

**TÉLÉCOM PARIS**



**IP PARIS**

**MODS203 - Data Analysis in Economics I: Collection and Visualization**

Professors Ulrich Laitenberger, Maxime Cornet, Pierre-François Darlas and  
Guillaume Thébaudin

## **Analysis of Restaurant Competition in Paris**

Daniel Jorge Deutsch  
José Lucas Barretto  
Kevin Felipe Kühl Oliveira  
Lucas Miguel Celing Agrizzi

Paris, January 24th 2021

# Summary

<b>1. Introduction and Background</b>	<b>5</b>
<b>2. Data Collection</b>	<b>6</b>
<b>2.1 Restaurants</b>	<b>7</b>
2.2 Real reviews	9
<b>2.3 Fake Reviews</b>	<b>10</b>
<b>3. Data Summary</b>	<b>12</b>
<b>4. Data Analysis</b>	<b>13</b>
4.1 Assumptions and Hypothesis	13
4.2 Using Data to Support the Hypothesis	13
4.3 Fake Review Characteristics	15
4.3 Review Fraud and its Economic Incentives	17
4.3.1 Effects of Competition on Review Fraud	17
4.3.2 Effects of Own-Reputation on Review Fraud	18
4.4 Review and Reviewer Content Analysis	21
<b>5. Conclusion</b>	<b>24</b>
<b>6. References</b>	<b>25</b>
<b>Annex</b>	<b>26</b>

# Figures

**Figure 1.** Initial page from Yelp

**Figure 2.** Restaurants distribution over Paris (along with their rates) - for confirming variability

**Figure 3.** HTTP request of the hidden API that retrieves the restaurant reviews

**Figure 4.** Pattern of the URL of the hidden API

**Figure 5.** Pattern of the URL of the fake reviews

**Figure 6.** Inspection of the HTML of the not recommended reviews page

**Figure 7.** Origin of filtered reviews

**Figure 8.** Total reviews in Paris per year.

**Figure 9.** Percentage of fake reviews over the years in Paris.

**Figure 10.** Distribution of reviews fake and reals.

**Figure 11.** Percentage of fake reviews per price.

**Figure 12.** Percentage of fake reviews per average price.

**Figure 13.** Scatter plots for Positive/Negative Review Fraud vs Competition Faced by Restaurants

**Figure 14.** Effect of Own Reputation and Competition on Negative Review Fraud

**Figure 15.** Effect of Own Reputation and Competition on Positive Review Fraud

**Figure 16.** What is polarity and subjectivity - Image from [3]

**Figure 17.** Polarity analysis

# Tables

**Table 1.** Description of restaurants.

**Table 2.** Description of reviews.

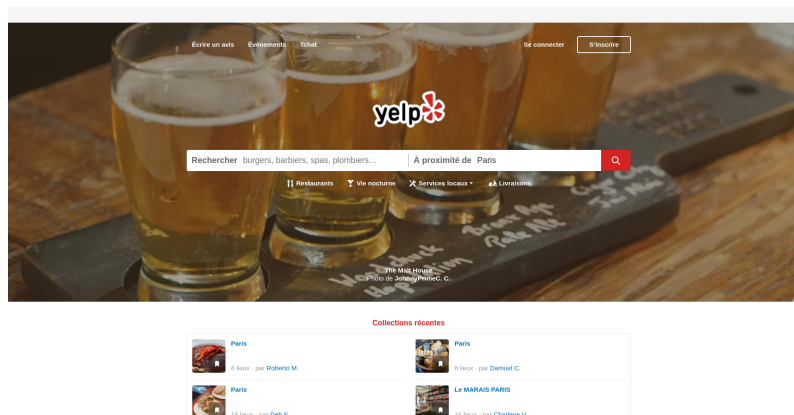
**Table 3.** Main statistics of the most important features.

**Table 4.** Statistics of review features.

# 1. Introduction and Background

Online reviews became a relevant source of information when shopping. Not only do users see what others think about products they bought but also how they evaluate places they visited. According to Statista Dossier on Online Reviews, published in 2019, 36% of global internet users aged between 25 to 34 years use online reviews for brand and product research. Furthermore, the same study reveals that 52% of the same population post reviews online. Specifically for the U.S. market, 19% of online users trust online customer reviews as much as personal recommendations. In that context, online reviews pose as important features regarding the user decision-making process.

For assessing and reviewing restaurants and other businesses (although less common in the latter case), users can make use of Yelp. Yelp is considered one of the most relevant platforms regarding restaurant evaluation, averaging more than 178 million unique visitors every month. According to the 2019 Online Reviews Survey, Yelp ranks right after Google on the tools consumers are likely to check before visiting a business.



