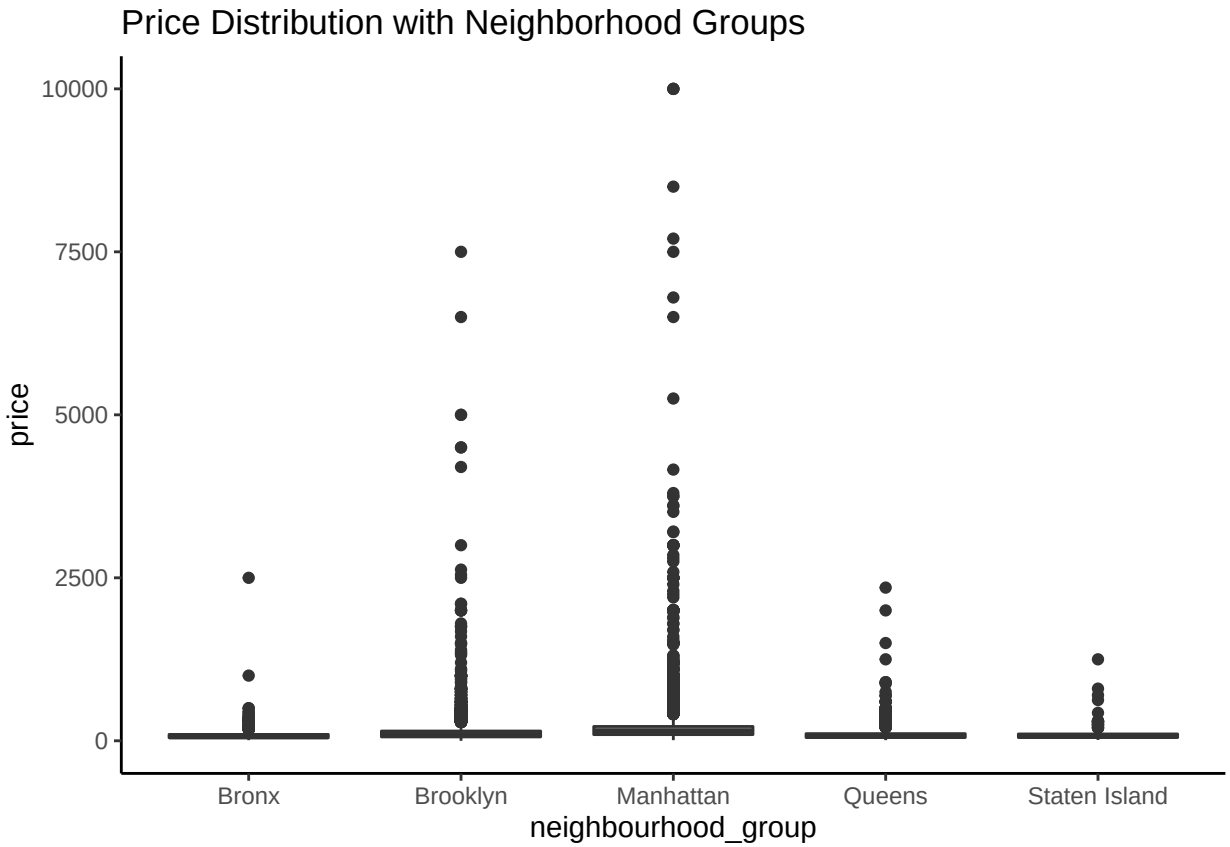


Figure 1: Boxplots show the large tail caused by high end properties available in NYC

