What listing features are associated with higher Airbnb demand from tourists in NYC and Asheville? (Part 1: NYC) by Sol Vloebergh

2025-04-15

Introduction

New York City stands as one of the most vibrant and visited destinations in the world, attracting millions of tourists each year with its rich culture, iconic landmarks, and diverse neighborhoods. As demand for flexible, authentic accommodations has grown, Airbnb has emerged as a popular alternative to traditional hotels, offering visitors the opportunity to experience the city from a more local perspective.

With thousands of listings available—varying widely in price, location, and property type—understanding the factors that drive guest demand has become critical for hosts seeking to succeed in this competitive market.

In this analysis, we focus on uncovering which listing features are associated with higher guest engagement in New York City. Using data from the Inside Airbnb platform, we carefully cleaned and prepared the dataset, engineered new variables to capture key attributes like pricing, availability, and proximity to tourist hotspots, and performed exploratory spatial analysis to visualize patterns across neighborhoods.

Building on these steps, we developed a logistic regression model to predict which listings are most likely to achieve high demand, as measured by review activity. Our findings provide insights into how strategic decisions regarding room type, pricing, and location can influence Airbnb performance in one of the most competitive short-term rental markets in the world.

1. Data Preparation and Feature Engineering

1.1 Selecting and Cleaning Key Variables

```
# Load raw data and select relevant columns
listings <- read.csv("listings.csv", stringsAsFactors = FALSE)
vars_to_keep <- c("id", "neighbourhood", "room_type", "price",
                  "availability_365", "number_of_reviews",
                  "reviews_per_month", "latitude", "longitude")
listings <- listings[, vars_to_keep]

# Clean price column (remove symbols and convert to numeric)
listings$price <- as.numeric(gsub("[$,]", "", listings$price))
```

We focused only on columns relevant to our analysis, such as price, room type, reviews, and availability. The price column was cleaned to remove dollar signs and commas and converted to numeric format.

1.2 Filtering Invalid or Inactive Listings

```
# Filter out listings with missing or zero values
listings <- listings[
  !is.na(listings$price) & listings$price > 0 &
  !is.na(listings$reviews_per_month) & listings$reviews_per_month > 0 &
  !is.na(listings$availability_365) & listings$availability_365 > 0, ]
```

To focus on meaningful activity, we removed listings with missing or zero values for price, reviews, or availability. This ensured that our analysis only included listings with active engagement.

1.3 Feature Engineering

```
# Log-transformed price variable to reduce skewness
listings$log_price <- log(listings$price)

# Binary indicator for high-demand listings
median_reviews <- median(listings$reviews_per_month)
listings$high_demand <- ifelse(listings$reviews_per_month > median_reviews, 1, 0)

# Remove incomplete rows
listings <- listings[complete.cases(listings), ]
```

To normalize the skewed price distribution, we created a log_price variable. We also introduced a high_demand binary indicator, where listings with above-median review frequency were labeled as high demand.

1.4 Defining Tourist Proximity

```
# Define a vector of tourist-heavy neighborhoods
tourist_neighs <- c("Midtown", "Harlem", "Williamsburg", "East Village",
                    "SoHo", "Lower East Side", "Upper West Side",
                    "Chelsea", "Greenwich Village")

# Flag listings in those neighborhoods
listings$near_tourist_area <- ifelse(listings$neighbourhood %in% tourist_neighs, 1, 0)

# Quick count of listings by tourist area
table(listings$near_tourist_area)
```

```
##
##     0     1
## 11718  3684
```

Since one of our main hypotheses is that proximity to tourist areas may influence demand, we created a binary variable near_tourist_area. This flag identifies listings in iconic neighborhoods such as Midtown, Harlem, or Williamsburg.