*Figure 1. Initial page from Yelp*

In Yelp, everyone can open an account and start posting reviews about businesses. Although this facilitates genuine users that intend to help others, it is also easy for users to create fake reviews. Previous works (ref. [1] and [2]) on the theme show that fake reviews usually serve as a tool for lowering competitors rating (case in which we call them ‘negative fake review’) and raising its own esteem in the review community (‘positive fake review’).

To minimize the effects of fake reviews on users' decisions, Yelp launched, in 2014, a fake review detector that tries to filter these kinds of influences and mark them as 'not recommended review'. It works by considering parameters such as the user's activity frequency, the content of reviews, and their relevance.

We profit from this segregation in types of reviews to collect data and further understand the economic incentives behind fake reviews as well as the key points to differentiate a fake review from a real one. In sum, we plan to answer the following questions in our work:

- What are the economic incentives of fake reviews?
  - What restaurants are more prone to receive fake reviews?
    - Does the current rating affect the propensity?
    - Does price range affect the propensity?
    - Does competition affect the propensity?
- How do fake and real reviews differentiate between them?
  - What is the profile of a fake reviewer?
    - Does it have more or fewer friends in its network?
    - Is it an active profile in the community?
  - How do fake and real reviews differentiate in shape?
  - How do fake and real reviews differentiate in content?
    - Are fake reviews more extreme?
    - Is this polarity only within the rating domain or also in the writing domain?

To answer these questions we relied on the use of data visualization (which can provide some insights on possible correlations), as well as regressions, hypothesis testing and natural language processing techniques.

After several different analyses, we've managed to find out that, under our hypothesis that restaurants may use review fraud as a way of obtaining competitive advantages, restaurants with a lower price range and a lower rating tend to have a slightly higher percentage of fake reviews. While analyzing the effect of competition on small restaurants, we've also discovered that an increase in competition may slightly increase restaurants' engagement in review fraud (positive review fraud for self-benefit and negative review fraud to diminish competitors' reputation). We also concluded, by analyzing the content of reviews, that filtered reviews are more polarized (either very good or very bad) when it comes to the given rating and to the written content, when compared to non-filtered reviews.

## 2. Data Collection

This stage of the project is responsible for collecting all the data used throughout the project. We divided this section into three stages: ingestion, processing, and enhancement.

For the ingestion, since this stage consumes a lot of time, we would only execute it once and the data retrieved would be directly saved onto a .csv (without any modifications). This way, any further treatment regarding the obtained dataset would refer to this .csv, without the necessity of re-ingesting all the data. Since our project englobes many different types of information and each one of them is available through different methods (scrape, API, and hidden API), we built an ingestion process for each source and saved them onto different .csv files.

Once we had the raw data of each ingestion process, we started the processing of them. The goal of the processing is to drop unnecessary columns of the datasets and make useful data more accessible.

Each dataset resulting from the processing was submitted to an enhancement process with the goal of adding and inferring useful data.

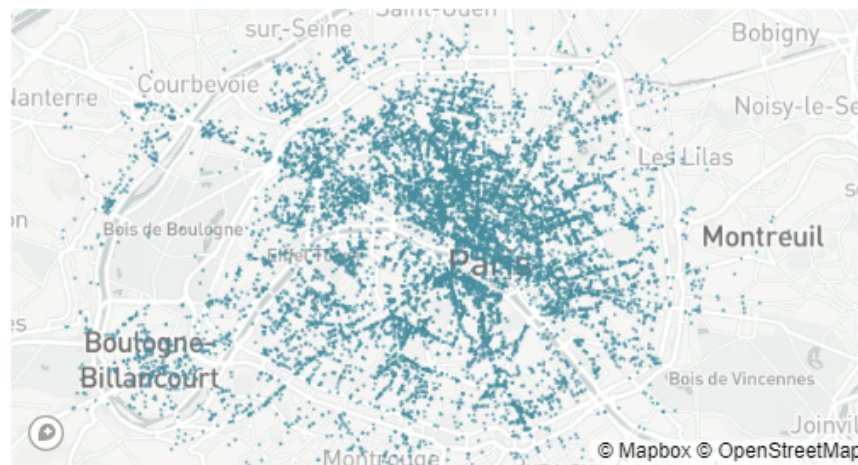
### 2.1 Restaurants

The goal here is to obtain a dataset containing as many Paris restaurants as possible. To do so we signed up for Yelp's developers API, which is available to everyone. Once you've registered, you receive an *api\_key* that must be sent in the headers of every request to the API in order to make sure only people with Yelp's developer account have access to their database.

