Airbnb

Group 8

Emely Callejas, Ashley Cortez,
Rithvik V Sourab, Angelica Verduzco

Case Background and Problem

Airbnb, launched in 2008 in San Francisco, and it has reshaped the hospitality industry. Homeowners can now lease their spaces to guests. This global platform acknowledges that renters across different regions have diverse preferences for property types and features.

In this case study, we focus on analyzing Airbnb's data from Miami, FL, and Paris, France.

Linear Regression Analysis

- Identify key factors influencing property occupancy rates in FL and France. Understand what key drivers help optimize the listings to improve occupancy and revenue

Topic Modeling of Review Texts

- Review texts from properties in Miami and Paris, analysis is aimed to uncover trends of high versus low property ratings. This will help guide potential improvements in service quality and customers satisfaction

Linear Regression Models

Airbnb Dataset Variables

Dependent Variable: Occupancy

Independent Variables:

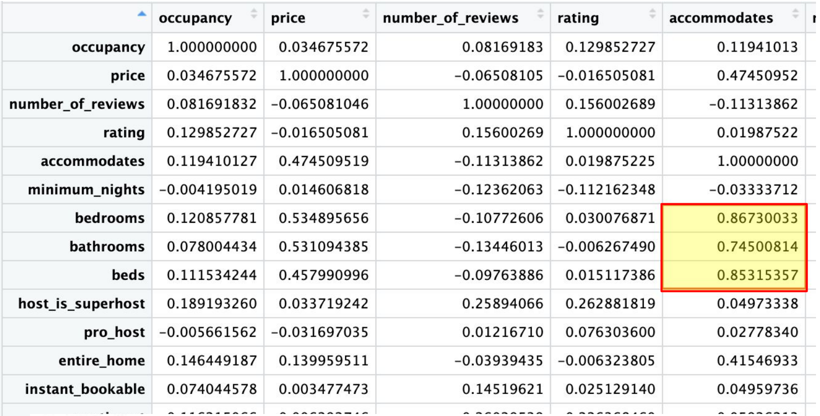
Miami:

log(price)
log(number of reviews +1)
Rating
log(accommodates)
Beds
Bedrooms
Bathrooms
log(minimum nights +1)
Host is superhost
Pro host
Entire home
Instant bookable
sentiment

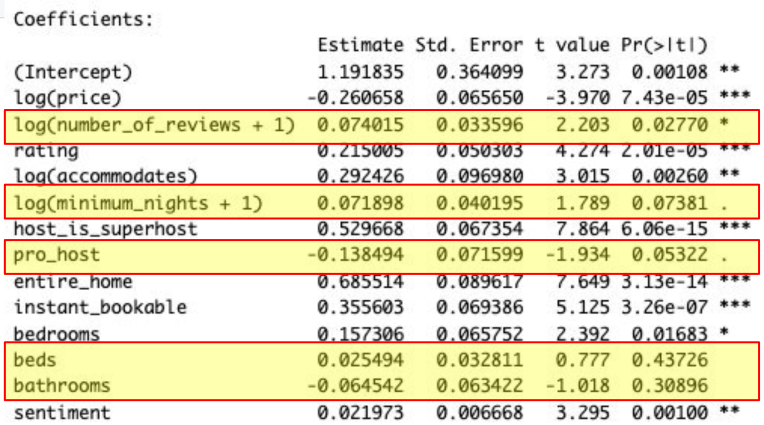
Paris:

log(price)
log(number of reviews +1)
Rating
log(accommodates)
log(minimum nights +1)
Host is superhost
Pro host
Entire home
Instant bookable
Beds
Bedrooms
Bathrooms
sentiment

Variables Omitted from Linear Regression - Miami



	occupancy	price	number_of_reviews	rating	accommodates
occupancy	1.000000000	0.034675572	0.08169183	0.129852727	0.11941013
price	0.034675572	1.000000000	-0.06508105	-0.016505081	0.47450952
number_of_reviews	0.081691832	-0.065081046	1.00000000	0.156002689	-0.11313862
rating	0.129852727	-0.016505081	0.15600269	1.000000000	0.01987522
accommodates	0.119410127	0.474509519	-0.11313862	0.019875225	1.00000000
minimum_nights	-0.004195019	0.014606818	-0.12362063	-0.112162348	-0.03333712
bedrooms	0.120857781	0.534895656	-0.10772606	0.030076871	0.86730033
bathrooms	0.078004434	0.531094385	-0.13446013	-0.006267490	0.74500814
beds	0.111534244	0.457990996	-0.09763886	0.015117386	0.85315357
host_is_superhost	0.189193260	0.033719242	0.25894066	0.262881819	0.04973338
pro_host	-0.005661562	-0.031697035	0.01216710	0.076303600	0.02778340
entire_home	0.146449187	0.139959511	-0.03939435	-0.006323805	0.41546933
instant_bookable	0.074044578	0.003477473	0.14519621	0.025129140	0.04959736

