MODS 203 REPORT

Data Analysis in economics :

Collection and visualization

Nicolas Jow, Tiffany Ly, Rémy Masbatin, Caroline Xia

INTRODUCTION

With the recent advent of the Internet and new technologies, companies of a new kind have emerged. They created a new type of economy that almost all Internet users are now part of : the sharing economy. This economic system is based on people sharing possessions and services, for free or not, and is usually organized through the Internet. Created in 2008, Airbnb is one of the most important companies of the sharing economy. As more and more people are travelling around the world, whether for vacation or business trips, the amount of users of this platform is growing exponentially, and so does the data that can be collected on the website. The analysis of these data would enable us to understand the different factors that set guidelines for lodgings' prices. This study would also highlight the mechanisms of this new economy of the Internet, which is likely to be the number one predominant economy in the future.

The objective of our project is to understand how the users define the price of their apartment or their house on the platform. Our team aims to find a correlation between prices of the lodgings and other factors, such as the number of rooms, beds, bathrooms, etc. But we also wish to determine if there is an influence of more complex variables, such as the number of comments, the ratings, the localisation, and more ethical variables like languages spoken by the host.

These variables are all the data we collected with our algorithm. We first ran a study using the coordinates of the lodgings so we could find whether the proximity to certain monuments had an impact on the price. In order to find a correlation between these and the price of each lodging, we used tables and a linear regression to illustrate what influence these characteristics of the lodging have on their price setting. In order to get the precise correlation coefficients, we used the OLS method in Python. The goal was to underline whether the price was globally impacted by some of these criteria.

BACKGROUND

Airbnb has been increasingly gaining popularity since 2008 due to its low prices and direct interactions with the local community. In the literature, we found the identification of the important variables that significantly influence room pricing on the Airbnb rental platform. It was also found that pricing depended on each city and its specificities. For Paris, a tourist city, the proximity of monuments or museums can play an important role.

Airbnb is an online platform that allows people to rent short term accommodation. This ranges from regular people with a spare bedroom, to property management firms who lease multiple rentals. However, the platform has expanded by partnering with car rental services, restaurants, entertainment and tour sites, among others to become an all-in-one travel site. They brand themselves as an online 'travel community', allowing guests to have a local experience in exotic locations. The platform itself is simple to use : you just have to type the city you want to visit, your travel dates, and your search result comes up.

In 2019, the market share of Airbnb on the market of online booking represented 23% of the available market, which makes it an important actor of this industry. This is due to the fact that Airbnb was the pioneer of online booking for private landlords. The platform's revenue has been drastically increasing since its creation, from $7 millions of annual revenue in 2010 to $900 millions in 2015, and $4.7 billions in 2019.

Airbnb is one of the most successful examples of the sharing economy platform, but is often criticized by regulators and policy makers. While, in theory, municipalities should regulate the emergence of Airbnb through evidence-based policy making, in practice, they engage in a false dichotomy: some municipalities allow the business without imposing any regulation, while others ban it altogether. That is because there is no evidence upon which to draft policies.

DATA COLLECTION STRATEGY

We collect our data directly via the website of Airbnb, and our study focuses on one city, Paris, where the demand and the supply are the highest in France. Using a web scraping method, we collect the price, but also other features and information related to the rooms and the hosts. For that, we use BeautifulSoup, requests, lxml and selenium packages in Python.

First, we use selenium in order to set some filters for our lodgings searches. We created a filtering function mimicking a user clicking on the different filters and options available on Airbnb, thus allowing the number of beds, bedrooms, or bathrooms to vary for example. It can also choose the type of lodging : an entire lodging, a private room, a hotel room, or a shared room. But the most important feature of that filtering function is that it enables us to type a minimal and maximal price. Since Airbnb does not allow the user to see more than 300 results in one search, by filtering a lot, we can have more precise searches and thus have more lodgings in our database -- and the price is the most discriminating filter to change. That filter function is used in a loop so we do not have to change the filters by ourselves: that way the filtering was entirely automated, and we can do searches with any filter combination possible .

Once the filtering is done and the main research page is loaded, the code’s “clicks” open the first displayed lodging page on a new tab, through the main research page. When the lodging page is loaded, our code combines BeautifulSoup, requests and lxml in order to harvest all the data we are interested in. We have to make sure the time.sleep is long enough so the page is correctly loaded ; and we also have to make the code scroll the page in order to get some of the information that only appears when the end of the page is reached, such as the coordinates of the lodging. Then, the algorithm closes the tab, and does the same for the second displayed lodging, the third, … until it reaches the end of the page. When all of the interesting information in a page is scraped, the code clicks on the next page, and starts all over again.

