Springboard Data Science Capstone Project - 2

Predicting Pricing for AirBnB Paris, France

MILESTONE REPORT

Carolyn Massa

May, 2020

Contents

1 Introduction/Null Hypothesis 3-7

2 Data Acquisition and Cleaning 7-10

3 Data Exploration 10-27

4 Data Preprocessing/Feature Engineering 27-29

5 Model Fitting 29-33

6 SHAP Analysis.....30-31

7 Null Hypothesis test 31-31

8 Text Analysis 32-32

9 Conclusions..... 33-35

10 Appendices.....35-40

Introduction – Problem Statement

Since I have lived in Paris, France I constantly hear the current Paris Mayor, Ann Hildago, discuss how AirBnB needs to be stopped as it is harming Paris rental rates and hotel prices. She suggests that both potential renters and hotels alike are damaged with the increase in Airbnb rentals. She asserts that tourists will use hotel less and those in need of housing are being priced out of the market. In a city that has a serious housing crisis, not unlike San Francisco, I had to see what is really going on. How much will AirBnB rates increase?

Problem: Will Airbnb rates continue to increase every year? What season is best for travelers to get the best rates? What factors contribute to the pricing of Airbnb listings?
Code to my project can be found here. Powerpoint here.

Null Hypothesis: Since many of my friends tell me they rarely listen to reviews when making a decision I will state this as my HO: ***"AirBnB Paris locations near the Top 10 attractions have little impact in the price of the listing"***

Methodology

1. Preprocessing and Cleaning using Python
2. Feature Extraction and Data Visualizations
3. NLP WordCloud to analyze the most effective words to attract guests
4. Analysis of various supervised learning methods such as XG Boost and the Spatial Hedonic Price Model (HPM)

Let's look at AirBnB Details

According the publication "The Local" in October, 2015 we review the sentiment:

527,821
That's the number of visitors to Paris who used Airbnb in the summer of 2014. That compares to just 144, 000 in the summer of 2009, which is why Paris has become the number one destination for Airbnb users, surpassing London, New York and Rio de Janeiro.

5 million
The number of tourists who have used the Airbnb site since 2008 to find accommodation in Paris. Of those, 2.5 million visitors used the site in 2015 alone which gives an idea of the sharp increase in popularity for the home-sharing website in recent months.

50,000
That's the number of apartments available to rent on Airbnb in Paris in 2015, compared to just 4,000 back in 2012.

2016 Regulation
On average, a Parisian host rents their property 26 nights per year through **Airbnb**. **France passed** a **law** in 2016 which limits owners to renting out their residences that they own for up to 120 days per year, while their primary residences do not currently have limits.

There is little doubt that AIRBNB has taken over Paris so I will use the years of 2015 (when listings increased) and the year of 2016 (when legislation was passed to regulation AirBnB listings) to check for insights.

Background, Problem Statement and Desired outcomes:

AirBnB Price Prediction – This study is an exploration into key factors that determine the pricing of an AIRBNB which allows home-owners and renters ('hosts') to put their properties ('listings') online, so that guests can pay to stay in them; whether they are entire apartments owned by the host or part of their own living space.

Hosts are expected to set their own prices for their listings or they can pay a fee to use a service () such as "Beyond Pricing" and others like "Smart Pricing" which will change pricing according to Supply and Demand and other typical factors such as holidays, special events and the like. Although Airbnb and other sites provide some general guidance, there are currently no free services which help hosts price their properties. Airbnb Hosts typically pay between 3 and 14% of their earnings to use the Airbnb platform. *Source here. These 3rd party algorithms will change each daily price around that base price on each day depending on day of the week, the season, how far away the date is, and other misc. factors. In order to truly maximize revenues and get pricing strategy correct, it is important to understand the factors involved.

Paris, France ranks as one of the most visited places in the World with highly ranked Museums, world renowned cuisine and impressive architecture- therefore high cost hotels. That being said, the number of hosts has grown exponentially since 2015 and competition is fierce. The goal of this analysis is to use machine learning to predict a type of model to learn how to forecast the best price, and also to explore Airbnb listing data, in order to help Airbnb hosts to maximize their earnings. It is important to note that Paris Proper is 41 square miles with 58,184 listings as of April of 2019 which essentially means There are 1,419 AirBnB per 1 square mile where New York has 6 Airbnb's per every square mile though it is significantly larger.

The only domain expertise I have is I have worked as an Analyst forecasting metals, elastomers, carbides and other regulated materials for the oil and gas domain. I did work as a front desk Manager for Le Meridien when I was just out of my undergrad program, which is a 4 star hotel so I do recall a bit about occupancy rates/timing factors as they relate to the hotel pricing strategies and owned my own tour business when I lived in Paris, France from 2018 to 2020.

Key Factors used and visualizations to achieve:

- Data description
- Locations on the map
- Room/Property Types
- Average prices for each neighbourhood
- Location listing amounts in different neighbourhoods (heatmap)
- Average location price in different neighbourhoods
- Amenities that contribute to the cost – which amenities matter the most?

2) To recommend Airbnb properties given certain criteria, and enable a more informed decision for the traveler:

- walking distance to one of the top attractions (within 2 miles)
- rating ==>8
- other user defined criteria i.e.
- number of beds
- number of bedrooms
- price range
- type of property
- if the lower priced Airbnb properties have less availability, this would affect the probability of finding a property for at that lower price range i.e. for a less than \$100/night budget
- the properties in the lower Arrondissements, those do those listings closest to the top attractions cost more on average compared to the properties in the higher arrondissements.

Null Hypothesis

H0 = "It makes no difference in Price for properties that are located near any of the top 10 attractions in Paris"
HA = “It makes a difference in the price of each AirBnB Rental if it is located within 2 miles of the Top 10 Paris City attractions.

Data Acquisition and Cleaning

I acquire datasets from one primary source and several supplementary sources.

Primary Source of AirBNB Data:

http://insideairbnb.com/get-the-data.html

Supplementary Sources of Data:

- 2018 top 10 attractions in Paris
- Haversine Formula
- I use data from April of 2019 to review the details and explore the data prior to making a comparison to the data from earlier or later times.

THE DATA SETS – From InsideAIRBNB.com: 4 primary data sets for EDA, 1 for Geographical data and 23 Datasets for Historical data and one dataset from TripAdvisor for Paris’ *“Top 10 Attractions”.*

The Calendar Data Set (See Appendix A)

Geography of Paris

To properly map my locations, I import the GEO dataset which contains the following 3 Variables:

	neighbourhood	neighbourhood_group	geometry
0	Batignolles-Monceau	None	MULTIPOLYGON (((2.29517 48.87396, 2.29504 48.8...
1	Palais-Bourbon	None	MULTIPOLYGON (((2.32090 48.86306, 2.32094 48.8...
2	Buttes-Chaumont	None	MULTIPOLYGON (((2.38943 48.90122, 2.39014 48.9...
3	Opéra	None	MULTIPOLYGON (((2.33978 48.88203, 2.33982 48.8...
4	Entrepôt	None	MULTIPOLYGON (((2.36469 48.88437, 2.36485 48.8...

We see below the concentrations of AIRBNB Rentals around Paris:

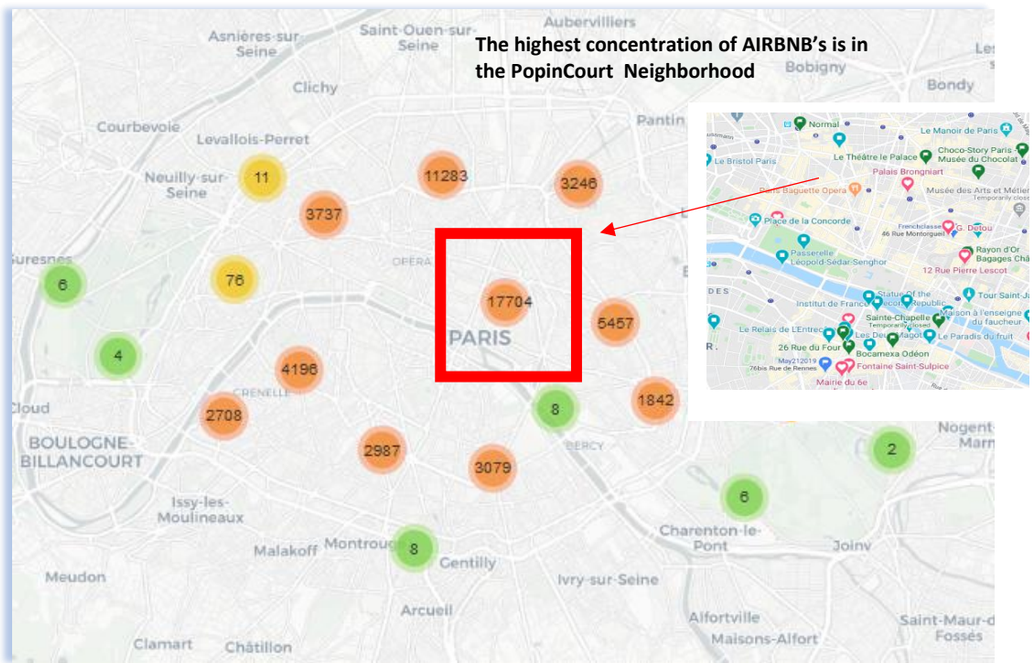


Fig1 Left

This folium map indicates clusters of listings throughout the City of Paris . More details on this will be revealed later in this study.

Data Wrangling & Missing Values:

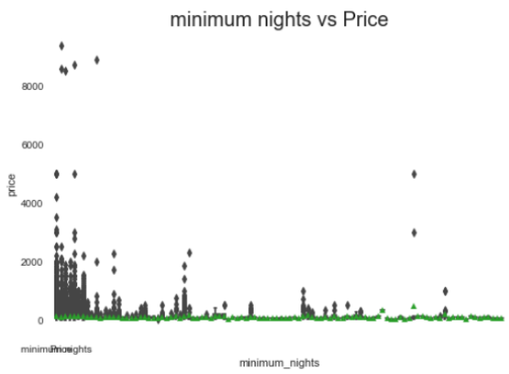
I have several missing values where I imputed them based on the following parameters:

If the 'beds' had a missing value, and 'bedrooms' had a valid value, 'beds' is set with the value of 'bedrooms'. If both 'bedrooms' and 'beds' are missing, both fields are set to 1, which is the average number of beds and bedrooms, and it would be safe to assume that if a property is being rented out in Airbnb, that there is at least one bed regardless if it is a Studio Apartment or One Bedroom.

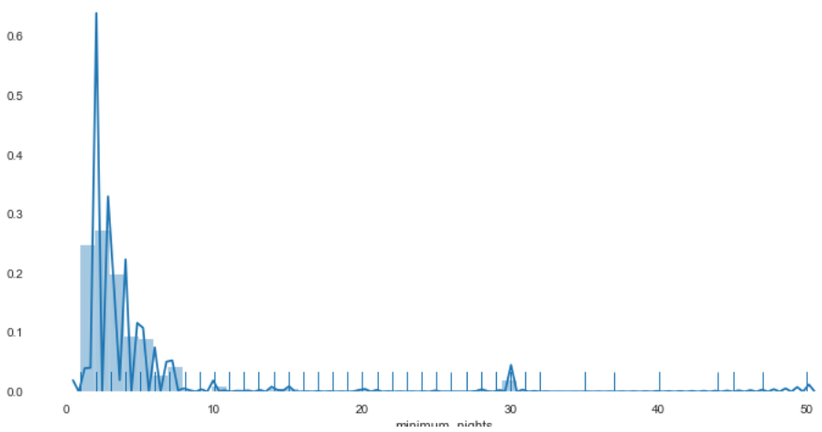
There were several records as well with missing 'scores' value, a new column 'rating_ind' was created to flag rated = 1, versus unrated = 0 records, so further analysis can be done between these two populations. The record was flagged as rated, if all the score values have been populated, and set to unrated, if at least one of the score values has not been populated.

Price Ranges and Minimum Nights

I notice that there are several outliers in both minimum nights (the least # of nights to be booked) and also in the price ranges There are listings with 999 minimum nights and listings with prices as high as $9,379 for a studio apartment in a lower cost neighborhood.



Figs 2 above: Outliers in the “minimum nights” that are required to book a rental



Since this is unusual, I decide to replace all listings over 50 days minimum nights with 15 days as this is the average and I also remove listings that are more than $2500 per night which removes 20 listings from my dataset ; I now have 58,164 to work with.

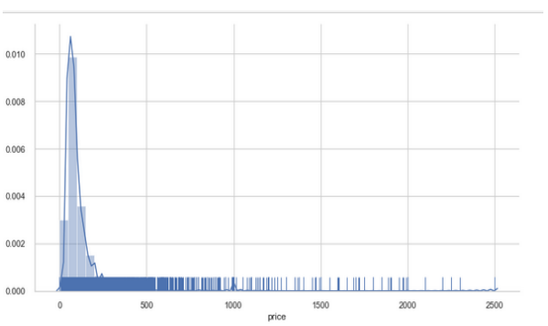


Fig 3 Left: We see a wide distribution in prices so this needed to be reduced to listings between 10 USD and 2500 per night

Fig 4 above: Now we see an improved distribution of rental prices.

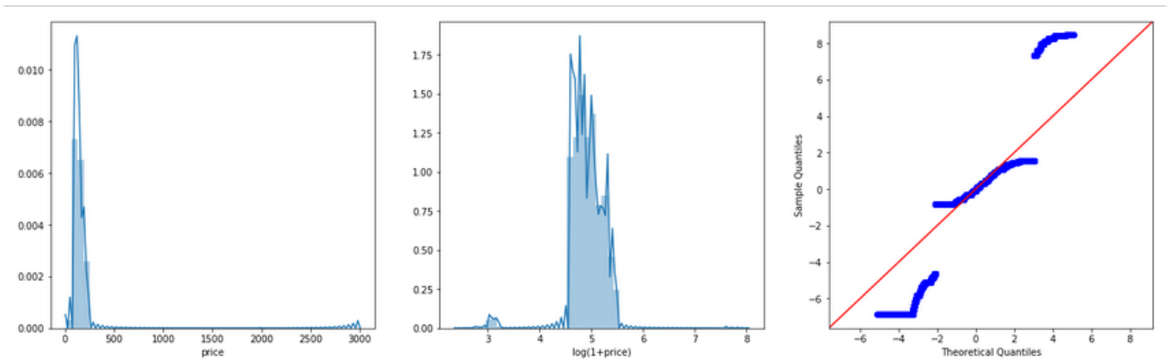


Fig 5 Above: We log our pricing to get a normal distribution

Fig 6: Here we review prices ranging up to 1200 to see the distribution to note the majority are between 20 and

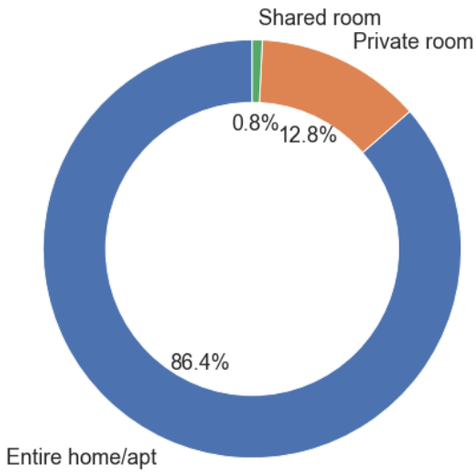
200$USD.

Fig 7: Here we review prices ranging up to 500 to see the distribution to note there are over 260 listings in the $1000 per night rate.

All About the Money?

So, since most travelers with parties of 2 or 4 usually consider prices as a key factor to decision making, let's look at pricing as it relates to a few of our other key variables.

First and foremost, what type of properties are there? What categories can we put them in? Below I run my 55,653 listings though the plotting process to give perspective to our choices:



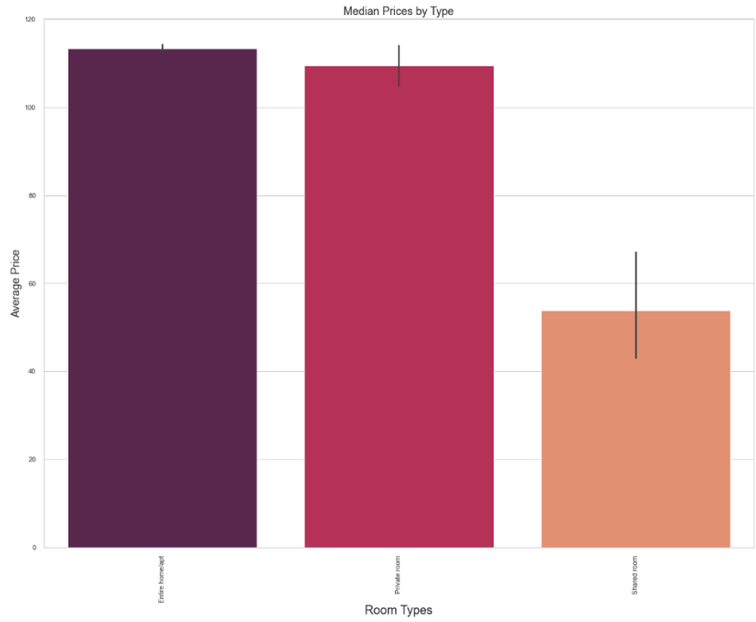


Fig 8a Left: We see the Entire home/apartment category makes up the majority of our listings which is rather odd considering the French Government passed legislation in 2016 to limit “entire dwellings” to 120 days out of the 365 calendar year from being rented.