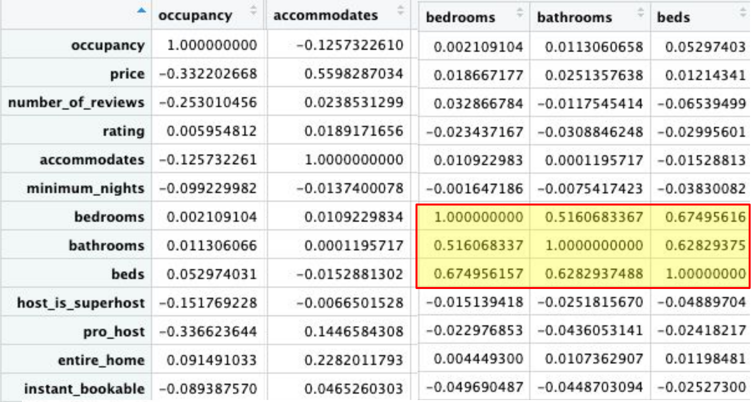


```
Coefficients:
                          Estimate Std. Error t value Pr(>|t|)
(Intercept)               1.191835   0.364099   3.273  0.00108 **
log(price)               -0.260658   0.065650  -3.970 7.43e-05 ***
log(number_of_reviews + 1) 0.074015  0.033596   2.203  0.02770 *
rating                    0.215005   0.050303   4.274 2.01e-05 ***
log(accommodates)         0.292426   0.096980   3.015  0.00260 **
log(minimum_nights + 1)   0.071898   0.040195   1.789  0.07381 .
host_is_superhost         0.529668   0.067354   7.864 6.06e-15 ***
pro_host                 -0.138494   0.071599  -1.934  0.05322 .
entire_home               0.685514   0.089617   7.649 3.13e-14 ***
instant_bookable          0.355603   0.069386   5.125 3.26e-07 ***
bedrooms                  0.157306   0.065752   2.392  0.01683 *
beds                      0.025494   0.032811   0.777  0.43726
bathrooms                -0.064542   0.063422  -1.018  0.30896
sentiment                 0.021973   0.006668   3.295  0.00100 **
```

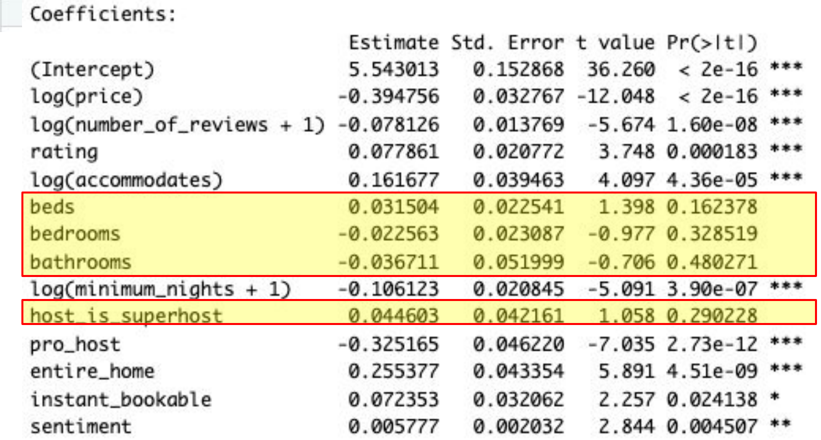
Steps Taken to Decide which Variables to Omit

1. **Correlation Matrix:**
   a. Multicollinearity between “accommodates” and “beds”, ”bathrooms”, and ”bedrooms”
2. **Checking Coefficients:**
   a. “beds” and “bathrooms” have no significance
   b. “minimum_nights” and “pro_host” had the second lowest significance among the variables initially included in the model
   c. After removing minimum nights and pro host, the significance of “number_of_reviews” decreased
3. **Variables Ultimately Removed:**
   a. Number_of_reviews
   b. Minimum_nights
   c. Pro_host
   d. Beds
   e. Bathrooms

Variables Omitted from Linear Regression - Paris



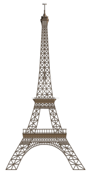
	occupancy	accommodates	bedrooms	bathrooms	beds
occupancy	1.000000000	-0.1257322610	0.002109104	0.0113060658	0.05297403
price	-0.332202668	0.5598287034	0.018667177	0.0251357638	0.01214341
number_of_reviews	-0.253010456	0.0238531299	0.032866784	-0.0117545414	-0.06539499
rating	0.005954812	0.0189171656	-0.023437167	-0.0308846248	-0.02995601
accommodates	-0.125732261	1.0000000000	0.010922983	0.0001195717	-0.01528813
minimum_nights	-0.099229982	-0.0137400078	-0.001647186	-0.0075417423	-0.03830082
bedrooms	0.002109104	0.0109229834	1.000000000	0.5160683367	0.67495616
bathrooms	0.011306066	0.0001195717	0.516068337	1.0000000000	0.62829375
beds	0.052974031	-0.0152881302	0.674956157	0.6282937488	1.00000000
host_is_superhost	-0.151769228	-0.0066501528	-0.015139418	-0.0251815670	-0.04889704
pro_host	-0.336623644	0.1446584308	-0.022976853	-0.0436053141	-0.02418217
entire_home	0.091491033	0.2282011793	0.004449300	0.0107362907	0.01198481
instant_bookable	-0.089387570	0.0465260303	-0.049690487	-0.0448703094	-0.02527300



```
Coefficients:
                          Estimate Std. Error t value Pr(>|t|)    
(Intercept)               5.543013   0.152868  36.260  < 2e-16 ***
log(price)               -0.394756   0.032767 -12.048  < 2e-16 ***
log(number_of_reviews + 1) -0.078126   0.013769  -5.674 1.60e-08 ***
rating                    0.077861   0.020772   3.748 0.000183 ***
log(accommodates)         0.161677   0.039463   4.097 4.36e-05 ***
beds                      0.031504   0.022541   1.398 0.162378    
bedrooms                 -0.022563   0.023087  -0.977 0.328519    
bathrooms                -0.036711   0.051999  -0.706 0.480271    
log(minimum_nights + 1)  -0.106123   0.020845  -5.091 3.90e-07 ***
host_is_superhost         0.044603   0.042161   1.058 0.290228    
pro_host                 -0.325165   0.046220  -7.035 2.73e-12 ***
entire_home               0.255377   0.043354   5.891 4.51e-09 ***
instant_bookable          0.072353   0.032062   2.257 0.024138 *  
sentiment                 0.005777   0.002032   2.844 0.004507 ** 
```

