Price Prediction of Upper East Side Airbnb Listings

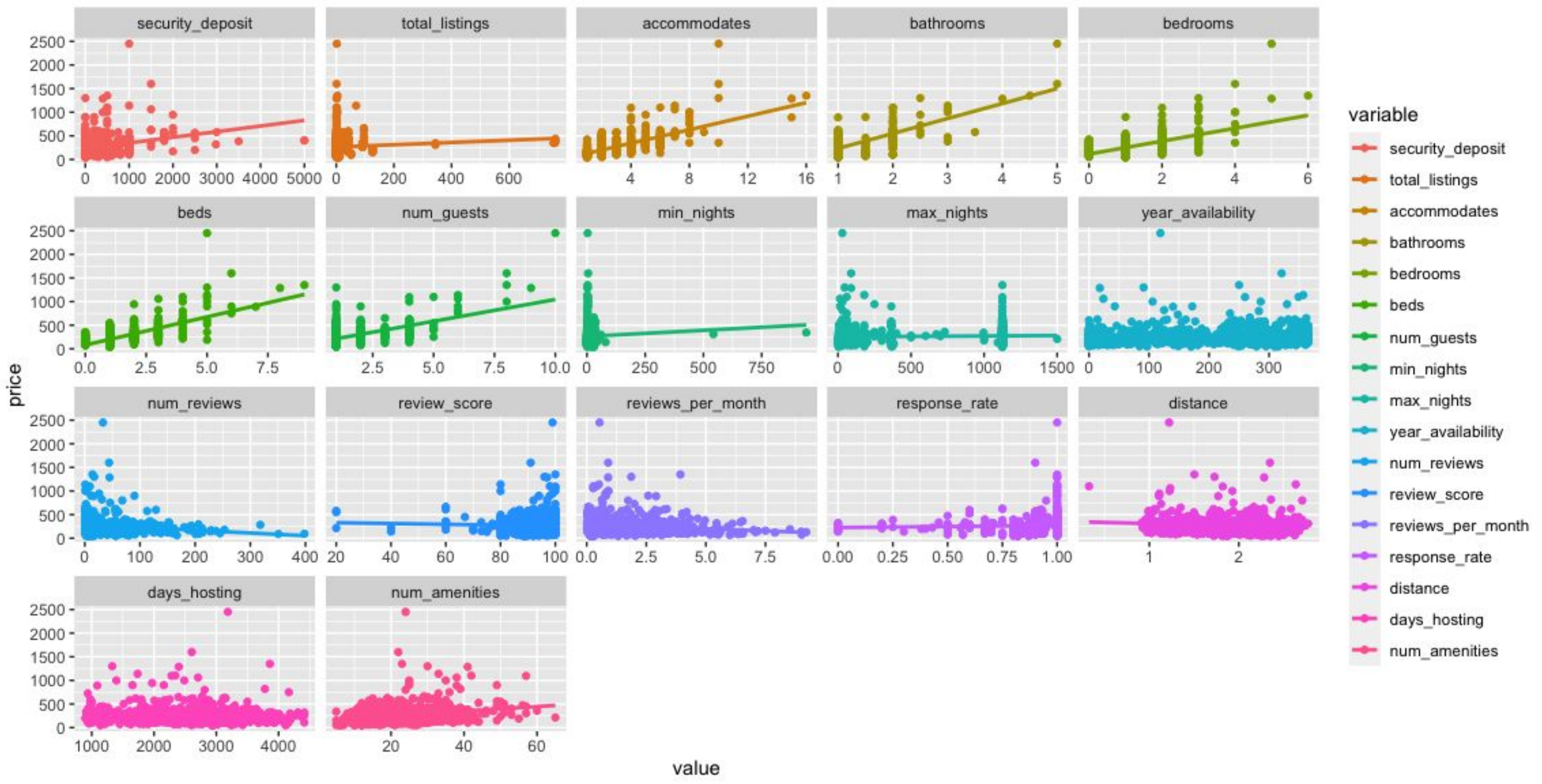
Armale Khan, Haeun Kim, Lakshmipriya Narayanan, Emily Painter