Fig 8 b Right: It is simple breakdown of property type by price range. An Entire home/apt is $113.39,a Private room is $109.48 & a Shared room is $53.81 on average.

What type of Dwellings are there to be book?

Fig(s) 9 Below: We take the 29 multiple types of properties and put them in 4 different categories which include our “unique experiences” such as staying in an igloo or treehouse, among other structures found in this dataset.

Different Prices by Neighborhood

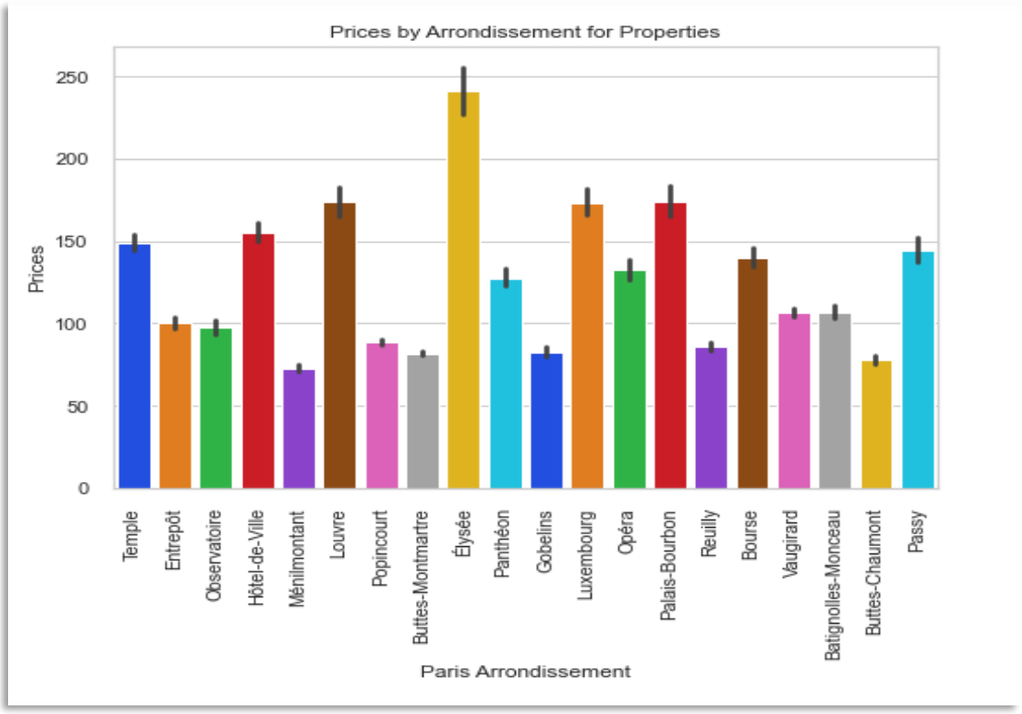


Fig 11 Above: The highest neighborhoods are, as expected, in the Le Elyseè as it has the highest concentrations of attractions. The lowest are to be expected as Menilmontant in the northeast part of Paris has much fewer attractions and is roughly a 30 minute bus/metro ride from the center of Paris. Below our map clearly tells us the price reduces as we go to the southeast and northeast of the city.

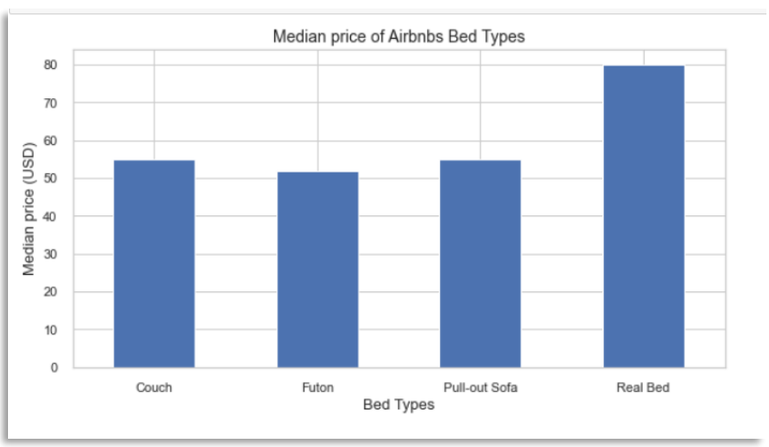


Fig 10 Left: A "Real Bed" in our rentals command a higher price and is has a median price of 80 USD versus a Futon which is 52 USD.

MEDIAN PRICES

Couch	55
Futon	52
Pull-out Sofa	55
Real Bed	80

It's a pity that we don't know the square feet for each location, we cannot calculate the price for each m2. However, we calculate the average price for different neighbourhood.

The top 3 neighbourhood are Elysée (240 dollars), Palais-Bourbon (182 dollars) and Louvre (178 dollars), the bottom 3 neighbourhood are Ménilmontant, Buttes-Chaumont and Gogelins, which might be because of the neighbourhood security, quality, etc. We find an interesting point: the price and locations' amount is negatively related, which means the better neighbourhood, the fewer the Airbnb locations and its price is more expensive than other areas with more Airbnb locations.

Fig 12 Below: A 5minute review of 4 maps created in Folium to show the density of Paris' AIRBNB listings as well as price averages in the 20 neighborhoods or "Arrondissements" as they are called in France. To achieve these 4 maps, I used the Haversine method which calculate the great circle distance between two points on the earth (specified in decimal degrees). I applied this formula using both a for loop and a lambda formula on my GEO location data on each listing.

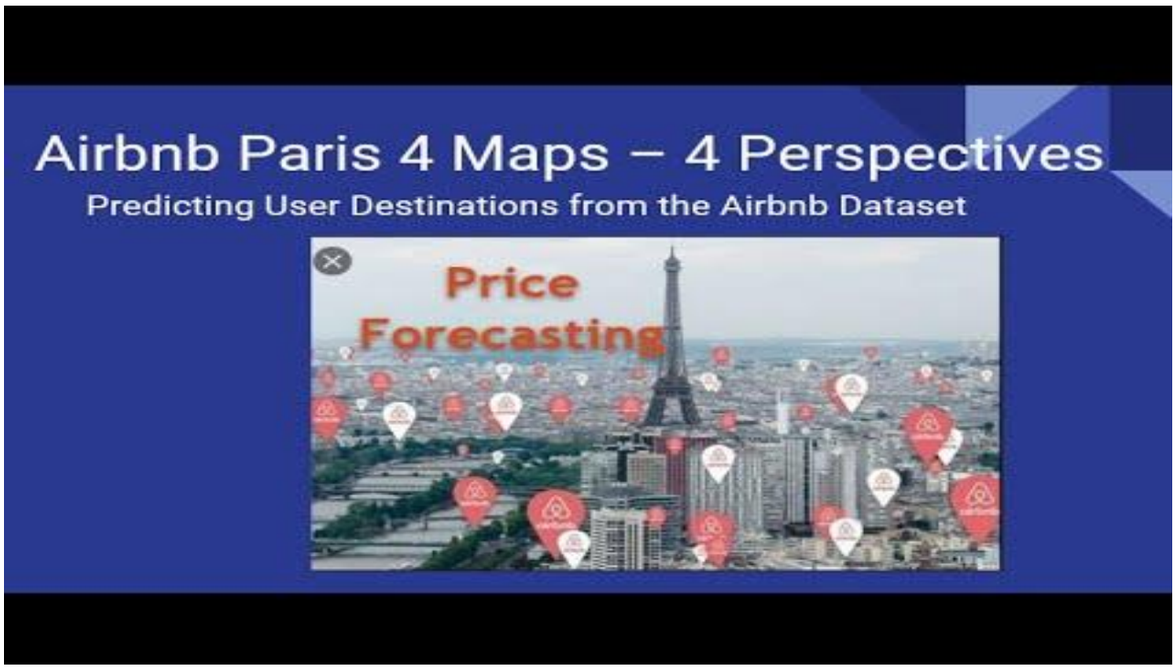


Fig 13 Left Below: We do notice higher “extra people” allowances for prime areas like Elyseè and Luxembourg

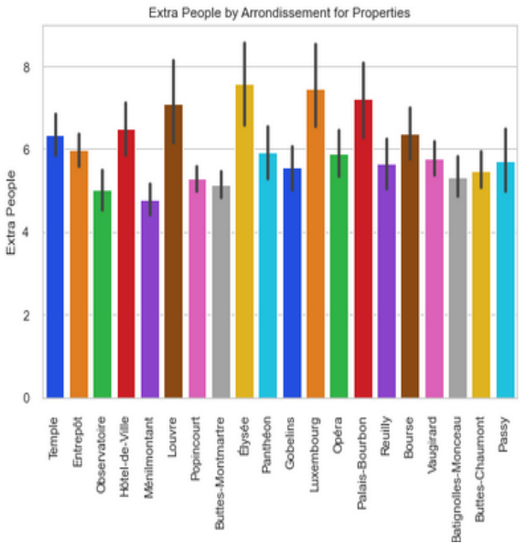
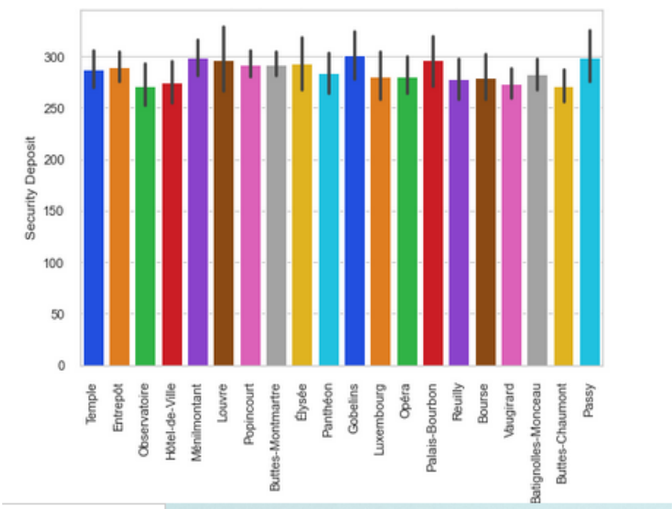


Fig 14 Below (right) The required security deposits do not vary drastically between neighborhoods though Passy (chic elegant area) has the same $300 requirement as Gobelins(Not so nice) area.

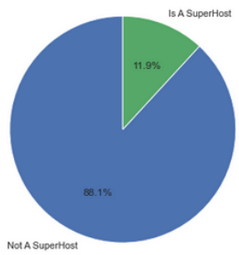


Let’s Meet Our Hosts

Let’s talk about what it takes to be called a “Super Host” in the Airbnb world.

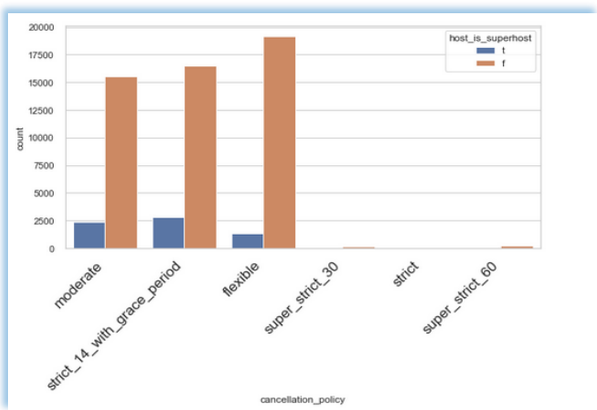
Airbnb awards the title of “Superhost” to a small percentage of its dependable hosts. This is designed as an incentive program that is a win-win for both the host, Airbnb, and their customers. The superhost increases their business in the form of higher bookings, and in turn, the customer gets improved service and Airbnb gets happy satisfied customers.

What are the requirements? Maintaining a review rate above 50%, a response rate above 90%, and other miscellaneous requirements. Here we investigate our dataset to see how the superhosts perform on two parameters : “Response rate” and “Ratings”. Both these variables range from 0 to 100.



First and foremost, what percent of our hosts have been bequeathed with this “prestigious title”? In Paris, as of April, 2019, 6,587 hosts of 48,923 hosts in Paris, France means that only 11.9% of the hosts have met the criteria set by AIRBNB.

Of these superior “Hostesses with Mostesses” what do they do correctly? What sets them apart? Let’s take a look.



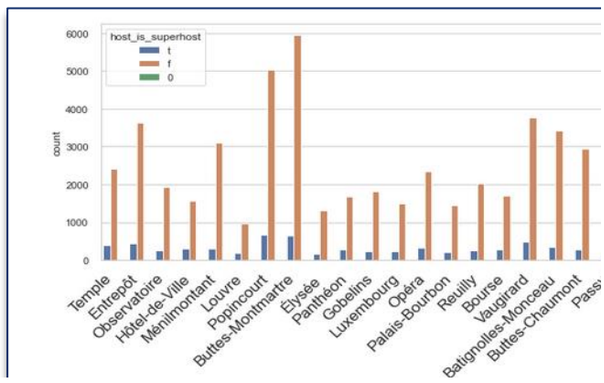
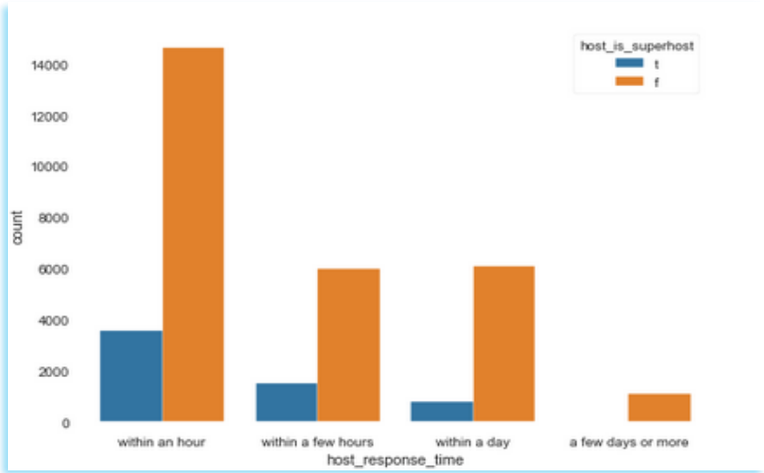


Fig 15 Above: most hosts choose The flexible option though super hosts choose the strict 14 with grace period as their choice

Fig 16 above: We notice that Popincourt and Buttes-Montmartre have the highest % of Super Hosts which is ironic as these are less costly and more “up and coming areas”

Fig 17 Right: We see that of the 5 choices the hosts that are SUPERHOSTS are more likely to reply within an hour to the request of the potential guest



superhost

.

What Are the Benefits of Airbnb ID Verification for Hosts and Guests?

Why does Airbnb verify ID and is it something that benefits all the stakeholders? For guests, the Airbnb ID verification process ensures that hosts are who they say they are. Given the risks that come along with renting properties without viewing them in-person, Airbnb verifies hosts' IDs as a means of establishing trust.

The verification process also helps to ensure that Airbnb and guests have someone who they can hold responsible in the event that a problem arises with a booking.

In the same way that the "Airbnb verify ID" process benefits guests, it also helps hosts. For hosts and Airbnb, ID verification helps to prevent fraud. Since Airbnb hosts rely on Airbnb to process and collect the payments from guests for bookings and experiences on their behalf, ID verification helps to ensure that all payments made via the platform are actually valid transactions.