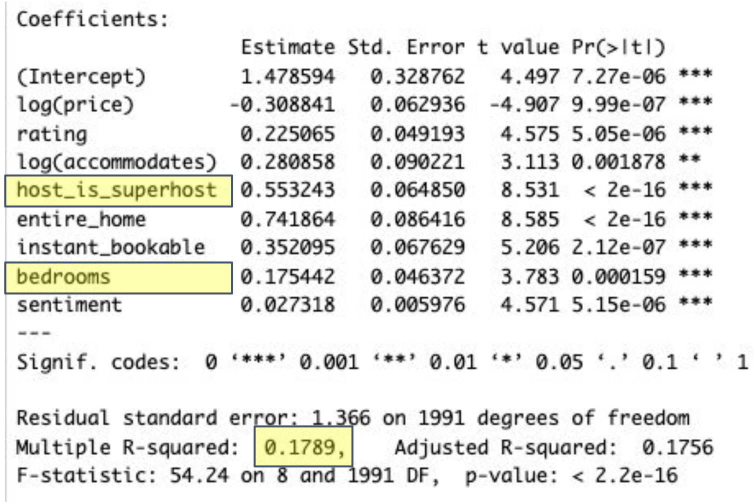
Steps Taken to Decide which Variables to Omit

1. **Correlation Matrix:**
   a. Multicollinearity between “beds”, ”bathrooms”, and ”bedrooms”
2. **Checking Coefficients:**
   a. “bedrooms”, “beds” and “bathrooms” have no significance
   b. “host_is_superhost” also had no significance
   c. After removing minimum nights and pro host, the significance of “number_of_reviews” decreased
3. **Variables Ultimately Removed:**
   a. Host_is_superhost
   b. Bedrooms
   c. Beds
   d. Bathrooms



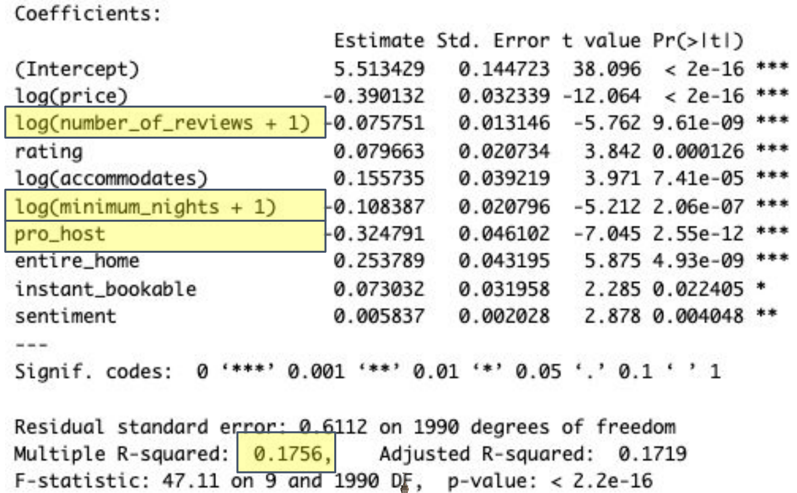
Variables Selected for Linear Regressions

Miami Regression

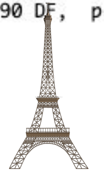


```
Coefficients:
                   Estimate Std. Error t value Pr(>|t|)    
(Intercept)        1.478594   0.328762   4.497 7.27e-06 ***
log(price)        -0.308841   0.062936  -4.907 9.99e-07 ***
rating             0.225065   0.049193   4.575 5.05e-06 ***
log(accommodates)  0.280858   0.090221   3.113 0.001878 ** 
host_is_superhost  0.553243   0.064850   8.531  < 2e-16 ***
entire_home        0.741864   0.086416   8.585  < 2e-16 ***
instant_bookable   0.352095   0.067629   5.206 2.12e-07 ***
bedrooms           0.175442   0.046372   3.783 0.000159 ***
sentiment          0.027318   0.005976   4.571 5.15e-06 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 1.366 on 1991 degrees of freedom
Multiple R-squared:  0.1789,	Adjusted R-squared:  0.1756 
F-statistic: 54.24 on 8 and 1991 DF,  p-value: < 2.2e-16
```

