



A web scraping project

By: TRAN Ngoc-Phung
Instructor: Professor Fabrice LE GUEL

March 2020

Overview

Airbnb is a platform of short-term rentals, experience and multi-day trips, originated in San Francisco from the idea of turning the founder's apartment into a bed and breakfast for short-term rentals. The first version of Airbnb, Airbedandbreakfast.com, officially launched on August 11, 2008 and eventually became a huge success as it met the needs of travelers struggling to find accommodation. In 2020, Airbnb has more than 4,500,000 listings in over 65,000 cities in 191 countries.¹

Legal Name	Airbnb, Inc.
Headquarters Regions	San Francisco Bay Area, West Coast, Western US
Founded Date	Aug 11, 2008
Founders	Brian Chesky, Joe Gebbia, Nathan Blecharczyk
Revenue	\$4,308,726,681 ²
Acquisitions	Airbnb has acquired 21 organizations. Most recent acquisition was Urbandoor on Aug 5, 2019.

¹ Crunchbase, "Airbnb", available at: <https://www.crunchbase.com/organization/airbnb> (accessed 23 March 2020).

² Airbnb: An Analyst's Guide (Part 1) - AllTheRooms Analytics. (2020), available at: <https://www.alltherooms.com/analytics/airbnb-ipo-going-public-revenues-business-model-statistics/> (accessed 23 March 2020).

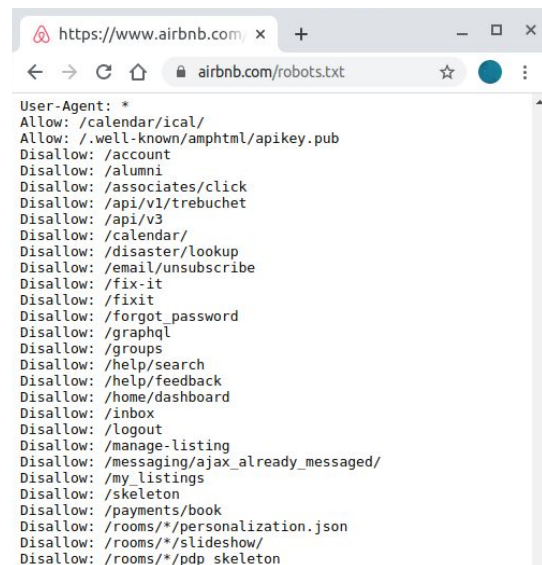
Economic aspect:

AirBnb is a pure multi-sided platform, charging both sides (hosts and guests) a “service fee.” The platform encourages the two sides to review one another, intensifying the indirect network effect between the two sides. There's a positive direct network effect between the guests, facilitated by the “referring” program, which helps users earn credit by sending a referral link to a potential customer. Meanwhile, the negative network effect between the hosts makes them competing by pricing strategies or non-pricing strategies (earning badges, encouraging their guests to write positive reviews). Though Airbnb was founded very recently, since 2011 the company began growing very rapidly³, becoming an increasingly important issue in both tourism and public policy.

Technical aspect:

Technically speaking, Airbnb is a web application built with Ruby on Rails, using many Javascript frameworks such as JQuery, React, AmplifyJS and MomentJS. The fact that Airbnb is a JavaScript-heavy site with data being loaded by AJAX makes Scrapy xpaths normally yield empty results. Moreover, Airbnb is a peculiar, dynamic target as the layout of the website might be changed depending on the session, as well as it gets updated very frequently.

Looking at the robots.txt file (<https://www.airbnb.com/robots.txt>), the site has many rules to follow (many pages or files the crawler can't request) but still, it allows crawlers to scrap simple listings; though the dynamic nature of it requires some more efforts while still respect the rules and avoid overload the site with too many requests.



```
User-Agent: *
Allow: /calendar/ical/
Allow: /.well-known/amphtml/apikey.pub
Disallow: /account
Disallow: /alumni
Disallow: /associates/click
Disallow: /api/v1/trebuchet
Disallow: /api/v3
Disallow: /calendar/
Disallow: /disaster/lookup
Disallow: /email/unsubscribe
Disallow: /fix-it
Disallow: /fixit
Disallow: /forgot_password
Disallow: /graphql
Disallow: /groups
Disallow: /help/search
Disallow: /help/feedback
Disallow: /home/dashboard
Disallow: /inbox
Disallow: /logout
Disallow: /manage-listing
Disallow: /messaging/ajax_already_messaging/
Disallow: /my_listings
Disallow: /skeleton
Disallow: /payments/book
Disallow: /rooms/*/personalization.json
Disallow: /rooms/*/slideshow/
Disallow: /rooms/*/pdp_skeleton
```

Airbnb's robots.txt file, accessed 23 March 2020

³ Griswold, A. (2018), “This new year’s, Airbnb got the hockey-stick growth that every startup envies”, Quartz, 3 January, available at: <https://qz.com/877080/airbnbs-growth-in-guests-on-new-years-is-the-hockey-stick-curve-that-every-startup-wants/> (accessed 23 March 2020).

