MODS207

Project in applied economics

The effect of competition on host-guest matching on Airbnb

Tiffany LY, Nicolas JOW

I - Introduction

In september 2017, the city of San Francisco introduced a Settlement Agreement with Airbnb. This policy established several conditions and regulations about short-term rental. The main rule was that host listings have to be now lawfully registered on the City's Short Term Residential Rental Registry for short-term rental, which includes complying with the administrative guidelines of the Official Short Term Rentals.

This particular amendment led to a general decrease of the number of listings offered on Airbnb, and may have caused changes in guests and hosts behaviour on the platform. Our study aims to highlight whether market conditions changes enabled hosts to select their favorite guest. Are the hosts more picky since the 2017 San Francisco policy ? Is there more discrimination in the matching process ? Is it preferable for a guest to have certain features in order to be picked ?

In order to answer these specific questions, we will use a dataset of 800.000 observations of information related to hosts and the dwelling they offer on the website : each row of the dataset corresponds to a transaction between a host and a guest. This dataset shows every transaction that has been made on Airbnb San Francisco between 2015 and 2019. Since we also have the identifiers of the different guests that posted comments about the lodgings, we will create our own new dataset of guests by browsing through the original dataset. Therefore, in order to harvest more details about the guests, we will scrape each guest of the dataset on the Airbnb website.

After describing and analyzing features of our dataset, such as ability to speak English or location, we figured out dependencies between features given by users on their profile and the number of comments they received from other users. The number of languages spoken has a positive effect on this number of comments, as well as the living place of the user, especially if the individual is american. Then, we analyzed the impact of these features depending on time.

II - Background

San Francisco policy is meant to fight against people who take advantage of the Airbnb system. Indeed, some people used Airbnb in order to turn apartments and houses into full-time Airbnb rentals, making it more difficult for residents to find housing. Since 2017, the Airbnb regulation requires all owners to register on a registry for a fee of $50. They must also live in their home for at least 275 days per year. Therefore, the rental period cannot exceed 90 days, and it has become more difficult for people to become Airbnb hosts in San Francisco.

In his study *Competition and Reputation in a Congested Marketplace: Theory and Evidence from Airbnb*, Michelangelo Rossi particularly focused on the relationship between the changes in the number of competitors and the role of reputation for sellers after the Airbnb regulation was effectively enforced in San Francisco in 2017. While we could think that having a lot of competition could pressure hosts to provide more effort and be more helpful to guests, it is actually not the case. Rossi found that there was a significant negative impact of the number of competitors on rating about hosts' effort, and thus reputation.

Indeed, this may seem counterintuitive, but even in high season, hosts can struggle to have guests. Thus, when there is a lot of competition, there is a discouraging effect, and hosts do not have the heart to make a lot of effort for their guests. Since there are already few chances that they will have a guest, their reputation will not really change anything. On the other hand, if there are less competitors, hosts are more likely to have guests. And therefore, their reputation will have greater value, since it could play in the decision of the guest to choose this host.

In our paper, instead of focusing on how the decrease of the number of hosts could affect hosts' reputation and efforts, we would like to study how the decrease changed the hosts' behaviour towards guests. Did they become more picky? Do they favor some types of guests? Did it change, or has it always been like that?

III - Data collection strategy

For our data collection, we will use the Airbnb website. In particular, we will scrape the information on the profile page of every guest who left a review and whose identifier is in the dataset of 800.000 rows we already have at our disposal, corresponding to every transaction that has been made in Airbnb San Francisco between 2015 and 2019. Each row of the dataset corresponding to a transaction between a host and a guest, we will have to go over each row and load the guest page to scrape it. Before being able to browse through the dataset, we will have to delete all the transactions without a review, since in this case the guest ID is not documented. We will also delete all duplicates so our code runs faster. After this step, our dataset has about 500.000 rows.

We will focus on the following personal data of the guests : **name, living place, job, languages spoken, number of hosts comments, content of comments, date of comments**, and if the **profile is verified or not**. If the guest has a verified profile, we will check the different features that are indeed verified (identity, phone number, email address, professional email address). Each observation of our final database should have all the features previously mentioned and only one host comment per observation. This means that we will have as many observations for a guest as the number of hosts comments the guest has on his profile.