However, despite presenting direct access to the data, the API has some limitations. The first problem encountered by the group was that the API limited the access to the first 1000 results of a given search. That is, if for a given set of parameters the number of results is greater than 1000, some entries would be lost. There is no way of setting the parameters to override this issue. What was implemented, to maximize the number of restaurants retrieved, was to iterate over the list of arrondissements and over the list of restaurant categories. This would possibly reduce the number of entries returned for a given query, allowing to retrieve more restaurants.

This choice of restaurant data extraction can be proved sensate observing the *figure* below, where a spatial plot of restaurants collected in Paris is presented. In this representation, one can note that restaurants are well distributed over the map. Some degree of concentration might be present around arrondissement centers (and in the city center, as well), but there are also a significant amount of businesses along borders and in decentralized regions.

### Distribution of Collected Restaurants in Paris



*Figure 2. Restaurants distribution over Paris (along with their rates) - for confirming variability*

With the given implementation, the group faced another important point, which was that a lot of restaurants were being repeated in two different sets of parameters. This mostly occurs because a restaurant can have multiple categories associated with it (can be ‘french’ and ‘cafe’, for example). We solved this by eliminating duplicates through the ID returned by the API.

Another challenge that was faced was the fact that each API account is only allowed to send 5000 requests to the API per day. That might seem a lot but, since we implemented a logic that sent 20 requests (each one returning 50 restaurants) for each (category, location) pair, it was not enough. The group estimated that it would need around 70.000 requests to iterate through both lists of arrondissements and categories. To avoid splitting out ingestion in several days, we decided to create as many developer accounts as necessary to complete the ingestion in one day. Then, we would only need to create a logic that would use a different *api\_key* once its requests limit was exceeded. This increased our collection pipeline to more than 70.000 entries a day, minimizing the time needed between tests and executions. The complete logic can be seen in *Annex 1*.

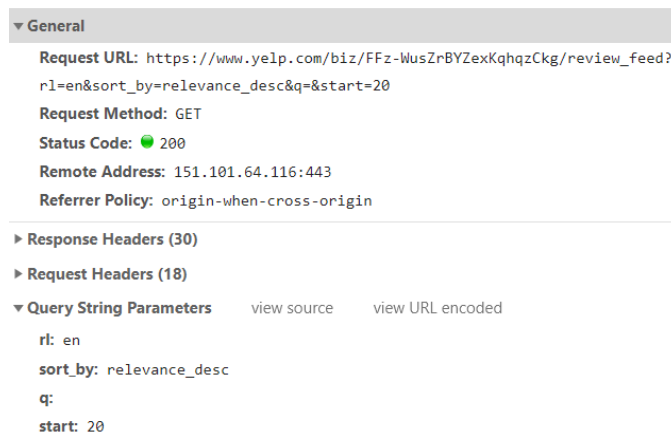


The API would return an array of businesses (restaurants) that match the filter criteria, in the format given by *Annex 2*. Once the ingestion of all restaurants was done, the *raw\_restaurants.csv* dataset had 13,184 restaurants (approximately one-third of the available restaurants in Paris).

After the ingestion, we processed the data because there were some columns not normalized which were difficult to work with, so we disregarded useless information and used a flatten function to extract only important information. We changed some data structures to better work with in the future and drop the other columns. Follow the steps that we made and the structure is in *Annex 3*. At this point, we extracted the coordinates from a restaurant from a list to columns with the latitude and longitude, changed the scale of prices from {*\$*, *\$\$*, *\$\$\$*, *\$\$\$\$*} to {1, 2, 3, 4}, dropped unnecessary columns like title and phone number and got the *arrondissement* from the address and the final structure is on *Annex 4*.

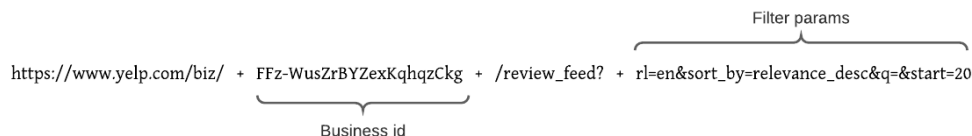
## 2.2 Real reviews

We noticed that Yelp's website sends the following HTTP request to list the reviews of a restaurant:



*Figure 3. HTTP request of the hidden API that retrieves the restaurant reviews*

One can easily see that the url used for the request has the following pattern:



