Predicting Airbnb Price by Neighborhood in New York City

Lucas Fonda, Neil Shah, Matthew Chen

1. INTRODUCTION

The motivation behind our research was to determine how the price per night of Airbnbs differed depending on where the Airbnb was located, and a variety of other variables. We have all had firsthand experiences with staying in Airbnbs and wondered if there were any significant factors that may change the actual price per night. Not only this, but we were interested in how the dynamics of the New York boroughs, with differing levels of development, and other factors had on the price of Airbnbs. Do things like population density and how developed the borough and the neighborhoods in general play a role in how much Airbnbs renters charge?

The five boroughs represented in New York are Brooklyn, Queens, Manhattan, The Bronx, and Staten Island, with populations 2,736,074, 2,405,464, 1,694,251, 1,472,654, and 495,747, respectively. The median house incomes greatly differ between each borough, where it is $85,071 for Manhattan, $37,397 for the Bronx, $56,942 for Brooklyn, $64,509 for Queens, and $79,201 for Staten Island. Just looking at a basic statistic such as median income price shows how dissimilar these boroughs are. We aimed to delve into actual statistics regarding Airbnb prices for these differing neighborhoods within the boroughs, as well as the boroughs themselves. Our actual hypothesis was this: Boroughs with higher development, meaning higher median household salaries, will on average have higher Airbnb price rates per night. Ones that are less developed will tend to have lower Airbnb price rates per night.

2. DATA SOURCE & METHODS

Our dataset came from Kaggle.com, and it was originally much larger than the final dataset we used for our analysis. The dataset is described on Kaggle.com as: “Listings, including full descriptions and average review score Reviews, including unique id for each reviewer and detailed comments Calendar, including listing id and the price and availability for that day.” The dataset had several missing values for a large amount of properties, so we removed any property with missing information. This still left us with about 17,000 observations so we were comfortable proceeding with this data. We then removed several variables that were not useful to us, some being the host’s name, latitudinal and longitudinal coordinates of the property, the country, and more. We kept the relevant variables such as number of reviews, minimum nights, construction year, neighborhood, and more. Finally, we had a very large number of properties with over 50 reviews, so we filtered the data to only include properties with at least 50 reviews.

We also made some adjustments to our variables when assessing model assumptions, and found that using log price instead of price helped to fit the data to a more normal distribution. Centering our variable for Minimum Nights and Number of Reviews also helped with this, so in our final analysis we will be using the centered variables and analyzing log price.

We felt that this data had a multilevel structure, since each individual property was nested inside the different neighborhoods in New York City. For this reason, we decided to analyze this data using a multilevel regression model. We experimented with different models, incorporating random intercepts and random slopes for certain predictors to help see if we could explain more variability in price per night of New York Airbnb's. We will describe the different models we tried using and how they performed in the following section.

3. RESULTS

Our exploratory analysis began with creating boxplots to examine the variability in Airbnb price per night across our different variables. From Figure 1 below, we found that there was lots of variability in price across the neighborhoods, so we wanted to explore the variability explained by the Neighborhood Group. This will be our level 2 variable, as it categorizes each of the neighborhoods into a larger subcategory of New York based on location.