In order to collect this data, we will use Python. As Airbnb's website does not have any API, we will use the two following libraries for the scraping : Beautifulsoup and Selenium. Selenium will enable us to navigate through the different guest profiles on Airbnb's website, and Beautifulsoup will harvest the data we are interested in.

The massive data scraping may drive us to be blocked by Airbnb's website.In order to avoid different forms of blocking, we will launch our scraping algorithm on a VPN. This way, the IP address will automatically change at each connection, and we will not be blocked by Airbnb. Another problem is that in order to scrape the guests' information, we need to be logged in on the Airbnb website : this is why we could not just use Beautifulsoup to load the HTML pages and scrape the data. Indeed, we need the Selenium webdriver to open a web page, so we can enter our identifiers (e-mail address and password) and fill in the CAPTCHA. Thus, the scraping will only be possible if we load each new guest page on the webdriver window.

In the end, one observational unit in the data will be as following :

- Guest name (string)

- Member ID (int)
- Living place (if not N/A, string of “City, State” if in the U.S.A, “City, Country” if not)
- Languages spoken (if not N/A, list of string [“Language 1”, “Language 2”]
- Job (if not N/A, string “Job”)
- Verified features (if not N/A, list of verified features [“Feature 1”, “Feature 2”])
- Number of hosts comments (i.e. if a host left them a comment, int)
- Content of host comments (list of string [“Comment 1”, “Comment 2”])
- Date of host comments (list of string [“Date 1”, “Date 2”])
- Number of guest comments (i.e. if they are also a host and a guest left them a comment, int)
- Content of guest comments (list of string [“Comment 1”, “Comment 2”])
- Date of guest comments (list of string [“Date 1”, “Date 2”])

In the end, we scraped 17.000 rows of the dataset, which represents 3,4% of our dataset. However, we did it by going through the start of the original dataset and sometimes starting from random rows of the dataset, so while it is not a relatively huge number of rows, we can still consider it representative.

We also proceeded to scrape another dataset, to compare the behavior of precise hosts that were on Airbnb before *and* after the San Francisco policy in 2017. In this dataset, we only have guests that have stayed at hosts that were active through the settlement of the policy. On this dataset, we have about 4.000 rows.

To analyse the data, we also had to prepare the data and clean it. In order to do that, we used the Pandas library. Since the data we scraped had a lot of “ornaments” (for example, the scraped name was “Bonjour, je m’appelle Britt” instead of just “Britt”), we had to clean almost all of our columns. Then, we had to divide in two lists the guests that had been accepted before 2017, and those who had been accepted after 2017, so we could establish the profiles of the people who were accepted in each case. We made it so guests that had stayed in San Francisco both before and after the policy appeared in both lists.

IV - Data Description

We have two datasets : one of 17.420 rows of guests from the original dataset of 800.000 rows, and one of 4.109 rows only made from the guests who stayed at hosts that were here before and after the San Francisco policy.

For the first dataset, we did not discriminate for this dataset and scraped all information of a guest as long as it was scrapable. We had 14.033 rows of guests who were accepted before 2017, and 3.128 who were accepted after 2017. We can describe some of our features in several tables :

	Before 2017	After 2017
No	63%	70%
Yes	37%	30%

Fig.1 : Percentage of people who mentioned their job on their profile

This table represents the percentage of people who mentioned their job on their profile, among guests who have been accepted in San Francisco Airbnbs before and after 2017. We can see that there are no big differences between both of the numbers, and that even after the policy, there are even less people who mentioned their job on their profile.

	Before 2017	After 2017
No	70%	74%
Yes	30%	26%

Fig.2 : Percentage of people who mentioned they spoke English on their profile

In the same fashion, this table represents the percentage of people who mentioned they spoke English on their profile. We see a result similar to the one we had above.

	Before 2017	After 2017
Not mentioned	78,6%	82,3%
1	12,5%	10,7%
2	5,9%	4,9%
More than 3	3,0%	2,1%

Fig.3 : Number of languages spoken

Now we studied the number of languages spoken by guests that we scraped. We see that most people didn't mention which languages they spoke, but even more in 2017 than before. Otherwise, a greater proportion of the guests accepted in 2017 indicated that they spoke more than one language.