Web scraping

First try with Scrapy

At first I tried to scrap the website with Scrapy, fully aware that the dynamic nature of the site could make it infeasible for Scrapy to handle, but at the same time I was also curious about how far I could go with Scrapy. Though the Scrapy Xpaths normally return empty results, it does work from time to time.

```
File Edit View Search Terminal Help
In [8]: response.xpath("//div[@class='_1ebt2xej']/text()").extract()
Out[8]: []

In [9]: response.xpath("//div[@class='_8ssblpx']/text()").extract()
Out[9]: []

In [10]: response.xpath("/html/body/div[3]/div/div[1]/main/div/div/div[2]/div/div[1]/div/div/div[2]/div/div/div[1]").extract()
Out[10]: []

In [11]: []
```

Scrapy Xpath results on 27 January 2020: empty data

```
def parse(self, response):
    for bnb in response.css('div._ylefn59'):
        yield {
            'name': bnb.css('div._6kiyebe div._1ebt2xej::text').extract_first(),
            'type': bnb.css('div._6kiyebe div._4ntfzh div._1q6rrz5::text').extract_first(),
            'rating': bnb.css('div._6kiyebe div._4ntfzh span._60dc7z span._60hvkx2 span._ky90pu0::text').extract_first(),
            '#total reviews': bnb.css('div._6kiyebe div._4ntfzh span._60dc7z span._60hvkx2 span._krbj::text').extract_first(),
            'maximum': bnb.css('div._6kiyebe div._1s7voim::text').extract_first(),
        }
```

Out[9]:

	name	type	rating	maximum
0	CARPE DIEM Hyper centre LA ROCHELLE classé ***	None	4.95	4 guests
1	Appartement T1Bis La Rochelle Centre avec Parking	Entire apartment	4.75	2 guests
2	Chambre	Private room	4.83	2 guests
3	Studio de charme en plein coeur de La Rochelle	None	4.88	2 guests
4	penthouse historic center	Entire apartment	4.73	4 guests
...
80	STUDIO RENOVE VIEUX PORT HYPER CENTRE AVEC WIFI	None	4.89	2 guests
81	Studio centre La Rochelle, proche gare	Entire apartment	4.41	2 guests
82	Studio La ROCHELLE, la Rochelière les Minimes	Entire apartment	4.60	2 guests
83	Studio cosy au calme et proche centre-ville	None	4.90	2 guests
84	Studio La Rochelle « les minimes » proche plage	None	5.00	2 guests

85 rows × 4 columns

Testing results on 29 January 2020: Scrapy works with response.css

```

class BasicSpider(scrapy.Spider):
    name = 'basic'
    allowed_domains = ['web']
    start_urls = ['https://www.airbnb.com/s/La-Rochelle/homes?query=La%20Rochelle']
    for x in range(0,280,20)
    ]

    def parse(self, response):
        item = ExodusItem()
        item['apppname']=response.xpath('//*[@class="_i24ijs"]/@href').extract()
        return item

```

Simple spider on 09 March 2020: the spider could get the list of rooms by using rel Xpath

```

'url': 'https://www.airbnb.com/rooms/19429245?location=La%20Rochelle&check_in=2020-07-04&check_out=2020-07-05&previous_page_section_name=1000&federated_search_id=f7032593-9cbb-4bbb-8a4c-c48afa2c4c39'}
2020-03-09 17:18:52 [scrapy.extensions.logstats] INFO: Crawled 10 pages (at 5 pages/min), scraped 8 items (at 5 items/min)
2020-03-09 17:18:59 [scrapy.core.engine] DEBUG: Crawled (200) <GET https://www.airbnb.com/rooms/8243759?location=La%20Rochelle&check_in=2020-07-04&check_out=2020-07-05&previous_page_section_name=1000&federated_search_id=f7032593-9cbb-4bbb-8a4c-c48afa2c4c39> (referer: None)
2020-03-09 17:18:59 [scrapy.core.scrapy] DEBUG: Scraped from <200 https://www.airbnb.com/rooms/8243759?location=La%20Rochelle&check_in=2020-07-04&check_out=2020-07-05&previous_page_section_name=1000&federated_search_id=f7032593-9cbb-4bbb-8a4c-c48afa2c4c39>
{'Name': [],
 'Price': [],
 'Rate': ['4.48'],

```

Spider results with Xpath after successfully retrieving the list of rooms: sometimes it works, sometimes it doesn't

These results gave the conclusion that though Scrapy Xpaths work from time to time, data loaded by AJAX impedes Scrapy accessing HTML elements, and the layout of the website could also be changed depending on the session. While the spider could get the list of rooms, scraping listing details may involve many sessions and the relative Xpath would change accordingly. As a result, Scrapy is not the ideal tool in this case, or at least, it needs to be combined with other tools.

Selenium

Selenium, written in Java, is a tool mainly used for automated testing, but it comes in handy for websites that Scrapy fails to scrap. Selenium was developed in 2004 by programmers and testers at ThoughtWorks, a software company in Illinois, United States. Its creator, Jason Huggins⁴, later came to work for Google for more than a year and is now the founder of Tapster Robotics, a company developing Tapster robot for testing iPhone and Android mobile applications. Over the years of development, Selenium now works well with popular programming languages, such as C#, Java, PHP, Python, Ruby,... and can run on Windows, Linux, macOS. It is open-source and is released under the Apache License 2.0.⁵

Selenium facilitates web scraping by simulating the browser. It opens a browser with a webdriver, then the crawler can mimic user behaviour with all the clicks, scrolls and pauses, while collecting the data. The process also allows us to observe if anything goes wrong, for example, a change in layout. One weakness of Selenium is that normally it takes time for the browser to fully load. However, in my opinion, this is also a useful way not to harm the website and not to get banned.

⁴ Huggins' LinkedIn profile. Available at: <https://www.linkedin.com/in/jrhuggins/> (accessed 23 March 2020).