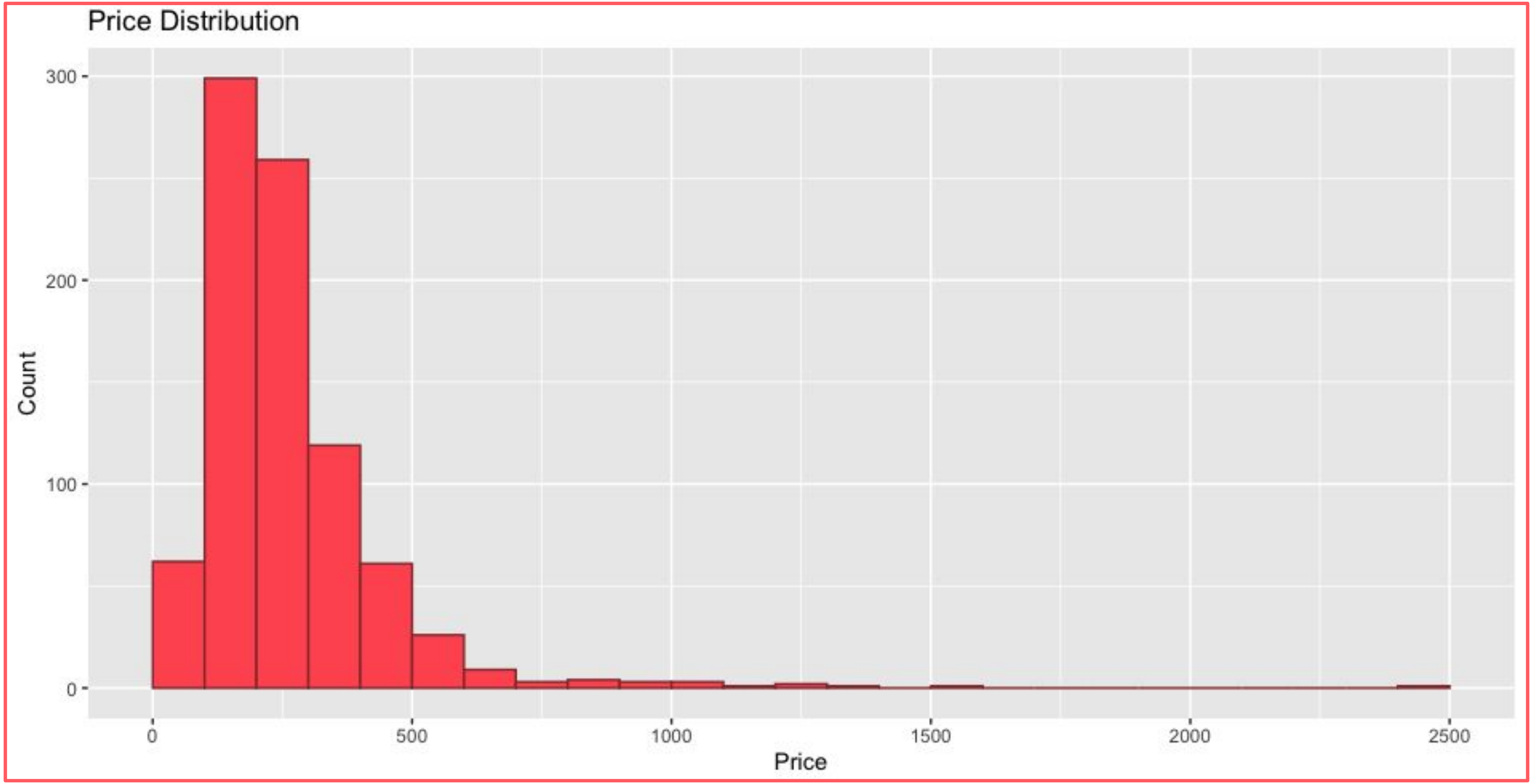
Introduction to the Data

- Snapshot taken in August of 2019 of Airbnb listings in New York City
- 48,864 observations and 106 attributes
- 62 attributes were either unusable character arrays, factor type, or boolean.
  - Selected few variables to use in data manipulation and removed the rest
  - Preferred strictly continuous predictor variables as well as our continuous response, price.
- Subsetted the data down to the Upper East Side neighborhood only
  - 854 observations after removing rows with NA values
  - Showed the most promise in modelling well of all the neighborhoods we tried
- **Response:** Price
- **Regressors:** Security Deposit, Total Listings, Accommodates, Bathrooms, Bedrooms, Beds, No. of Guests, Minimum Nights, Maximum Nights, Year Availability, No. of Reviews, Review Score, Reviews Per Month, Response Rate, Distance, Days Hosting, No. of Amenities

- ***Analysis Goal*** : Our ultimate goal is to identify features affecting the price of a one night stay.

Exploratory Analysis


security_deposit
total_listings
accommodates
bathrooms
bedrooms
beds
num_guests
min_nights
max_nights
year_availability
num_reviews
review_score
reviews_per_month
response_rate
distance
days_hosting
num_amenities
price
value
variable



