Predictors of Airbnb Prices in New York City

Jonathan Chow, Vamsi Ravilla, Jessica Rawls, & Sarah Nastasi

4/25/2022

Georgia Tech

Background

- Online marketplace where people can list their apartments, homes, condos, and vacation properties for rentals
- Used by travelers for a variety of reasons
  - Potential for less expensive stays than a hotel
  - Can provide a family stay in a larger space
  - Can provide a unique experience more immersive in the destination's culture

Motivation

- Identify and quantify factors that influence Airbnb prices
- Help vacationers determine what Airbnb would best fit their budget
- Help interested listers determine the best price to list their property

Data Set

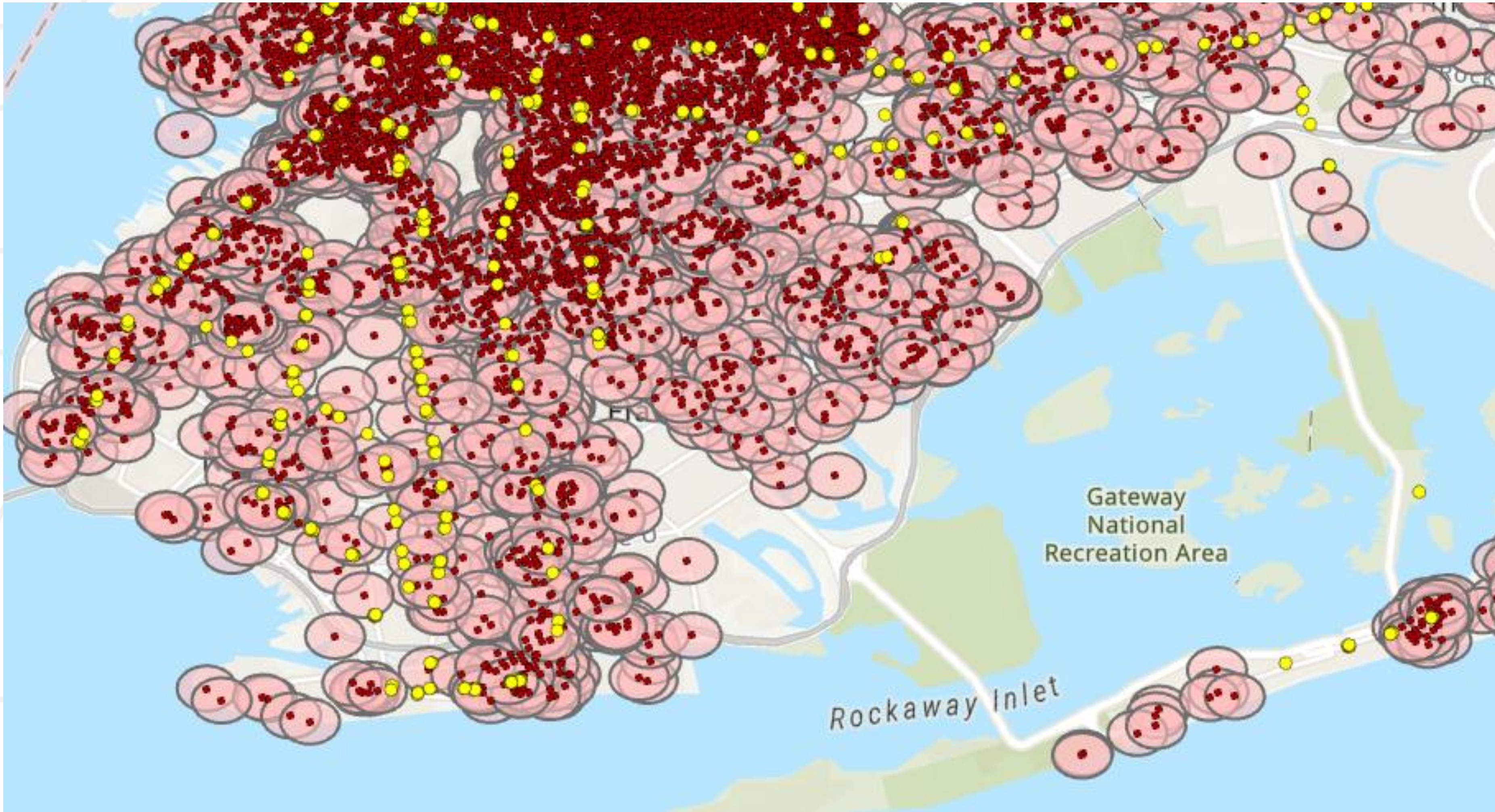
- Kaggle provided the data set
- ~50,000 datapoints for Airbnb listings in New York City, New York, in 2019
- 16 variables in the raw data set
  - ID
  - Name
  - Host ID
  - Host Name
  - Neighborhood Group
  - Neighborhood
  - Latitude
  - Longitude
  - Room Type
  - Price
  - Minimum Night Stay
  - Number of reviews
  - Date of last review
  - Reviews per month
  - Number of host listings
  - Yearly availability