	Before 2017	After 2017
Mean	2.3	4.0
Standard deviation	2.9	2.2
Median	8	10

Fig.4 : Information on number of comments

For this last table, we focused on the information we could get about the number of comments. It is important to specify here that the number of comments is at most equal to 10, because we didn't scrape the comments in the section "See more".

We can see that before and after 2017, the mean of the comments increased. But this could be explained as people got more comments over time, thanks to the expansion of the sharing economy worldwide. We also see that the median is very high, in both cases. This illustrates that guests on Airbnb have either a lot of comments, or almost no comments at all.

Maybe the San Francisco policy pushed hosts to post comments more often, as they know the market has become pickier. But it is too difficult to tell if the law on short-term rentals really impacted the market with only this information.

For the second dataset, we only chose guests that were here before and after 2017. We only scraped guests who stayed at these hosts. We had 3.160 rows of guests who were accepted before 2017, and 1.151 who were accepted after 2017. We can describe some of our features in the same sort of tables :

	Before 2017	After 2017
No	67%	77%
Yes	33%	23%

Fig.5 : Percentage of people who mentioned their job on their profile

	Before 2017	After 2017
No	71%	76%
Yes	29%	24%

Fig. 6 : Percentage of people who mentioned they spoke English on their profile

	Before 2017	After 2017
Not mentioned	78,6%	84,3%
1	12,6%	8,9%
2	5,8%	4,5%
More than 3	3,0%	2,3%

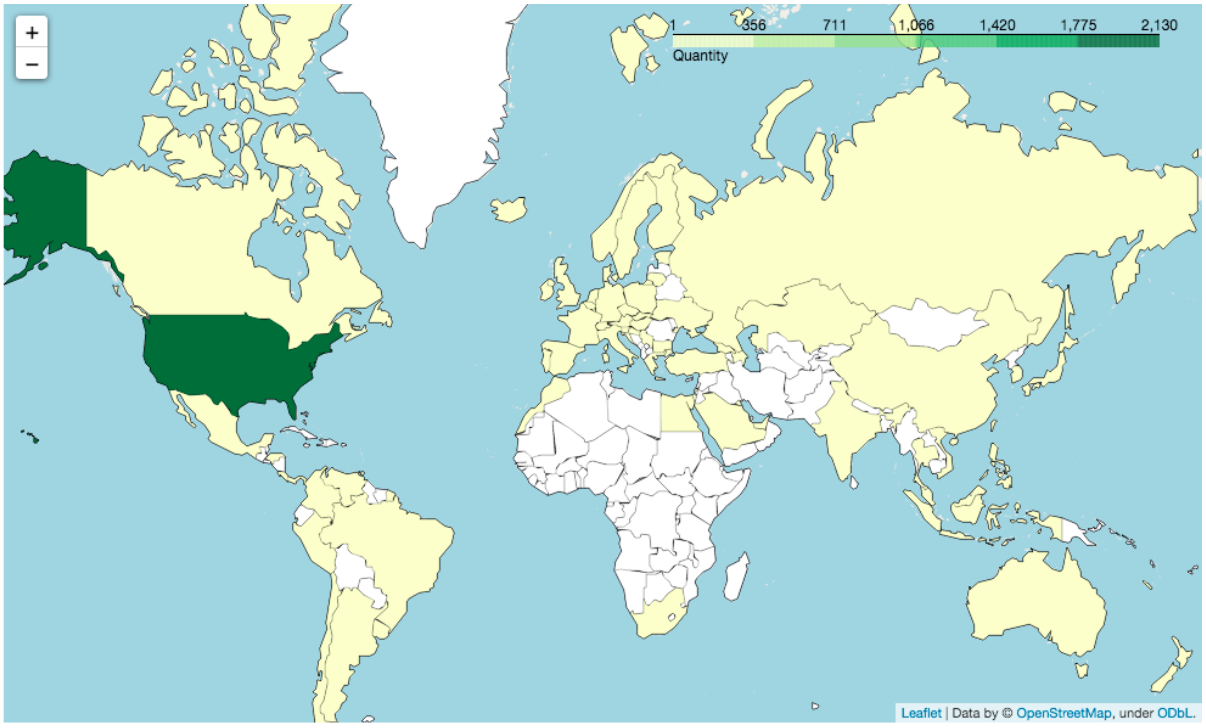
Fig. 7 : Number of languages spoken

	Before 2017	After 2017
Mean	1.9	4.3
Standard deviation	2.9	2.4
Median	8	10

Fig. 8 : Information on number of comments

Concerning the four tables above, we studied the same features as in Fig. 1, 2, 3, 4. The results are approximately the same as the ones we got in the first study, with the dataset of 17.000 rows.

