Airbnb Market Insights

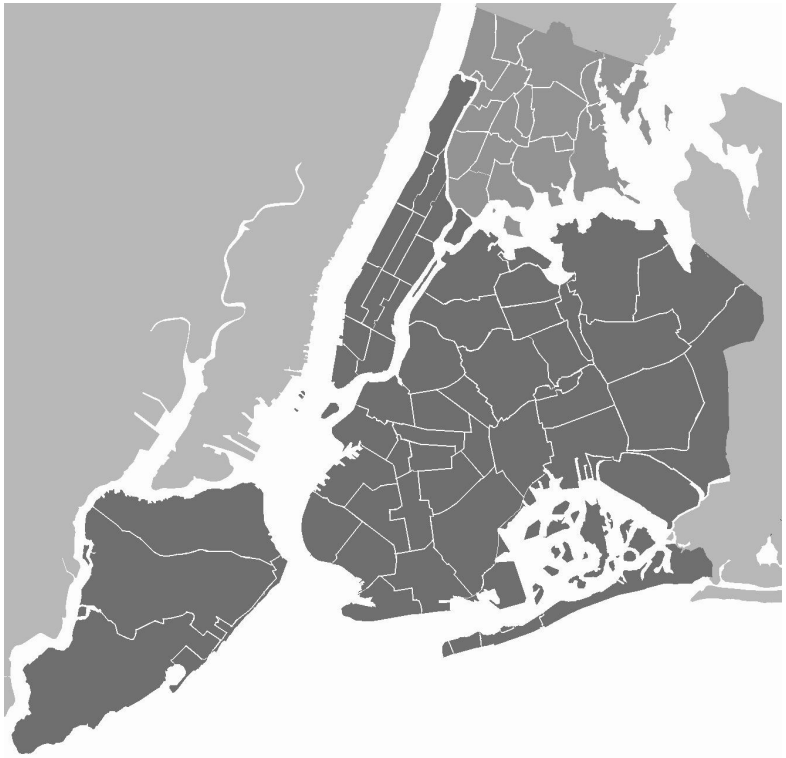
Group 3
By: Hao Yang Lin, Ariadna Sandoya
Arjun Talapatra, Benjamin Novik

Dataset

Since 2008, guests and hosts have used Airbnb to expand on traveling possibilities and present more unique, personalized way of experiencing the world. This dataset describes the listing activity and metrics in NYC, NY for 2019.

Variables: p = 16 columns

Sample size: 48,895 observations



File Name: AB_NYC_2019.csv
https://www.kaggle.com/datasets/dgomonov/new-york-city-airbnb-open-data?select=AB_NYC_2019.csv

Introduction: The Scientific Questions

1) Does the type of room (e.g., private room vs. entire home/apartment) influence the number of reviews per month?
2) Do properties in certain neighborhoods receive significantly more reviews than others monthly & overall?

Expected Findings

1. Neighborhoods with high tourist traffic (e.g., Times Square -> Manhattan) will receive more reviews.

2. Areas with limited Airbnb supply (e.g., outer boroughs -> Staten Island) receive fewer reviews.

Variable Summary

Room Count:

- Entire homes/apartments: 25,409 listings
- Private rooms: 22,326 listings
- Shared rooms: 1,160 listings
- Total: 48,895

Neighborhood Groups (Boroughs):

- Manhattan: 44% of listings
- Brooklyn: 41% of listings
- Others: 15% in Queens, Bronx, Staten Island

Average Price:

- Entire homes/apartments have the highest average price at $212.
- Private rooms average at $89.8.
- Shared rooms are the least expensive, averaging $70.1.

Average Reviews per Month:

- Private rooms have a slightly higher average number of reviews per month (1.14), followed closely by shared rooms (1.07) and entire homes/apartments (1.05).
- This slight difference may indicate that private rooms tend to receive more frequent reviews, potentially because they are more affordable and accessible.

Data Cleaning Process

1. Identifying and Handling Missing Values

Missing Data: Detected missing values in key variables (e.g., price, reviews_per_month).

2. Outlier Detection and Treatment

Extreme Values: Applied Interquartile Range (IQR) Method to detect outliers in price and other numerical columns.