Figure 1: Boxplot of Price by Neighborhood

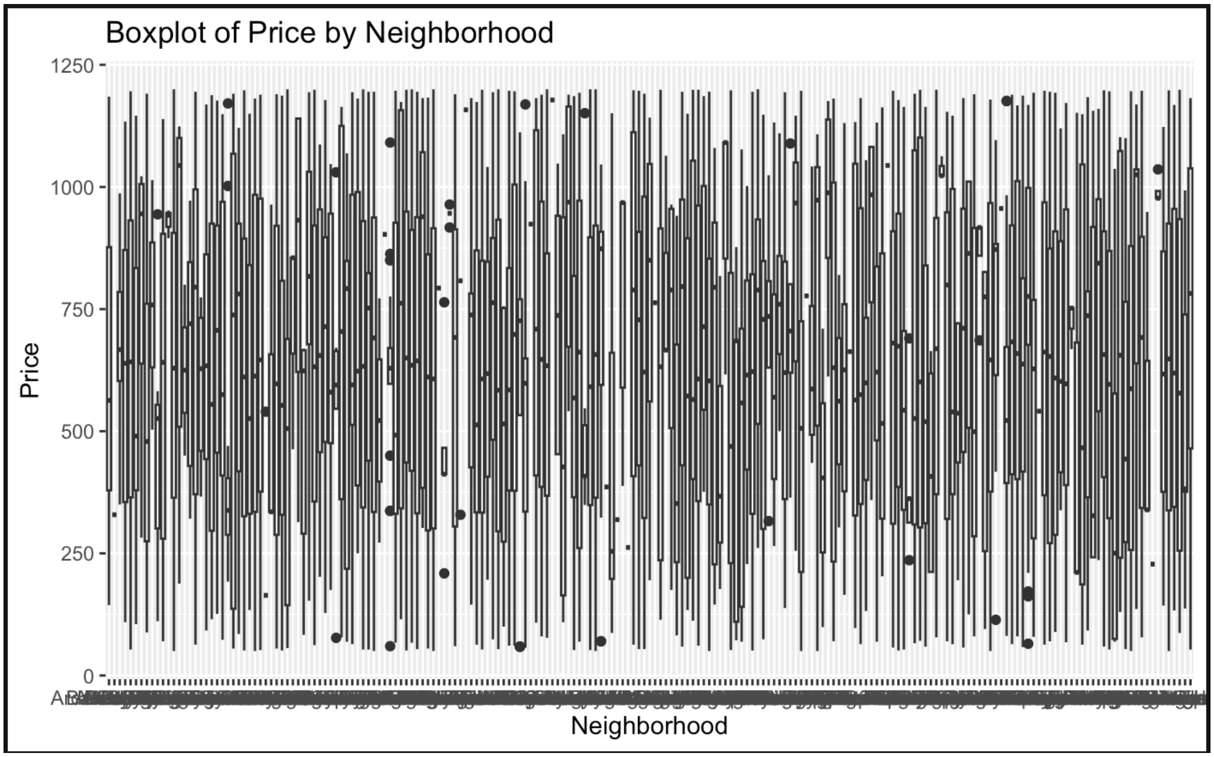


Figure 2: OLS lines of Price by Construction Year across Neighborhood

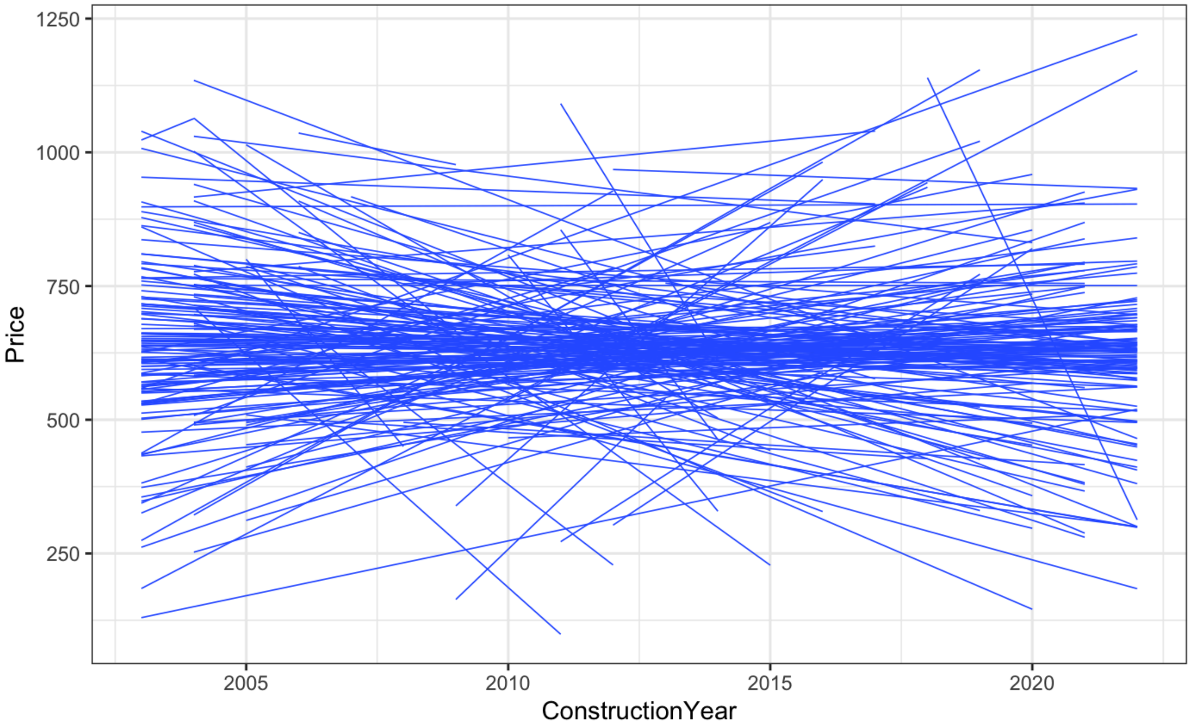
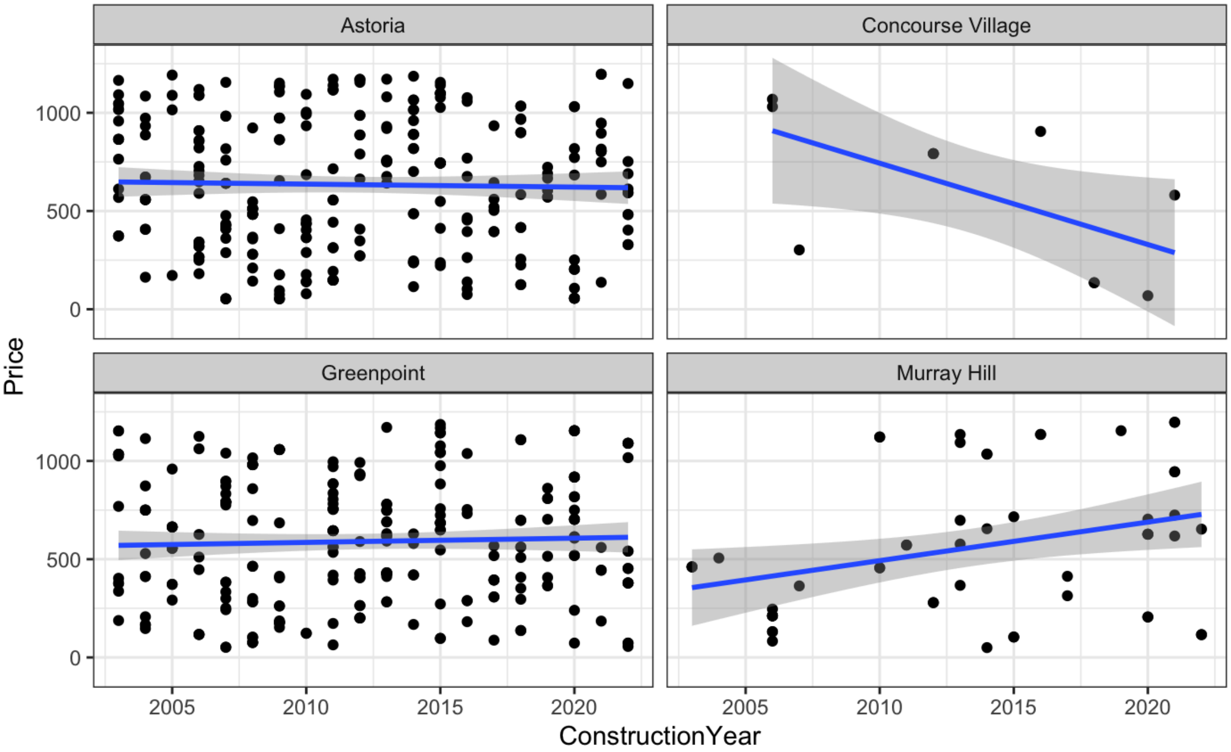


Figure 2 represents the ordinary least squares lines for construction year vs price, with each neighborhood representing a line. The graph illustrates the change in Airbnb prices among different neighborhoods based on the respective construction years of the properties. Some neighborhoods show a negative correlation between prices and construction year, while others exhibit a positive correlation. Consequently, we aim to delve deeper into understanding the variability in prices attributed to the construction year factor. It appears that construction year may be a significant level 1 factor, after viewing the graph. As well, neighborhood may be a significant level 2 factor, as the slopes differ greatly depending on what neighborhood you are in.

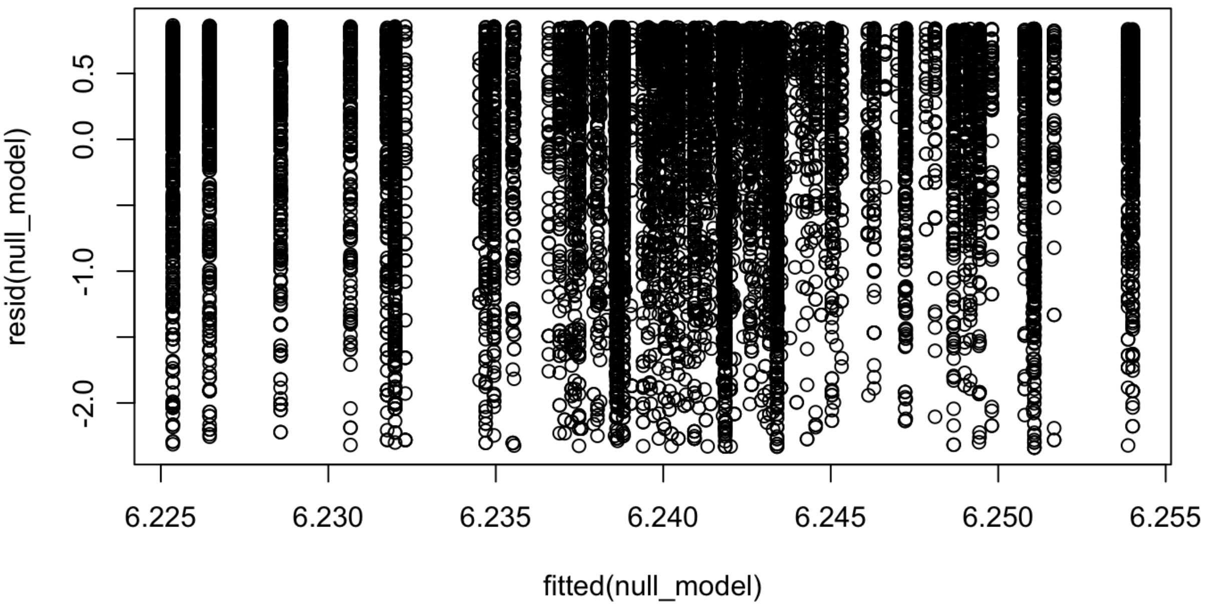
Figure 3: Price by Construction Year across four Neighborhoods