Out of 15,402 listings, 3,684 (24%) were located in tourist-heavy neighborhoods, while 11,718 (76%) were located outside of these zones. This variable allows us to compare listing performance between these two groups in later visualizations.

```
## 1.5 Descriptive Statistics for Key Variables

# Summary statistics
summary(listings$price)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##     7.0    87.0   133.0   180.9   207.0 10271.0
```

```
summary(listings$availability_365)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##     1.0   136.0   246.0   230.1   337.0   365.0
```

```
# Room type distribution
table(listings$room_type)
```

```
##
## Entire home/apt      Hotel room    Private room     Shared room
##            8816             108            6421              57
```

```
prop.table(table(listings$room_type)) * 100
```

```
##
## Entire home/apt      Hotel room    Private room     Shared room
##      57.2393196       0.7012076      41.6893910       0.3700818
```

Before moving into modeling, we explored a few summary statistics to better understand the overall structure of the cleaned dataset. The average listing price was around $143, and listings were available for about 212 days per year, on average. Most properties were either entire homes (57%) or private rooms (42%). These patterns gave us early insight into how key features like price, availability, and room type vary across listings. With this context in place, we moved forward to quantify how these characteristics relate to listing demand using a logistic regression model.

```
# Final Data Set
summary(listings)
```

```
##        id            neighbourhood       room_type             price
##  Min.   :6.848e+03   Length:15402       Length:15402       Min.   :    7.0
##  1st Qu.:2.960e+07   Class :character   Class :character   1st Qu.:   87.0
##  Median :5.349e+07   Mode  :character   Mode  :character   Median :  133.0
##  Mean   :4.255e+17                                         Mean   :  180.9
##  3rd Qu.:8.623e+17                                         3rd Qu.:  207.0
##  Max.   :1.348e+18                                         Max.   :10271.0
##  availability_365 number_of_reviews reviews_per_month    latitude
##  Min.   :  1.0    Min.   :   1.00   Min.   :  0.010   Min.   :40.50
##  1st Qu.:136.0    1st Qu.:   4.00   1st Qu.:  0.190   1st Qu.:40.68
```

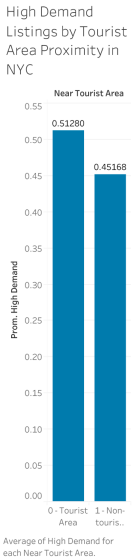
```
##  Median :246.0    Median :  18.00   Median :  0.500   Median :40.72
##  Mean   :230.1    Mean   :  49.27   Mean   :  1.158   Mean   :40.73
##  3rd Qu.:337.0    3rd Qu.:  58.00   3rd Qu.:  1.490   3rd Qu.:40.76
##  Max.   :365.0    Max.   :2749.00   Max.   :117.980   Max.   :40.91
##    longitude        log_price      high_demand     near_tourist_area
##  Min.   :-74.25   Min.   :1.946   Min.   :0.0000   Min.   :0.0000
##  1st Qu.:-73.98   1st Qu.:4.466   1st Qu.:0.0000   1st Qu.:0.0000
##  Median :-73.95   Median :4.890   Median :0.0000   Median :0.0000
##  Mean   :-73.94   Mean   :4.920   Mean   :0.4982   Mean   :0.2392
##  3rd Qu.:-73.92   3rd Qu.:5.333   3rd Qu.:1.0000   3rd Qu.:0.0000
##  Max.   :-73.71   Max.   :9.237   Max.   :1.0000   Max.   :1.0000
```

```
write.csv(listings, "listings_cleaned_NYC.csv", row.names = FALSE)
```

After cleaning and transforming the Airbnb listings data, we saved the resulting dataset as a CSV file to use it for visualization in Tableau. The cleaned dataset includes only the relevant variables and newly created fields such as log_price, high_demand, and near_tourist_area, ensuring that the visualizations in Tableau accurately reflect the analysis objectives. By exporting this version, we maintain consistency between the statistical analysis and the interactive visual exploration.