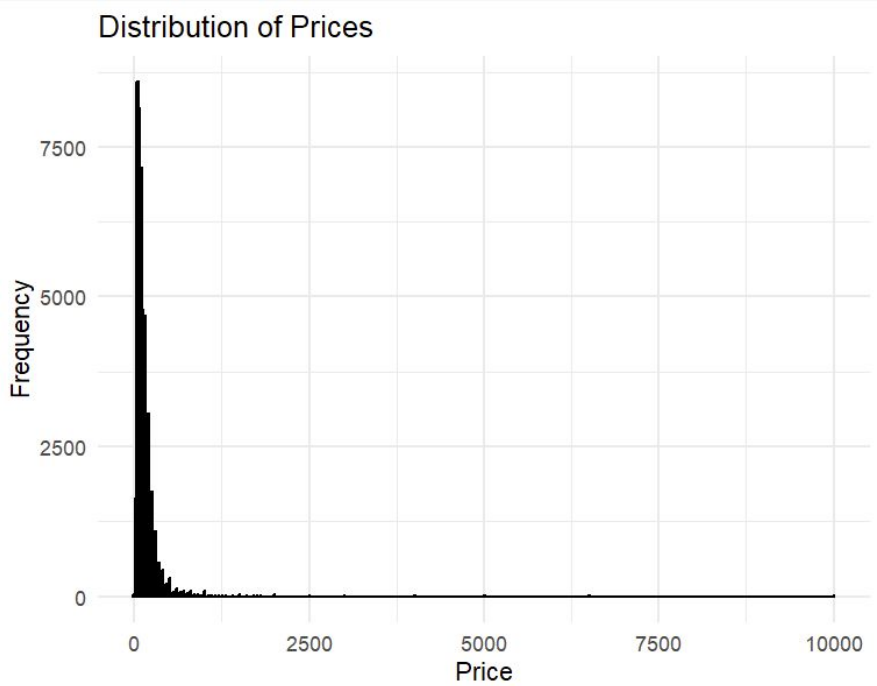
3. Data Transformation

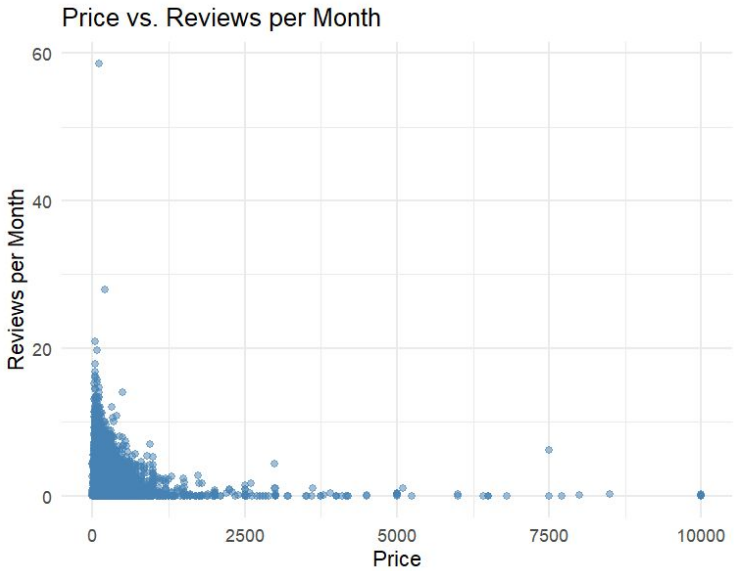
Skewed Data: Performed log transformation on price to handle right-skewed distribution.