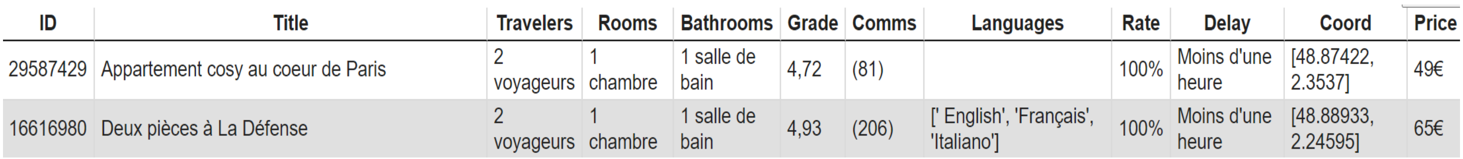
We met some major issues while scraping the data. For instance, position and names of buttons to filter the page would periodically change during the web scraping, causing crashes. Besides, the ID of classes periodically changes, so we had to update our code several times. Also, Airbnb tried to prevent us from scraping multiple times, by IP banning our laptops and therefore denying access to their website. Sometimes they also required us to log in the page before we could see the lodging, also causing the code to crash.

Finally, once the data is scraped, we have to clean it. We use Pandas library to manipulate our data. Since most of the content in the DataFrame are strings, we have to convert them to integers or float so the data is usable. In addition, we have to be careful and delete any duplicates so the analysis will not be erroneous later.

We chose an arbitrary night (18th to 19th of February), and scraped the available lodgings at this date. It is difficult to say how many Airbnb lodgings are currently available : the website only indicates when there are more than 300 lodgings available in the area. According to the website insideairbnb.com, there are about 60 000 Airbnb lodgings in Paris, however this website also counts lodgings that are no longer available. Moreover, with the current COVID-19 epidemic, a great part of the demand -- and thus, offer -- has collapsed, making it harder to find diverse lodgings.

DATA DESCRIPTION

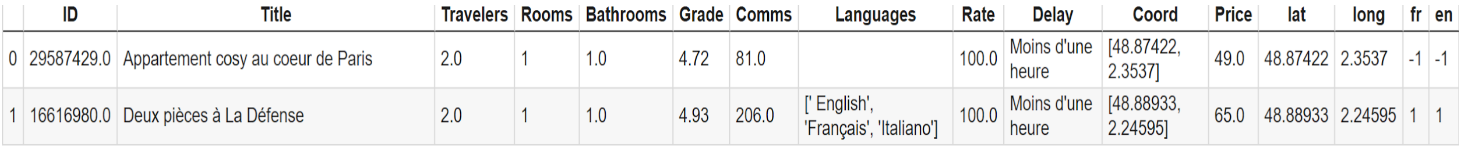
In the end, we managed to scrape a dataset of 3990 units. Raw units, before the cleansing phase, are presented like this :



ID	Title	Travelers	Rooms	Bathrooms	Grade	Comms	Languages	Rate	Delay	Coord	Price
29587429	Appartement cosy au coeur de Paris	2 voyageurs	1 chambre	1 salle de bain	4,72	(81)		100%	Moins d'une heure	[48.87422, 2.3537]	49€
16616980	Deux pièces à La Défense	2 voyageurs	1 chambre	1 salle de bain	4,93	(206)	[' English', 'Français', 'Italiano']	100%	Moins d'une heure	[48.88933, 2.24595]	65€

Fig.1 : A raw sample of our dataset

There are twelve features for each lodging : an ID in order to delete the duplicates, the title of the lodging given by the host, the number of travelers it can welcome, the number of rooms and bathrooms, the grade, the number of comments, the languages spoken by the host, the rate and delay at which they answer the comments, the coordinates of the lodging, and finally the price. When one of these could not be scraped, the data is NaN. After cleaning the dataset, units are presented like this :



	ID	Title	Travelers	Rooms	Bathrooms	Grade	Comms	Languages	Rate	Delay	Coord	Price	lat	long	fr	en
0	29587429.0	Appartement cosy au coeur de Paris	2.0	1	1.0	4.72	81.0		100.0	Moins d'une heure	[48.87422, 2.3537]	49.0	48.87422	2.3537	-1	-1
1	16616980.0	Deux pièces à La Défense	2.0	1	1.0	4.93	206.0	[' English', 'Français', 'Italiano']	100.0	Moins d'une heure	[48.88933, 2.24595]	65.0	48.88933	2.24595	1	1

Fig.2 : Clean sample of our dataset

So that the number of travelers, rooms, bathrooms, grades, comments, rate, and price are integers or floats and not strings. Further columns that are not shown here, such as booleans according to whether a host spoke a certain language or not, or the arrondissement that we calculated thanks to the coordinates, were also added for analysis purposes.

