**Airbnb Room Price Prediction**

**Xiaofei Wu**

1. **Introduction**
2. **Project overall**

This report analyze an Airbnb dataset to predict the price (in $US) for a night stay. through Linear mixed effect model with log transformation. Before all model were build, the exploratory data analysis (EDA) is used to analyzing data sets to summarize their main characteristics, mostly with visual methods. The model will be evaluated from the root mean squared error (RMSE), where y hat is the natural log of the predicted price for (in $US) for a night stay and y is the natural log of the actual price value. Linear mixed effect model with log transformation leads to the least of 0.365945. Then, to share the results from analysis, the report will contain the interpretation. In the end, limitation and future work are talked to for future analysis and improvement.

1. **Background**

## The data is collected by Tom Slee from the public Airbnb web site without logging in. Tom Slee’s Report: How Airbnb hid the facts in New York City inspired me to looking into the Airbnb room price in New York City and see if I can find the pattern of price setting for Airbnb hosts.

1. **Evaluation**

The goodness of fit will be evaluated on the root mean squared error. RMSE is defined as: 图片包含 物体

描述已自动生成

1. **Data source**

The data set is all from Tom Slee’s website (To learn more about the data source, please visit <<http://tomslee.net/airbnb-data-collection-get-the-data>>).

My analysis contains ‘room\_id’, ‘host\_id’ to identify an Airbnb host, ‘room\_type’ as One of “Entire home/apt”, “Private room”, or “Shared room”. ‘Borough’, ‘neighborhood’(a subset of the city that is smaller than a borough), ‘reviews’(number of reviews from the room), ‘overall satisfaction’ (a star rate higher indicating more satisfaction), ‘accommodates’(the number of guests a listing can accommodate), ‘bedrooms’, the price (in $US) for a night stay, ‘latitude’ and ‘longitude’.

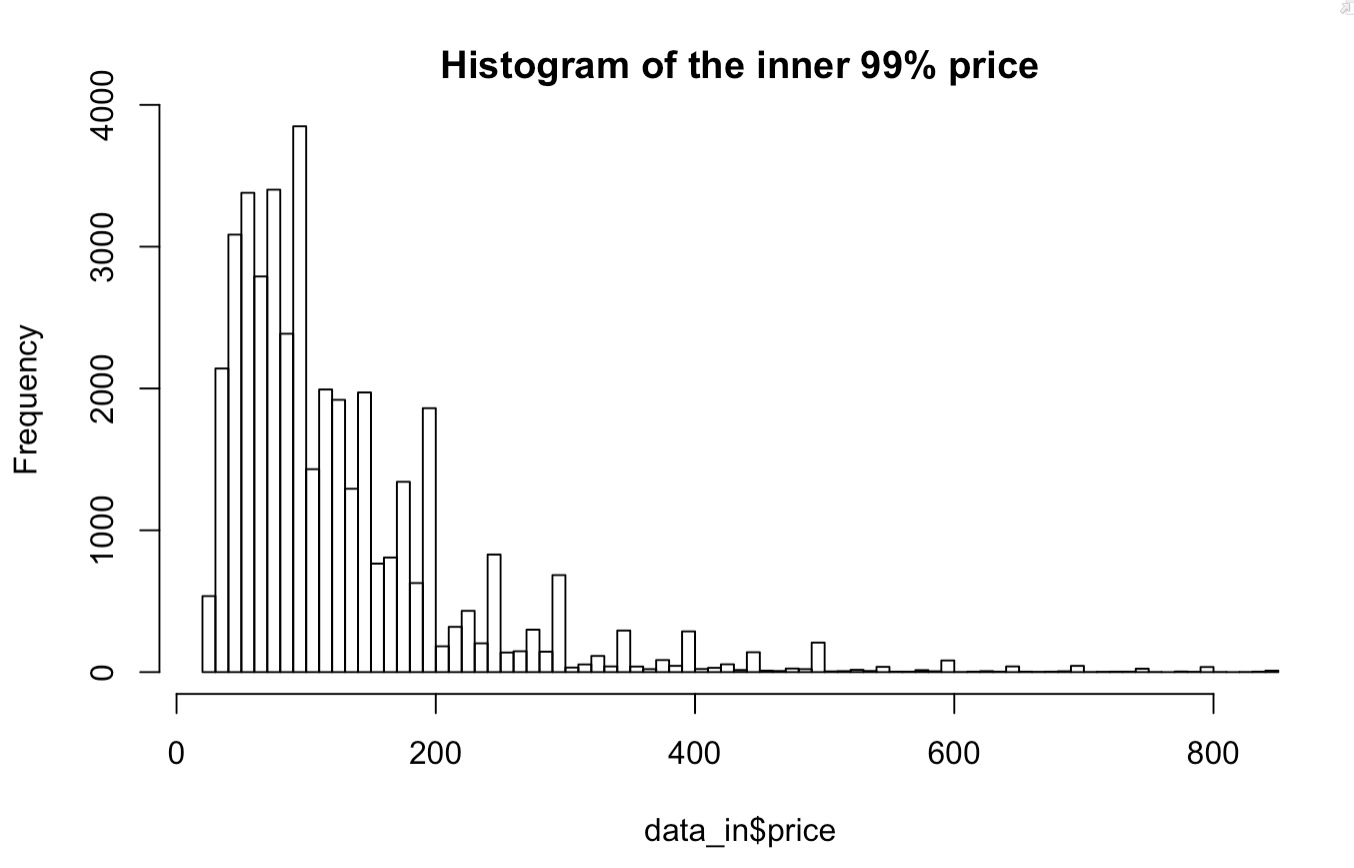
I delete columns for country and city since it is all from New York, as well as columns with null value such as ‘minstay’ and ‘bathroom’.