Price Distribution
300
200
100
0
Count
0
500
1000
1500
2000
2500
Price

Variable Selection and Model Fitting

```
Call:
lm(formula = price ~ security_deposit + total_listings + accommodates +
    bathrooms + bedrooms + beds + num_guests + min_nights + max_nights +
    year_availability + num_reviews + review_score + reviews_per_month +
    response_rate + distance + days_hosting + num_amenities,
    data = airbnb)

Residuals:
    Min      1Q  Median      3Q     Max
-388.46  -56.95   -9.75   52.26  966.31

Coefficients:
                    Estimate Std. Error t value Pr(>|t|)
(Intercept)       -1.921e+02  5.168e+01  -3.718 0.000214 ***
security_deposit   7.481e-02  7.352e-03  10.176  < 2e-16 ***
total_listings     1.814e-01  5.659e-02   3.204 0.001404 **
accommodates       2.637e+01  3.644e+00   7.237 1.04e-12 ***
bathrooms          1.820e+02  1.150e+01  15.831  < 2e-16 ***
bedrooms           2.571e+00  6.815e+00   0.377 0.706129
beds               2.568e+01  5.907e+00   4.347 1.55e-05 ***
num_guests         1.743e+01  4.130e+00   4.220 2.71e-05 ***
min_nights        -1.792e-01  9.769e-02  -1.834 0.066966 .
max_nights        -5.599e-03  7.152e-03  -0.783 0.433924
year_availability  1.279e-01  2.898e-02   4.412 1.16e-05 ***
num_reviews       -1.817e-01  9.745e-02  -1.865 0.062600 .
review_score       1.117e+00  4.279e-01   2.609 0.009236 **
reviews_per_month -6.152e+00  3.040e+00  -2.024 0.043335 *
response_rate      1.661e+01  2.417e+01   0.687 0.492165
distance          -2.671e+01  8.235e+00  -3.243 0.001228 **
days_hosting      -5.418e-03  5.275e-03  -1.027 0.304663
num_amenities      1.198e+00  3.982e-01   3.008 0.002709 **
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 102.2 on 836 degrees of freedom
Multiple R-squared:  0.7034,    Adjusted R-squared:  0.6974
F-statistic: 116.6 on 17 and 836 DF,  p-value: < 2.2e-16
```

Full Model

```
Call:
lm(formula = price ~ security_deposit + total_listings + accommodates +
    bathrooms + beds + num_guests + min_nights + year_availability +
    num_reviews + review_score + reviews_per_month + distance +
    num_amenities, data = airbnb)

Residuals:
    Min      1Q  Median      3Q     Max
-388.84  -58.02   -9.11   51.76  967.08

Coefficients:
                    Estimate Std. Error t value Pr(>|t|)
(Intercept)       -1.924e+02  4.577e+01  -4.205 2.89e-05 ***
security_deposit   7.417e-02  7.299e-03  10.161  < 2e-16 ***
total_listings     1.831e-01  5.623e-02   3.256 0.001175 **
accommodates       2.669e+01  3.448e+00   7.740 2.85e-14 ***
bathrooms          1.834e+02  1.100e+01  16.681  < 2e-16 ***
beds               2.650e+01  5.740e+00   4.616 4.52e-06 ***
num_guests         1.754e+01  4.112e+00   4.265 2.23e-05 ***
min_nights        -1.771e-01  9.729e-02  -1.820 0.069137 .
year_availability  1.251e-01  2.844e-02   4.400 1.22e-05 ***
num_reviews       -2.226e-01  8.847e-02  -2.517 0.012035 *
review_score       1.097e+00  4.249e-01   2.581 0.010030 *
reviews_per_month -4.309e+00  2.682e+00  -1.607 0.108500
distance          -2.752e+01  8.158e+00  -3.374 0.000776 ***
num_amenities      1.223e+00  3.949e-01   3.098 0.002013 **
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 102.1 on 840 degrees of freedom
Multiple R-squared:  0.7026,    Adjusted R-squared:  0.698
F-statistic: 152.6 on 13 and 840 DF,  p-value: < 2.2e-16
```

Reduced Model

```
Analysis of Variance Table

Model 1: price ~ security_deposit + total_listings + accommodates + bathrooms +
    beds + num_guests + min_nights + year_availability + num_reviews +
    review_score + reviews_per_month + distance + num_amenities
Model 2: price ~ security_deposit + total_listings + accommodates + bathrooms +
    bedrooms + beds + num_guests + min_nights + max_nights +
    year_availability + num_reviews + review_score + reviews_per_month +
    response_rate + distance + days_hosting + num_amenities
  Res.Df     RSS Df Sum of Sq      F Pr(>F)
1    840 8755084
2    836 8730831  4     24253 0.5806 0.6768
```