Feature Engineering

- Desire to capture locality factors not provided in the raw data set
- Geographic variables were able to be extracted using the longitude & latitude of Airbnb listings

Table 1. Feature Engineered Variables

Measurement	Value
Access to Transportation System	Number of subway stations within 0.25 miles
Proximity to tourist attractions	Number of top 10 tourist attractions within 1 mile
Manhattan 1-mile buffer	0 or 1 for whether listing is within 1-mile buffer

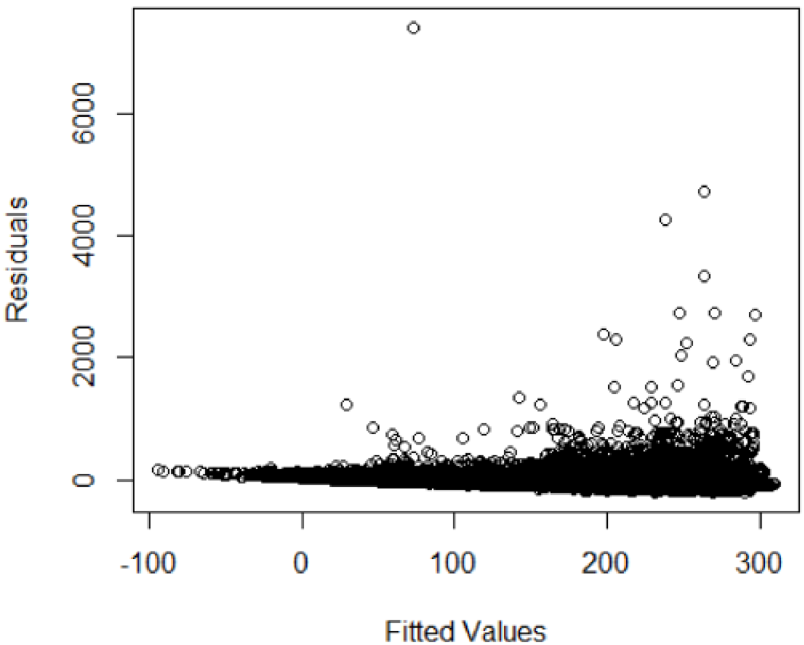


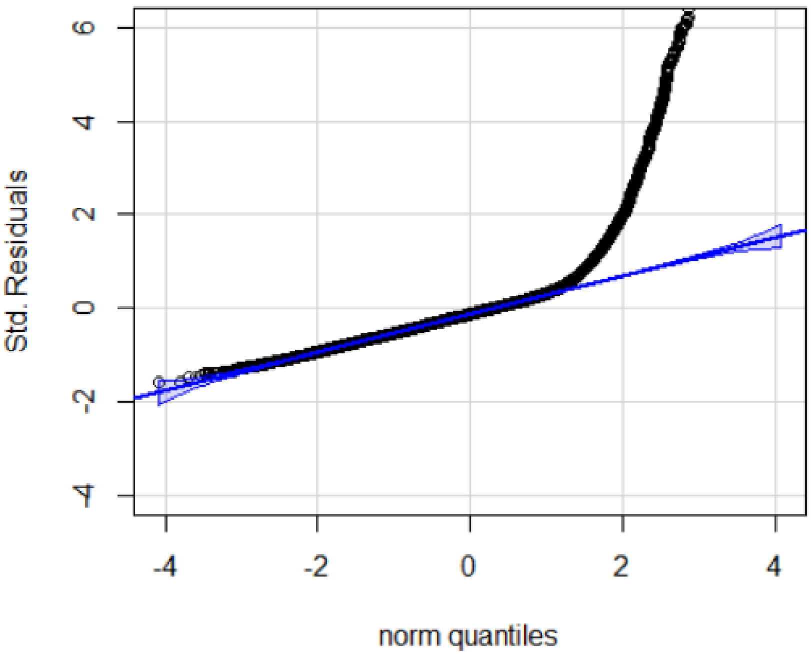
Subway Entrances
AirBnB Listing
AirBnB 0.25 mi buffer

Figure 1. Calculation of Airbnb Subway Access

Model Iteration 1 – Raw data

```
Call:
lm(formula = price ~ Flat + Sharing + Staten + Brooklyn + Queens +
    Bronx + n_reviews + min_nights + revs_pm + availability +
    host_list_count, data = dataUncat)
```