Consistency Checks

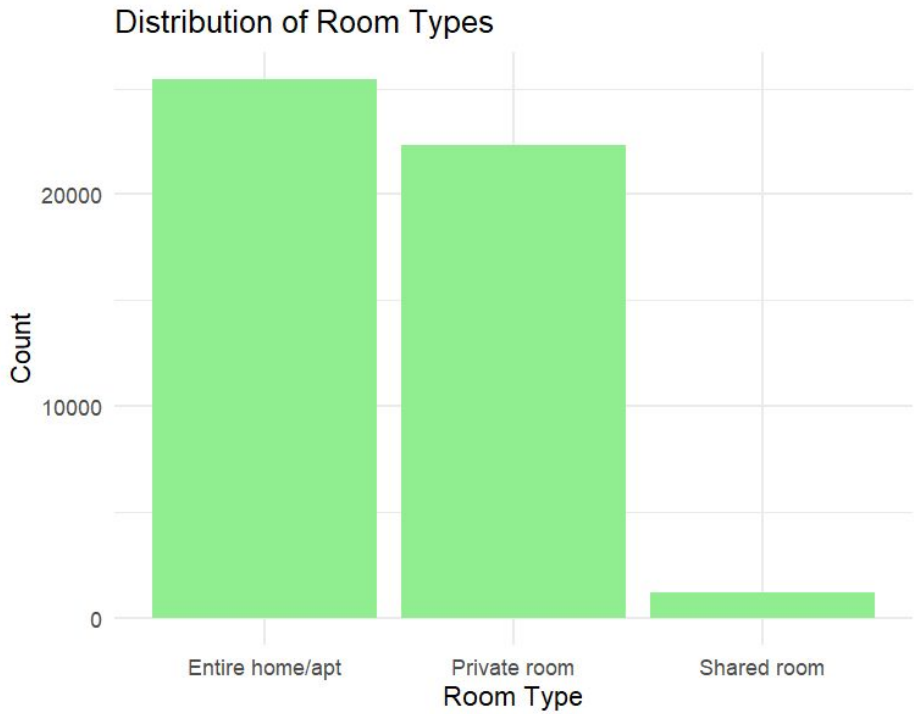
Verified consistent formatting and standardization in key categorical fields (e.g., room_type).

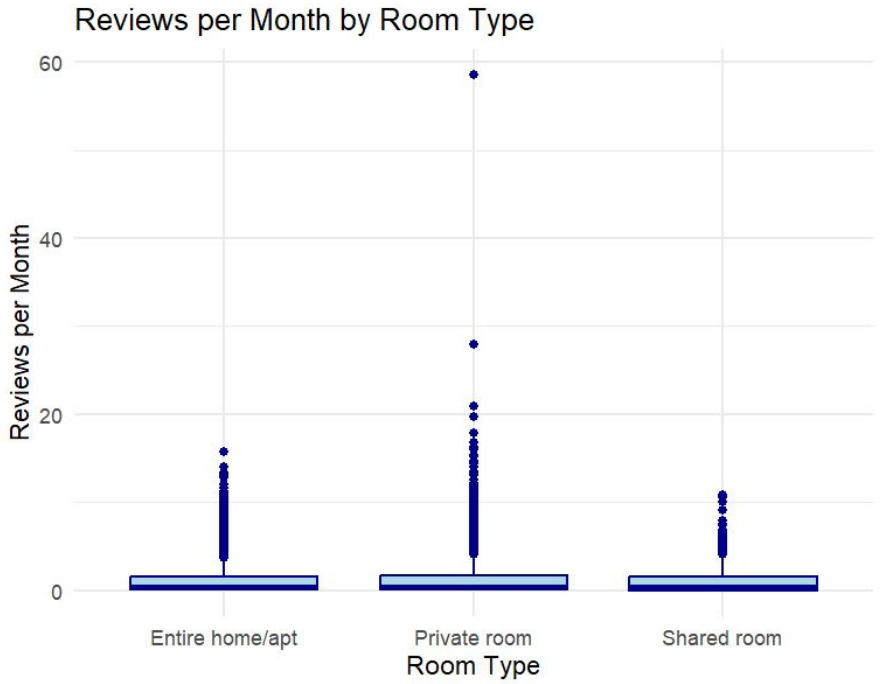
Exploratory Data Analysis





Exploratory Data Analysis (continue…)





