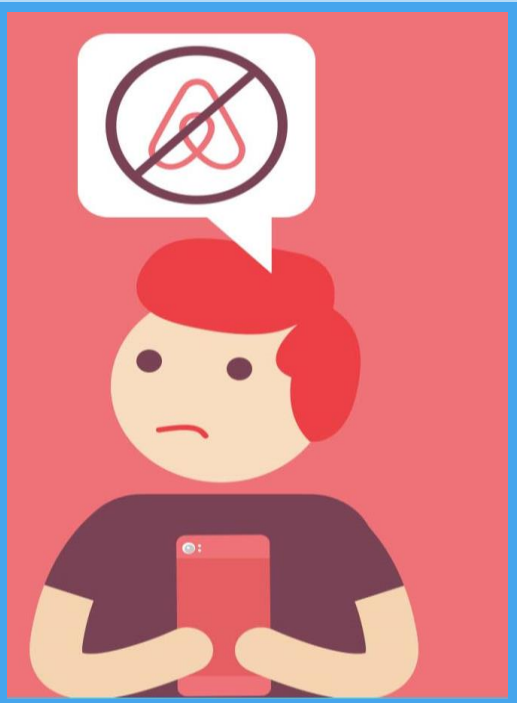
PREDICTING PRICING FOR AIRBNB PARIS, FRANCE

Carolyn Massa
Carolyn@revinformatics.com

June.2020
CODE FOUND HERE

PROBLEM STATEMENT :

- 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?



METHODOLOGY:

1. Preprocessing and Cleaning using Python
2. Feature Extraction
3. NLP of listing descriptions for top 10% and bottom 10%
4. Analysis of various supervised learning methods such as XG Boost and Hedonic Spatial Model
5. SHAP Interpretation
6. Conclusions/Future Analysis

EXPLORE AND ANSWER THE FOLLOWING QUESTIONS:

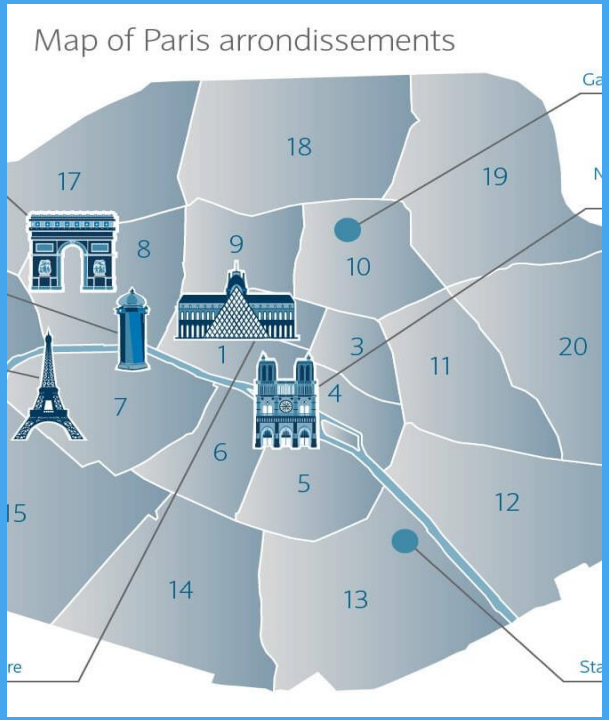
- How does each location influence the property rental price?
- Do reviews affect the pricing?
- Can a traveler stay close to the main attractions such as the Eiffel Tower and the most popular museums such as the Louvre and Musee D'Orsay?
- What other features drive the price of an AirBnB rental property?

- i.e. price vs location, price vs distance from a top attraction, price vs. access to transportation and perhaps convention centers as there are many conventions in September

WHICH FEATURES ARE IMPORTANT TO TRAVELERS? WHAT DRIVES PRICE?

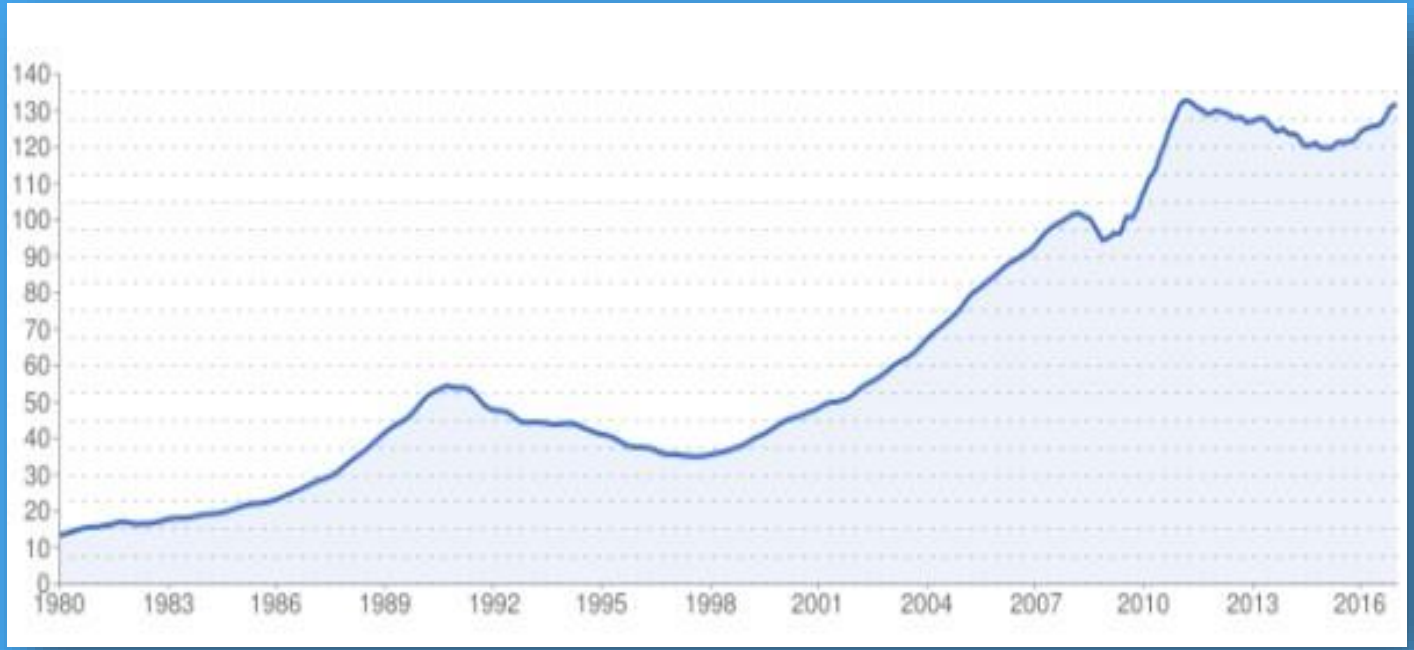
- walking distance to one of the top attractions (within 2 miles)
- rating >= 8
- other user defined criteria i.e. cleanliness, views, etc.
- number of beds
- number of bedrooms
- price range
- type of property
- Neighborhood

WHICH AREAS HAVE INCREASED THE MOST? WHICH SEASONS ARE HIGHEST?



- According to INSEE, the peak season for hotel bookings in Paris are the months of April, May, September, October and December.
- Starting in 2015 to AirBnB listings were at approximately 8,000 and increased to over 60,000 in 2019. (source: Athena Advisers)
- Paris has 20 distinct neighborhoods or Arrondisements and a few are markedly higher ; we examine each using geopandas in python

PRICING INTO 2020 -THE UPWARD RENTAL TENDS CONTINUE



AIRBNB EATS PARIS

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.

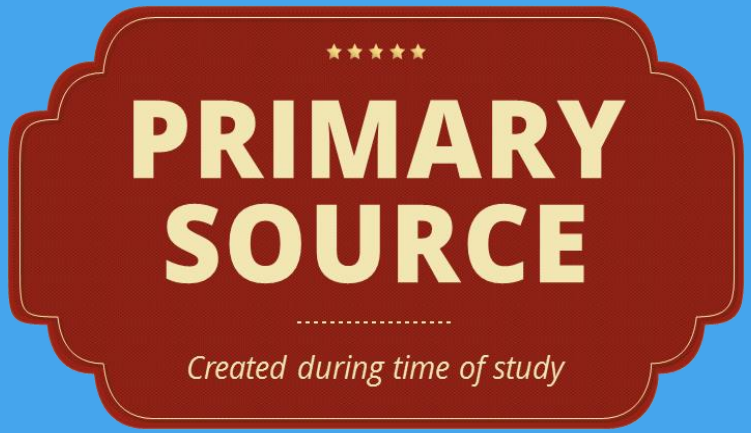
Hosts are expected to set their own prices for their listings or they can pay a fee to use a service () such as "Beyond Pricing" and others like "Smart Pricing" which will change pricing according to Supply and Demand and other typical factors such as holidays, special events and the like.

BACKGROUND, PROBLEM STATEMENT AND DESIRED OUTCOMES:

Although Airbnb and other sites provide some general guidance, there are currently no free services which help hosts price their properties. 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 that contribute to the cost and how to price a property to maximize revenue and be competitive.

BACKGROUND, PROBLEM STATEMENT PRIMARY SOURCE OF CAPSTONE DATA:

http://insideairbnb.com/get-the-data.html - INSIDE AIR BNB is an independent, non-commercial set of tools and data that allows you to explore how Airbnb is being used in cities around the world and creates a discussion also for those concerned about the socio-economic impact of AirBNB.



CONSIDERATIONS

- Data description
- Locations on the map
- What type of room are they?
- Average price for each neighborhood?
- Location amount in different neighborhood (heatmap)
- Average location price in different neighborhood
- Amenities that contribute to the cost – which amenities matter the most?
- walking distance to one of the top attractions (within 2 miles)
- rating == 10
- 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

OUR 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

We acquire datasets from one primary source and a few 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
- The month of April of 2019 is used review the details and explore the data prior to making a comparison to the data from earlier or later times.

The Calendar Data Set:
This data set has 58184 x 365 observations (prices from each calendar day x list ings) 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).

LISTING COUNT BY NEIGHBORHOOD

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 neighborhood.

neighborhood	#Listings	neighborhood	#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

AIRBNB CONCENTRATIONS IN PARIS

We see below the concentrations of AIRBNB Rentals around Paris:

DATA WRANGLING - IMPUTION

If the 'beds' had a missing value, and 'bedrooms' had a valid value, 'beds' is set with the value of 'bedrooms'.

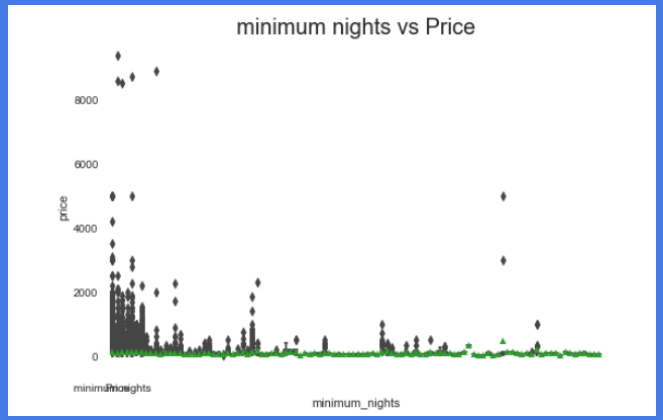
If both 'bedrooms' and 'beds' are missing, both fields are set to 1, which is the average number of beds and bedrooms, and it would be safe to assume that if a property is being rented out in Airbnb, that there is at least one bed regardless if it is a Studio Apartment or One Bedroom.

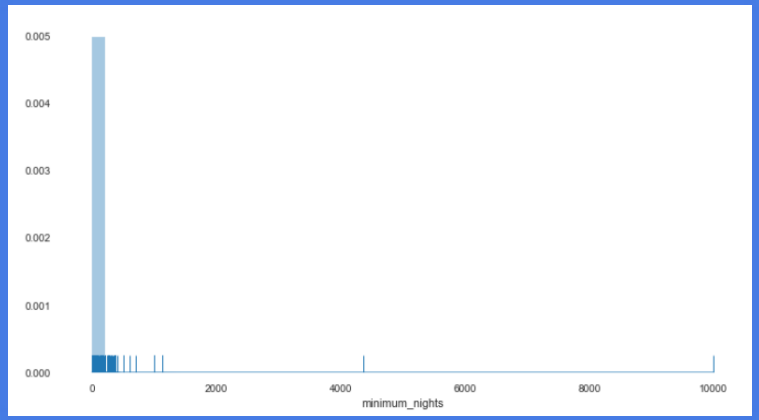
There were several records as well with missing 'scores' value, a new column 'rating_ind' was created to flag rated = 1, versus unrated = 0 records, so further analysis can be done between these two populations.

M

DATA WRANGLING - IMPUTION

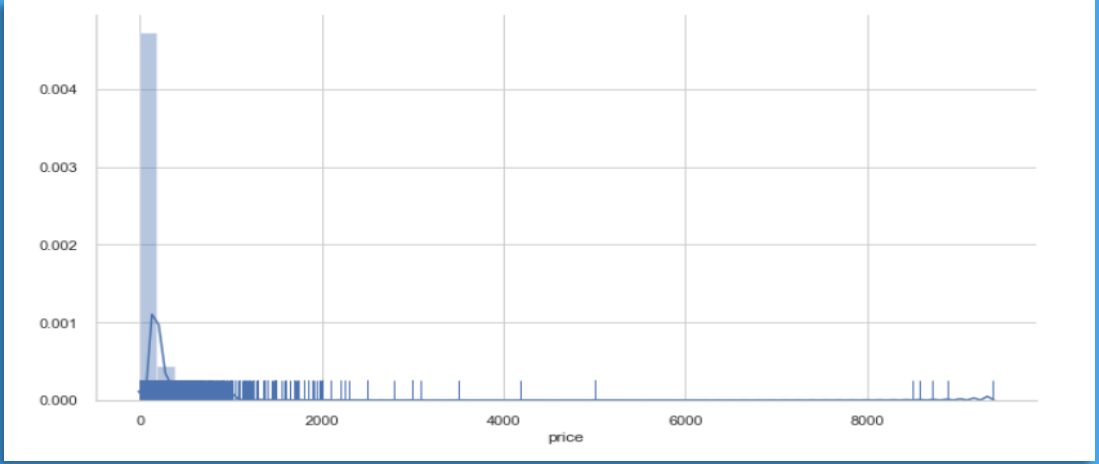
There are listings with 999 minimum nights and listings with prices as high as $9,379 per night for a studio apartment in a lower cost neighborhood





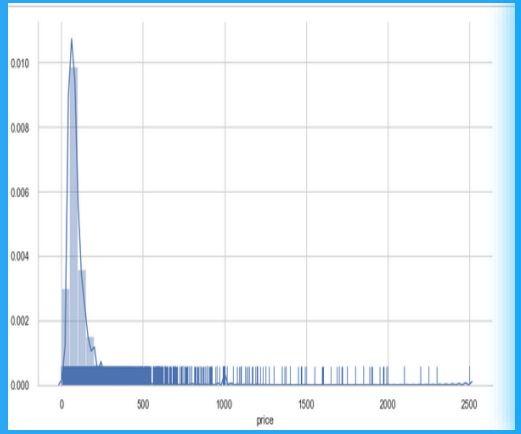
DATA WRANGLING - IMPUTION

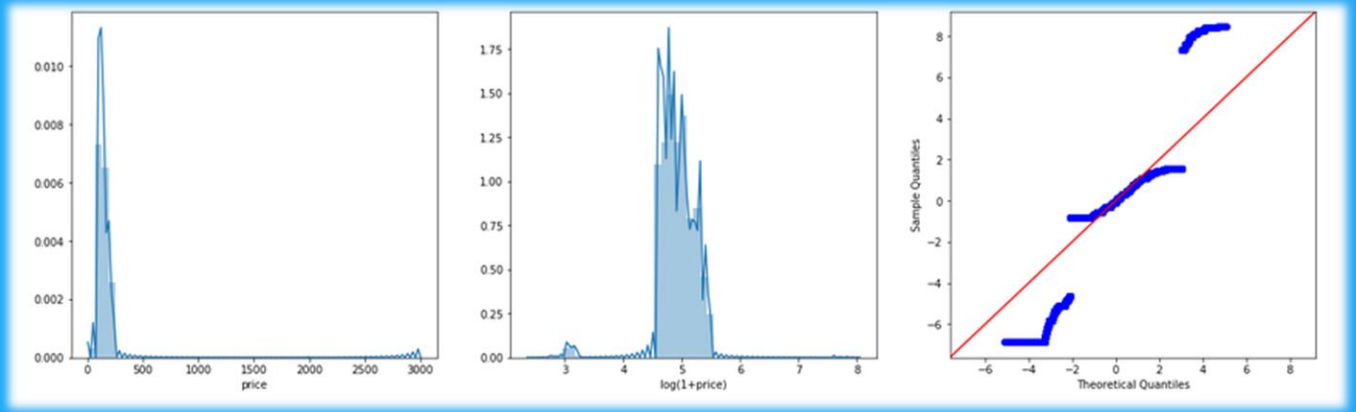
Since these are unusual I decide to replace all listings over 15 days minimum nights with 15 days as this is the average and remove listings that are more than $2500 per night which removes 20 listings from my dataset ; now have 58,164 to work with.