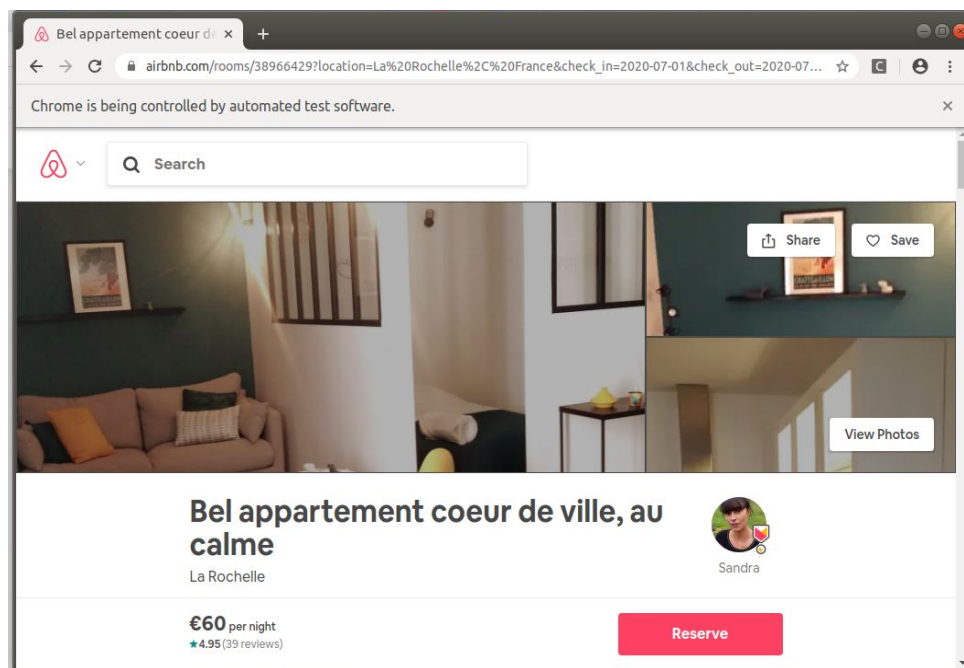
⁵ Wikipedia, Selenium. Available at: [https://en.wikipedia.org/wiki/Selenium_\(software\)](https://en.wikipedia.org/wiki/Selenium_(software)) (accessed 23 March 2020).

Data acquisition & data cleaning

Firstly, I tried to scrap Airbnb listings in La Rochelle, then used the same code to scrap listings in Paris as well. The idea is to enlarge the dataset, as Airbnb would normally limit the maximum number of listing pages to 17. The second reason is to see the difference between the two, as well as to indicate the factors affecting the price in these two cities and to evaluate if they are the same.

Though the first listing was successfully acquired by a Scrapy spider, with Selenium. `next_button.click()` is an interesting feature that could be used to click the next page automatically and grab listings. Selenium opens a driver on the screen so we could observe the layout and intervene if something goes wrong.

As Airbnb gets layout updates quite frequently, the image below shows the layout displaying on the 19th and 21st of March 2020, the two dates that La Rochelle and Paris listings' data in this project were scrapped.



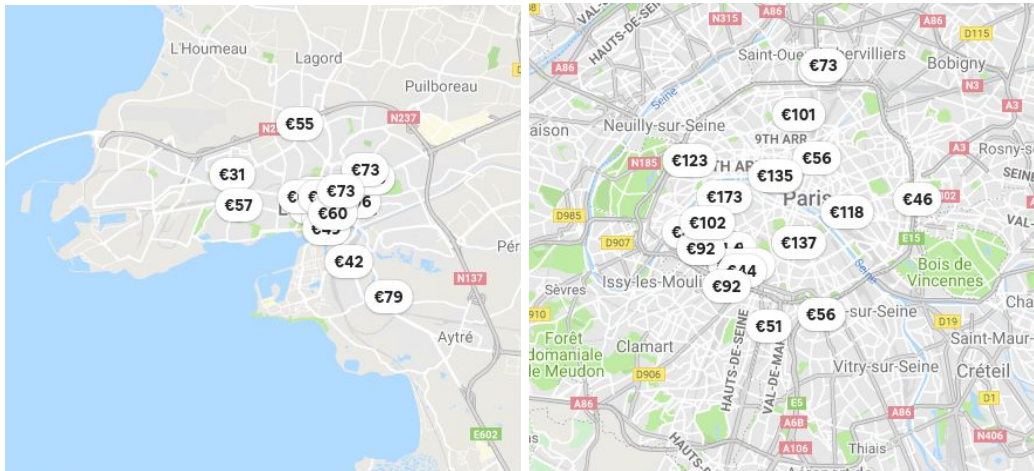
The layout of the locations' pages

The next step is to acquire data with the following variables:

- the location's title given by the host: **house_title**
- type of the location: **house_type**
- price for a night (with date of search being fixed): **price**
- username of the host: **user**
- the number of people who gave their reviews on the location: **no_of_ratings**
- rating of the location, on a scale of 5: **rating**
- other house details (number of bedrooms, beds, baths): **house_detail**

Thereafter, in the data cleaning process, house_detail are separated into four different variables:

- Maximum number of guests the location could accommodate: **no_of_guest**
- Number of bedrooms (studio is manipulated to 1): **house_size**
- Number of beds: **no_of_beds**
- Number of baths: **no_of_baths**



Airbnb listings in La Rochelle and in Paris

Data cleaning also involves fixing and manipulating the data that is incorrectly formatted (for example, "half-bath" to 0.5, locations noted as "studio" are given the value of 1 in "bedroom") or deleting observations missing the price.

```
In [41]: airbnbdfClean.dtypes
Out[41]: house_title      object
house_type      object
price           float64
user            object
no_of_ratings    int64
rating          float64
house_detail     object
house_details    object
no_of_guest      float64
house_size       object
no_of_beds       object
no_of_baths      object
house            float64
beds             float64
bath             float64
dtype: object
```