We can summarize the most important numerical features of our dataset with the following table:

	Mean	Standard deviation	Median
Price (€)	122.5	156.6	75.0
Travelers	3.4	2.6	2.0
Grades (out of 5)	4.6	0.4	4.67
Comments	47.3	77.7	16.0

Fig.3 : Descriptive table of some of our features

In particular, the mean price is 122.5€, however the median is 75€ and the standard deviation relatively high in comparison to the mean. It means that there are few expensive lodgings, however those who are are very much so. The same goes for comments : there are a lot of lodgings without comments, and there is a handful that have a paramount number of comments. The grades, though, seem pretty centered around 4.6.

ANALYSIS

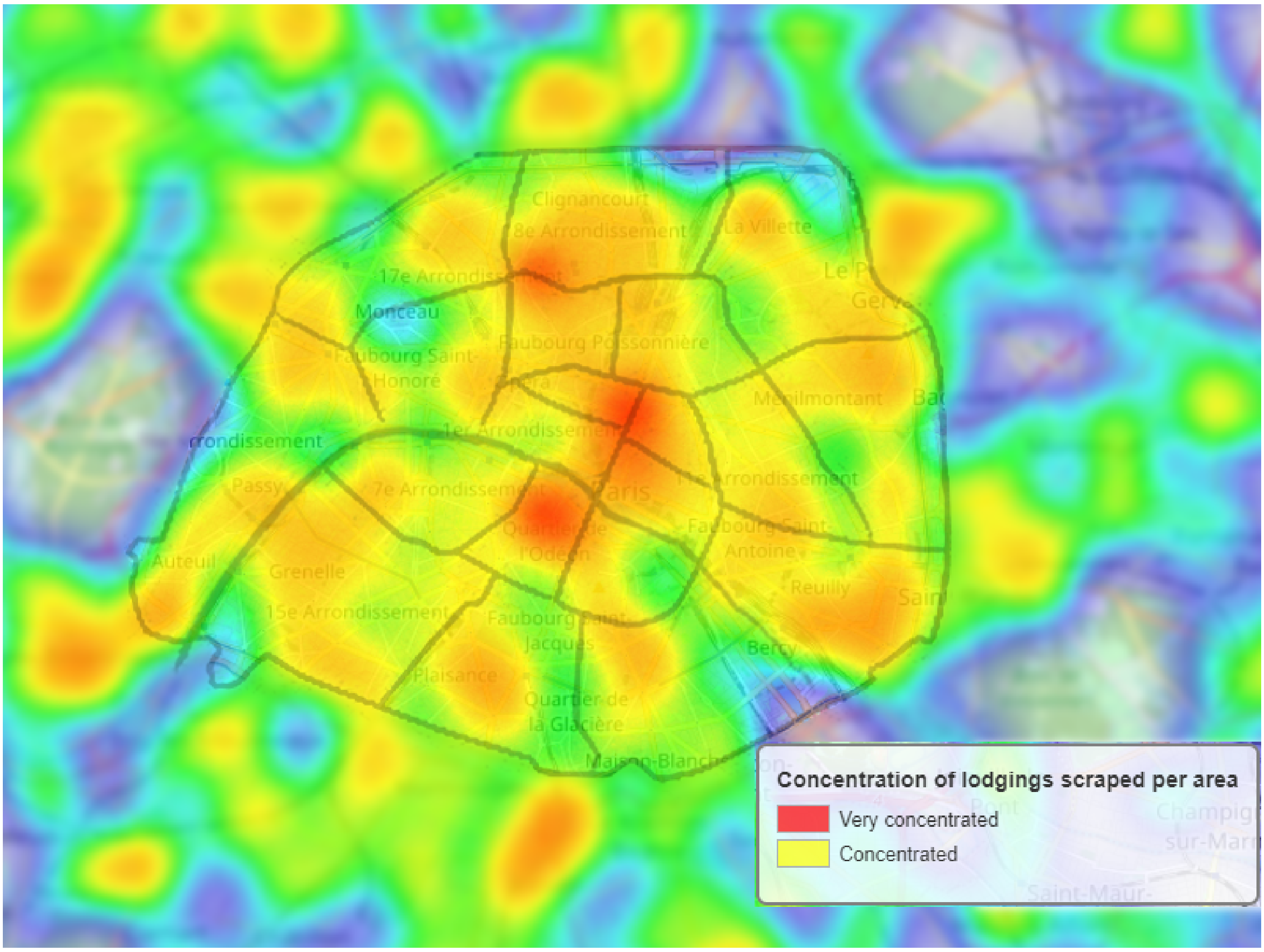


Fig.4 : Repartition of scraped lodgings in Paris

This heatmap of Paris depicts the concentration of lodgings offered on Airbnb that we scraped. We can see that there are areas very concentrated in lodgings, such as the 6th arrondissement, also known as Quartier Latin (Quartier de l'Odéon) , or in the first and second arrondissement near Le Louvre. Those areas are full of places for tourists and therefore a lot of people rent their lodgings there. People are more likely to rent their lodging as there surely will be a lot of demands.