In Figure 3, our beliefs were further confirmed in seeing that the variability in Airbnb price per night seems to change across construction year, depending on which neighborhood we are looking into. Along with Neighborhood at level 2 and Construction Year at level 1, we also felt like adding the level 1 variables of centered Minimum Nights and centered Number of Reviews would help explain some variability in price. Our idea was that a house with lots of positive reviews would have a higher price, or lower price if there are many negative reviews. Additionally, we thought it would be cheaper per night if the minimum required nights to stay in the Airbnb was higher.

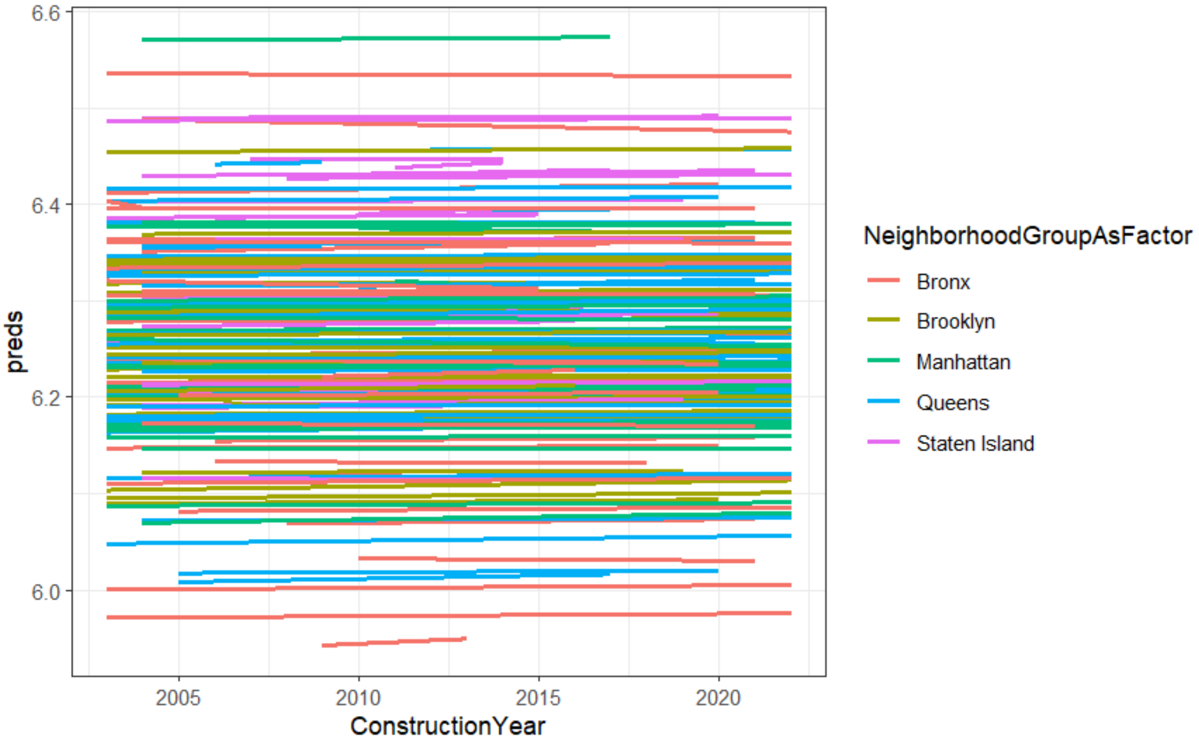
First, we created our null model which contains random intercepts for the individual neighborhoods. The response is the log price for the airbnb's per night. As discussed in the data source and methods section, we used a log transformation in order to satisfy the model assumptions. τ^2 is about 0.0004 while σ^2 is about 0.53. Our ICC is almost equal to zero (0.0007). A low ICC value suggests that there is not much variation between the groups compared to the variability within the groups. In context, in this model there is very little variation that is explained by the neighborhoods, or the level 2 variable. This may change after adding our level 1 variables, as they may explain some of the variation at the neighborhood level.

Figure 4: Residual vs. Fitted Plot for Null Model



Based on the figure 4 above, it appears that the linearity and equal variance assumptions are met. The residuals appear to be scattered quite randomly, and there does not appear to be any curved pattern. With our null model not explaining much variability in price per night, we decided to add the level 1 variables, that we thought were useful, into the next model. We hoped to see our predictors of Construction Year, Minimum Nights, and Number of Reviews. In this new model, we included centered Minimum Nights, centered Number of Reviews, Construction Year, and Neighborhood Group as fixed effects, while also keeping the random intercepts across Neighborhood. We also added random slopes for Construction Year in this model, as some neighborhoods have positive slopes and some have negative slopes in Figure 2 from above. Looking at a graph of our model, Figure 5 below, we see that the predicted price of an Airbnb does not change much across construction year, regardless of the neighborhood that the property is in.

Figure 5: Predicted Price by Construction Year Across Neighborhood for Random Slopes Model



After noting that the random slopes for construction year were not valuable to us, we decided to remove them from the model. In doing this, we found that Construction Year was not explaining much level 1 or level 2 variability and it had an insignificant p-value. As a result, we dropped it from the model entirely.

Table 1: Summary Output for Random Intercepts Model including Construction Year

	Estimate	t value
Intercept	6.244	-
NumReviews_cen	5.509e-05	0.672
MinNights_cen	-4.629e-04	-0.936
NeighborhoodGroupBrooklyn	7.950e-04	0.025
NeighborhoodGroupManhattan	-1.745e-02	-0.537
NeighborhoodGroupQueens	8.378e-03	0.246
NeighborhoodGroupStatenIsland	1.040e-01	1.825

τ^2 is about 0.0001 while σ^2 is about 0.53.

We were able to draw some conclusions from our final model, the output is shown in Table 1 above. For every one review increase in the average number of reviews, there is a predicted increase of 5.51e-05 in the log price per night of an Airbnb, when the number of minimum nights are at their average and after adjusting for the different boroughs. For every one night increase in the average number of minimum nights stayed, there is a predicted decrease of 4.63e-04 in the log price per night of an Airbnb, when the number of reviews are at their average and after adjusting for the different boroughs. After keeping the number of reviews of the Airbnb at its average, as well as the minimum nights that you must stay at its average, the predicted log price when staying in Brooklyn is 7.95e-04 higher than staying in the Bronx. After keeping the number of reviews of the Airbnb at its average, as well as the minimum nights that you must stay at its average, the predicted log price when staying in Manhattan is 1.75e-02 lower than staying in the Bronx. After keeping the number of reviews of the Airbnb at its average, as well as the minimum nights that you must stay at its average, the predicted log price when staying in Queens is 8.38e-03 higher than staying in the Bronx. After keeping the number of reviews of the Airbnb at its average, as well as the minimum nights that you must stay at its average, the predicted log price when staying in Staten Island is 7.95e-04 higher than staying in the Bronx.

