★★★★★

BIA 672 Final Presentation

Mohammed Abdul Aasim

Part - I

Company Information

★★★★★

- Founded: 2008, San Francisco, CA
- Online marketplace connecting hosts and guests for short-term stays and experiences
- Global Reach: 220+ countries and regions, 4M+ hosts, 150M+ users
- Offerings:
  - Vacation rentals (rooms, homes, unique stays)
  - Airbnb Experiences (tours, events)
  - Airbnb for Work (corporate travel)
- Revenue Model: Service fees from bookings (hosts and guests)
- "Create a world where anyone can belong anywhere"

What is the Problem?

Problem Statement: Our goal is to analyze how location affects occupancy rates for Airbnb listings in Paris. While some neighborhoods experience consistently high bookings, others may struggle due to pricing, competition, or lack of tourist appeal.

Key Question to Address:

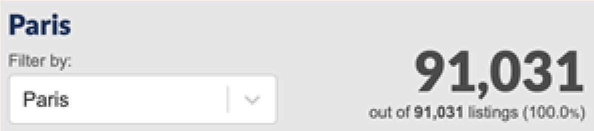
- Which neighborhoods in Paris have the highest and lowest Airbnb occupancy rates?

Additional Questions that can be answered through this group project:

- Are occupancy rates correlated with proximity to popular tourist attractions?
- Are some neighborhoods oversaturated with listings, leading to lower occupancy due to competition?
- How do different room types compare in terms of popularity and preference?

Location

- Paris, France
- 91,031 listings (based on insideairbnb)
- Known as the "City of Light" and the "City of Love"
- One of the most visited cities in the world, attracting **over 30 million tourists annually**



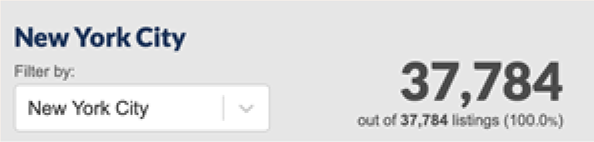
Paris

Filter by:

Paris

91,031

out of **91,031** listings (100.0%)



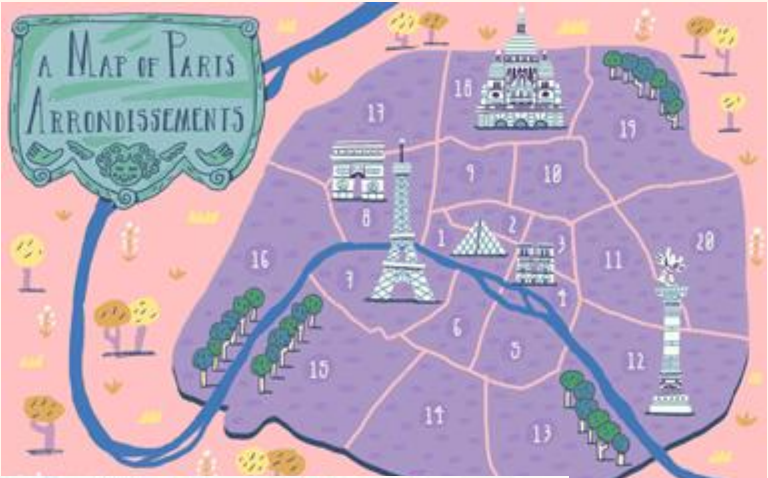
New York City

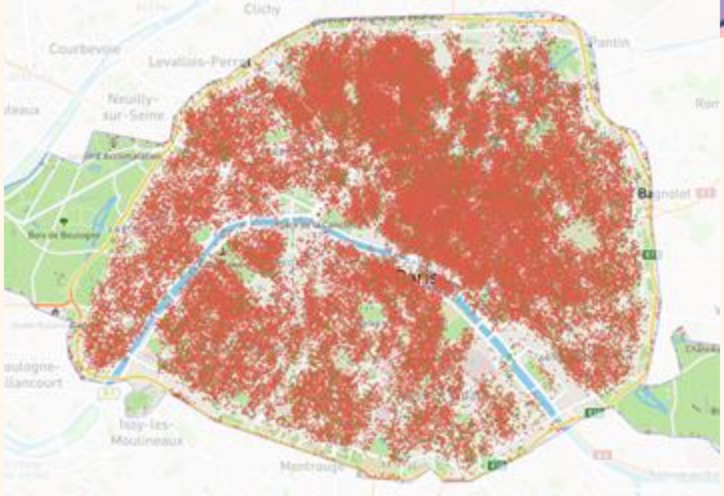
Filter by:

New York City

37,784

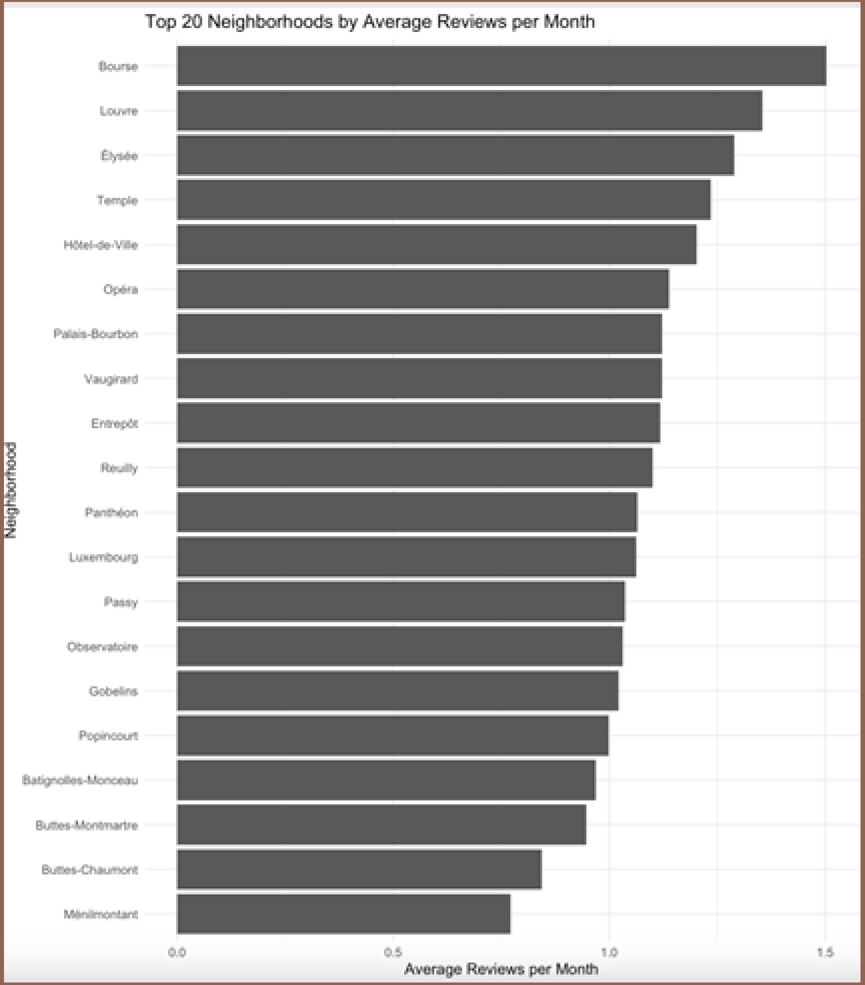
out of **37,784** listings (100.0%)



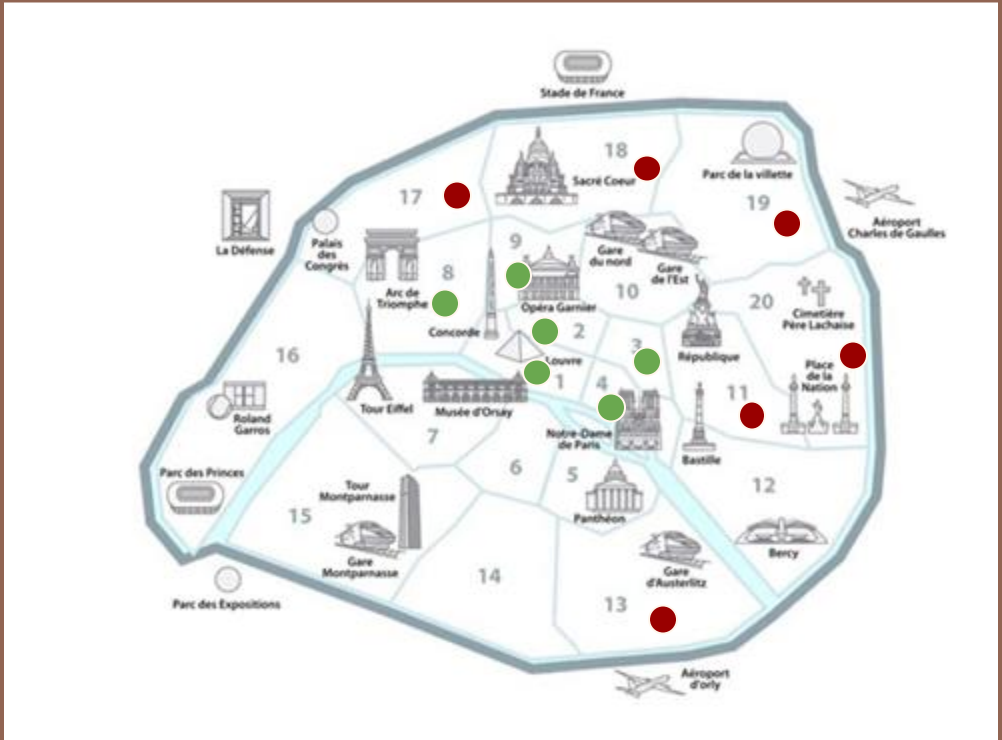


Part - II

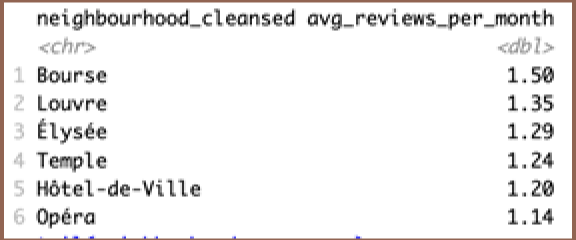
Highest and Lowest activity in the neighborhoods, based on reviews