Across the globe, all hosts and guests are screened against regulatory, sections, and terrorist watch lists. Airbnb may pass along information to banks, financial institutions and law enforcement agencies to facilitate investigations that require the involvement of Airbnb. These investigations may involve tax, money laundering, sanctions laws, and criminal investigations. Source IMGS

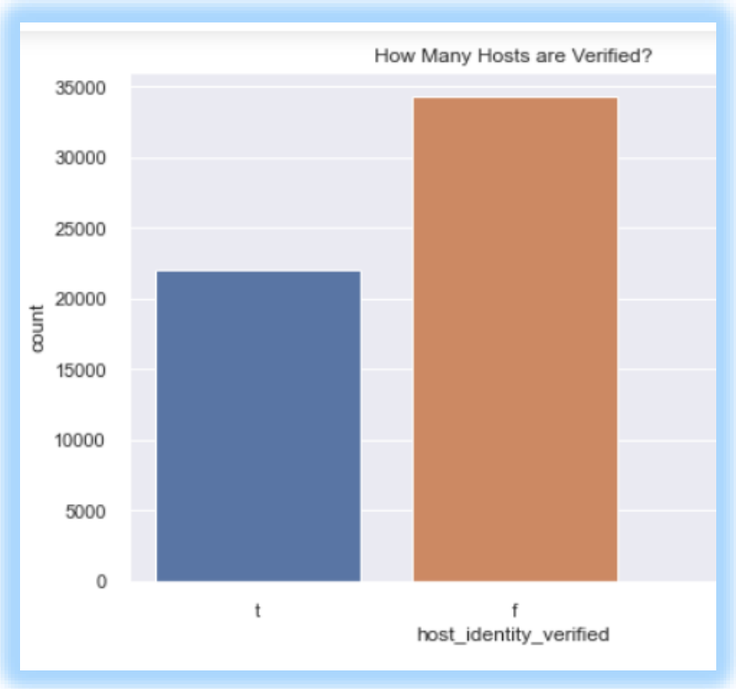


Fig 18 Above: We notice that only 39% of the total hosts have had their identity verified. This leads us to believe that perhaps companies are managing the properties and Airbnb has not identified the individual employee who is managing the property.

Nothing beats Experience

Fig 19 Above: Insights: We notice that those hosts with more than 5 years of experience (as of 2019) run 66% of the locations of that 66% the hosts with 8 years' experience have 11% of listings

Do Ratings Matter?

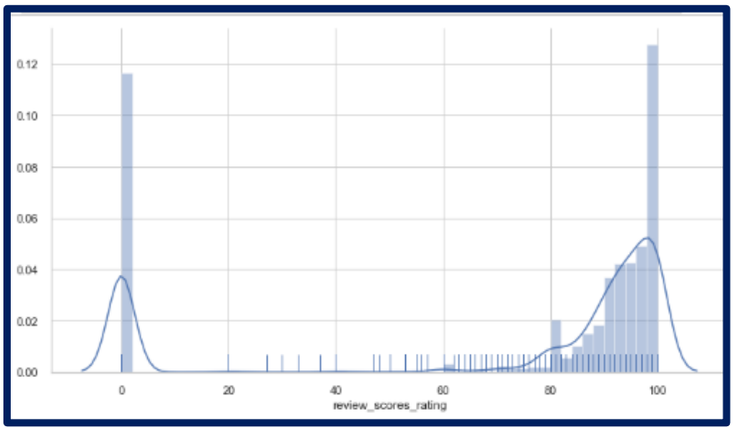


Fig 20 Left

We immediately notice that the majority of our ratings are between 80 and 100% on a scale from 0 to 100%.

Of the 56,961 properties and possible ratings 13,741 went unrated which is 24% of guests who preferred to not rate the property . We may be able to take that as a positive as often people will leave a review if they are not happy though that is my opinion.

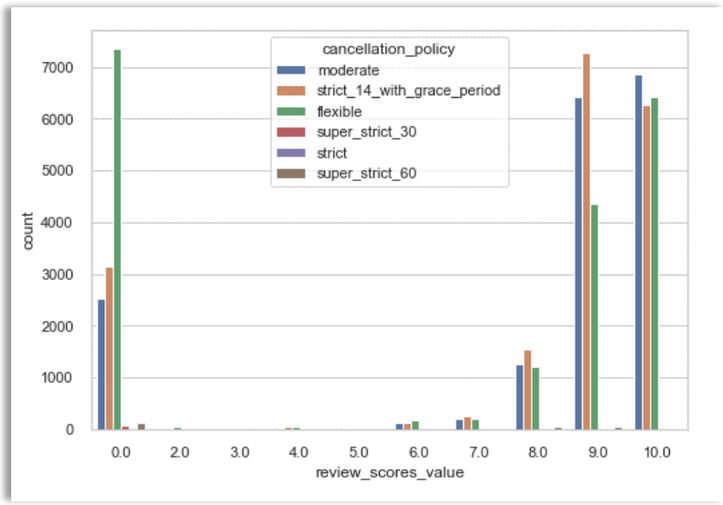


Fig 21 Left Cancellation Policies

These 6 policies range from Flexible (the most lenient) to Super Strict 60.

For example: A flexible cancellation policy is the most lenient option for guests. Under the Flexible cancellation policy, guests are eligible for a full refund when canceling a reservation at least 14 days before check-in. Super Strict which means the guest must cancel 60 days in advance to receive just 50% of the accommodation fees back.

Do Ratings Matter? (cont'd)

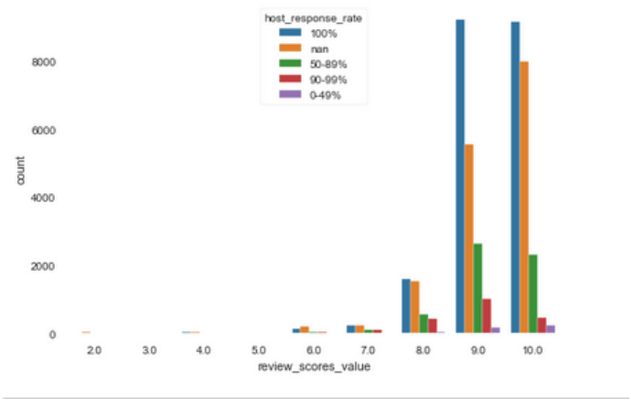


Fig 22 Above Here we compare the host response rate to the ratings; we see a high response rate with the higher ratings

Fig 23 Here we compare the average price by score to notice the higher score correlates to the higher price

Fig 24 Below: We notice that higher priced properties tend to not have ratings; we guess that perhaps they are priced too high and have fewer bookings.

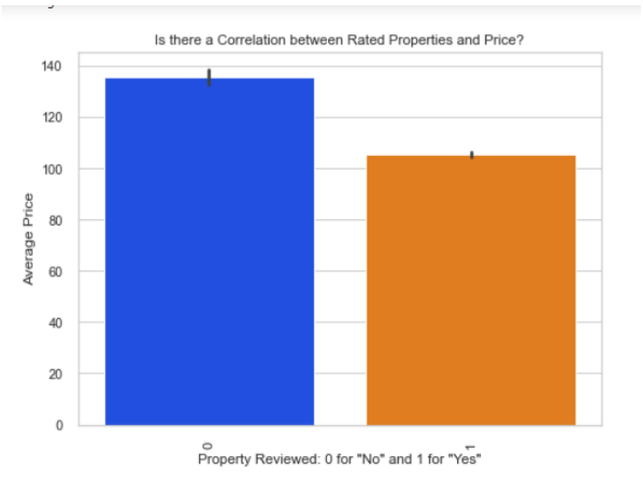
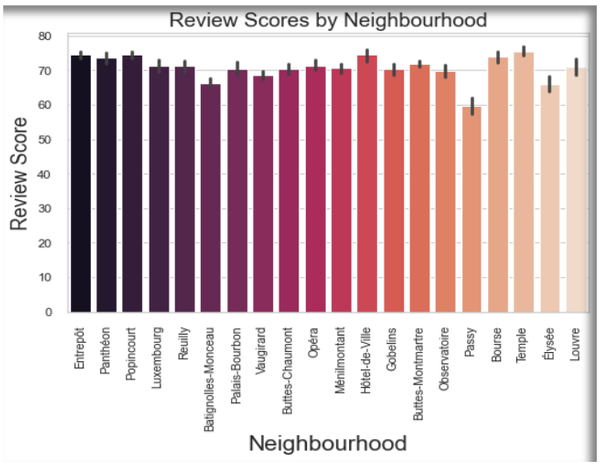


Fig 25 Below: We see higher ratings in the Temple, Popincourt and Hotel de Ville neighbourhoods and the lowest in the Passy area.



Do Ratings Matter? (con't)

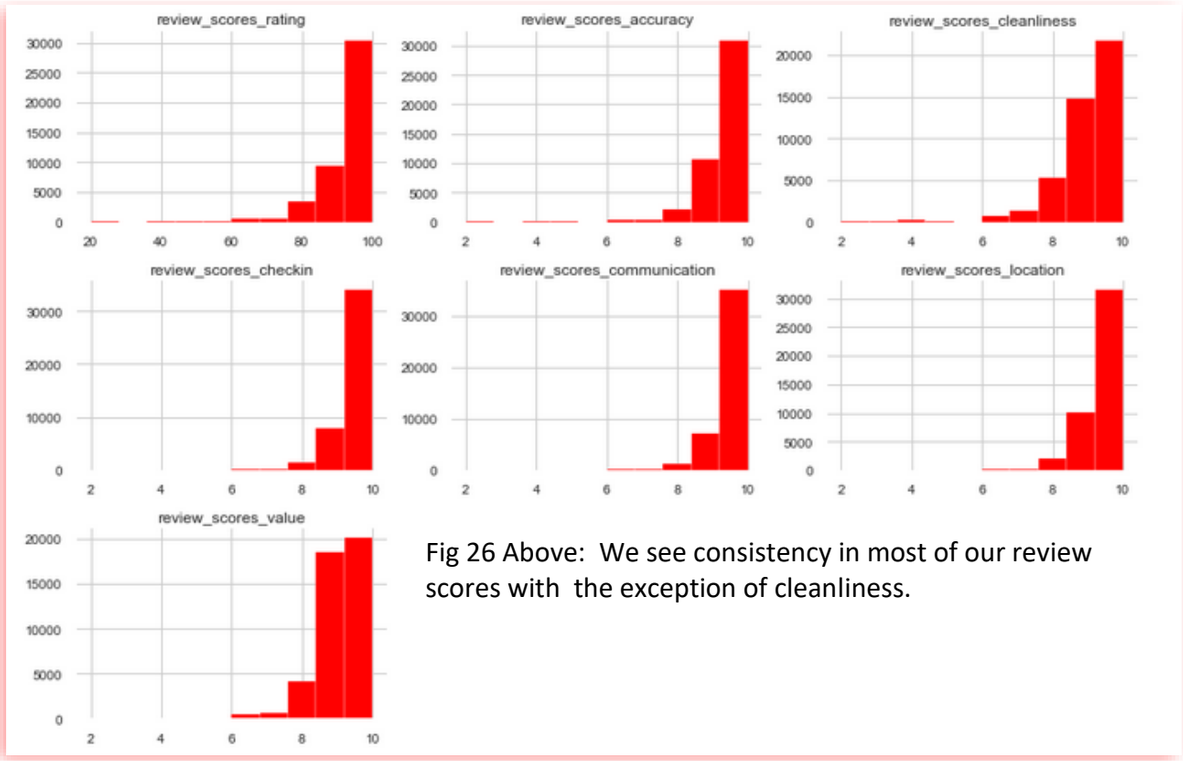
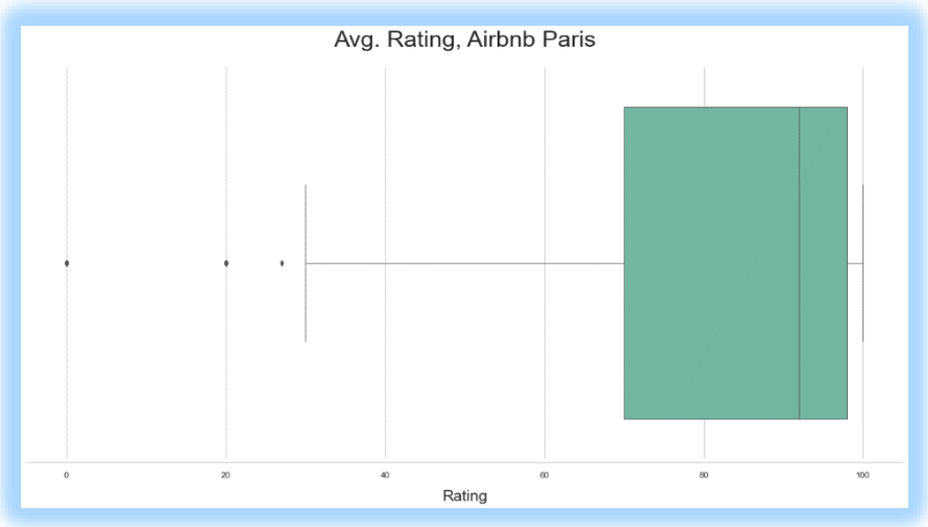


Fig 26 Above: We see consistency in most of our review scores with the exception of cleanliness.

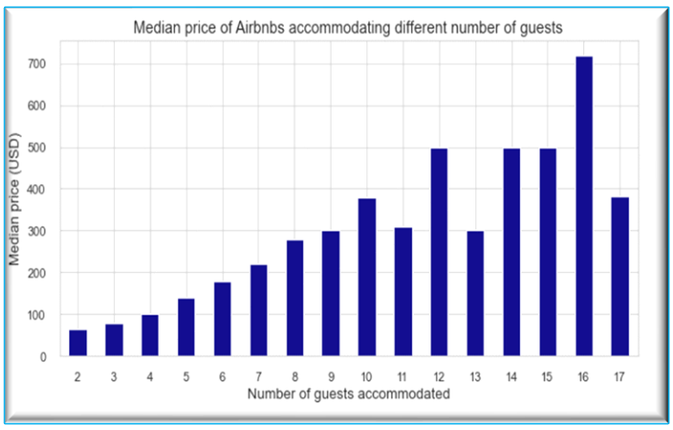


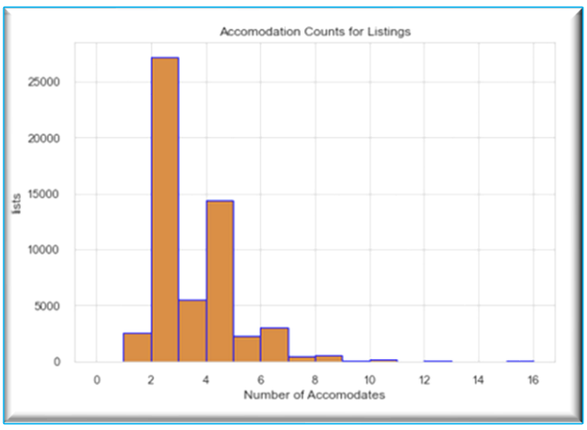
To put the date into perspective we have the highest number of listings in the Buttes-Montmartre area and the lowest number of listings in the Bourse neighbourhood.

neighbourhood	#Listings	neighbourhood	#Listings
Batignolles-Monceau	3789	Gobelins	2061
Bourse	1980	Hôtel-de-Ville	1859
Buttes-Chaumont	3225	Louvre	1152
Buttes-Montmartre	6590	Luxembourg	1734
Entrepôt	4071	Ménilmontant	3401
Observatoire	2171	Popincourt	5698
Opéra	2686	Reuilly	2269
Palais-Bourbon	1640	Temple	2809
Panthéon	1954	Vaugirard	4250
Passy	1549	Élysée	1473

Fig 27 Above: Here we see the Average rating in a Box Plot with the majority falling between 70 and 90%.

Other Factors that contribute to Pricing of AirBnb Rentals





Fig(s)28 a Above Left and 28 b Above Right: We see that over 50% of our listings accommodate 2 persons and the highest priced listings accommodate 16 people and properties that accommodate 12,14 and 15 persons are equal in median price.

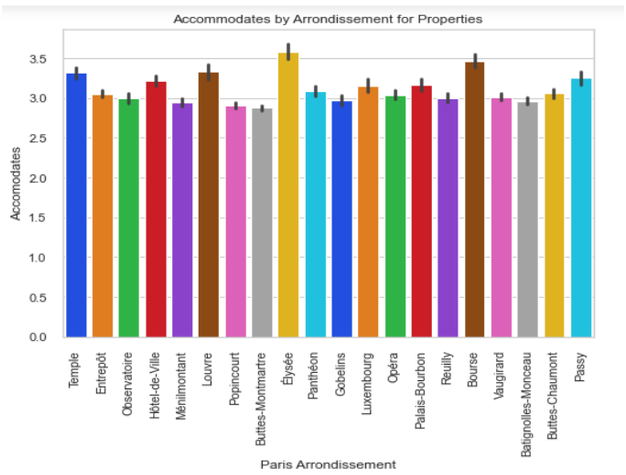
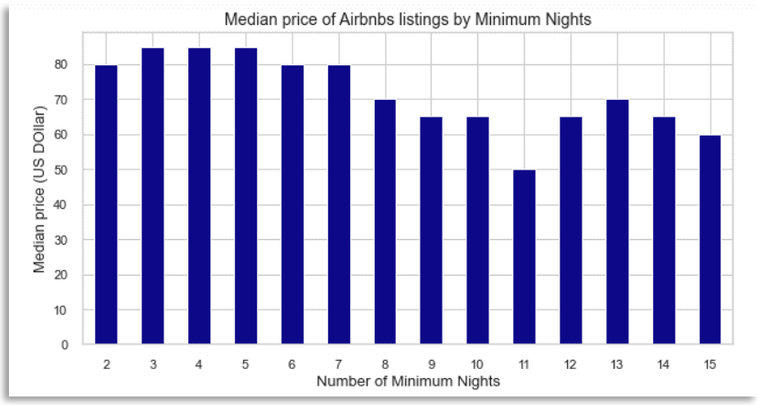


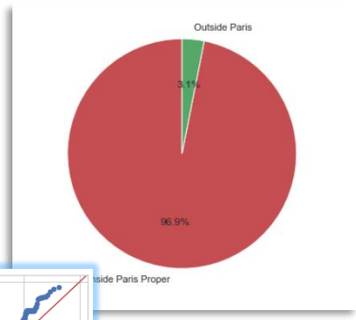
Fig 29 Left: Le Elysee is the highest priced neighbourhood and Popincourt/Montmartre are the lowest.