72.19% of the unexplained level 2 variation in the null model is explained by our final model. 0.013% of the unexplained level 1 variation in the null model is explained by our final model. In context, this does make sense as none of our level 1 variables that we put into the model are significant. It appears that adding neighborhood group as a fixed effect is effective in explaining quite a bit of the level 2 variation in the log price per night. The log price per night varies greatly depending on what neighborhood you are in. It is also quite evident from the figure that the neighborhood groups have a large impact on

Figure 6 below shows us that our final model does an adequate job of meeting all model assumptions. Our data is roughly linear, has equal variance, and the random neighborhood effects follow a normal distribution. Our data does not perfectly fit a normal distribution, evident in the residual vs. fitted plot of our final model, but we proceeded with the analysis even with this violation. In Figure 7, the effects plot of centered Minimum Nights and Number of Reviews show that there is much less variability in our log price response at the smaller values of both predictors. This is because we have many more observations at these levels and can model the variability much better with many points, compared to only a handful of observations at larger values. We also see that on average, the log price of Airbnbs is higher in Staten Island, which coincides with Staten Island having the highest t-value in our final model.

Final Model Equations:

Level 1: $log(price)_{ij} = \beta_{0j} + \beta_1 cen\_MinNights_j + \beta_2 cen\_NumReviews_j + \varepsilon_{ij}$

$\varepsilon_{ij} \sim N(0,\sigma^2)$

Level 2: $\beta_{0j} = \beta_{00} + \beta_{01} NeighborhoodGroup + u_{0j}$

$u_{0j} \sim \mathrm{N}(0,\tau^2\ )$

Figure 6: Collection of Plots that Check Model Assumptions

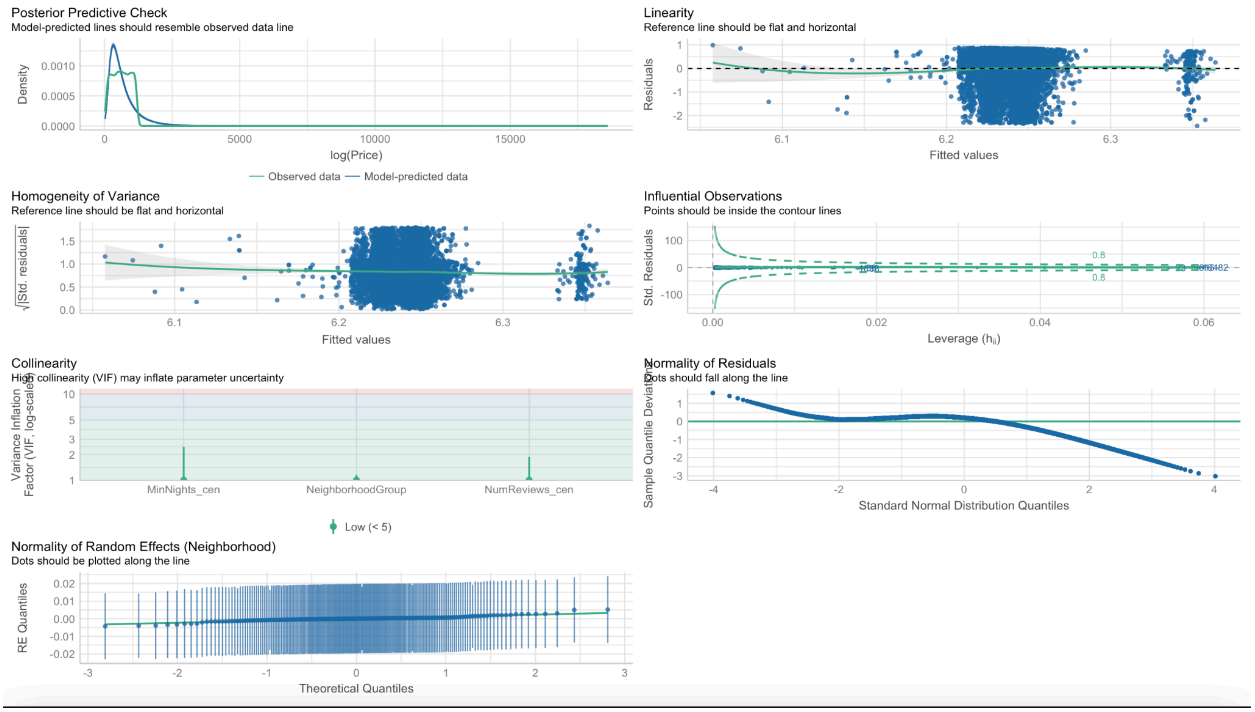
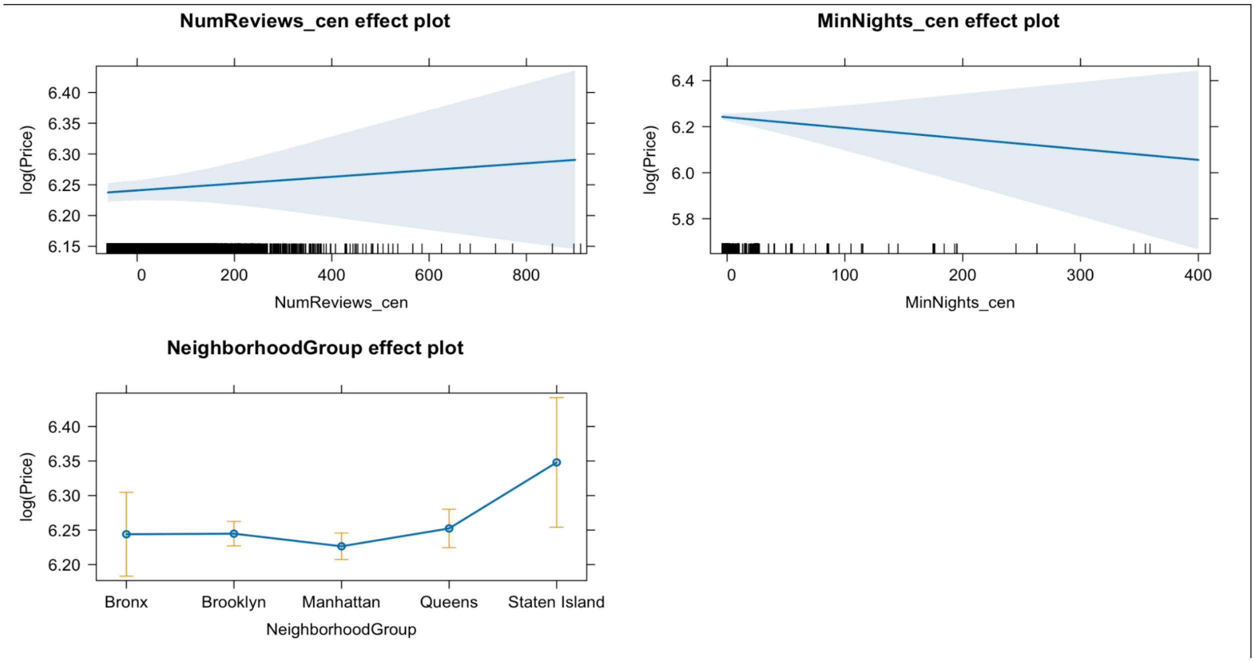


Figure 7: Effects Plots of Final Model



4. DISCUSSION

Our research question aimed to discover how the different boroughs and other factors impacted the price per night of Airbnbs. From the results, it appears that there were obvious differences in the boroughs relationship with price for Airbnbs. Not only this, but other factors appeared to play a role in helping explain the differences between the neighborhoods, specifically the number of reviews for the Airbnbs and the minimum number of nights that you must stay.