LOGGING THE PRICE RANGE

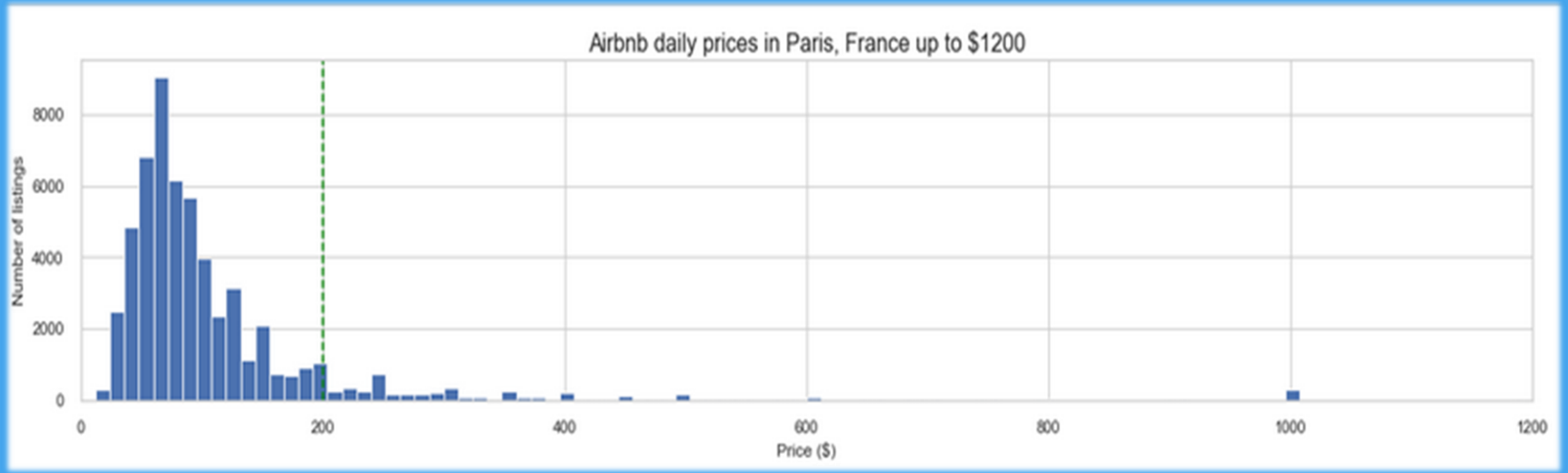
We see a wide distribution in prices so this needed to be reduced to listings between 10 USD and 2600 per night. We log our pricing to get a normal distribution





NORMAL DISTRIBUTION

Here we review prices ranging up to 1200 to see the distribution to note the majority of listings are between $20 and $200 USD

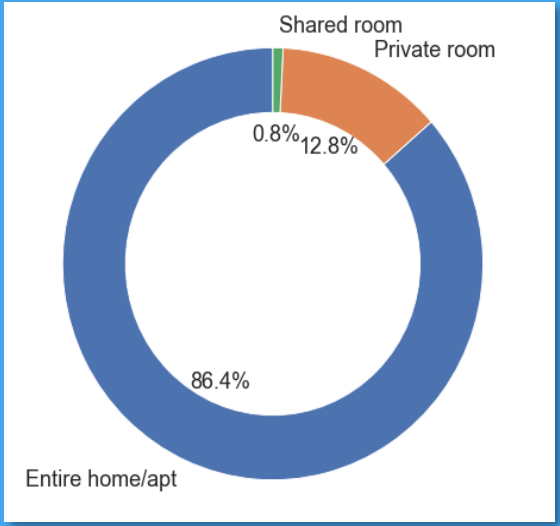


PRICE DISTRIBUTION

Here we review prices ranging up to $500 to see the distribution to note there are about 300 listings at $1000 per night and the majority less than $500

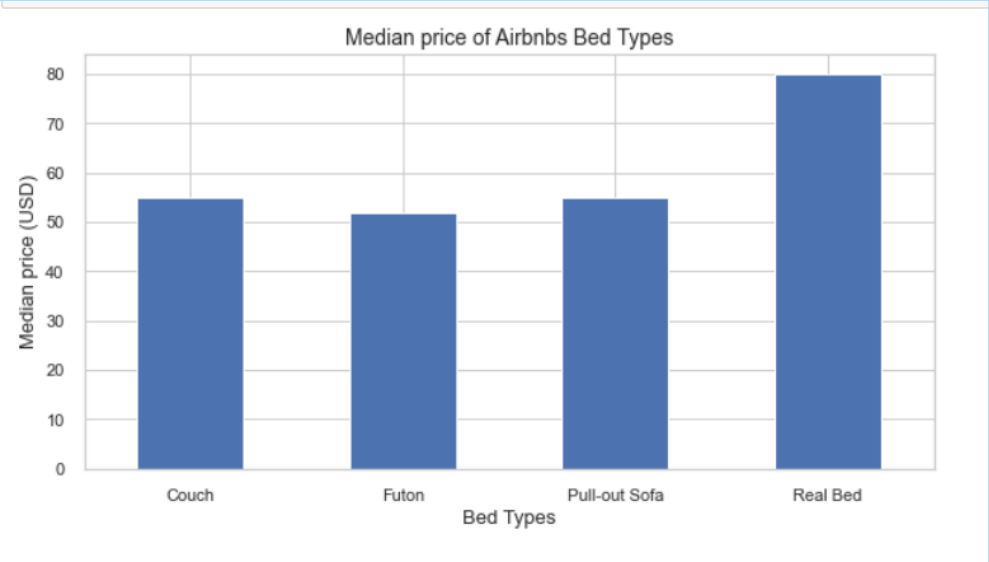
PROPERTY CHOICES

- I run my 55,653 listings though the plotting process to give perspective to our choices and how each type of listing breaks down. We find mostly Entire homes and apartments are offered.



LET'S CATEGORIZE OUR 29 DWELLING TYPES

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.

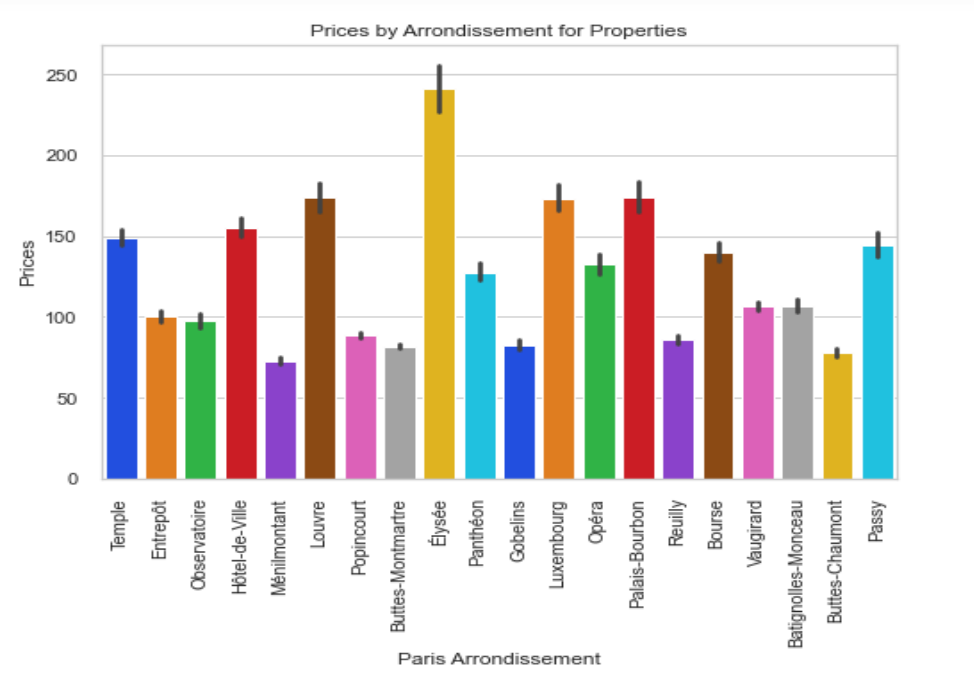


PRICE BY BED TYPE

MEDIAN PRICES

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

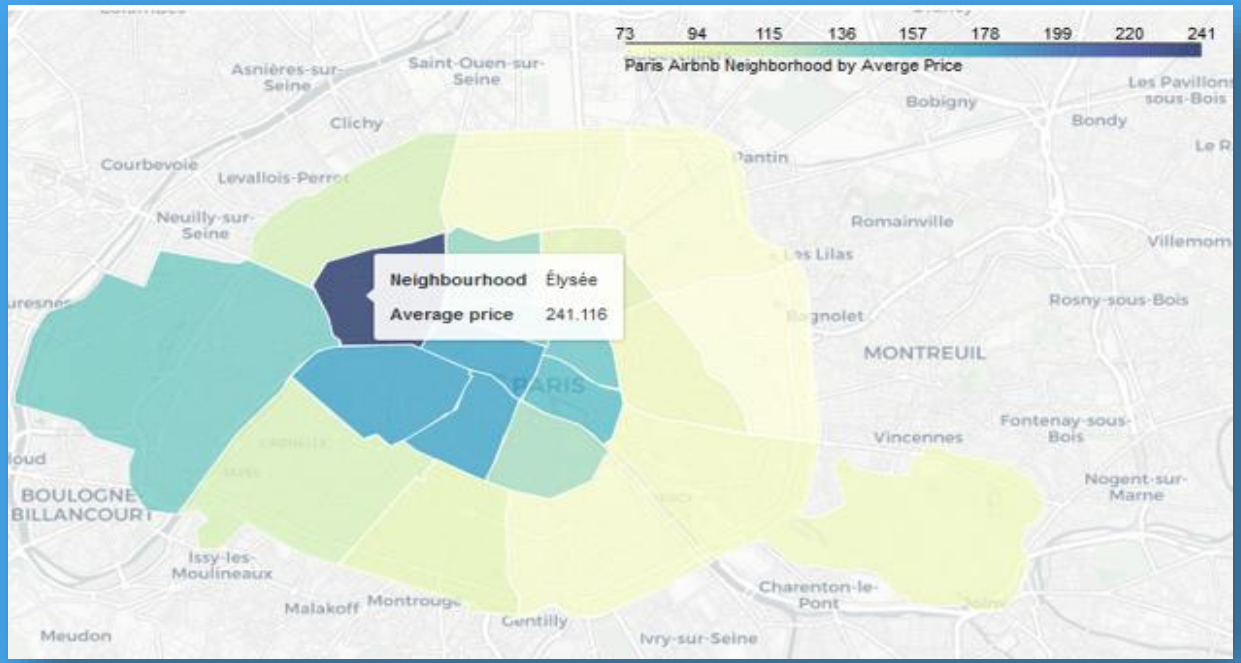
AVERAGE PRICES BY NEIGHBORHOOD



The highest neighborhoods are as expected as in the **Le Elyseè** we have the highest concentrations of attractions. The lowest are **Menilmontant** in the northeast area of Paris which has much fewer attractions and is roughly a 30 minute bus/metro ride from the center of Paris.

PRICE BY CITY REGIONS

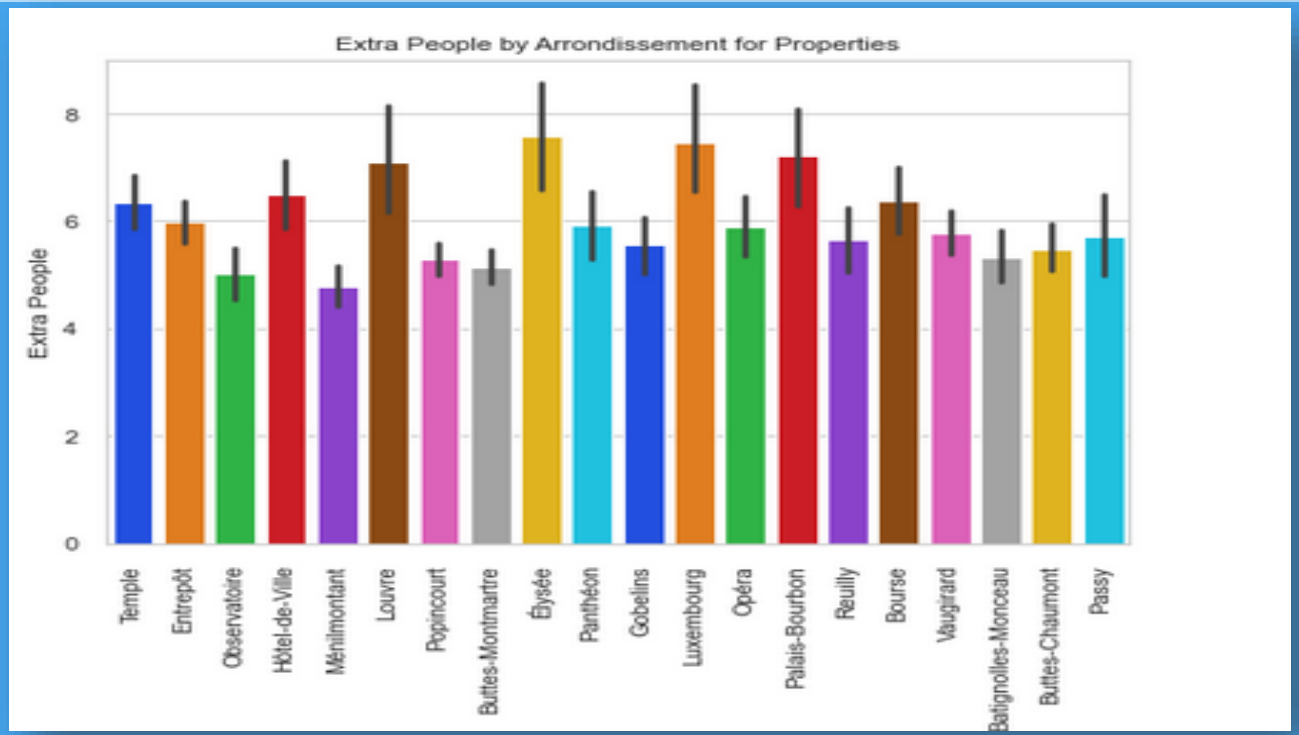
Below our map clearly tells us the price reduces as we go to the southeast and northeast of the city.



4 DIFFERENT MAPS – A 5 MINUTE VIDEO

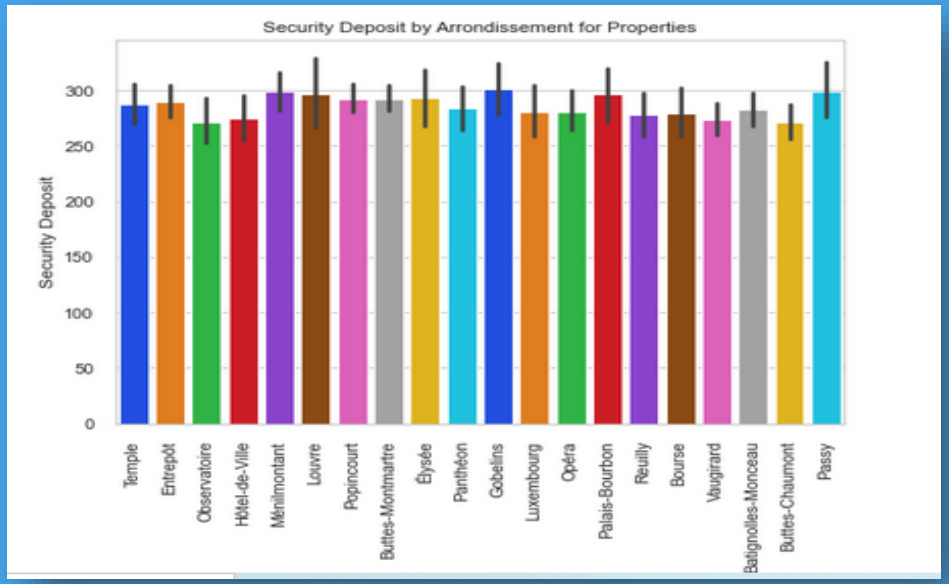
LET'S MEET OUR HOSTS

We have more "extra people" allowances for prime areas like Elyseè and Luxembourg



THE NEIGHBORHOOD PRICE DIFFERENCES

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

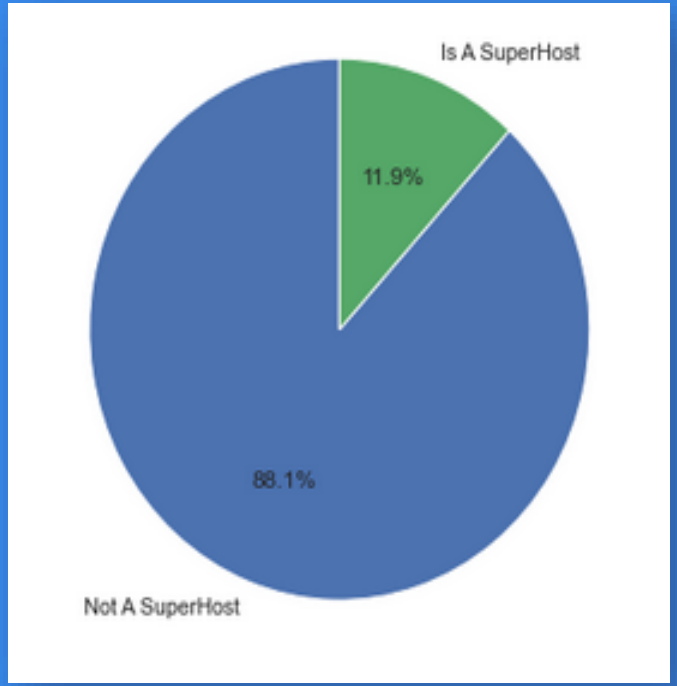


LET'S MEET OUR HOSTS

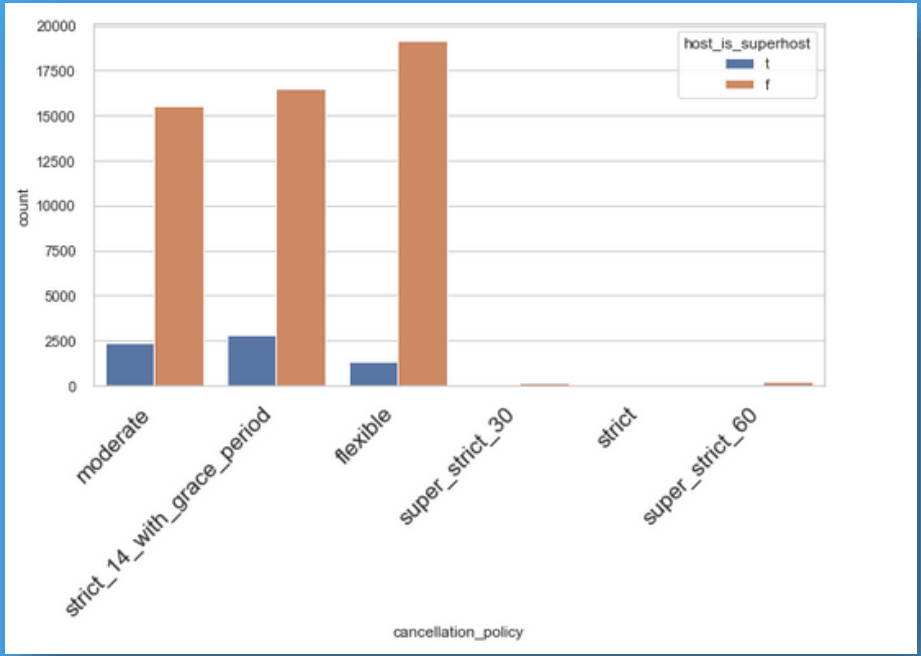
First of all what percent of our hosts have been bequeathed with this "prestigious title"?

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.

What do they do correct? What sets them apart? A few criterion are that the hosts maintain a 90% response rate and reviews above a 50% rating among others.