Fig 30 Right: The highest priced listings are those have set their minimum nights at 3,4 & 5.



Zip Code Reduction – The Borders of Paris Listings

The zipcodes of “Paris Proper” are 75001 to 75021 so I separate out these codes from my groupings to get most of my listings , 55,653 out of 58,184 are in the Paris zipcodes and 1,788 are just on the Paris borders. I separate these into two groups “*Inside Paris”* and *“Outside Paris”.*



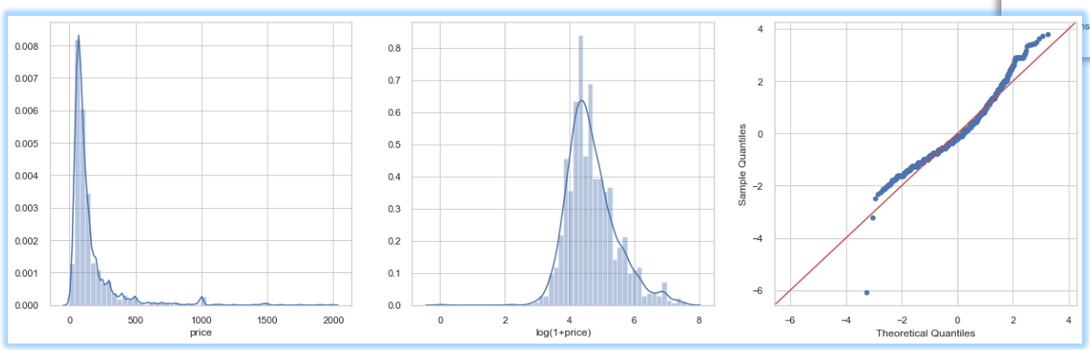


Fig 31 Above: Outside Paris Price Distribution Chart: This has been logged transformed to display a “Standard Di stribution” and the prices are similar to outside Paris, as 75.2% of the listings not having Paris Proper Zipcodes ar e found in the west part of Paris which is a historically higher priced area. The average price for an Apartment is 147 US$ while “INSIDE Paris” it is 110US$. For a shared room “Outside Paris” is 71US$ and Inside Paris it is 58US $.

Fig 32 Below: Passy Neighbourhoods

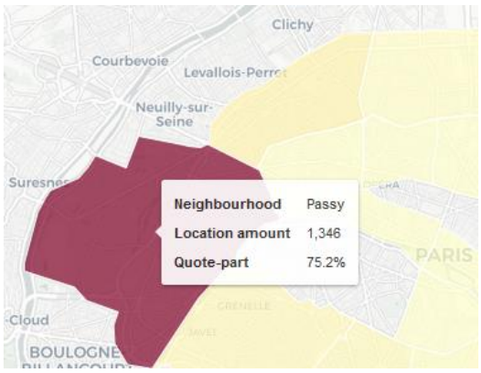


Fig 33 Below: Price by Rental Type

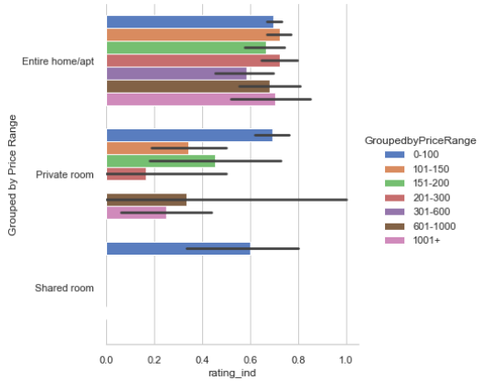
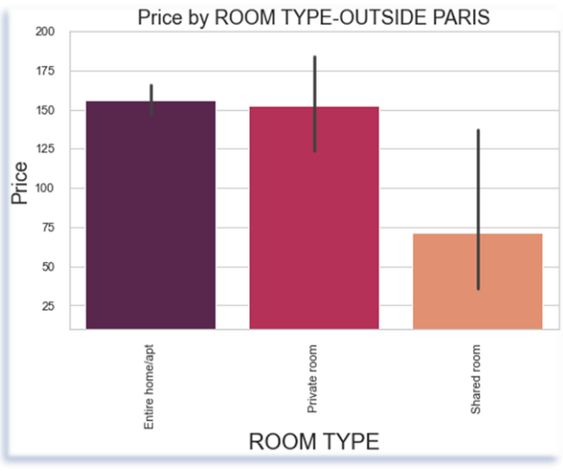
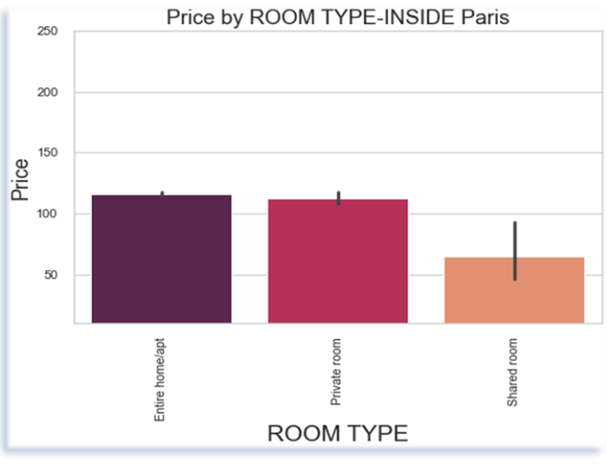


Fig 34 Below: Average price of room type *OUTSIDE PARIS*



Entire home/apt 156.101307
Private room 152.419753
Shared room 71.687500

Fig 35 Below: Average price of room type *INSIDE PARIS*



Entire home/apt 113.395738
Private room 109.489400
Shared room 53.818182

Fig 36 Left We see the average price by Property type for outside Paris.

We can see the "outlier" is the VILLA which is just one property that is priced at over $1100 per night

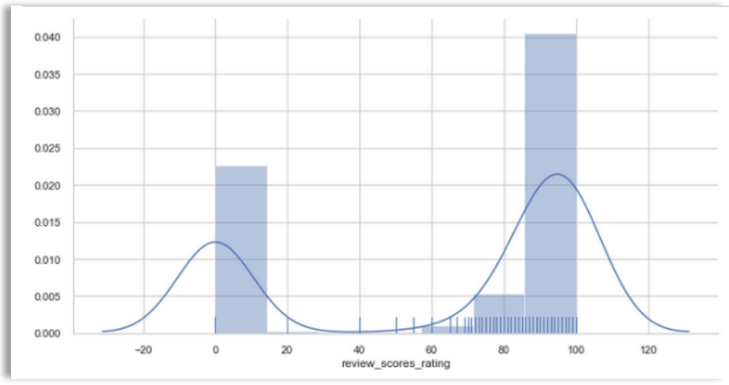


Fig 37 Above: We notice a wider distribution of rating scores as there are more between 60% and 80% which are attributed to the location factor according to the ratings categories.

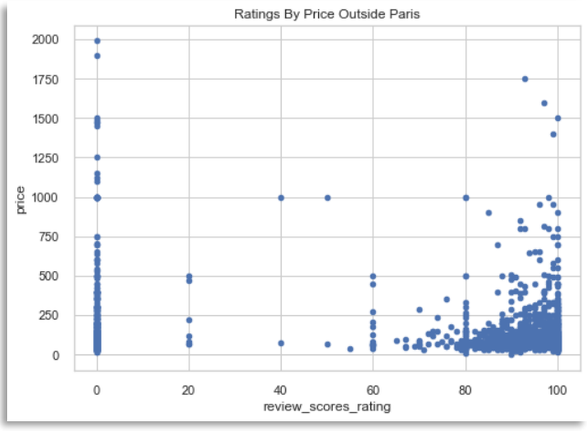


Fig 38 Above: ***Outside Paris***

We notice larger fluctuations in reviews as we are comparing average price to ratings and it looks like the higher priced properties had lower reviews.

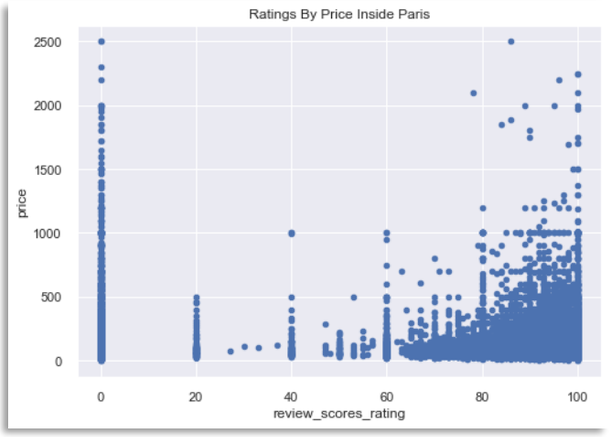


Fig 39 Above ***Inside Paris***

We notice higher concentrations of reviews between 80 and 100%

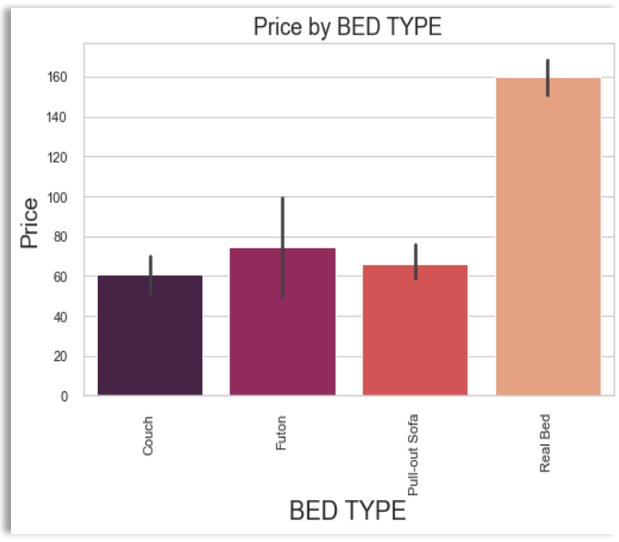


Fig 40 Left the average bed type with the average price; the "real bed" is priced about $30 higher than its equivalent in "Inside Paris". The other 3 types are roughly the same.

Price by Average

Couch	60.86
Futon	74.50
Pull-out Sofa	66.14
Real Bed	159.85

Amenities

So very many amenities ! Which ones matter the most out of the 178+ that are offered? Can having a balcony actually make a difference in price versus having all the day -to- day living items such as a Washing Machine/Dryer, a Kitchen and a Coffee Maker? We resolve these issues by reviewing what the top 20 amenities are, the bottom 20 and those that are the top 20 for the most expensive properties then decisions are made from these results. I had to consider how to eliminate the "Curse of Dimensionality" (essentially too many features to have meaningful outcomes) so I started to reduce & reduce!

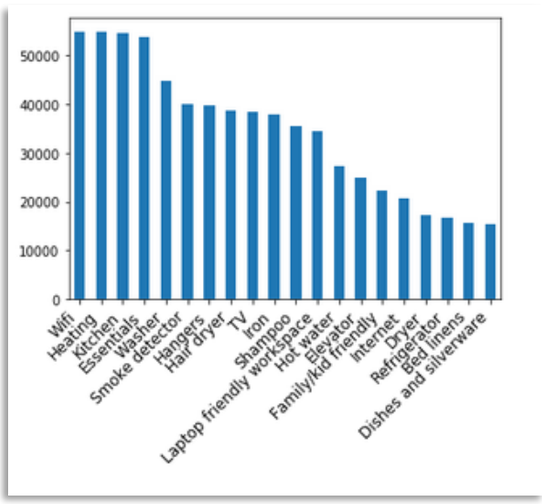


Fig. 41 Above Top 20 Amenities that are the most offered

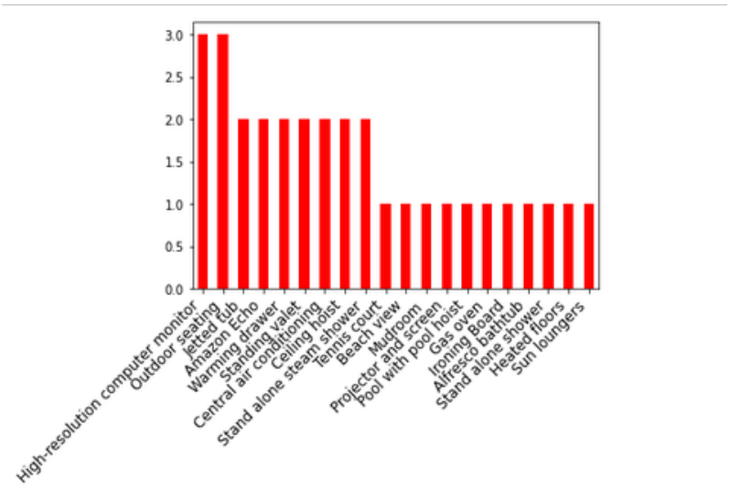


Fig 42 Above 20 Amenities that are least offered

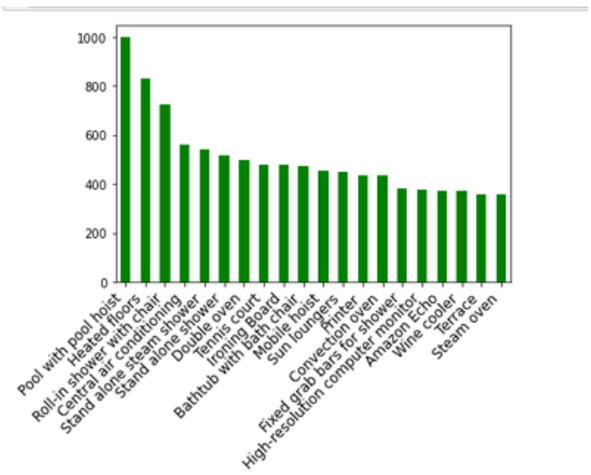


Fig 43 Above Top 20 Amenities for Highest Properties

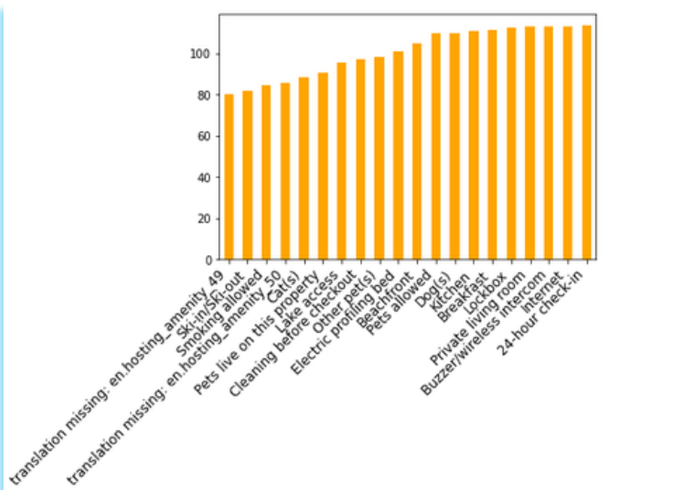


Fig 44 Above: Top 20 Amenities for the lowest Priced Priced properties

Exploratory Data Analysis of the "Calendar" DataSet:

I first sought a general overview of the # of listings that were occupied by the time frame starting with April 9, 2019 to April 8, 2020. Below you can see data analyzed from the Calendar data set.