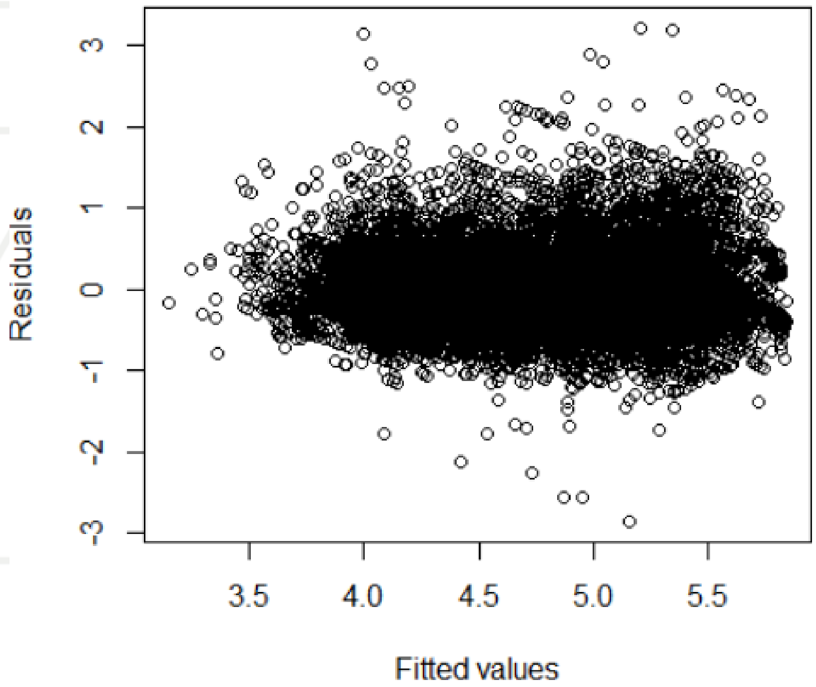


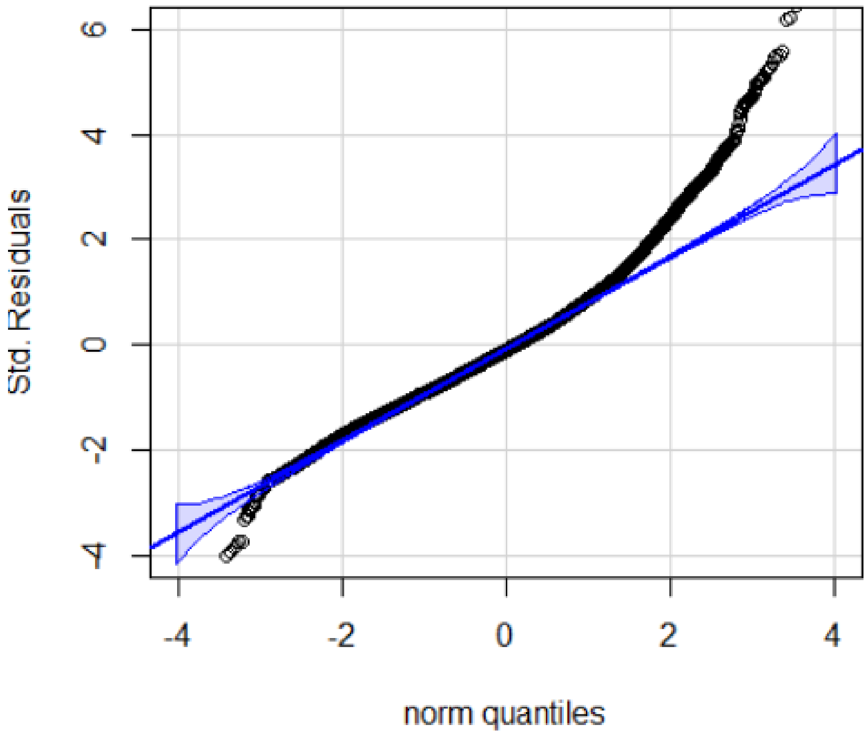
- Non-constant variance
- Non-normal behaviour
- Poor fit for the data
  - R2 = 0.2243
  - adj-R2 =0.2239

Log transform of price
Add more variables

Model Iteration 2 – Stepwise BIC

```
Call:
lm(formula = log(price) ~ Flat + subway + Queens + Brooklyn +
    Bronx + NYC + availability + Sharing + Staten + n_reviews,
    data = trainData)
```