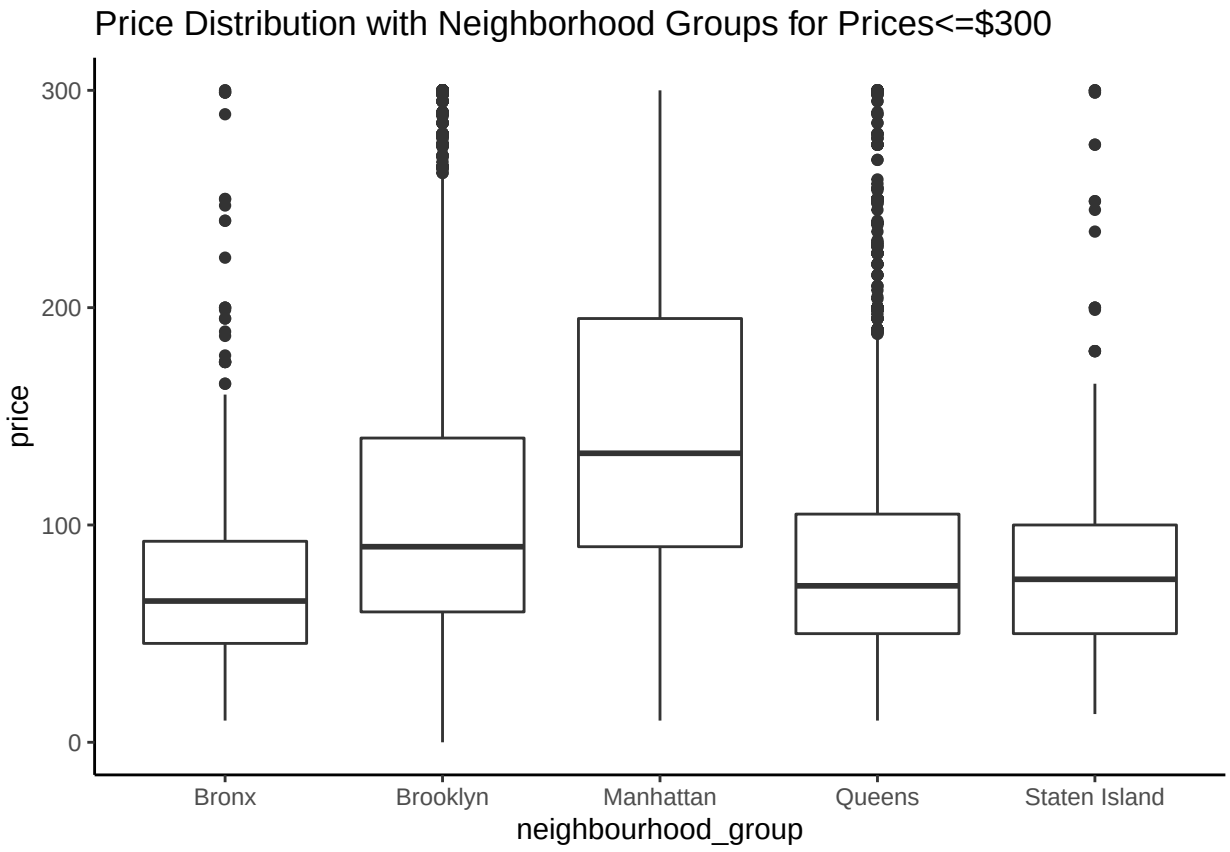


Figure 2: Boxplots show the large tail caused by high end properties available in NYC

Each neighbourhood group has 32 to 51 neighborhoods. For example, Allerton and Baychester are neighborhoods in the Bronx.

Table 4: Neighbourhood Summary

neighbourhood_group	n_neighbourhoods	n_properties
Bronx	48	717
Brooklyn	47	12939
Manhattan	32	13788
Queens	51	3604
Staten Island	40	240

AirBNB offers three categories of rooms: Private Room, Shared Room, and Entire Home/Apt. Each category has a different price point with "Shared Room" being the least expensive option and "Entire Home/Apt" being the most. Figure 3 shows the number of properties in each category.

