AirBnB Rental Pricing Guide - Paris, France

Sarah Jordan
Capstone 2 Final Report
March 24, 2018

Report Objective

It can be difficult to break into the AirBnb world as a host; there are thousands of rentals in large cities, and not much guidance on how hosts should price their rentals to optimize bookings and to stand out from the crowd. I will use AirBnb rental data in Paris, France to help Parisian AirBnb hosts understand how much they should charge based on rental size, amenities, and location in order to actually increase their monthly growth rate and get more bookings.

Recommendations

1. Rentals with higher numbers of reviews tend to get more reviews. Get renters to review your rental, and use the following two recommendations to help up your compound monthly growth rate for bookings/reviews.
2. If you have a rental that is a shared room, consider a significant price drop (~$20 per night). Shared-room rentals with high CMGRs tend to be much cheaper than those with low CMGRs.
3. To up your rentals per month and compound monthly bookings growth rate, try decreasing the price your rental a little bit, especially if you are located in the inner arrondissements (1-8) or arrondissement 16. Rentals with high compound monthly growth rates tend to be cheaper than rentals that are not seeing as much growth, particularly in these arrondissements. A price drop of just ~$2-$8 per night could help make your rental more competitive.
4. AirBnb should conduct a more robust study, as outlined in the ‘Future Research’ section, to help make more confident and concrete pricing recommendations based on a more in-depth list of rental features.

The Data

I. Datasets

I used the following datasets for my analysis:

- Paris AirBnB data from 2016-2017, source
- Paris neighborhood dataset, source

II. Data Wrangling

Paris AirBnB Data:

The AirBnB data was scraped from AirBnb.com at different dates. Each file represents a different audit date. In this report, we explore both the data from a single audit, as well as data from audits over time in 2016 and 2017. All data wrangling can be found here.

1. Single Audit:
    - *Removing unnecessary columns:* Certain columns, such as borough and city were filled with entirely null values. I deleted these columns. There were also columns with information that was irrelevant to my analysis, such as the longitude and latitude of the rental. I also removed these columns.

2. Audits Over Time:
    - *Concatenating multiple CSV files*: the data was scraped at a variety of different dates. I concatenated all of the data collected between 2016 and 2017 into a single dataframe so that I could look at changes in the number of reviews over time.
    - *Imputing null values:*
        i. Bedrooms: I filled in the missing values with the mean number of bedrooms, rounded to the nearest integer.
        ii. Overall Satisfaction: for any listing that had 0 reviews, I filled in 0 for the overall_satisfaction value since rentals cannot be rated without a review.
    - *Dropping rows with null values:*
        i. Overall Satisfaction: There were still a few remaining overall_satisfaction rows that had null values. I opted to drop these instead of imputing values because of the distribution of values: most listings have an overall satisfaction of about 4-5 stars, but many also have a 0 rating because they have not been reviewed. Taking an average would give a rating that

doesn't fit in with the rest of the data, and since there are not a ton of instances like this, I would rather just delete these.

ii. Price: Price will serve as a response variable, so I did not want to impute values. There were not many null values so I dropped the instances where price was null.

iii. Accommodates: There were also not many instances with null 'accommodates' values, so I opted to drop these as well.

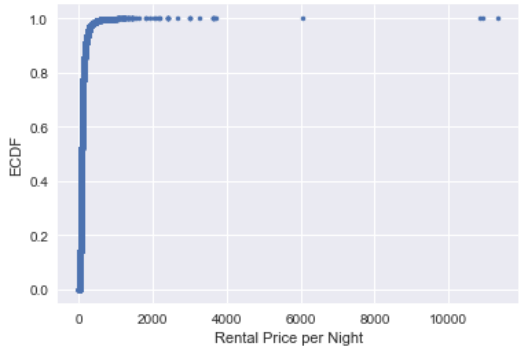
- *Removing outliers:* rentals that had extremely high numbers of bedrooms or people they could accommodate were removed from the data to exclude extreme cases.