Data types after cleaning

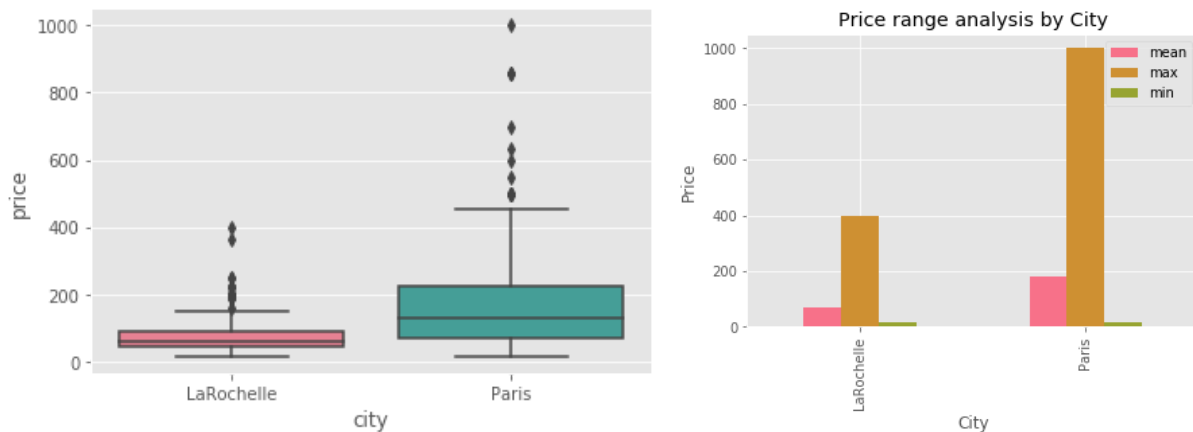
Analysis

A very short introduction to the two cities: La Rochelle is a coastal city in southwestern France and capital of the Charente-Maritime department, with the population of 75,735 people. Paris, on the other hand, is the capital and the most populous city of France with 2,187,526 habitants.⁶

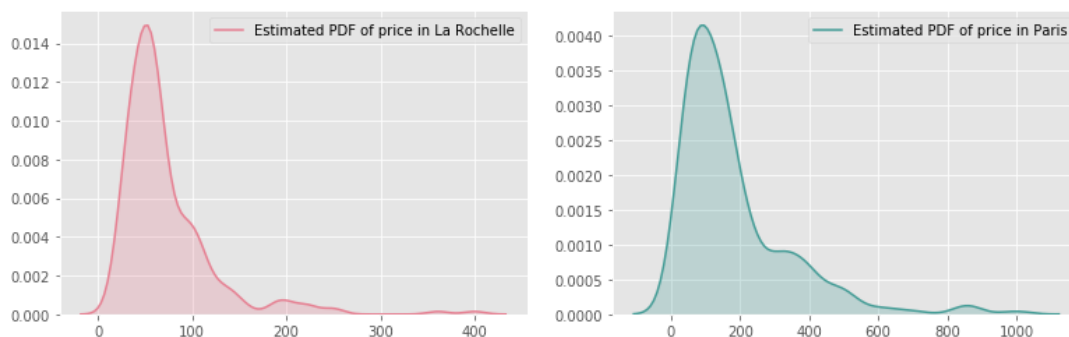
The data analysis is to compare factors influencing Airbnb prices and to see how price varies depending on different types of location in La Rochelle and in Paris.

Price Range

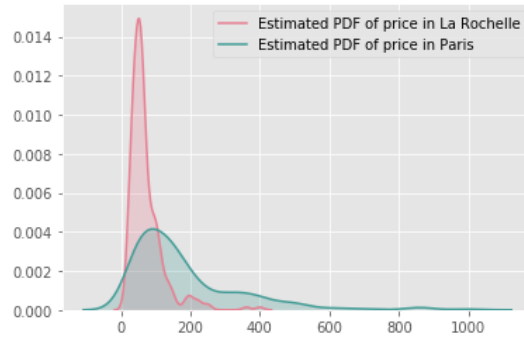
The boxplot below shows that locations in Paris have higher prices than La Rochelle, as predicted, averaging around €182 and €72 respectively. The maximum price of a night in Paris could cost up to €1000, 2.5 times higher than the maximum price in La Rochelle. Despite the pricing distance, in both cities, one person can find a place with only €15 - €16 for a night.



The kernel density estimate of price below gives the idea of where the locations are priced on the price range.

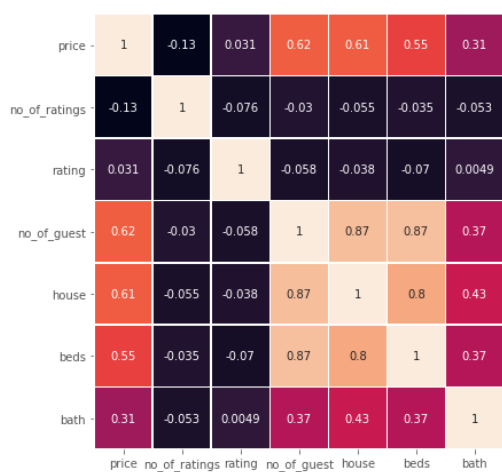


⁶ INSEE, Populations légales 2017. Available at: <https://www.insee.fr/fr/statistiques/4269674?geo=COM-17300> (accessed 23 March 2020).