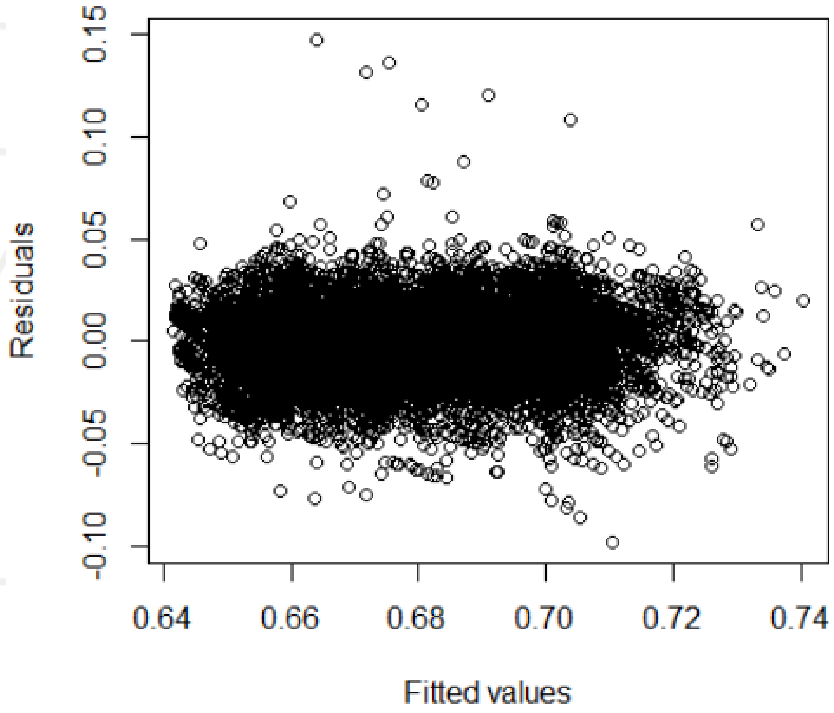


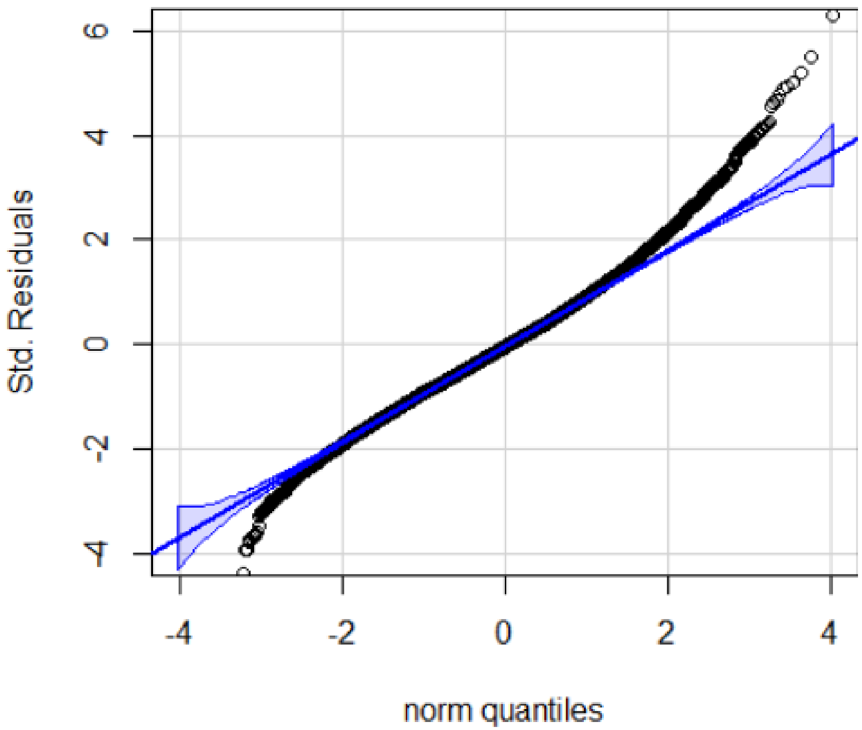
- Predictive performance
    - $R2 = 0.5623$, Adj. $R2 = 0.562$
- Slight linear increase in variance
- Non-normality at right tails

Box-Cox transform