LET'S MEET OUR HOSTS CANCELLATION POLICIES



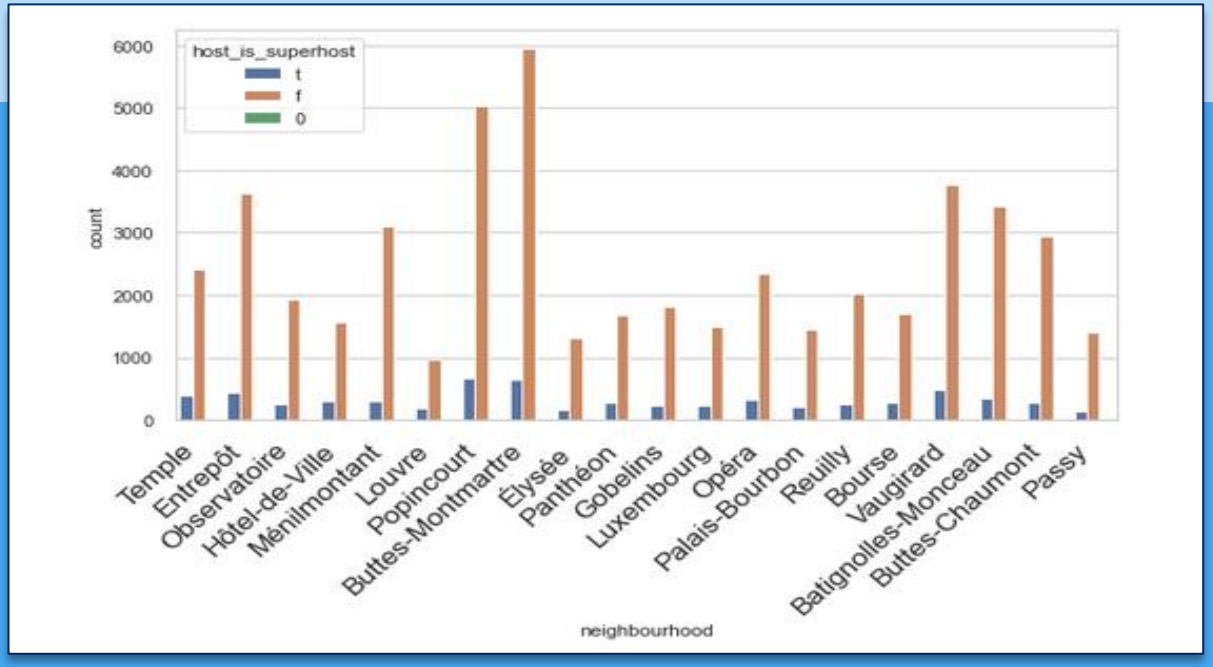
Most hosts choose The **"flexible"** option though more **super hosts** choose the strict 14 with grace period

The least popular is super strict 60 where

The renter must cancel

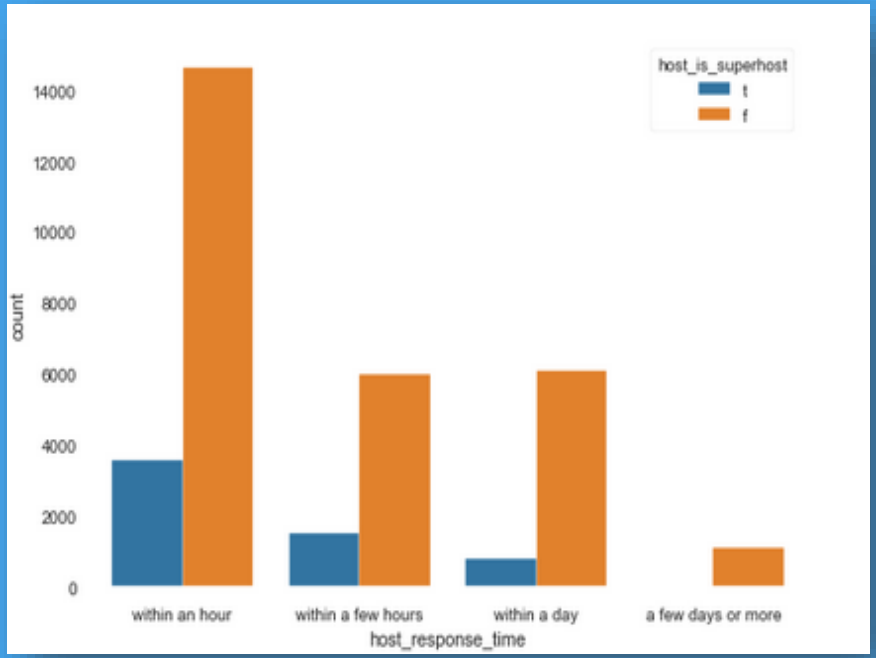
60 days in advance to get a 50% refund.

LET'S MEET OUR HOSTS



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 but they have more listings

LET'S MEET OUR HOSTS



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

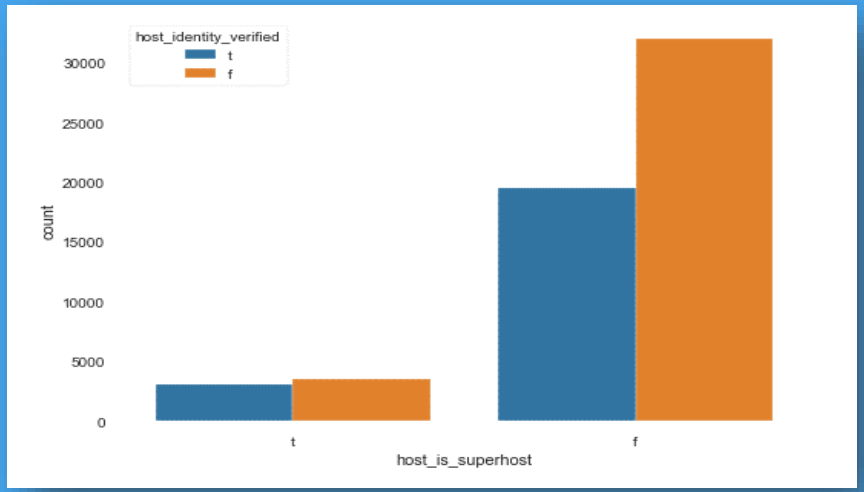
WHAT ARE THE BENEFITS OF AIRBNB ID VERIFICATION FOR HOSTS AND GUESTS?

Since Airbnb hosts rely on Airbnb to process and collect the payments from guests for bookings and expe-riences 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 facili-tate investigations that require the involvement of Airbnb. These investigations may involve tax, money laundering, sanctions laws, and criminal investigations.

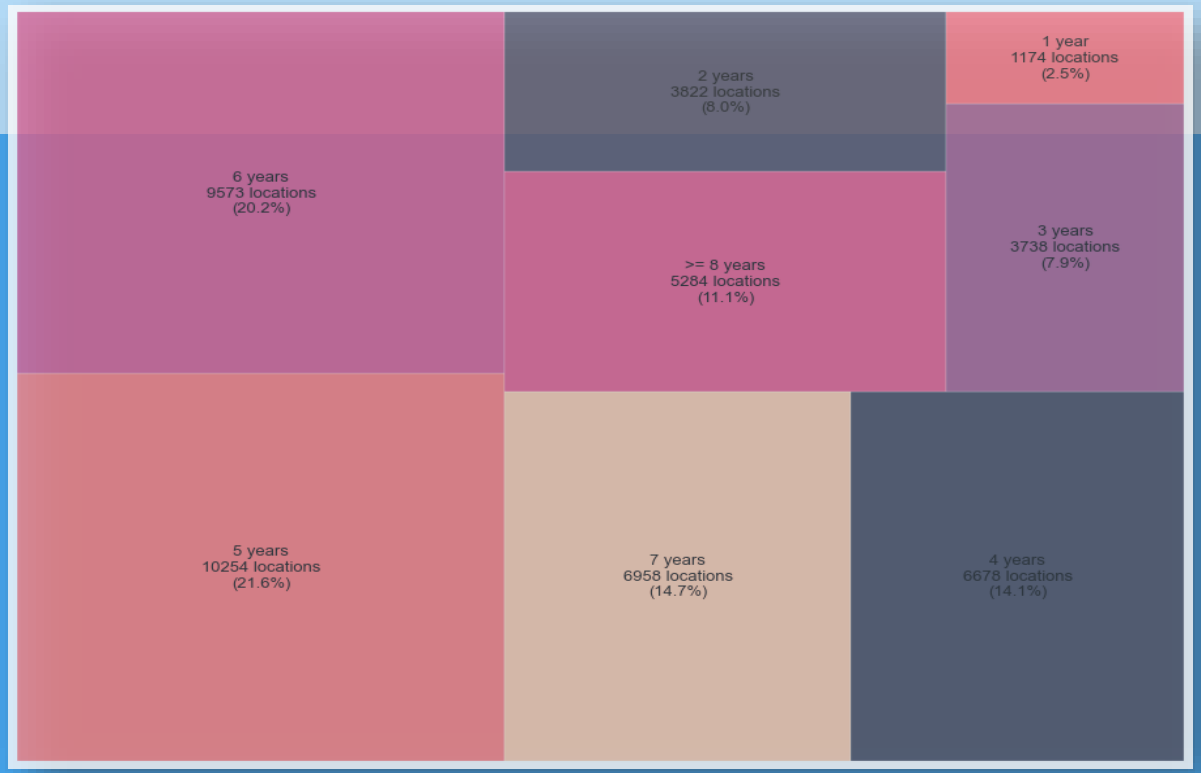
.

WHAT ARE THE BENEFITS OF AIRBNB ID VERIFICATION FOR HOSTS AND GUESTS?



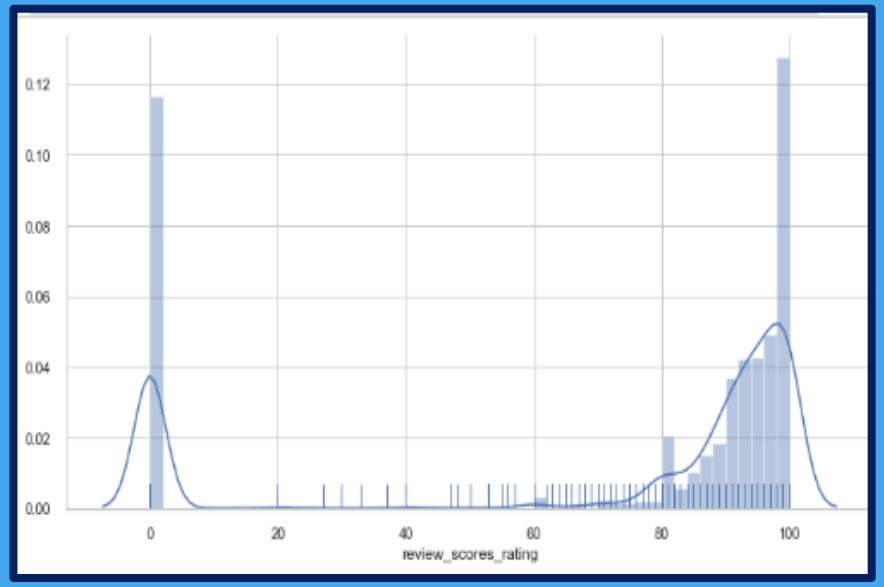
We notice that of the 11.9% of hosts who are superhosts less than half have had their identity verified. This leads us to believe that perhaps companies are managing the properties and Airbnb has not identified the individual employee who is managing the property.

NOTHING BEATS EXPERIENCE



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?

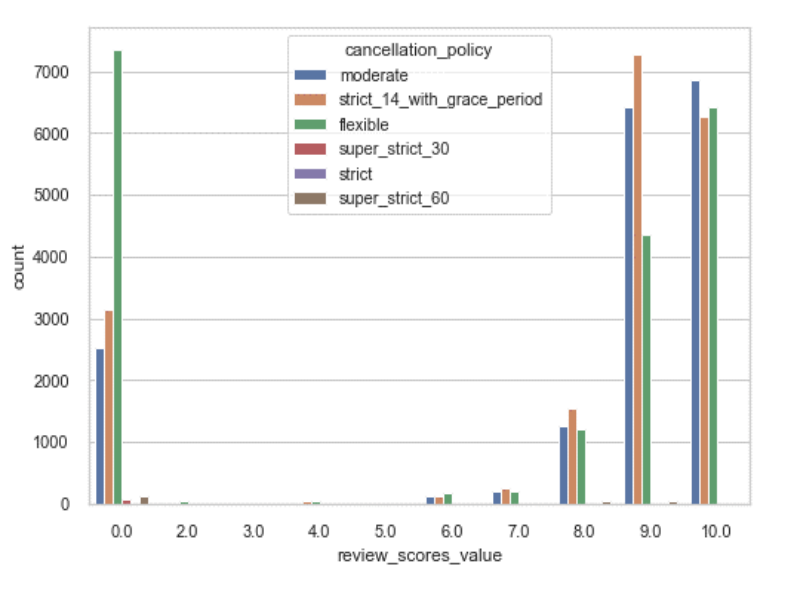


We immediately notice that the majority of our ratings are between 80 and 100% on a scale from 0 to 100%.

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 an opinion.

CANCELLATION POLICIES



Cancellation Policies

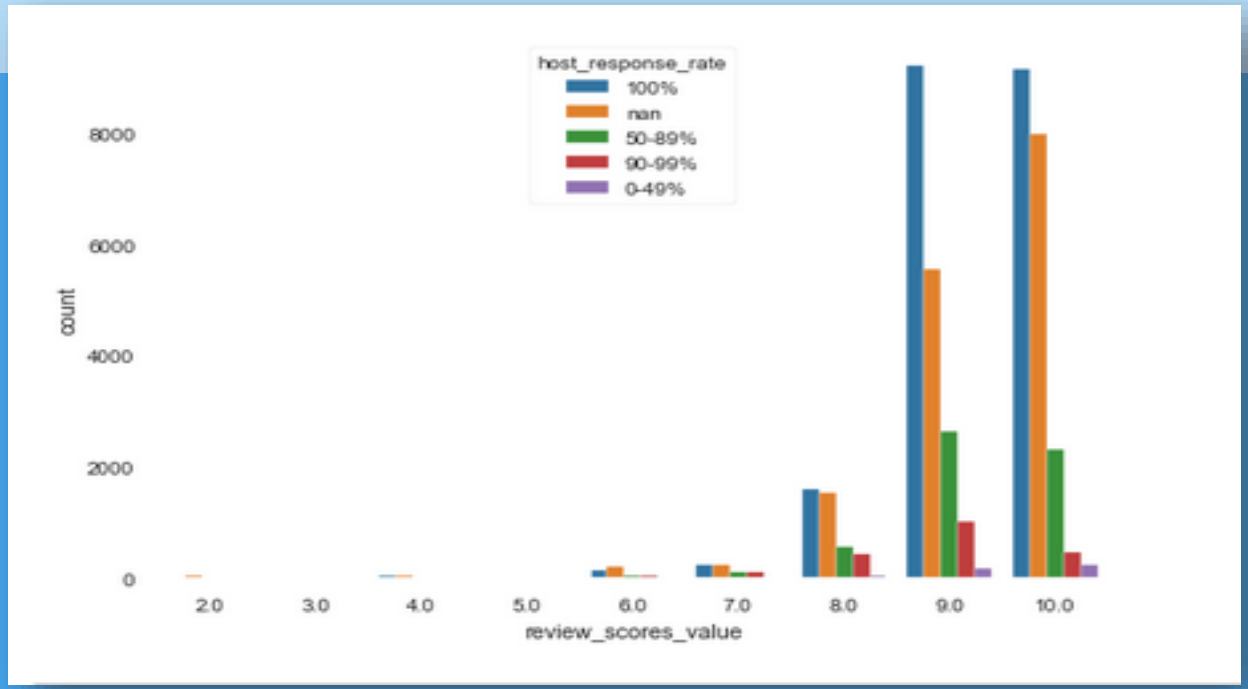
These 6 policies range from Flexible (the most lenient) to Super Strict 60.

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 30 or 60 days in advance to receive just 50% of the accommodation fees back.

DO RATINGS MATTER?



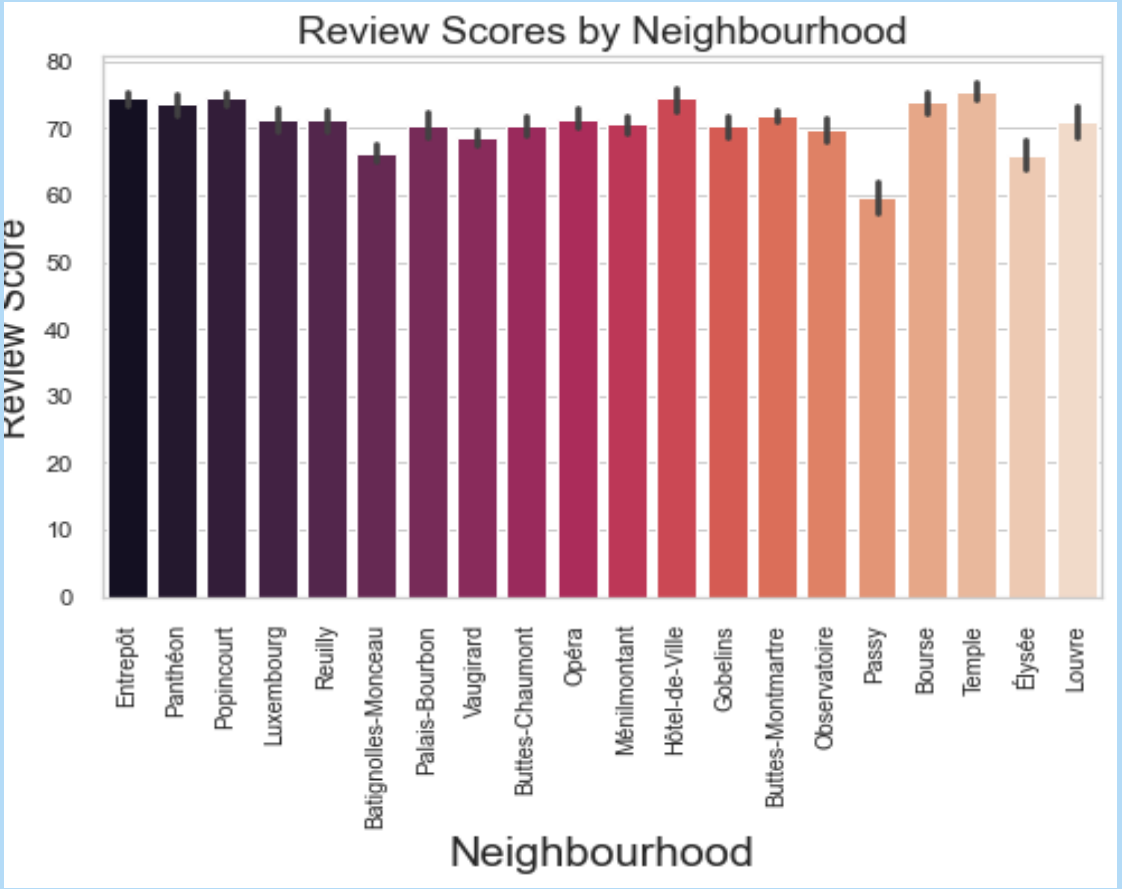
Here we compare the host response rate to the ratings; we see a high response rate with the higher ratings

DO RATINGS MATTER?

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

DO RATINGS MATTER?