Paris Neighborhood Data:

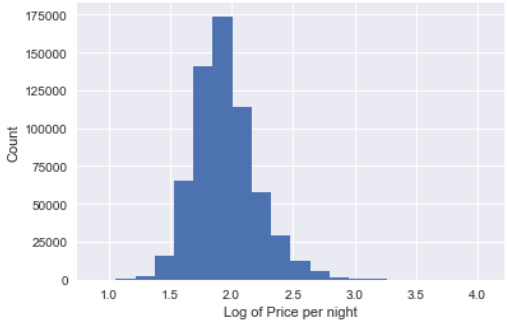
- *Removing unnecessary columns:* I only needed information on each neighborhood and its corresponding arrondissement. I pulled just these two columns from the data.
- *Merging with Paris AirBnB data:* I merged this dataframe with my AirBnB data (both from a single audit and audits over time) on the neighborhood column.

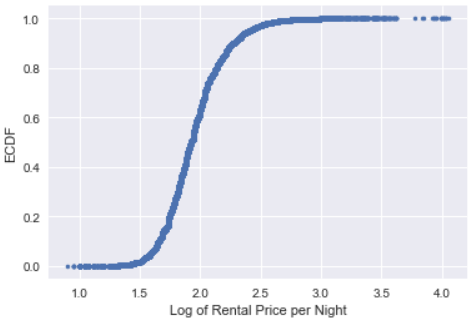
Variable Transformations

The distribution of rental prices is heavily skewed; most rentals cost between about $50-$125 per night. However there are a fair number of rentals with sky high prices: up to $11,323 per night. The skew in the data can be seen below:



I decided to take the log of the rental price to see if this would give a more normal distribution of pricing data, and found that it makes sense to transform the response variable in this way, as exhibited below:



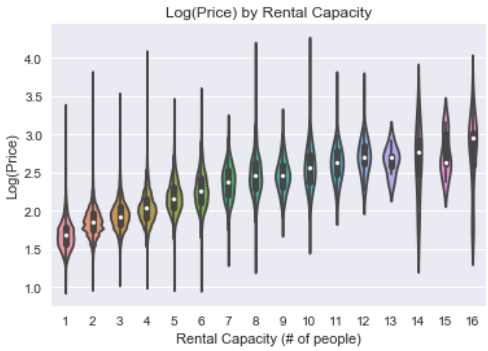


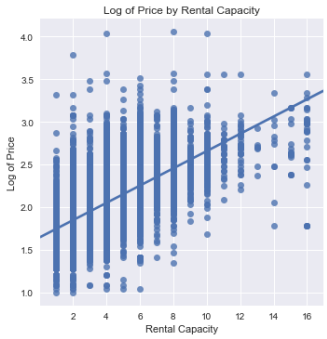
Single Audit AirBnB Data

I. Exploratory Data Analysis

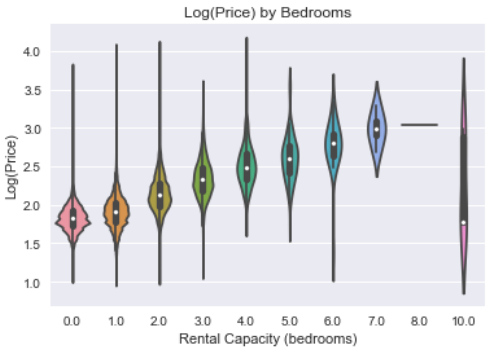
Relationship Between Rental Features and Price - code

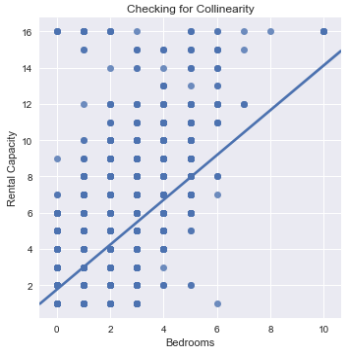
Rental Capacity: The price tends to go up as the number of people a rental can accommodate goes up:



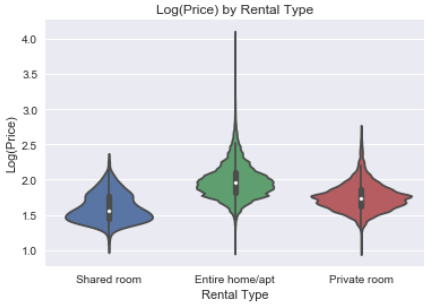


Number of Bedrooms: The same can be said of the price and the number of bedrooms. We can also see that the number of bedrooms tends to increase as the rental capacity increases.

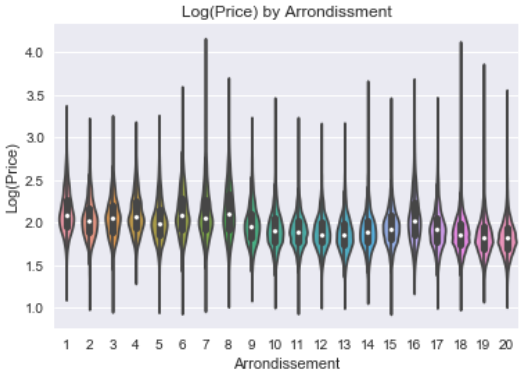




Rental Type: Rentals of an entire home or apartment tend to be the most expensive, then private rooms, followed by shared rooms.



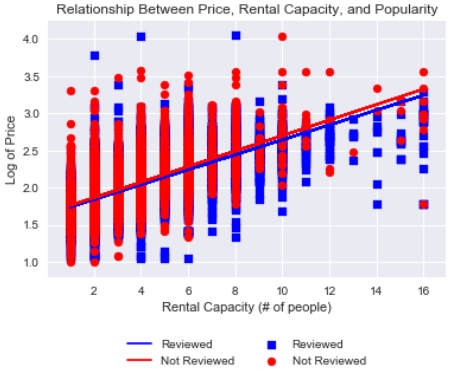
Arrondissement: arrondissements 1-7 have slightly higher nightly rental prices than the other arrondissements; this makes sense, as these arrondissements are at the city center, and have major attractions such as the Eiffel Tower, the Louvre, Notre Dame, the Arc de Triomphe, etc.



Relationship Between Rental Features and Popularity - code

AirBnb says that ~70% of people who stay in a rental end up reviewing the rental, so it can be thought of an approximate proxy for total stays in a rental. We see that the price levels out within a fairly steady range as the number of reviews increases, indicating that there may be 'sweet spot' rental price (or range of prices) that leads to popularity.

I split the data between rentals that had been reviewed, and those that had not, and I found that rental prices tend to be slightly higher among rentals that have not been reviewed yet.



I also split the data between rentals that had a ‘typical’ number of reviews (those below 75th percentile number of reviews of the rental review data) and those that had a high number of reviews (those in the 75th percentile of review numbers). The range of prices we see with a high number of reviews becomes smaller, and we can see that the price increases less per person accommodated in a rental when there are a higher number of reviews.

II. Linear Regression

In order to quantify the differences in pricing between popular rentals (those with a high number of reviews) and those that have fewer bookings (those with a low number of reviews), we will divide the data by their number of reviews and create two multiple linear regression models between rental features and the price. By comparing the coefficients of these models, we can provide concrete recommendations for AirBnB hosts on how to price their rentals for maximal bookings. All code for this section can be found here.

Feature Selection

Here, the response variable is log of price. The input variables are the number of people a rental can accommodate, the number of bedrooms, the rental type (entire home/apartment, private room, or shared room), and arrondissement.