V - Analysis



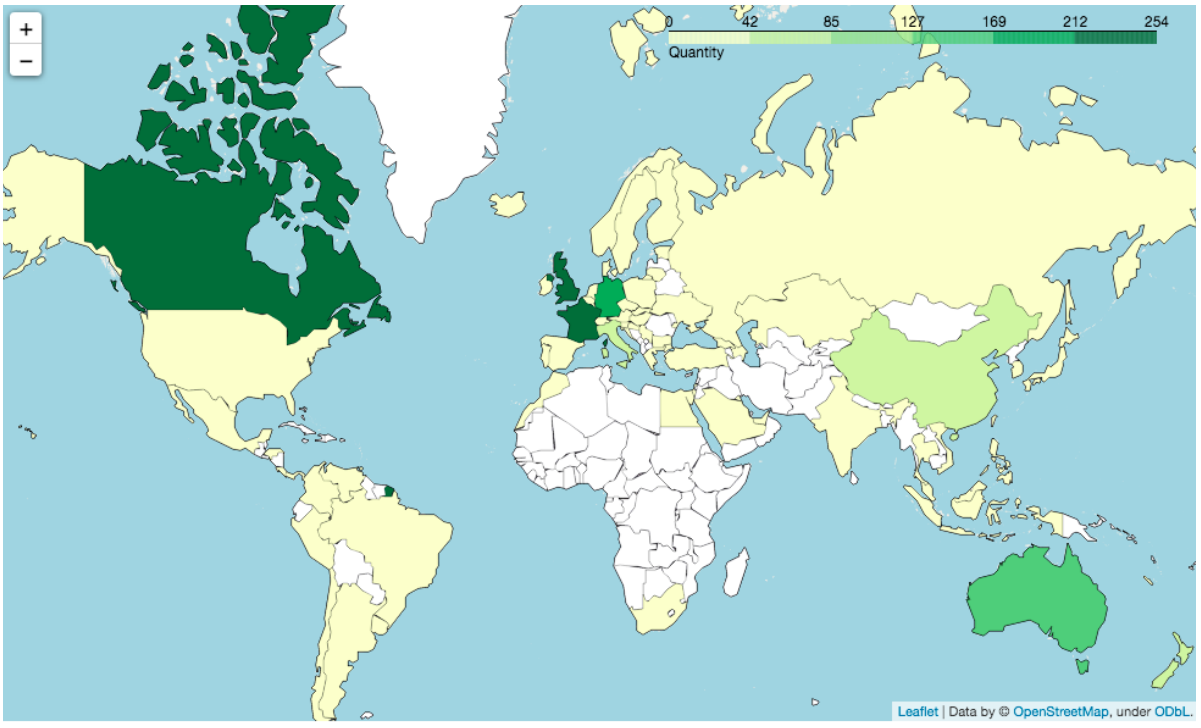


Fig 9.1, 9.2 : Where are the guests coming from?

On these two maps, we drew the maps of the countries the guests are coming from, using the dataset of 4.000 guests. The first map shows the repartition if we count the United States. We can see that there are so many guests that come from the U.S. that all countries are in yellow while the U.S. is in dark green. The U.S. has 2130 out of 4.109 guests and is the top country the guests are coming from, followed by

Canada with 254 guests. The second map shows what the map is like if we don't count the U.S., so that we can study more meticulously from which countries the guests are from. We can see that the only greener countries are Canada, the United Kingdom, and France, then Germany, Australia, and we have paler green for New Zealand and China. Thus, we see that most guests come from Western countries. Yellow countries are the ones that had at least one guest that came from it. We can see that most countries are in yellow, except for most of Africa and Middle East.