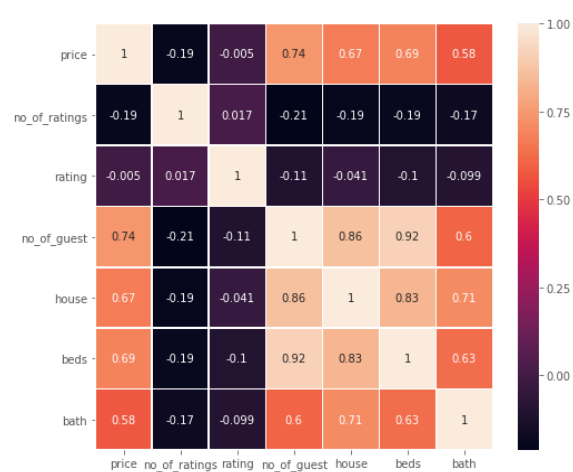


What drives price?

From the correlation matrix, we can see that factors such as the number of bedrooms, beds, and guests drive price significantly. In both cities, the number of baths is less correlated with price, especially in La Rochelle. Maximum number of guests the location could accommodate is the most correlated factor to price, rating 0.62 in La Rochelle and 0.74 in Paris.



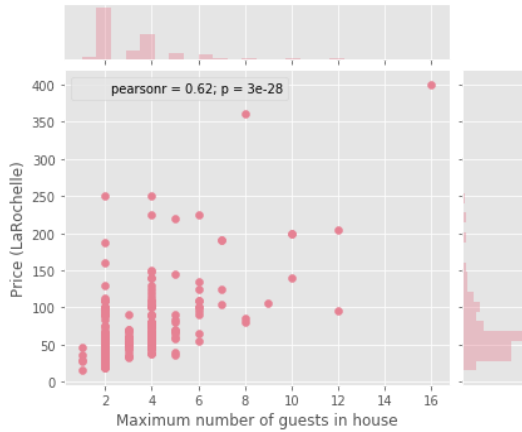
La Rochelle listings: Correlation Matrix



Paris listings: Correlation Matrix

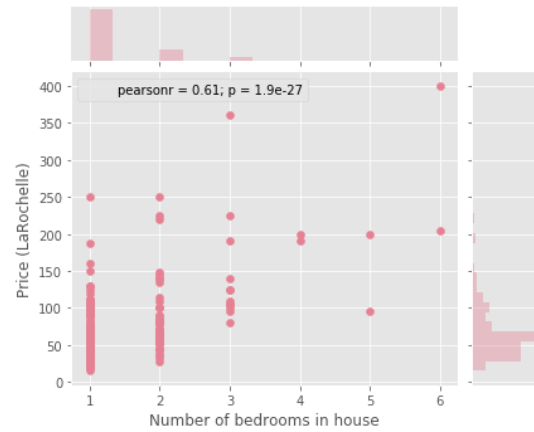
Pearsonr indicates the Pearson correlation of these variables. In these four cases, the p-value is extremely small, it means that there is significant linear relation between each pair of variables analysed above.

La Rochelle: Pearson correlation between price and maximum number of guests



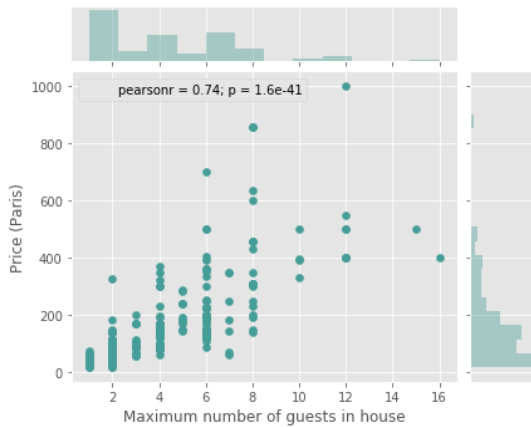
Pearson correlation of these two variables is 0.62, p-value is very small which indicates that there is a significant linear relation between Maximum number of guests and Price in La Rochelle

La Rochelle: Pearson correlation between price and size of house



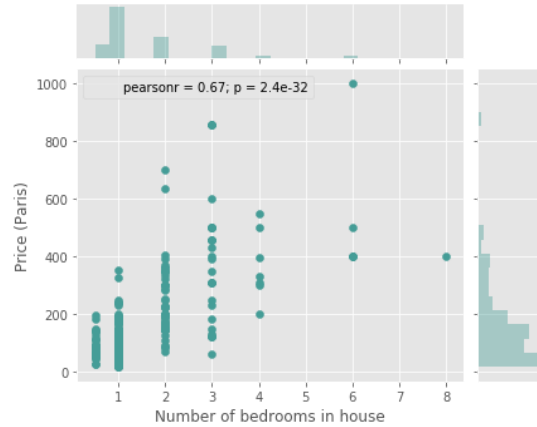
Pearson correlation of these two variables is 0.61, p-value is very small which indicates that there is a significant linear relation between number of bedrooms and Price in La Rochelle