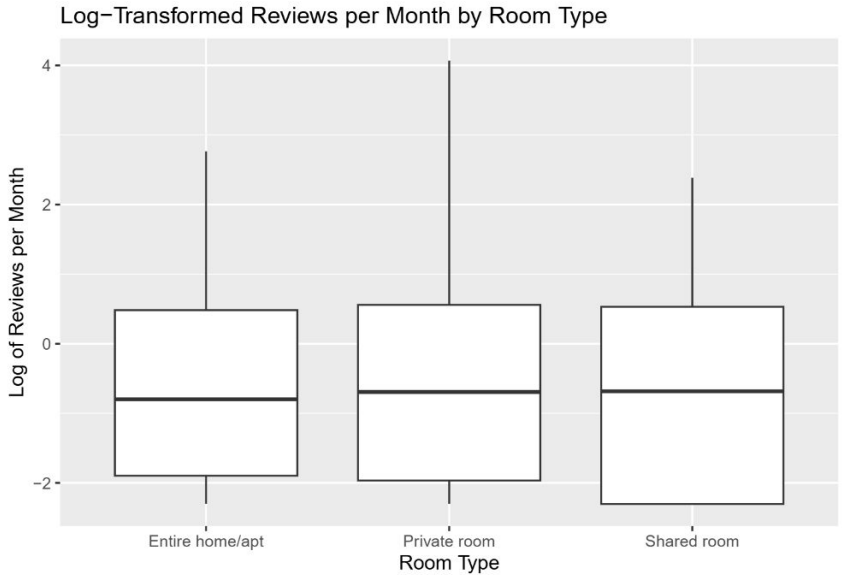
Hypothesis Test 1: ANOVA

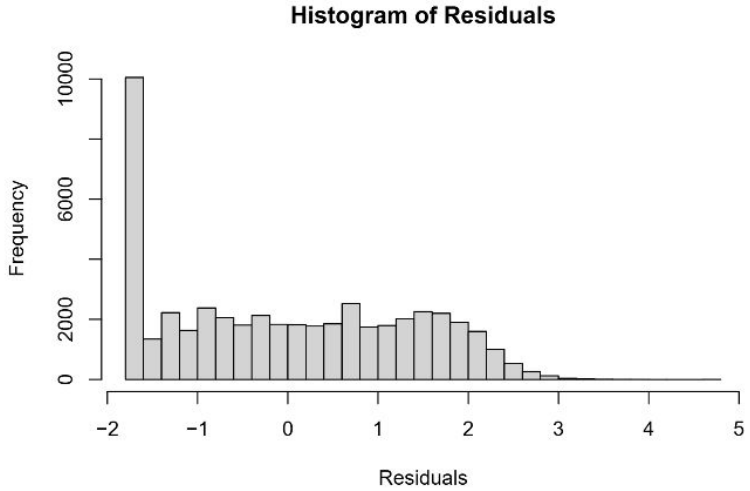
Null hypothesis: Room type has no effect on reviews/month

Alternative hypothesis: Room type has effect on reviews/month - at least one room type has a mean number of reviews per month that is significantly different from the others

```
##                Df Sum Sq Mean Sq F value   Pr(>F)
## room_type       2    114   57.23   22.45 1.79e-10 ***
## Residuals   48892 124629    2.55
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Pr(>F) and F value: reject the null hypothesis because less than significance level of 0.05.

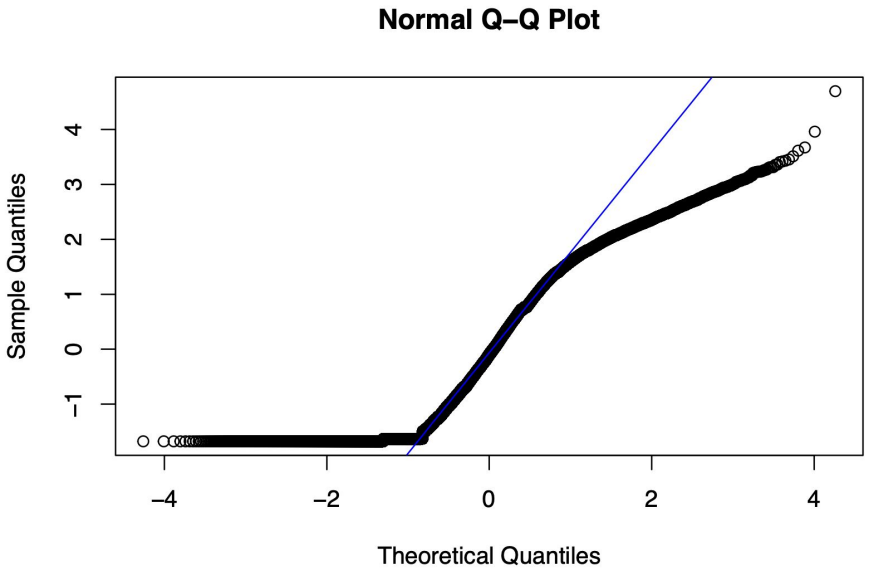