*Figure 4. Pattern of the URL of the hidden API*

Our goal here is to obtain all the reviews written in #nglish or french (our analysis will only consider those) for each one of the collected restaurants. But the task isn't as simple as sending one request per restaurant and saving the retrieved data, we also need to consider the following:

- The above request uses pagination and the server only returns up to 20 reviews per page, so we need to send several requests per restaurant. After each request, we should increase the parameter *start* by 20 so we get the next page. We repeat that logic until we've collected all the reviews of the restaurant;
- Yelp has a security system that blocks our IP from using the website if the server registers unusual behavior from our part, i.e. if we send several requests in a row we get blocked from Yelp. To avoid that we should use a VPN service that changes the VPN we are connected in once we get blocked;

The logic used to collect restaurant's reviews explained above is illustrated in *Annex 5*.


After running the code we were able to gather 226,143 reviews that were directly saved (without any modification) onto the *raw\_reviews.csv* dataset with the following structure of *Annex 6*.

After this, we processed the data to extract business information and drop unnecessary columns like photos, user link, etc, and the inal structure is in *Annex 7*.

## 2.3 Fake Reviews

To collect the non-recommended reviews (which we will be calling from now on fake reviews), the process was a bit different. Yelp's developer API didn't retrieve any information about the falsehood of the review and we also weren't able to find anything about it in any hidden API, so the only way we could gather this data was by scraping the website.

Firstly we must pay attention to the url we want to scrape and understand its patterns:

`https://www.yelp.fr/not_recommended_reviews/ + l-as-du-fallafel-paris`  
  
Business alias

*Figure 5. Pattern of the URL of the fake reviews*

Once we have the URL, we should inspect the page to understand the patterns of the HTML. Through the inspection we were able to identify the HTML tags that contain the information we wanted and how we should extract them.

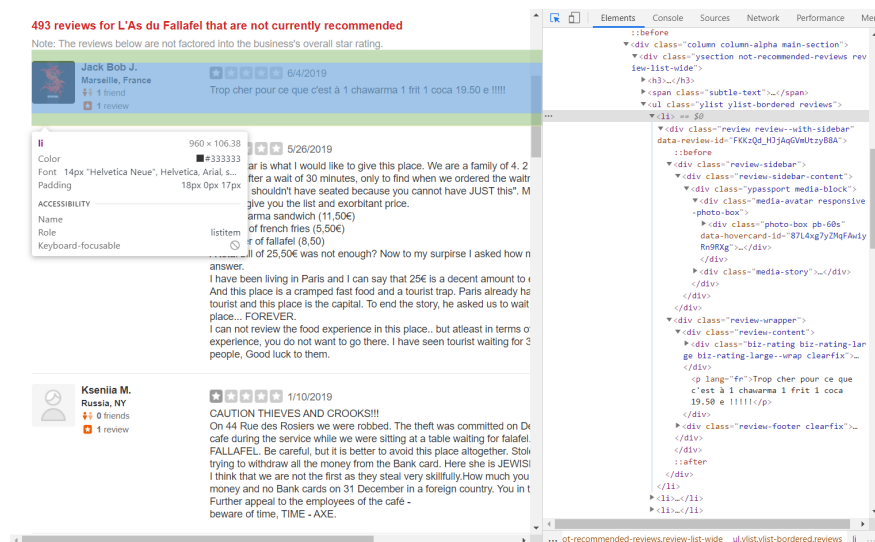


Figure 6. Inspection of the HTML of the not recommended reviews page

Having the HTML data extraction part figured out, we started facing the same difficulties we encountered when doing the ingestion of reviews: pagination and getting blocked from Yelp's server. Luckily, since we had already solved those problems before, it was easy to adapt the logic to this case, which resulted in the diagram of Annex 9.

As one can notice from the diagram of Annex 9, we decided to only collect the fake reviews that appeared on the first page of the not recommended reviews URL of each restaurant. That was done because we noticed that the ingestion of fake reviews required a lot of computational time because of the parsing of the HTML. This decision limited the analysis of section 5.2 to consider only small restaurants since famous restaurants tend to have way more than 10 fake reviews.

Once we ran the code, we were able to collect 32,869 fake reviews (approximately 80% of the total 40,966 available for the collected restaurants). These reviews were directly saved (without any modification) onto the *raw\_freviews.csv* dataset with the following structure of the Annex 10.

After obtaining the raw data we ran an enhancement code in which we dropped unnecessary columns and changed the way we accessed some of the data, and also created non-trivial information more accessible. The result of the enhancement can be found in Annex 11.