Paris Regression

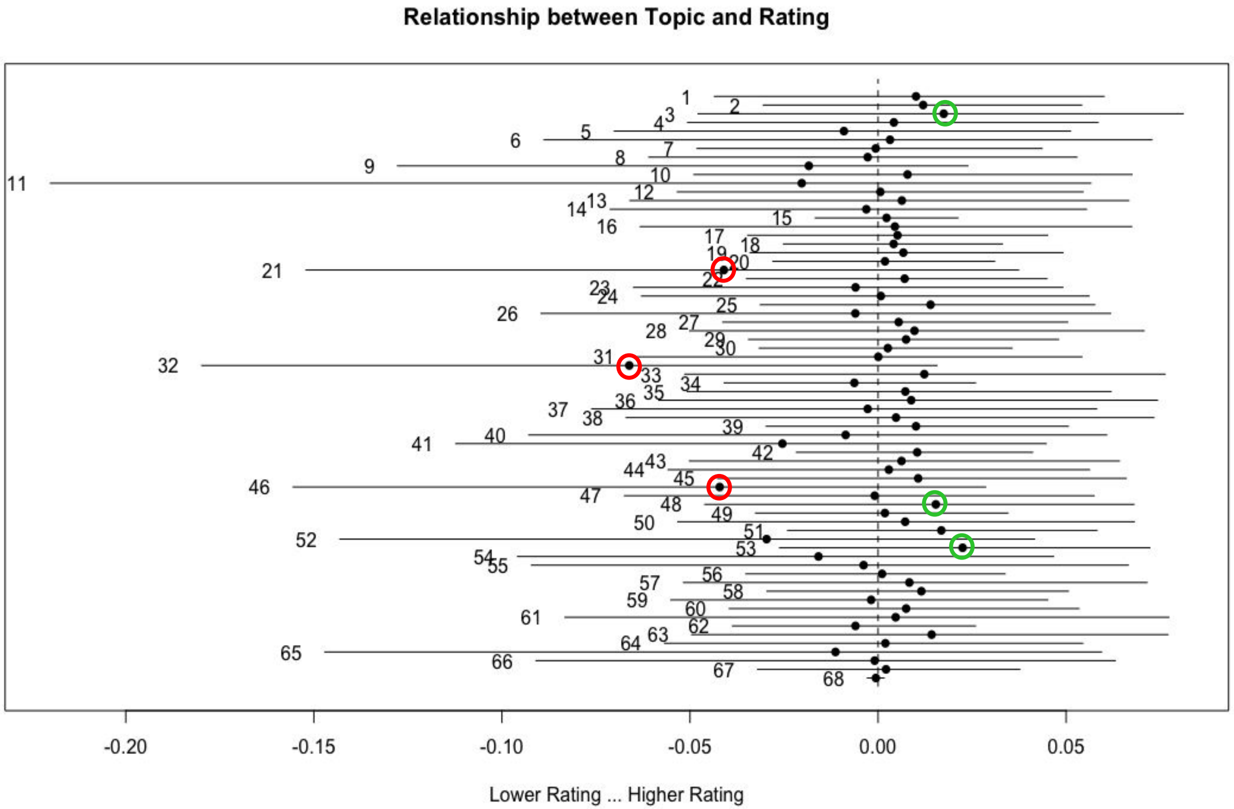


```
Coefficients:
                        Estimate Std. Error t value Pr(>|t|)    
(Intercept)             5.513429   0.144723  38.096  < 2e-16 ***
log(price)             -0.390132   0.032339 -12.064  < 2e-16 ***
log(number_of_reviews + 1) -0.075751   0.013146  -5.762 9.61e-09 ***
rating                  0.079663   0.020734   3.842 0.000126 ***
log(accommodates)       0.155735   0.039219   3.971 7.41e-05 ***
log(minimum_nights + 1) -0.108387   0.020796  -5.212 2.06e-07 ***
pro_host               -0.324791   0.046102  -7.045 2.55e-12 ***
entire_home             0.253789   0.043195   5.875 4.93e-09 ***
instant_bookable        0.073032   0.031958   2.285 0.022405 *  
sentiment               0.005837   0.002028   2.878 0.004048 ** 
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.6112 on 1990 degrees of freedom
Multiple R-squared:  0.1756,	Adjusted R-squared:  0.1719 
F-statistic: 47.11 on 9 and 1990 DF,  p-value: < 2.2e-16
```