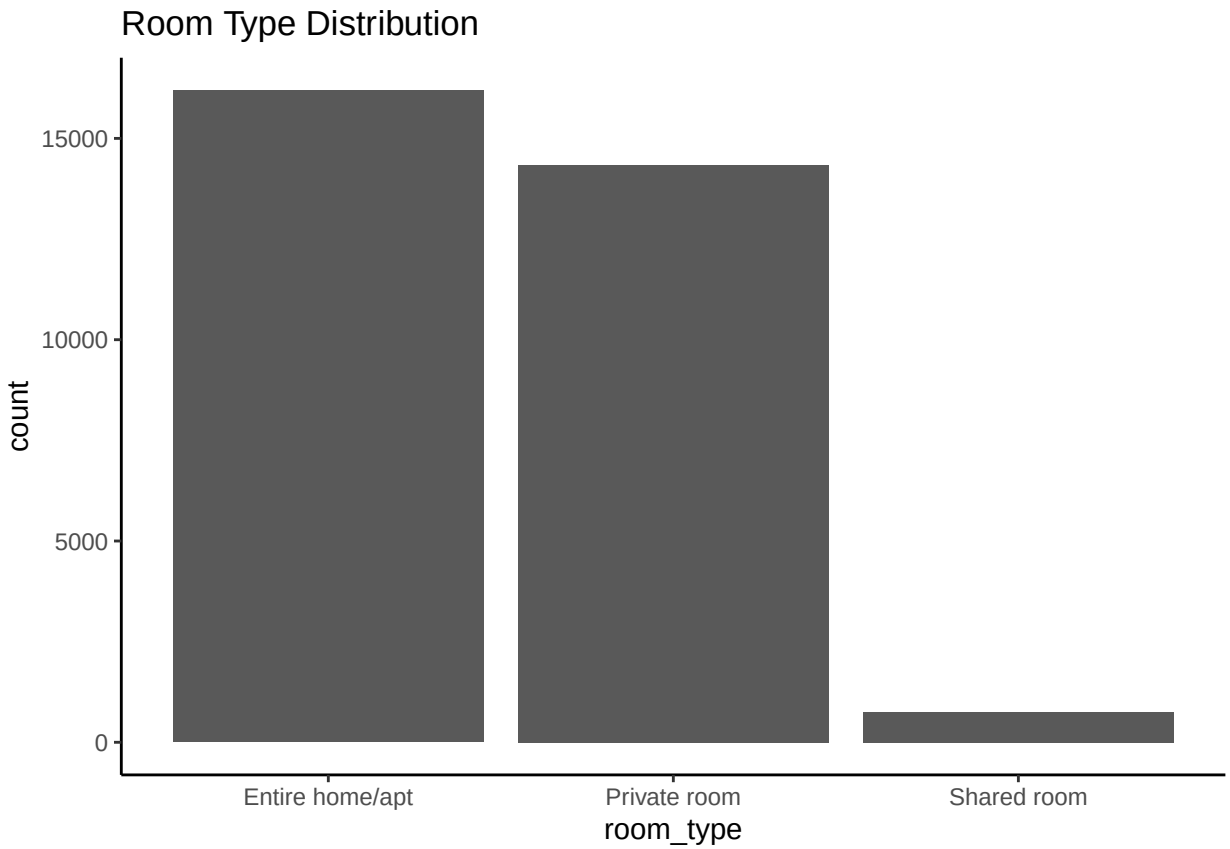


Figure 3: Shared Rooms are the least common type of listing.

The Minimum Nights shows the minimum number of nights the guest is required to book in order to stay at the property. Typically minimums are low at 1-5 nights with the exception of monthly stays. A few properties have minimum night requirements outside of those ranges.