Paris: Pearson correlation between price and maximum number of guests



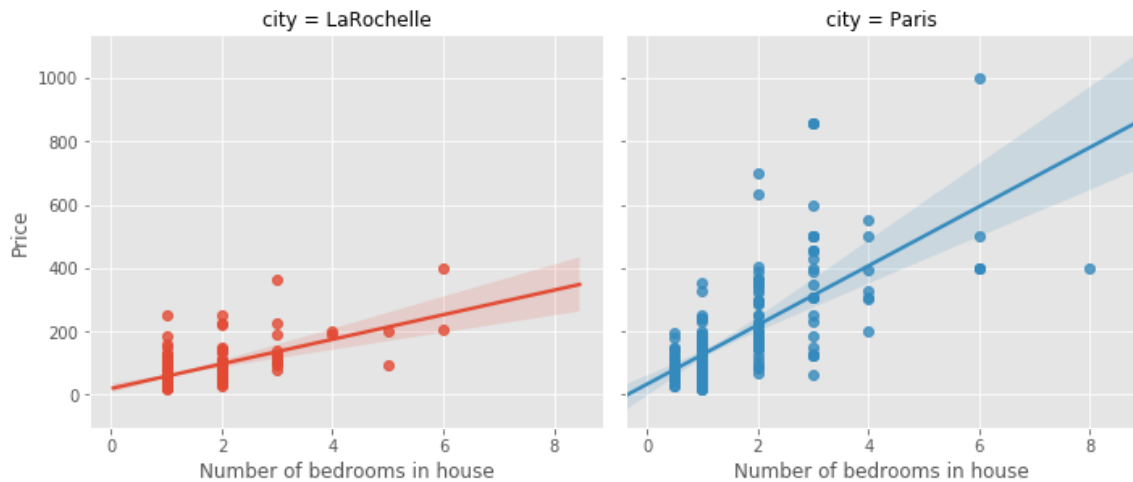
Pearson correlation of these two variables is 0.74, p-value is very small which indicates that there is a significant linear relation between Maximum number of guests and Price in Paris

Paris: Pearson correlation between price and size of house

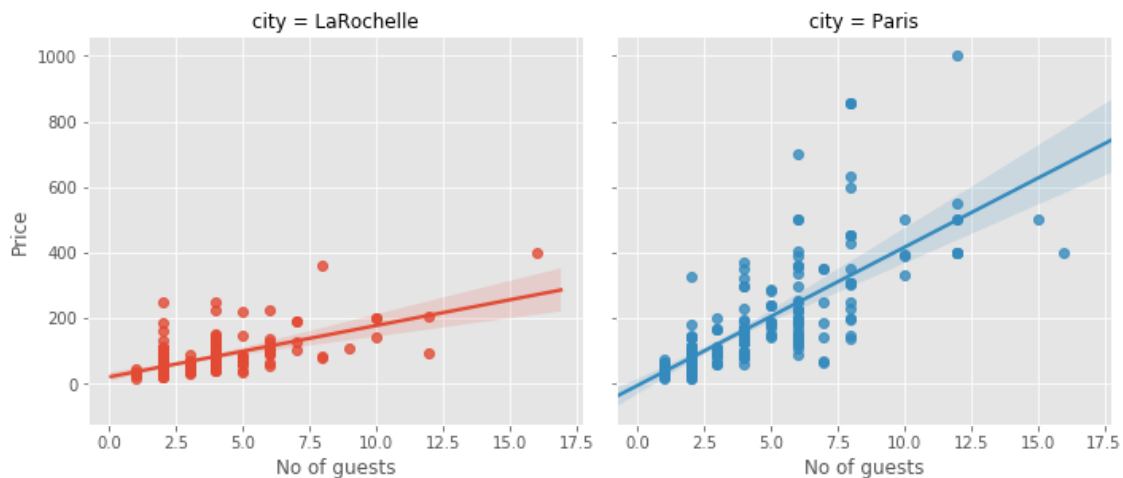


Pearson correlation of these two variables is 0.67, p-value is very small which indicates that there is a significant linear relation between number of bedrooms and Price in Paris