Model Iteration 3 – Box-cox BIC model

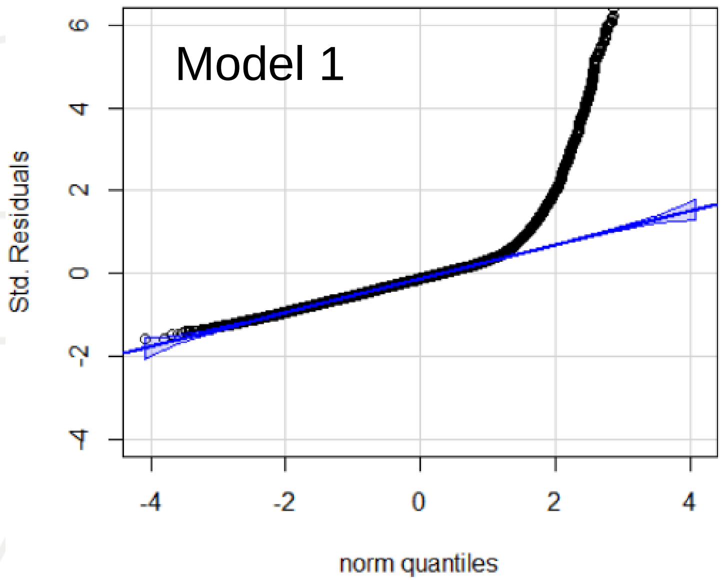
```
Call:
lm(formula = log(price)^(-0.25) ~ Flat + subway + Sharing + Queens +
    Brooklyn + Bronx + NYC + availability + Staten + n_reviews,
    data = trainData)
```