Highest and Lowest activity in the neighborhoods, based on reviews Visualized

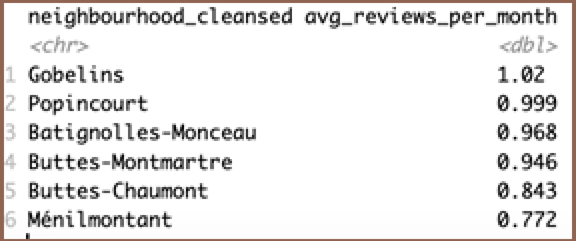


Highest



```
  neighbourhood_cleansed avg_reviews_per_month
  <chr>                                  <dbl>
1 Bourse                                  1.50
2 Louvre                                  1.35
3 Élysée                                  1.29
4 Temple                                  1.24
5 Hôtel-de-Ville                          1.20
6 Opéra                                   1.14
```

Lowest

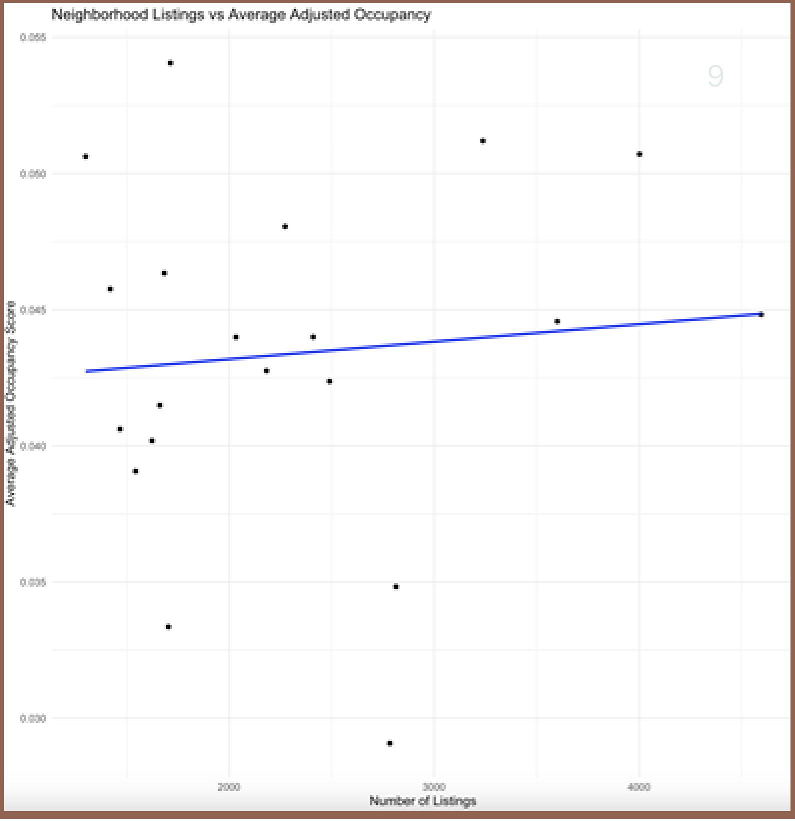


```
 neighbourhood_cleansed avg_reviews_per_month
 <chr>                                  <dbl>
1 Gobelins                               1.02
2 Popincourt                             0.999
3 Batignolles-Monceau                    0.968
4 Buttes-Montmartre                      0.946
5 Buttes-Chaumont                        0.843
6 Ménilmontant                           0.772
```

Highest: 2nd, 1st, 8th, 3rd, 4th, 9th | | Lowest: 13th, 11th, 17th, 18th, 19th, 20th

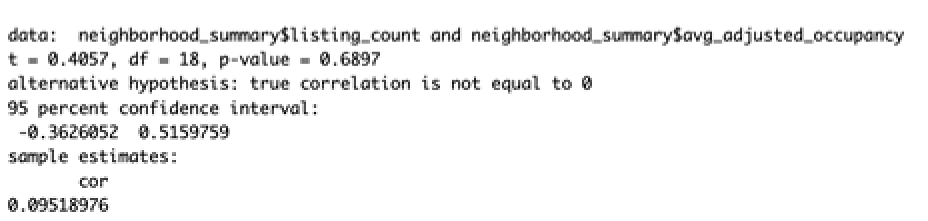
Number of Listings Vs Average Adjusted Occupancy

- There is a very slight positive trend (more listings = slightly higher occupancy).
- But the points are very spread out → no strong clear pattern visually.



- No strong evidence that the number of listings affects occupancy.
- More listings in a neighborhood does NOT significantly hurt or help occupancy.

- Correlation (r) = 0.0952 | Very weak positive relationship
- p-value = 0.6897 | Not statistically significant (way > 0.05)
- 95% confidence interval = -0.36 to 0.52 | Includes 0, meaning no real correlation

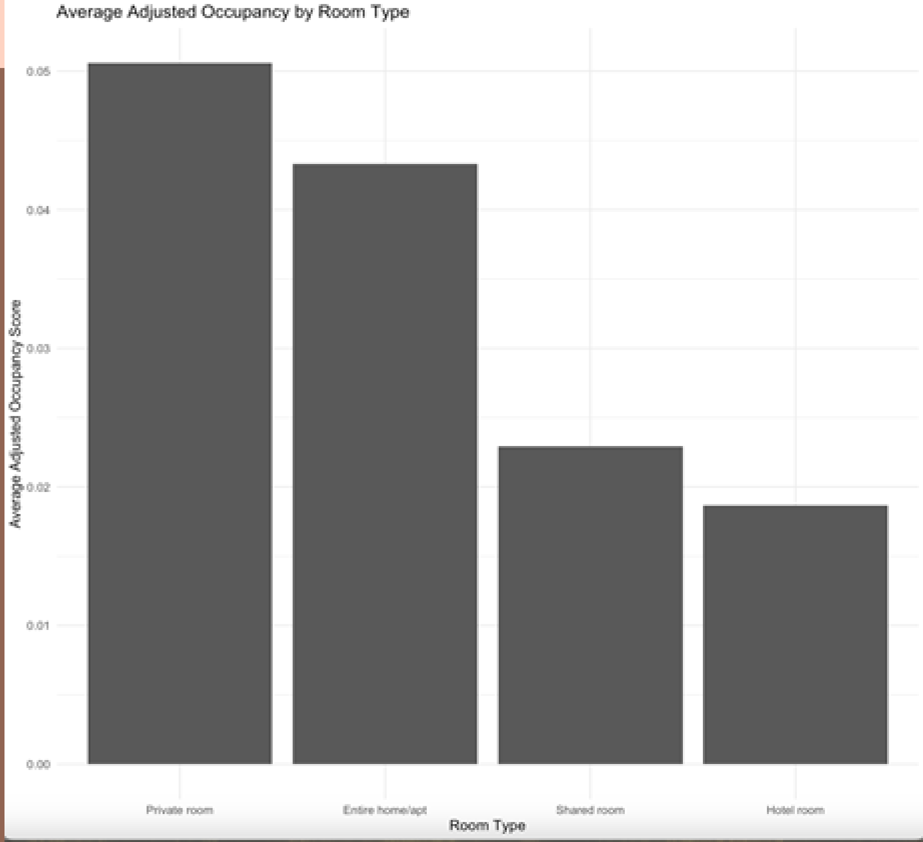


```
data:  neighborhood_summary$listing_count and neighborhood_summary$avg_adjusted_occupancy
t = 0.4057, df = 18, p-value = 0.6897
alternative hypothesis: true correlation is not equal to 0
95 percent confidence interval:
 -0.3626052  0.5159759
sample estimates:
       cor
0.09518976
```