ANOVA

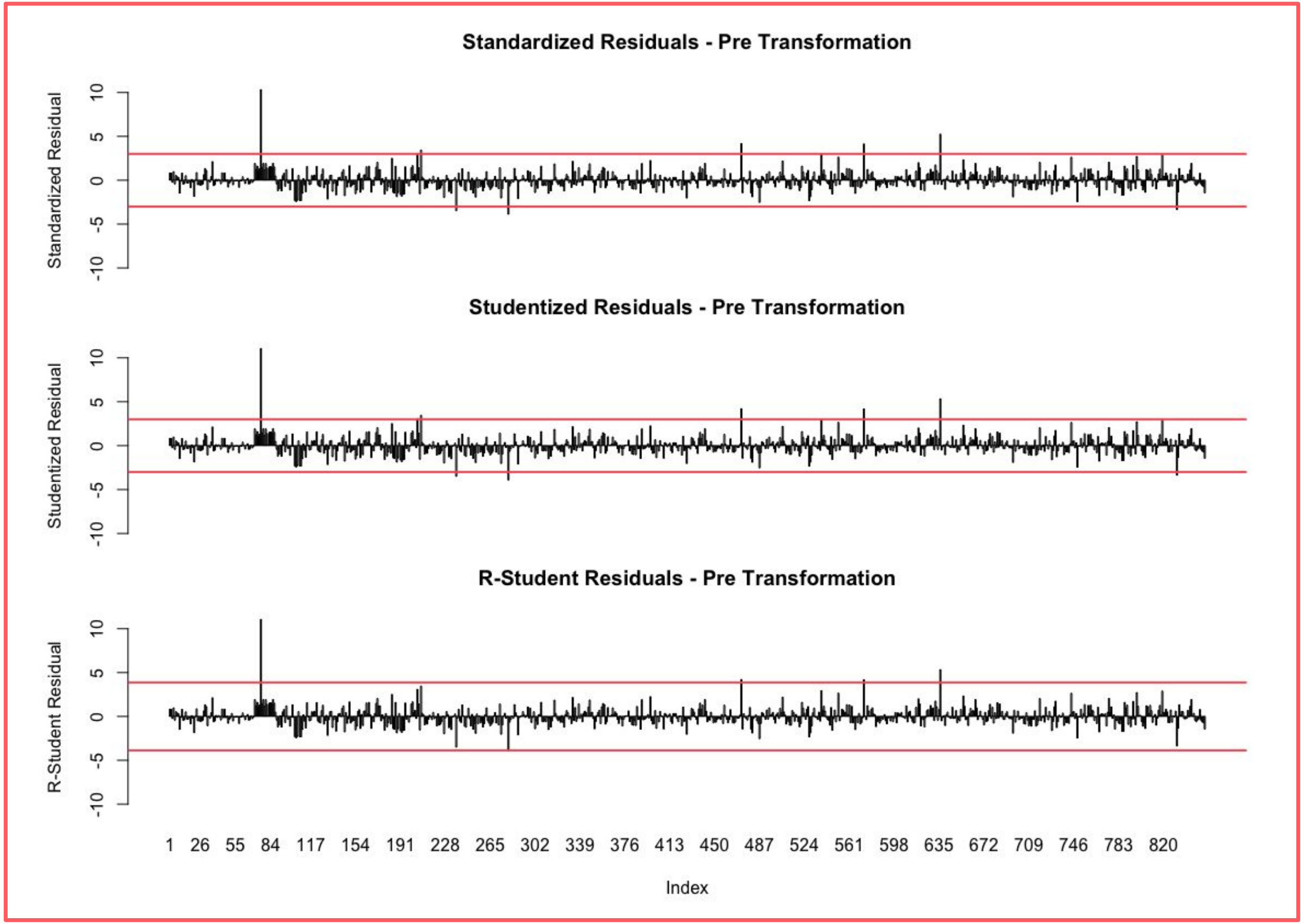
Our ANOVA table comparing our full model versus our reduced model confirmed that the dropped variables—bedrooms, max_nights, response_rate, and days_hosting —do not contribute significantly to the model and that our reduced model is a better fit for our data.

Multicollinearity

```
 security_deposit    total_listings      accommodates         bathrooms              beds        num_guests        min_nights
         1.193696          1.184074          2.888087          1.538691          2.822806          1.681442          1.159204
year_availability       num_reviews      review_score reviews_per_month          distance     num_amenities
         1.159794          1.390661          1.085663          1.427472          1.041987          1.109605
```

Output from the vif() function revealed no evidence of a multicollinearity problem in our model.

Residual Analysis



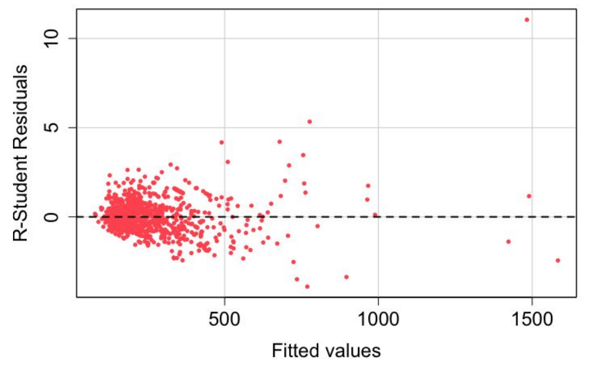
- Significant Residuals
  - 9 standardized
  - 9 studentized
  - 5 R-student
- One observation with a residual >10 worthy of investigation
- The index with a residual >10 is highly unusual in the y-space, with a price over $800 greater than the second highest price in our data.