The Models

Using OLS from statsmodels, I fit several multiple linear regression models on the features listed above. I made these regression models in pairs to compare rentals with reviews above and below a certain threshold. These pairs are listed below:

- *Reviewed vs. Not:* Rentals with 0 reviews and those with 1 or more reviews
- *50th percentile threshold;* Rentals with 8 or fewer reviews and those with 9 or more
- *75th percentile threshold:* Rentals with 22 or fewer reviews and those with 23 or more
- *95th percentile threshold:* Rentals with 82 or fewer reviews and those with 83 or more

These models all have rather low R^2 values (between 0.53 and 0.61), so anyone looking to predict their price using these models should take caution; we can see that there is a spread in observed versus predicted values for these models:

These graphs show the actual price against the predicted price for the model for regressions on reviewed vs. non-reviewed rentals and is representative of all of the regression models built. The model does okay, but the real value in these models are not the predictive power, but rather the comparative power.

If we calculate a z-score between the coefficients of these models, we can see (a) see which coefficients have statistically significant differences and (b) see where the real differences are in pricing between rentals that are popular (have high numbers of reviews) and those that are not (low reviews).

Findings:

Based on z-scores of comparison coefficients between our pairs of regressions, we find that with everything else held constant…

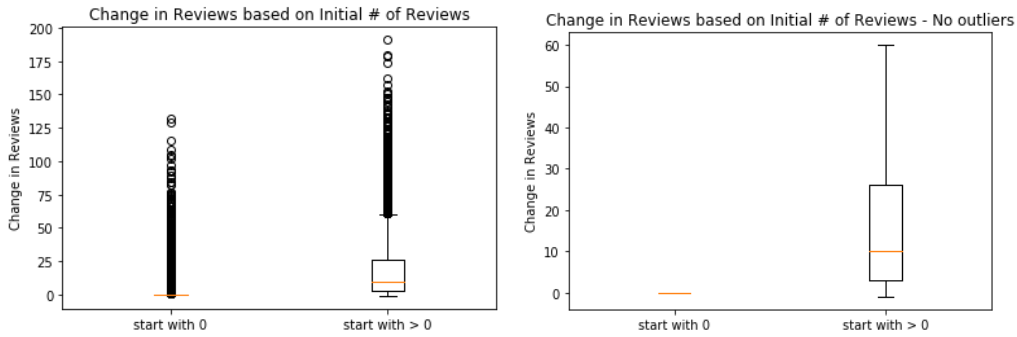
- Rentals with lower numbers of reviews tend to be more expensive per person they can accommodate.
- Rentals in arrondissements 1, 8, and 16 tend to be priced higher in rentals that have lower numbers of reviews.
- Private rooms and shared rooms tend to be more expensive in rentals with lower numbers of reviews

AirBnB Audits Over Time

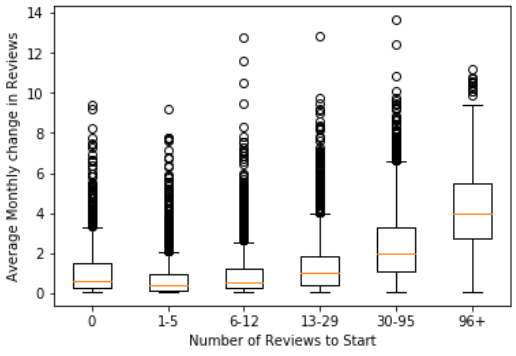
While an analysis of a single month of data is interesting, and can provide some interesting insights on pricing, we are lacking a critical element in our analysis: time. Some of the rentals during the audit we analyze may have just joined AirBnb, and they only have a low number of reviews because they are new, not because their pricing scheme is flawed. There are multiple audits of rentals at various dates, so now we can investigate how change in rentals (and by extension, change in popularity) over time is related to rental features and rental price. All the code for this analysis can be found here.

I. Exploratory Data Analysis

Exploring the data, we can see that rentals that start with no reviews tend to have a smaller change in reviews than rentals that start with rentals.