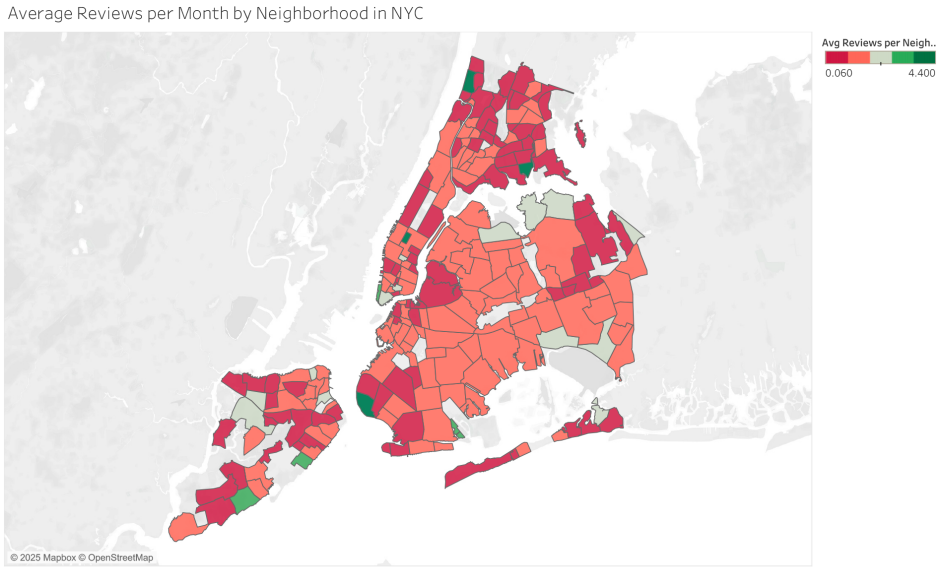
2. Exploratory Spatial Analysis: Mapping Neighborhood Boundaries

2.1 Bar Chart: High Demand Listings by Tourist Area Proximity in NYC



This bar chart compares the proportion of high-demand listings between tourist and non-tourist neighborhoods in New York City. The results show that approximately 51% of listings in tourist areas have above-median review activity, compared to 45% in non-tourist areas. While the difference is moderate, it supports the hypothesis that location near popular attractions is positively associated with Airbnb listing performance.

2.2 Choropleth Map: Average Reviews per Month by Neighborhood in NYC

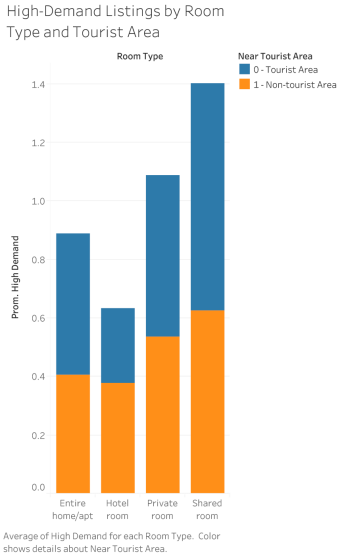


To better understand the spatial distribution of demand for Airbnb listings in New York City, we created a choropleth map in Tableau that visualizes the average number of monthly reviews per listing across neighborhoods. Using the GeoJSON file containing the geometries of 233 NYC neighborhoods, we joined this spatial dataset with our cleaned Airbnb listings data based on neighborhood names. A fixed-level calculated field was used to aggregate reviews_per_month by neighborhood, ensuring accurate spatial averaging. We also created a second calculated field to count the number of listings in each neighborhood, which appears in the tooltip.

In the map, neighborhoods are color-coded on a diverging scale from red to green, where red represents lower review activity and green indicates higher engagement. The majority of neighborhoods fall in the lower to mid-range of average monthly reviews (between 0.5 and 2), with only a few neighborhoods—mostly in parts of Brooklyn and northern Manhattan—reaching higher levels (above 4 reviews per month). Blank areas correspond to neighborhoods with no listings after data filtering and were retained to preserve spatial context.

This visualization reveals that high Airbnb demand is concentrated in just a few neighborhoods, while most areas experience moderate or low levels of guest activity.

2.3 Grouped Bar Chart: High-Demand Listings by Room Type and Tourist

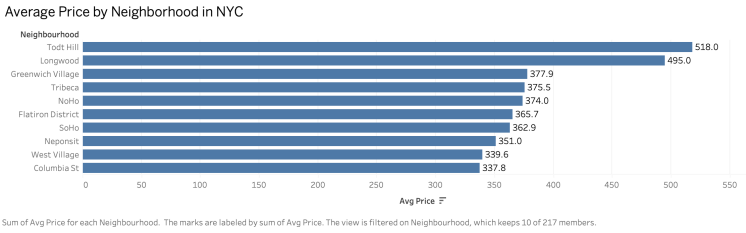


To explore how both room type and location influence Airbnb listing performance, we created a grouped bar chart showing the percentage of high-demand listings by room_type, separated by whether or not the listing is located in a tourist-heavy neighborhood. The high_demand variable was defined as listings with a reviews_per_month value above the citywide median. The results reveal that shared and private rooms in non-tourist areas tend to have the highest proportion of high-demand listings, with values exceeding 70%. In contrast, hotel rooms and entire homes near tourist zones show more moderate demand. This suggests that affordability and listing type may outweigh location when it comes to driving guest engagement, particularly in dense urban markets like New York City.