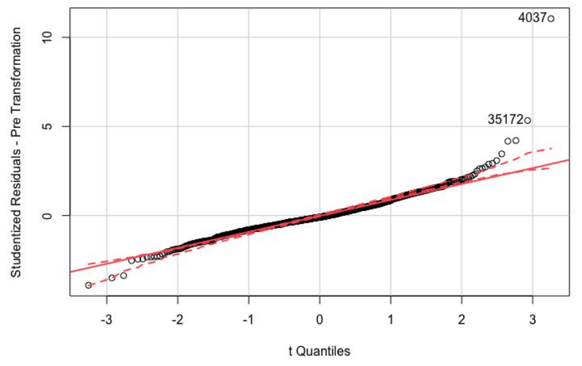
Transformation

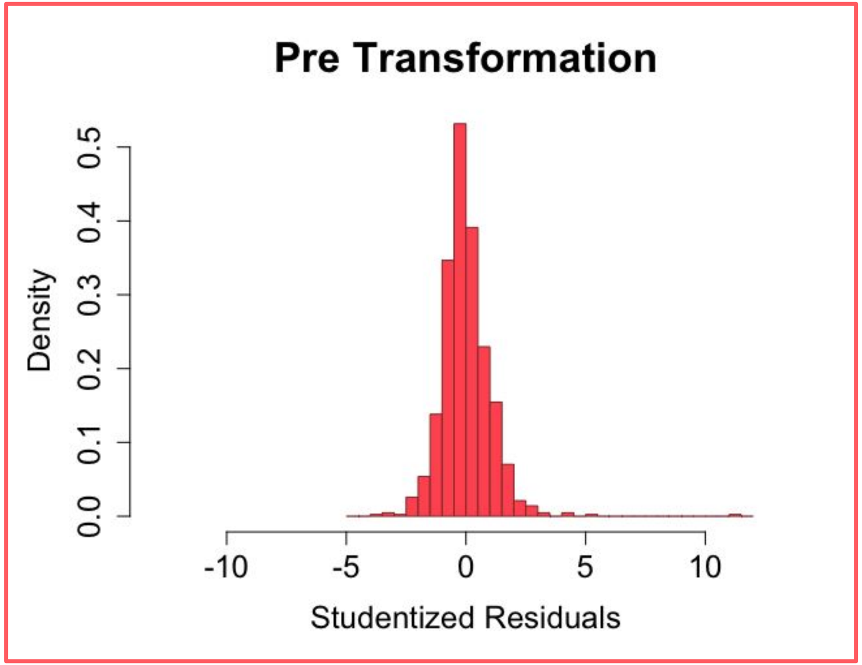
Funnel shape in the plot of the residuals against the fitted values shows non constant error variance– meaning a transformation on the response variable is necessary.

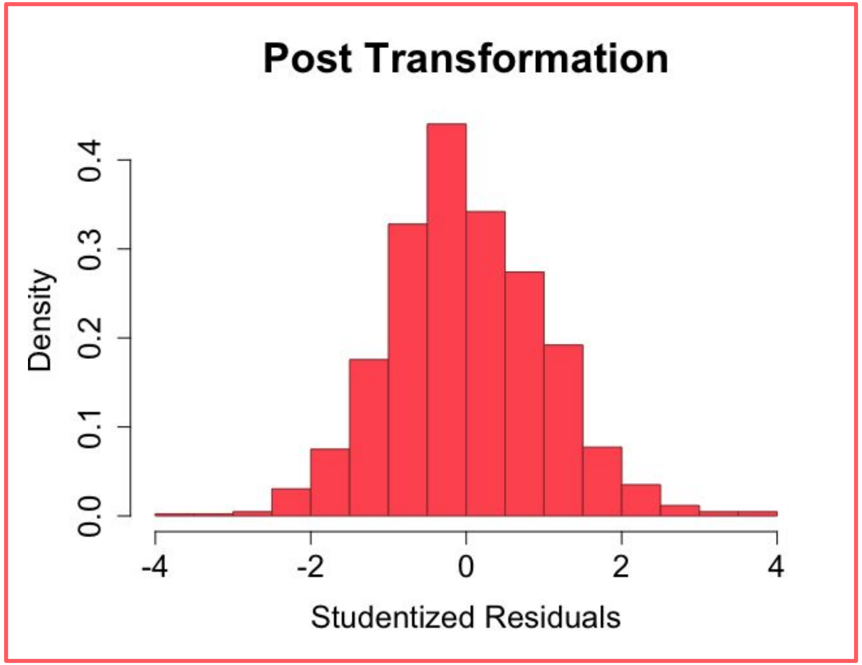
Nonlinearity in QQ plot also suggests a transformation is needed.

We perform a square root transformation on price and re-evaluate our model’s adequacy.