We've learned that there is a lot of variation within each borough, and that there was no clear pattern on Airbnb prices in each of them. This being said, our original hypothesis seemed to be slightly confirmed as we saw that Staten Island had higher predicted prices than the other boroughs and it had the second highest median salary. Manhattan had the largest median salary, but Manhattan has a population of roughly 1.6 million people while Staten Island has about 500,000 residents. This large difference in population size can explain why Staten Island was the borough with the highest predicted price in our model. We did some additional research into these boroughs, and found that the cost of living in Staten Island is lower than Manhattan and Brooklyn(Roadmap Moving). We found it interesting that the Airbnb rental price is high when the cost of living is lower. Staten Island has been investing a lot more money into building attractions, which may indicate that they are looking to drive up the cost of living in the area.

Some limitations of our study are, obviously, that we cannot make larger conclusions about Airbnb prices outside of the New York City area. Our study does not represent a broader analysis of the relationship between our predictor variables on price per night because we don't have data outside of New York City. Another noteworthy point is that the centered variables for Minimum

Nights and Number of Reviews have VIF values around 2, but we did not think this was a major concern.

Our analysis was able to explain roughly 72% of the variation at the neighborhood level, which could be improved but we were satisfied with having explained at least this much of the variability in price per night. Our model did not explain any variability at the individual property level, so there is work to be done there. We also did not have any significant predictors in our model, based off of the t-values of each variable in our final model. To improve this report in the future, we would like to consider more variables that describe each property, such as number of rooms. We would hope that some of these would help explain variability in price per night at the individual property level.

5. APPENDIX

Our code for creating the dataset "airbnb2":

```
airbnb <- read.csv("/Users/matthewchen/Downloads/Airbnb_Open_Data 2 (1).csv", header = TRUE)

airbnb2 <- airbnb %>%
  filter(NumReviews > 50)
```

Our Final model:

```
modelbnb5 <- lmer(log(Price) ~ NumReviews_cen + MinNights_cen + NeighborhoodGroup + (1| Neighborhood), data = airbnb2, REML = F)
```

Description of variables in the final model:

NumReviews_cen: a level 1 centered variable of the number of reviews an individual Airbnb has.

MinNights_cen: a level 1 centered variable of the minimum nights' requirement in an individual Airbnb.

Neighborhood Group: The five boroughs where all the neighborhoods listed are coming from. It is a level 2 variable.
Neighborhood: A level 2 unit that lists out which neighborhood an individual Airbnb belongs to.

log(Price): predicted price per night of an individual Airbnb that has been log-transformed.

Description of variable that has been used in the other models:

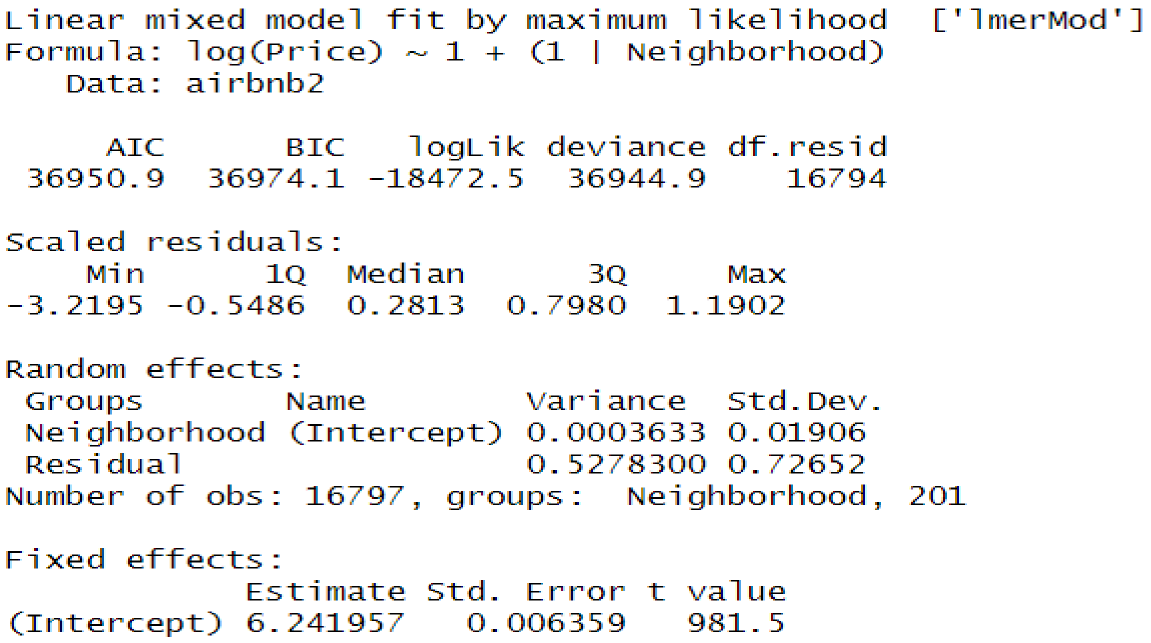
Construction Year: the year that a home used for Airbnb has been built.

Link to the data file being used after removing several variables:
https://docs.google.com/spreadsheets/d/1_LTXKQ92-sLNBjrXwEh6HZ3GoKWe4qOX0s2D__mAja0/edit?usp=sharing

Link to the website of the original data file being downloaded:

https://www.kaggle.com/datasets/arianazmoudeh/airbnbopendata

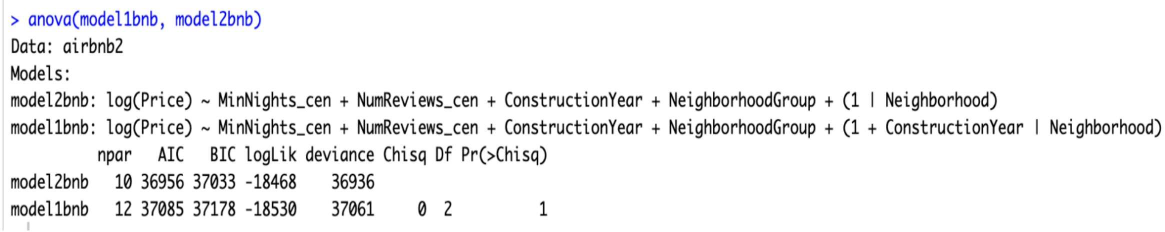
Null model:



```
Linear mixed model fit by maximum likelihood  ['lmerMod']
Formula: log(Price) ~ 1 + (1 | Neighborhood)
   Data: airbnb2

     AIC      BIC   logLik deviance df.resid
 36950.9  36974.1 -18472.5  36944.9    16794

Scaled residuals:
    Min      1Q  Median      3Q     Max
-3.2195 -0.5486  0.2813  0.7980  1.1902

Random effects:
 Groups       Name        Variance  Std.Dev.
 Neighborhood (Intercept) 0.0003633 0.01906
 Residual                 0.5278300 0.72652
Number of obs: 16797, groups:  Neighborhood, 201

Fixed effects:
            Estimate Std. Error t value
(Intercept) 6.241957   0.006359   981.5
```