The heatmap shows that some areas are more represented than others. We also wanted to take a look at any other missing data so that we get a better insight of the data we scraped.

	Travelers	Rooms	Bathrooms
1	10%	78%	88%
2	43%	13%	9%
3	9%	4%	1%
4	22%	2%	1%
5+	16%	3%	1%
Total	100%	100%	100%

Fig.5 : Repartition of the different features within the Airbnb lodgings in Paris

This table presents the proportion of lodgings that can accept 1 to 5 travelers, has 1 to 5 rooms and 1 to 5 bathrooms. The first observation that is really striking, is that the huge majority of the lodgings only have one room and one bathroom. This statement meets our expectations, as most apartments in Paris are very small because of the density of the population. Only 12% percent of the Airbnb lodgings offered have 2 bathrooms or more, and only 9% have more than 2 rooms. The vast majority of the lodgings offered on Airbnb in Paris are not houses, but rather small flats.

Concerning the number of travelers admitted in each lodgings, almost 50% of them accept 2 travelers. As Paris is reputed for being one of the most romantic cities in the world, this is not surprising that most lodgings are suited for two people, who are generally couples. Furthermore, more than one third (38%) of the flats allows more than 3 people to stay, whereas, only 10% only accept 1 traveler maximum. This shows that Paris aims to attract couples or families, more than lone travelers.

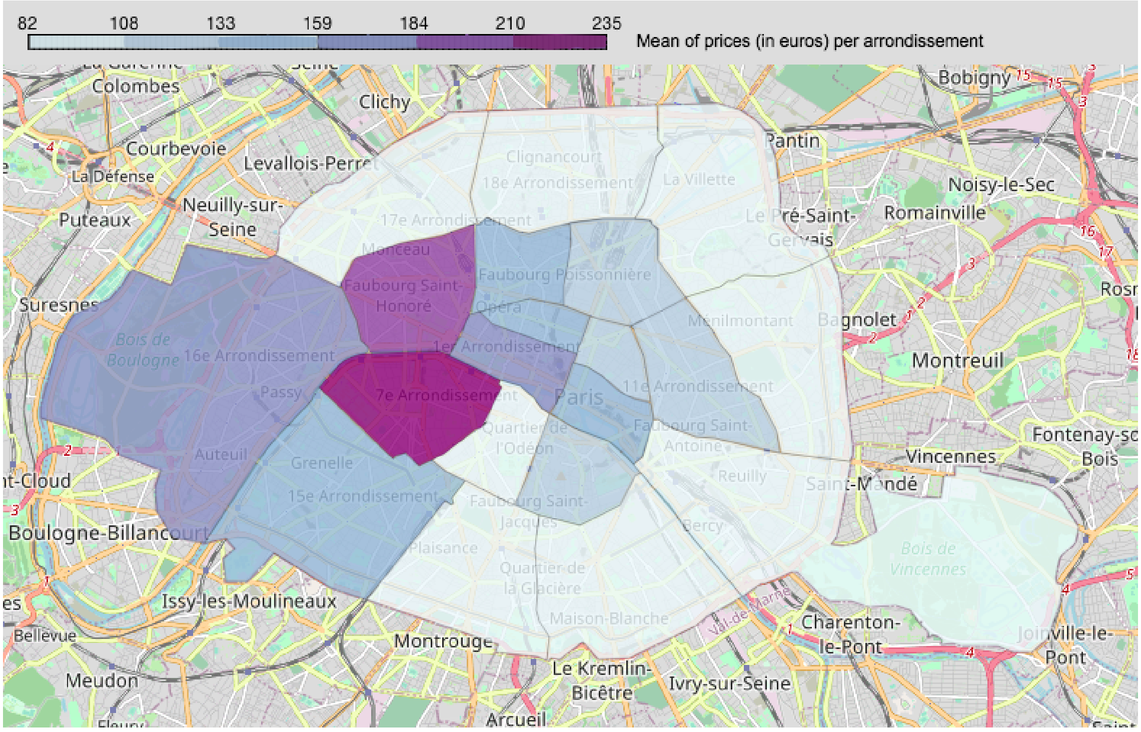


Fig.6 : Average prices per district

This second map completes the first one, by showing the average price in euros of all the lodgings for each district. The cheapest districts (from 82€ per night) are colored in a very pale blue, whereas the expensive ones (up to 235€ per night) are colored in purple .

We can notice here that the most costly district is the 7th, which is an expected result, as it is where the Eiffel Tower is located. The second most expensive one is the 8th, just above the 7th. Here, the prices can be explained by the presence of the world-known avenue : the Champs-Elysées.