Index	listing_id	date	available	price	adjusted_price	minimum_nights	maximum_nights
21236880	33338090	2020-04-04	f	$65.00	$65.00	1.0	322.0

This data set had 21,235,880 observations and 7 Data Points

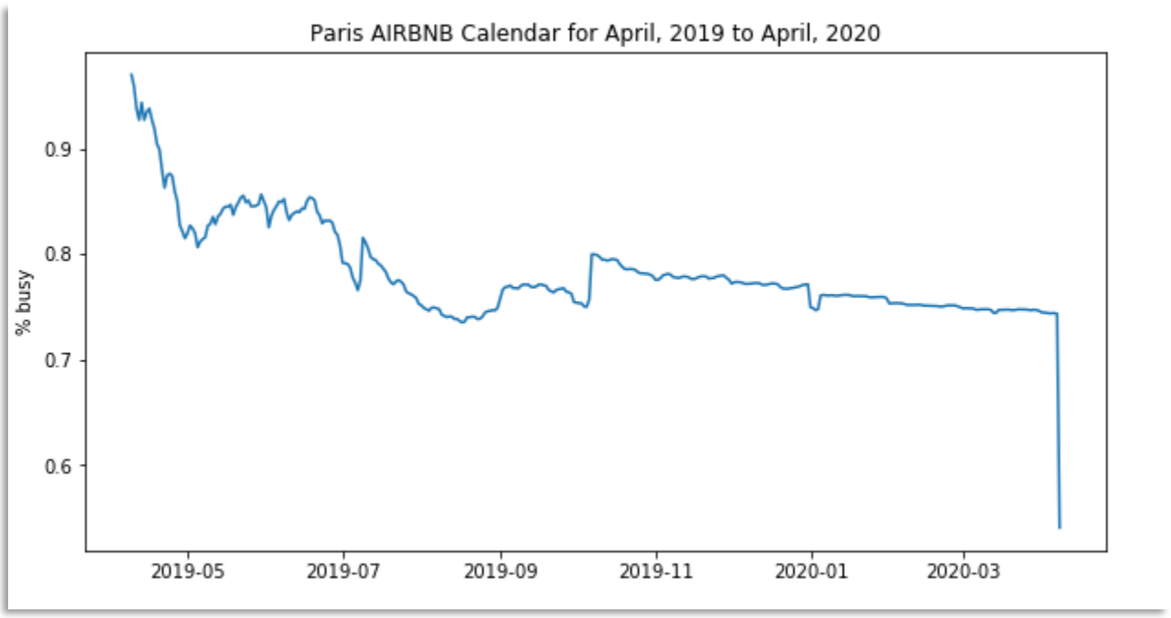


Fig 45: We notice a dip from April of 2019 to May of 2019 then a slight increase in both September of 2019 and November of 2019 then the obvious decline when France closed is borders in mid-March.

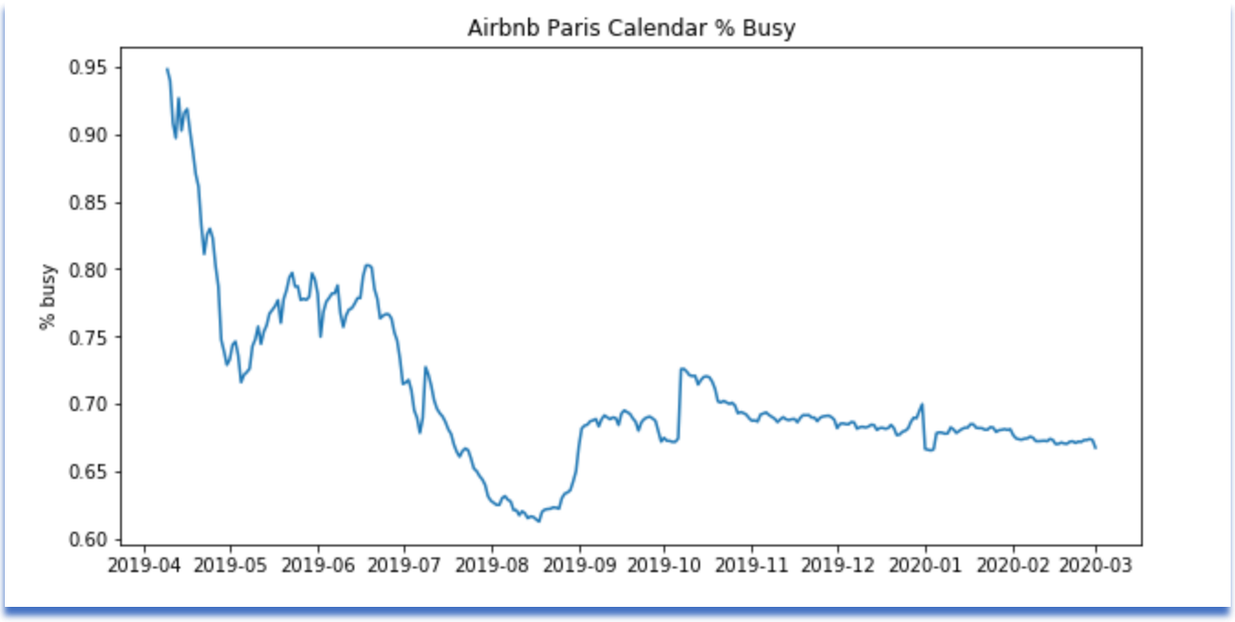
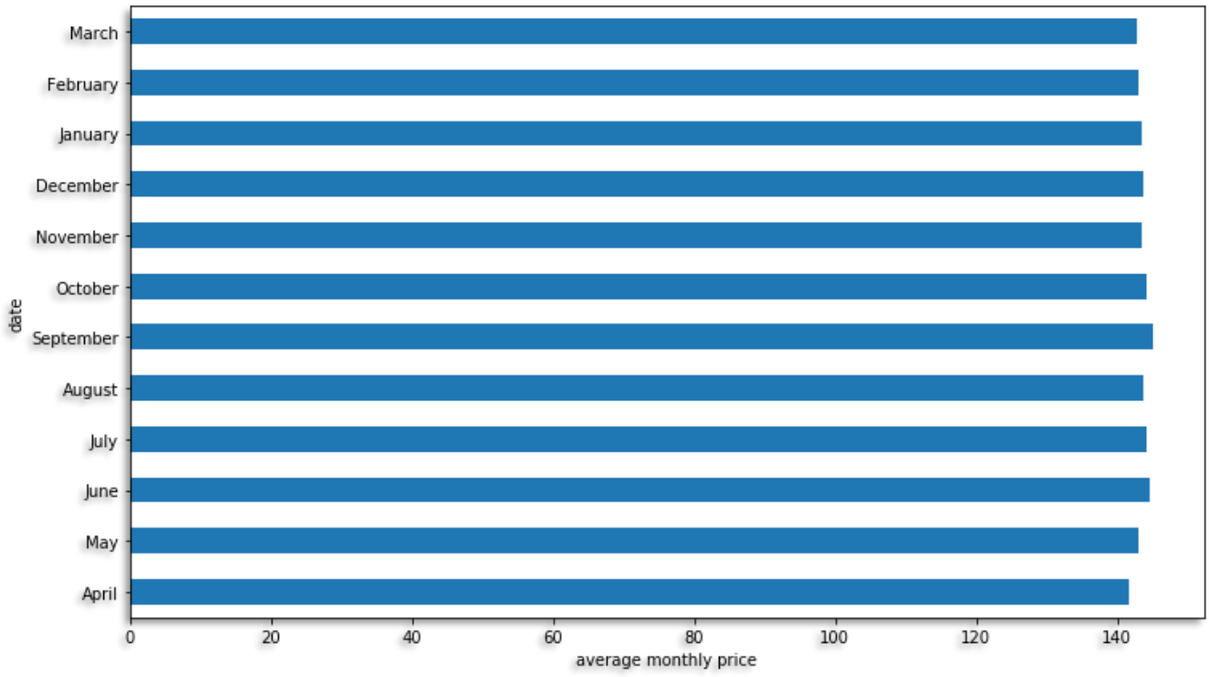


Fig 46: I adjust my dates to: April 9, 2019 to March 1, 2020 to better understand the pre-COVID19 rental scenario we notice a steep decline in July and August as this is when the French population typically take their 8 week vacations.



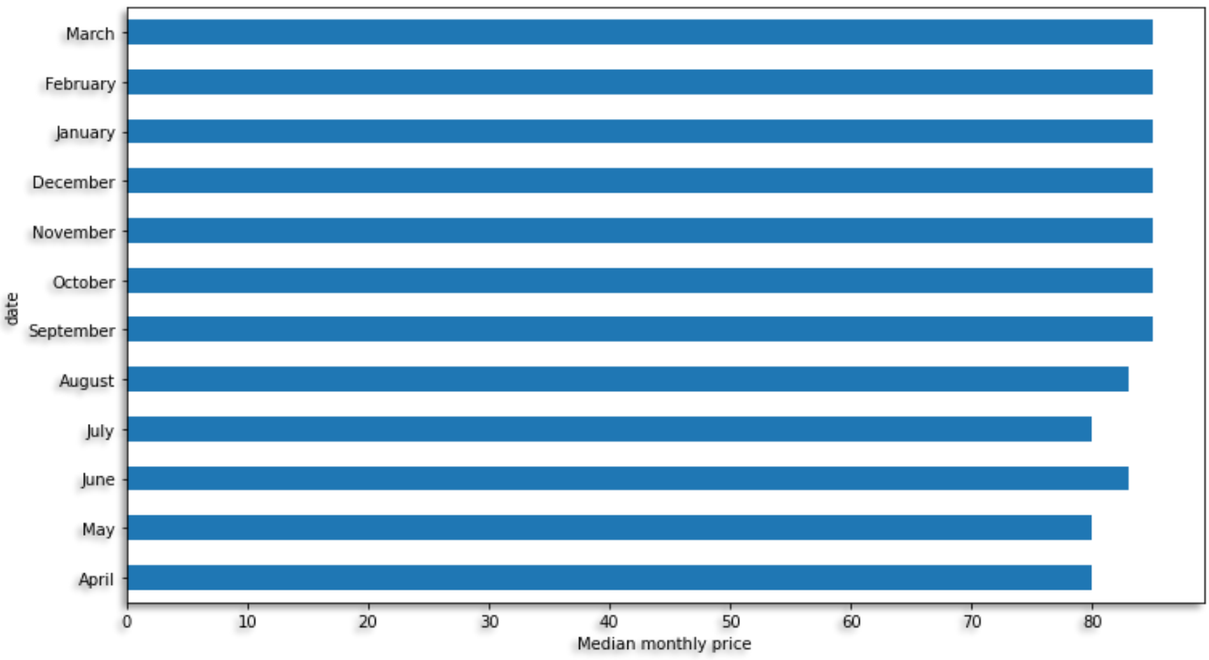
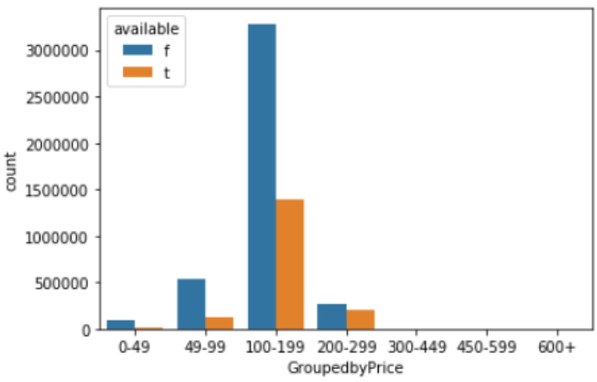


Fig 47a We had an overall average of $119.42 with slight increases in September and slight decreases in April.
Fig 47b: We found the median price to be $85 USD per night with decreases in April ,May, July and August.



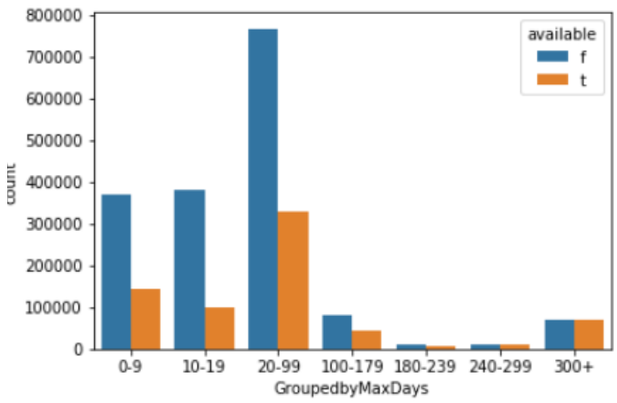
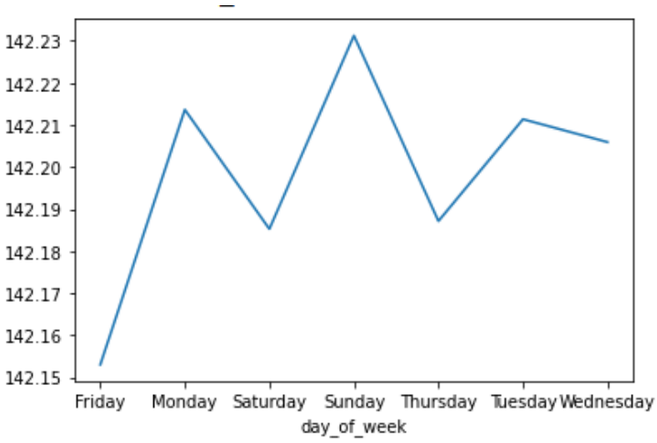
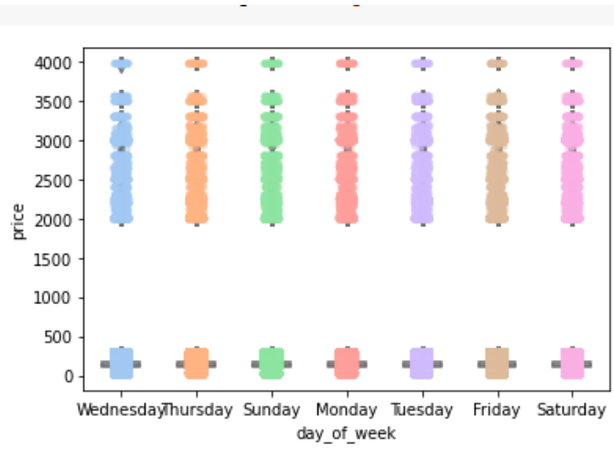


Fig 48a & 48 b Above: We review our grouped prices by availability then the 58,184 properties by their maximum days that each property can be rented





Fig(s)49a & 49b Above: We review our average daily prices and do not notice larger changes on the weekend which is typical of listings.

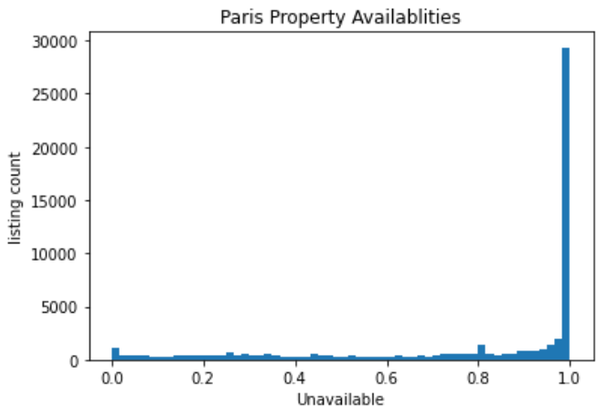
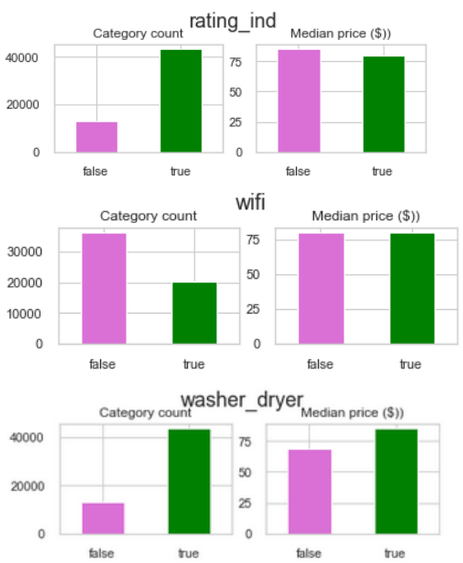


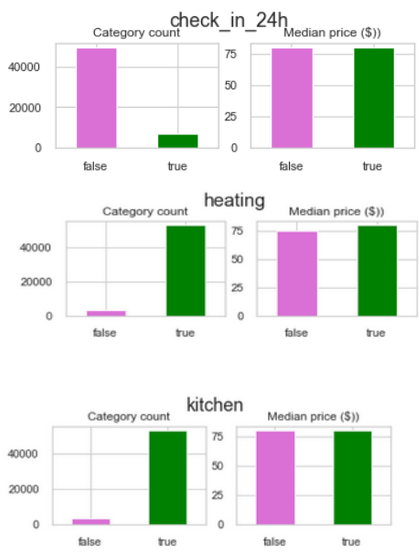
Fig. 50 Left Here we see that the majority of listings (67%) are not available 365 days out of the year.