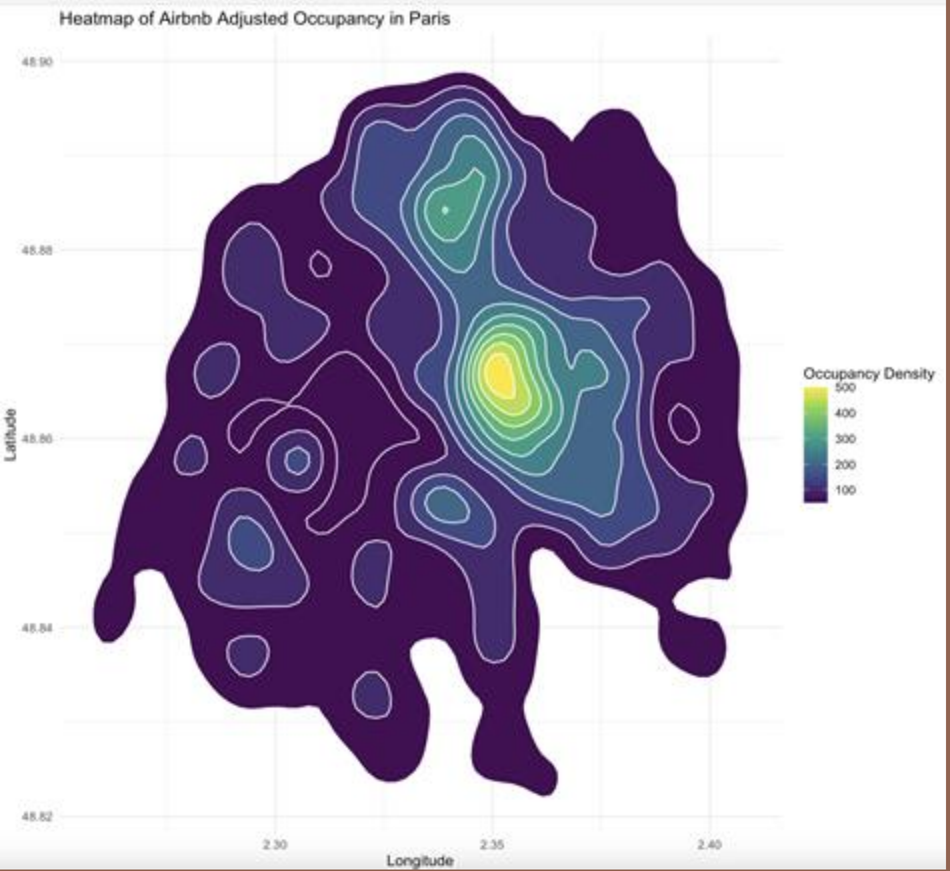
Room Types

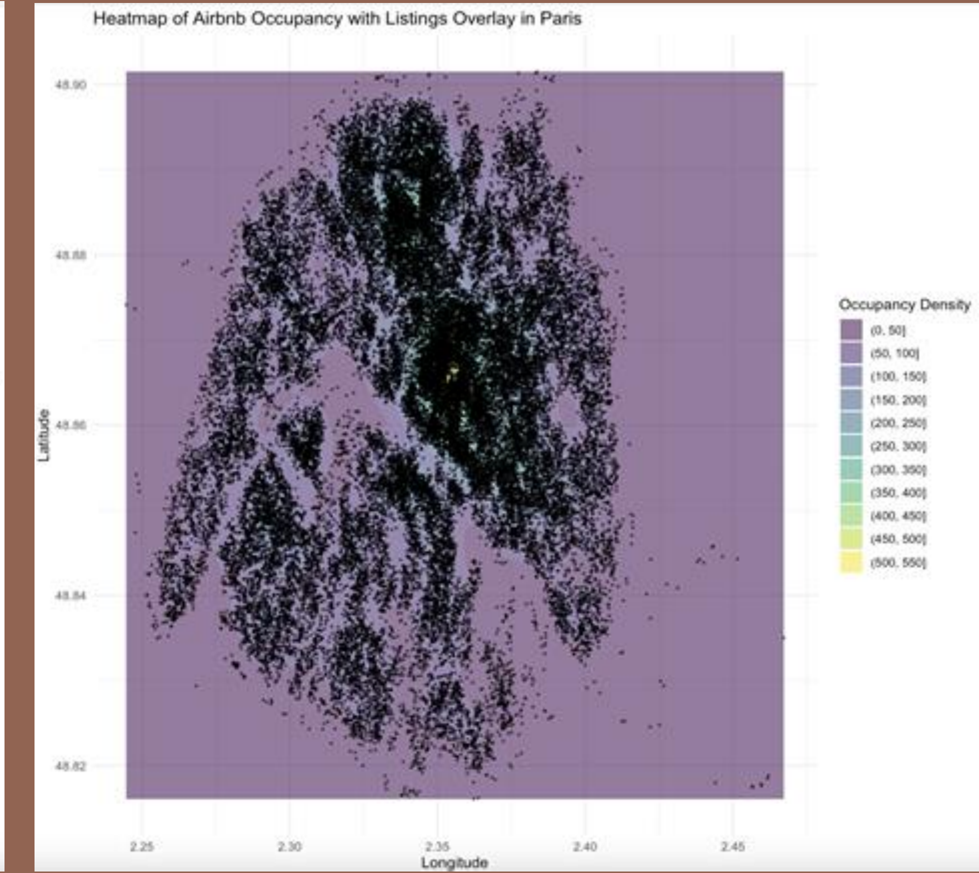
- Private rooms have the highest average occupancy rate among all Airbnb listings in Paris
- Followed by entire homes/apartments.
- Shared rooms and hotel rooms perform significantly worse.
- This suggests that affordability combined with privacy is highly valued by guests, and listings offering private spaces are more likely to achieve higher occupancy.

This matches broader Airbnb trends, guests often balance privacy and price when choosing.



HeatMaps
Eiffel Tower's:Latitude: 48.8584° N and Longitude: 2.2945° E





Statistical Analysis

Do Neighborhoods and Room Types Affect Airbnb Occupancy in Paris?

One-Way ANOVA:

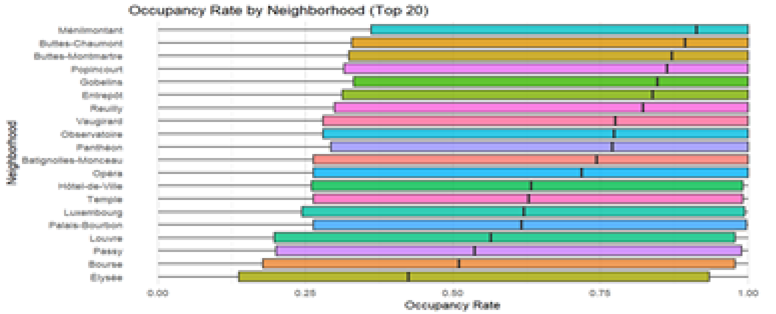
Test: occupancy_rate ~ neighbourhood_cleansed

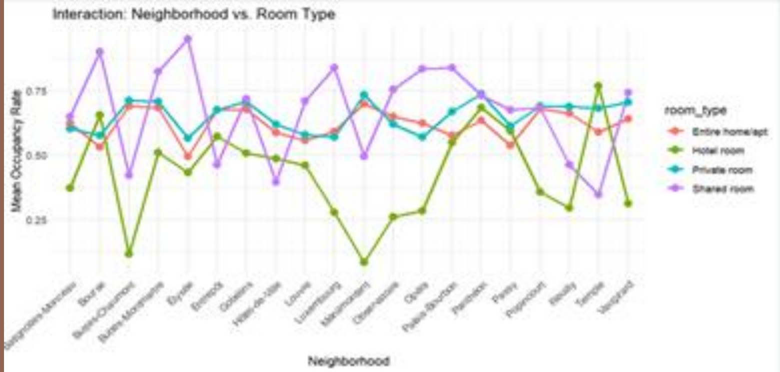
- F(19, 91011) = 106.8, $p < 2e-16$
- $\eta^2 = 0.0218$ → Small effect

Two-Way ANOVA (With Interaction):

Test: occupancy_rate ~ neighbourhood_cleansed & room_type

- All terms significant:
  - ○ Neighborhood ($\eta^2 = 0.0218$)
  - ○ Room Type ($\eta^2 = 0.0023$)
  - ○ Interaction ($\eta^2 = 0.0036$)





Post-Hoc: Tukey HSD Test:

- Élysée significantly outperforms Ménilmontant, Gobelin, Popincourt.
- Helps identify top-performing zones for marketing/pricing.

Kruskal-Wallis Test: Confirms differences across neighborhoods

(p < 2.2e-16)

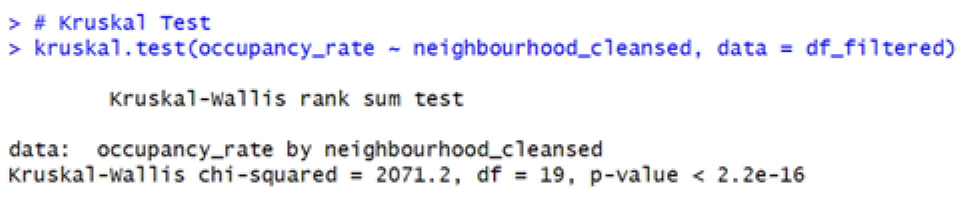
Q-Q Plot: Minor tail deviations – normality is acceptable

Boxplot:

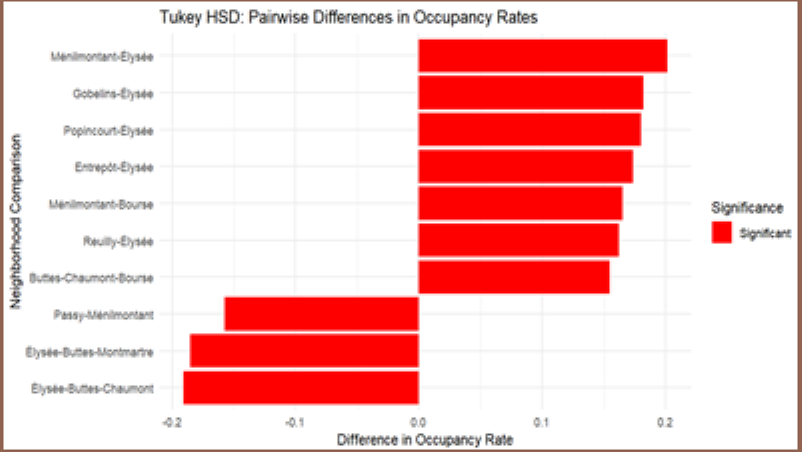
- Élysée: High & consistent occupancy
- Buttes-Chaumont, Ménilmontant: High variability