Topic Models & Word Clouds

Topic Modeling - Miami



Positive Rating Topics

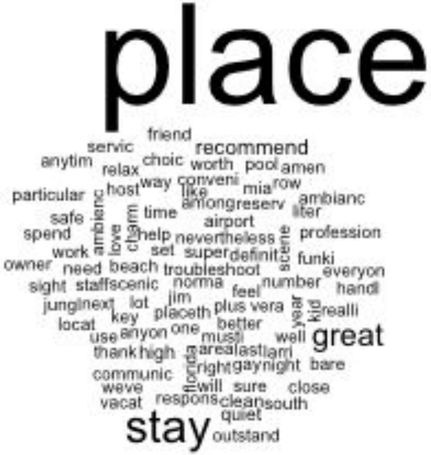
- Topic = 53
- Topic = 3
- Topic = 51

Negative Rating Topics

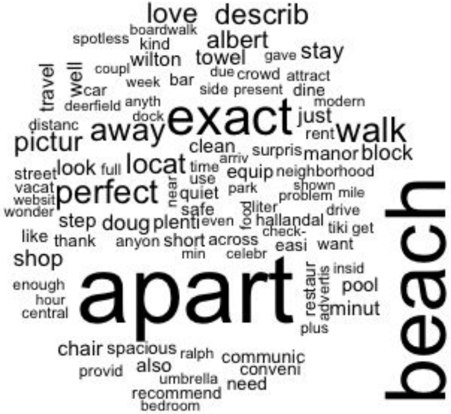
- Topic = 32
- Topic = 46
- Topic = 21

Positive Rating Word Cloud - Miami

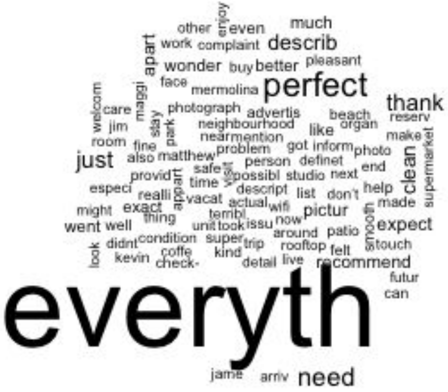
Topic 53



Topic 3

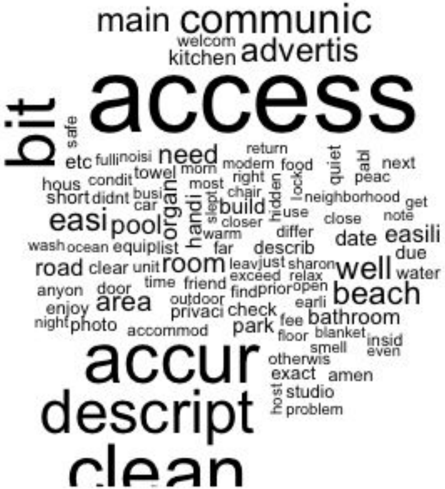


Topic 51

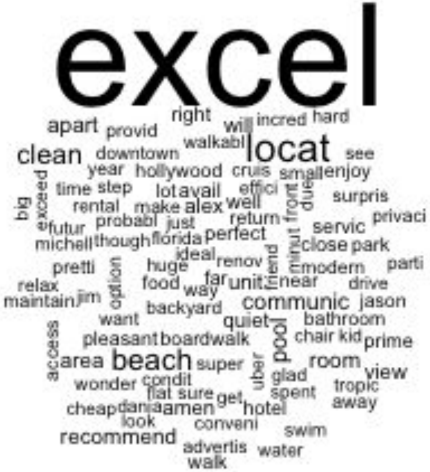


Negative Rating Word Cloud - Miami

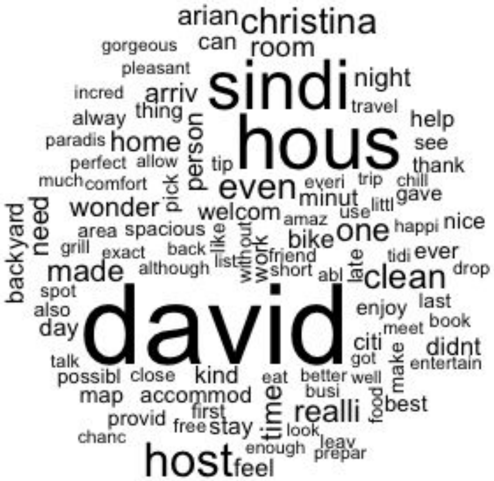
Topic 32



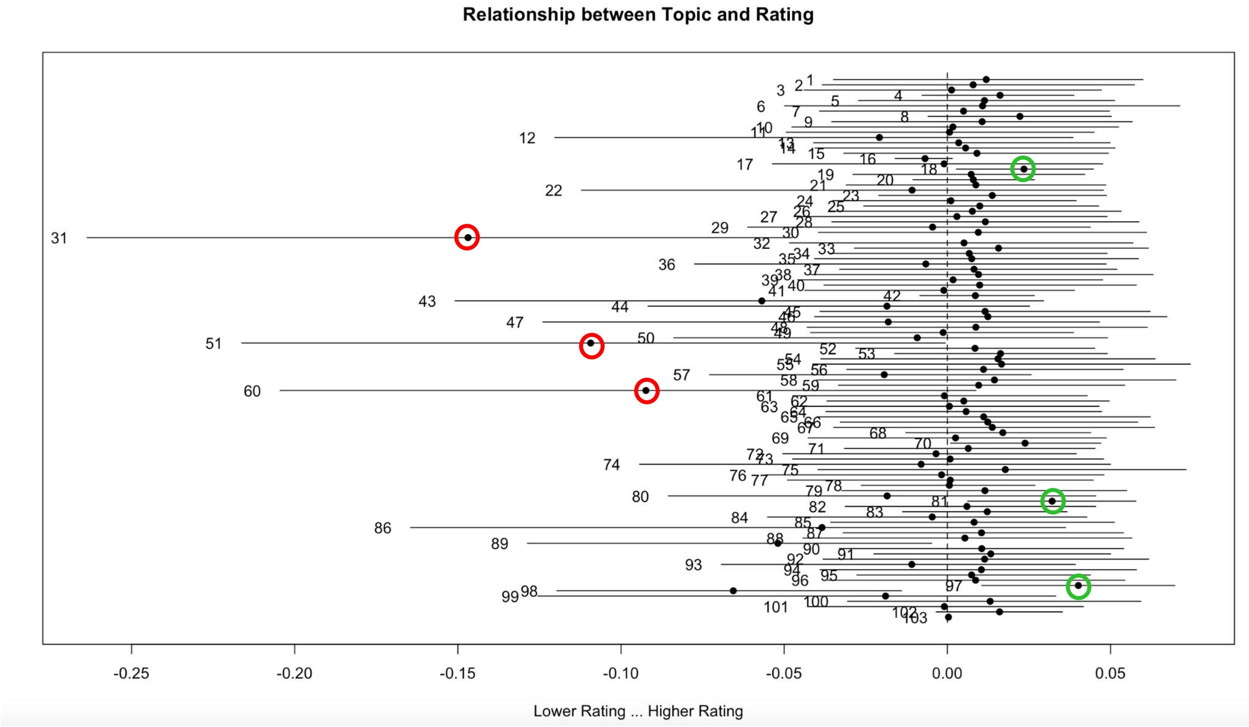
Topic 46



Topic 21



Topic Modeling - Paris

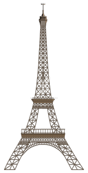


Positive Rating Topics

- Topic = 97
- Topic = 81
- Topic = 18

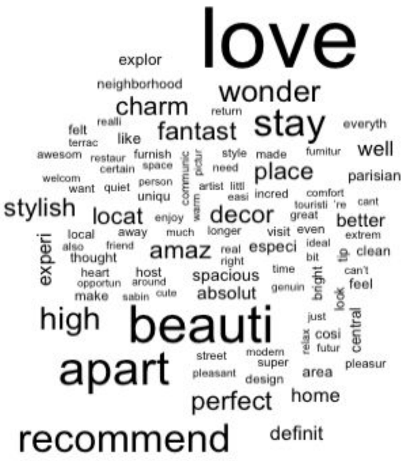
Negative Rating Topics

- Topic = 31
- Topic = 51
- Topic = 60

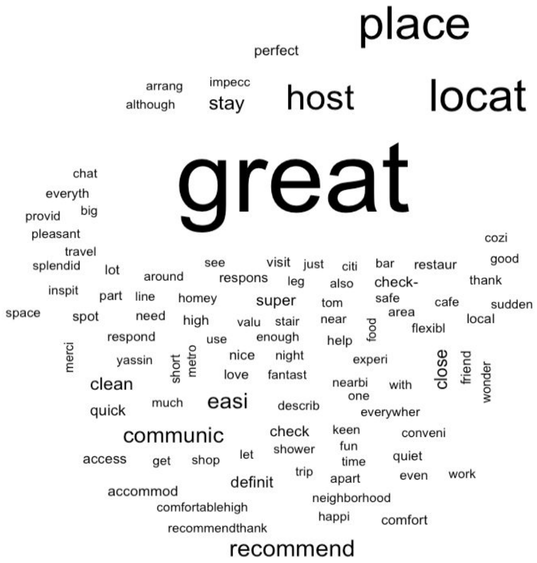


Positive Rating Word Cloud - Paris

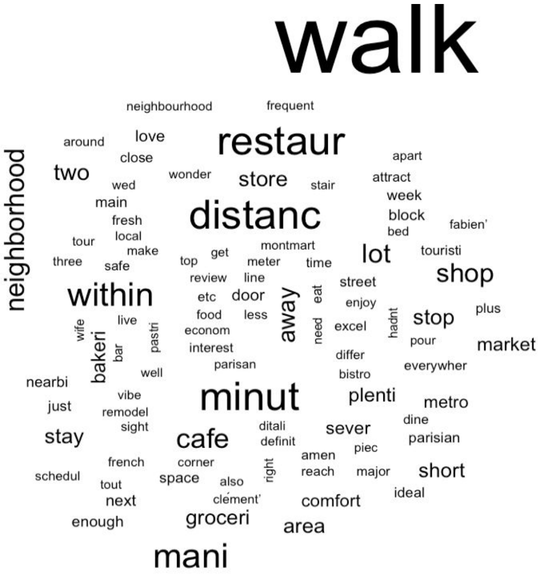
Topic 97



Topic 81

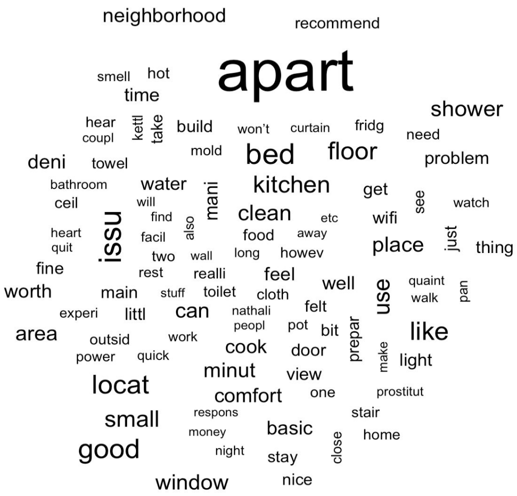


Topic 18

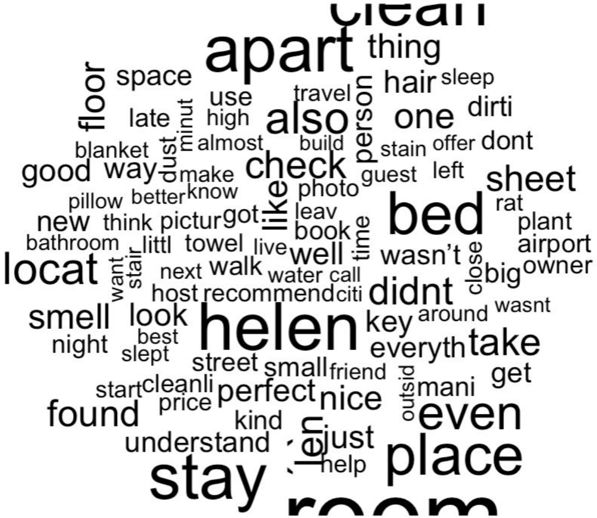


Negative Rating Word Cloud - Paris

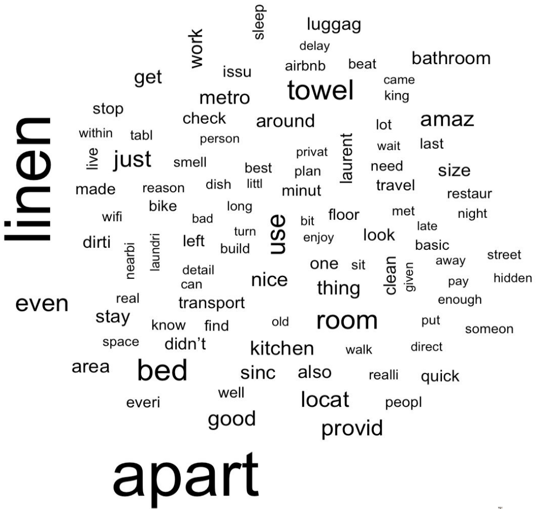
Topic 31