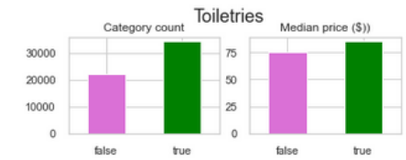
Maximum Nights: The maximum nights the Owner is willing to rent it though Paris has restrictions to 120 days per year if the property is the Owner's primary resident so we can "guess" that is the listing is on the market for more than 120 days it is classified as a share or it is simply "illegal".

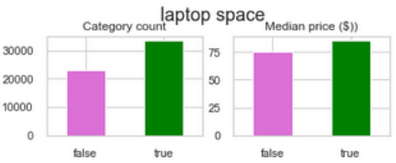
Further Research and Analysis

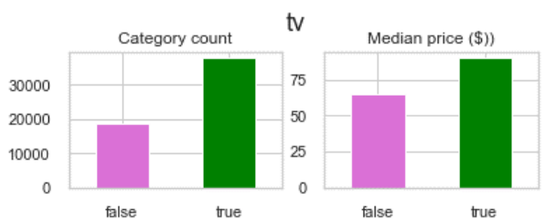
Fig 51 Below: I rate each amenity to verify their relevance and decide which ones are worth retaining in the dataset.

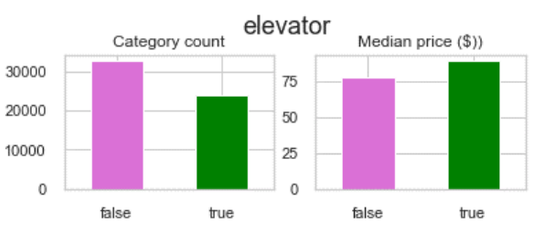


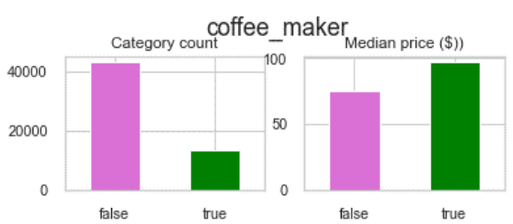


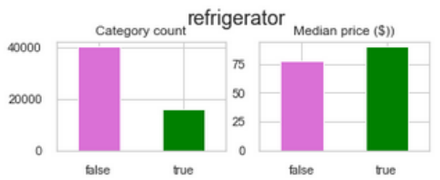


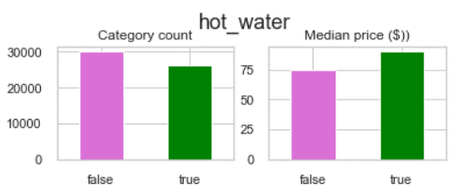


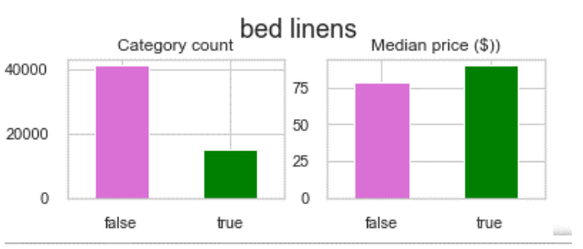












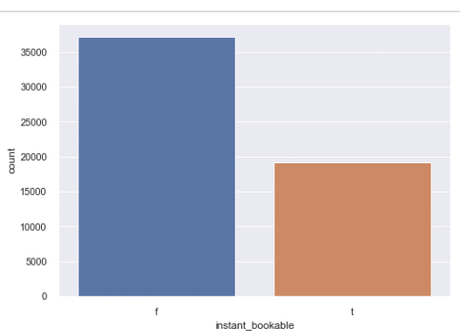


Fig 52 Left

The majority of listings can be booked directly from the Airbnb website without the host's prior approval

AirBnB Explosion – I review 23 data sets from 2015 to 2019 to determine the price increases

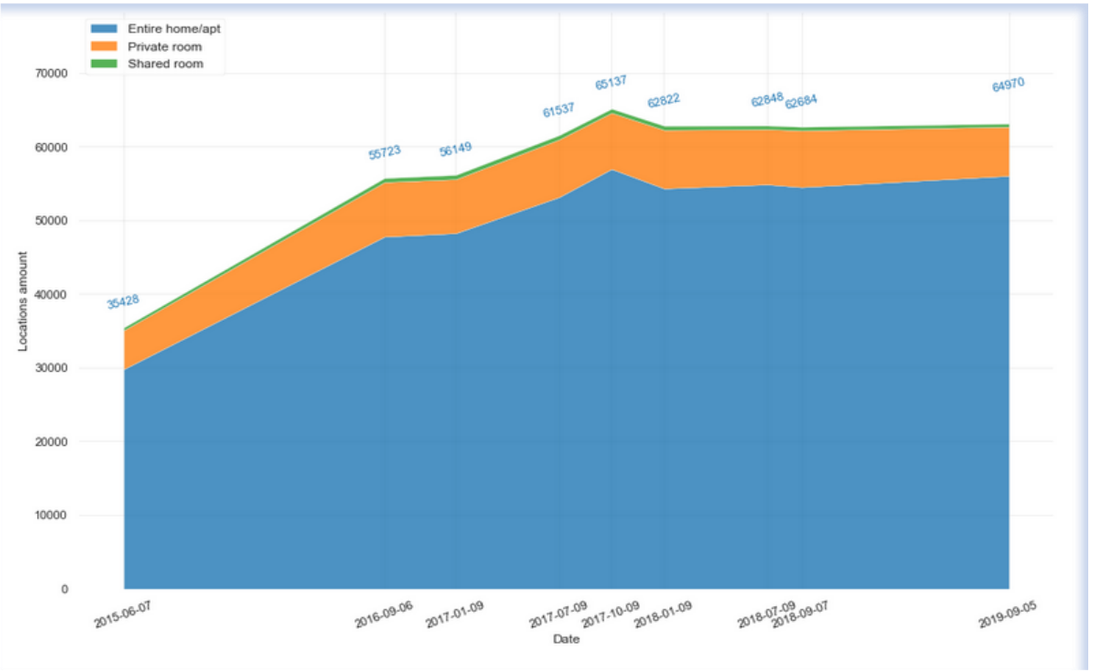


Fig 53 We notice immediately the growth of AIRBNB locations since June 7, 2015 from 35,428 to 66,212 in September 5, 2019. This is a 115% increase in just 4.5 years. From June 2015 to just September 2016 we see an increase from 35,428 to 55,723 which is a 57% increase.

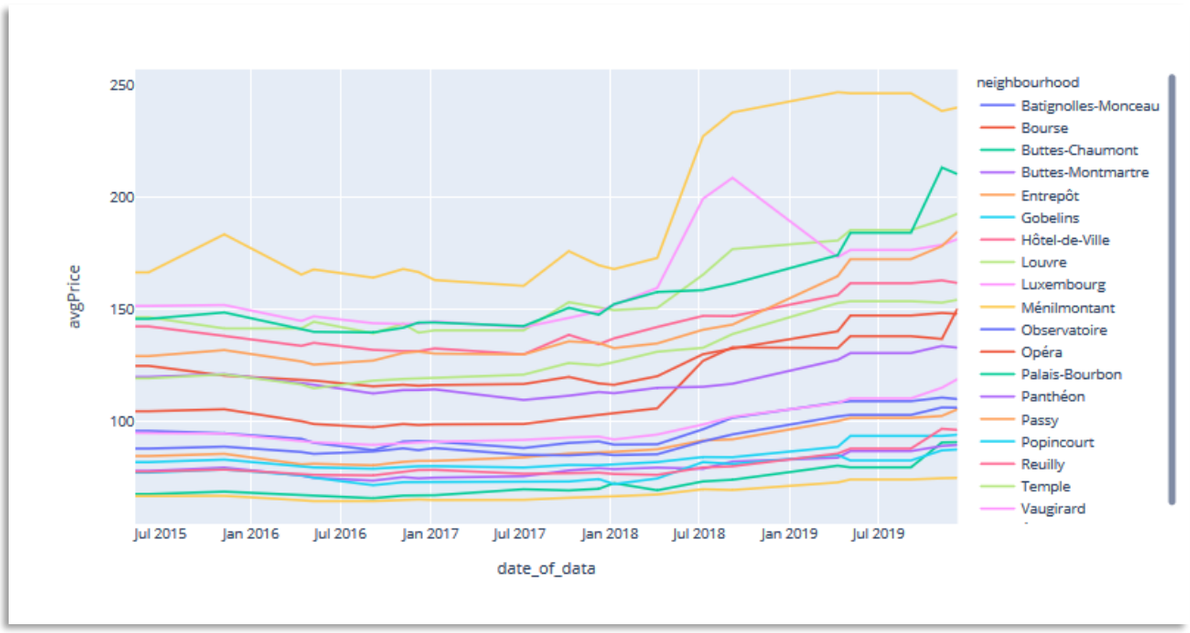
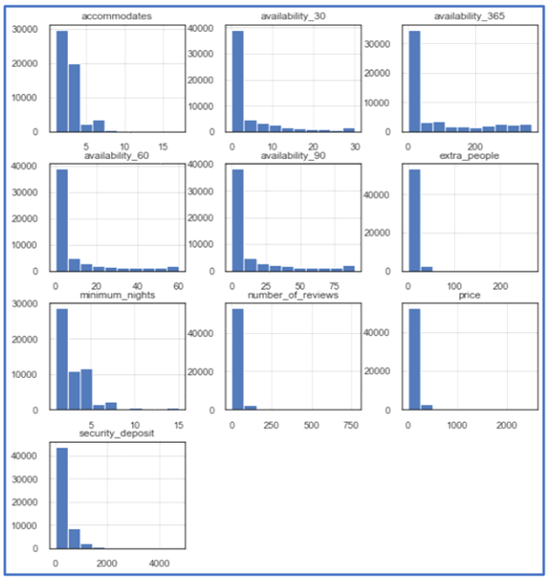
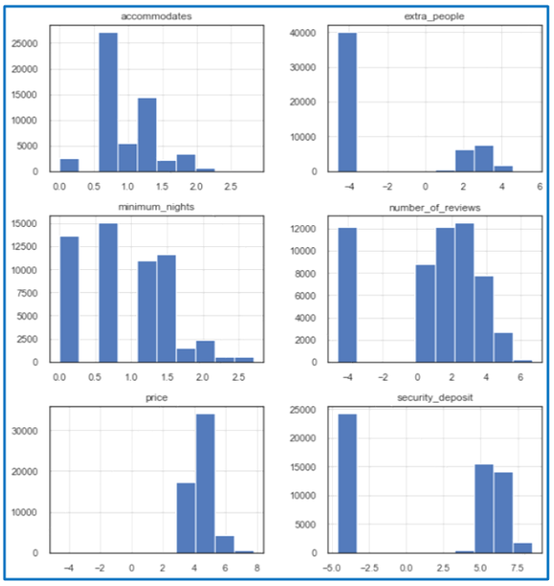


Fig 54 Above: Here we notice that 3 neighborhoods Buttes Chaumont, Opera, Entrepot have had 50%+ increases since 2015.

Fig 55 Correlation Heatmap – See Appendix B

Fig. 56 Below Left = unlogged numeric variables & Fig 57 Right: logged to adjust to a normal distribution





I selected 98 Variable to be processed into my two models. They can be viewed here (line 21).

The Models

I chose *two models* to analyze my variables to determine the key factors that price our Airbnb rentals. I also use SK Learn "Standard Scaler" to scale my data which standardizes a feature by subtracting the mean and then scaling to unit variance. Unit variance means **dividing** all the values by the **standard deviation**

Model 1: Spatial Hedonic Price Model (HPM)

The hedonic model involves regressing observed asking-prices for the listing against those attributes of a property hypothesized to be determinants of the asking-price. It comes from hedonic price theory which assumes that a commodity, such as a house can be viewed as an aggregation of individual components or attributes (Griliches, 1971). Consumers are assumed to purchase goods embodying bundles of attributes that maximize their underlying utility functions (Rosen, 1974).

Standardization:

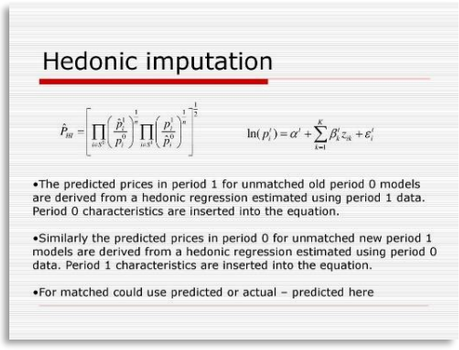
$$z = \frac{x-\mu}{\sigma}$$

with mean:

$$\mu = \frac{1}{N}\sum_{i=1}^{N}(x_i)$$

and standard deviation:

$$\sigma = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(x_i-\mu)^2}$$



In addition to the characteristics of the Airbnb listings, we add location features as they have been shown to be important factors in influencing the price. Ideally, Lagrange multiplier tests should be conducted to verify if there is spatial lag in the dependent variable and therefore a spatial lag model (see this post for spatial regression using Pysal) is preferred for estimating a spatial HPM. We are using a conventional OLS model for hedonic price estimation that includes spatial and locational features, but not a spatial lag that accounts for spatial dependence.

So, the first explanatory variables are the listings characteristics (accommodates, property type, etc) and our second group of explanatory variables based on spatial and locational features are "close to an attraction" or 'not close to an attraction" which indicates, for example, how far an Airbnb is from the Eiffel Tower or the Louvre Museum.

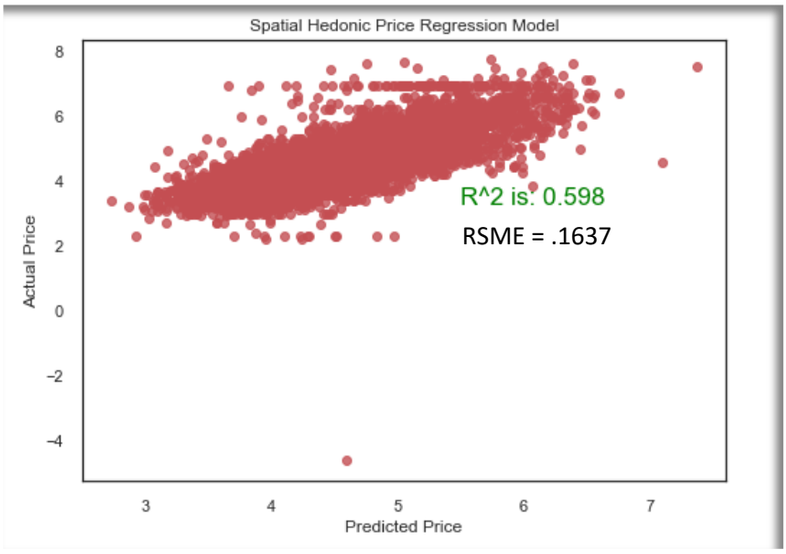


Fig 58 Left

These are my results:

Training RMSE: 0.1677
Validation RMSE: 0.1637

Training r2: 0.5968
Validation r2: 0.598

Note 59 See Appendix D – I run my model with 4 different levels of the 'alpha' parameter

Improving our model

We can try using Ridge Regularization in attempt to reduce the ***less*** important features. Ridge Regularization is a process which ***shrinks the regression coefficients*** of less important features. It essentially penalizes a number of features in a model in order to only keep the most important features. The Ridge Regularization model uses a parameter such as its "alpha" , which will control the strength of the regularization. Alpha measures the value where the regression line crosses the y-axis

We can loop though difference degrees of ALPHA to see how it improves our results. When alpha is 0, the regression produces the same coefficients as a linear regression. When alpha is very large, all coefficients are zero.

The new results are below of my Hedonic Price Model again after removing what I thought of as unimpactful data points :

Review_Score_Accuracy by 0 to 10 and "neighbourhoods" and making 5 different "alpha" adjustments did not improve my model score. The score indicates that my model explains 59% of the variations in price.

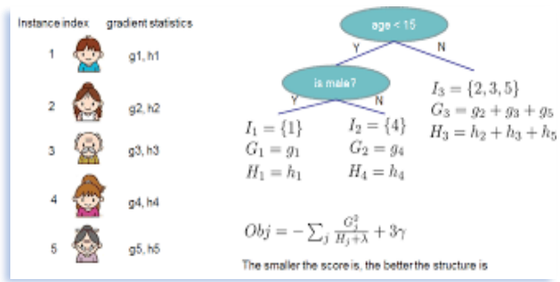
Training RMSE: 0.1688

Validation RMSE: 0.1649

Training r2: 0.5942

Validation r2: 0.5952

Model 2: XG Boost



Apart from its superior performance, a benefit of using ensembles of decision tree methods like gradient boosting is that they can automatically provide estimates of feature importance from a trained predictive model.