1. **Exploratory Data Analysis(EDA)**
2. **Firstly group the data by prices, which is the outcome, and create plots.**

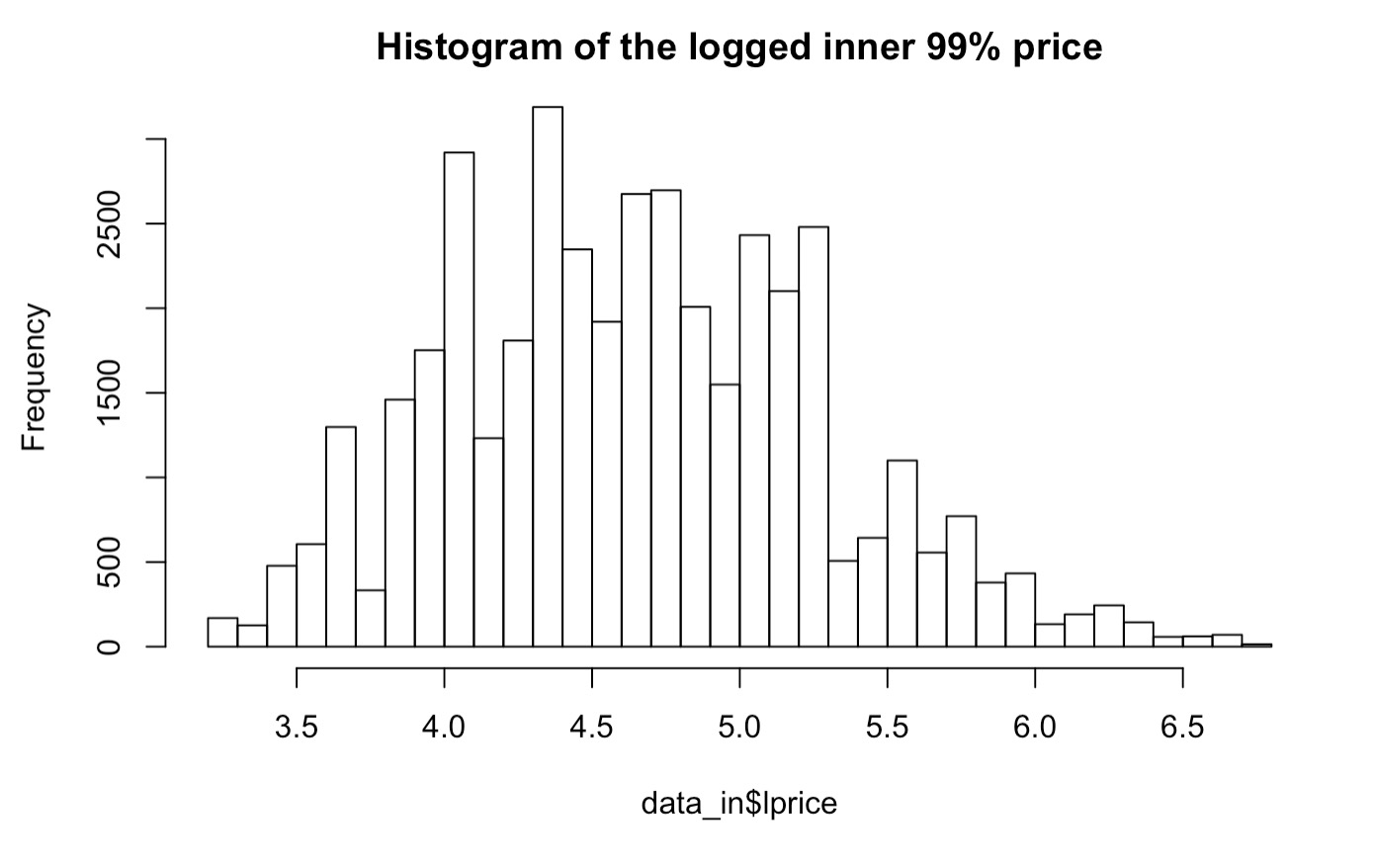
图片包含 屏幕截图

描述已自动生成

The prices are not normally distributed, with max price 9999 and min price 0. I want to avoid extreme price values and use the inner 99% price for prediction.



I took log transformation on the price and get a distribution with max 6.745236 and min 3.218876. The minimum price for price in the middle 99% is 25, so I avoided 0.



It seems like a normal distribution, so I created QQ plot to check:

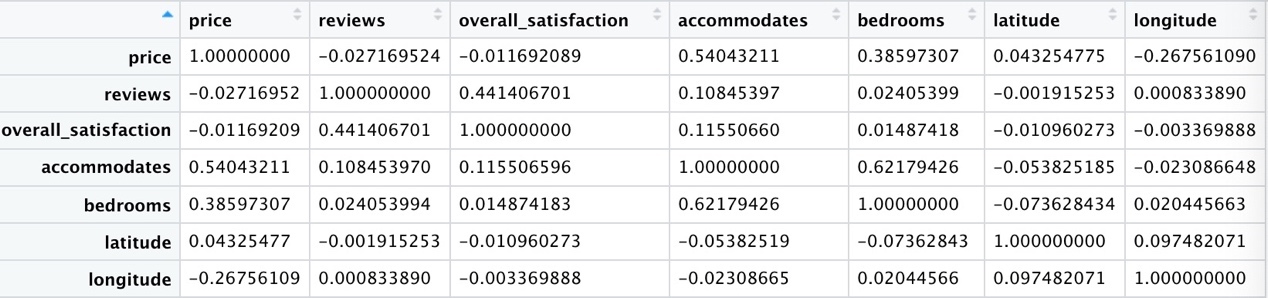
图片包含 屏幕截图

描述已自动生成

It indicates that the inner 99% target after log transformation approximately follows normal distribution.

1. **Analyze the correlations between explanatory and response variables**

First I want to see the correlations between all numerical variables.