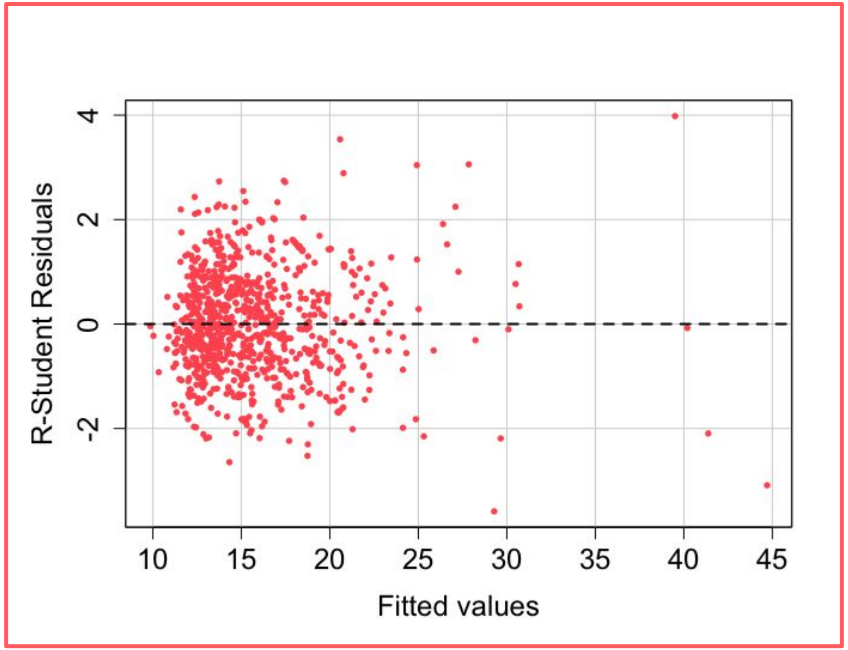


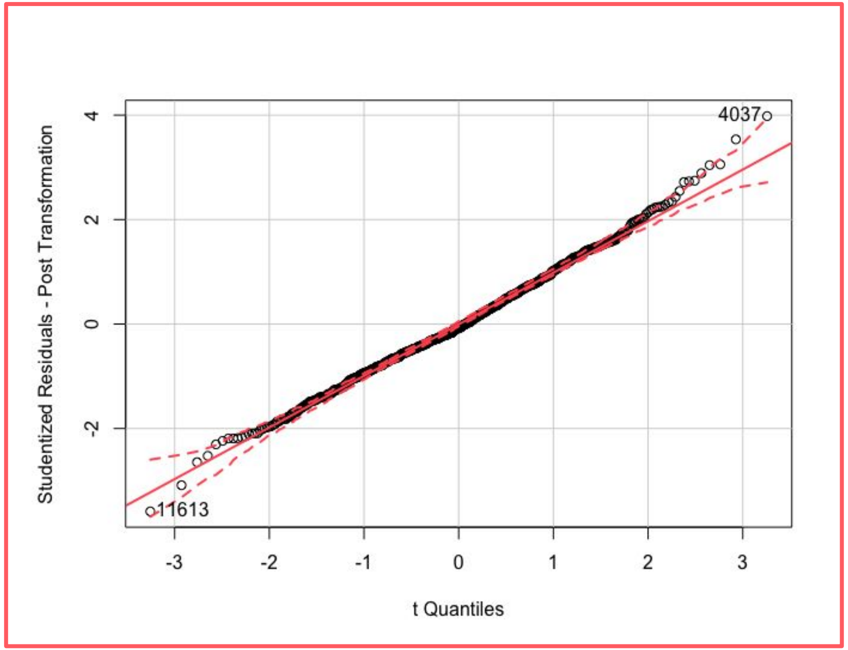




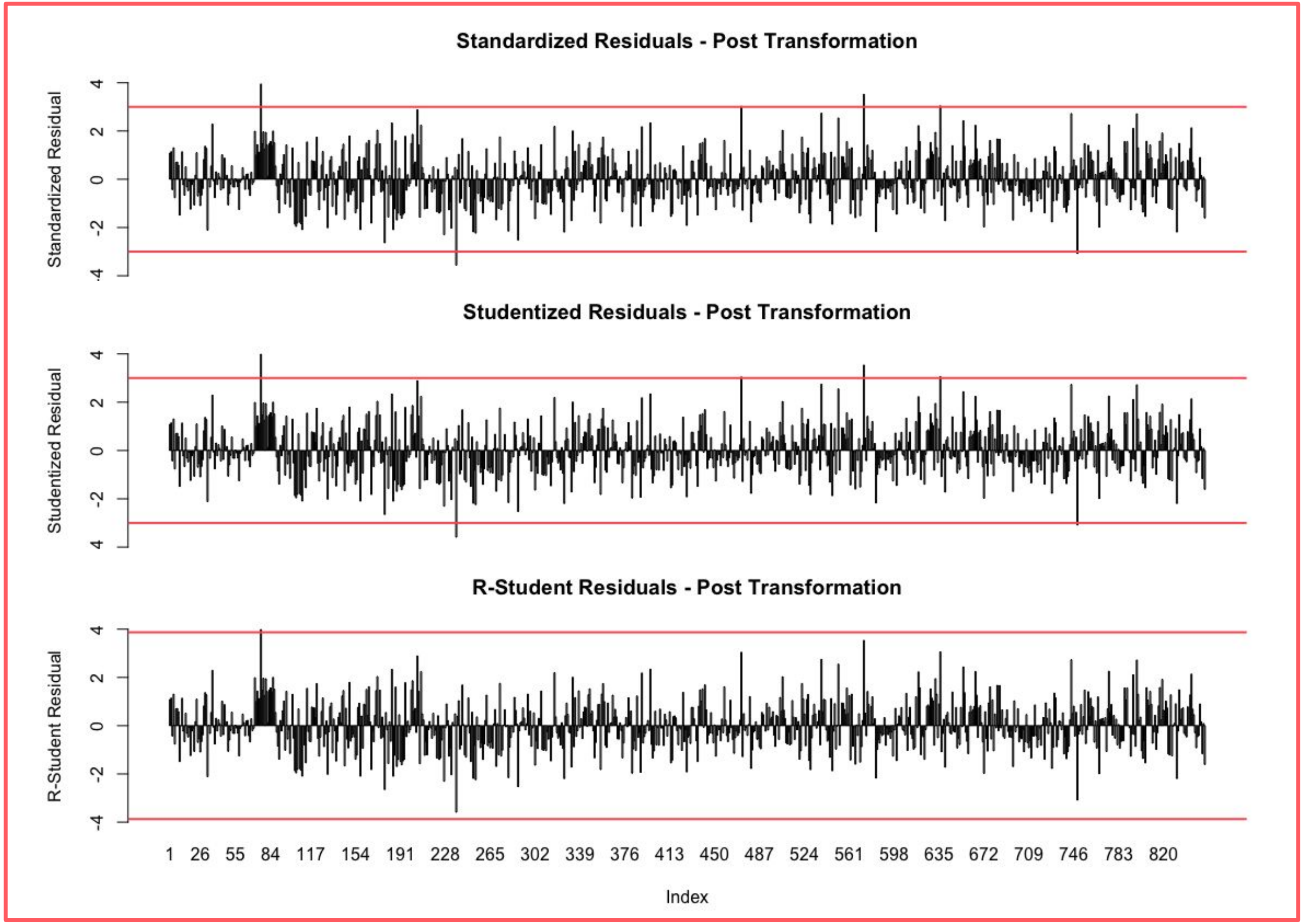


Through the data transformation, we can normalize our distribution of residuals. The second plot looks the better fit for normality. The most