This map also clearly points out that the cost per night for the lodgings decreases progressively as we move away from the center of Paris. Even with the presence of famous places and monuments, such as the Sacré Coeur, the 17th and the 18th district (the two most northern arrondissement of Paris) are far cheaper than the 7th. This way, we can compute that the 7th district is 286.6% more expensive than the cheapest district of Paris on average.

As a consequence, we can infer from that observation that lodgings that are near to a very famous monument, at a global scale, can have their price set almost three times higher than others.

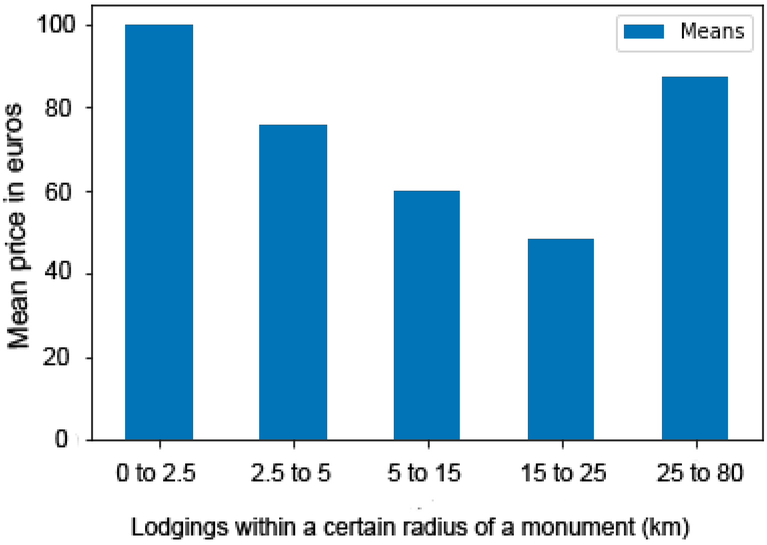


Fig.7 : Mean prices of the lodgings within a certain radius of a monument

Through that graphic, we can indeed show the impact on price when having a famous monument in the surroundings of the lodging. Here, we took some of Paris' most important monuments : the Tour Eiffel, Notre-Dame de Paris, the Arc de Triomphe, and Montmartre, and we calculated the distance between each lodging and the closest of these four monuments. For this graphic, we removed luxury lodging that we scraped, since they were very few but weighed a lot because of their very high prices and could erroneously influence our findings, however this study still has 3507 observations. We found that the closest we were to a monument, the more expensive the lodgings were. As the radius increases, the mean price decreases slowly, until it reaches a radius of 25 to 80 km. However, at these distances, we are no longer within Paris itself. Thus, the attractivity and price of these lodgings are not based on their proximity to the monuments, but other features like the size or a special equipment (swimming pool, jacuzzi...) for example, hence explaining why the mean price increases.

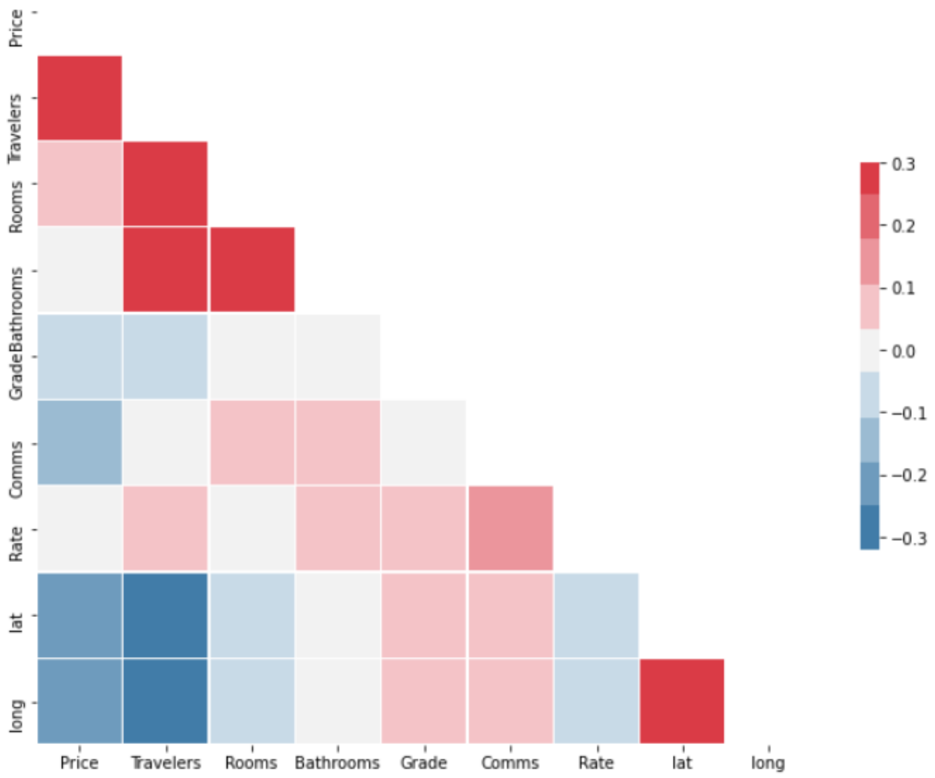


Fig.8 : Correlations between characteristics of the lodgings

This correlations chart points out what are the influences of the number of travelers, the number of rooms, the number of bathrooms, the grade, the number of comments, the rate of response, the latitude and longitude. The squares are colored red when the correlation coefficient is positive, and blue when negative.