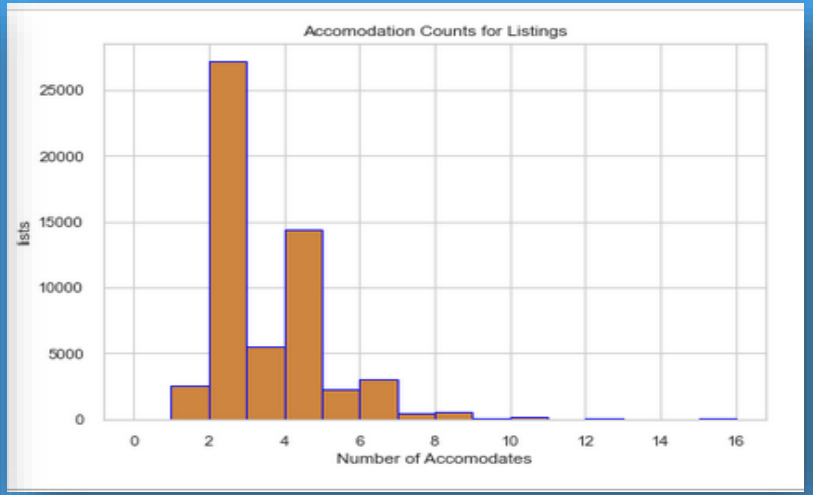
We see higher ratings in the Temple, Popincourt and Hotel de Ville neighbourhoods and the lowest in the Passy area.



DO RATINGS MATTER? (CON'T)

We see consistency in most of our review scores with the exception of cleanliness.

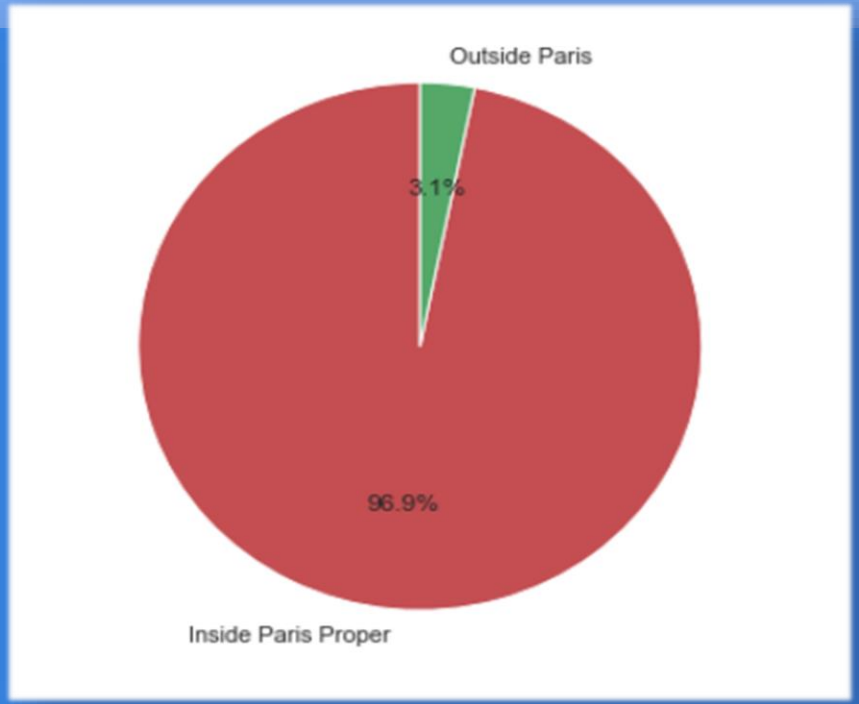
OTHER FACTORS THAT CONTRIBUTE TO PRICING OF AIRBNB RENTALS



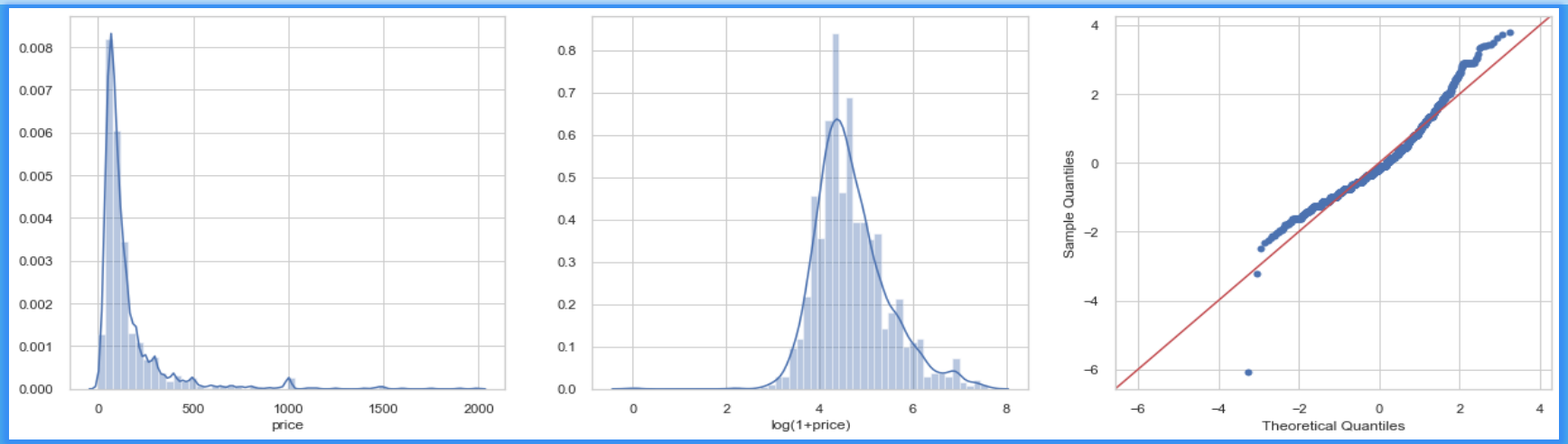
We see that over 50% of our listings accommodate 2 persons and the highest priced listings accommodate 16 people though the properties that accommodate 12,14 and 15 persons are equal in median price.

ZIP CODE REDUCTION – THE BORDERS OF PARIS LISTINGS

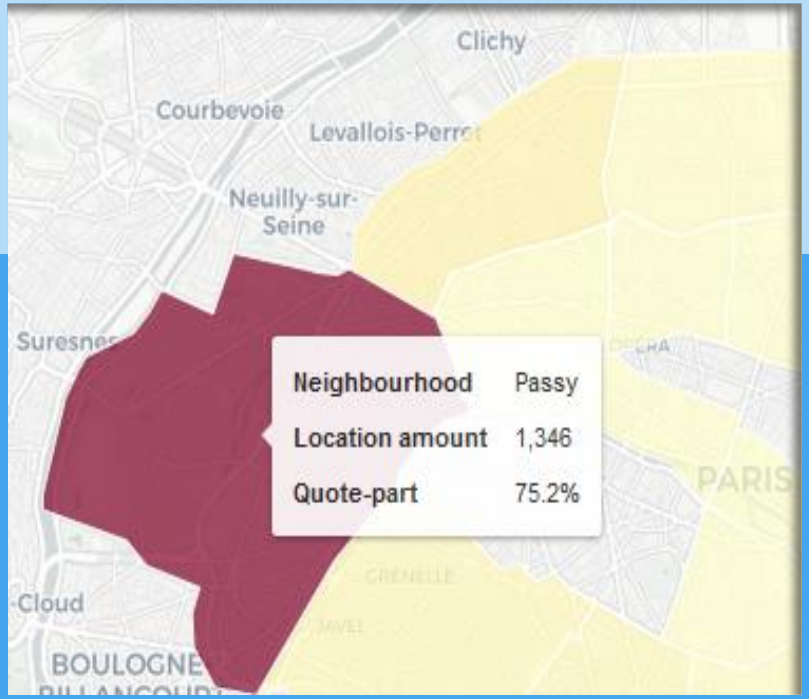
The zipcodes of “Paris Proper” are 75001 to 75021 so I separate out these codes from my groupings to get most of my listings , 55,653 out of 58,184 are in the Paris zipcodes and 1,788 are just on the Paris borders. I separate these into two groups “Inside Paris” and “Outside Paris”.



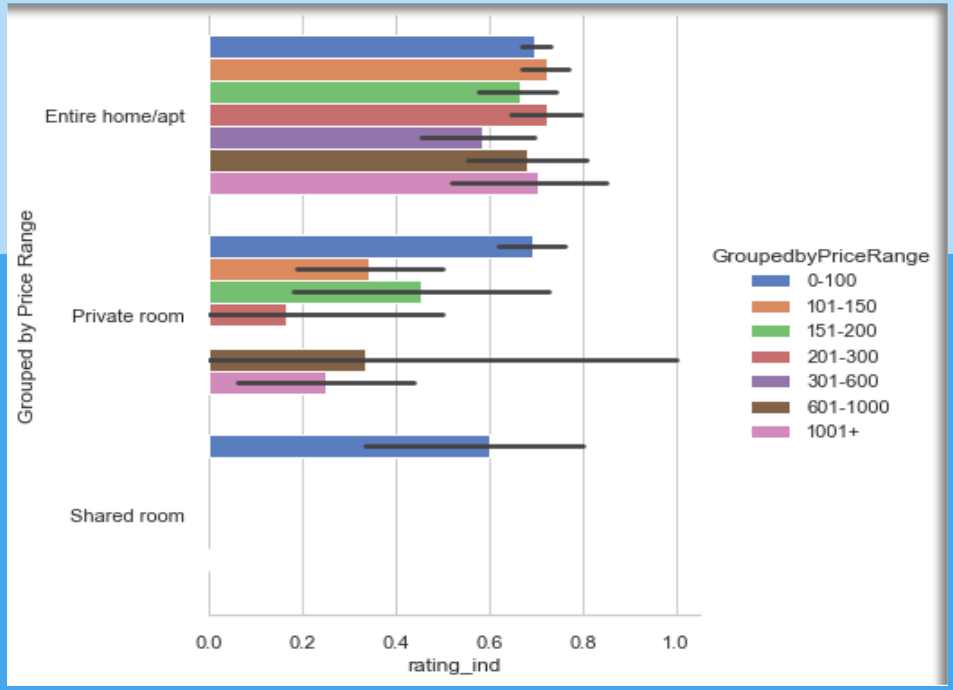
“JUST” OUTSIDE PARIS LISTINGS - DIFFERENCES



Outside Paris Price Distribution Chart: This has been logged transformed to display a “Standard Distribution” and the prices are similiar to outside Paris, as 75.2% of the listings not having Paris Proper zipcodes are found in the west part of Paris which is a historically higher priced area.



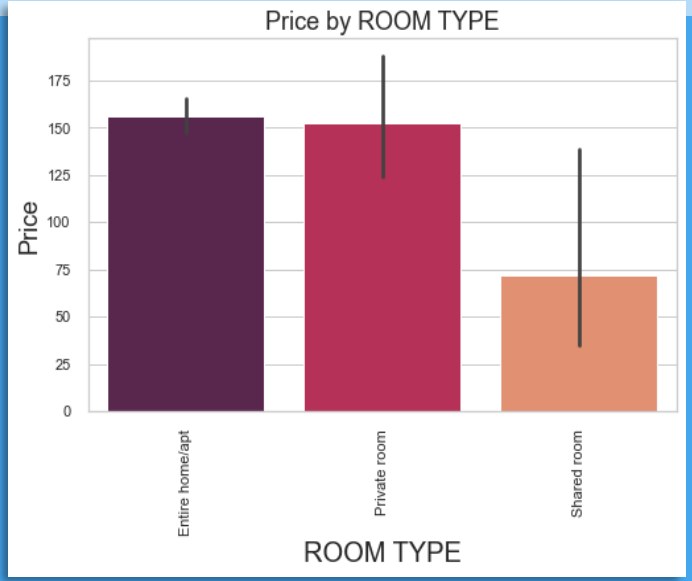
Passy Neighbourhoods (Western Paris)

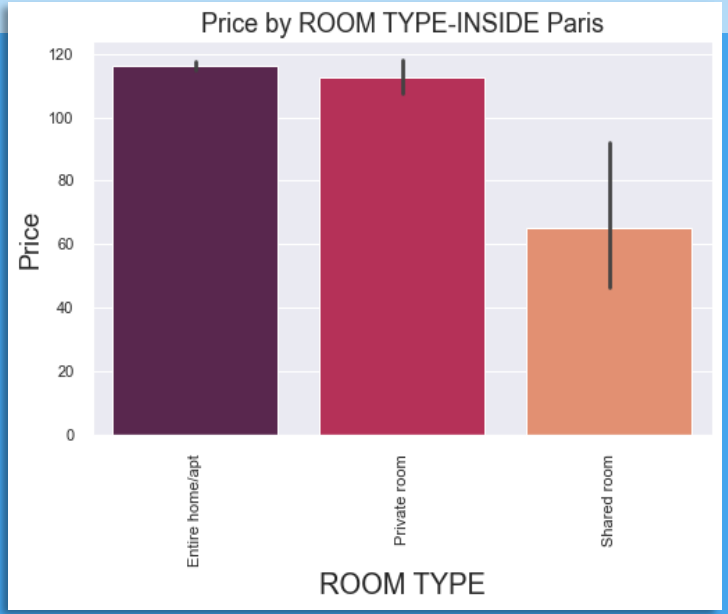


Above: Price by Rental Type

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$.

ZIP CODE REDUCTION – THE BORDERS OF PARIS LISTINGS





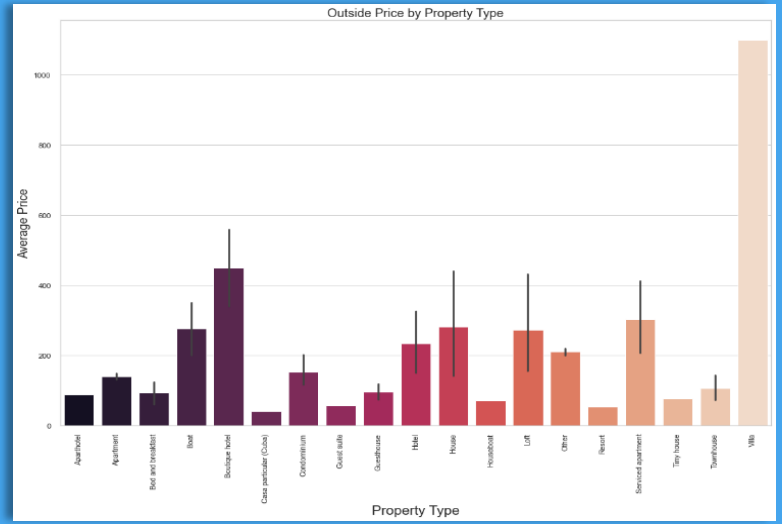
Average price of room type OUTSIDE PARIS

Entire home/apt 156.101307
Private room 152.419753
Shared room 71.687500

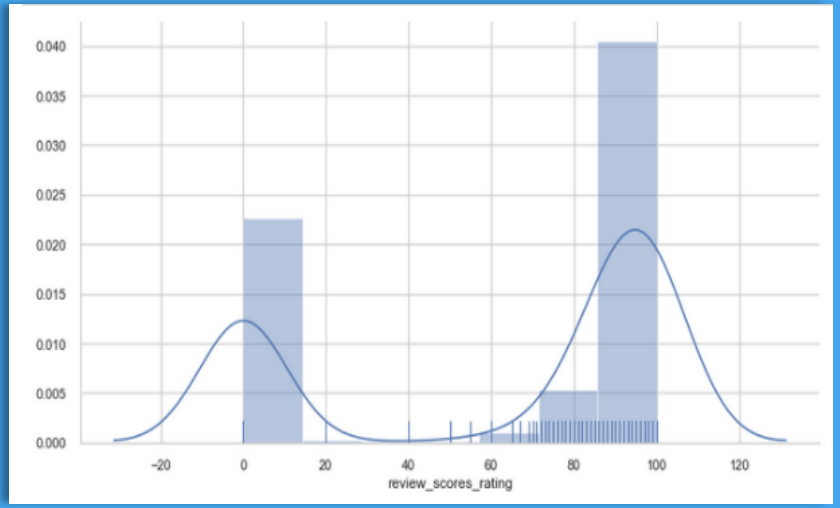
Average price of room type INSIDE PARIS

Entire home/apt 113.395738
Private room 109.489400
Shared room 53.818182

ZIP CODE REDUCTION – THE BORDERS OF PARIS LISTINGS

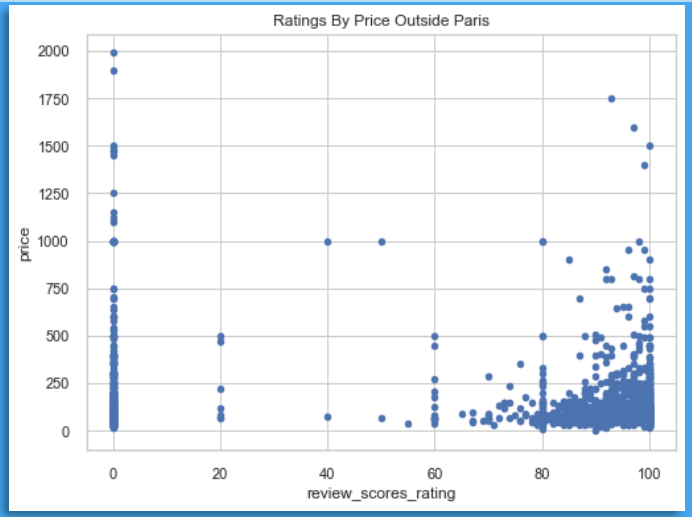


We see the average price by Property type for outside Paris. We can see the “outlier” is the VILLA which is just one property that is priced at over $1100 per night



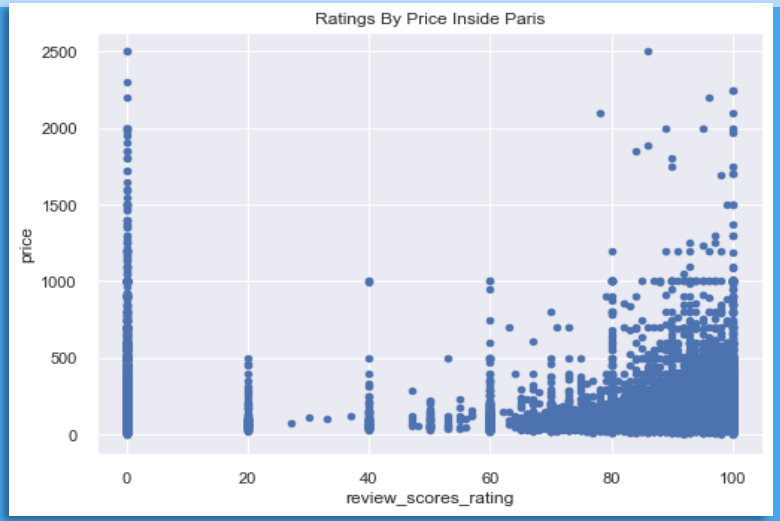
We notice a wider distribution of rating scores as there are more between 60 and 80 which are Attributes to the location factor according to the ratings categories.