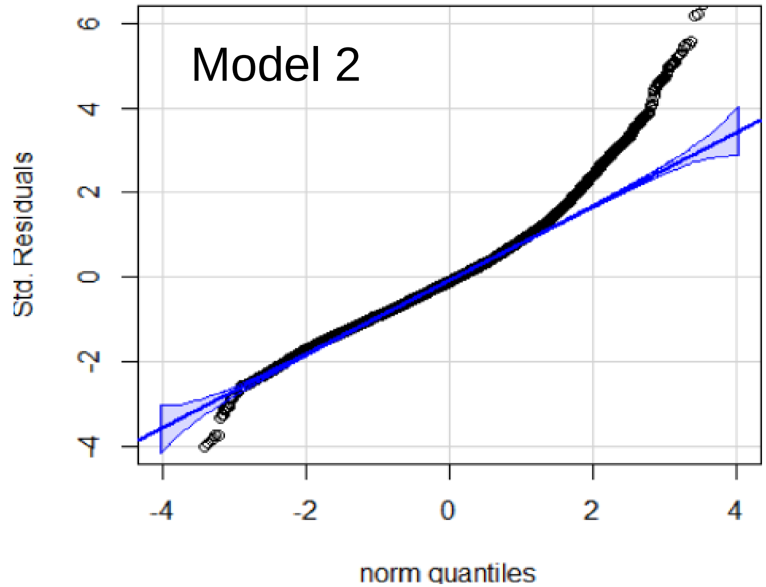


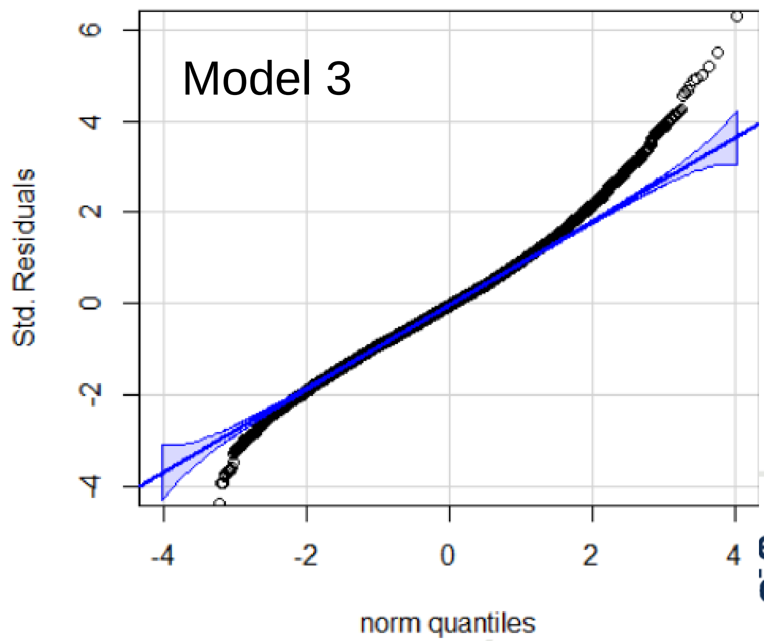


- Predictive performance
  - R2 = 0.579, Adj. R2 = 0.5787
- Constant variance of residuals
- Improved normality – but heavy right tail persists

Non-normality at high prices







- Improvement across 3 models
- Significant non-normality exists
  - Current features/variables are locality focused
  - What about property features?
  - Dataset does not provide any property specific information
    - Ex. Rooms, area, facilities
- Explore high price property attributes