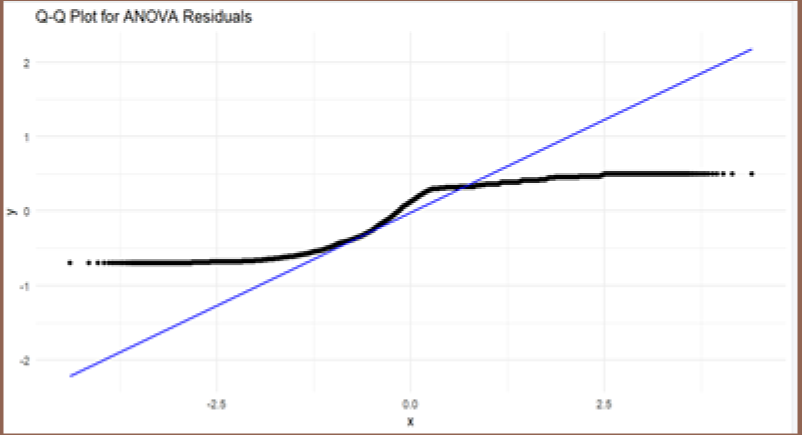
Interaction Plot:

- Shared rooms = highly location-sensitive
- Private & entire homes = stable
- Hotel rooms = inconsistent



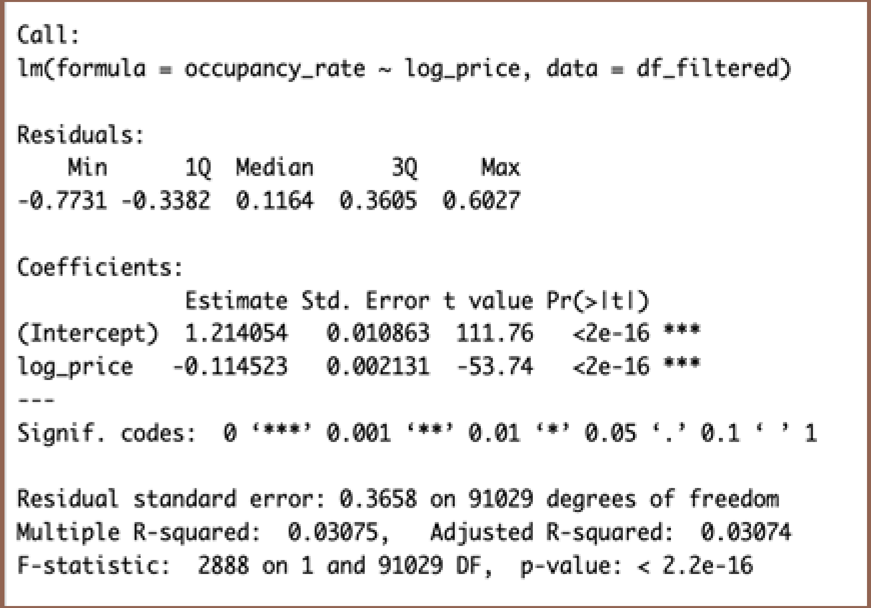
```
> # Kruskal Test
> kruskal.test(occupancy_rate ~ neighbourhood_cleansed, data = df_filtered)

	Kruskal-Wallis rank sum test

data:  occupancy_rate by neighbourhood_cleansed
Kruskal-Wallis chi-squared = 2071.2, df = 19, p-value < 2.2e-16
```



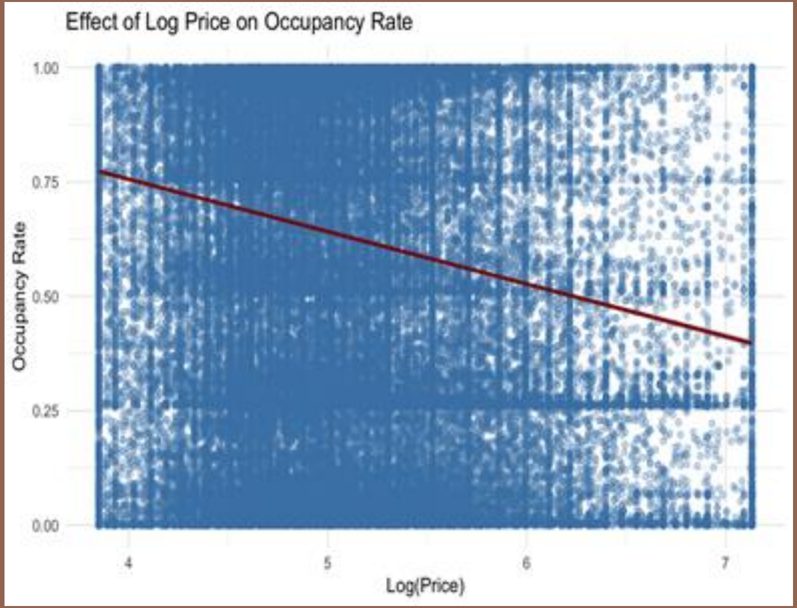


Correlation & Regression Analysis

Does increasing price lower bookings?

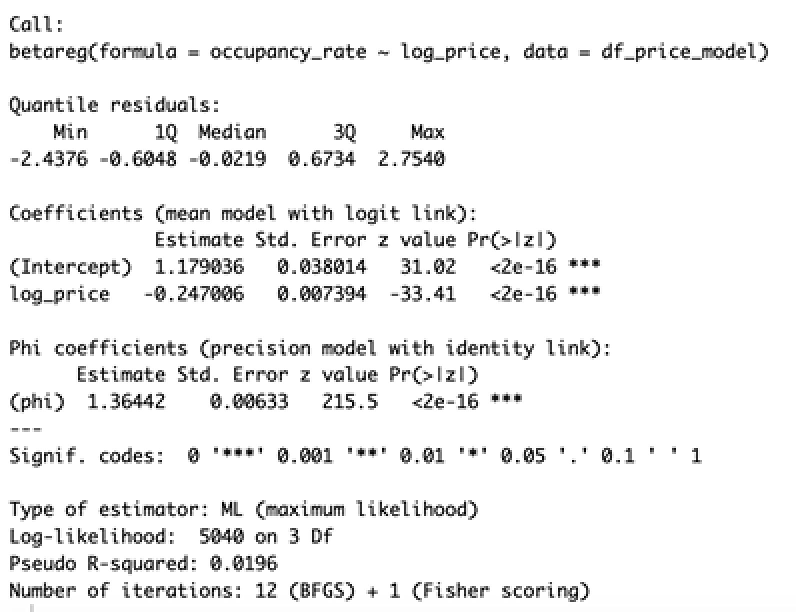


```
Call:
lm(formula = occupancy_rate ~ log_price, data = df_filtered)

Residuals:
    Min      1Q  Median      3Q     Max 
-0.7731 -0.3382  0.1164  0.3605  0.6027 

Coefficients:
             Estimate Std. Error t value Pr(>|t|)    
(Intercept)  1.214054   0.010863  111.76   <2e-16 ***
log_price   -0.114523   0.002131  -53.74   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.3658 on 91029 degrees of freedom
Multiple R-squared:  0.03075,	Adjusted R-squared:  0.03074 
F-statistic:  2888 on 1 and 91029 DF,  p-value: < 2.2e-16
```

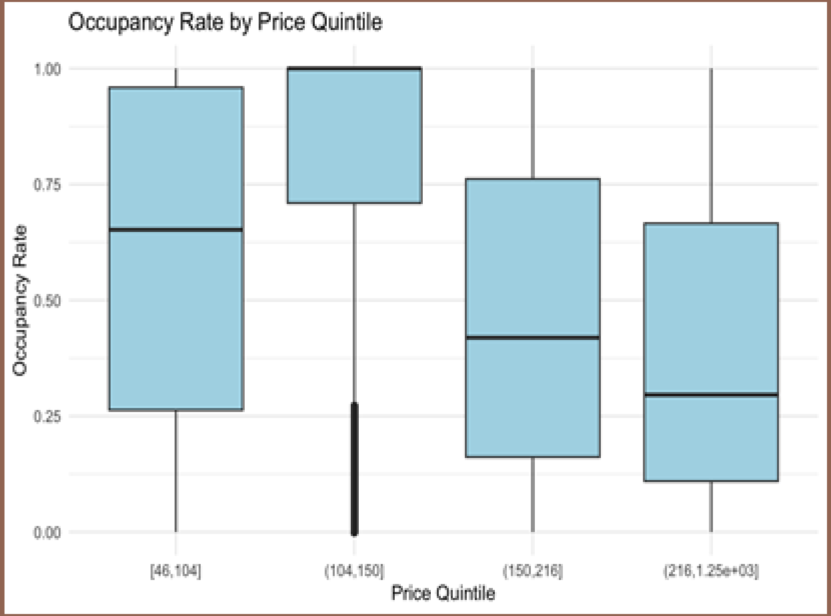


Yes — the analysis shows that as log(price) increases, occupancy rate significantly decreases.