2.4 Bar Chart: Average Price by Neighborhood in NYC

```r
# Calculate average price by neighborhood
avg_price_neigh <- listings %>%
  group_by(neighbourhood) %>%
  summarise(avg_price = mean(price, na.rm = TRUE)) %>%
  arrange(desc(avg_price))

# Save to CSV for Tableau visualization
write.csv(avg_price_neigh, "avg_price_by_neighbourhood.csv", row.names = FALSE)
```



To better understand how pricing varies across neighborhoods, we created a summary table that calculates the average listing price in each NYC neighborhood. This summary was exported and visualized in Tableau to show pricing disparities across boroughs and highlight potential hotspots of higher-cost accommodations.

The bar chart above shows the ten neighborhoods in New York City with the highest average Airbnb listing prices. Leading the list is Todt Hill with an average price of $518, followed closely by Longwood at $495. These neighborhoods stand out as luxury areas with significantly higher nightly rates compared to the citywide average.

Other high-price areas include well-known upscale Manhattan neighborhoods such as Greenwich Village, Tribeca, NoHo, and SoHo, all averaging over $360 per night.

This pricing pattern highlights the geographic disparity in Airbnb costs across NYC and provides valuable context for understanding how location influences both price and potential demand. Hosts in these areas may benefit from the prestige and centrality of their listings, while guests may pay a premium for convenience and ambiance.

3. Predictive Modeling: Logistic Regression to explore High-Demand Listings

Following our visual analysis, we developed a logistic regression model to further explore how listing characteristics relate to demand. This model builds on the patterns observed in Tableau by examining how variables like log price, room type, availability, and tourist proximity contribute to the likelihood that a listing is high demand.

```
# Logistic regression model
model <- glm(high_demand ~ log_price + room_type + availability_365 + near_tourist_area,
             data = listings, family = binomial)
summary(model)
```

```
##
## Call:
## glm(formula = high_demand ~ log_price + room_type + availability_365 +
##     near_tourist_area, family = binomial, data = listings)
##
## Coefficients:
##                           Estimate Std. Error z value Pr(>|z|)
## (Intercept)             -1.6417992  0.1498270 -10.958  < 2e-16 ***
## log_price                0.4030600  0.0285734  14.106  < 2e-16 ***
## room_typeHotel room     -0.7505696  0.2144894  -3.499 0.000466 ***
## room_typePrivate room    0.5899683  0.0386139  15.279  < 2e-16 ***
## room_typeShared room     1.5537926  0.3137840   4.952 7.35e-07 ***
## availability_365        -0.0022147  0.0001494 -14.819  < 2e-16 ***
## near_tourist_area       -0.3507921  0.0395775  -8.863  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 21352  on 15401  degrees of freedom
## Residual deviance: 20771  on 15395  degrees of freedom
## AIC: 20785
##
## Number of Fisher Scoring iterations: 4
```

The results show that listings with higher log-transformed prices are significantly more likely to be high demand, with a coefficient of 0.403 ($p < 0.001$), suggesting that as price increases (on a log scale), so does the likelihood of receiving above-median review activity.

Room type also matters: compared to entire homes (the baseline), hotel rooms (0.750), private rooms (0.589), and especially shared rooms (1.553) are all positively associated with higher demand, and all are statistically significant ($p < 0.001$).

Interestingly, availability_365 has a small but negative coefficient (-0.0022, $p < 0.001$), implying that simply being available more days does not necessarily increase demand. One of the most notable findings is that listings in tourist-heavy neighborhoods show a negative relationship with high demand (-0.350, $p < 0.001$). This could suggest that competition is higher in those areas, or that lesser-known neighborhoods are capturing more interest from certain travelers.