Generally, feature importance provides a "score" that indicates how useful or valuable each feature was in the construction of the boosted decision trees within the model. The more an attribute is used to make key decisions with decision trees, the higher its relative importance.

This importance is calculated explicitly for each attribute in the dataset, allowing attributes to be ranked and compared to each other.

Importance is calculated for a single decision tree by the amount that each attribute split point improves the performance measure, weighted by the number of observations the node is responsible for. The performance measure may be the purity (Gini index) used to select the split points or another more specific error function.

The "feature importance" values are then averaged across the decision trees within the model. For more detailed information on how feature importance is calculated in boosted decision trees, see this answer in https://stats.stackexchange.com/questions/162162/relative-variable-importance-for-boosting

THE XG BOOST MODEL

My initial run of my XG Boost model did not produce excellent results as I used the default parameters which outputted these numbers: See Appendix C

I now decide to tune my XG Boost Model to see if it makes a difference in my results.

I choose the parameters below and the numbers according to the approximations that were a result of a CV Grid search performed.

- ✓ n_estimators = Number of trees one wants to build; I choose 200
- ✓ learning_rate= Rate at which our model learns patterns in data. After every round, it shrinks the feature weights to reach the best optimum value . I choose .1
- ✓ max_depth= Determines how deeply each tree is allowed to grow during any boosting round. I choose 6
- ✓ colsample_bytree = Percentage of features used per tree. I choose .7
- ✓ gamma= Specifies the minimum loss reduction required to make a split. I choose .2

Name	Description	Domain
learning rate	Step size shrinkage used in model update	(0.005, 0.2)
subsample	Subsample ratio of the training instances used for fitting the individual tree	(0.8, 1)
max_leaves	Maximum number of nodes to be added	(10, 200)
max_depth	Maximum depth of a tree	(5, 30)
gamma (γ)	Minimum loss reduction required for further partition	(0, 0.02)
colsample_bytree	Subsample ratio of features/columnsused for fitting the individual tree	(0.8, 1)
min_child_weight (w_{mc})	Minimum weights of the instances required in a leaf	(0, 10)

Below is my 2nd run of XG Boost with the above parameters

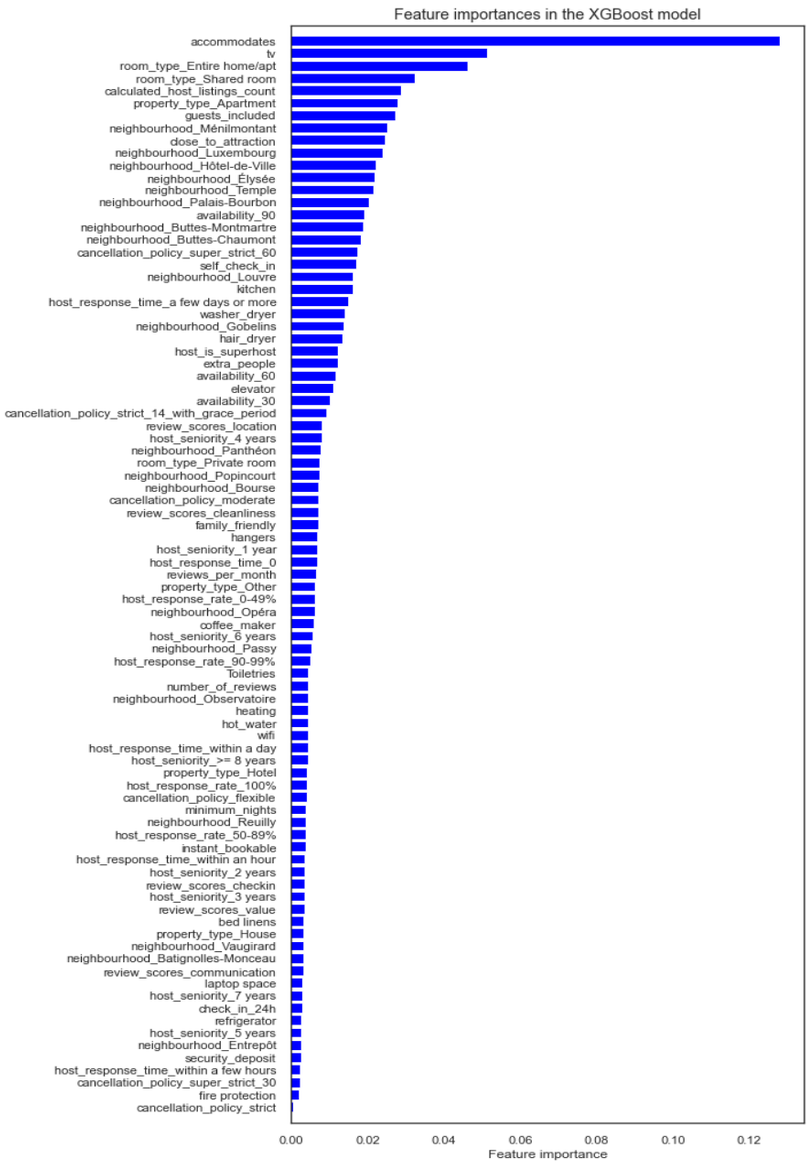


Fig 61 Left Here is my new Feature Importance with review_scores and neighborhood removed in hopes of improving my model:

- ✓ How many people the property accommodates
- ✓ A tv in the rental
- ✓ Room type-Apartment
- ✓ How many other listings the host has (and whether they are a multi-listing host)?
- ✓ Guests included
- ✓ Neighborhood Menilmontant
- ✓ Close to Attraction
- ✓ Neighborhood-Luxembourg
- ✓ Neighbourhood-Hotel De Ville
- ✓ Neighbourhood-Buttes Montmartre

NEW RESULTS:

Training MSE: 0.0962
Validation MSE: 0.1325

Training MSE: 0.0962
Validation MSE: 0.1325

Training r2: 0.7688
Validation r2: 0.6747

Model Interpretation with ELI5 – see Appendix D

SHAP INTERPRETATIONS

SHAP (SHapley Additive exPlanations) by Lundberg and Lee (2016)[41] is a method to explain individual predictions. SHAP is based on the game theoretically optimal Shapley Values. Read more here.

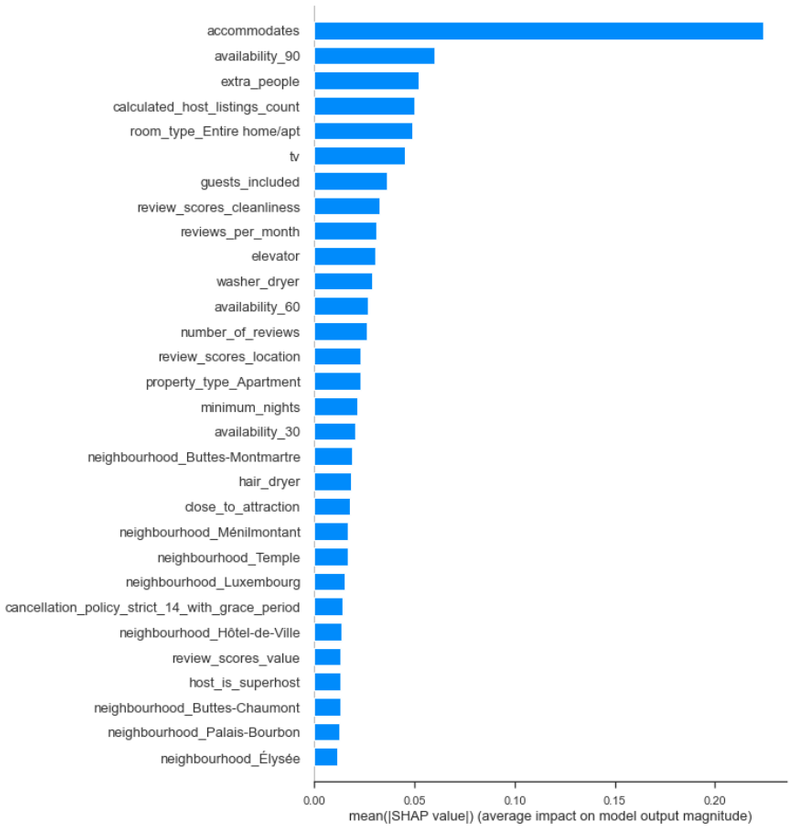


Fig 62 above: The SHAP Feature Importance Results puts the individual neighborhoods toward the bottom of its priorities and elevated "reviews_scores_cleanliness" to the top 10.

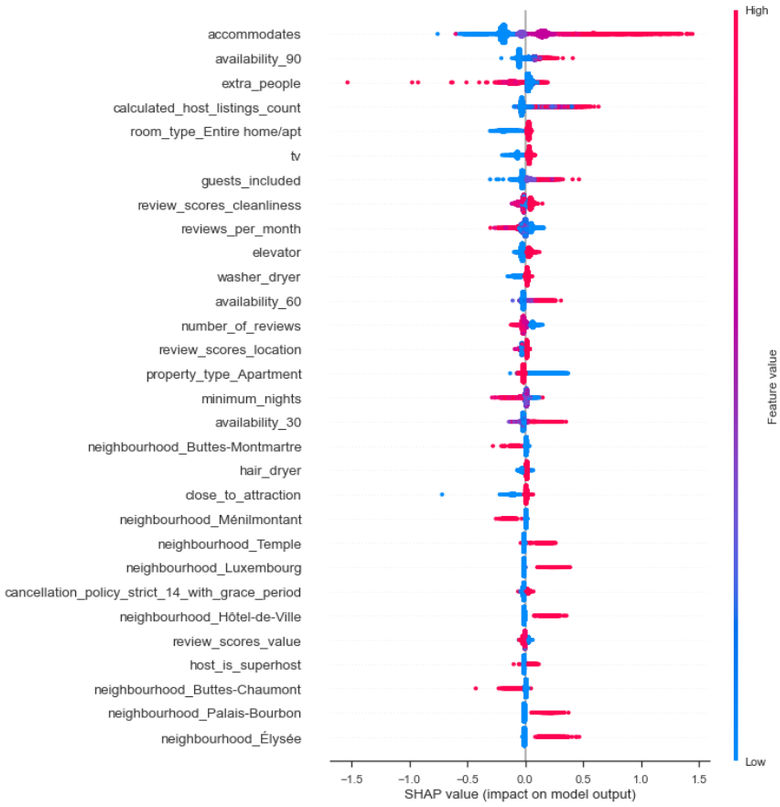


Fig 63 left:

SHAP Dependence Plots While a SHAP summary plot gives a general overview of each feature a SHAP dependence plot shows how the model output varies by feature value. Note that every dot is listing with its features, and the vertical dispersion at a single feature value results from interaction effects in the model. The feature used for coloring is automatically chosen to highlight what might be driving these interactions.

The Shap summary also shows data point distribution and provides visual indicators of how feature values affect predictions. ***Here red indicates higher feature value, blue indicates lower feature value.*** On the x-axis, higher SHAP value to the right corresponds to higher prediction value (more likely listing gets booked), lower SHAP value to the left corresponds to lower prediction value (less likely listing gets booked).

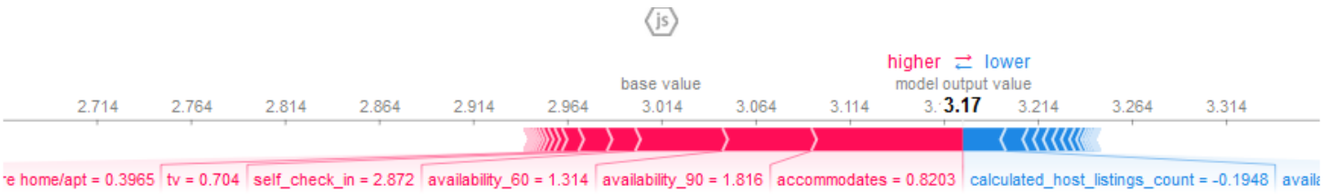


Fig 64 above:

For the 1st example, we can see that there is a base value (bias term) of 3.014, with features in red pushes that value to the right, and features in blue pushes that value to the left, with a combined output of 3.17. Therefore, the effect of the top feature is quantified on the prediction with local accuracy. This particular AirBnb listing has a number of features values (accommodates, self_check_in, and availability_90) that contribute to its outcome. The price of this listing is 156 USD and if we take the log base 5 of 156 we get 3.13 so we are close in our forecast as shown above.

Fig 65 above: SHAP force plot can be used to explain individual predictions. For example, we can see that there is a base value (bias term) of 3.014, with features in red pushes that value to the right, and features in blue pushes that value to the left, with a combined output of 3.17. Therefore, the effect of top feature is quantified on the prediction with local accuracy. The particular listing has a number of features values (accommodates, room_type_Entire home/apt, tv, and availability_90, self_check_in) that contribute to its outcome. The price of this listing is 87 USD and if we take the log base 5 of 87 we get 2.77 so we are close in our forecast as shown above.

Fig 66 above: The price of this listing is 185 USD and if we take the log base 5 of 185 we get 3.24 so we are close in our forecast as shown above.

Null Hypothesis:

My N0: ***"AirBNB Paris locations near the Top 10 attractions have little impact in the price of the listing"***

The **Chi-square test** is intended to **test** how likely it is that an observed distribution is due to chance. It is also called a "goodness of fit" statistic, because it measures how well the observed distribution of data fits with the distribution that is expected if the variables are independent.

Below I performed my **Chi Squared Test** to test my Null Hypothesis above.

We see from the results below the "p-value" is less than our significance of .05 so we reject the HO and go with the HA which is "It ***makes*** a difference in the price of each AirBnB Rental if it is located within 2 miles of the *Top 10 Paris City attractions.*

The contingency_tables below are comparing rentals near the TOP 10 List of Paris Attractions and those NOT near the TOP 10 List of Paris Attractions

Observed Values

```
[[ 2  5]
 [ 0  2]
 [ 2 12]
 ...
 [ 0  2]
 [ 0  1]
 [ 0  3]]
```

Expected Values

```
[[ 0.55318394  6.44681606]
 [ 0.15805255  1.84194745]
 [ 1.10636788 12.89363212]
 ...
 [ 0.15805255  1.84194745]
 [ 0.07902628  0.92097372]
 [ 0.23707883  2.76292117]]
```

Degree of Freedom:- 1
chi-square statistic:- 1960.8378847696392
critical_value: 3.841458820694124
p-value: 0.0
Significance level: 0.05
Degree of Freedom: 1
chi-square statistic: 1960.8378847696392
critical_value: 3.841458820694124
p-value: 0.0
Reject H0,There is a relationship between 2 categorical variables
Reject H0,There is a relationship between 2 categorical variables

I see there is a relationship as the p-value is below my .05 significance level so I reject my H0 and accept my HA which is "It ***makes a difference*** in the price of each AirBnB Rental if it is located within 2 miles of the Top 10 Paris City attractions.."

Text Analysis

Here I use NLK which is a leading platform for building Python programs to work with human language data. It provides easy-to-use interfaces to over 50 corpora and lexical resources such as WordNet, along with a suite of text processing libraries for classification, tokenization, stemming, tagging, parsing, and semantic reasoning, wrappers for industrial-strength NLP libraries, and an active discussion forum. Source: https://www.nltk.org/

The first WordCloud I create is derived from taking the top 10% of review scores then the bottom 10% according to "Ratings_By_Range". I achieve this by creating a function to apply NTLK to the property descriptions in efforts to improve the way a listing is written and gain a little marketing insight.

TOP 10% of Reviews

Fig 67 below: The 1st WordCloud is to review the most word usages for the *top 10% of ratings* of our list.