****

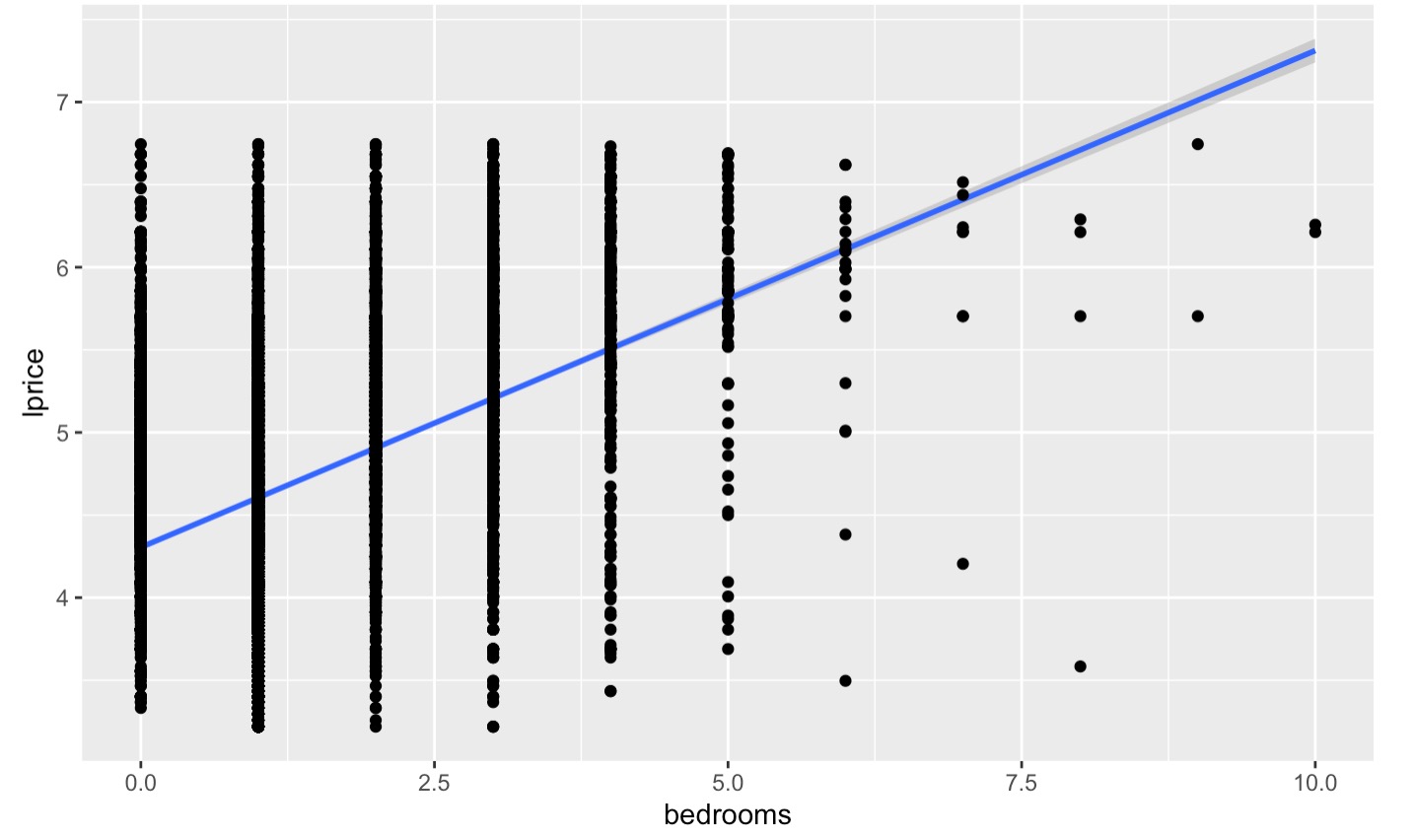
From the table, accommodates and bedrooms have apparently positive correlations with price, it makes sense that bigger room has higher price. Longitude has negative correlation with price, I guess it is caused by different overall room price in different boroughs. Overall satisfaction has positive correlation with review number.

1. **Analyze from the Room characteristic: “room\_type”, "accommodates", "bedrooms","property\_type".**
2. accommodates & bedrooms

There is one room with 50 bedrooms, I delete it and draw the following plots.

图片包含 物体

描述已自动生成



CONCLUSION: There is linear correlation between price and accommodates or bedrooms, consider accommodates and bedrooms as fixed effects. Adding (accommodates + bedrooms) in the linear mixed models. However, the price has a wide range for any fixed accommodates and bedrooms value, some random effect should be considered towards to these two fixed predictors.