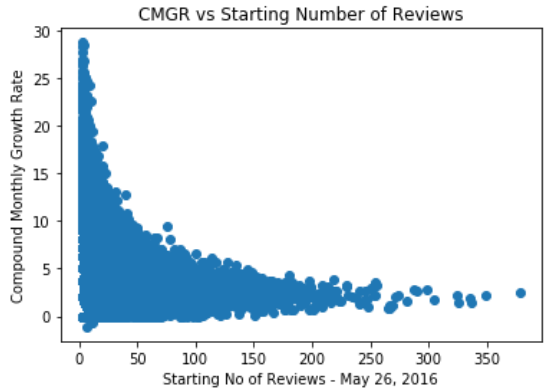


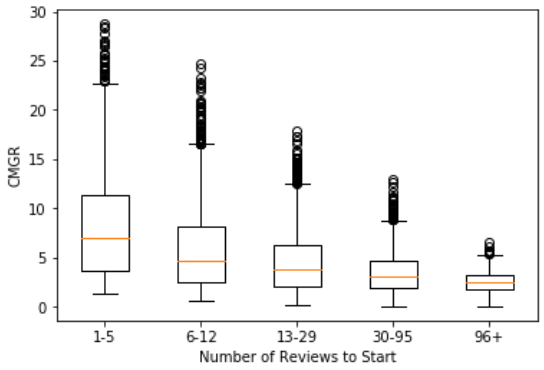
This is true throughout the data. When we break the data into segments by the number of reviews at the start of the series of audits by percentiles (0 reviews, lower 25%, 25%-50%, 50%-75%, 75%-95%, and 95% plus of reviews to start), we can see that rentals with more reviews to start tend to have higher average monthly increases in rentals as well.



It is interesting to then look at the compound monthly growth rate (CMGR) for these rentals between May 26, 2016 and July 25, 2017. We can see that rentals that start with a low number of reviews tend to have a wider range of CMGRs, which is to be expected as small changes will

result in higher growth rates for rentals that start with a low number of reviews. This can be useful for our analysis, though; if rentals with higher reviews also get more reviews per month, looking at what helps rentals have high CMGRs when they have few rentals can tell us how hosts can increase their number of reviews and break into the elite group of hosts that get the majority of rentals each month.





II. Inferential Statistics

ANOVA

We want to check if the difference in the average monthly change in reviews is actually statistically significant. First, we test if the difference between no reviews and having reviews is statistically significant using one-way ANOVA, testing the null hypothesis:

$$H_0 : \mu_{notreviewed} = \mu_{reviewed}$$

We get a p-value of <0.01, so we can reject the null hypothesis with <99% confidence and say that rentals that start with reviews have a higher mean average monthly change in reviews.

Tukey's Test

Next, we used Tukey's Test to see if there is a statistically significant difference between each of our datasets (which is data binned based on their starting number of reviews). Tukey's tells us that there is a statistically significant difference between each of these groups, with rentals with 1-5 reviews to start seeing the smallest average changes and rentals with 95+ reviews to start seeing the highest average changes.

III. Linear Regression

Next, we would like to be able to quantify the differences in pricing between rentals with high and low compound monthly review growth rates. Since we can see that rentals with a high number of reviews tend to get more reviews per month, the best chance that a host has to break into that group is to up their compound monthly growth rate to quickly get to a point where they have a high number of reviews, which is clearly an enticing rental feature of AirBnBs for renters.

Feature Selection

Our features will be the same as for the single audit regression: the response variable is log of price. The input variables are the number of people a rental can accommodate, the number of bedrooms, the rental type (entire home/apartment, private room, or shared room), and arrondissement.