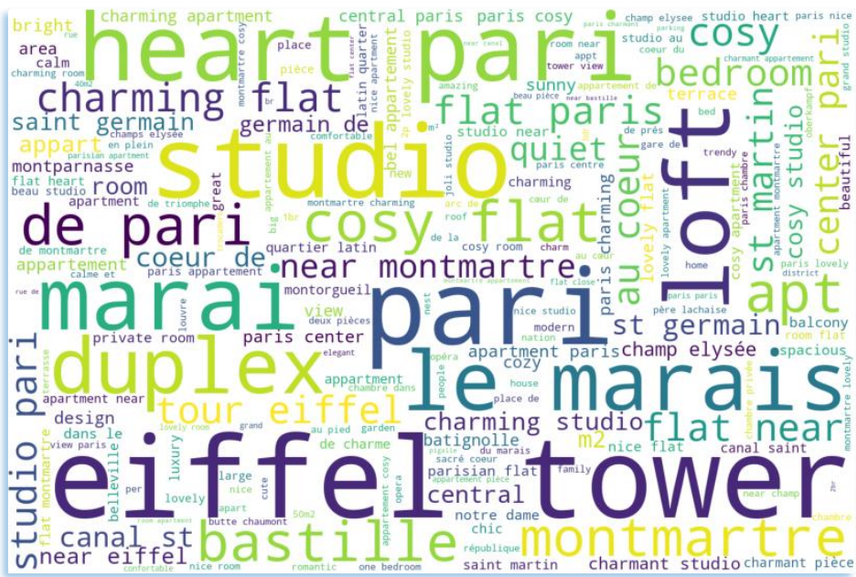
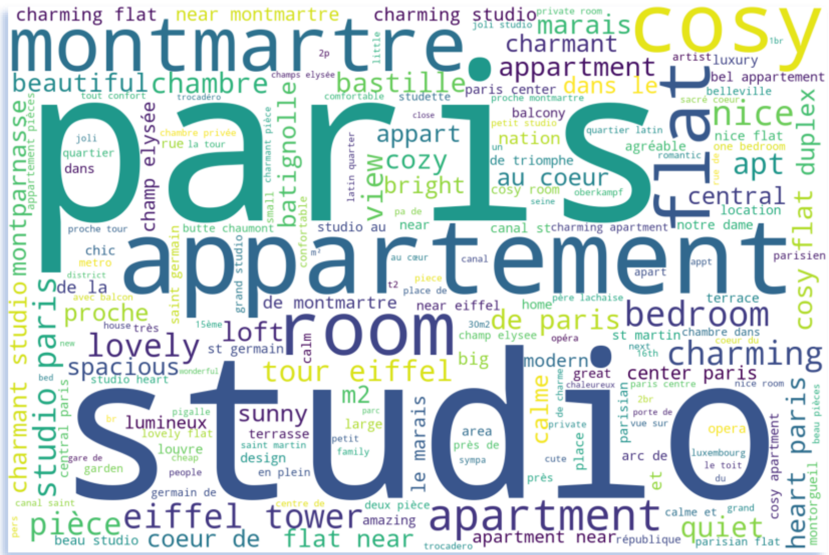


Fig 68 below: A WordCloud of listings in the *bottom 10%* according to “Ratings_By_Range”.



We notice that the Eiffel Tower, Heart of Paris, is used more often in the top 10% of listings as this can be interpreted that location is the KEY contributing factor to these listings. The bottom 10% emphasizes the “style” of the facilities such as it is a “STUDIO, and APARTMENT, or a BEDROOM and placed less emphasis on the location.

Conclusions:

Through this exploratory data analysis and visualization project, I gained several interesting insights into the Airbnb rental market. Below I will summarize the answers to the questions that I sought answers at the beginning of the project:

How do prices of listings vary by location? What localities in Paris are rated highly by guests? The Elysee has the most expensive rentals compared to the other “arrondissements”. Prices are higher for rentals closer to the Top 10 City Attractions. Rentals that are rated higher like the Elysee and Opera arrondissements also have higher prices. There are a few outliers in that the OUTSIDE Paris has some high priced and larger rentals.

How does the demand for Airbnb rentals fluctuate across the year and over years? The demand (assuming that it can be inferred from the number of reviews) shows an increasing pattern - demand increases from April to May, then drops slightly in July and August. In general, the demand for Airbnb listings has been steadily increasing over the years.

Are the demand and prices of the rentals correlated? Average prices of the rentals increase across the year, which correlates with demand. However, the prices show a slight increase in September and October as opposed to July and August, which is counterintuitive since these are prime vacation months in many countries. Prices are only very slightly higher on average on Fridays and Saturdays, compared to the other days of the week.

What are the different types of properties in Paris? Do they vary by neighborhood? There are more than 29 different types of listings in Paris. The ratio of the type of listings to total numbers varies by arrondisement. The Elysee and Outside Paris tend to have property types that are larger and can accommodate a higher number of people.

What makes a host a Super host? Ratings and Response rates tend to have a direct correlation with a host being ‘promoted’ to the status of the Super host. However, there are other factors too that makes someone a super host as the not all hosts with high ratings and response rates were superb hosts.

Do regular hosts and super hosts have different cancellation and booking policies? Both have similar cancellation and booking policies.

Are there any common themes that can be identified from the free-text section of the reviews? What aspects of the rental experience do people like and what aspects do they abhor? There are certain words such as words such as “quiet”, “walkable”, “clean”, “spotless” that are associated with the word “comfortable” demonstrating the importance of environment, location and cleanliness. Words associated with “uncomfortable” include “cramped”, “crowded”,“small”, “stuffy” “cluttered” which indicate that lack of space is one of the most common complaints.

Limitations

- There was not a clear way to confirm the booking of each airbnb other than through reviews and what was in my calendar dataset as this contained only bookings and not confirmed “stays”. Hence, there was an assumption made, particularly in the demand and supply section of the report to understand the booking trends. We assume that the best rates are when the rental availability is low which is July and August. We also assume trends will follow what our figure #54 tells us Buttes Chaumont, Opera, Entrepot have increased over 50% since 2015 and will continue to do so the recent pandemic may decrease or stagnate rates for the next 3 or 4 years according to economists.
- There was random sampling done while performing the user review analysis due to memory limitations. We assume that our random sample is representative of the whole population.
- There were certain features such as acceptance_rate, monthly_price and descriptions that either contained missing values or values in the free-text format that was not easy to work on and hence were dropped from our analysis. We need to use better quality/more accurate data which includes the actual average prices paid per night.
- There was a sufficient supply of rentals in the $100 price range throughout the year

Besides gaining interesting insights into the Airbnb rental market in Paris, we acquired several technical and soft skills along the way. Dealing with multiple data formats helped us strengthen our skills in data manipulation and cleaning. We learned how to work on different Python frameworks and libraries, particularly Folium to create interactive visualization and XG Boost to better understand its features. In discussing my project with other French businesspeople, I was provided insight into certain aspects of my project as well as to the lack of verification for AIRBNB hosts and the like. We also learnt how to effectively deploy Github and other version control systems while working on this project. Finally getting to know SHAP interpretations was a “real treat” as it is able to simply reflect the outputs of tree models in an elegant way.

Future Analysis

I want to expand our analysis to multiple European cities and compare patterns and trends amongst these cities. From the insights we have derived, we would also like to build predictive models using different features from the dataset. Lastly, we hope to implement the visualizations and techniques used in this project to many other fields and datasets.

APPENDICES

Appendix A

This data set has 58184 x 365 observations (prices from each calendar day x listings) which equate to 21,236,885 total values and 7 Variables (data points)

This will tell us how often and when each property is available to rent and its price (base price) and adjusted price(each listing manager can adjust their price according to supply and demand). We will also know how many days and nights the listing manager can rent out the space per year.
Please note that any property not occupied by the owner in Paris France cannot be legally rented out more than 120 days per calendar year as I have mentioned earlier.

Index	listing_id	date	available	price	adjusted price	minimum_nights	maximum_nights
21236880	33338090	2020-04-04	f	$65.00	$65.00	1.0	322.0

Description of each Variable:

- Listing_id: This is the number that is connected to each individual listing (discrete)
- Date: This indicates the date of the active listing time (Datetime)
- Available: This indicates if the property is available on the given date (categorical , Boolean t/f data)
- Price: Price per night for number of included guests (continuous data)
- Adjusted Price: price that can be adjusted as related to supply and demand (continuous)
- Minimum Nights: Listing with high value of minimum nights are likely sublettings (discrete value)
- Maximum Nights: Most of the values are above 30 days suggesting it's used as an open bracket (discrete value)

The Summary Listings have *58184 observations and 16 Variables.*

Variables with Type and Description	
id- discrete	room_type - categorical
name - name of host	price - continuous
host_id - discrete	minimum_nights - discrete
host_name - discrete	number_of_reviews - continuous
neighbourhood_group -name of neighborhoods	last_review - datetime
neighbourhood - name of neighborhood	reviews_per_month - continuous
latitude - continuous	calculated_host_listings_count - # of properties each host manages; continuous
longitude - continuous	availability_365 - if the property is available 365 days of the year - categorical

The Detailed Listing details have 58184 observations and 106 Variables.

For details of these 106 Variables please review my code here.

I merge the **two data sets** to begin my analysis **of 58184 observations** and the following 44 Variables:
(I will further describe each Variable during my analysis) Review Variables here.

Appendix B

I review my variables and determine which ones need to be removed and I decide on my "review_score accuracy" Variable to reduce noise in my data.


instant_bookable
host_is_superhost
guests_included
extra_people
accommodates
review_scores_cleanliness
review_scores_checkin
review_scores_communication
review_scores_location
review_scores_value
price
minimum_nights
number_of_reviews
reviews_per_month
calculated_host_listings_count
availability_30
availability_60
availability_90
availability_365
check_in_24h
wifi
heating
washer_dryer
kitchen
Toiletries
hangers
laptop space
family_friendly
fire protection
refrigerator
elevator
bed linens
coffee_maker
hot_water
tv
self_check_in
flair_dryer
dryer
security_deposit
close_to_attraction
host_seniority_1 year
host_seniority_2 years
host_seniority_3 years
host_seniority_4 years
host_seniority_5 years
host_seniority_6 years
host_seniority_7 years
host_seniority_>= 8 years
host_response_time_0
host_response_time_a few days or more
host_response_time_within a day
host_response_time_within a few hours
host_response_time_within an hour
host_response_rate_0-49%
host_response_rate_100%
host_response_rate_50-89%
host_response_rate_90-99%
host_response_rate_unknown
neighbourhood_Batignolles-Monceau
neighbourhood_Bourse
neighbourhood_Buttes-Chaumont
neighbourhood_Buttes-Montmartre
neighbourhood_Entrepôt
neighbourhood_Gobelins
neighbourhood_Hôtel-de-Ville
neighbourhood_Louvre
neighbourhood_Luxembourg
neighbourhood_Ménilmontant
neighbourhood_Observatoire
neighbourhood_Opéra
neighbourhood_Palais-Bourbon
neighbourhood_Panthéon
neighbourhood_Passy
neighbourhood_Popincourt
neighbourhood_Reuilly
neighbourhood_Temple
neighbourhood_Vaugirard
neighbourhood_Elysée
room_type_Entire home/apt
room_type_Private room
room_type_Shared room
property_type_Apartment
property_type_Hotel
property_type_House
property_type_Other
cancellation_policy_flexible
cancellation_policy_moderate
cancellation_policy_strict
cancellation_policy_strict_14_with_grace_period
cancellation_policy_super_strict_30
cancellation_policy_super_strict_60
review_scores_accuracy_0.0
review_scores_accuracy_2.0
review_scores_accuracy_3.0
review_scores_accuracy_4.0
review_scores_accuracy_5.0
review_scores_accuracy_6.0
review_scores_accuracy_7.0
review_scores_accuracy_8.0
review_scores_accuracy_9.0
review_scores_accuracy_10.0
0.75
0.50
0.25
0.00
−0.25
−0.50
−0.75

Appendix C

Below I review the contributing factors to price along with our related scores; the coefficient report does not show many significant points as most of the values do not have a P-value less than .05 other than the two highlighted in yellow; accommodates and minimum nights.

Dep. Variable:	price	**R-squared (uncentered):**	0.016
Model:	OLS	**Adj. R-squared (uncentered):**	0.011
Method:	Least Squares	**F-statistic:**	3.357
Date:	Tue, 30 Jun 2020	**Prob (F-statistic):**	5.51e-22
Time:	08:27:53	**Log-Likelihood:**	-49350.
No. Observations:	16909	**AIC:**	9.886e+04
Df Residuals:	16829	**BIC:**	9.948e+04
Df Model:	80		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
instant_bookable	0.0365	0.039	0.940	0.347	-0.040	0.113
host_is_superhost	0.0419	0.038	1.096	0.273	-0.033	0.117
guests_included	0.0333	0.042	0.787	0.431	-0.050	0.116
extra_people	-0.0518	0.040	-1.305	0.192	-0.130	0.026
accommodates	0.2851	0.044	6.535	0.000	0.200	0.371
review_scores_cleanliness	-0.0387	0.175	-0.222	0.825	-0.381	0.303
review_scores_checkin	0.1696	0.286	0.594	0.553	-0.390	0.729
review_scores_communication	-0.1960	0.294	-0.666	0.506	-0.773	0.381
review_scores_location	0.1430	0.232	0.617	0.538	-0.312	0.597
review_scores_value	0.0396	0.235	0.168	0.866	-0.421	0.500
minimum_nights	-0.0823	0.038	-2.147	0.032	-0.157	-0.007
number_of_reviews	-0.1768	0.102	-1.729	0.084	-0.377	0.024
reviews_per_month	-0.0430	0.049	-0.876	0.381	-0.139	0.053
calculated_host_listings_count	0.0008	0.042	0.018	0.985	-0.081	0.082
availability_30	-0.0065	0.113	-0.058	0.954	-0.228	0.214
availability_60	0.3294	0.224	1.469	0.142	-0.110	0.769
availability_90	-0.1860	0.170	-1.094	0.274	-0.519	0.147
check_in_24h	-0.0580	0.039	-1.487	0.137	-0.134	0.018
wifi	0.0232	0.042	0.558	0.577	-0.058	0.105
heating	-0.0149	0.036	-0.410	0.682	-0.086	0.056
washer_dryer	0.0689	0.038	1.790	0.073	-0.007	0.144

Appendix D

ELI5 is a Python package which helps to debug machine learning classifiers and explain their predictions in an easy to understand an intuitive way. It is perhaps the easiest of the three machine learning frameworks to get started with since it involves minimal reading of documentation! Let's look at some intuitive ways of model interpretation with ELI5 on our classification model.

Typically for tree-based models ELI5 does nothing special but uses the out-of-the-box feature importance computation methods which we discussed in the previous section. By default, **'gain'** is used, that is the average gain of the feature when it is used in trees.

ELI5 does this by showing weights for each feature depicting how influential it might have been in contributing to the final prediction decision across all trees. The idea for weight calculation is described **here**; ELI5 provides an independent implementation of this algorithm for *XGBoos*t and most scikit-learn tree ensembles which is definitely on the path towards model-agnostic interpretation but not purely model-agnostic like LIME.

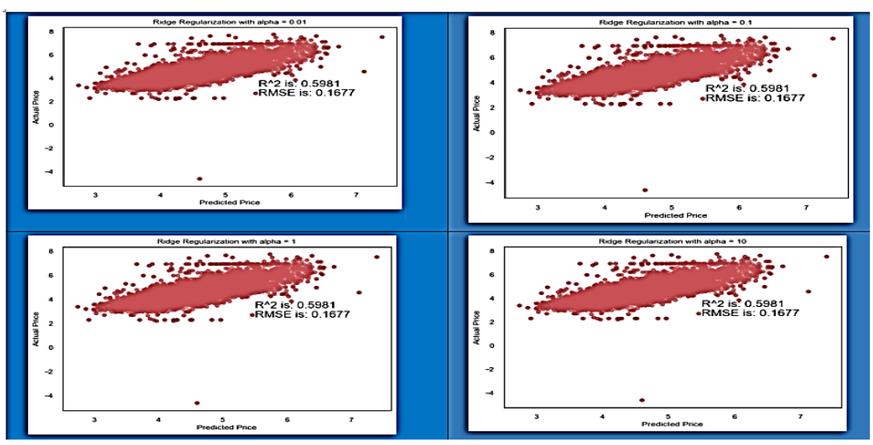
Typically, the prediction can be defined as the sum of the feature contributions + the "bias" (i.e. the mean given by the topmost region that covers the entire training set)

I used the EL15 to produce a Feature Weight list below:

Weight	Feature
0.1278	accommodates
0.0512	tv
0.0463	room_type_Entire home/apt
0.0323	room_type_Shared room
0.0288	calculated_host_listings_count
0.0279	property_type_Apartment
0.0272	guests_included
0.0250	neighbourhood_Ménilmontant
0.0244	close_to_attraction
0.0240	neighbourhood_Luxembourg
0.0220	neighbourhood_Hôtel-de-Ville
0.0219	neighbourhood_Élysée
0.0214	neighbourhood_Temple
0.0202	neighbourhood_Palais-Bourbon
0.0192	availability_90
0.0189	neighbourhood_Buttes-Montmartre
0.0182	neighbourhood_Buttes-Chaumont
0.0173	cancellation_policy_super_strict_60
0.0171	self_check_in
0.0162	neighbourhood_Louvre

Appendix E

I ran my model again with revised "alpha" values using 5 degrees of Alpha: [.01,.1,1,10,10], but did not receive better results.



Appendix F: My initial run of XG BOOST did not produce the results desired either with are between .70 and 100% of variations explained :

Training MSE: 0.1538
Validation MSE: 0.1544

Training r2: 0.6302
Validation r2: 0.621

The score indicates that my model explains only 62% of the variations in price and the features weighted as below.