# 3. Data Summary

Once the data collection process is finished we can analyze what we managed to gather.  
For the restaurants, we had:

Feature	Mean	Mode	Std	Min	Max
review_count	17.448	1	42.514	1	1811
freview_count	3.109	0	7.301	0	492
treview_count	20.562	1	48.557	1	2303
freview_pct	0.168	0	0.188	0	0.941
rating	3.724	4	0.861	1	5
price	2.497	3	0.682	1	4
coordinates.latitude	48.862	48.853	0.025	48.686	50.731
coordinates.longitude	2.335	2.343	0.040	1.945	4.240
arrondissement	10.153	9	5.393	1	20
category	-	'french'	-	-	-

*Table 1. Description of restaurants.*

And for the reviews:

Feature	Mean	Mode	Std	Min	Max
rating	3.723	5	1.282	1	5
totalPhotos	0.521	0	1.467	0	35
user.reviewCount	113.157	1	291.108	1	14429
user.friendCount	100.195	0	770.496	0	22811
comment.language	-	fr	-	-	-
user.country.code	-	FRA	-	-	-
has_img	-	True	-	-	-

*Table 2. Description of reviews.*

# 4. Data Analysis

## 4.1 Assumptions and Hypothesis

Much of the following work has the intention of understanding human behavior inside Yelp's environment, and its real-world consequences. We propose some hypotheses for the collected data and make use of inference and some heuristics to assess correlations.

The main goal throughout the subsequent sections is to build a common ground on the incentives which play a relevant role on reviewing fraud and how those can impact businesses outside the digital world. We will also address the effects of review and reviewer characteristics on the filtering of the review.

One important hypothesis that we will use throughout the analyses is that restaurants can obtain competitive advantages if they successfully commit review fraud, by improving its own reputation or worsening its competitors'.

It's clear that the filtered reviews on Yelp are not all really fake. In fact, it is known that many of those filtered reviews are actually not fake reviews, even though Yelp's algorithm filtered them. This happens for many reasons, the main one being that Yelp's algorithm is not perfect, it may classify fake reviews as real and vice-versa.

Since it is out of the scope of this project to work on this fraud detection algorithm, we will lean on the hypothesis that filtered reviews are fake, however we acknowledge that filtered reviews do have a bias. For instance, if a restaurant with a high average grade contains many fake reviews, it is likely that most of the fake reviews will be biased towards a higher rating, due to flaws in the filtering system.

## 4.2 Using Data to Support the Hypothesis

By summarizing some key statistics of the collected data, we can already see that approximately 16.8% of all reviews are fake ones. The goal of this section is to analyse which restaurant characteristics are associated with a higher/smaller proportion of fake reviews.

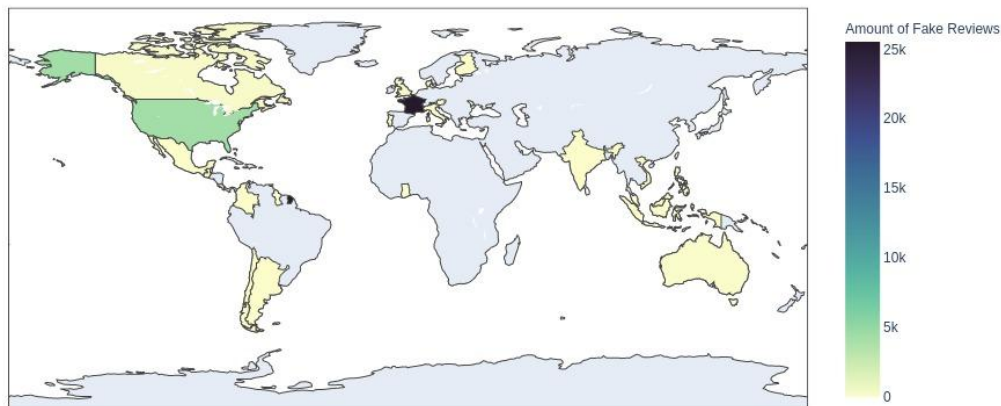
Feature	Mean (std)	Feature	Mean (std)
Number of total reviews	20.561	Rating of restaurants	3.723

per restaurant	(48.557)		(0.861)
Number of real reviews per restaurant	17.447 (42.513)	Price of restaurant	2.497 (0.68)
Number of fake reviews per restaurant	3.108 (7.300)	Percentage of fake review per restaurant	0.168 (0.188)

*Table 3. Main statistics of the most important features.*

To sustain our hypothesis that restaurants can obtain competitive advantages by committing review fraud, an important verification is whether most filtered reviews are local or not. As shown in the image below, we can see that most filtered reviews are located in France, and thus, this hypothesis cannot be neglected.