extreme residuals are now no larger than 4.



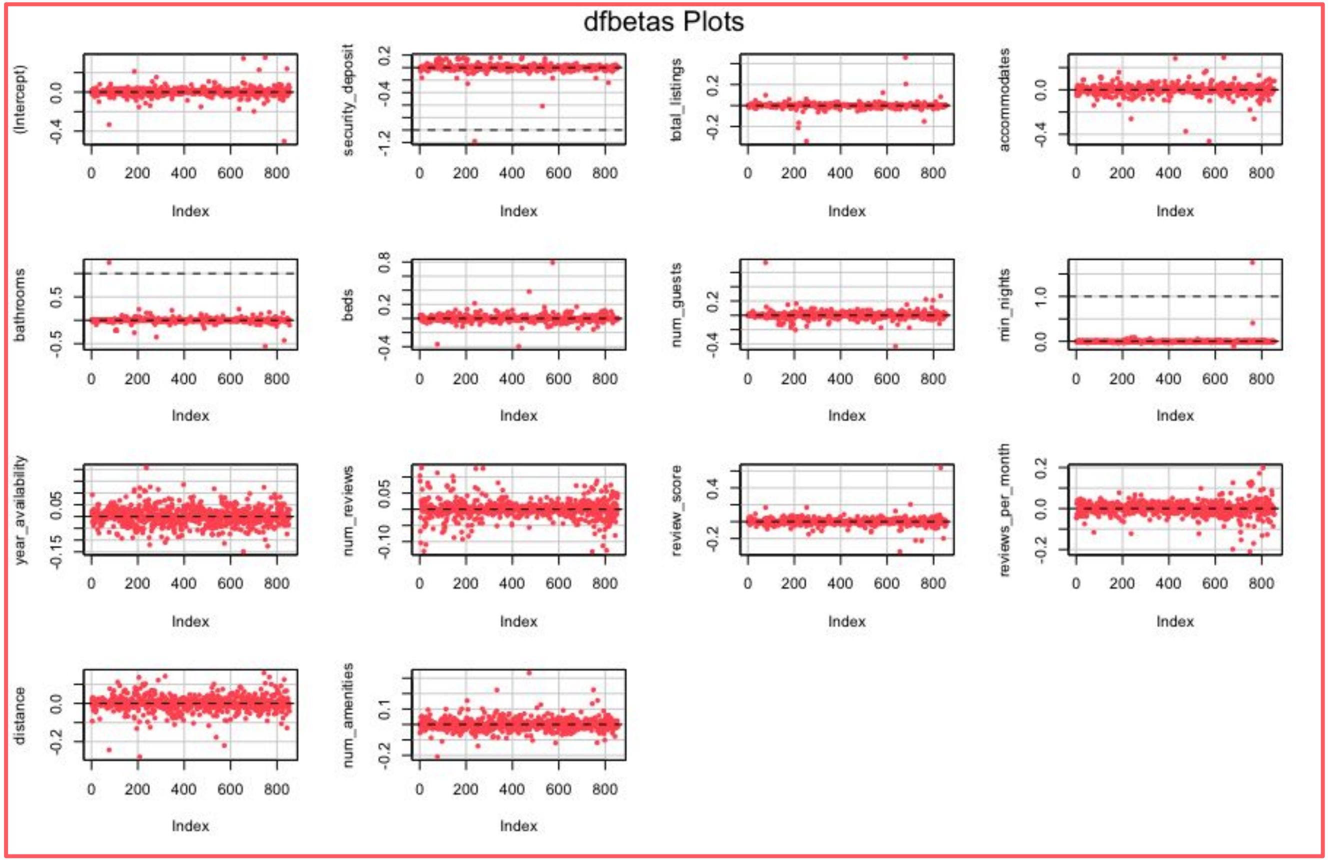


Our non constant error variance issue appears to be solved, as does our normality discrepancy. The residual plot no longer exhibits funnelling and the QQ plot is approximately linear.