- The regression coefficient for log_price is -0.1145, meaning higher prices are associated with lower booking rates.
- This negative effect is highly statistically significant (p-value < 2e-16).
- The scatter plot with the trend line visually confirms this downward trend in occupancy as prices rise.



```
Call:
betareg(formula = occupancy_rate ~ log_price, data = df_price_model)

Quantile residuals:
    Min      1Q  Median      3Q     Max
-2.4376 -0.6048 -0.0219  0.6734  2.7540

Coefficients (mean model with logit link):
             Estimate Std. Error z value Pr(>|z|)
(Intercept)  1.179036   0.038014   31.02   <2e-16 ***
log_price   -0.247006   0.007394  -33.41   <2e-16 ***

Phi coefficients (precision model with identity link):
      Estimate Std. Error z value Pr(>|z|)
(phi)  1.36442    0.00633   215.5   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Type of estimator: ML (maximum likelihood)
Log-likelihood:  5040 on 3 Df
Pseudo R-squared: 0.0196
Number of iterations: 12 (BFGS) + 1 (Fisher scoring)
```



Yes — Beta regression confirms a strong negative relationship between log(price) and occupancy rate.

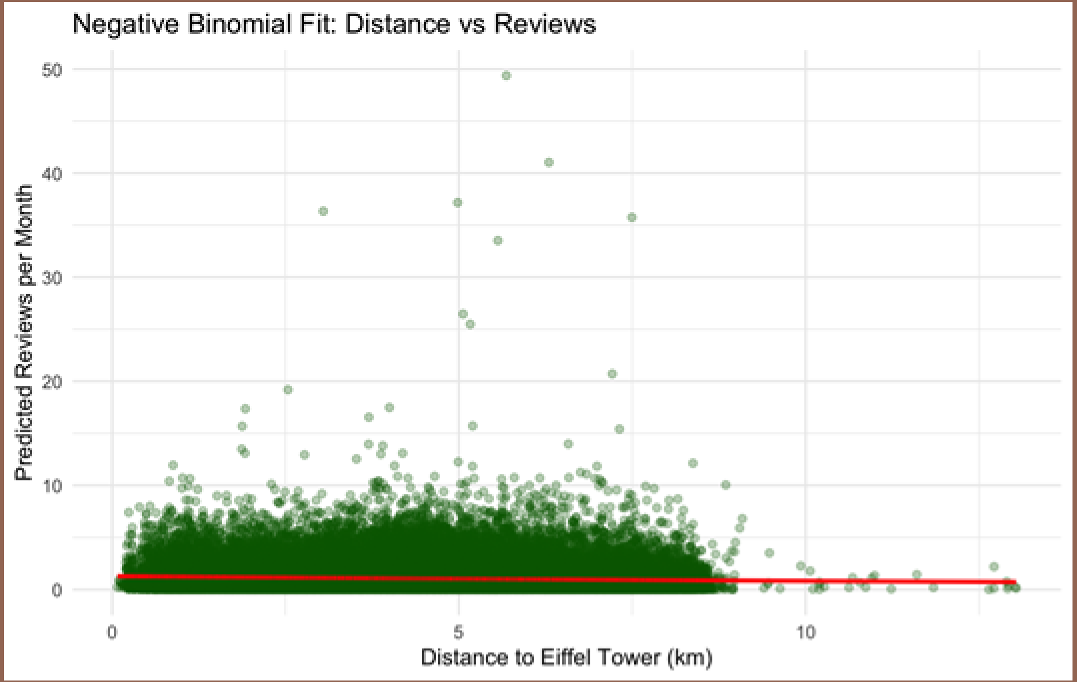
- The coefficient for log_price is -0.247, statistically significant ($p < 2e-16$).
- The boxplot shows lower occupancy rates in higher price quintiles, especially above $216.
- As prices rise, bookings drop — but the effect size varies across price levels.

Correlation & Regression Analysis

Does closer distance lead to more bookings?

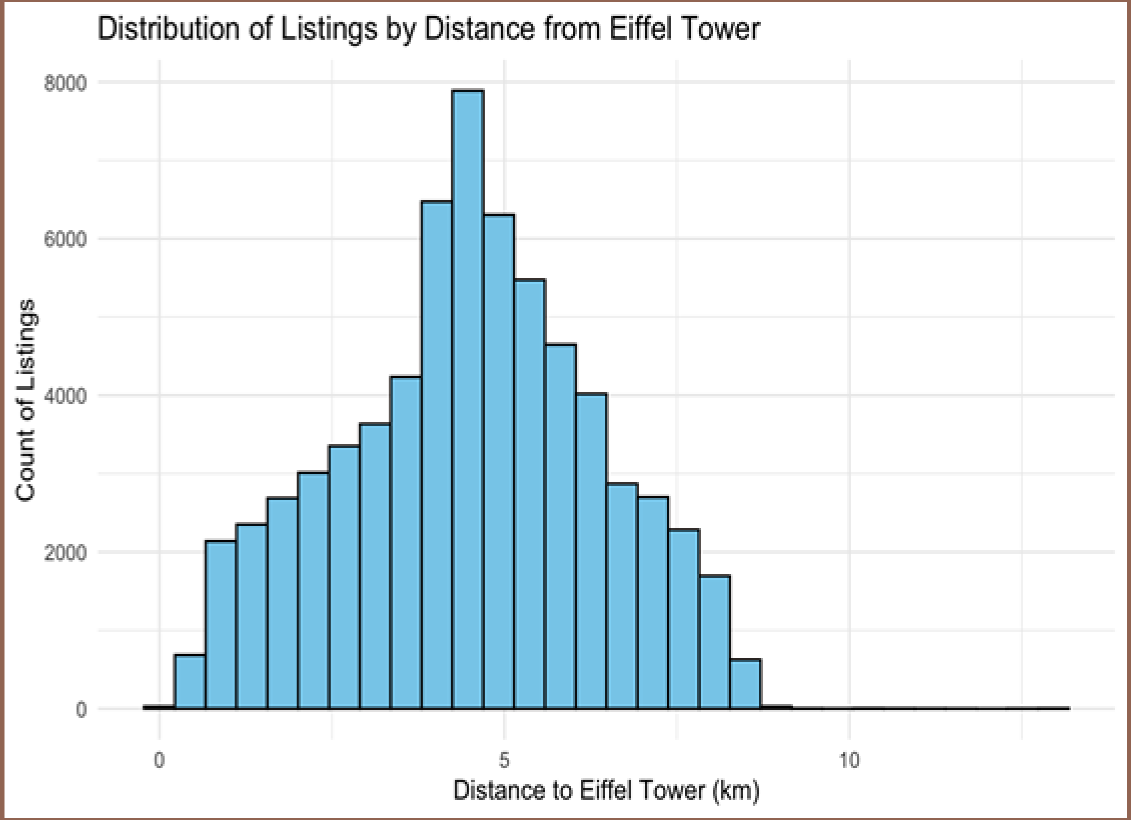
Yes — Listings closer to the Eiffel Tower receive more bookings.

- The negative binomial model shows a significant negative effect of distance (coef = -0.0439, p < 2e-16).
- The red line in the plot shows predicted bookings decrease as distance increases.
- Distance matters, but it explains only a small portion of variation in bookings (pseudo R² ≈ 0.005).



Distribution of Listings by Distance

- Most Airbnb listings are located between 3 to 6 km from the Eiffel Tower.
- This provides a reliable sample for evaluating how distance impacts bookings.
- Ensures that our model's negative relationship between distance and reviews per month is not skewed by sparse data at extreme distances.

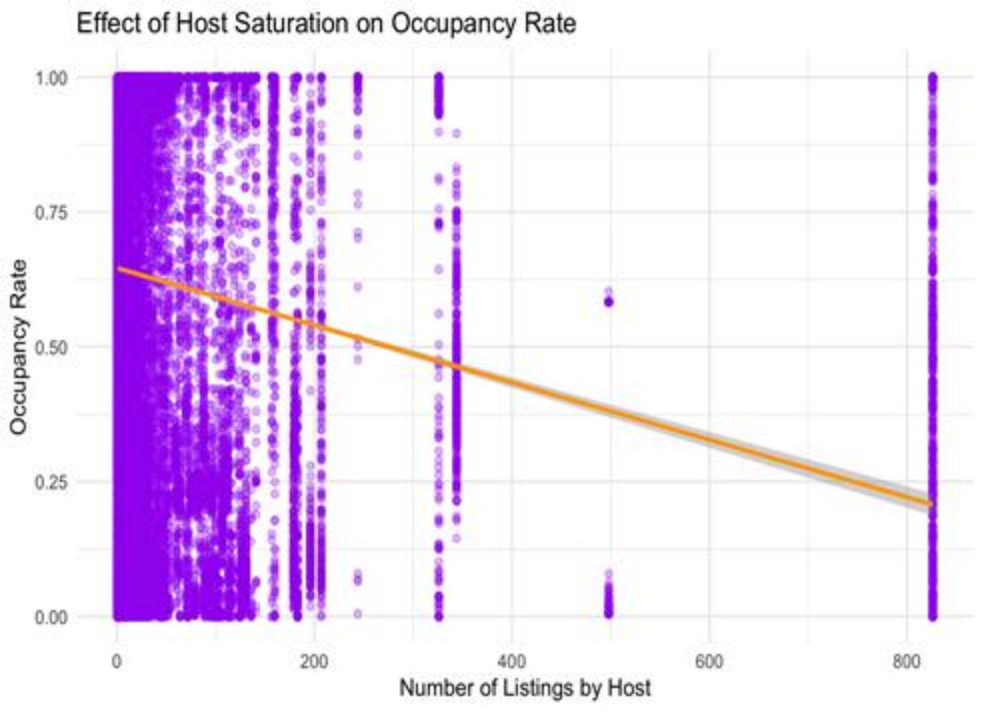


Correlation & Regression Analysis

Does too much competition reduce occupancy?