ZIP CODE REDUCTION – THE BORDERS OF PARIS LISTINGS



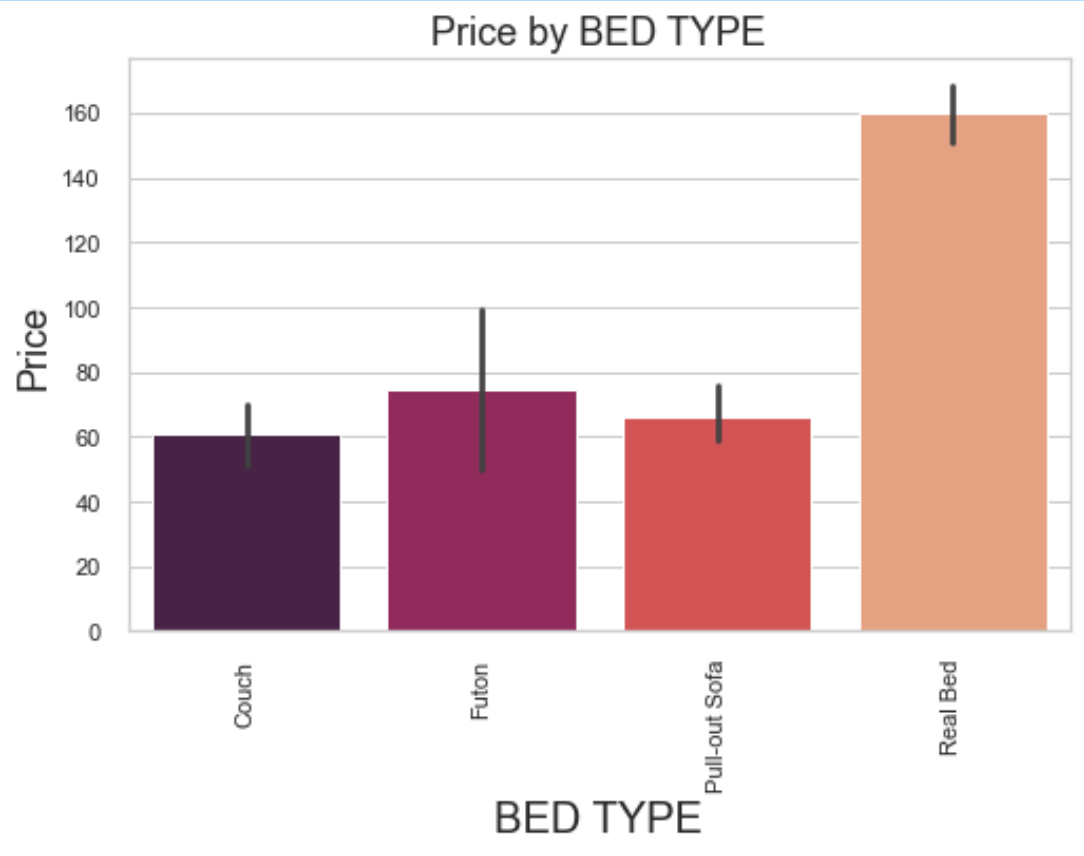
Outside Paris

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.



Inside Paris

We notice high concentrations of reviews between 80 and 100%



THE BORDERS OF PARIS LISTINGS

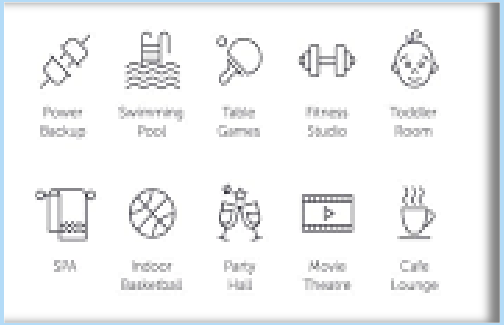
Fig 40 Left the average bed type with the average price; the “real bed” is priced about $30 higher than its equivalent in “Inside Paris”. The other 3 types are roughly the same.

Price by Average

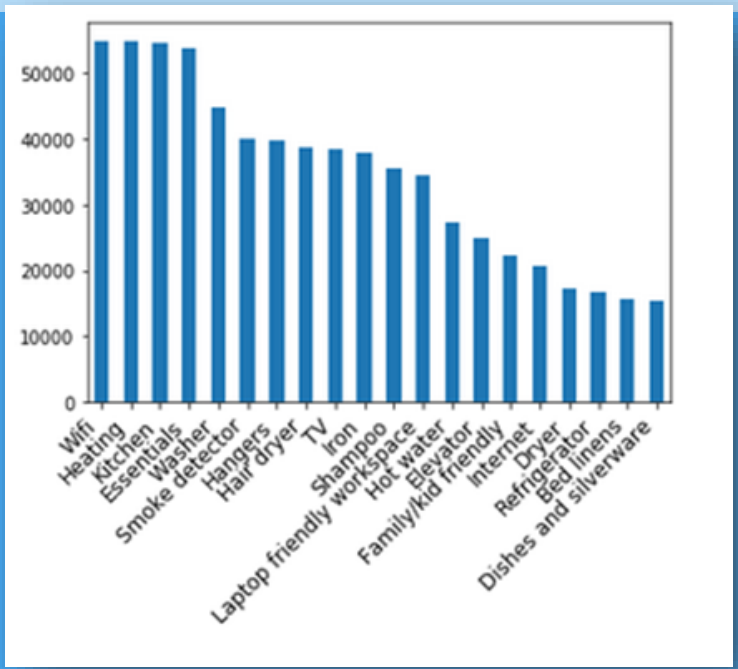
Couch	60.86
Futon	74.50
Pull-out Sofa	66.14
Real Bed	159.85

AMENITIES

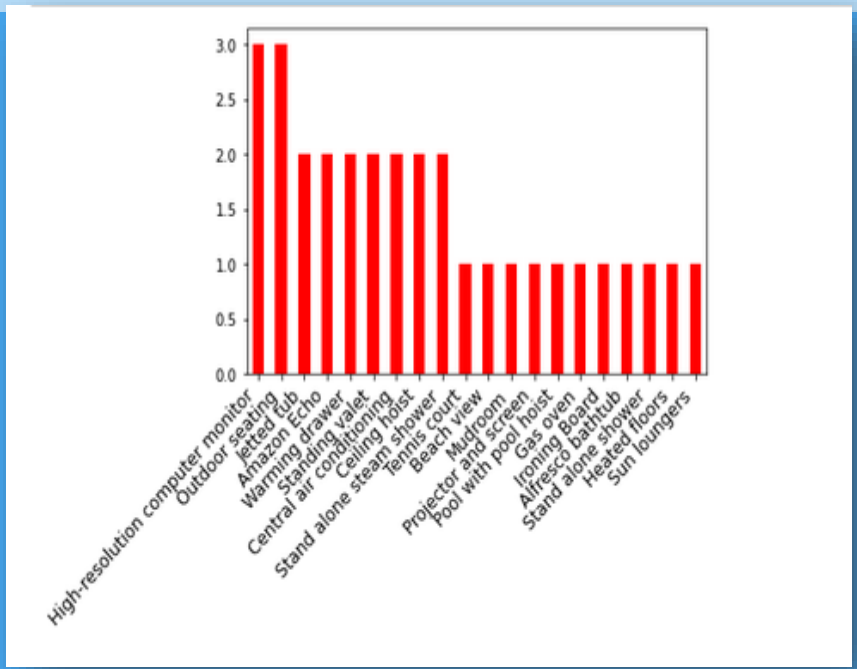
- So very many amenities ! Which ones matter the most out of the 178+ that are offered? Can having a balcony actually make a difference in price versus having all the day to day living items such as a Washing Machine/Dryer, a Kitchen and a Coffee Maker? We resolve these issues by reviewing what the top 20 amenities are, the bottom 20 and those that are the top and bottom 20 for cost and then decisions are made from these results. This also reduces the “curse of dimensionality” as we have sparse data as the amenities become less common.



AMENITIES

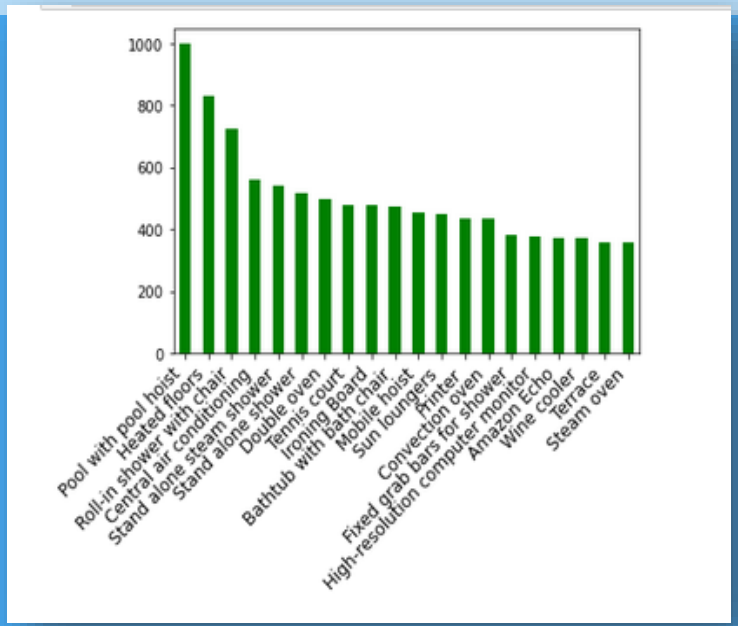


Top 20 Amenities the most offered

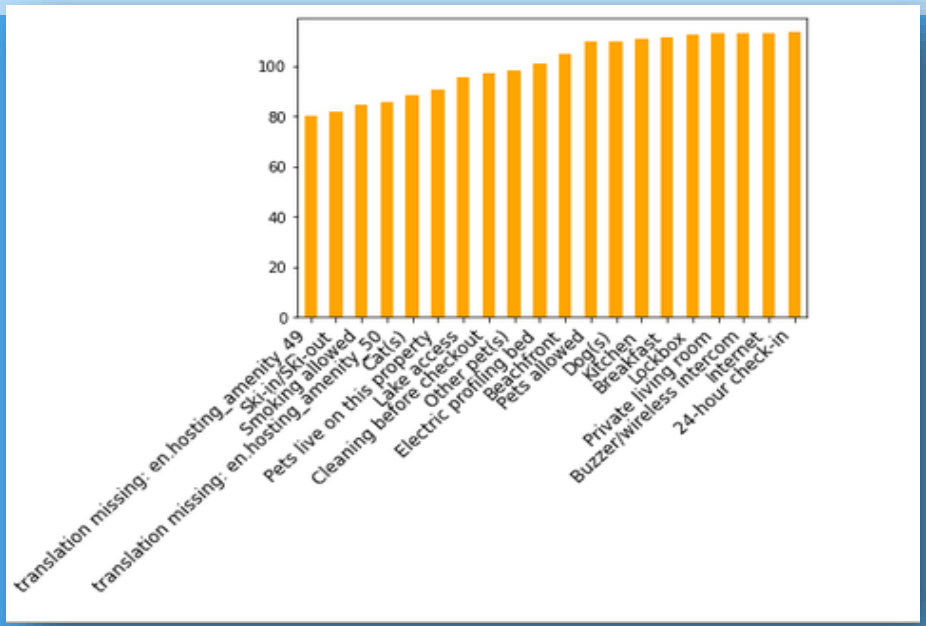


20 Amenities that are least offered

AMENITIES



Top 20 Amenities the most offered by the Most Pricey Properties



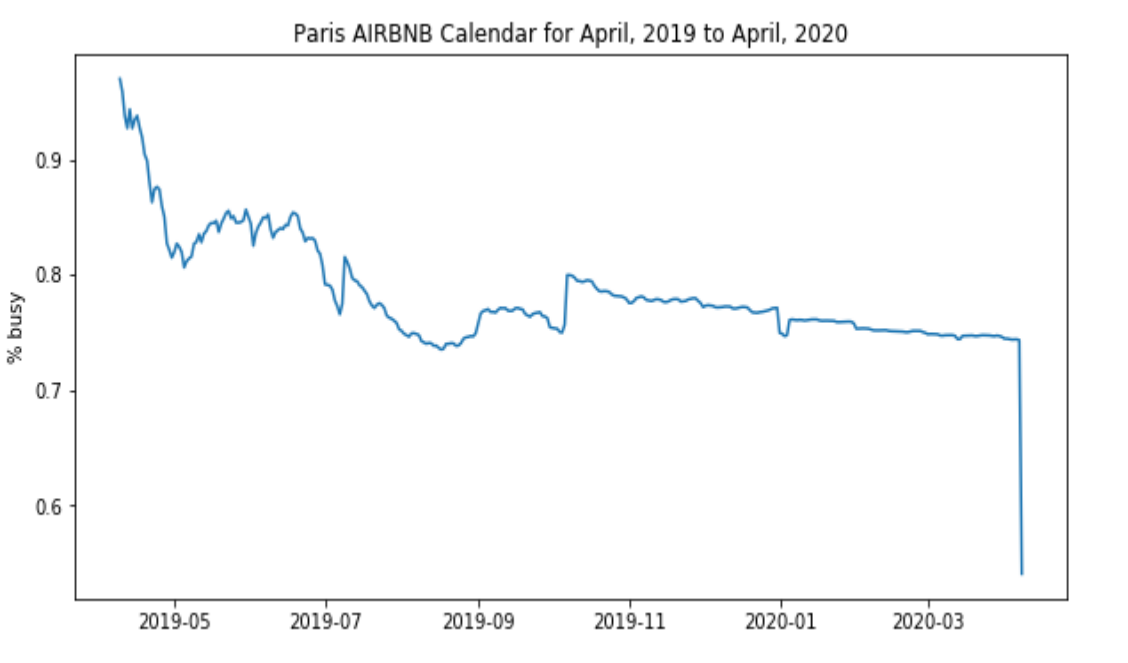
20 Amenities that are offered by the cheapest properties

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. This data set had 21,235,880 observations and 7 Data Points. The main purpose of this dataset is to review the availability and bookings for the month.

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

EXPLORATORY DATA ANALYSIS:



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

EXPLORATORY DATA ANALYSIS:

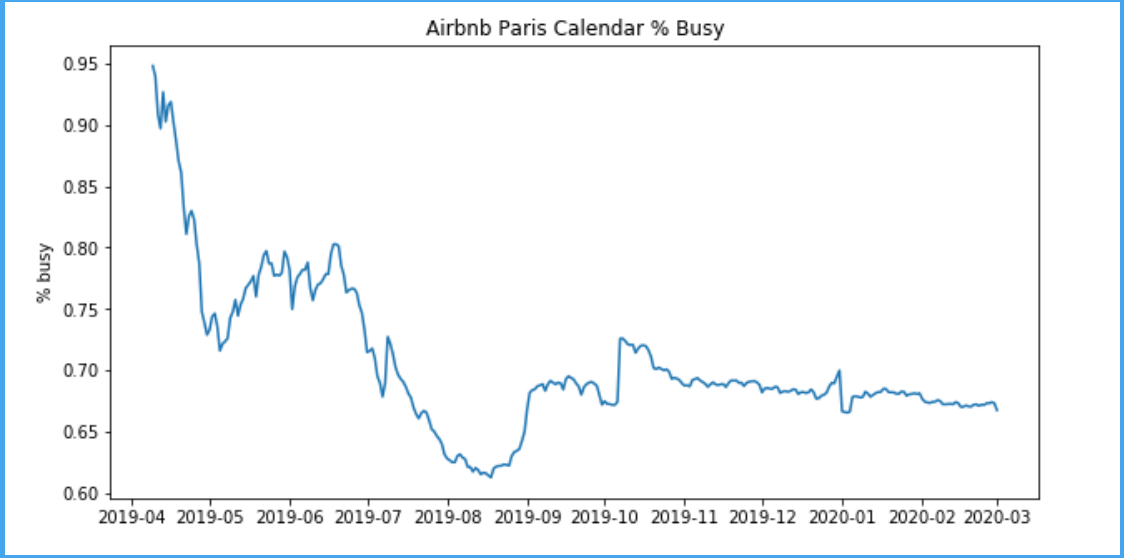
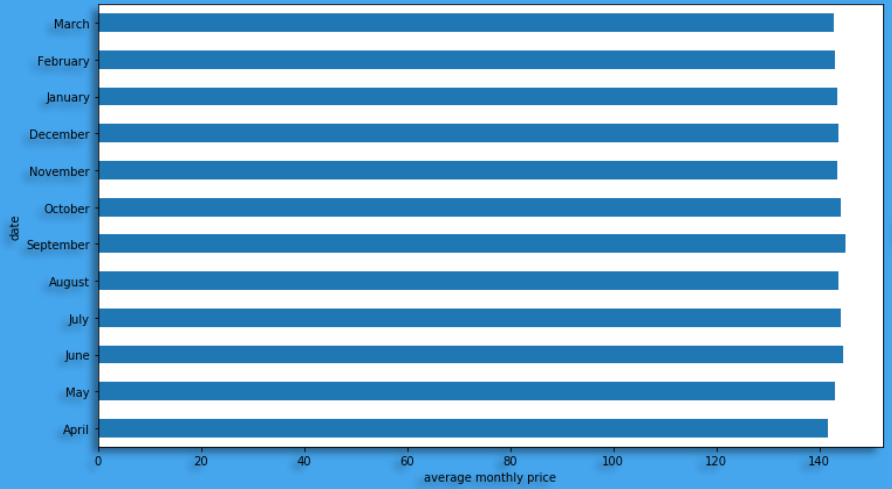
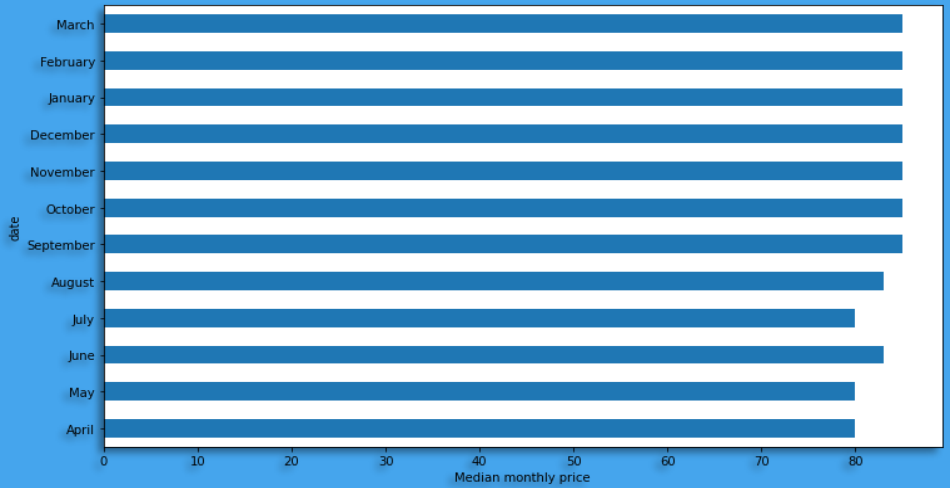


Fig 2: I adjust my dates to: April 9, 2019 to March 1, 2020 to better understand the pre-COVID19 rental scenario we notice a steep decline in July and August as this is when the French population typically take their 8 week vacations.