Price analysis by size of the house

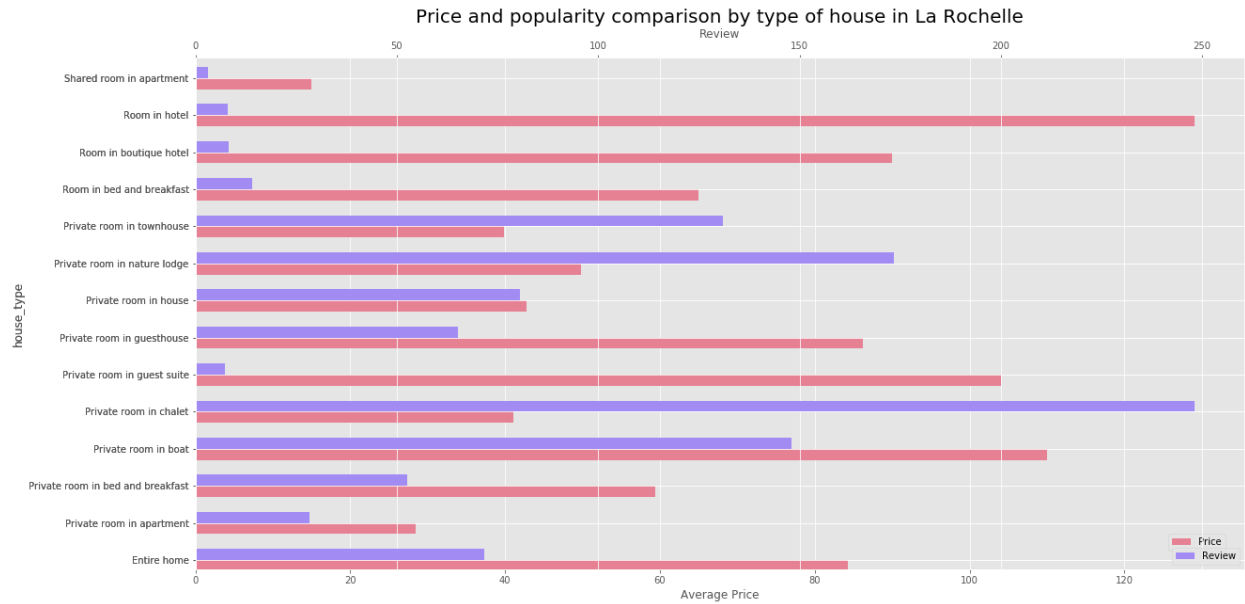


Price analysis by number of guest

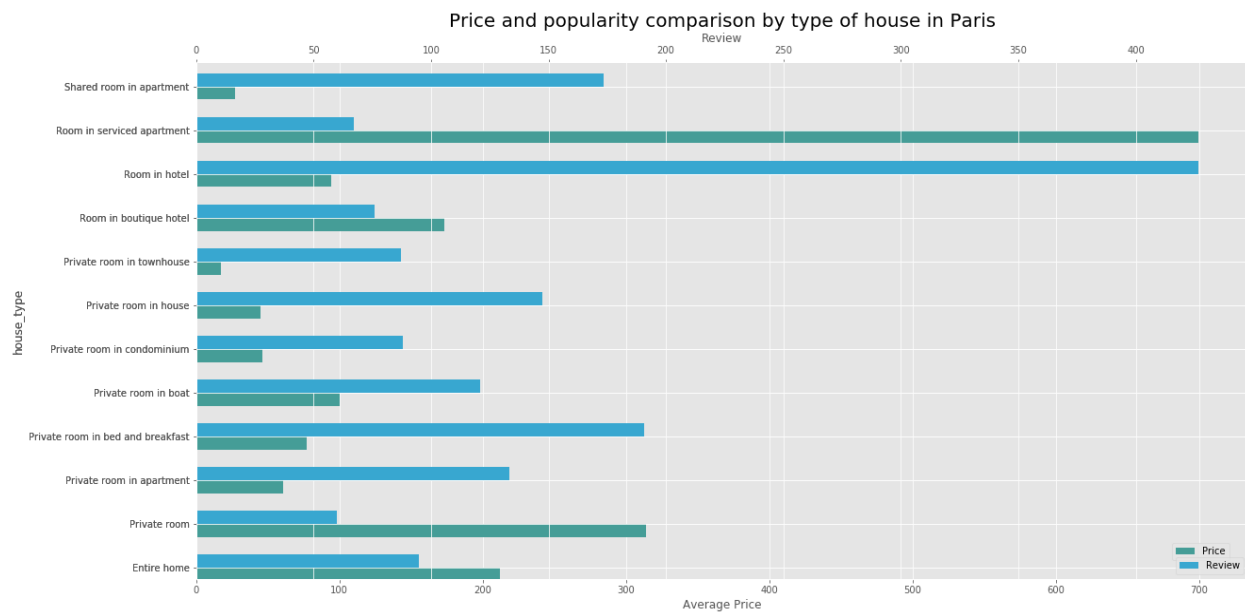


Next, looking at the relation between price, popularity and house type, we can see some interesting points. Popularity here is defined by the number of ratings that were left for the location, not the star-rating that location has got.

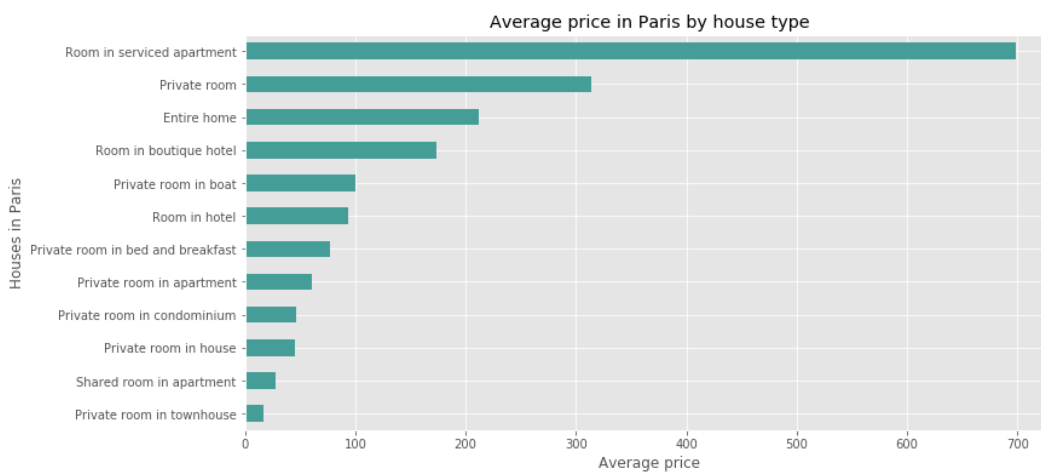
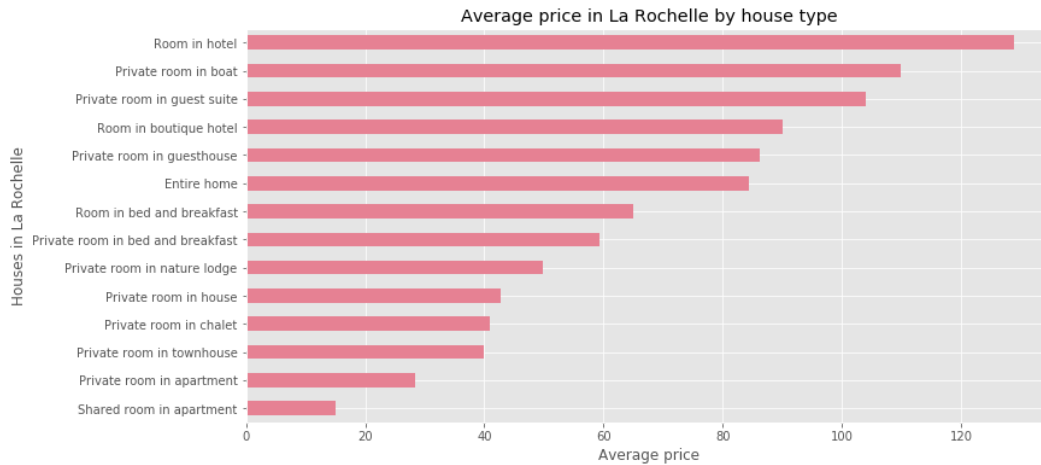
In La Rochelle, "Private room in chalet" is the most popular type of short-term renting with the average price of €41, the second-most popular type is "Private room in nature lodge," costing around €50 on average. "Shared room in apartment" is the least sought-after type with the average price of just €15. "Private room in guest suite" is also not so desirable, pricing at €104 for a night. Also, in La Rochelle, "Room in hotel" is not so popular, this is a point worth noting in the two graphs. A reason could be the highest average price among all the offers, but another reason could be the nature of short-term renting, i.e business travelers (which Paris accommodates a lot more often) would choose differently than a family on vacation.