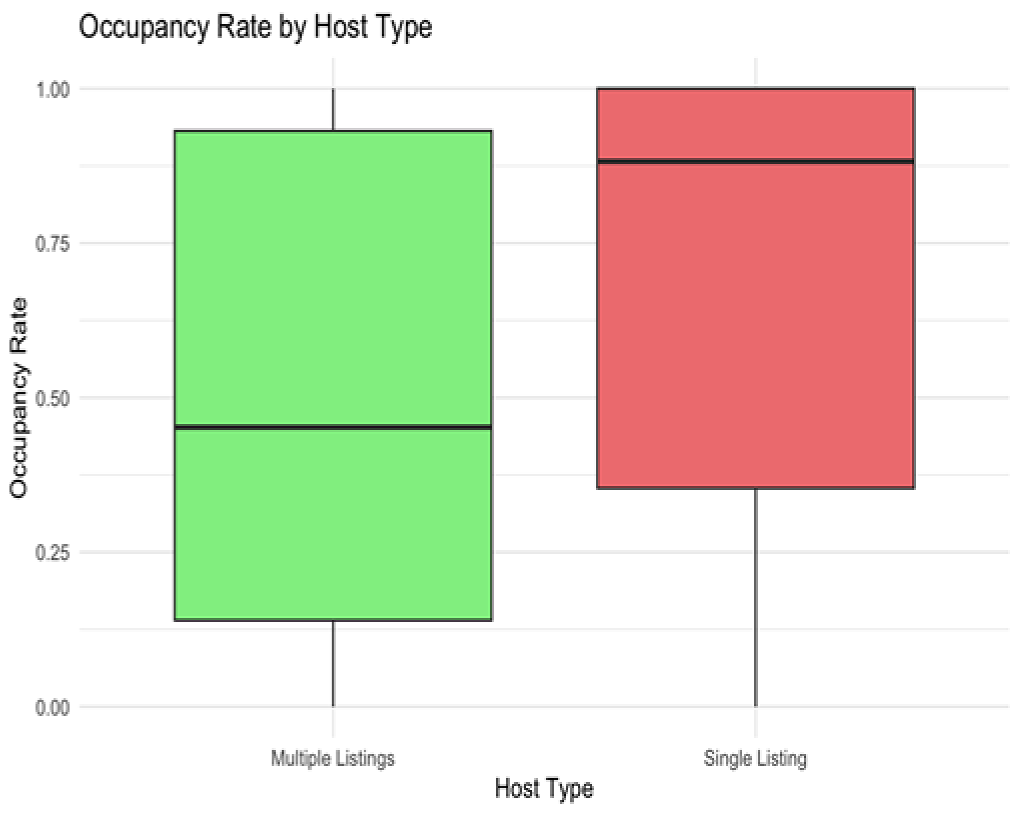
Yes — Hosts with more listings tend to have lower occupancy rates.

- Both linear and beta regression show a significant negative relationship between host listing count and occupancy.
  - Linear coef: -0.00053, Beta coef: -0.00109 (both p < 2e-16)
- The scatterplot (right) shows occupancy declines as host listing count increases.
- Suggests over-saturation or diluted attention may reduce booking performance.



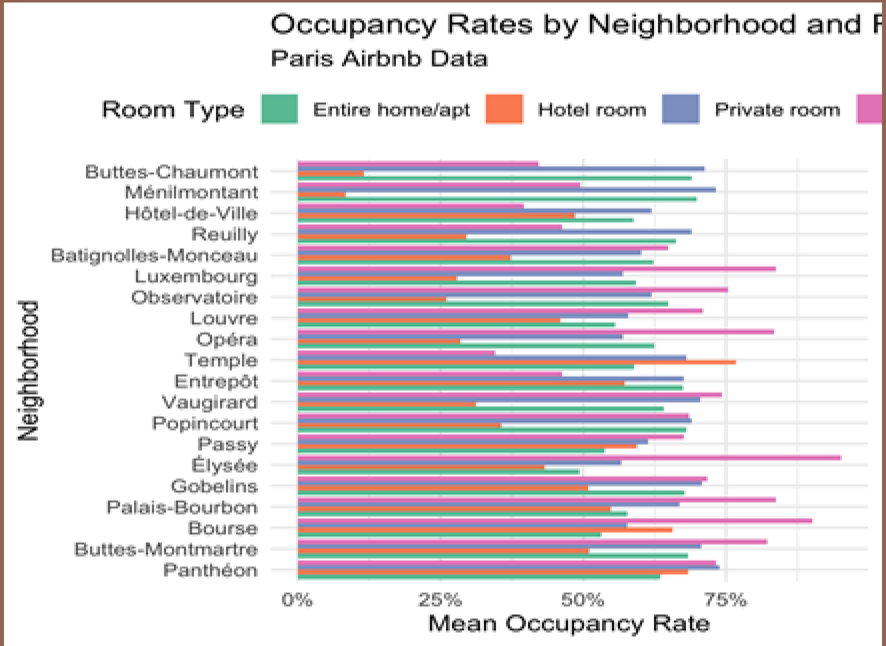
Insight:

- Single-listing hosts have a higher median occupancy rate than those with multiple listings.
- Multi-listing hosts show more variability and generally lower occupancy.
- This supports the idea that increased competition from the same host can reduce performance per listing, possibly due to less personalized attention or market dilution.



Mean Occupancy Rate

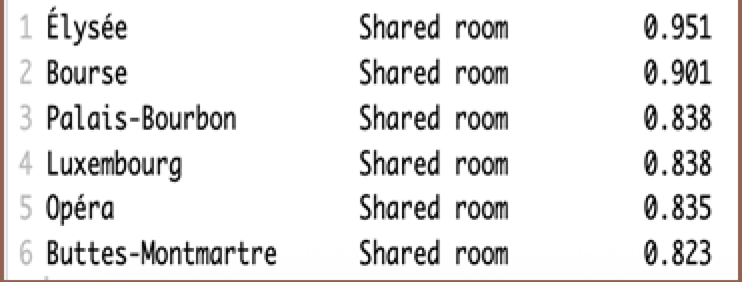
- Entire Home : Highest occupancy (~75%) in tourist zones like Louvre & Opéra which is ideal for premium pricing.
- Private Rooms : Steady demand (~50%) in residential areas (Ménilmontant) which could be optimized for long-term stays.
- Shared Rooms : Niche high-occupancy (~80% in Bourse). These usually target budget travelers and interns.
- Hotel rooms : Lowest occupancy (~25%). Tourists avoid in luxury areas (Élysée).



Mean Occupancy Rate

Top 5 investment opportunities for hosts

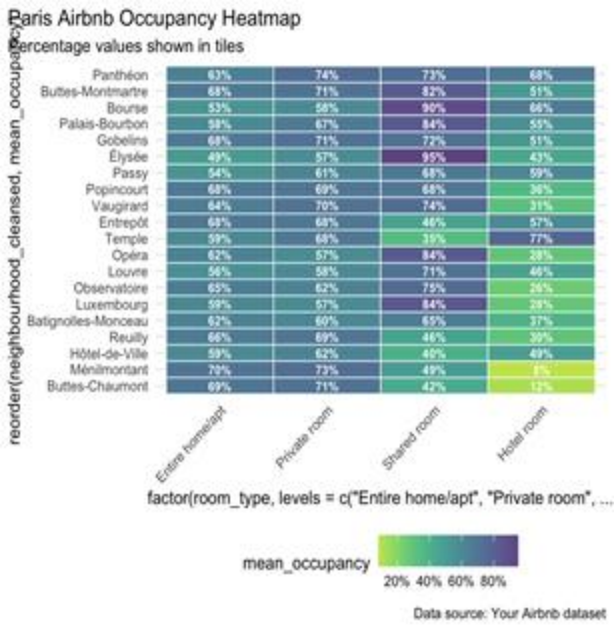
- Shared Rooms have had an advantage over the other types of rooms in these places.
- Elysee, Opera, Palais-Bourbon and luxembourg have been occupied the most because of its close proximity to tourist attractions.
- Bourse's high occupancy rate is likely due to the financial district attracting work/trips.
- Even though there is high demand in Elysee, there is low supply of rooms in Elysee.



1	Élysée	Shared room	0.951
2	Bourse	Shared room	0.901
3	Palais-Bourbon	Shared room	0.838
4	Luxembourg	Shared room	0.838
5	Opéra	Shared room	0.835
6	Buttes-Montmartre	Shared room	0.823

Heat Map of Occupation

- Elysee (95%), Bourse (90%), and Opera (84%) show unusually high shared-room demand—likely from business travelers and luxury seekers.
- Menilmontant (8%) and Buttes-Chaumont (12%) have disastrous hotel occupancy
- Buttes-Montmartre (84%) and Louvre (75%) perform best for entire homes, but shared rooms compete closely (82% vs. 84%).
- Bourse’s 90% occupancy suggests untapped demand for budget-friendly work stays (interns, contractors).