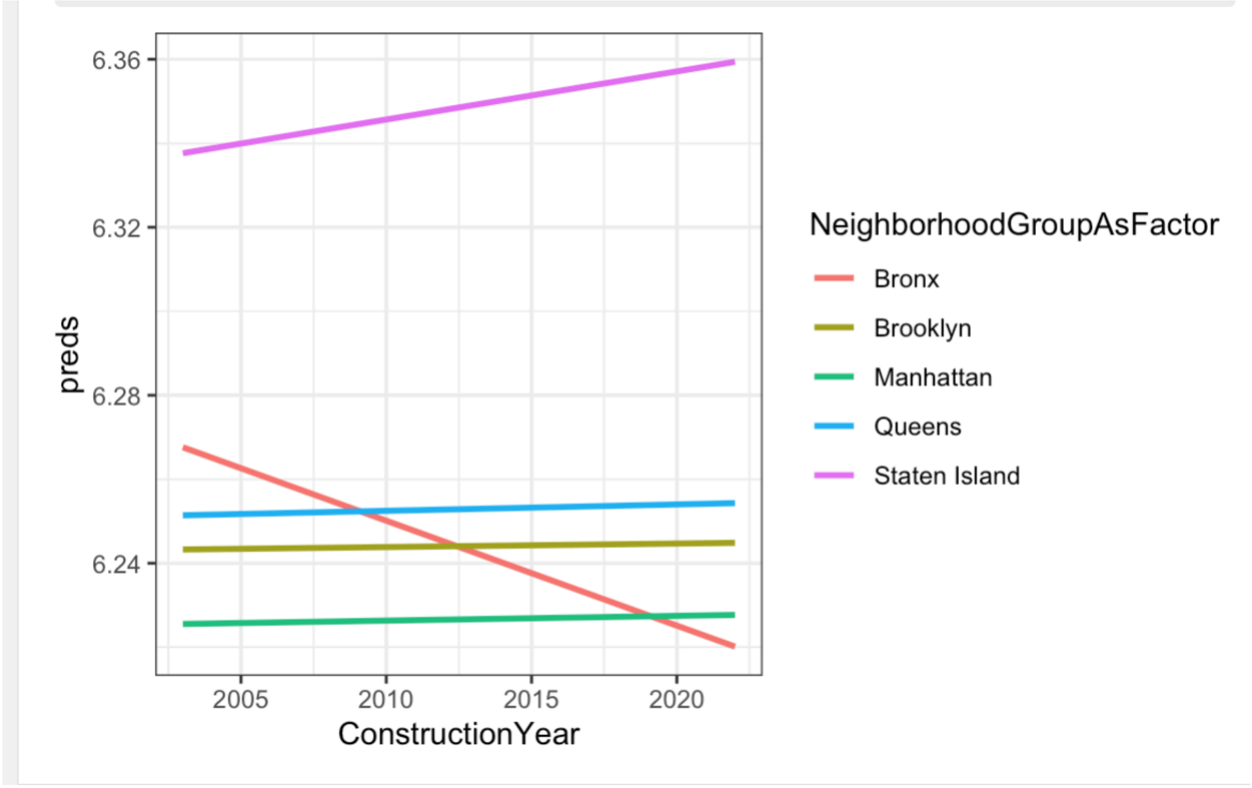
The above is our null model with the neighborhood as the random intercept. As we have seen in the above graph, there are no other variables in the model other than the random intercept of neighborhoods and there is more variation at level 1 than at level 2 (0.5278300 vs. 0.0003633). Our goal is to find a model that could explain the most variation at level 2 compared to the null model.

Comparison of two models with likelihood ratio test:

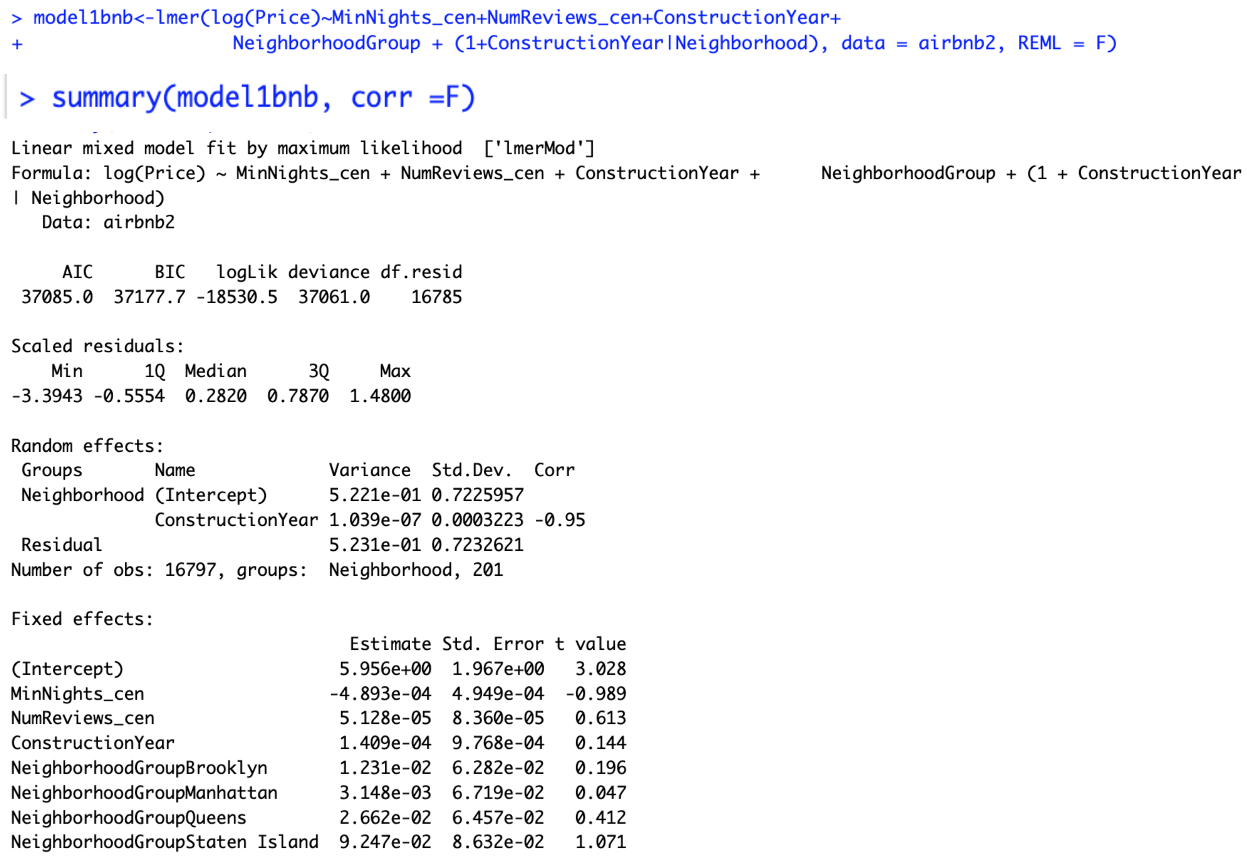


```
> anova(model1bnb, model2bnb)
Data: airbnb2
Models:
model2bnb: log(Price) ~ MinNights_cen + NumReviews_cen + ConstructionYear + NeighborhoodGroup + (1 | Neighborhood)
model1bnb: log(Price) ~ MinNights_cen + NumReviews_cen + ConstructionYear + NeighborhoodGroup + (1 + ConstructionYear | Neighborhood)
          npar   AIC   BIC logLik deviance Chisq Df Pr(>Chisq)
model2bnb   10 36956 37033 -18468    36936                    
model1bnb   12 37085 37178 -18530    37061     0  2          1
```