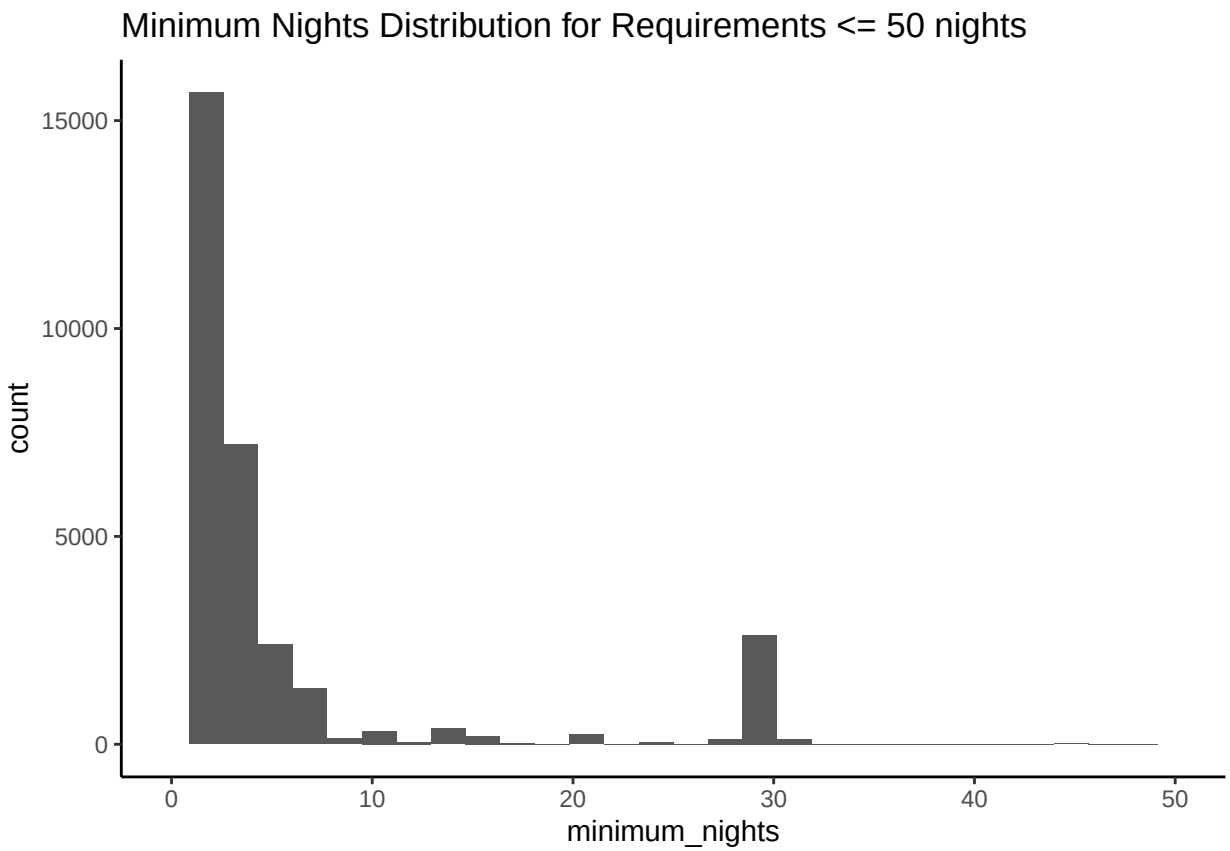


Figure 4: Minimum Night requirements are generally less than a week with a few properties with much higher requirements

Project Goal: The goal of this project is to improve the RMSE of price predictions currently based on the average price for NYC AirBNB properties.

Key Steps:

The key steps we take in order to derive the algorithm include the following:

1) *Create a Train and Test set:* We create a test, training, and validation set using the AirBNB data to cross validate the model without overtraining.

2) *Clean Data:* We check for NA fields in the data.

3) *Explore Insights:* We analyze Data correlations between elements of the cleaned data to help pinpoint where to focus the algorithm.

4) *Model Algorithms:* We try different algorithms on the training set.

5) *Evaluate Performance:* We use the algorithm to predict prices for the test set and then compare the predictions to the actual prices in the test set.

6) *Finalize Model:* The model producing the lowerst RMSE becomes our final model.

7) *Validate Results:* We test the final model by predicting prices for the validation set and comparing the predictions to the actual prices in the validation set.