*Amount of Fake Reviews Made by Each Country*



*Figure 7. Origin of filtered reviews*

We can see in *Figure* the number of reviews in Yelp year by year: the platform was growing steadily, peaked in 2015, and started to lose a bit of its importance with the market entrance of other players like TripAdvisor and Google, and also due to litigious problems. These litigious problems were about restaurants gaming on the platform to defame their competitors with fake reviews. At the beginning of 2014, Yelp made public their filter of fake reviews in order to discourage review fraud, publicly exposing users and restaurants who were caught cheating on the platform. The graph below shows a dip in filtered reviews after this act - which discouraged the malicious users.

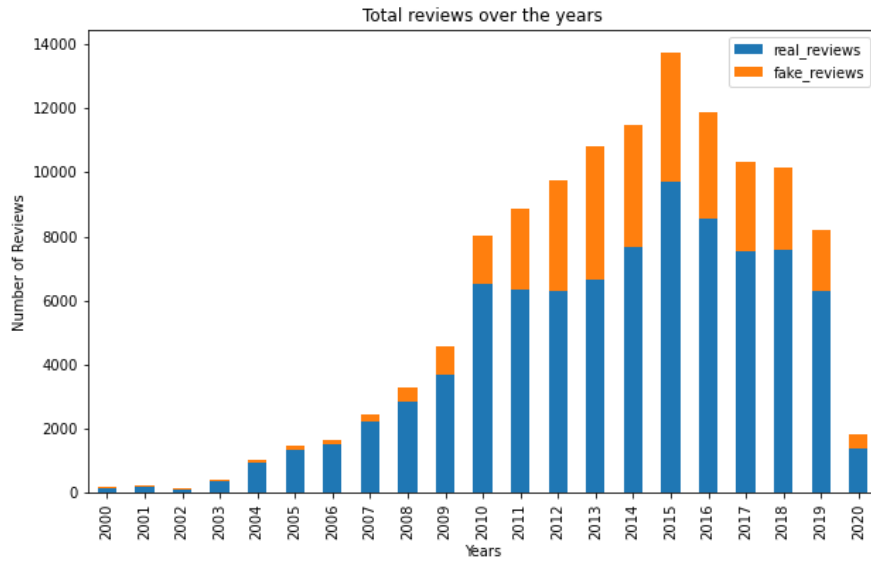


Figure 8. Total reviews in Paris per year.