It clearly appears that the more travellers are staying in the lodging, the higher they will have to pay for the stay. But there is also an interesting fact we noticed : although it is a weak negative correlation, it seems that the cheaper the apartment, the better the grades. Maybe low pricing lowers the expectations of the travellers, and thus makes them less demanding. This phenomenon could explain why the grades are higher for low-cost lodging. We can also notice that there will be no multicollinearity issues if we use those predictor variables, as they are not strongly correlated.

No. Observations : 977

Df Residuals : 968

Model : OLS

R-squared : 0.307

Adj. R-squared : 0.302

	Coeff	Std err	t	P>|t|
Intersect	1806	2396	0.75	0.451
Travelers	18.81	1.17	16.10	0.000
Bathrooms	-15.46	2.79	-5.54	0.000
Rooms	9.81	2.94	3.34	0.001
Answer rate	-0.03	0.12	-0.26	0.789
Comments	-0.1	0.03	-3.86	0.000
Grade	-0,36	6.76	-0.054	0.957
Latitude	-28.21	48.63	-0.58	0.562
Longitude	-161.1	32.77	-4.92	0.000

Fig.9 : OLS regression of the price with different features : number of travelers, bathrooms, rooms, answer rate, number of comments, grade, latitude and longitude

We can see thanks to an OLS regression whether the different features have a statistical significance over the price. The features tested are the number of travelers, the number of bathrooms, rooms, the answer rate, the number of comments and the grades, in that order in the OLS regression. We have 977 observations, as we had to remove all the units that had missing data that were sometimes not given on the website of Airbnb. The determination coefficient is equal to 0.29, meaning our model can explain 29% of the pricing variance (this model is not appropriate for predicting potential prices that meet certain criteria). Even if the R square is low, we can still draw important conclusions about how changes in the predictor values are associated with changes in the response value since we have statistically significant predictors.The p-value of the features "number of travelers", "number of bathrooms", "number of rooms" and "number of comments" are inferior to 0.05. Hence, these features are statistically significant to the price and indicates strong evidence that we can reject the null hypothesis. The rate of response and the grade, though, have a p-value superior to 0.05, meaning these features are not statistically significant to the setting of the price . In the end, we can see that the feature “number of travelers” has the most important coefficient at 17.80, and is positive. Thus, longitude is the feature that has the most important impact on the setting of the price : lodgings located in west areas are the most expensive ones. On the opposite, the number of comments has a coefficient of -0.1, meaning it has little ruling over the pricing. If we take a look at the intersect value (1806)

A further step to check the reliability of the model is to compute the residuals (differences between fitted values and the observed ones).Using residual plots, we can assess whether the observed error (residuals) is consistent with random error.