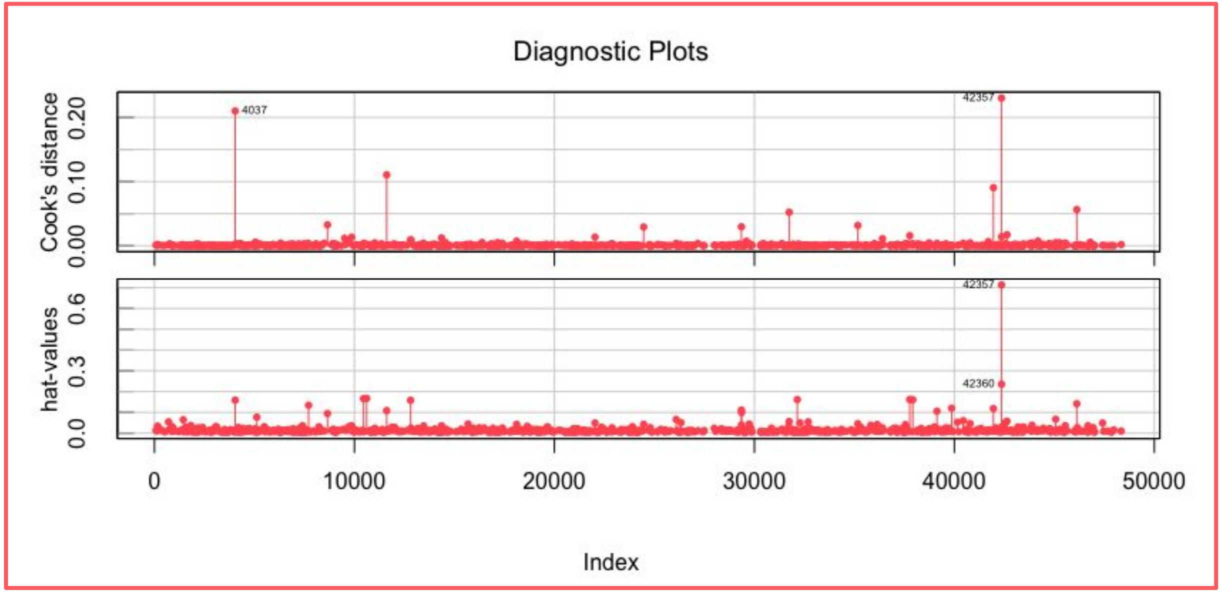


- Significant Residuals
  - 6 standardized
  - 6 studentized
  - 1 R-student
- Vast improvement over the spread of residuals prior to transformation
- The index that originally had a residual >10 is still significant for all three scaling methods, but is now just <4, which is much more reasonable.

Influential Analysis



- Three potentially influential points
  - 1 security deposit
  - 1 bathrooms
  - 1 min nights
- Observation 11613 is tied for highest security deposit at $5,000
- Observation 4037 is tied for highest number of bathrooms at 5; it is also the highest price listing in the data
- Observation 42357 has the highest number of minimum nights at 941



Observation 42357 and 42360 are highly unusual in the min_nights field, with average minimum night stays in the hundreds. Upon further investigation, we found these two listings are through the apartment rental company Sonder. Sonder typically requires longer stays than the average Airbnb listing given the properties are owned by a company rather than an individual, which could explain why these two data points are so remarkable.

Influential Measure	COVRATIO	Cook's D	DFFITS	Hat Values
No. of Potentially Influential Observations	57	0	18	31

Conclusion

Pre transformation, our model came out to be,

Price = -192.40 + 0.07(**security_deposit**) + 0.18(**total_listings**) + 26.70(**accommodates**) + 183.40(**bathrooms**) + 26.50(**beds**) + 17.54(**num_guests**) - 0.18(**min_nights**) + 0.13(**year_availability**) - 0.22(**num_reviews**) + 1.10(**review_score**) - 4.31(**reviews_per_month**) - 27.52(**distance**) + 1.22(**num_amenities**)

Post transformation, our final model is,

Sqrt(Price) = 6.05 + .0020(**security_deposit**) + 0.0051(**total_listings**) + 0.94(**accommodates**) + 2.76(**bathrooms**) + 0.63(**beds**) + 0.26(**num_guests**) - 0.0044(**min_nights**) + 0.0038(**year_availability**) - 0.0081(**num_reviews**) + 0.0191(**review_score**) - 0.15(**reviews_per_month**) - 0.73(**distance**) + 0.04(**num_amenities**)

We found that security deposit, total listings, accommodates, bathrooms, beds, number of guests, min_nights, year availability, number of reviews, review score, reviews per month, distance, and number of amenities were the most influential features in identifying the price of a one night stay in an Airbnb in Upper East Side NYC in August of 2019. In the future, we would hope to analyze data from other cities and compare which features are more significant in which regions.