Additionally, the model's AIC was 20,785.18, suggesting a good balance between model complexity and predictive ability. While AIC values are primarily used for model comparison, the result here supports that our selected predictors provide a strong and efficient structure for modeling high-demand listings.

```
# Odds ratios and 95% confidence intervals
exp(coef(model))
```

```
##           (Intercept)             log_price   room_typeHotel room
##             0.1936313             1.4963966             0.4720976
## room_typePrivate room  room_typeShared room      availability_365
##             1.8039312             4.7293727             0.9977878
##     near_tourist_area
##             0.7041301
```

```
exp(confint(model))
```

```
## Waiting for profiling to be done...
```

```
##                           2.5 %    97.5 %
## (Intercept)           0.1442708 0.2595748
## log_price             1.4150933 1.5828218
## room_typeHotel room   0.3061298 0.7117860
## room_typePrivate room 1.6726317 1.9459848
## room_typeShared room  2.6218078 9.0570462
## availability_365      0.9974952 0.9980798
## near_tourist_area     0.6515176 0.7608636
```

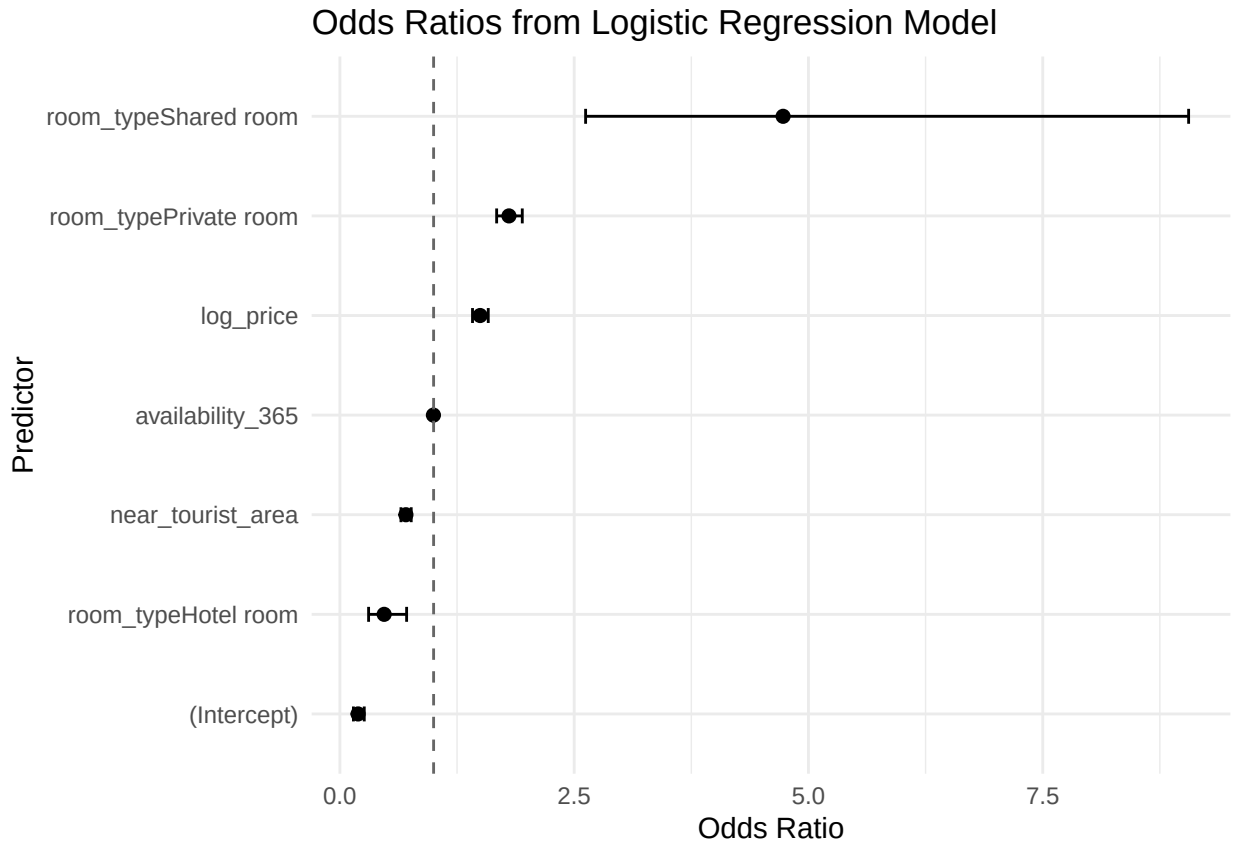
The odds ratios confirmed that listings with higher log prices and certain room types—especially shared and private rooms—are more likely to be in high demand. For instance, shared rooms are over 4.7 times more likely to be high demand compared to entire homes. On the other hand, hotel rooms are associated with lower demand. Notably, listings in tourist-heavy areas are about 35% less likely to be high demand, possibly due to higher competition or market saturation in those zones.

```
# Odds ratio plot with 95% CI
coef_table <- summary(model)$coefficients
odds <- exp(coef(model))
conf <- exp(confint(model))
```

```
## Waiting for profiling to be done...
```

```r
odds_df <- data.frame(
  Variable = rownames(coef_table),
  OR = odds,
  Lower = conf[, 1],
  Upper = conf[, 2]
)

# Visualizing Odds Ratios
ggplot(odds_df, aes(x = reorder(Variable, OR), y = OR)) +
  geom_point(size = 2) +
  geom_errorbar(aes(ymin = Lower, ymax = Upper), width = 0.15) +
  geom_hline(yintercept = 1, linetype = "dashed", color = "gray40") +
  coord_flip() +
  labs(
    title = "Odds Ratios from Logistic Regression Model",
    x = "Predictor",
    y = "Odds Ratio"
  ) +
  theme_minimal()
```