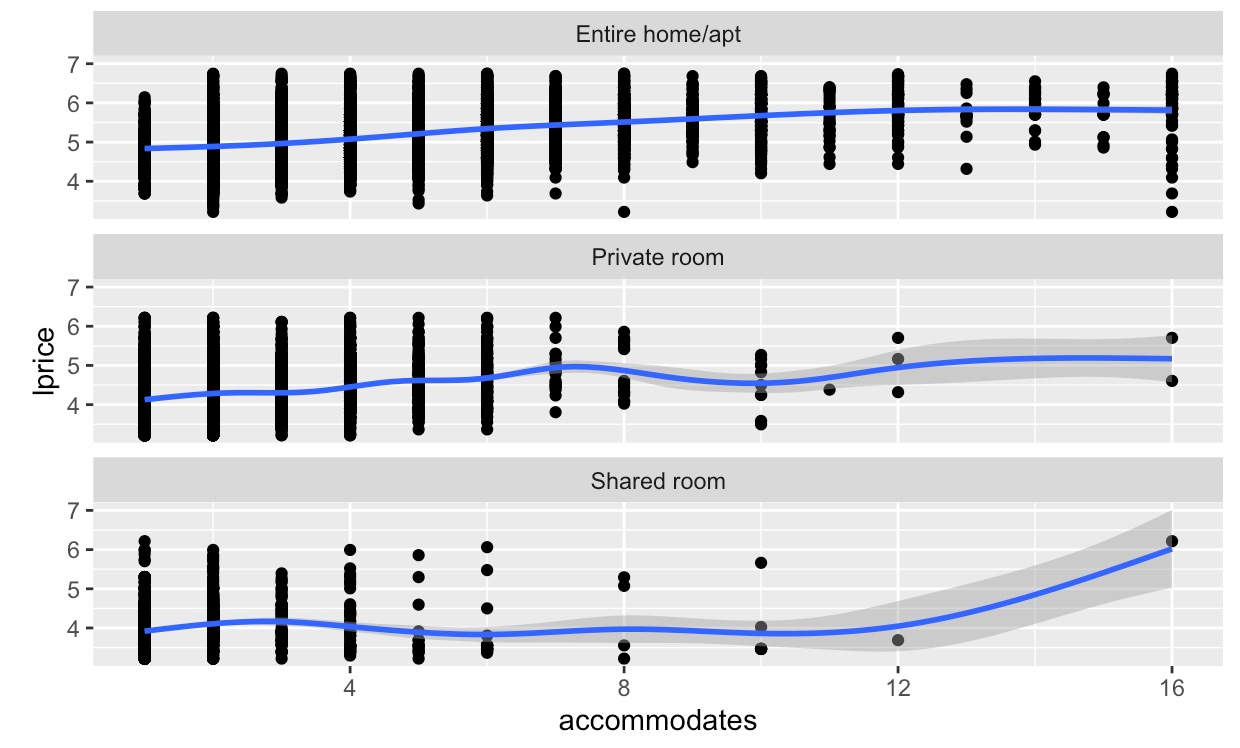
1. room\_type图片包含 屏幕截图

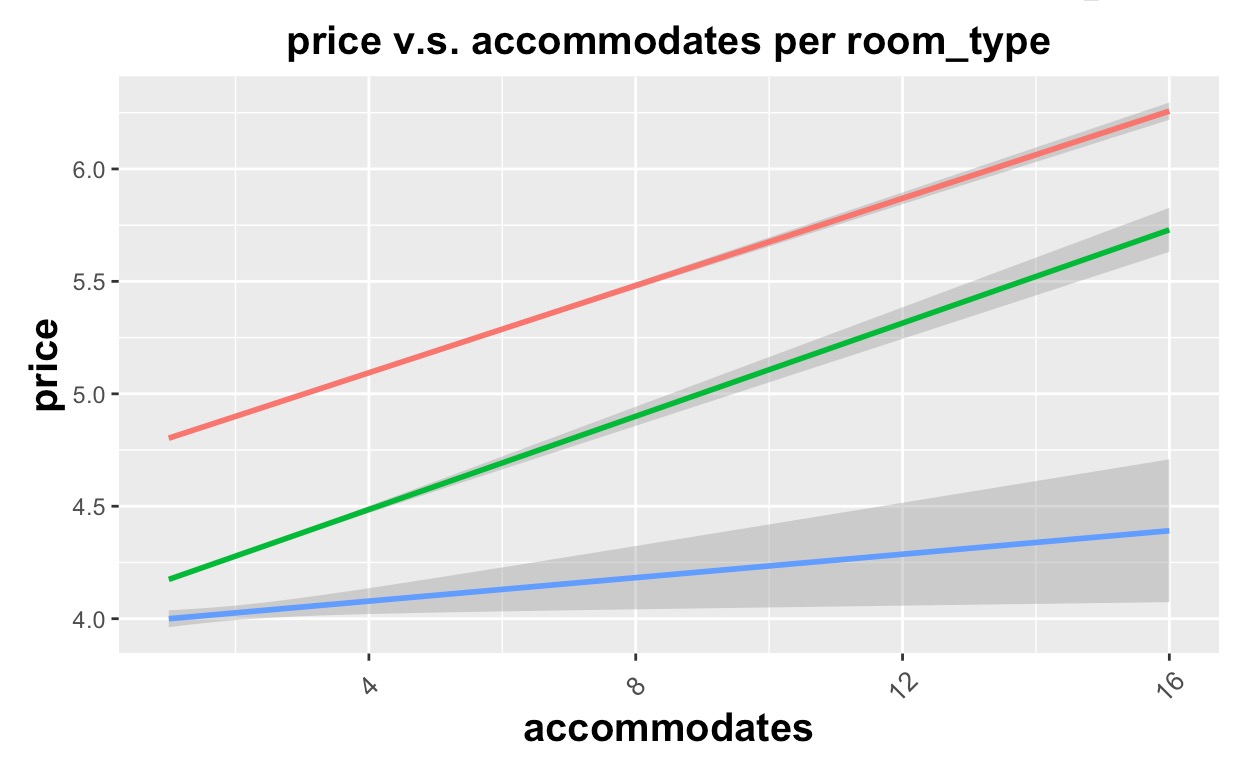
   描述已自动生成

图片包含 屏幕截图

描述已自动生成CONCLUSION: different room\_type has different price distribution, consider room\_type in the model.

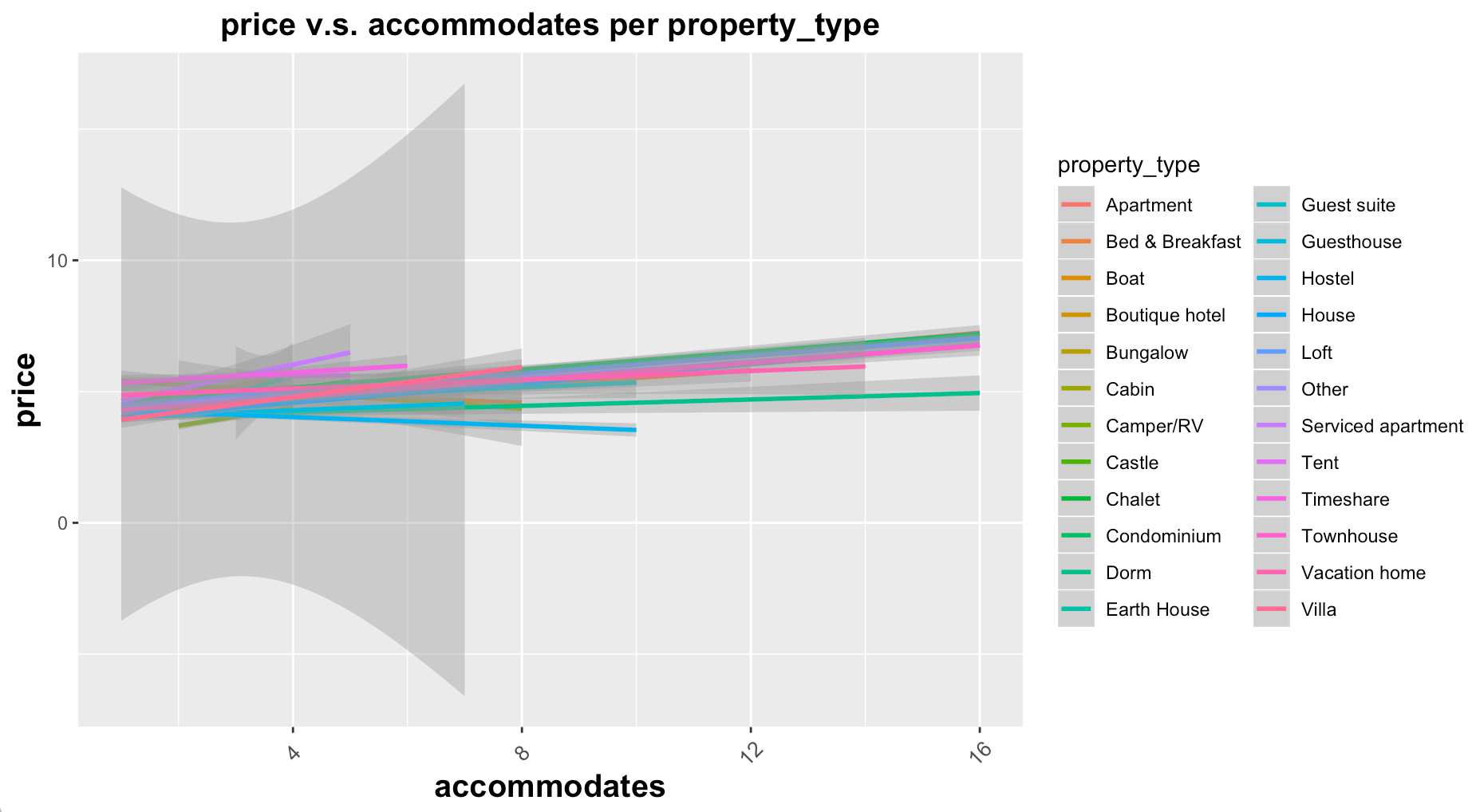
1. Room types & accommodates:





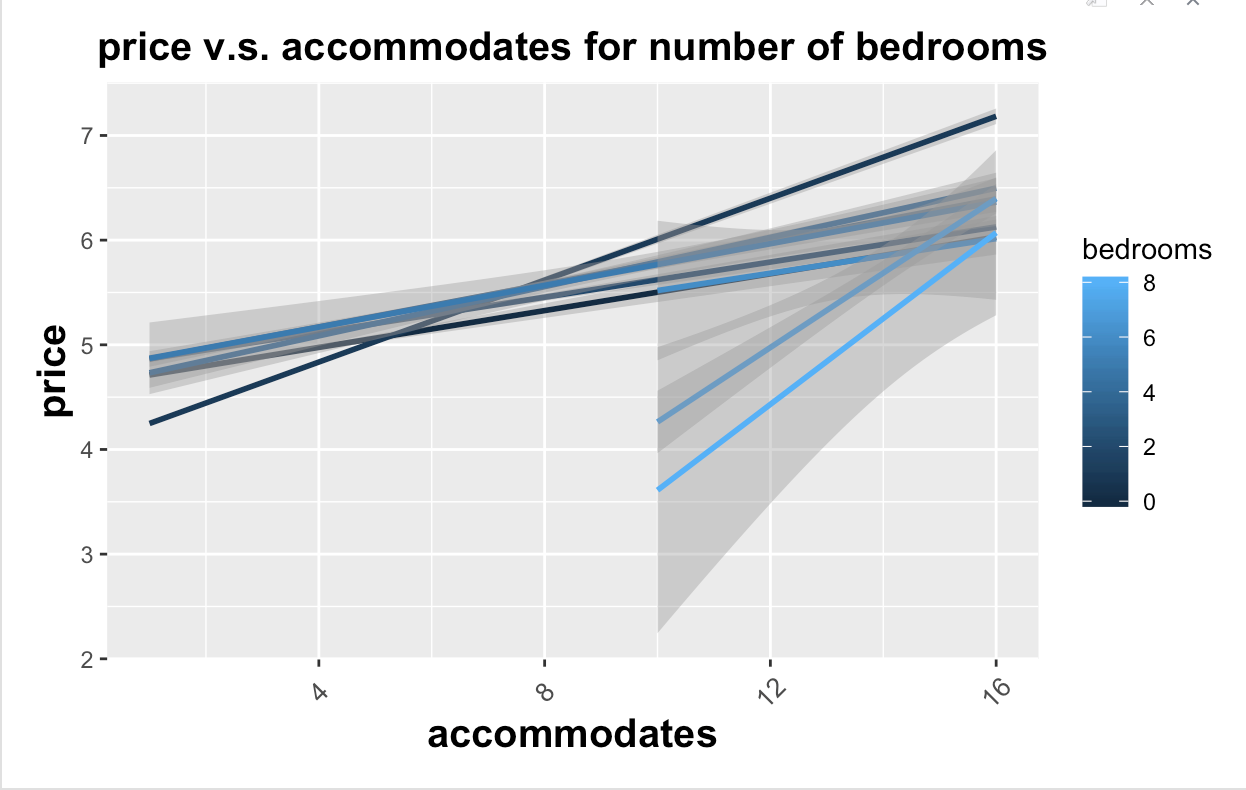
CONCLUSION: The effect (slope) of accommodates on price differ from room\_type, Adding (1+accommodates|room\_type) in the model.

1. Property\_type & Accommodates



CONCLUSION: No significant effect (slope) of accommodates on price differ from Property\_type.

1. Bedroom & Accommodates



CONCLUSION: The effect (slope) of accommodates on price differ from bedrooms, but we should either add bedroom as a fixed effect or as a random effect (1+accommodates|bedrooms) in the model.

1. **Analyze geometrical: borough, neighborhood, latitude, longitude**

图片包含 屏幕截图

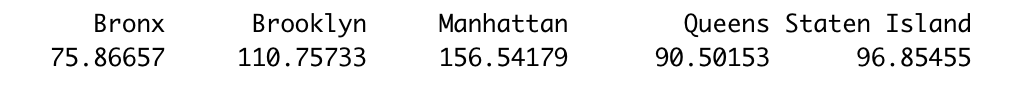
描述已自动生成

图片包含 屏幕截图

描述已自动生成

CONCLUSION: My assumption “borough causes the correlation between borough and price” seems to be right. The longitude for boroughs differ a lot, the slopes of longitude towards price did not show any signal or differ by boroughs. Different boroughs has apparent different price.

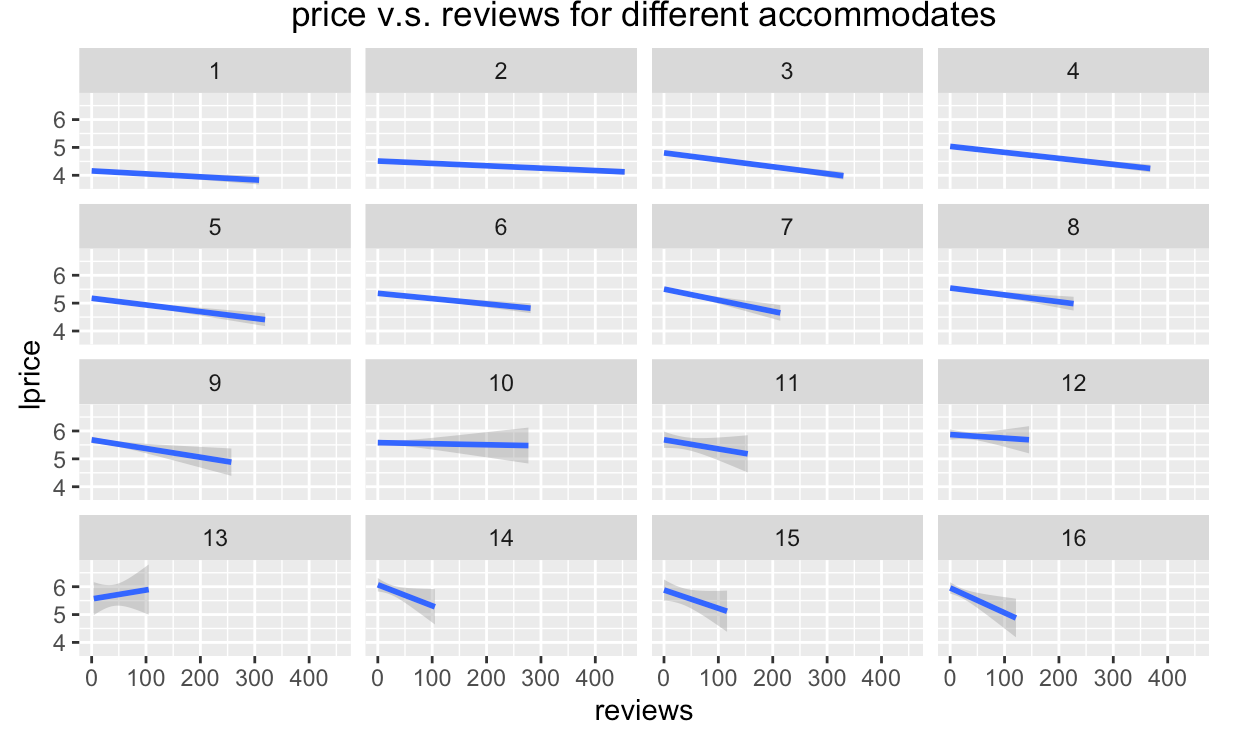
Tabulate the neighborhood with the average price：