Word cloud of high-end properties

- High-cost features such as rooftop, luxury, penthouse, backyard, 3/4 BR not captured by dataset
- Many properties over $1000 per night are commercial spaces

Misfit with remaining data

Georgia Tech

Model 4 – Non luxury stepwise BIC

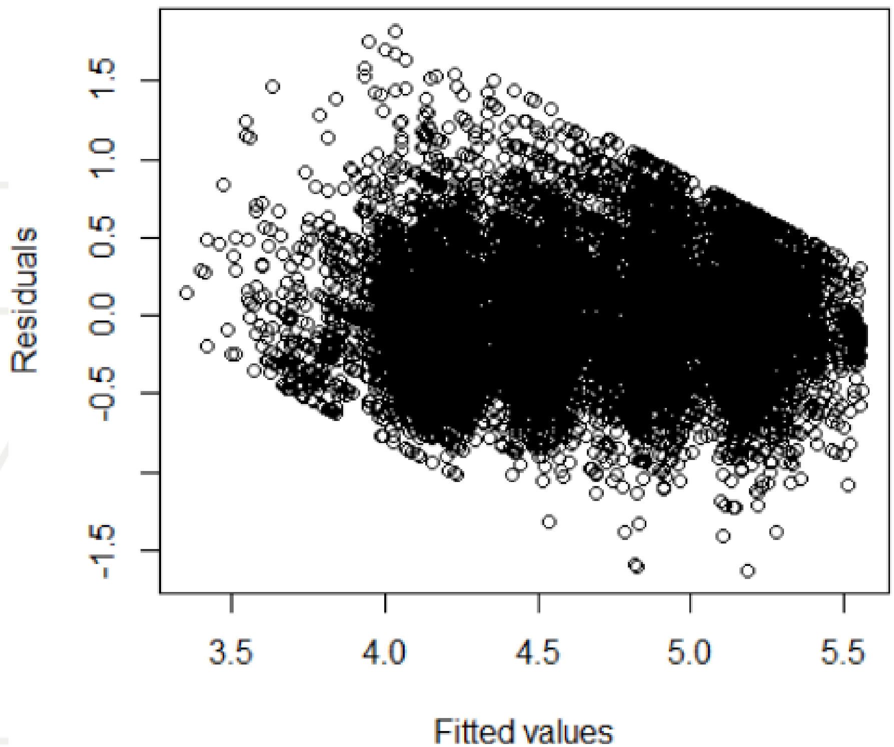
```
Call:
lm(formula = log(price) ~ Flat + subway + Sharing + Queens +
    Brooklyn + Bronx + NYC + Staten + availability, data = trainData)

Residuals:
     Min       1Q   Median       3Q      Max
-1.63261 -0.24828 -0.01506  0.22842  1.82213

Coefficients:
               Estimate Std. Error t value Pr(>|t|)
(Intercept)   4.384e+00  7.923e-03  553.41   <2e-16 ***
Flat          7.097e-01  5.941e-03  119.47   <2e-16 ***
subway        1.018e-02  5.378e-04   18.92   <2e-16 ***
Sharing      -4.377e-01  1.824e-02  -24.00   <2e-16 ***
Queens       -4.085e-01  9.605e-03  -42.53   <2e-16 ***
Brooklyn     -2.921e-01  7.319e-03  -39.91   <2e-16 ***
Bronx        -5.945e-01  1.739e-02  -34.19   <2e-16 ***
NYC           2.356e-01  9.661e-03   24.39   <2e-16 ***
Staten       -4.892e-01  2.641e-02  -18.52   <2e-16 ***
availability  4.387e-04  2.397e-05   18.30   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.3741 on 16666 degrees of freedom
Multiple R-squared:  0.5816,    Adjusted R-squared:  0.5814
F-statistic:  2574 on 9 and 16666 DF,  p-value: < 2.2e-16
```