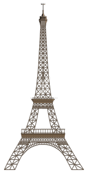


Topic 51



Topic 60





Conclusions

Business Recommendations

Miami Market

<u>Linear Regression:</u>

- Strongest Coefficient: Entire Home
- Differences: Superhost Status and amount of Bedrooms

<u>Topic Modeling:</u>

- Positive Ratings for listings close to the beach and other attractions. Also valued cleanliness.

With the linear regression and topic modeling, we can recommend the following:

- Provide incentives to hosts for receiving higher reviews, so that there are more Super Hosts in the area.
- Focus advertising efforts on listings for entire homes that are closer to the beach and tourist attractions
- Partner with local businesses to offer discounts to nearby Airbnb guests
- Offer discounted rates to guests for booking a longer stay at an entire home listing

Paris Market

<u>Linear Regression:</u>

- Strongest Coefficient: Price
- Differences: Pro Host Status, Number of Reviews, Minimum Nights

<u>Topic Modeling:</u>

- Positive Ratings were for the stylish and beauty of the stay for the people

With the linear regression and topic modeling, we can recommend the following:

- Focus on enhancing the guest experience to capitalize on positive ratings related to the stylishness and beauty of stays.
- Given that price is the strongest coefficient in linear regression analysis, consider optimizing pricing strategies to maximize occupancy and profitability. This could involve adjusting prices based on demand fluctuations, competitor analysis, and seasonal trends.
- Additionally, consider offering discounts or promotions to incentivize bookings during off-peak periods

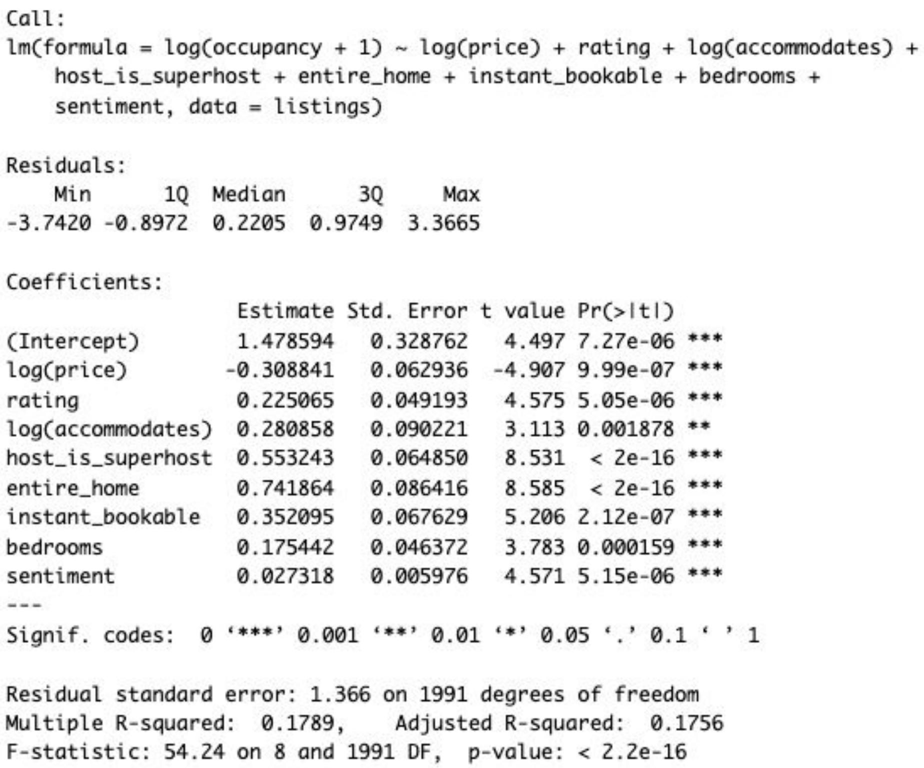
Research Project Recommendations

- Extended Demographic Analysis:
  Study different age groups, travel reasons, and economic backgrounds in Miami and Paris to customize Airbnb listings.
- Comparative Analysis of Host Statuses:
  Look into how Superhost and Pro Host statuses affect bookings and guest satisfaction in various locations.
- Price Elasticity of Demand Study:
  Further explore how changes in price impact bookings in Paris to find the best pricing strategies for higher occupancy and revenue.
- Customer Journey Mapping:
  Track the guest experience from searching for a rental to after their stay, highlighting areas for improvement.