Methods/Analysis

Data Cleaning:

1) Check data for NA values: The data is fairly complete with the reviews the only section with NA because many properties have not been reviewed.

```
                     id                            name
                      0                               0
                host_id                       host_name
                      0                               0
     neighbourhood_group                   neighbourhood
                      0                               0
               latitude                       longitude
                      0                               0
              room_type                           price
                      0                               0
         minimum_nights                number_of_reviews
                      0                               0
            last_review                reviews_per_month
                      0                            6483
```

calculated_host_listings_count availability_365 0 0

1) Create a test, training, and validation set: We use the caret package to partition the data first into a validation and train set, then we further partition the train set into a test and train set.

```
test_index<-createDataPartition(data$price,times=1,p = .2, list=FALSE)

validation_set<-data[test_index,]
train_set<-data[-test_index,]

test_index<-createDataPartition(train_set$price,times=1,p = .2, list=FALSE)

test_set<-train_set[test_index,]
train_set<-train_set[-test_index,]
```

Data Exploration, Visualization, and Insights:

Location Effect: With the common real estate mantra being "Location, Location, Location", we expect location data to have a large impact on room prices. As we see from figures 1 and 2 above, Manhattan has the highest average price, with Brooklyn a close second and the other three about equal. However, the high variability within each group shows that other factors also play a large role in price.

Length of Stay Effect: We expect length of stay to have a negative impact on price since longer stays typically have a discount over the regular price. Figure 5 show the length of stay compared to price faceted by Neighborhood Group for listings less than $300. In order to control for confounding from the room type, we filter the data for Private Rooms only. We see the data has a weak negative correlation (most visibly seen in Queens).