This plot visualizes the estimated odds ratios and their 95% confidence intervals for each predictor in the logistic regression model. Values above 1 indicate a positive association with high demand, while values below 1 indicate a negative association. For example, shared rooms have the highest odds ratio, reinforcing their strong link to demand, while hotel rooms are less likely to be high demand compared to entire homes. The dashed line at 1 serves as a reference—predictors with intervals that do not cross this line are considered statistically significant.

This visual summary complements our earlier interpretation and helps identify which variables most strongly influence listing performance. It also helps stakeholders quickly grasp which factors most strongly influence listing visibility, guiding data-driven decisions for Airbnb hosts.

```
# Likelihood Ratio Test for model fit
anova(model, test = "Chisq")
```

```
## Analysis of Deviance Table
##
## Model: binomial, link: logit
##
## Response: high_demand
##
## Terms added sequentially (first to last)
##
##
##                   Df Deviance Resid. Df Resid. Dev  Pr(>Chi)
## NULL                                15401      21352
## log_price          1   24.217     15400      21327 8.607e-07 ***
## room_type          3  268.160     15397      21059 < 2.2e-16 ***
## availability_365   1  208.788     15396      20850 < 2.2e-16 ***
## near_tourist_area  1   79.157     15395      20771 < 2.2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

To assess the overall model fit, we performed a likelihood ratio test comparing the full logistic model to a null model with no predictors. The results showed a significant reduction in deviance ($p < 0.001$) for each of the predictors—log-transformed price, room type, availability, and tourist proximity—indicating that each variable contributes meaningfully to explaining variation in listing demand. This confirms that the model is not only interpretable, but also statistically robust in terms of overall fit.

4. Conclusion

This section of the project focused on understanding which Airbnb listing features are associated with higher demand in New York City by tourists, using reviews per month as a proxy for visibility. After cleaning the dataset and creating key variables such as log_price, high_demand, and near_tourist_area, we explored spatial patterns and modeled the impact of listing characteristics using logistic regression.

To complement the modeling, we visualized the geographic distribution of both demand and pricing. The choropleth map revealed that high average review activity is concentrated in only a few neighborhoods—mainly in parts of Brooklyn and Manhattan—while most areas show moderate or low guest engagement. Similarly, the bar chart of average price showed that upscale areas like Todt Hill, Greenwich Village, and Tribeca have the highest nightly rates, often exceeding $350.

However, our regression model indicated that price and location alone do not fully explain listing performance. Listings offering private or shared rooms—particularly in non-tourist-heavy areas—were significantly more likely to be in high demand. Additionally, availability had a modest but statistically significant negative effect, suggesting that simply being available for more days does not guarantee greater guest engagement. Interestingly, being located in a tourist-heavy neighborhood was also negatively associated with demand, potentially due to market saturation and competition.

Overall, the findings suggest that Airbnb success in NYC is not solely driven by location or luxury pricing. Instead, listings that balance affordability, room type, and strategic availability can outperform more expensive properties in central areas. Hosts can increase demand by optimizing features that are within their control, regardless of neighborhood prestige.