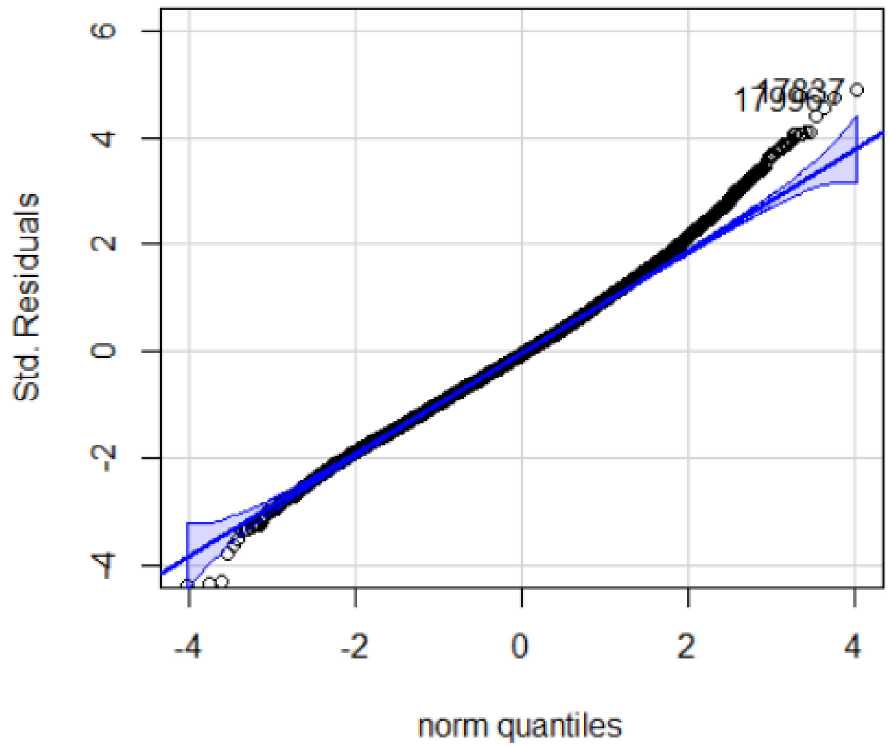
- Around 4.5% of the top properties removed
  - Price > $350
- Bottom 0.5% properties also removed
  - Price < $25

- All variables are significant
- VIF < 5 for all variables
- Improved prediction

Georgia Tech

Model 4 – Residual analysis





- Residuals mean around zero - Linearity holds good
- All residuals are bounded within a horizontal band – Constant variance
- Normal behaviour – only slight deviations at tails
- Good predictive power – R2 = 0.58

Georgia Tech

Model Iterations Summary

Model	Adj. R-squared (Test data)	Number of Variables
Raw / Only Kaggle data	0.21	11
BIC stepwise with feature engineering	0.56	10
Box-Cox transformation	0.57	10
Random Forest	0.62 (train)	28
Non-Luxury listings	0.58	9

Model 4 is the final model:

- All coefficients are significant
- No multicollinearity
  - VIF < 5 for all input variables
- Cook's distance << 1 for all points
  - Around 1% points are greater than 4/N
  - Possibly due to lack of property specific variables in the data

$$\boxed{\begin{aligned} &log(price) \sim Flat + subway + Sharing \\ &+ Queens + Brooklyn + Bronx + NYC \\ &+ Staten + availability \end{aligned}}$$

Conclusions

- Locality plays a significant role
  - Neighbourhood
  - Accessibility
  - Proximity to Manhattan
- Property specific attributes play significant role

- Surprisingly, tourist information plays

Future Research

- Examine how variables change across different cities
- Analyze how higher-income and luxury Airbnbs behave differently than averaged priced listings

Thank you

Questions?

Jonathan Chow, Chakri Ravilla, Jessica Rawls, &
Sarah Nastasi

4/25/2022

Georgia Tech