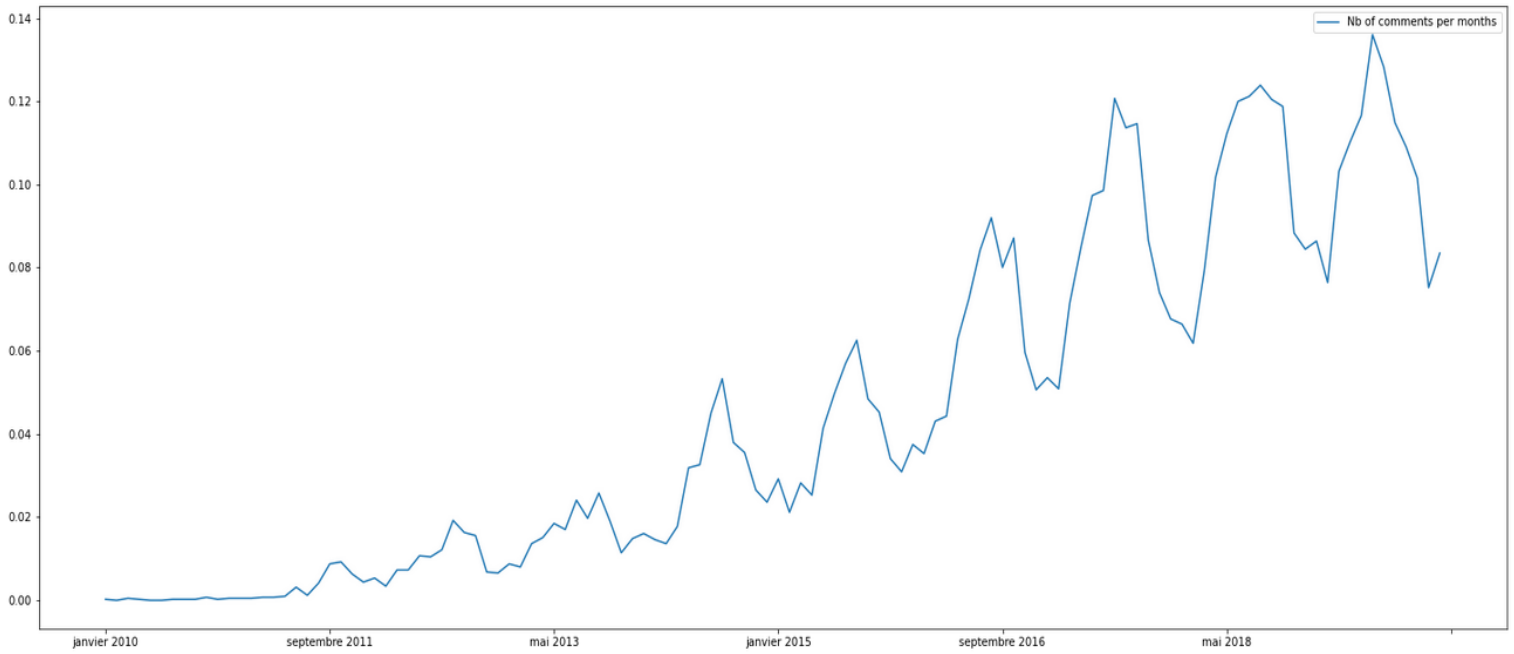


Fig. 10 : Evolution of the number of comments between 2010 and 2020 for guests who reviewed hosts that were present before and after 2017

This graph presents the total number of comments posted on the guest profiles we scraped. Each bar of this histogram corresponds to one month, from november 2010 to june 2021.These results were computed on the 4000 guests dataset.

We observe here that the number of comments increases over the years, until reaching a peak around 2018-2019. We can easily say that this rise is due to the development of Airbnb and the sharing economy in general.

There is no visible impact of the new San Francisco policy at the end of 2017 on this figure. But the policy on short-term rentals doesn't seem to have impacted significantly the number of comments posted on guests profiles.

We therefore looked for other variables that may illustrate more the consequences of the 2017 San Francisco law. In order to do that, we did a few regressions between the number of comments and other features, to see if some of these features had an

impact on the assessment of the guest (supposing that if a guest has a lot of comments, then that guest is appreciated).