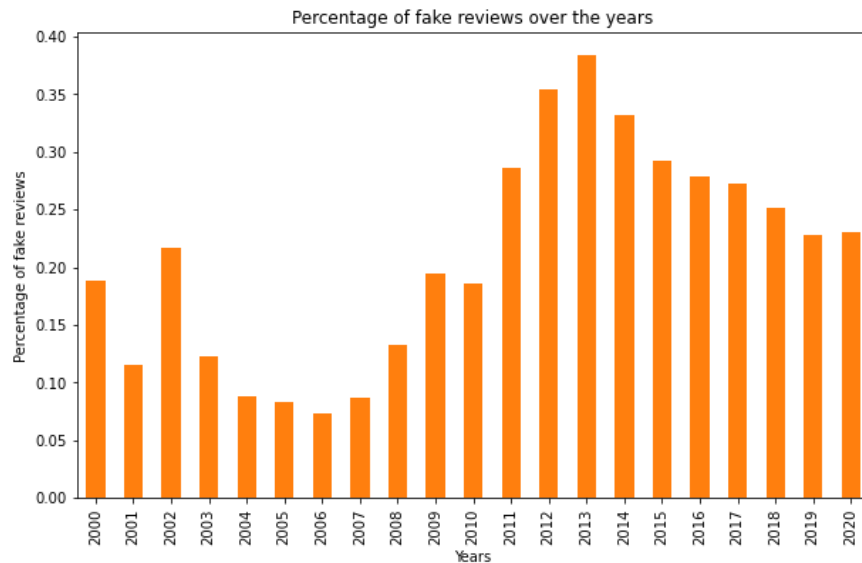


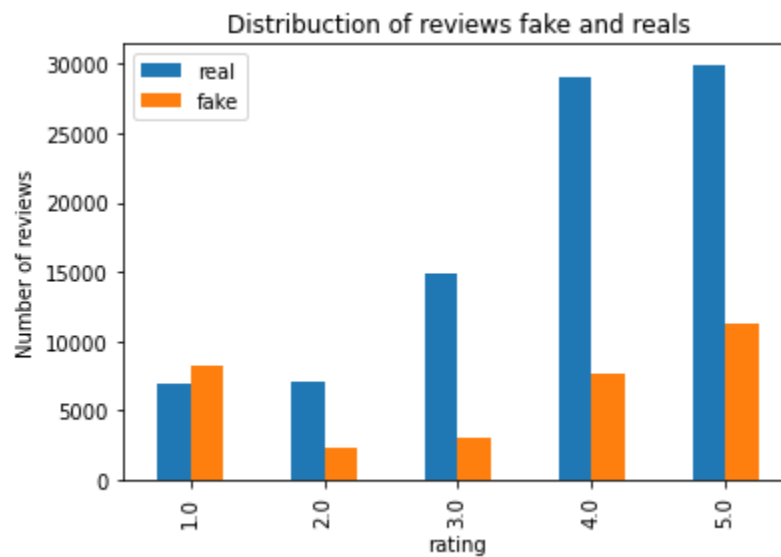
Figure 9. Percentage of fake reviews over the years in Paris.

This reduction is an indicator that, indeed, Yelp’s algorithm is filtering false reviews - otherwise, there would be no reason for this dip after the public disclosure of fraudulent users.

### 4.3 Fake Review Characteristics

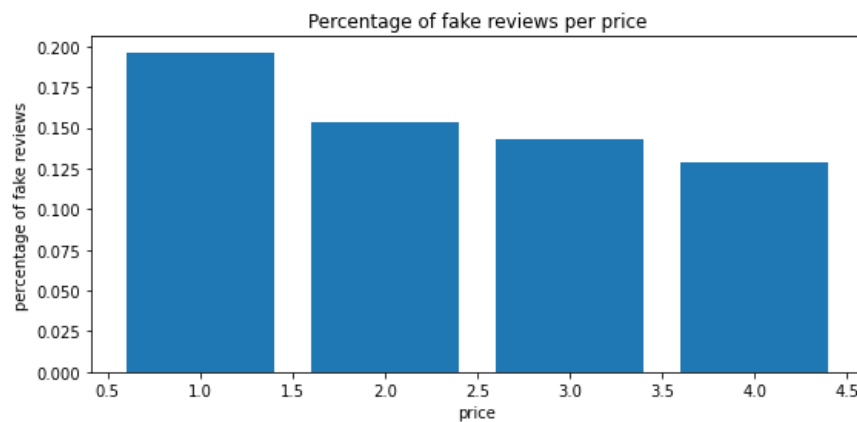
Here, we want to understand restaurants and review characteristics that indicate a higher/lower percentage of frauds.

By analyzing the distribution of fake and real reviews, we can see that filtered reviews are more polarized. This means that fake reviews are a lot more concentrated in the extremes: either they are extremely negative or extremely positive.



*Figure 10. Distribution of reviews fake and reals.*

We can also check the percentage of fake reviews by different types of restaurants. The chart below shows us that restaurants with a lower price range tend to have a slightly higher percentage of fake reviews.



*Figure 11. Percentage of fake reviews per price.*

As for the average rating, it is more difficult to make a statement. However, we can see in the chart below that, in general, the proportion of fake reviews tends to decrease as the restaurant rating increases. This could suggest that higher rated restaurants are less likely to engage in or receive fraudulent reviews.



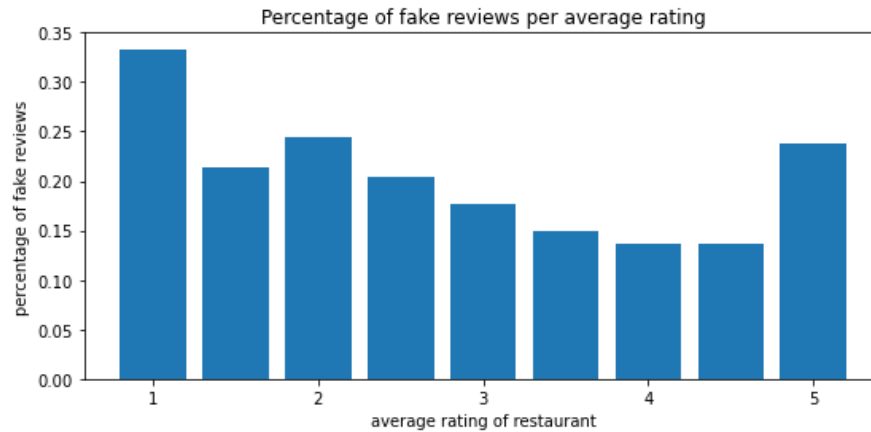


Figure 12. Percentage of fake reviews per average price.