The Models

Using OLS from statsmodels, I fit several multiple linear regression models on the features listed above. I made these regression models in pairs to compare rentals with CMGRs above and below a certain threshold and/or in pairs to compare rentals with a change in reviews above and below a certain threshold. The data were also binned into segments based on their starting number of reviews; this can help hosts look at their current number of reviews, consider their goals, and see what they can do to get the growth rate they desire. The segments were broken down as follows:

- *Change in Reviews:*
    - No change in reviews vs. change in reviews
    - No change in reviews vs. change in reviews for rentals starting with 0 reviews
- *Compound Monthly Growth Rate:* each of these models was created in pairs, with rentals where the compound monthly growth rate was below the 50th percentile and those where it was above the 50th percentile:
    - Rentals starting with 1-5 reviews
    - Rentals starting with 6-12 reviews
    - Rentals starting with 13-29 reviews
    - Rentals starting with 30-95 reviews
    - Rentals starting with 96+ reviews

Once again, these models all have rather low R^2 values (between 0.53 and 0.62), so anyone looking to predict their price using these models should take caution. The value in the models will once again be comparative rather than predictive.

Findings

Change in Reviews:

Rentals that do not change in their number of reviews over the course of fourteen months…

- Tend to be less expensive per bedroom than rentals that gained reviews
- Are more expensive for all rental types (entire houses/apartments, private rooms, and shared rooms) than rentals that gained reviews
- Tend to be less expensive than rentals that gained reviews in the outer arrondissements (#s 9-18)

Of rentals that start with 0 reviews, there are no statistically significant differences in rental features between those that increase their number of reviews and those that remained at 0 reviews.

Compound Monthly Growth Rate:

There were most differences in these groups between the bottom 50% of CMGRs and the top 50% of CMGRs, so we will focus on those regression pairs.

Of rentals that start with 1-5 reviews, those with lower CMGRs:

- Tend to be less expensive per bedroom than those with high CMGRs
- Tend to be more expensive in arrondissements 6, 9, 10, 16, and 18 than comparable rentals with higher CMGRs

Of rentals that start with 6-12 reviews, those with lower CMGRs:

- Tend to be less expensive per bedroom than those with high CMGRs
- Tend to be more expensive in arrondissements 1, 3, and 16 than comparable rentals with higher CMGrs

Of rentals that start with 13-29 reviews, those with lower CMGRs:

- Tend to be more expensive per bedroom than those with high CMGRs
- Tend to be more expensive in almost every arrondissement than those with higher CMGRs

Of rentals that start with 30-95 reviews, those with lower CMGRs:

- Tend to be far more expensive per person accommodated than those with high CMGRs
- Tend to be more expensive for all rental types (entire home/apt, private room, shared room) than those with higher CMGRs

Of rentals that start with over 96 reviews, those with lower CMGRs:

- Tend to be less expensive per person than rentals with higher CMGRs
- Tend to be less expensive for shared rooms than those with higher CMGRs

Discussion: Recommendations

1. ***For hosts with 0 reviews***: There are no statistically significant differences between rental features or rental pricing in the data to capture why and how hosts get their first few reviews/bookings. Since renters with more reviews tend to get more rentals per month, if you do manage to get a booking, encourage people who stay with you to review their experience. If you're struggling to get that first booking, do a trial booking of your rental with friends or family and get them to leave a review. Getting your first few reviews the board will likely be your biggest hurdle in getting more bookings each month.
2. ***For hosts with fewer than 95 reviews:*** If you're trying to up your compound monthly rental booking rate, try reducing the price of your rental, especially if you have fewer than 30 reviews and you are in the inner arrondissements (1-8) or the 16th arrondissement. Many hosts likely get too excited by the proximity of their rental to major attractions in these arrondissements; however, it appears that renters tend to opt for the cheaper options.
3. ***For hosts with 96+ reviews:*** At this point, there are fewer statistically significant differences between rentals with high growth rates and those with lower growth rates, and these rentals actually tend to be priced higher per person they can accommodate and for shared rooms than comparable rentals with lower growth rates. Certain well established rentals may have qualities we have not captured in the data here (quality of photos in the posting, desirable location, etc).

Removing Price Outliers

While the models created in the Single Audit AirBnb Data and AirBnb Audits Over Time sections are helpful in capturing where pricing faux pas may occur, it would be helpful to have some tangible pricing recommendations for hosts on *how much* they should be increasing or reducing their rental price based on certain rental features.