Equation :

Nb of coms on guests profile = 0.0598*Nb of languages spoken + constant

	Coefficient	Standard error	p-value
Constant	0.6111	0.024	0.000
Number of languages spoken	0.0598	0.020	0.003

Fig.11 : Number of comments on guests profile depending on number of languages spoken

In that first OLS regression, we studied the dependency between the number of comments on guests profiles before 2017 depending on the number of languages spoken. We made sure that we took only the number of comments before the review. This study was made on the second dataset of 4.000 rows, where we discriminated guests. We see that the coefficient correlating the number of comments and languages spoken is of 0.0598 and is positive, meaning that the more languages a person speaks, the more comments this person gets. The p-value is equal to 0.003, which is inferior to 0.01, meaning that the value is reliable at the 5% percent level, and even at the 1% level, so we do not reject the hypothesis that “the coefficient correlating the number of comments and the number of languages spoken is at 0.0598”.

This could mean that a guest might be more appreciated if they speak more than one language, since they could be able to speak with the hosts that prefer to speak in their language. However, it could also mean that if a guest speaks a lot of languages, they are more likely to travel a lot and thus have a lot of comments from the Airbnb where they stayed.

Equation :

Nb of coms on guests profile = 0.0599*Nb of languages spoken -0.0004*ability to speak english + constant

	Coefficient	Standard error	p-value
Constant	0.6111	0.024	0.000
Number of languages spoken	0.0599	0.030	0.045
Ability to speak English	-0.0004	0.073	0.996

Fig. 12 : Number of comments on guests profile depending on number of languages spoken and ability to speak English

Now in this OLS regression, we studied the number of comments depending on the number of languages spoken and the ability to speak English, represented by a binary variable 1 (spoken) or 0 (not spoken).

We see that the coefficient for the languages spoken is 0.0599, which is positive, and the p-value is equal to 0.045, which means that the value is reliable at 5%, but not at 1% anymore. This value is less trustworthy than what we had earlier.

When we look at the second variable, we can see that the coefficient is negative at -0.0004, which would mean that if a guest does not speak English, he will have more likely a higher number of comments than someone who does speak English. This result seems to be counterintuitive, and when we look at the p-value, we can see that it is at 9,96%. Thus, this p-value is not reliable at 5%, and we can therefore reject the hypothesis “the coefficient correlating the number of comments and the ability to speak English is at -0.0004”.

Equation :

Nb of coms on guests profile = 0.0649*Nb of languages spoken + 0.0614*Guest living in the US + constant

	Coefficient	Standard error	p-value
Constant	0.6111	0.024	0.000
Number of languages spoken	0.0649	0.020	0.002
Guests living in the U.S.	0.0614	0.045	0.171

Fig.13 : Number of comments on guests profile depending on number of languages spoken and whether guest is living in the U.S.

In this third OLS, we studied the number of comments on guests profiles depending on the number of languages spoken and whether the guest is living in the U.S., represented by a binary variable (1 if yes, 0 if no).

Now, we can see that the coefficient for the number of languages spoken is positive at 0.0649, with a p-value of 0.2%. Thus, the value is reliable at 5% and even 1%, so we can accept the hypothesis for this value of the coefficient. It means that the more languages are spoken by a guest, the more comments there will be.

The coefficient of the binary variable indicating whether the guest lives in the United States is positive at 0.0614 with a p-value of 1.71%, so the value is also reliable at 5% (but not at 1%). We can also accept the hypothesis for this value of the coefficient. It means that if a guest is American (or lives in the U.S.), they will have more comments than if not.

The coefficients are almost equal, meaning that the two variables have a similar impact on the number of comments on guests' profiles. As said before, the number of languages could play on the number of comments, since hosts could particularly appreciate people that speak their language. If they live in the U.S., they could also have more comments because Americans could be more likely to leave comments on their hosts' dwelling and since comments are often reciprocated, hosts could also leave comments.

In order to see the impact of time on the number of comments, we created variables based on the review date of each reviewer in the original dataset. To do so, we used the month and the year of review dates, and transformed them into binary variables. For the review date '*12/06/2017*' for instance, the created variable is called '*062017*'. Then, for each reviewer, the variable '062017' has been set to 1 if they ever posted a review during this month, and 0 otherwise. And so on, for all the other dates between 01/2010 and 06/2021.

Therefore, each variable is a vector of length 4000 (one component for each user). With this set of new variables, we decided to compute the OLS between the number of comments per profile.