MONTHLY PRICES

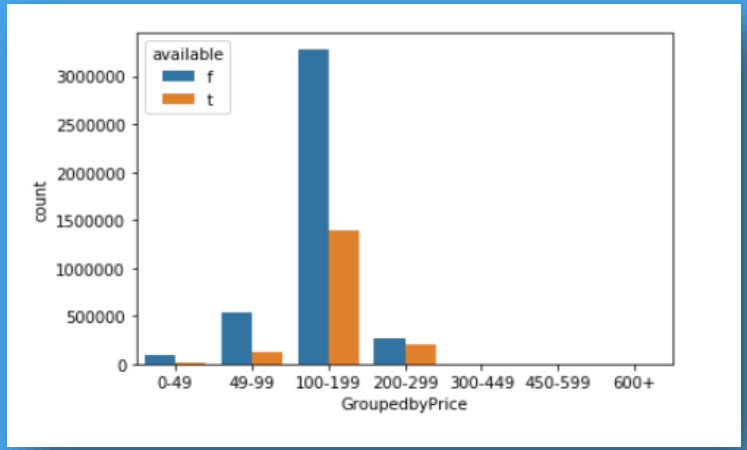


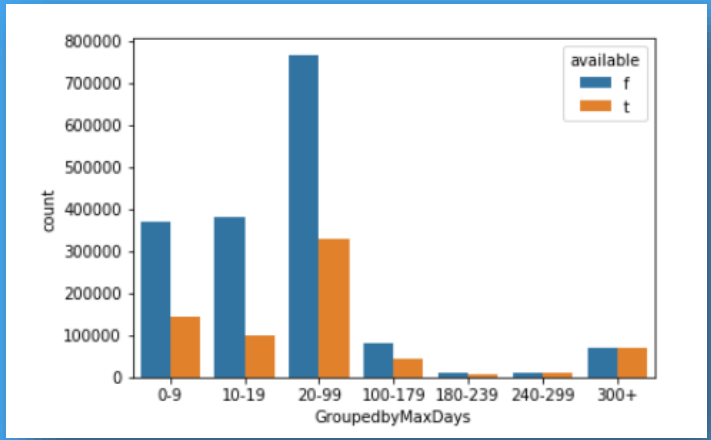


I found an overall average of $119.42 with slight increases in September and slight decreases in April. I found the median price to be $85 USD per night with decreases in April ,May, July and August

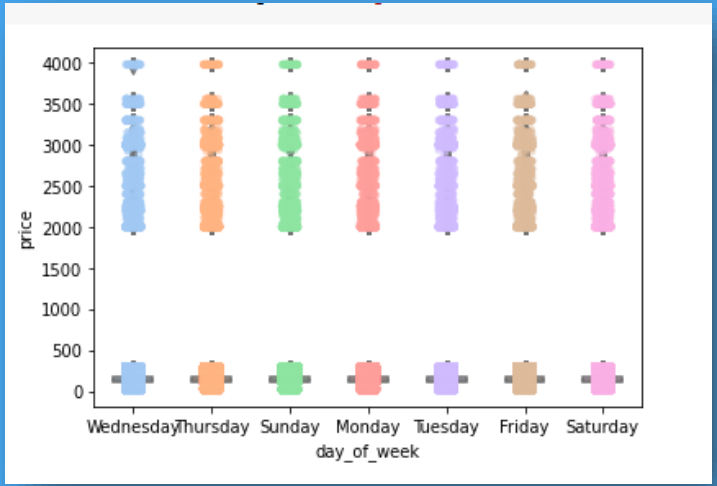
EXPLORATORY DATA ANALYSIS:

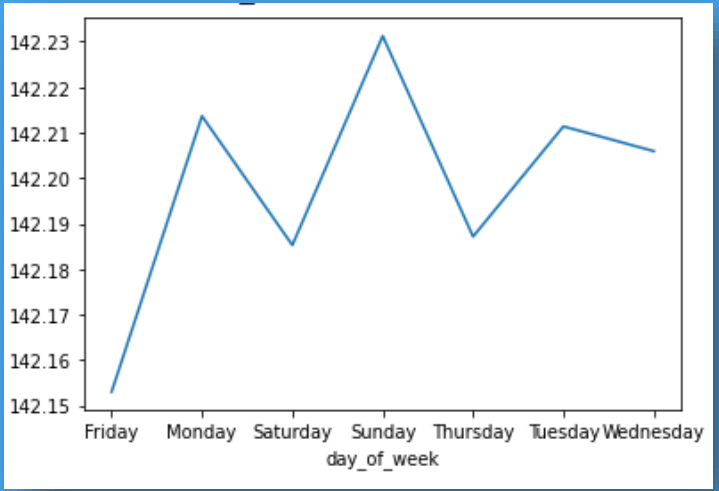
We see the majority of prices are in the 100 – 199 USD range and the majority of listings are available between the 20 and 99 day range.





EXPLORATORY DATA ANALYSIS:

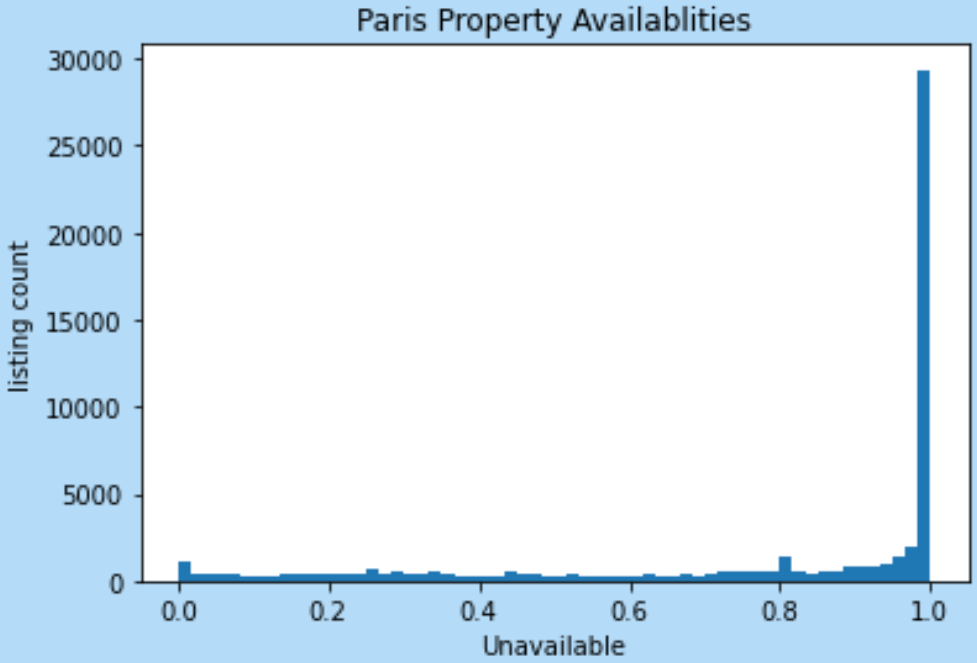




Fig(s) Above: We review our average daily prices and do not notice larger changes on the weekend which is typical of listings.

EXPLORATORY DATA ANALYSIS:

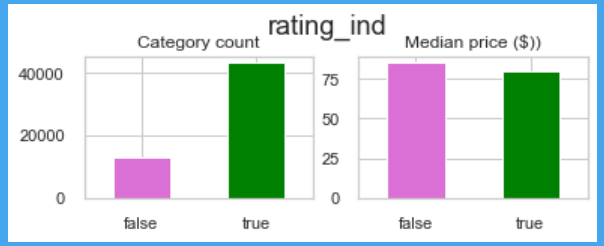
Here we see that the majority of list-ings (67%) are not available 365 days out of the year.

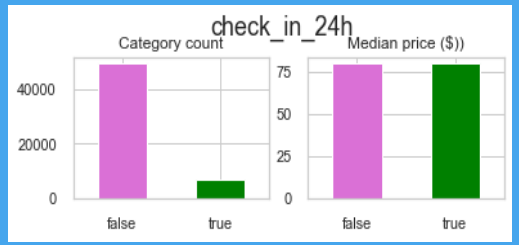


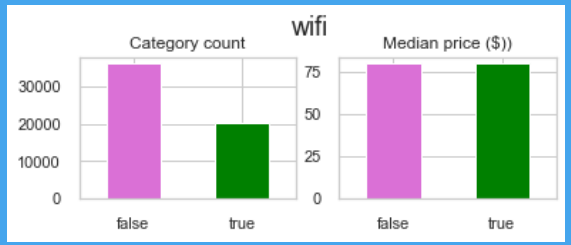
Maximum Nights: The maximum nights the Owner is willing to rent it though Paris has restrictions to 120 days per year if the property is the Owner's primary resident so we can "guess" that is the listing is on the market for more than 120 days it is classified as a share or it is simply "illegal".

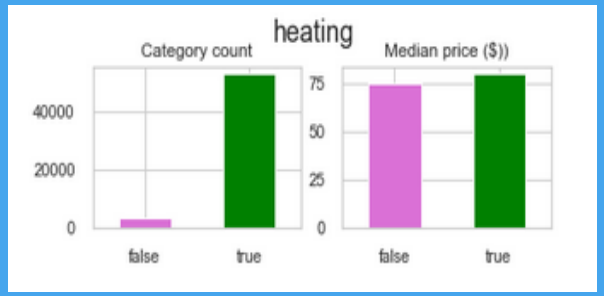
FURTHER RESEARCH AND ANALYSIS

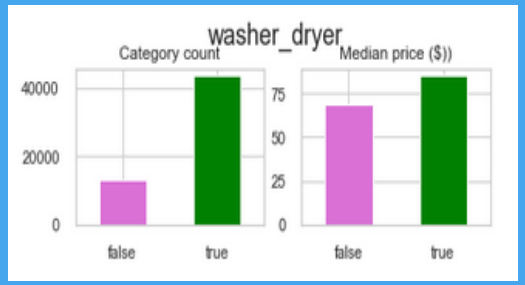
Below I rate each amenity to verify their relevance and decide which ones are worth retaining in the dataset.