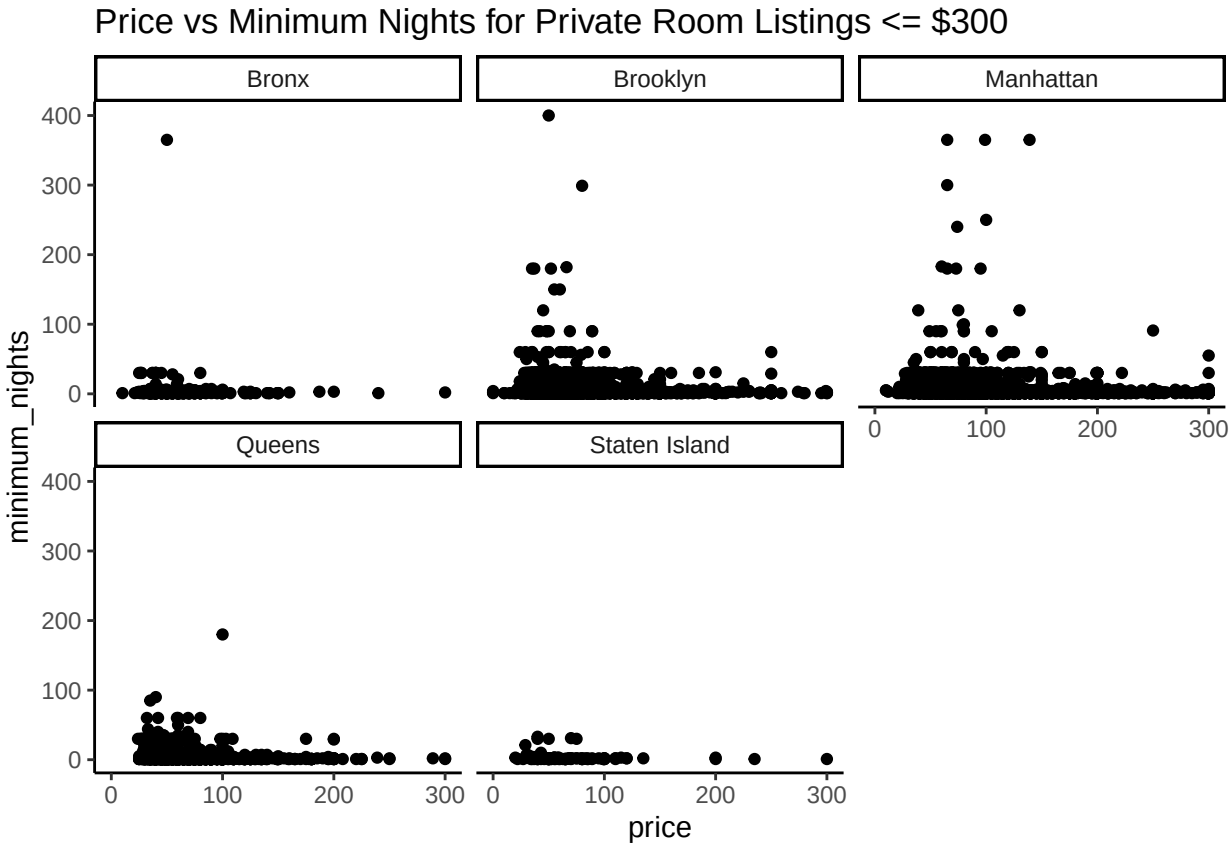


Figure 5: The data has a weak negative correlation (most visibly seen in Queens).

Room Type Effect: We expect the room type to affect the price with guests willing to spend more for an Entire Home/Apt than they are willing to spend on a Shared Room. Figure 6 shows boxplots with Entire Home/Apt having the highest mean price and Shared Room having the lowest. However, these boxplots also show variability due to other factors with some private rooms pricing higher than the mean for an Entire home/apt. We filter the data for Queens to limit some of the confounding due to location.