Here are the coefficients plotted on the same graph, after removing all the coefficients equal to 0 (it still remains coefficients that are very close to 0) :

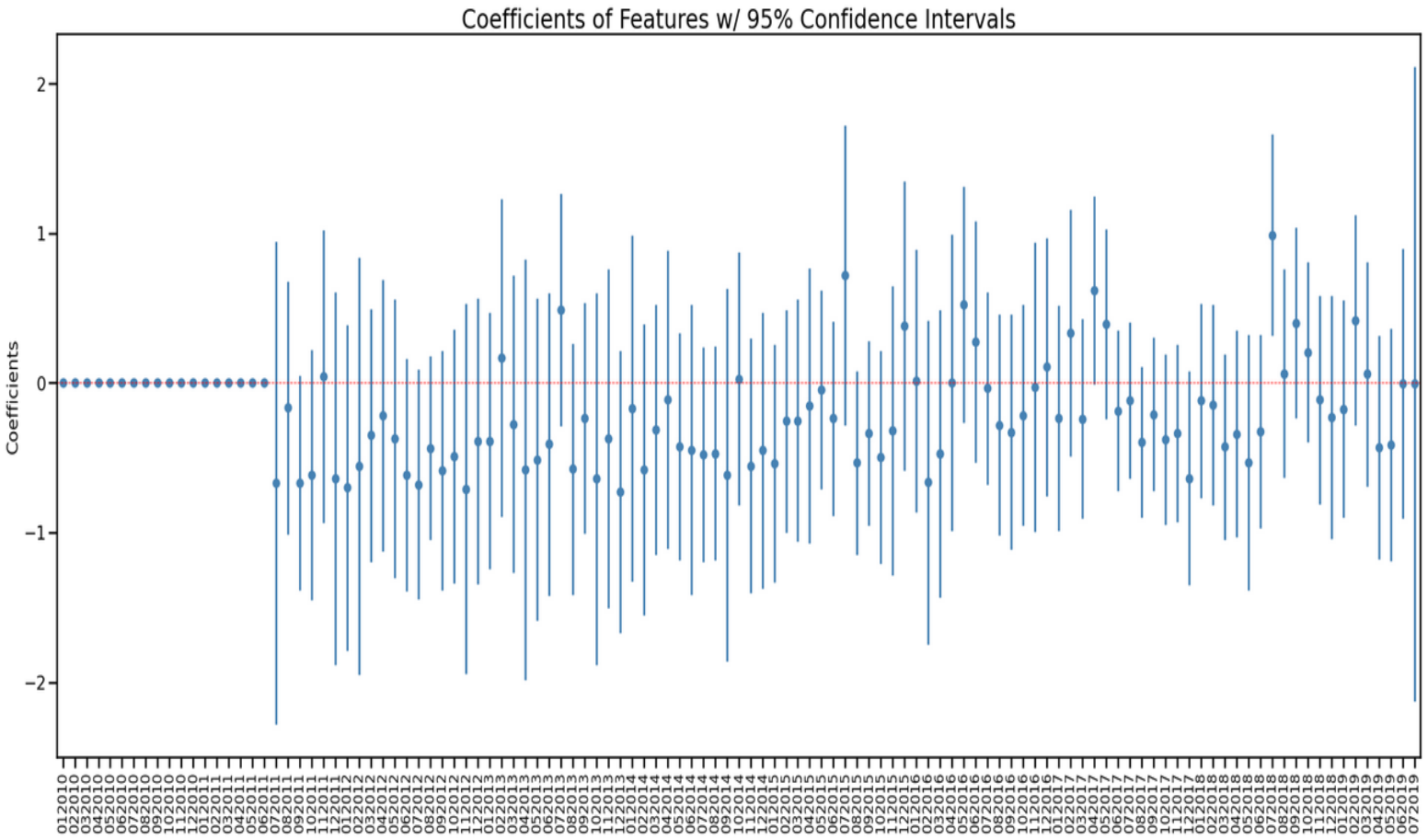


Fig.14 : Coefficient of number of comments depending on the date with 95% confidence interval

We observe that most coefficients are negative before 2016 (only 5 positive points before january 2016), and after the start of 2016, there are much more positive

coefficients. This could be explained by the high expansion of Airbnb during this period.

However, in 2017, the coefficients become negative. This might indicate the impact of the San Francisco law on the users. As many guests left the platform, and those who stayed were maybe less likely to post comments on guests profiles, because of the growing competition.

We also decided to compute the OLS between the english_speaker variable and these created variables, in order to see if the policy impacted the number of english speakers. But the results do not enable us to draw precise conclusions, as we do not see clearly consequences on the coefficients after 2017.