Occupancy Rate Ranking

Top Performers:

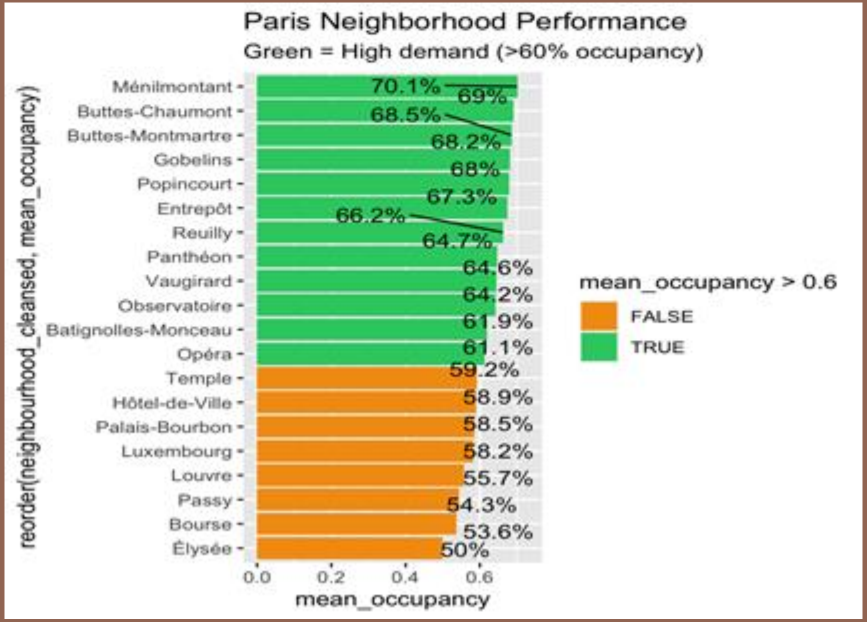
- The top 5 performers have charms like nightlife experience, tourist hotspots with high demand for entire homes and suburban charm that attracts long-term stays.
- Some of these are within University proximity which increases demand for private rooms.

Mid Tier Performers:

- Places like Louvre, Palais-bourbon charge more as they are in close proximity to tourist attractions.
- Passy, Bourse have people who are on business trips/work as it is close by financial district

Underperformer

- Elysee being the only underperformer here as it has prices which often leads to low occupancy.



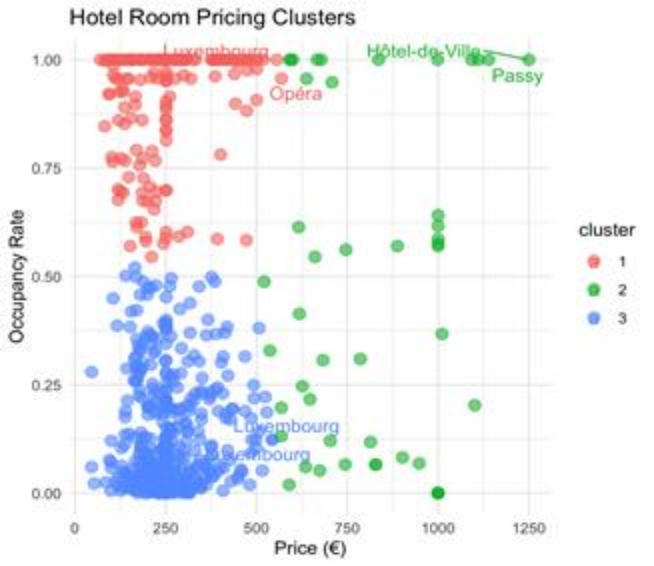
Hotel Room Pricing

Pricing

- Cluster 1(red): This cluster has a low price and high occupancy.
- Cluster 2(blue): This cluster has a moderate price and moderate level of occupancy.
- Cluster 3(green): This cluster has low occupancies and high price.

Strategies

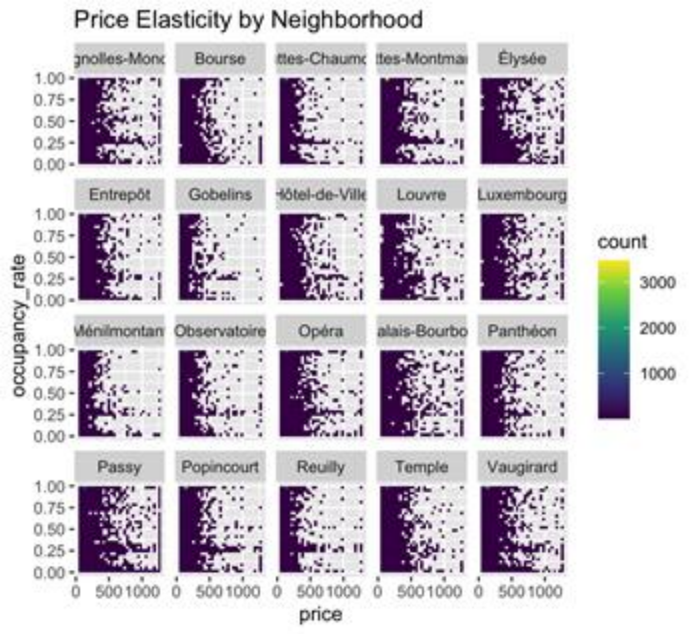
- For cluster 1, Increase prices 15-20% (high demand, underpriced).
- For cluster 2, they could optimize their pricing structure.
- For cluster 3, they could either decrease prices or convert to private rooms.



Price Elasticity

Price elasticity of demand measures how sensitive demand like occupancy rate is to price changes.

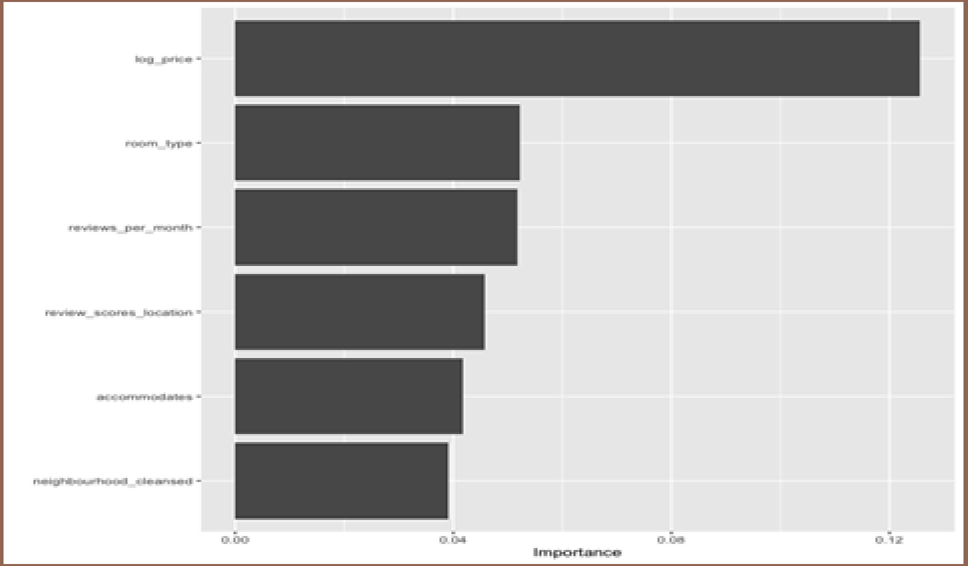
- Vertical patterns : Elysee show dark vertical streaks which indicates that occupancy stays high even as prices increase.
- Horizontal patterns : Passy shows dark horizontal streaks which tells us that occupancy drops sharply when prices exceed.
- Balanced patterns : Luxembourg has occupancy declining gradually with increase in pricing.



Variable Importance Plot

The relative influence of each variable on our model's predictions, Higher bars equals to greater impact.

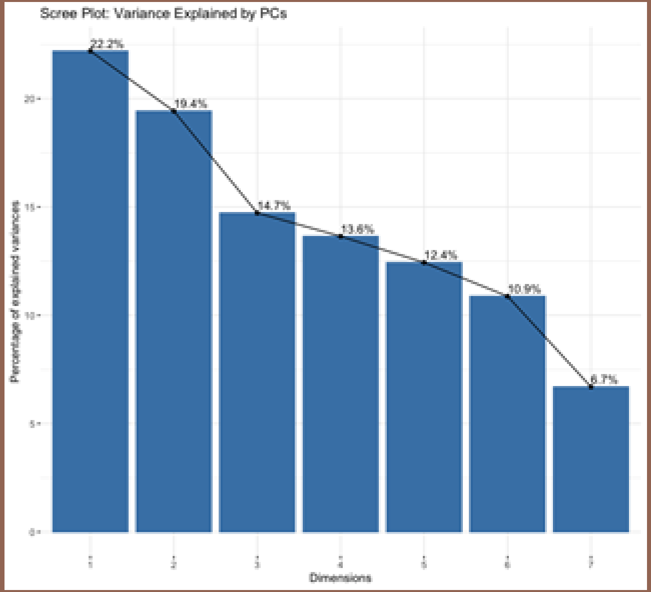
- Log_price has the highest importance as small prices changes disproportionately which affect bookings.
- Reviews_per_month creates an impact within consumers, if there are frequent reviews these boost visibility which help in bringing more eyes.
- Review_scores_location also has an impact as guests prioritize location quality over amenities.