We are comparing the model with random slopes of construction year to the model that does not have random slopes of construction year using the likelihood ratio test. Due to a very large p-value, we have found out that the random slope of the construction year is not very useful to be added to the model.

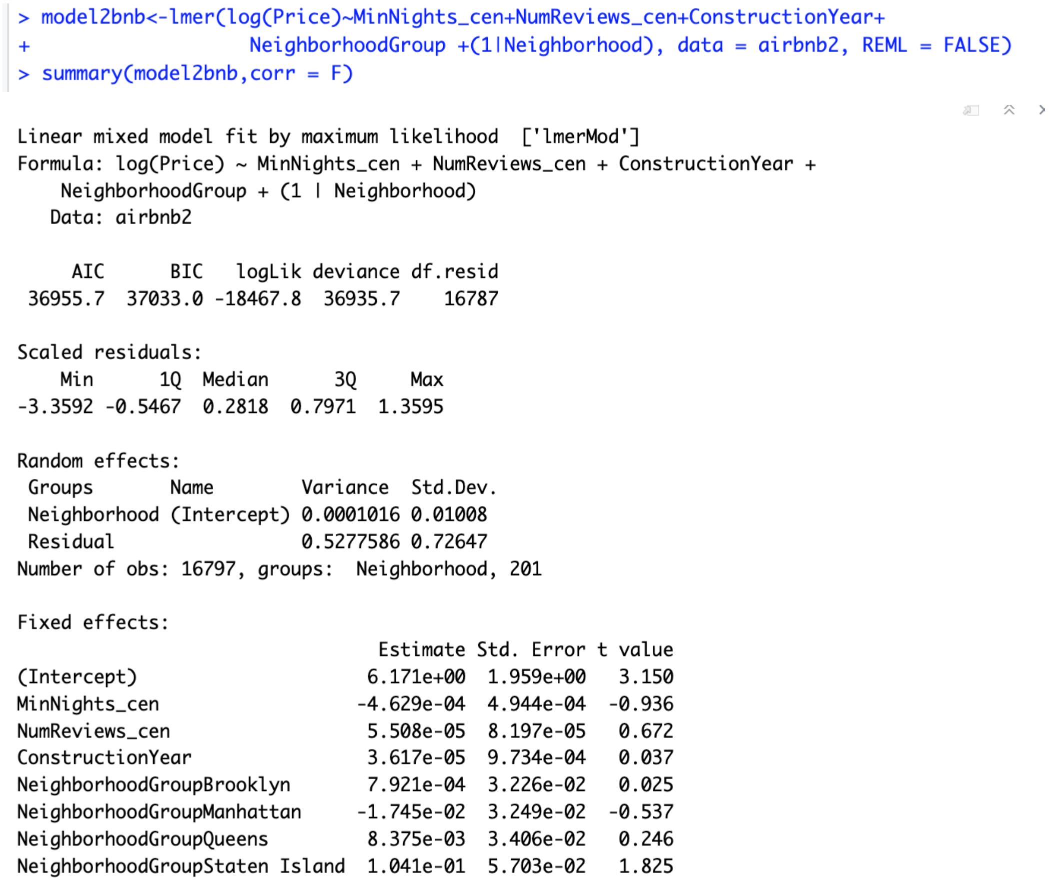


Even though the random slopes of the construction year are not statistically significant, we still wanted to see how the mean slope of all the random slopes of the construction year in a borough differs from borough to borough. From the graph above, we have seen that the mean slope of all the random slopes of the construction year in Bronx and Staten Island have different behavior than the other three boroughs. Staten Island has the largest increase in the mean slope because it is a very popular destination for travelers due to its low population density and large living space while having a low cost of living. Bronx has experienced huge gentrification over the past two decades due to being a very poor borough and more middle-class Airbnb owners are staying away from richer neighborhoods by moving into gentrified neighborhoods that were formerly very poor in living quality so the predicted price per night for an individual Airbnb in Bronx has decreased substantially with the increased in construction years over the past two decades.



```
> model1bnb<-lmer(log(Price)~MinNights_cen+NumReviews_cen+ConstructionYear+
+                    NeighborhoodGroup + (1+ConstructionYear|Neighborhood), data = airbnb2, REML = F)

> summary(model1bnb, corr =F)

Linear mixed model fit by maximum likelihood  ['lmerMod']
Formula: log(Price) ~ MinNights_cen + NumReviews_cen + ConstructionYear +      NeighborhoodGroup + (1 + ConstructionYear
| Neighborhood)
   Data: airbnb2

     AIC      BIC   logLik deviance df.resid
 37085.0  37177.7 -18530.5  37061.0    16785

Scaled residuals:
    Min      1Q  Median      3Q     Max
-3.3943 -0.5554  0.2820  0.7870  1.4800

Random effects:
 Groups       Name             Variance  Std.Dev.  Corr
 Neighborhood (Intercept)      5.221e-01 0.7225957
              ConstructionYear 1.039e-07 0.0003223 -0.95
 Residual                      5.231e-01 0.7232621
Number of obs: 16797, groups:  Neighborhood, 201

Fixed effects:
                             Estimate Std. Error t value
(Intercept)                 5.956e+00  1.967e+00   3.028
MinNights_cen              -4.893e-04  4.949e-04  -0.989
NumReviews_cen              5.128e-05  8.360e-05   0.613
ConstructionYear            1.409e-04  9.768e-04   0.144
NeighborhoodGroupBrooklyn   1.231e-02  6.282e-02   0.196
NeighborhoodGroupManhattan  3.148e-03  6.719e-02   0.047
NeighborhoodGroupQueens     2.662e-02  6.457e-02   0.412
NeighborhoodGroupStaten Island  9.247e-02  8.632e-02   1.071
```

Even though the random slopes of construction years have shown us some information regarding the differences in price per night between the different boroughs, we decided to not add the random slopes of construction years to the model because it does not explain any variation at level 2 compared to the null model (0.5221 vs. 0.0003633).