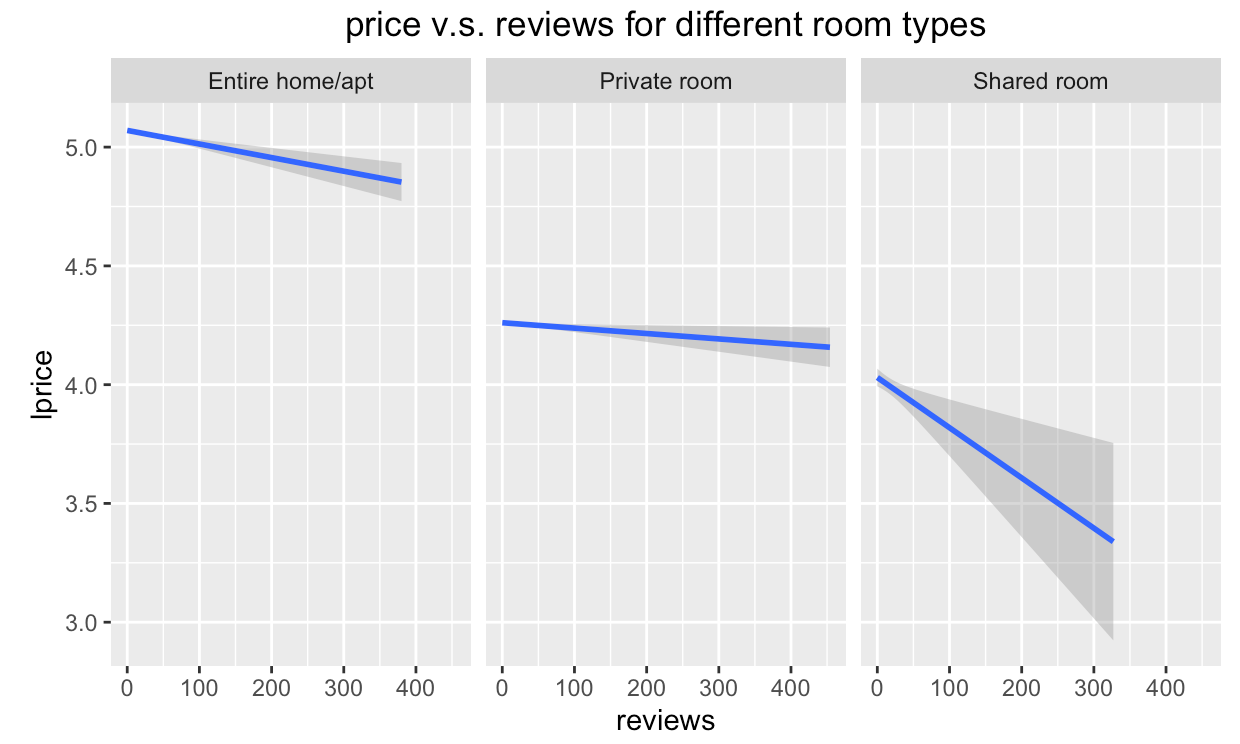


1. **Analyze review: reviews, overall satisfaction**
2. reviews

图片包含 物体

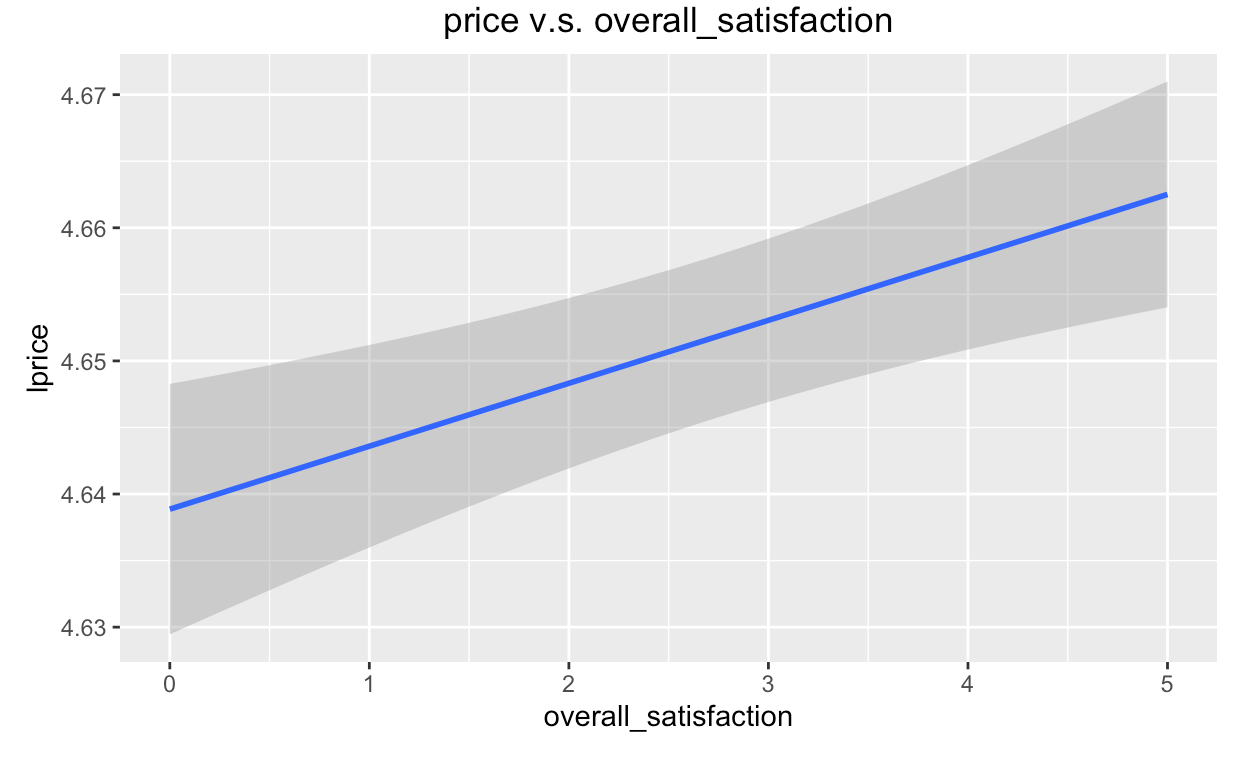
描述已自动生成





CONCLUSION: more reviews imply lower price in general, for various accommodates and various room types.

b) overall satisfaction



图片包含 屏幕截图

描述已自动生成

CONCLUSION: Higher overall satisfaction imply higher price in general, however, for any specific room types, the slope of overall satisfaction are not very significant, which means when we groups outcome by room type, it is questionable to say the overall satisfaction could imply the price.