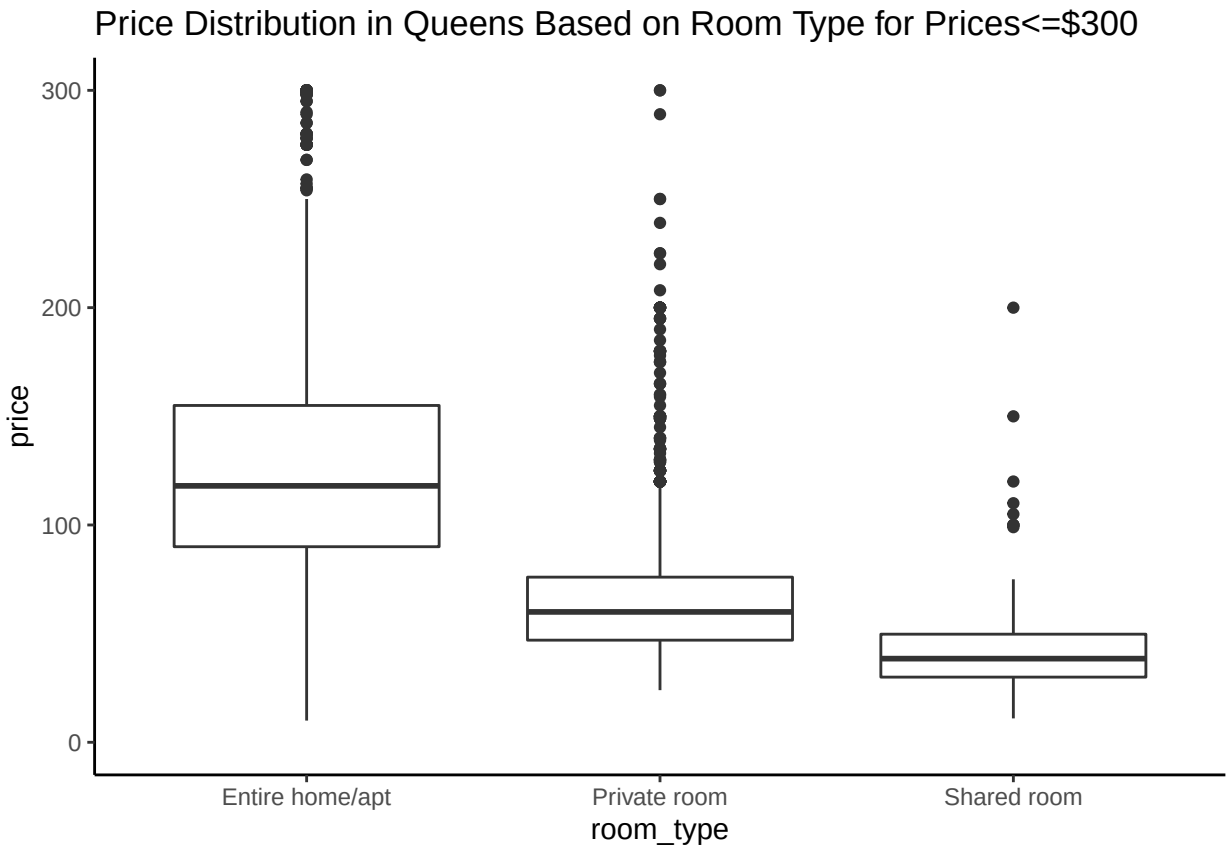


Figure 6: A Shared room has the lowest average price while an Entire Home/Apt has the highest.

Modeling Approach:

Base Model: The current base model predicts the price of each property based on the overall average price. Using this model, our Base RMSE for the test set is 193.80.

```
mu<-mean(train_set$price)
Base_RMSE<-RMSE(test_set$price, mu)
```

Linear Regression Model: We add location, room type, and length of stay affects using a linear regression model.

```
train_lm<-train(price ~ room_type+neighbourhood_group+minimum_nights,
          data = train_set, method = "lm")
preds<-predict(train_lm, test_set)
RMSE(test_set$price, preds)
```

We adjust this equation to try different combinations of predictors and different variables for location (latitude/longitude - neighbourhood is not used because many of the neighbourhoods have insufficient data to create a model) with the following results:

Table 5: RMSE Results Summary

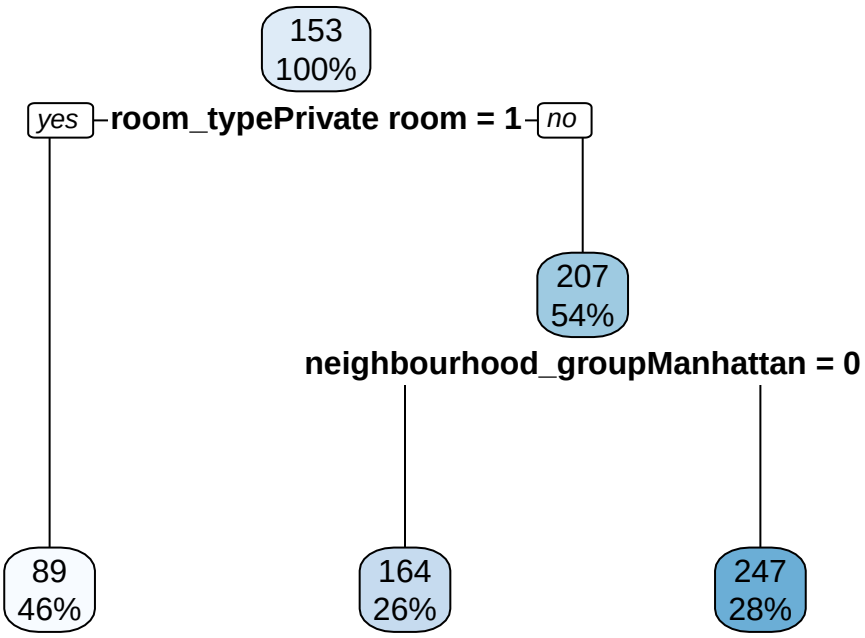
method	RMSE
Base Model	193.8022
Room Type Effect	184.6769
Room Type + Minimum Nights Effect	184.6209
Room Type + Minimum Nights + Latitude/Longitude Effect	182.8018
Room Type + Minimum Nights + Neighbourhood Group Effect	182.1630

RPart Model: We also try an RPART model to create a tree of effects.

```
train_rpart<-train(price ~ room_type+neighbourhood_group+minimum_nights,
                   data = train_set, method = "rpart")
preds<-predict(train_rpart, test_set)
RMSE_RPART<-RMSE(test_set$price, preds)
```

The rpart model gives the following tree:

```
rpart.plot(train_rpart$finalModel)
```



The model uses room type and neighbourhood group to determine price with an RMSE of 183.12, which is still higher that the RMSE of our previous model.