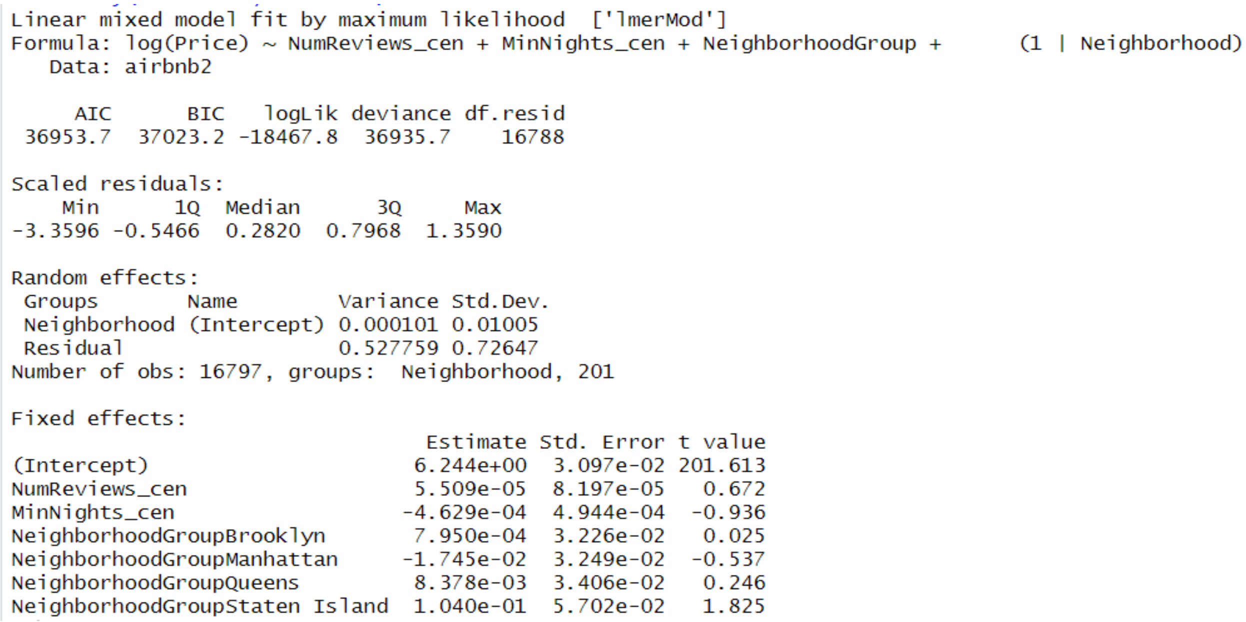


```
> model2bnb<-lmer(log(Price)~MinNights_cen+NumReviews_cen+ConstructionYear+
+                   NeighborhoodGroup +(1|Neighborhood), data = airbnb2, REML = FALSE)
> summary(model2bnb,corr = F)
```

```
Linear mixed model fit by maximum likelihood  ['lmerMod']
Formula: log(Price) ~ MinNights_cen + NumReviews_cen + ConstructionYear +
    NeighborhoodGroup + (1 | Neighborhood)
   Data: airbnb2

     AIC      BIC   logLik deviance df.resid
 36955.7  37033.0 -18467.8  36935.7    16787

Scaled residuals:
    Min      1Q  Median      3Q     Max
-3.3592 -0.5467  0.2818  0.7971  1.3595

Random effects:
 Groups       Name        Variance  Std.Dev.
 Neighborhood (Intercept) 0.0001016 0.01008
 Residual                 0.5277586 0.72647
Number of obs: 16797, groups:  Neighborhood, 201

Fixed effects:
                                 Estimate Std. Error t value
(Intercept)                     6.171e+00  1.959e+00   3.150
MinNights_cen                  -4.629e-04  4.944e-04  -0.936
NumReviews_cen                  5.508e-05  8.197e-05   0.672
ConstructionYear                3.617e-05  9.734e-04   0.037
NeighborhoodGroupBrooklyn       7.921e-04  3.226e-02   0.025
NeighborhoodGroupManhattan     -1.745e-02  3.249e-02  -0.537
NeighborhoodGroupQueens         8.375e-03  3.406e-02   0.246
NeighborhoodGroupStaten Island  1.041e-01  5.703e-02   1.825
```

After taking out random slopes for the construction year, we decided to see if our new model would have better performance. Even though the model without the random slopes explains more variation in level 2 than the null model (0.0001016 vs. 0.0003633), we believe that the variable Construction year is not necessary due to a low t-value of 0.037.

Final Model:



```
Linear mixed model fit by maximum likelihood  ['lmerMod']
Formula: log(Price) ~ NumReviews_cen + MinNights_cen + NeighborhoodGroup +      (1 | Neighborhood)
   Data: airbnb2

     AIC      BIC   logLik deviance df.resid 
 36953.7  37023.2 -18467.8  36935.7    16788 

Scaled residuals: 
    Min      1Q  Median      3Q     Max 
-3.3596 -0.5466  0.2820  0.7968  1.3590 

Random effects:
 Groups       Name        Variance Std.Dev.
 Neighborhood (Intercept) 0.000101 0.01005 
 Residual                 0.527759 0.72647 
Number of obs: 16797, groups:  Neighborhood, 201

Fixed effects:
                                Estimate Std. Error t value
(Intercept)                    6.244e+00  3.097e-02 201.613
NumReviews_cen                 5.509e-05  8.197e-05   0.672
MinNights_cen                 -4.629e-04  4.944e-04  -0.936
NeighborhoodGroupBrooklyn      7.950e-04  3.226e-02   0.025
NeighborhoodGroupManhattan    -1.745e-02  3.249e-02  -0.537
NeighborhoodGroupQueens        8.378e-03  3.406e-02   0.246
NeighborhoodGroupStaten Island 1.040e-01  5.702e-02   1.825
```

This is the final model we settled on for the report. We settled for this model because it explains far more variation at level 2 than the null model (0.000101 vs. 0.0003633) compared to the model with the construction year that we have used (0.0001016 vs. 0.0003633).

Works Cited:

"Pros and Cons of Moving to Staten Island." *Dumbo Moving and Storage,* https://dumbomoving.com/blog/pros-and-cons-of-moving-to-staten- island/. Accessed 4 Dec. 2023.

"Pros and Cons of Moving to Staten Island." *Roadmap Moving,* https://www.roadwaymoving.com/blog/moving-to-staten-island/. Accessed 4 Dec. 2023.

Bronck, Jonas. “Poverty in the Bronx: A Deep-Rooted and Pervasive Problem.” *The Bronx Daily,* 10 Feb. 2023,

https://bronx.com/poverty-in-the-bronx-a-deep-rooted-and-pervasive-problem/.

Cohen, Michelle. “Study Finds Bronx Residents Most in Danger of Housing Displacement Due to Gentrification.” *6sqrt,* 7 Mar. 2017,

https://www.6sqft.com/study-finds-bronx-residents-most-in-danger-of-housing-displacement-due-to-gentrification/

“Income and Taxes New York City (NYC) Median Household Income 2017 Estimates.” *NYCdata,*

https://www.baruch.cuny.edu/nycdata/income-taxes/med_hhold_income.htm. Accessed 10 Dec. 2023.