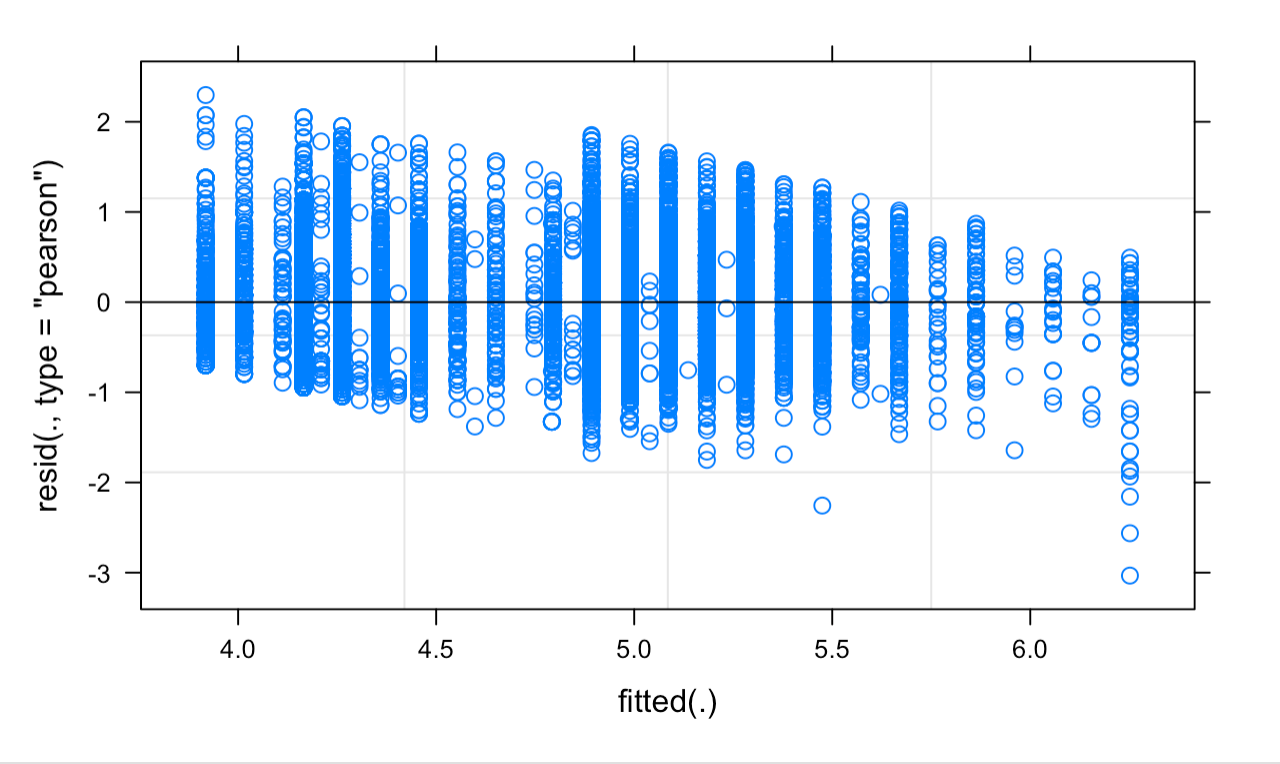
1. **Model**
2. **Fit linear mixed model group by room types. [RMSE = 0.4524676]**

**图片包含 文字

描述已自动生成**

The goodness of fit will be evaluated on the root mean squared error. RMSE is 0.4524676.

Check residuals using plot(MD1) to see if the **linear mixed model** follows assumptions (1. The errors have constant variance. 2. The errors are independent. 3. The errors are Normally distributed.**)**



Residuals decrease on the right side.

图片包含 屏幕截图

描述已自动生成

QQ plot show the residual follows the normal distribution except the two ends.

1. **Fit linear mixed model group by neighborhood. (including the effect of borough) [RMSE = 0.4356327]**

图片包含 文字

描述已自动生成

Check Residuals:

图片包含 文字, 地图

描述已自动生成

图片包含 地图, 文字

描述已自动生成

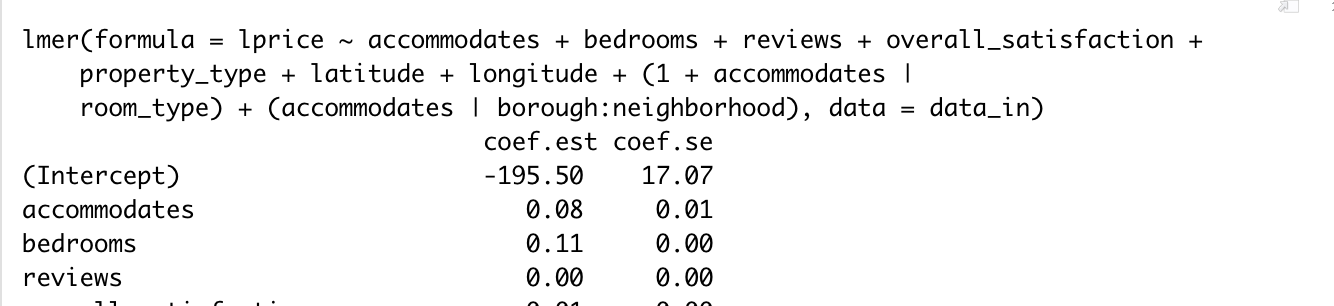
QQ plot show the residual follows the normal distribution except the two ends.