airbnb

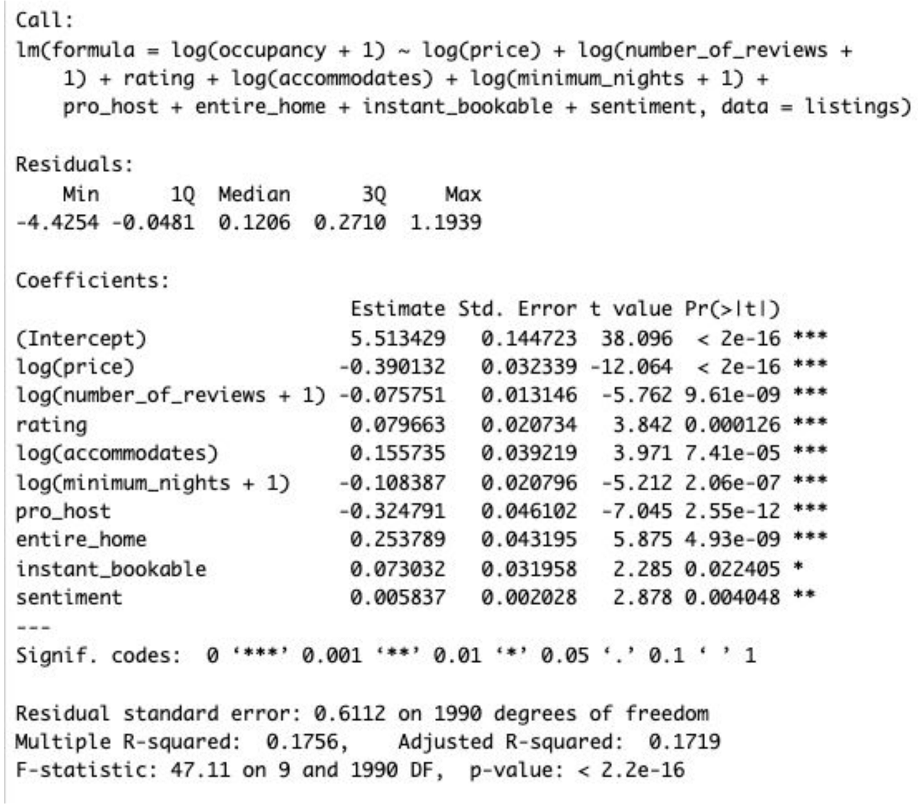
Appendix

Linear Regression Model - Miami



```
Call:
lm(formula = log(occupancy + 1) ~ log(price) + rating + log(accommodates) +
    host_is_superhost + entire_home + instant_bookable + bedrooms +
    sentiment, data = listings)

Residuals:
    Min      1Q  Median      3Q     Max
-3.7420 -0.8972  0.2205  0.9749  3.3665

Coefficients:
                   Estimate Std. Error t value Pr(>|t|)
(Intercept)        1.478594   0.328762   4.497 7.27e-06 ***
log(price)        -0.308841   0.062936  -4.907 9.99e-07 ***
rating             0.225065   0.049193   4.575 5.05e-06 ***
log(accommodates)  0.280858   0.090221   3.113 0.001878 **
host_is_superhost  0.553243   0.064850   8.531  < 2e-16 ***
entire_home        0.741864   0.086416   8.585  < 2e-16 ***
instant_bookable   0.352095   0.067629   5.206 2.12e-07 ***
bedrooms           0.175442   0.046372   3.783 0.000159 ***
sentiment          0.027318   0.005976   4.571 5.15e-06 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 1.366 on 1991 degrees of freedom
Multiple R-squared:  0.1789,    Adjusted R-squared:  0.1756
F-statistic: 54.24 on 8 and 1991 DF,  p-value: < 2.2e-16
```

Linear Regression Model - Paris



```
Call:
lm(formula = log(occupancy + 1) ~ log(price) + log(number_of_reviews +
    1) + rating + log(accommodates) + log(minimum_nights + 1) +
    pro_host + entire_home + instant_bookable + sentiment, data = listings)

Residuals:
    Min      1Q  Median      3Q     Max
-4.4254 -0.0481  0.1206  0.2710  1.1939

Coefficients:
                         Estimate Std. Error t value Pr(>|t|)
(Intercept)              5.513429   0.144723  38.096  < 2e-16 ***
log(price)              -0.390132   0.032339 -12.064  < 2e-16 ***
log(number_of_reviews + 1) -0.075751   0.013146  -5.762 9.61e-09 ***
rating                   0.079663   0.020734   3.842 0.000126 ***
log(accommodates)        0.155735   0.039219   3.971 7.41e-05 ***
log(minimum_nights + 1) -0.108387   0.020796  -5.212 2.06e-07 ***
pro_host                -0.324791   0.046102  -7.045 2.55e-12 ***
entire_home              0.253789   0.043195   5.875 4.93e-09 ***
instant_bookable         0.073032   0.031958   2.285 0.022405 *
sentiment                0.005837   0.002028   2.878 0.004048 **
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.6112 on 1990 degrees of freedom
Multiple R-squared:  0.1756,	Adjusted R-squared:  0.1719
F-statistic: 47.11 on 9 and 1990 DF,  p-value: < 2.2e-16
```