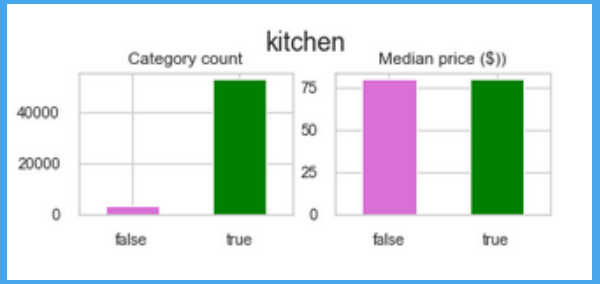




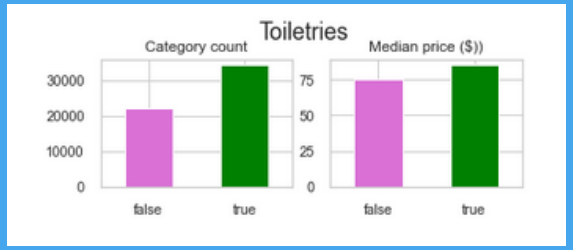


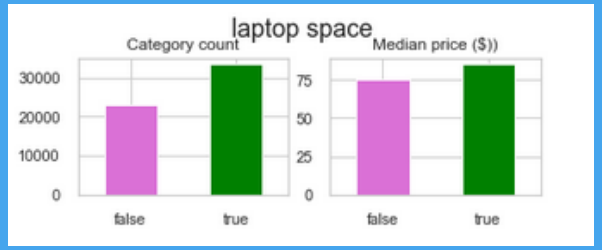


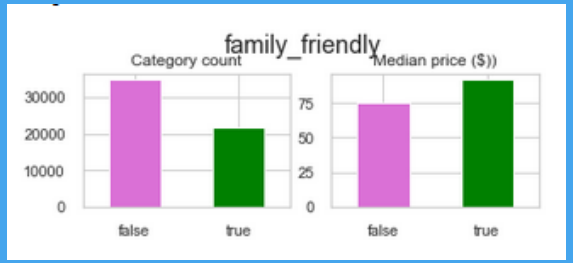


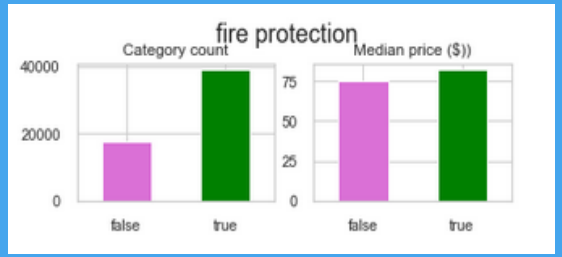


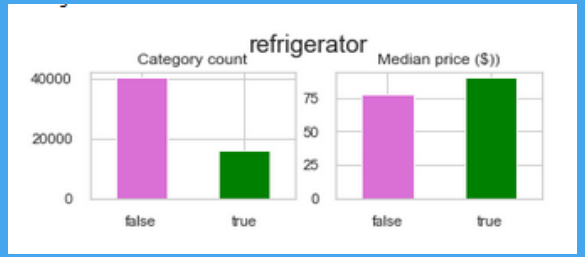
FURTHER RESEARCH AND ANALYSIS



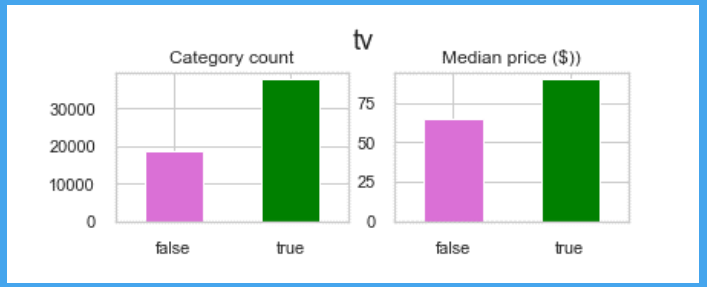


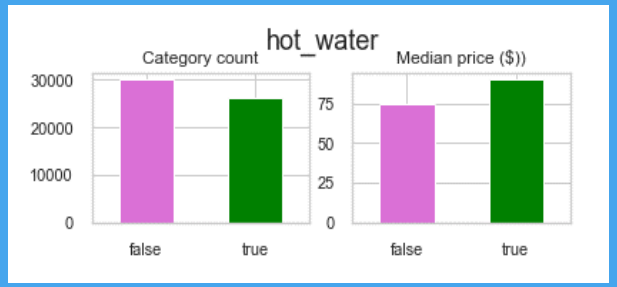


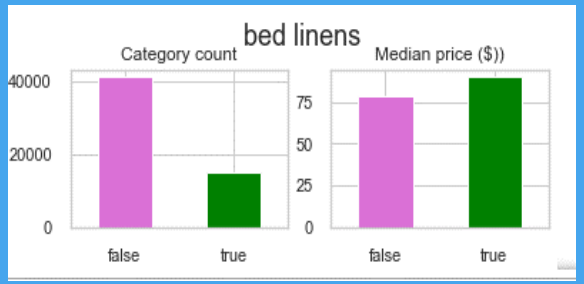


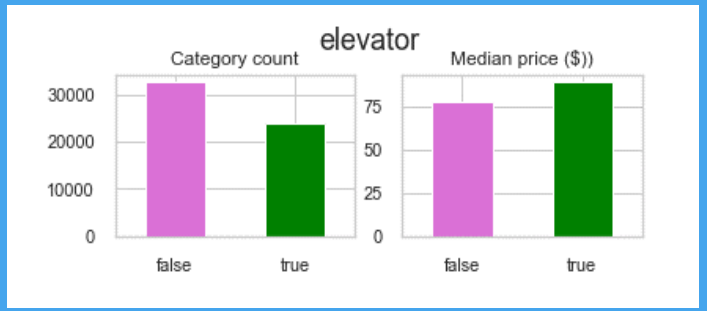


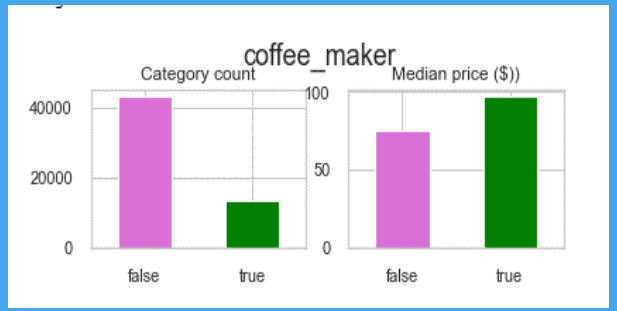
FURTHER RESEARCH AND ANALYSIS



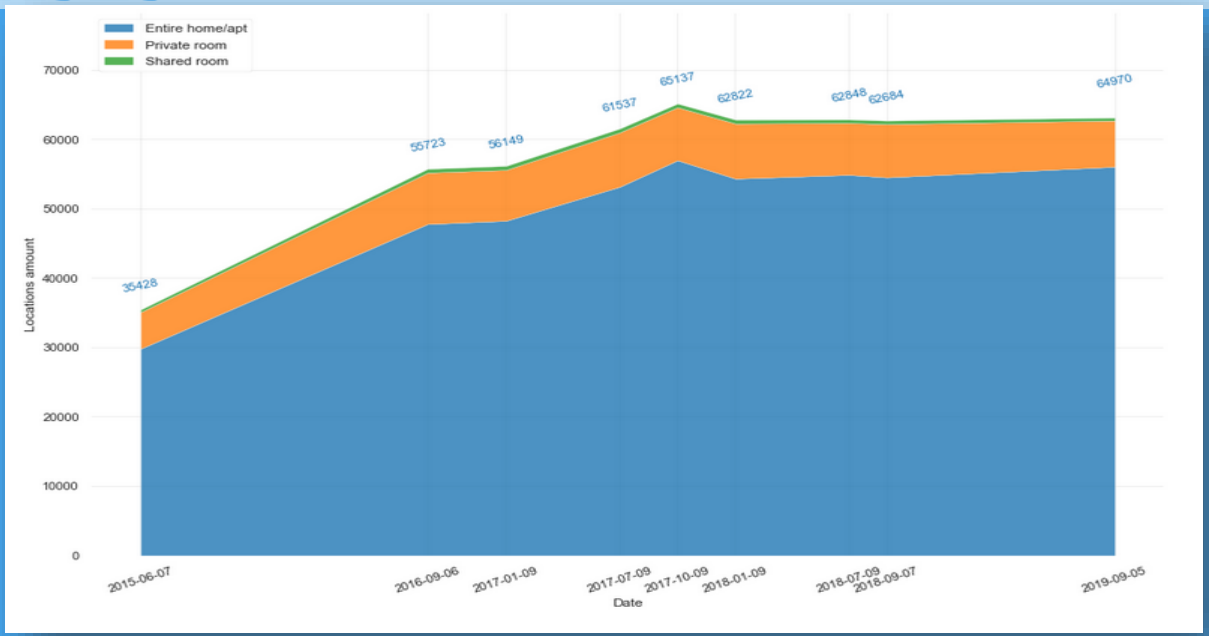






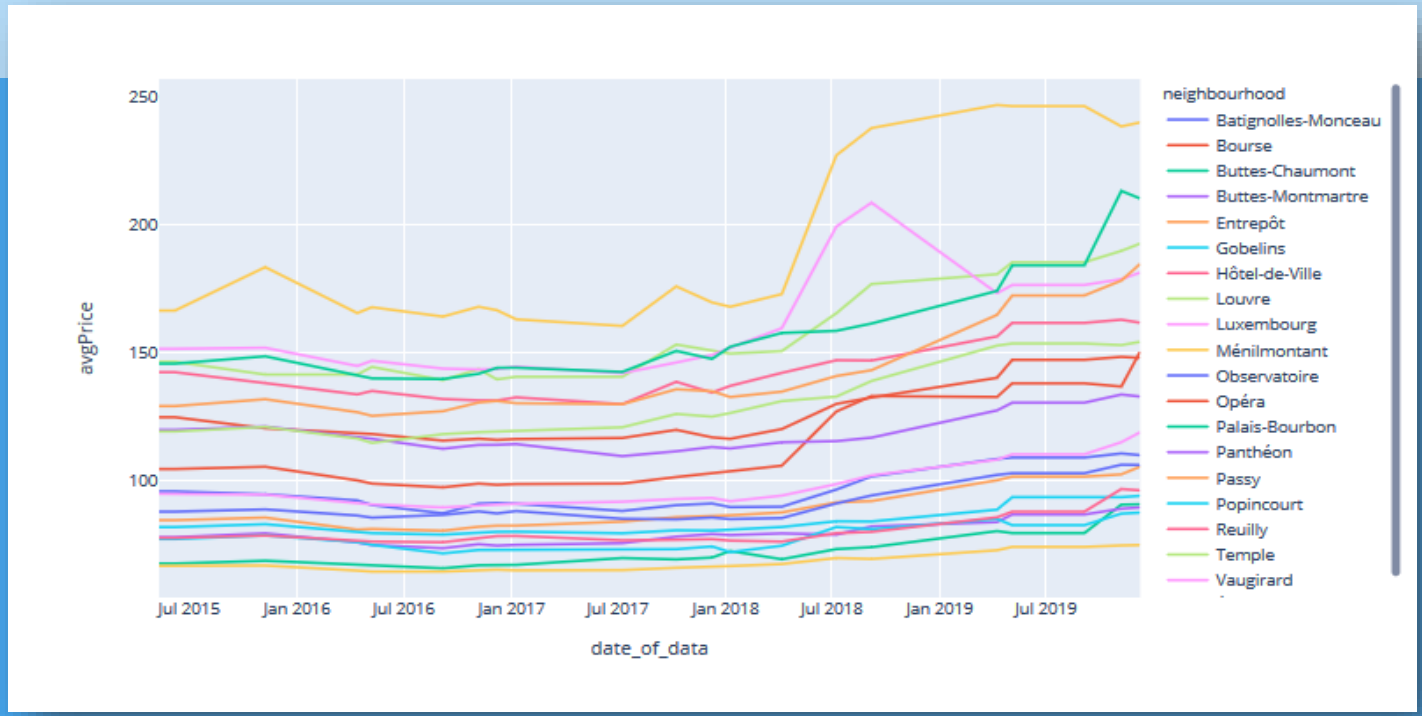


AIRBNB EXPLOSION- REVIEW 23 DATA SETS FROM 2015 TO 2019 TO DETERMINE PRICE INCREASES



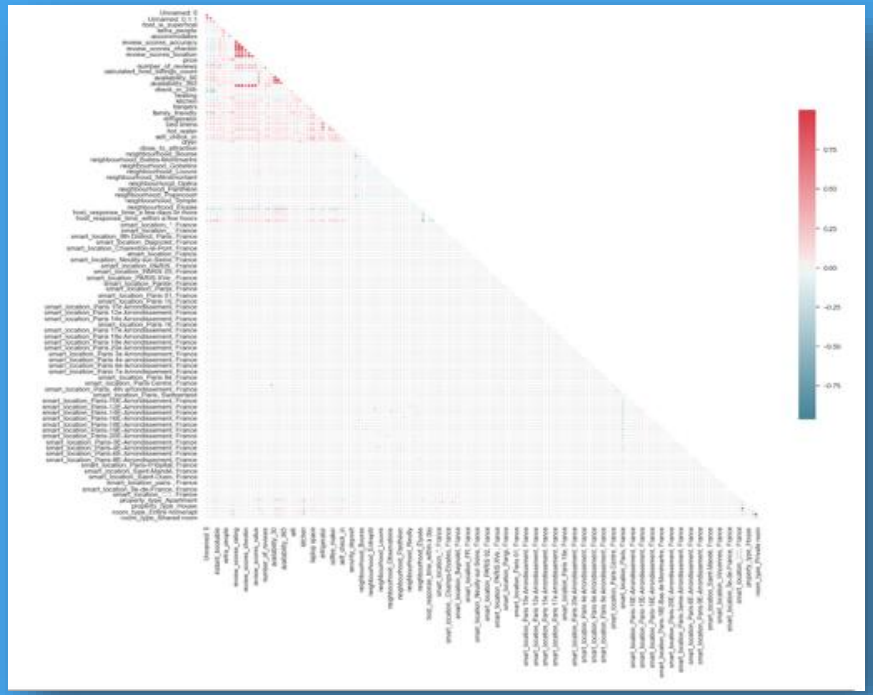
We notice immediately the growth of AIRBNB locations since June 7, 2015 from 35,428 to 66,212 in September 5, 2019. This is a 115% increase in just 4.5 years.

20 DIFFERENT PRICE MOVEMENTS



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

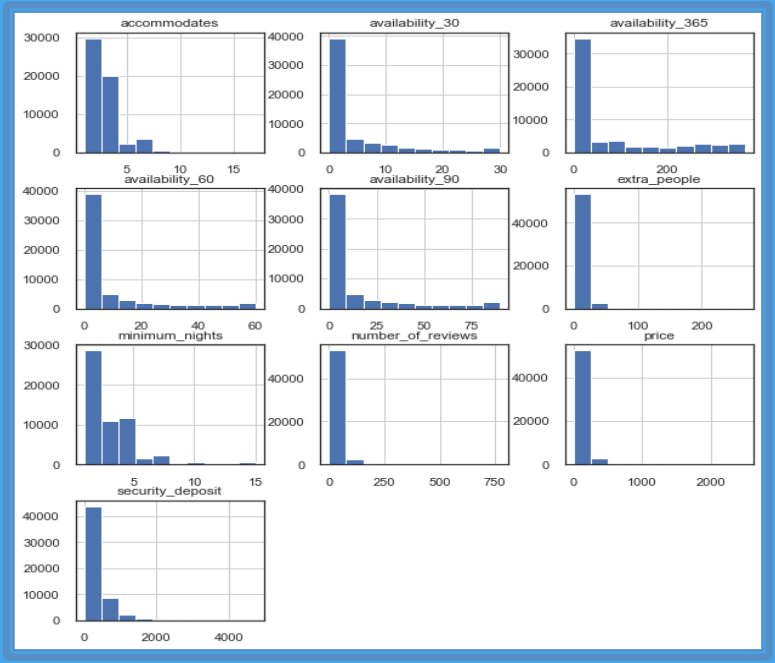
MULTICOLLINEARITY – REDUCING THE NOISE

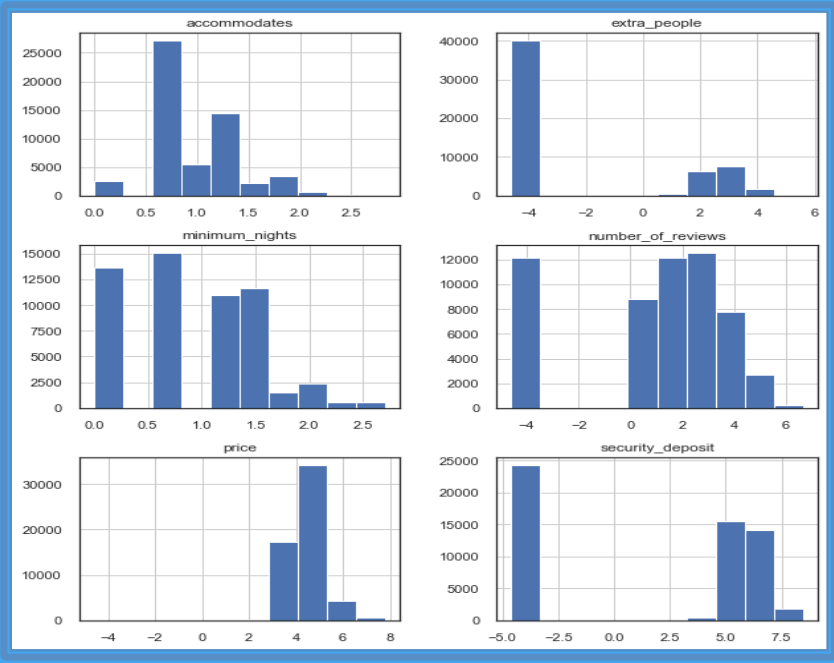


Let’s Review our Variables and determine which ones need to be removed and I decide on removing both neighborhood and re-view_score accuracy to reduce noise in my data.

AIRBNB EXPLOSION

Unlogged numeric variables changed to logged variables to adjust to a normal distribution





THE 88 VARIABLES FOR 2 MODELS

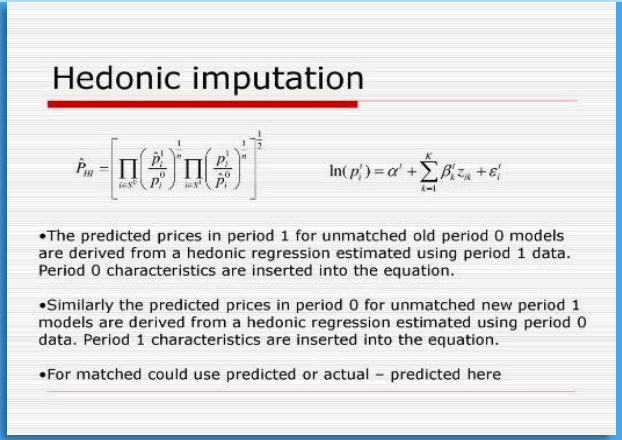
Variables used for my 2 Models				
host_is_superhost	guests_included	extra_people	accommodates	review_scores_cleanliness
review_scores_communication	review_scores_location	review_scores_value	minimum_nights	number_of_reviews
calculated_host_listings_count	availability_30	availability_60	availability_90	availability_365
wifi	heating	washer_dryer	kitchen	Toiletries
laptop space	family_friendly	fire protection	refrigerator	elevator
coffee_maker	hot_water	tv	self_check_in	hair_dryer
security_deposit	close_to_attraction	host_seniority_1 year	host_seniority_2 years	host_seniority_3 years
host_seniority_5 years	host_seniority_6 years	host_seniority_7 years	host_seniority_>= 8 years	host_response_ti
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_90-99%	neighborhood_Batignolles-Monceau	neighborhood_Bourse	neighborhood_Buttes-Chaumont	neighborhood_Buttes-Montmartre

THE 88 VARIABLES FOR 2 MODELS

Variables used for my 2 Models				
neighborhood_Entrepôt	neighborhood_Gobelins	neighborhood_Hôtel-de-Ville	neighborhood_Louvre	neighborhood_Luxembourg
neighborhood_Ménilmontant	neighborhood_Observatoire	neighborhood_Opéra	neighborhood_Palais-Bourbon	neighborhood_Panthéon
neighborhood_Passy	neighborhood_Popincourt	neighborhood_Reuilly	neighborhood_Temple	neighborhood_Vaugirard
neighborhood_Élysée	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_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		

MODEL 1: SPATIAL HEDONIC PRICE MODEL (HPM)

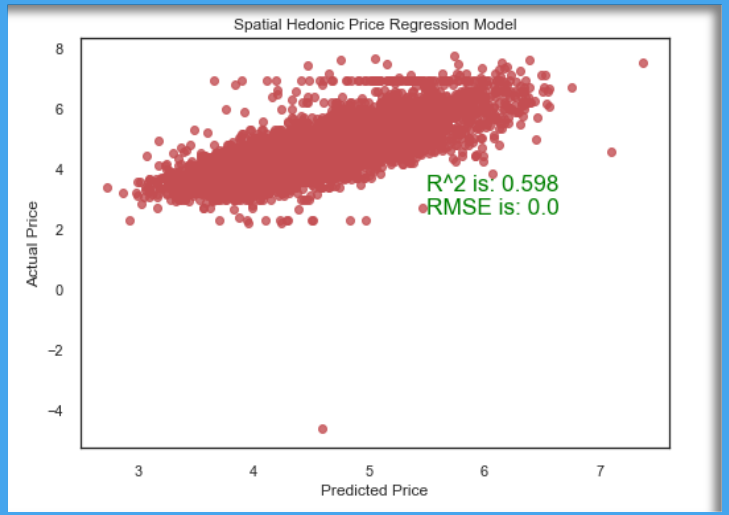
The “Hedonic” model involves regressing observed asking-prices for the listing against those attributes of a property hypothesized to be determinants of the asking-price. It comes from hedonic price theory which assumes that a commodity, such as a house can be viewed as an aggregation of individual components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



We are only 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.

MODEL 1: SPATIAL HEDONIC PRICE MODEL (HPM)

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, etc.



These are my results for my untuned model

Training RMSE: 0.1677

Validation RMSE: 0.1637

Training r2: 0.5968

Validation r2: 0.598

I ran my model a 2nd time using 5 levels of alpha [.01,.1,1,10,10] but received the same outcome.

MODEL 1: SPATIAL HEDONIC PRICE MODEL SUMMARY REPORT

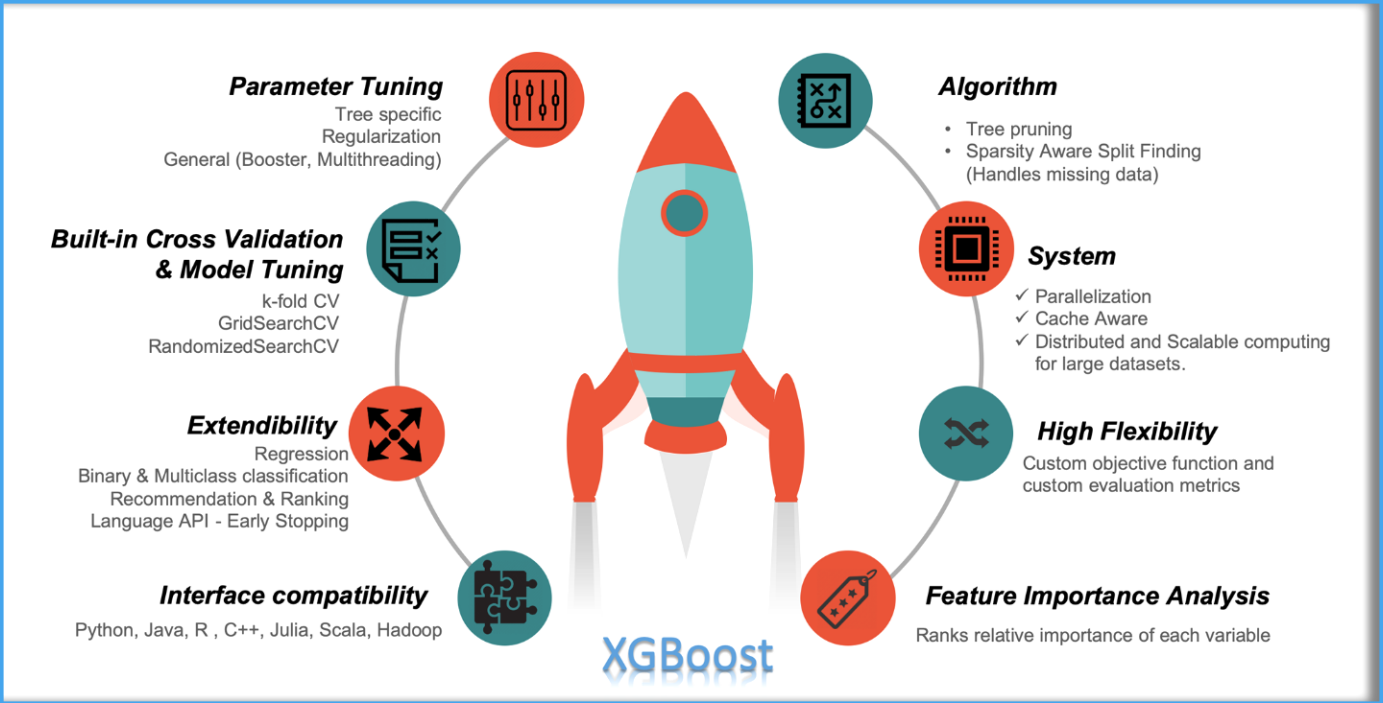
Dep. Variable:	**price**	**R-squared (uncentered):**	**0.016**
Model:	OLS	Adj. R-squared (uncentered):	0.011
Method:	Least Squares	F-statistic:	3.357
Date:	Tue, 30 Jun 2020	Prob (F-statistic):	5.51e-22
Time:	08:27:53	Log-Likelihood:	-49350.
No. Observations:	16909	AIC:	9.886e+04
Df Residuals:	16829	BIC: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

We review the contributing factors to price along with our related scores.

	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

HSP MODEL FEATURE REPORT: WE SEE ACCOMMODATES, AVAILABILITY IN 60 DAYS, REVIEW SCORES FOR CHECK IN AND LOCATION, AND WASHER/DRYER ARE AMONG THE TOP 5 FOR IMPORTANCE IN PRICING

MODEL 2: XG BOOST – LET'S DIVE INTO IT



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. XGBoost parallelizes the sequential process of generating trees.

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.