1. **Fit linear mixed model group by both neighborhood and room types. (including the effect of borough) [RMSE = 0.3715812]**

**图片包含 文字

描述已自动生成**

1. **Fit linear mixed model group by both neighborhood and room types, adding more fixed variables. (including the effect of borough) [RMSE = 0.3675805]**

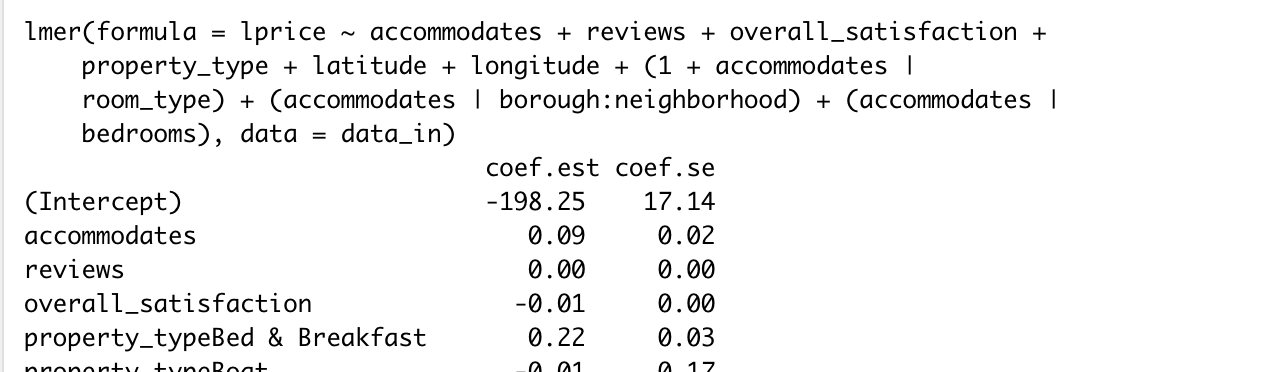
****

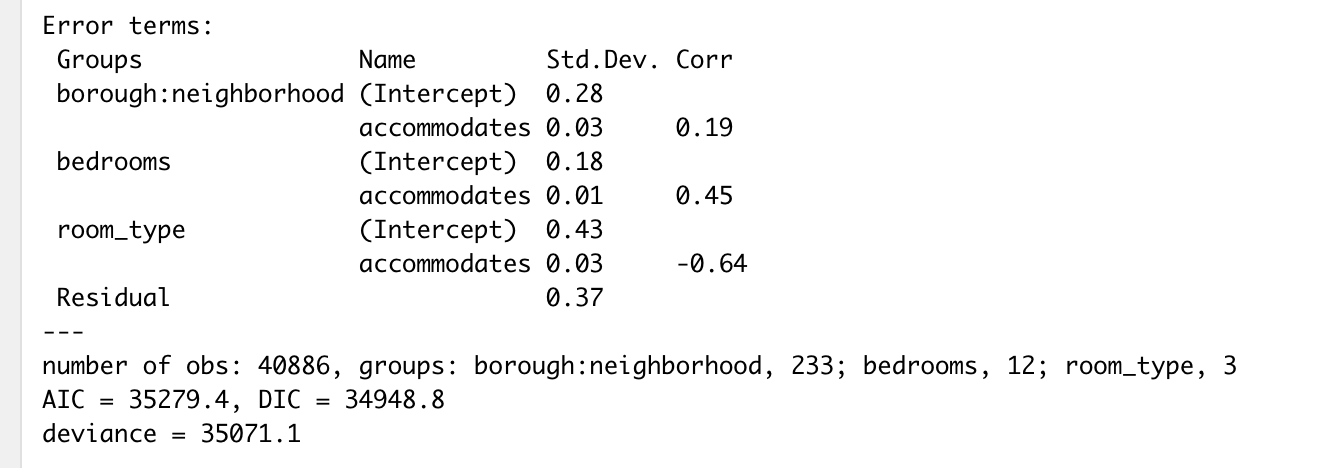
**…**

**图片包含 屏幕截图

描述已自动生成**

1. **Fit linear mixed model group by both neighborhood, room types and bedrooms, adding more fixed variables. (including the effect of borough) [RMSE = 0.365945]**

****

****

**Check residuals:**

**图片包含 文字, 地图

描述已自动生成**

**图片包含 屏幕截图, 地图

描述已自动生成**

1. **Interpretation**

**linear mixed model group by both neighborhood, room types and bedrooms.**

|  |  |  |  |
| --- | --- | --- | --- |
| **Predictors** | **+/-** | **estimate** | **t value** |
| **accommodates** | **+** | **8.602e-02** | **4.941** |
| **reviews** | **-** | **7.918e-04** | **-12.298** |
| **overall\_satisfaction** | **-** | **8.328e-03** | **-9.475** |
| **longitude** | **+** | **2.534e+00** | **-14.699** |

|  |  |  |
| --- | --- | --- |
| **Random effects** | **Intercept/Slope to accommodates** | **Variance** |
| **borough:neighborhood** | **Intercept** | **0.0810395** |
| **bedrooms** | **Intercept** | **0.0324233** |
| **room\_type** | **Intercept** | **0.1887975** |
| **Residual** |  | **0.1349951** |

**Total random variance: 0.435**

The best fit model use fixed predictors: accommodates + reviews + overall\_satisfaction + property\_type + latitude + longitude, grouped by room types, boroughs, neighborhoods, and bedrooms.

From the tables, we can see that the most significant predictor in the final model is accommodates, reviews, overall satisfaction and longitude. Larger accommodates and longitude determined higher price, on the other side, higher reviews or overall satisfaction imply lower price. This result roughly follows my assumptions in the EDA.

From the perspective of random effects, both room types and borough:neighborhood explain a lot of the variance.

1. **Discussions:**
2. **Assessment of the result**

Linear mixed effect model with log transformation leads to the least of 0.365945. The outcome gives me some explanations on the pricing for Airbnb in New York city.

1. **Limitations**

There are more data about Airbnb pricing online, due to time concerns, I only used this dataset. Also, as I know, there are some professional managers that help hosts to manage the house, for different type of managers, there could be pattern exists when they pricing the room. I did not get these information from this dataset. However, due to time concerns, I will keep this thoughts for future.

1. **Future directions**
2. **Time:** The date and time could be considered in the model if I got the corresponding data. For example, some Airbnb hosts raise price on weekend, lower the price if the room is not booked late that day. Date and time could be a huge impact on the final transaction price.
3. **Multiple Dataset**：there is a great website <<http://insideairbnb.com/>> contains more dataset about Airbnb. Also, it contains some sophisticated research on Airbnb data through which I could get deeper understanding of the Airbnb business. Then consider new perspective for this project.
4. **Acknowledgement**

I would like express many thanks to Professor Masanao Yajima for helping me gaining all these modeling knowledge.

1. **Reference**
2. Data source: <http://tomslee.net/category/about>
3. Tom Slee’s Report: How Airbnb hid the facts in New York City: <http://insideairbnb.com/how-airbnb-hid-the-facts-in-nyc/>