The RMSE is the lowest using the lm model with room type, minimum nights, and neighbourhood group (RMSE = 182.16) so this becomes our final model. This is a 6% improvement from the base RMSE of 193.80.

Results

Model Results: Our final model uses the following formula to predict AirBNB prices:

```
train_lm$finalModel
```

```
##
## Call:
## lm(formula = .outcome ~ ., data = dat)
##
## Coefficients:
##                           (Intercept)              `room_typePrivate room`
##                             160.6854                              -111.8938
##              `room_typeShared room`         neighbourhood_groupBrooklyn
##                            -134.2347                               22.1368
##       neighbourhood_groupManhattan           neighbourhood_groupQueens
##                              80.2729                                6.2729
## `neighbourhood_groupStaten Island`                         minimum_nights
##                               2.8896                                0.1934
```

The final model gives an RMSE of 182.16 which is a 6% improvement from the base model's 193.80 RMSE. On average, our model differs from the actual price by $182. If we look at the top 10 greatest errors in prediction, we get the result in Table 6 and 7.

We see that the errors occur due to the skewed data from luxury properties. The largest error is a luxury townhouse in Greenwich Village that rents for $6,000 per night. There is also an event space that is likely much larger than the other properties that rents for $5,000 per night. In order to tease out these differences we would need to add additional information to the dataset. This could be acheived by creating a star grading system similar to hotel stars that indicates the level of luxury the property offers along with square footage and number of beds. Another option could be to add address data to this data set and match it to a Zillow Zestimate to gauge the level of luxury and add in square footage and number of bedrooms. Based on a quick analysis where we split the properties into 5 groups based on price, we expect we could improve the RMSE by at least 30% by adding a luxury metric up with 5 groupings.

Table 6: Top Error Summary

name	neighbourhood_group	minimum_nights
Luxury townhouse Greenwich Village	Manhattan	1
Midtown Manhattan great location (Gramacy park)	Manhattan	30
4-Floor Unique Event Space 50P Cap. - #10299B	Manhattan	1
Beautiful private Brooklyn room with kitchenette	Brooklyn	114
Greenwich Village Townhome with Private Garden!	Manhattan	30
LUXURIOUS 5 bedroom, 4.5 bath home	Manhattan	1
Amazing Chelsea 4BR Loft!	Manhattan	30
Modern Townhouse for Photo, Film & Daytime Events	Manhattan	1
Columbus Circle and Park Views	Manhattan	1
Ultimate 50th Floor Downtown Penthouse - 4000SqFt	Manhattan	2

Table 7: Top Error Summary Cont.

room_type	price	pred
Entire home/apt	6000	241.15169
Entire home/apt	5100	246.75943
Entire home/apt	5000	241.15169
Private room	4200	92.97265
Entire home/apt	3900	246.75943
Entire home/apt	2999	241.15169
Entire home/apt	2995	246.75943
Entire home/apt	2900	241.15169
Entire home/apt	2695	241.15169
Entire home/apt	2250	241.34506

Model Performance:

We run the model on our validation set in order to determine performance.

```
preds<-predict(train_lm, validation_set)
RMSE(validation_set$price, preds)
```

```
## [1] 282.4124
```

Our final RMSE is **282.41** compared to 290.25, an improvement of 3%.

Conclusion

Summary: Though our model does improve the predictions compared to the base by 3%, additional information will need to be added to the data in order to improve the model and make it useful for giving meaningful predictions about price. Currently, the model will error by $282 on average when making predictions. Most of the error occurs with high priced luxury properties.

Recommendations: Incorporate the following information into the data: 1) A new metric such as a star rating system, in order to differentiate the quality of these properties 2) Information about size and number of beds.

This data could be added by adding questions when users list the property or by incorporating data such as the Zestimate and number of bedrooms from Zillow matching on address.