```
##
##  Asymptotic one-sample Kolmogorov-Smirnov test
##
## data:  residuals(anova_reviews)
## D = 0.10299, p-value < 2.2e-16
## alternative hypothesis: two-sided
```

ANOVA Results & Graphs

Q- Q Plot (Log Transformation)



- Residuals in the middle range align well with the diagonal line but there is deviation at the tails
- Hence, will trim data to include middle 10 percent of the data to remove outliers and improve normality

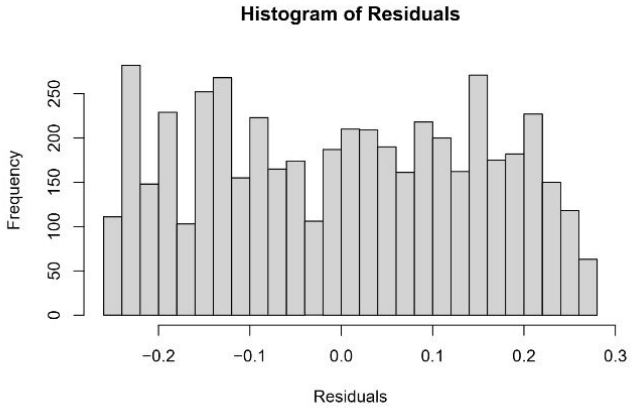
Anova(log_reviews_per_month ~ room_type)

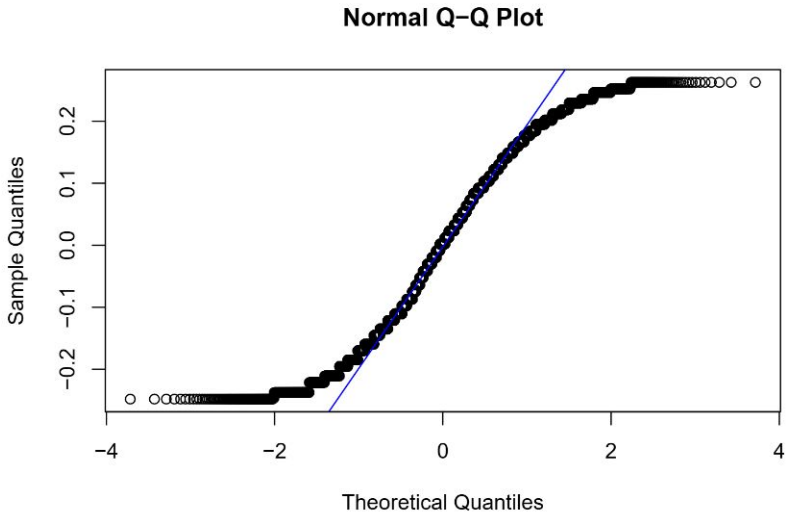
	Df	Sum Sq	Mean Sq	F value	Pr(>F)	
room_type	2	21	10.368	5.895	0.00276	**
Residuals	48892	85987	1.759			

Signif. codes: 0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Trimming the data (Middle 10% Quantile)

```
##                Df Sum Sq Mean Sq F value Pr(>F)
## room_type       2   0.14 0.06877   3.071 0.0465 *
## Residuals    4936 110.53 0.02239
```






```
##
##  Shapiro-Wilk normality test
##
## data:  residuals(anova_middle)
## W = 0.95233, p-value < 2.2e-16
```