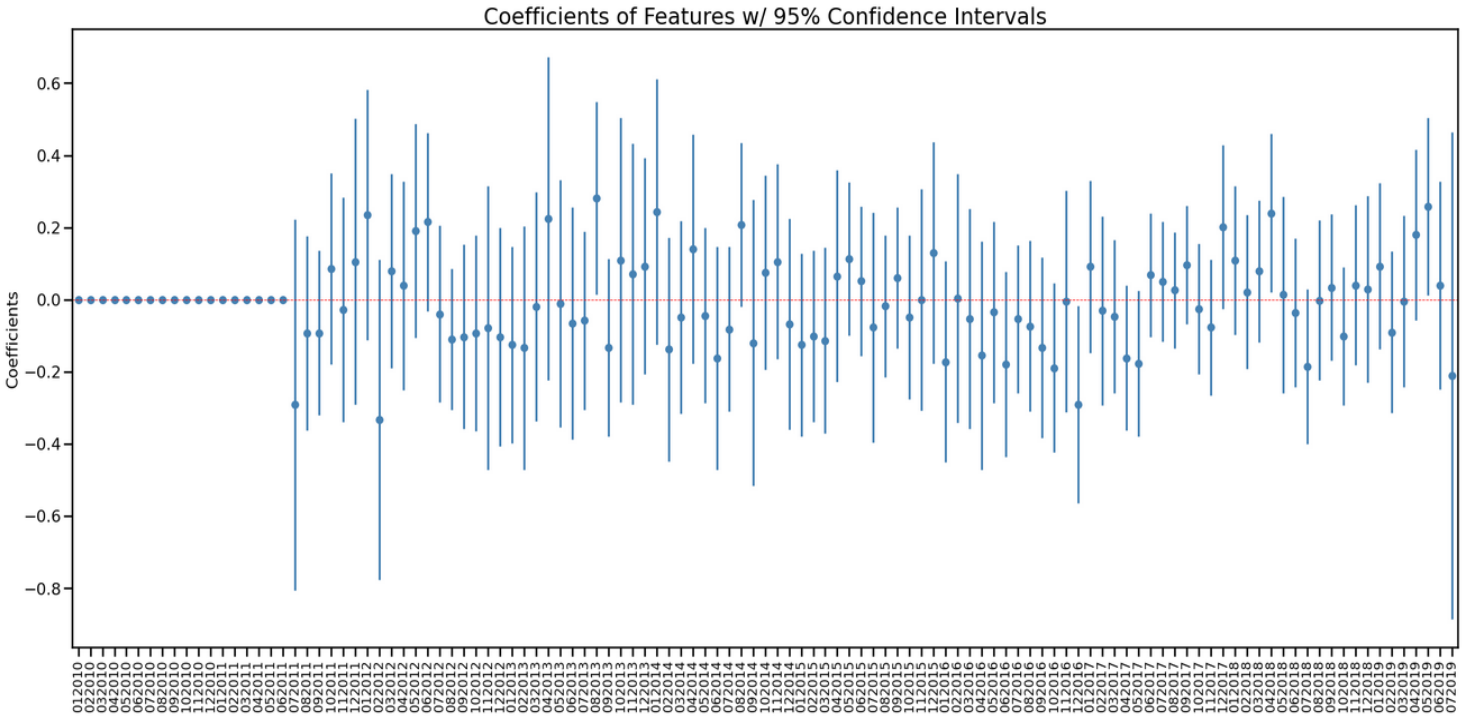


Fig.15 : Coefficient of binary variable english_speaker depending on the date with 95% confidence interval

VI - Conclusion

The study we conducted allowed us first to compute how many users account among 17.000 people survived to the Settlement Agreement. Approximately 4.000 people posted comments both before and after the law on short-term rentals.

We observed on these 4000 guests the influence of languages, ethnicity and time over the number of comments they received. After analysis, we did not conclude that the host market has become pickier.

However, the Coronavirus crisis made the interpretations of our different results more difficult. Some observations can be both interpreted as a consequence of the 2017 San Francisco law as well as the health crisis.

We also tried to regress prices and the average rating obtained by the host over the variables we scraped. It did not really give us some interpretable results so we decided not to show them in this study.