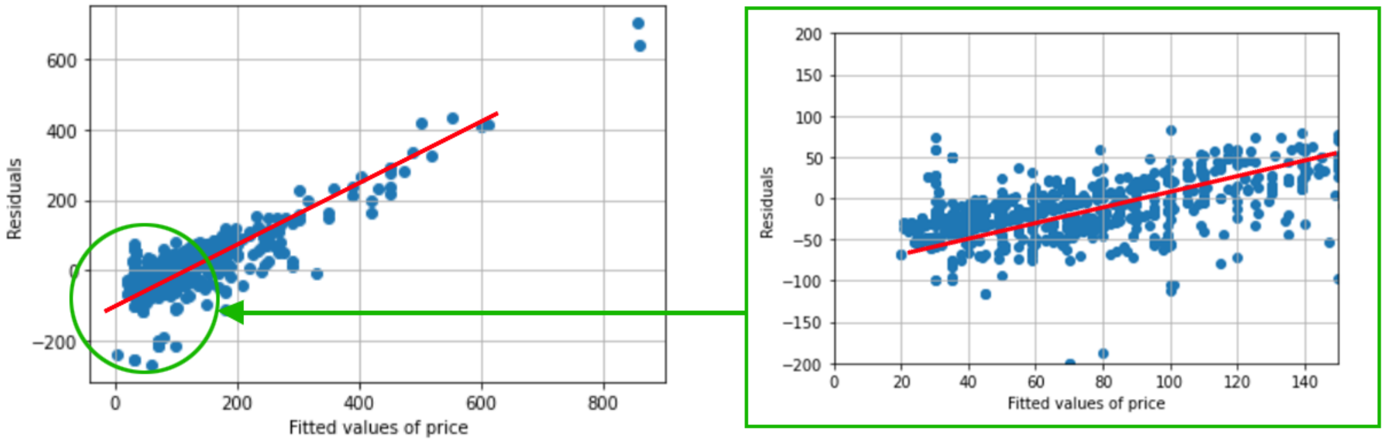


Fig.10 : Residuals of the previous OLS regression

The mean of the residual is$1.5 * 10^{-12}$: the residuals are centered on 0 (which is a good point). However, in the OLS context, random errors are assumed to produce residuals that are normally distributed. Therefore, the residuals should fall in a symmetrical pattern and have a constant spread throughout the range. The residuals should not include any predictive information. Though, we can see a clear tendency of residuals to go uphill as the price of lodgings increases.

The non-random pattern of residuals makes us think that the predictor variables of the model are not capturing some explanatory information that is "leaking" into the residuals. We think that there are two possibilities : maybe we are missing a variable, or perhaps we are missing an interaction between terms already in the model. Since variables are not strongly correlated as we saw before, we can draw the conclusion that it's likely that we miss a significant predictor variable to validate the model.

CONCLUSION

This project enabled us to understand the main mechanisms that rule the lodgings pricing on Airbnb, in an important metropolis as Paris.

First, the price highly depends on the localisation, whether there are monuments or famous places near the apartment. Especially in the center (and west-center) of the city, the prices become higher. The concentration of the lodgings offered on Airbnb also increases in touristic districts. Our team succeeded in finding correlations and dependencies between the different main characteristics of each lodgings, such as the price, the number of rooms and bathrooms, and the evaluation of the travellers.

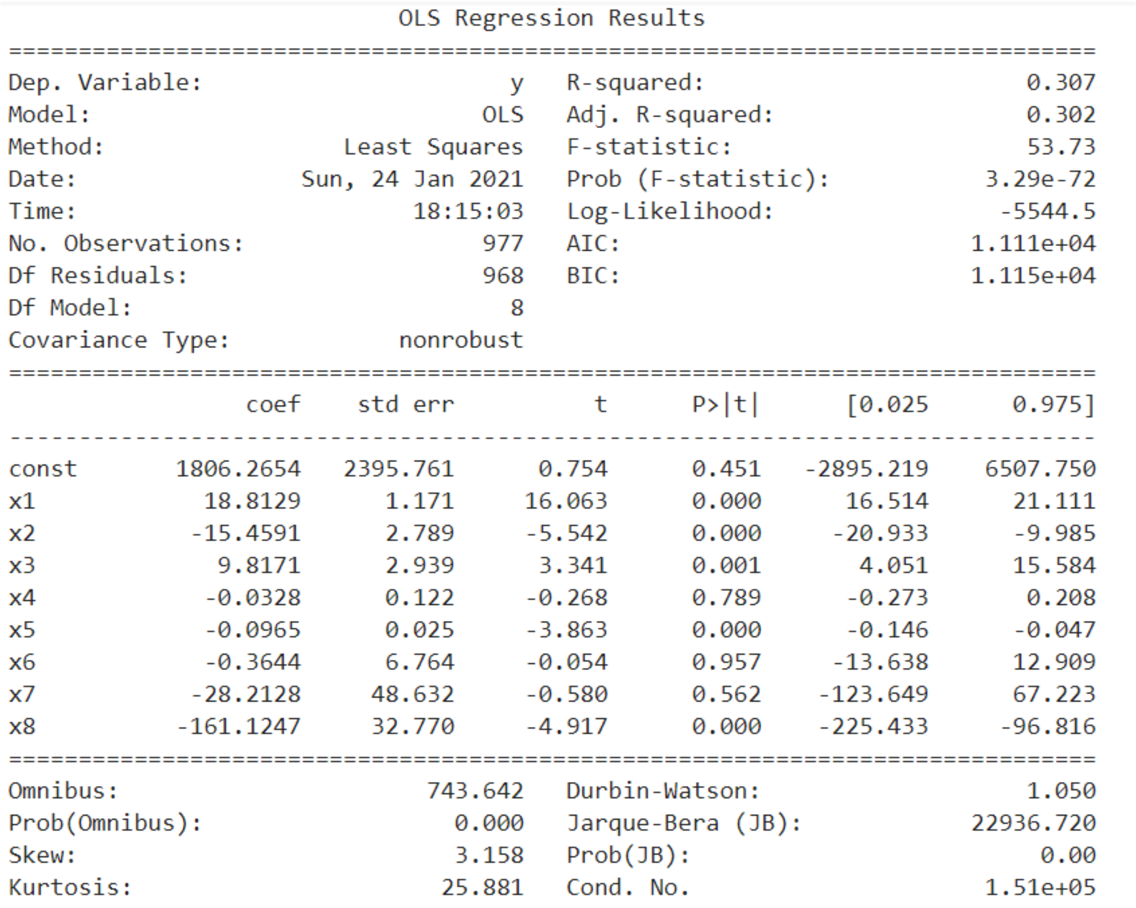
The models could have been improved by increasing the size of our database and also by retrieving more explanatory features (maybe the floor of the lodgings) for the price regression. The scraping took us a very long time, which has limited the amount of data we managed to harvest.