```
## Levene's Test for Homogeneity of Variance (center = median)
##         Df F value Pr(>F)
## group    2  1.2658 0.2821
##       4936
```

Normality Assumptions:

- ANOVA assumes residuals follow normal distribution
- Trimming data showed improved normality
- Histogram of Residuals after transformation reduced skewness
- **Kolmogorov-Smirnov Test:** confirmed deviations from normality for residuals in the initial data from p-value
- **Shapiro-Wilk Test**: still indicated deviations from perfect normality
- However have very large dataset, can visually see approximately normal residuals in the center range of histogram, and used Levene's test to show homogeneity of variances: hence can bypass normality assumption

Hypothesis Conclusions:

- **ANOVA Test:** ANOVA shows there is significant association between room type and reviews per month, which remains after the transformation, although at a weaker strength
- The **F-statistic** and **p-values** from ANOVA confirm that at least one room type has a mean number of reviews that significantly differs from the rest
- Despite having residuals deviate from normality, even after the transformations. Levene's Test supports the use of ANOVA

Hypothesis Test 2: Kruskal-Wallis

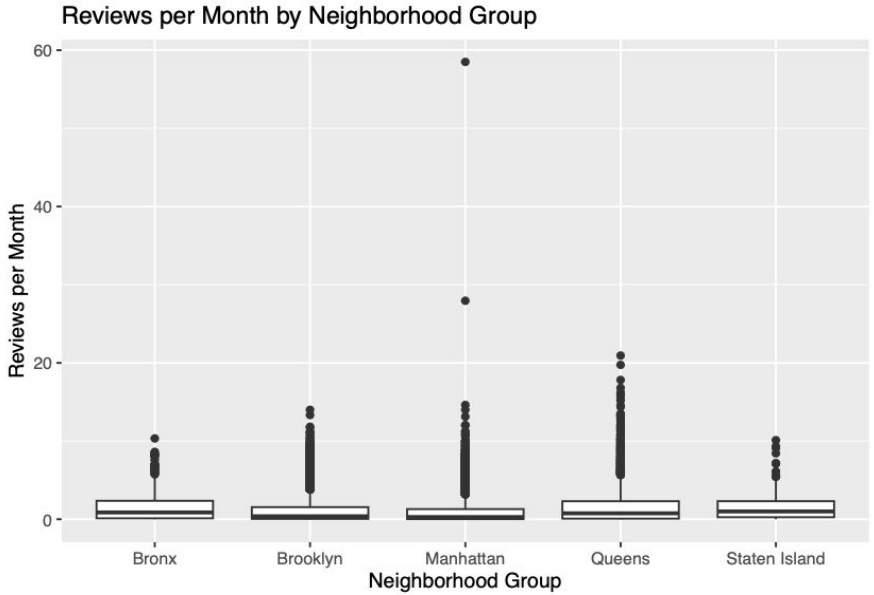
Null Hypothesis (H_0): The neighborhood of the property has no significant effect on the number of reviews received.

Alternative Hypothesis (H_1): The neighborhood of the property significantly affects the number of reviews received.

Summary Statistics for the 5 Boroughs

```
## # A tibble: 5 x 4
##   neighbourhood_group count avg_reviews_per_month median_reviews_per_month
##   <chr>               <int>                 <dbl>                    <dbl>
## 1 Bronx                1091                 1.48                      0.87
## 2 Brooklyn            20104                 1.05                      0.38
## 3 Manhattan           21661                 0.977                     0.28
## 4 Queens               5666                 1.57                      0.76
## 5 Staten Island         373                 1.58                      1
```

It can be noted that The Bronx, Queens, and Brooklyn have the highest review rate per month, however share a smaller percentage of the overall review count.



The box plot shows that Queens, The Bronx and Staten Island have a tighter spread with little to no outliers when compared to Manhattan and Brooklyn. It is important to note that some listing receive little no engagement while others receive a large number of reviews compared to others.

```
## Kruskal-Wallis chi-squared = 587.12, df = 4, p-value < 2.2e-16
```

Conducted a Kruskal-Wallis Test to determine whether or not there are statistically significant differences between the boroughs review frequencies. Given that the p-value is significantly less than .001, we can conclude a rejection of the null hypothesis, since it can be determined that there are significant differences in review frequencies between some boroughs.

Hypothesis Conclusion

It was determined from the The Kruskal-Wallis that we reject the null hypothesis, telling us that review rates vary significantly across the five boroughs.

Airbnb hosts could benefit from the location of their properties. All in all neighborhood does matter when it comes to review rates. We reject the null hypothesis and accept the alternative, the neighborhood of the property significantly affects the number of reviews received.