While in Paris, "Room in hotel" is the most popular type with the average price of €94, the second-most popular type is "Private room in bed and breakfast," significantly less popular than "Room in hotel", costing around €77 on average. "Private room" is the least sought-after type with the average price of €313. "Room in serviced apartment" is also not so desirable, pricing up to €699 for a night.

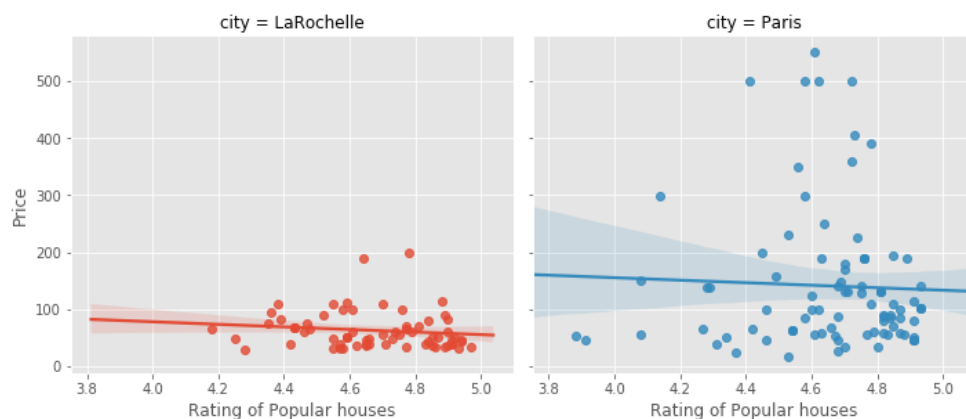


Another reason why the most costly places are not popular I could think of, is that the majority of people who can pay the highest price wouldn't tend to have time to leave reviews.

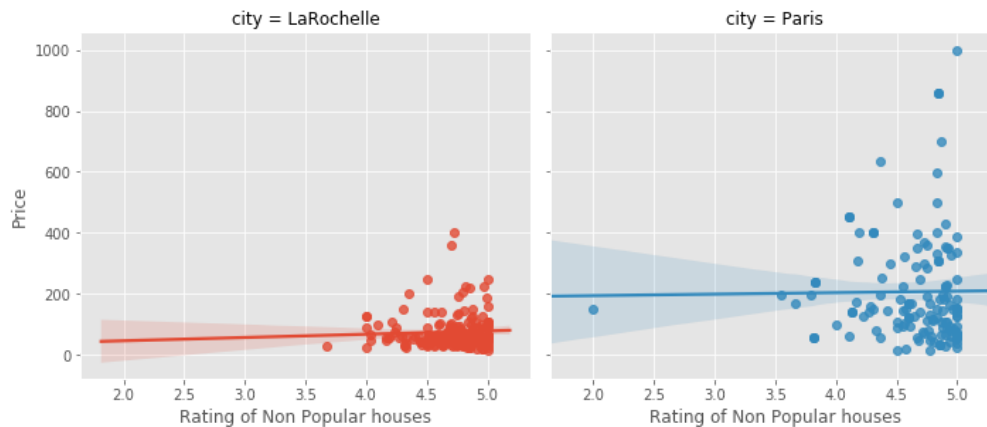


About price and popularity, the regression lines below show that the popularity of a place doesn't drive its price on Airbnb. It means the predefined notion that better service comes with extra price seems to be wrong on this platform. One can probably find a better-rated place to rent with the same or even lower price, in Paris or in La Rochelle.

Price & Popularity relationship



Price & Non Popularity relationship



Conclusion

Based on the analysis above, we can come to several conclusions:

- The price of Airbnb listings in Paris and in La Rochelle has a significant positive linear relationship with the size of the location (number of bedrooms and beds), as well as maximum number of guests, while the linear relationship between price and the number of baths is less significant.
- The higher number of reviews doesn't drive price on Airbnb. This indicates that short-term renters could find a highly-rated place at a relatively affordable price.
- Whether in Paris or in La Rochelle, short-term renters can find a place from only €15 a night. Interestingly, the cheapest places (costing less than €18 a night) in both cities have very high ratings, ranging from 4.5 to 5 stars on a scale of 5. However, as the geographical factor (exact location) is excluded, therefore, it doesn't mean that the cheapest price comes with the most comfortable place depending on the purpose of the trip.

Limitations

As mentioned above, the current dataset doesn't include the exact location of the listings, as Airbnb doesn't disclose the listing addresses until payment is confirmed. I think there would be a way to scrap latitude and longitude data, but still, I'm not sure if that action would be considered acceptable by Airbnb and not fall into the "restricted area" of the platform.

Project on Github

All the code, analysis and results can be found at: <https://github.com/tnp1606/ISE>