Principal Component Analysis

Principal component analysis (PCA) is a linear dimensionality reduction technique with applications in exploratory data analysis, visualization and data preprocessing.

- PC1(22.2%) : Captures the high performing listings like high review, strong reviews, elevated number of reviews. These indicate which ones are frequently booked , well rated-properties etc.
- PC2(19.4%) : Captures affordable, always-available listing and also pricier and selectively available properties.
- PC3+ : These capture Niche patterns like less nights, room types.



PCA Biplot

Biplot combines variable vectors and listing positions to show how the variables correlate and how listing cluster.

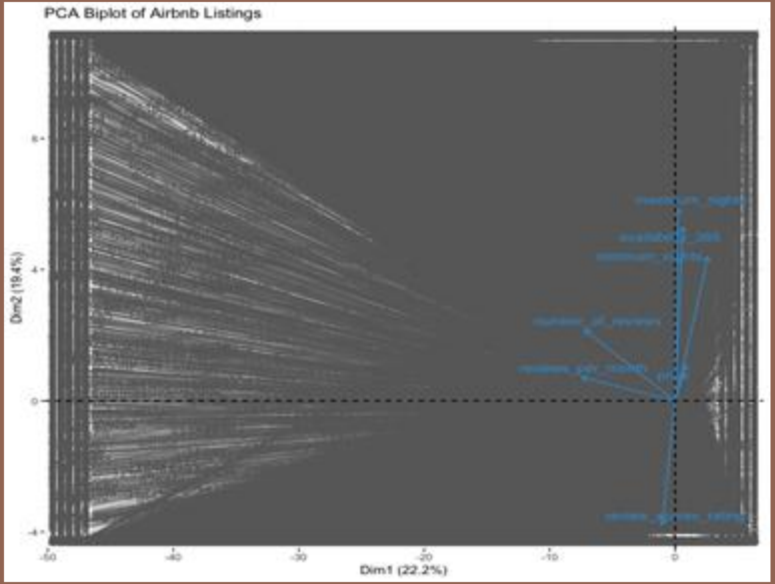
- reviews_per_month + number_of_reviews point strongly right → Drive PC1 (22.2%).
- availability_365 points downward → Year-round availability correlates with lower ratings
- price points left (negative PC1) → Higher prices reduce occupancy.
- Vertical spread → Influence PC2 (19.4%) more than PC1.

Clusters

- Top-Right: High-review, high-occupancy listings (align with review vectors).
- Bottom-Left: Pricey, always-available listings (often with lower ratings).
- Middle: Balanced properties (moderate prices, seasonal availability).

Strategies

- To boost popularity one would have to increase reviews and ratings to move towards the right side.
- To balance price and availability one would have to avoid the bottom left side.



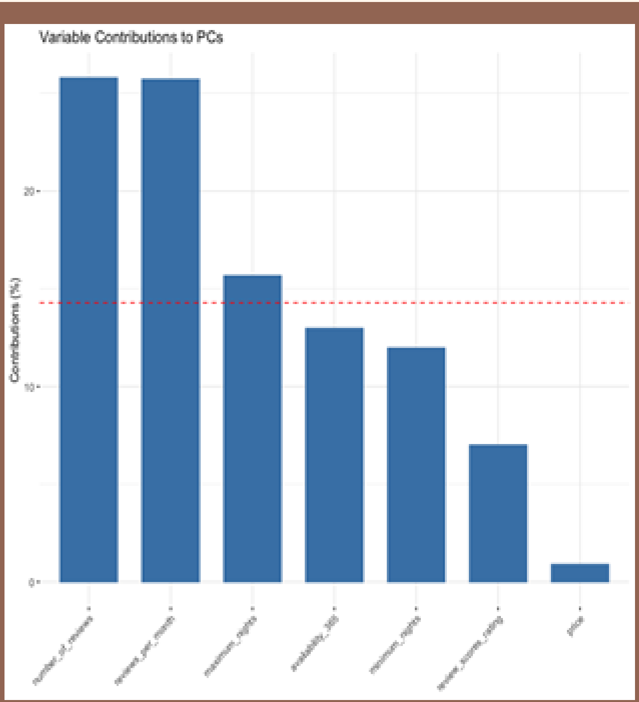
Variable Contributions to PC

From these graphs, we can see that :

- Review_per_month are the number 1 driver of listing popularity.
- Number_of_review means more trust for the customers.
- Year- round availability might hurt ratings
- Higher prices reduces the occupancy of customers.
- Review_scores_ratings contribute less than expected as guests prioritize quantity of reviews over average rating.

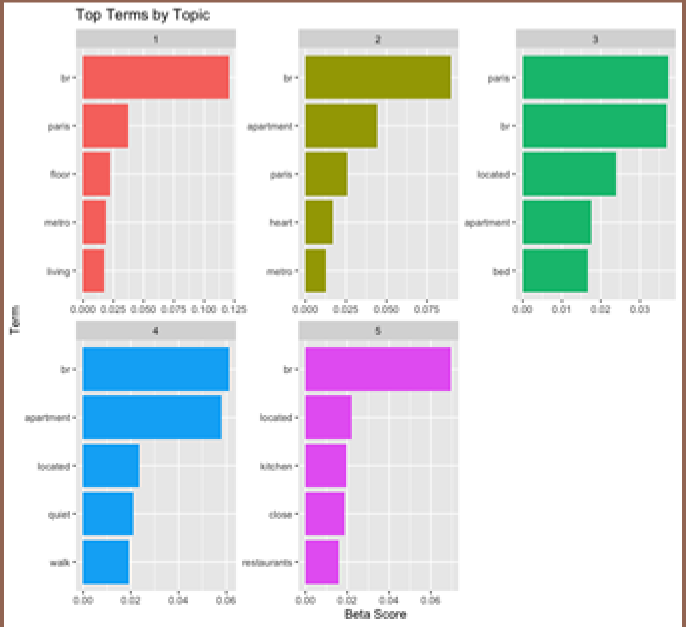
Strategies

- Focus on improving review_per_month and also the pricing structure.
- “Always available” listings often underperform.

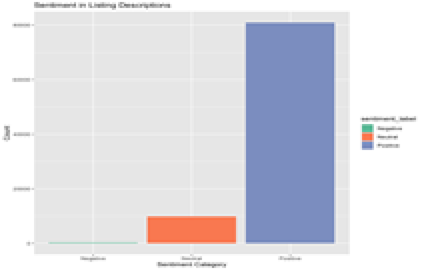


Terms by Topic

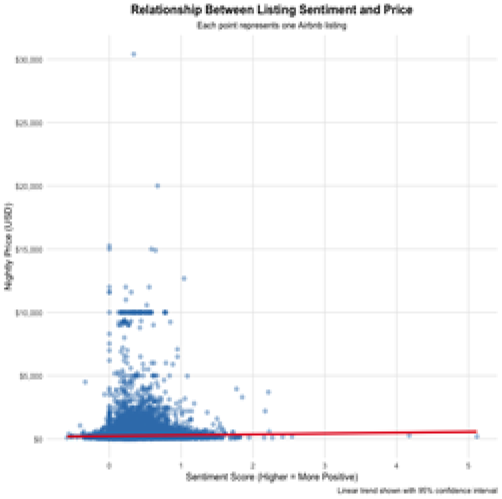
- **"br" (bedroom)** is the most frequently mentioned term across all topics, indicating its high importance in guest reviews.
- Terms like **“paris”, “metro”, and “located”** highlight the strong focus on **location and accessibility**.
- Guests often mention **amenities** such as **“kitchen”, “closet”, and “restaurants”**, suggesting the value placed on convenience.
- Words like **“quiet”** and **“walk”** reflect guest preferences for peaceful surroundings and walkable neighborhoods.



Sentimental Analysis



- Positive Correlation: The graph shows a positive relationship between the sentiment score of listings and their prices. As sentiment increases, prices tend to rise.
- Confidence Interval: The overall trend is upward, there is some variability in the data. The wider the interval, the less certain the exact relationship may be in certain regions.
- Data Spread: The trend line, showing that while sentiment generally correlates with price, other factors likely influence pricing as well.



Key Takeaways from the Project ★★★★★

This project demonstrates full application of BIA 672 concepts through a data-driven analysis of Airbnb listings in Paris. We applied **cross-sectional analysis** to explore occupancy patterns, used **regression models** to predict booking behavior, and implemented **PCA and clustering** for market segmentation. We also performed **text mining, topic modeling, and sentiment analysis** to extract insights from listing descriptions. Collectively, these techniques enabled us to assess pricing elasticity, identify neighborhood-level saturation, and understand guest sentiment covering all major topics from **basic data analysis to advanced predictive modeling and big data decision tools**.

Part - III

Recommended Decisions

★★★★★

- **Optimize Room Type Strategy:**
  Focus on promoting *private and shared rooms*, which have shown the highest occupancy rates (~80% in Bourse and ~50% in residential areas). Convert underperforming *hotel rooms* (25% occupancy) into private/shared spaces to maximize returns.

- **Target High-Demand Areas with Low Supply:**
  Neighborhoods like *Élysée* and *Opera* have high demand but limited listings. Encourage new hosts or expansion in these zones to capture unmet demand and drive revenue growth.

- **Capitalize on Proximity to Attractions:**
  Promote listings within 3 km of major attractions like the Eiffel Tower, as they consistently receive more bookings and offer higher earning potential.

THANK YOU

Group 3