With more data, we could also make a deeper study on the influence of the ethnicity of the host over the price. We wanted to analyze precisely if a host who spoke certain foreign languages would set his price lower compared to the other hosts. But unfortunately, not enough lodgings in our database have a host who speaks these languages. We therefore didn't manage to make a legit analysis out of it.

REFERENCES

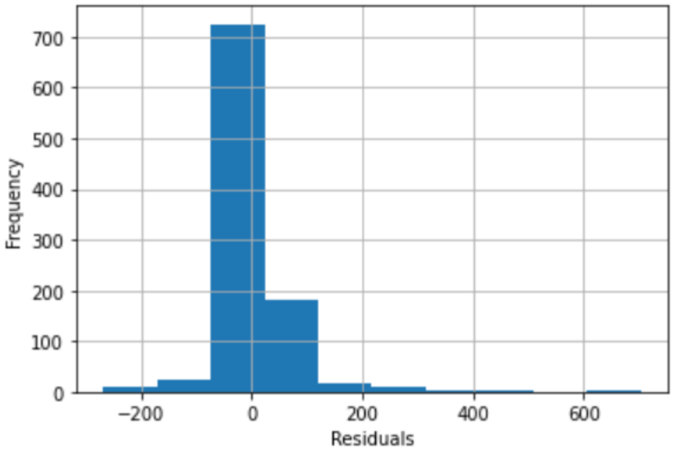
1. “*Who Benefits from the ‘Sharing’ Economy of Airbnb?*” Quattrone, Giovanni, et al. WWW ’16 Proceedings of the 25th International Conference on World Wide Web, Pages 1385-1394 (2016)
2. “*Do airbnb host listing attributes influence room pricing homogeneously?*” Manojit Chattopadhyay and Subrata Kumar Mitra, Indian Institute of Management Raipur (2019)
3. *"Use of dynamic pricing strategies by Airbnb hosts"* Chris Gibbs, Daniel Guttentag, Ulrike Gretzel, Lan Yao, Jym Morton, International Journal of Contemporary Hospitality Management, Vol. 30 Issue: 1, pp.2-20 (2018)
4. *“Key Factors Affecting the Price of Airbnb Listings:A Geographically Weighted Approach”* Zhihua Zhang, Rachel J. C. Chen, Lee D. Han, and Lu Yang, University of Tennessee (2017)

ANNEX



```
                            OLS Regression Results                            
==============================================================================
Dep. Variable:                      y   R-squared:                       0.307
Model:                            OLS   Adj. R-squared:                  0.302
Method:                 Least Squares   F-statistic:                     53.73
Date:                Sun, 24 Jan 2021   Prob (F-statistic):           3.29e-72
Time:                        18:15:03   Log-Likelihood:                -5544.5
No. Observations:                 977   AIC:                         1.111e+04
Df Residuals:                     968   BIC:                         1.115e+04
Df Model:                           8                                         
Covariance Type:            nonrobust                                         
==============================================================================
                 coef    std err          t      P>|t|      [0.025      0.975]
------------------------------------------------------------------------------
const       1806.2654   2395.761      0.754      0.451   -2895.219    6507.750
x1            18.8129      1.171     16.063      0.000      16.514      21.111
x2           -15.4591      2.789     -5.542      0.000     -20.933      -9.985
x3             9.8171      2.939      3.341      0.001       4.051      15.584
x4            -0.0328      0.122     -0.268      0.789      -0.273       0.208
x5            -0.0965      0.025     -3.863      0.000      -0.146      -0.047
x6            -0.3644      6.764     -0.054      0.957     -13.638      12.909
x7           -28.2128     48.632     -0.580      0.562    -123.649      67.223
x8          -161.1247     32.770     -4.917      0.000    -225.433     -96.816
==============================================================================
Omnibus:                      743.642   Durbin-Watson:                   1.050
Prob(Omnibus):                  0.000   Jarque-Bera (JB):            22936.720
Skew:                           3.158   Prob(JB):                         0.00
Kurtosis:                      25.881   Cond. No.                     1.51e+05
```

The entire table for the OLS regression from the analysis



The frequency of residuals of the OLS regression (centered around zero)