MODEL 2: XG BOOST – TUNING THE PARAMETERS

Since my initial XG Boost run with default parameters did not return strong results I decide to tune my model. 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

MODEL 2: XG BOOST – UNTUNED MODEL

The top 10 most important features are:

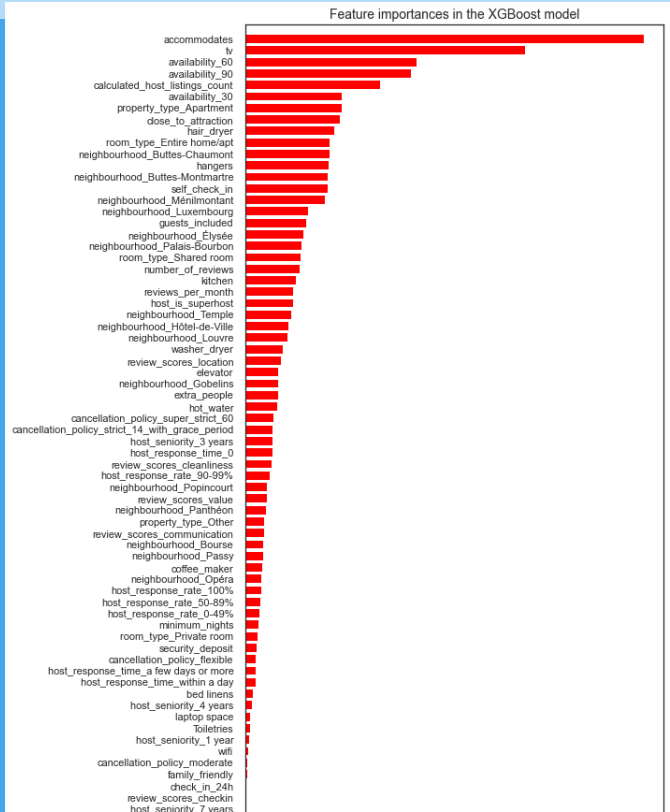
- How many people the property accommodates
- A TV in the rental
- How many days are available to book out of the next 365, 60,30,90
- How many other listings the host has (and whether they are a multi-listing host)
- Has a dryer
- Self check in
- Room Type of Entire Apartment
- Close to Attraction
- Neighborhood of Menil Mentmontant
- Guests included

It is not surprising that the most important feature is how many people the property accommodates, as that's one of the main things you would use to search for properties with in the first place. It is also not surprising that features related to location and reviews are in the top ten.

Training MSE: 0.1518 Training r2: 0.635

Validation MSE: 0.1522 Validation r2: 0.6262

or "The score indicates that my model predicts 62% of the variations in price."
The MSE is how close the fitted line is to the data points so the lower the better.



MODEL 2: XG BOOST – THE TUNED VERSION

I decide to tune my model using the following parameters suggestions from my 3 fold Cross Validation run:

- ✓ 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

MODEL 2_A: XG BOOST – TUNED VERSION

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
- How many days are available to book out of the next 365, 60,30,90
- How many other listings the host has (and whether they are a multi-listing host)
- Has a dryer
- Self check in
- Room Type of Entire Apartment
- Close to Attraction
- Guests included

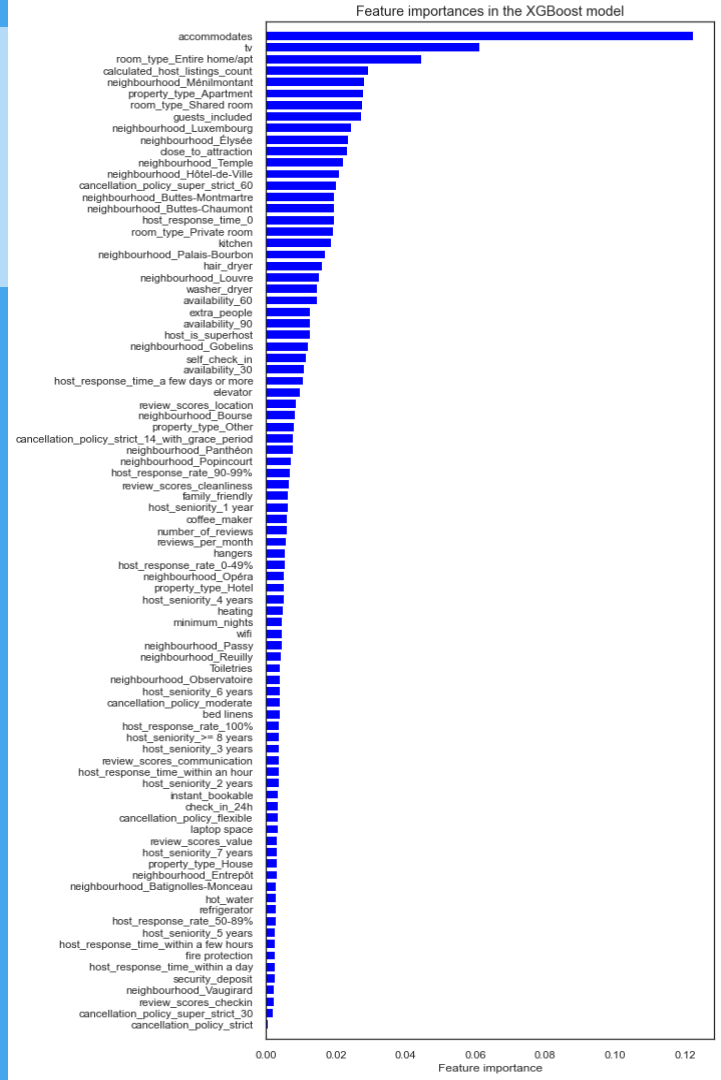
Training MSE: 0.0962

Validation MSE: 0.1325

Training r2: 0.7688

Validation r2: 0.6747

.



XG BOOST ELI5 INTERPRETER

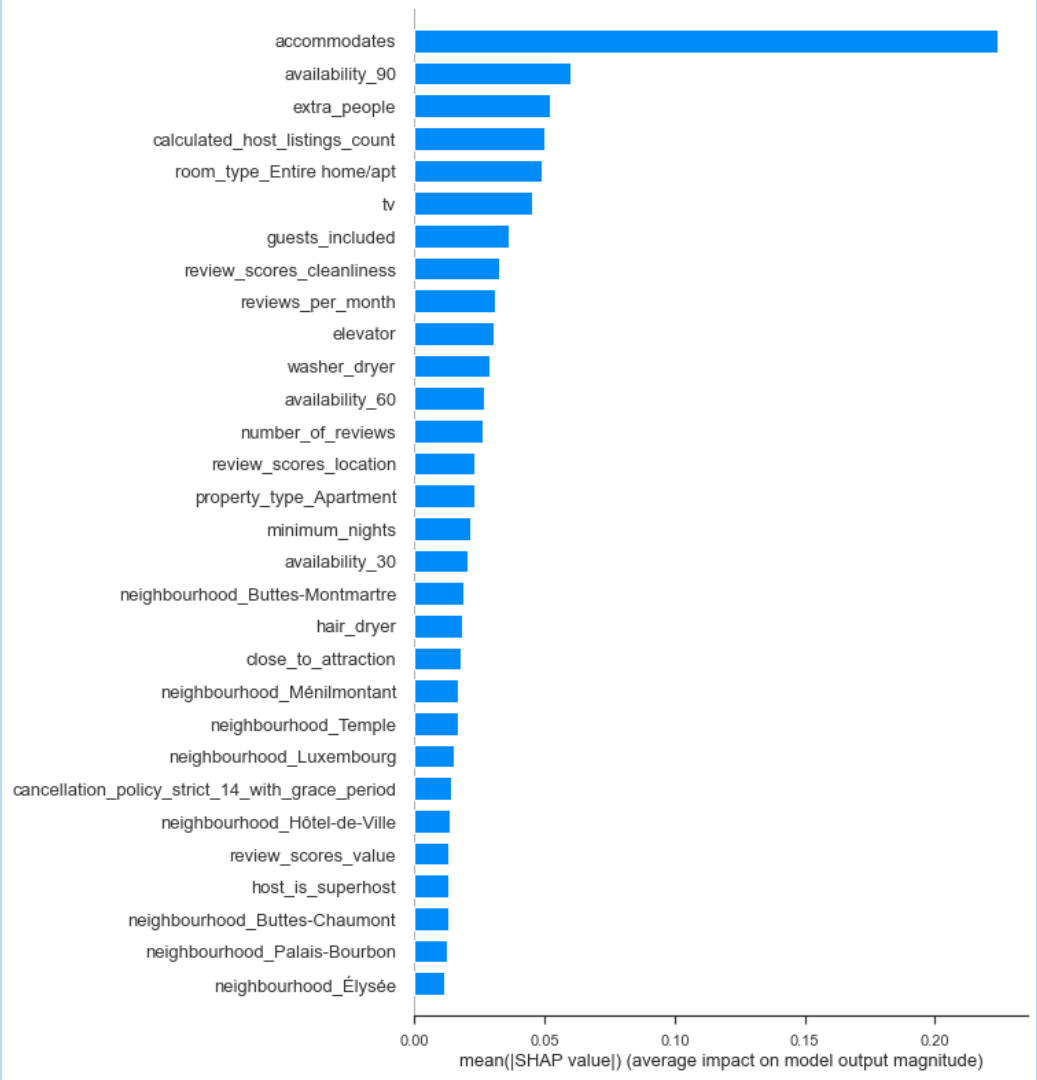
ELI5 does this by showing weights for each feature depicting how influential it might have been in contributing to the final prediction decision across all trees. The idea for weight calculation is described **here**; ELI5 provides an independent implementation of this algorithm for XGBoost and most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

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	neighborhood_Ménilmontant
0.0244	close_to_attraction
0.0240	neighborhood_Luxembourg
0.0220	neighborhood_Hôtel-de-Ville
0.0219	neighborhood_Élysée
0.0214	neighborhood_Temple
0.0202	neighborhood_Palais-Bourbon
0.0192	availability_90
0.0189	neighborhood_Buttes-Montmartre
0.0182	neighborhood_Buttes-Chaumont
0.0173	cancellation_policy_super_strict_60
0.0171	self_check_in
0.0162	neighborhood_Louvre

THE SHAP INTERPRETER

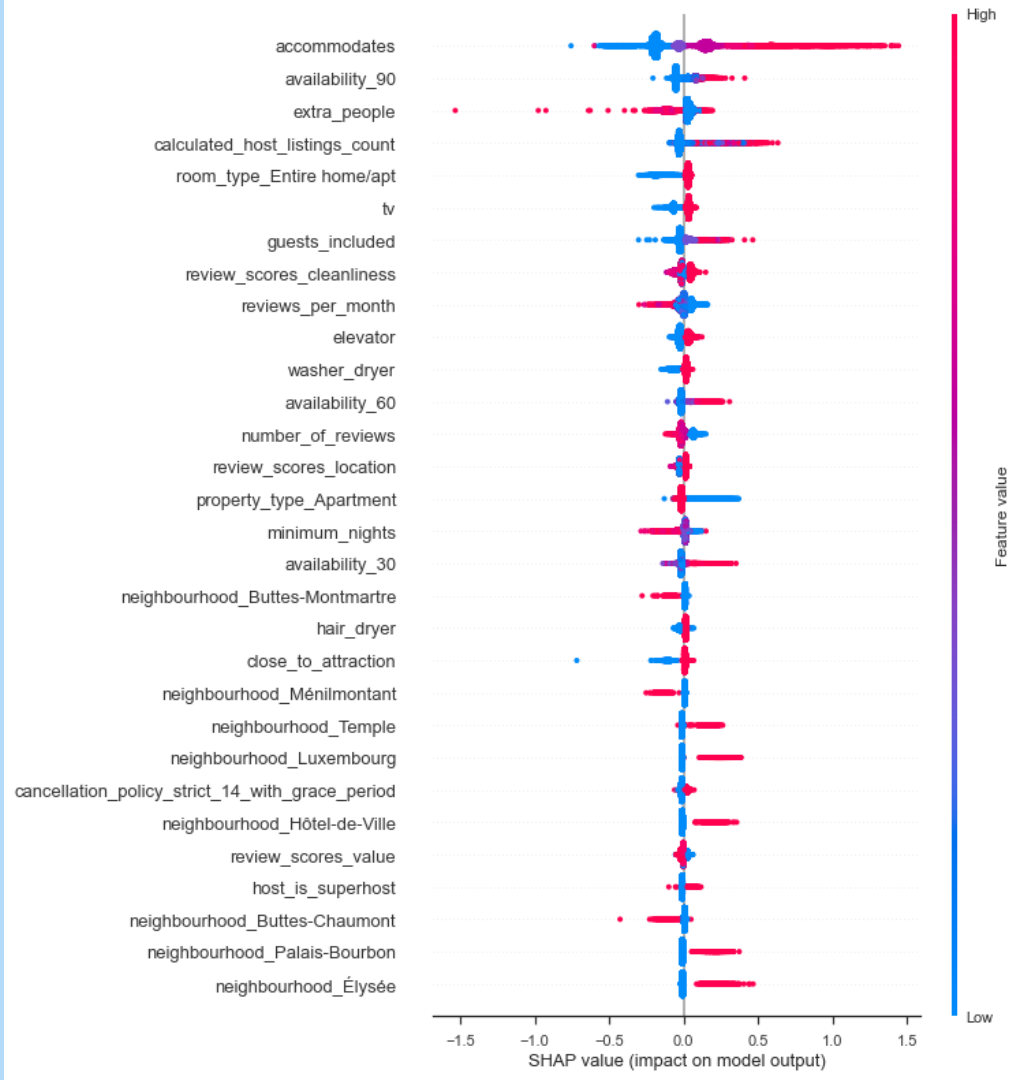
SHAP (Shapley Additive explanations) by Lundberg and Lee (2016)[41] is a method to explain individual predictions. *SHAP* is based on the game theoretically optimal Shapley Values. ... Second, *SHAP* comes with many global interpretation methods based on aggregations of Shapley values.

Left: The SHAP interpretation of feature impact- it places the neighbourhoods toward the bottom



SHAP SUMMARY MODEL

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 above 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).



SHAP FORCE PLOTS

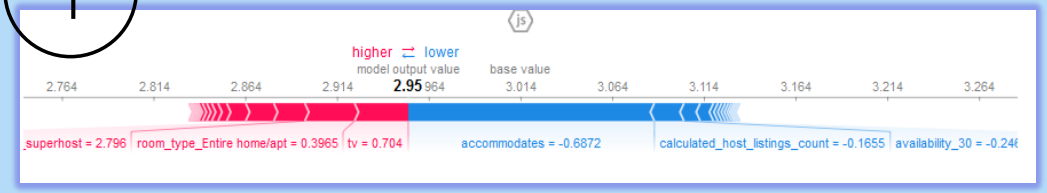
SHAP force plot can be used to explain individual predictions.

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 top feature is quantified on the prediction with local accuracy. This particular 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 70 USD and if we take the log base 5 of 70 we get 2.63 so we are close in our forecast as shown above.

Ex 2 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.

Ex 3The 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.

1



2

3

NULL HYPOTHESIS:

I will state this as 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 a 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 **“Properties near the Top 10 Parisian attractions are a determining factor of price”**

NULL HYPOTHESIS:

The contingency_tables below are comparing rentals near the TOP 10 and those NOT near the TOP 10

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

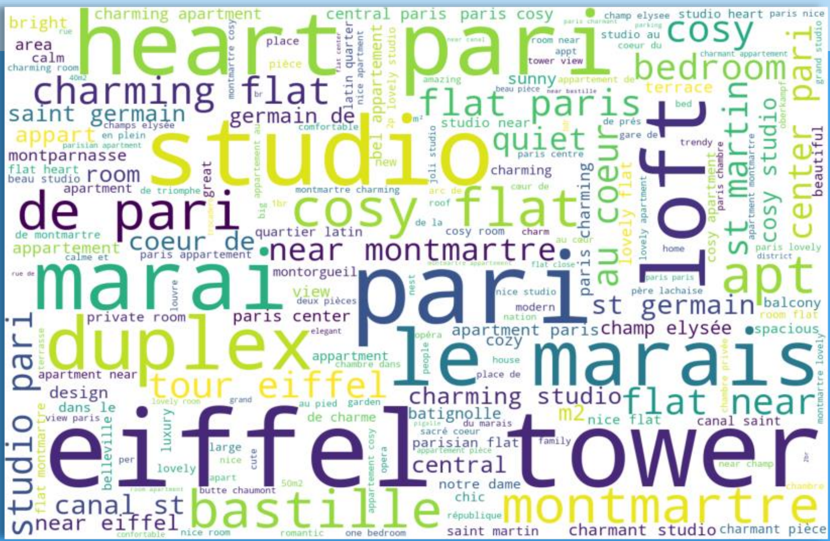
Reject H0,There is a relationship between 2 categorical variables

Reject H0,There is a relationship between 2 categorical variables

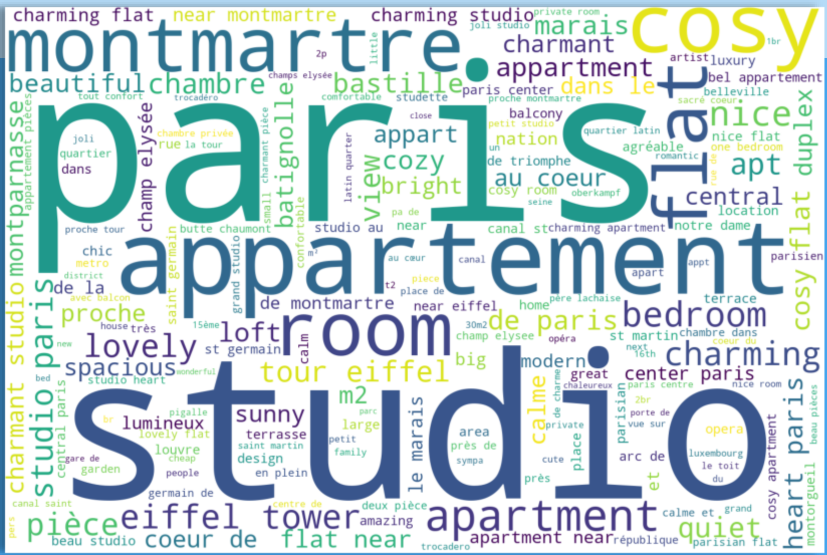
TEXT ANALYSIS

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/

WHAT WORDS MATTER



The 1st wordcloud is to review the most word usages for the top 10% of ratings of our list. This tells us people loved the location!



The 2nd wordcloud is to review the most word usages for the bottom 10% of ratings of our list.This tells us people spoke more about the type of listing..a studio, an apartment, etc

CONCLUSIONS

Through this exploratory data analysis and visualization project, we gained several interesting insights into the Airbnb rental market. Below we will summarize the answers to the questions that we 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

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 a 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.

CONCLUSIONS

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 neighborhood tends 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.

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 slide 67 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 description 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.

FUTURE ANALYSIS

We want to expand our analysis to multiple European cities and compare patterns and trends amongst these cities.

Better utilize Shap interpretations to gain further insights and tune our model

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. Perhaps a current prospective or current AirBnb host can utilize a refined model to replace a paid pricing service.

LIMITATIONS

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 business people they provided insight into certain aspects of my project as well as to the lack of verification for AIRBNB hosts and the like.

We also learned how to effectively deploy Github and other version control systems while working on this project.