In this section, we remove outliers from the price to see if fitting a regression on rental features and price will create a stronger pricing model based on rental transformation. All code for this section is here.

I. Linear Regression

Feature Selection

Here, the response variable is price (once outliers have been removed). The input variables are the number of people a rental can accommodate, the number of bedrooms, the rental type (entire home/apartment, private room, or shared room), and arrondissement.

The Models

Using OLS from statsmodels, I fit several multiple linear regression models on the features listed above. The segments were broken down as follows:

- *Change in Reviews:*
    - Below 50th percentile change in in reviews vs. above 50th percentile change in reviews
- *Compound Monthly Growth Rate:*
    - Above 50th percentile of CMGR vs. below 50th percentile of CMGR

These models have even smaller R^2 values (between ~0.4), so any pricing recommendations taken from these models should be taken with caution. However, they do give a ballpark idea of how one might want to adjust their pricing based on their rental features.

Findings

Change in Reviews:
Rentals with a high change in reviews:

- Have a statistically significant difference in the pricing of shared room rentals: based on this model, rental price decreases by $24 in rentals with a small change in reviews whereas rental price decreases by $46 in rentals with a high change in reviews.
- Are less expensive in arrondissements 1, 3, and 6 than those with a lower increase in reviews. In our model, being in any given arrondissement changes the price of the rental by a certain amount. The differences between the increases in price between rentals with large vs small changes in reviews are outlined below:
    - Arrondissement 1: $43 for large change, $51 for small change
    - Arrondissement 3: $37 for large change, $41 for small change
    - Arrondissement 6: $30 for large change, $30 for small change

Compound Monthly Growth Rate:
Rentals with high CMGRS:

- Have a statistically significant difference in the pricing of shared rooms and private rooms from rentals with low CMGRS

    - Shared Rooms: rental price decreases by $39 for rentals with low CMGRs and decreases by $51 for rentals with high CMGRs
    - Private Rooms: rental price decreases by $22 for rentals with low CMGRs and decreases by $24 for rentals with high CMGRs
- Tend to be cheaper per person:
    - Price increases by $9 per person in rentals with low CMGRs vs $8 per person in rentals with high CMGRs
- Are cheaper in arrondissements 3, 5, 6, 10, 11, 16, 17, 18, and 19 than rentals with low CMGRs; the difference between rental prices with high and low CMGRs is between $1 and $6 in these arrondissements.

Future Research

As mentioned earlier in the report, the R^2 values of the models created here are relatively low, so they do not hold much predictive power, but rather are helpful in finding key levers in how pricing and rental features are associated with high or low rental booking growth rates. While this is valuable, it would also be helpful to create a more robust model that can more accurately predict price based on rental features to provide a more concrete pricing model for hosts.

The following rental features should be analyzed to captures more than just rental capacity, rental type, and general rental location and help create more concrete recommendations for AirBnb hosts:

- In-Depth Amenities Analysis: is there wifi, washer/dryer, hot tub, pools, terraces, etc? How many beds are there?
- Robust Location Analysis: how close/accessible is this rental to major attractions, the metro, the airport?

Final Recommendations

1. Rentals with higher numbers of reviews tend to get more reviews. Get renters to review your rental, and use the following two recommendations to help up your compound monthly growth rate for bookings/reviews.
2. If you have a rental that is a shared room, consider a significant price drop (~$20 per night). Shared-room rentals with high CMGRs tend to be much cheaper than those with low CMGRs.
3. To up your rentals per month and compound monthly bookings growth rate, try decreasing the price your rental a little bit, especially if you are located in the inner arrondissements (1-8) or arrondissement 16. Rentals with high compound monthly growth rates tend to be cheaper than rentals that are not seeing as much growth,

particularly in these arrondissements. A price drop of just ~$2-$8 per night could help make your rental more competitive.

4. AirBnb should conduct a more robust study, as outlined in the ‘Future Research’ section, to help make more confident and concrete pricing recommendations based on a more in-depth list of rental features.