### 4.3 Review Fraud and its Economic Incentives

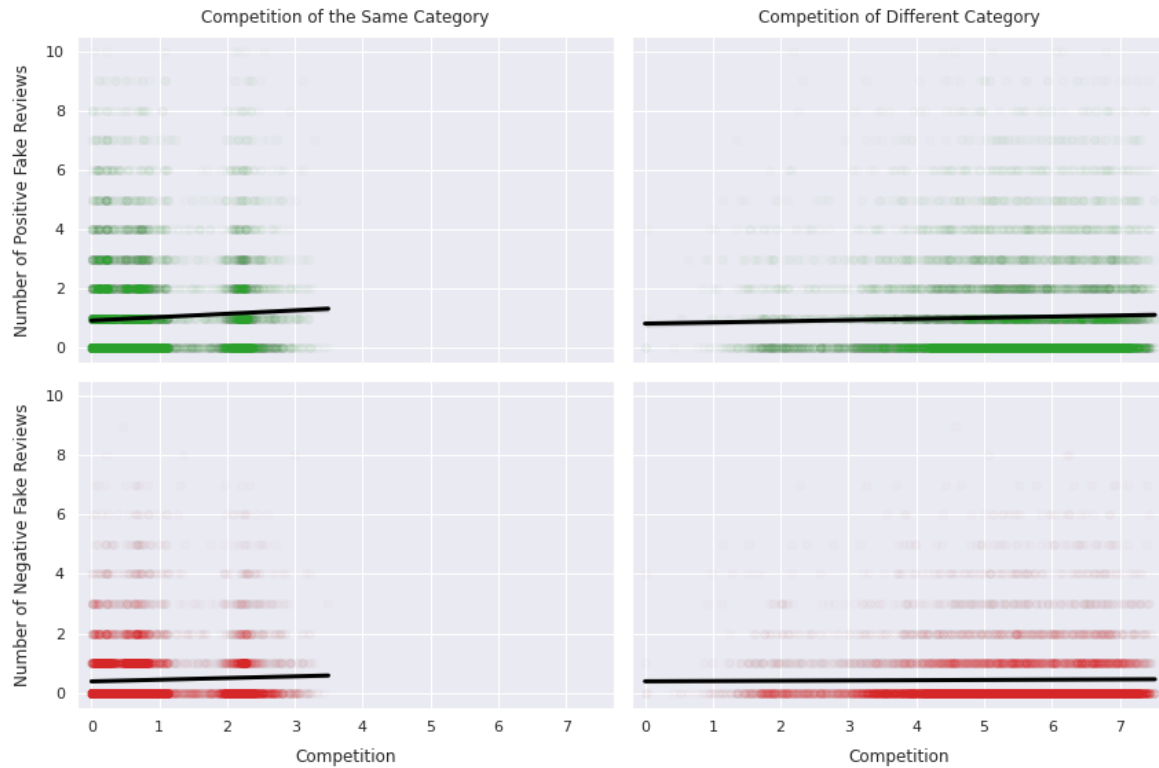
The goal of this section is to better understand the economic motivations behind review frauds. With that in mind, we decided to split our analysis into two main groups: negative fake reviews (rating  $\leq 2$ ) and positive fake reviews (rating  $\geq 4$ ). This aims to classify fake review motivations accordingly with our hypothesis - negative fake reviews are associated with restaurants that are trying to lessen its competitors' reputation, while positive fake reviews are associated with restaurants that are trying to improve their own reputation. Due to limitations in the data ingestion process, all the following analyses should not be considered for restaurants with a large number of total reviews.

We considered two main metrics that influence the number of positive or negative fake reviews that a restaurant has engaged in or received, respectively: the competition faced by the restaurant and its reputation.

#### 4.3.1 Effects of Competition on Review Fraud

The competition metric was calculated in a similar manner as in *Fake It Till You Make It: Reputation, Competition, and Yelp Review Fraud* [1]. As described in the article, the competition between two restaurants can be described as an inversely correlated function of the distance between them. We also decided to split competition between restaurants that share the same categories and restaurants that don't share categories.

After calculating the competition metrics, we observe that it has a small positive effect on the number of negative/positive fake reviews that restaurants receive, whether it is competition of the same category or of different categories.



*Figure 13. Scatter plots for Positive/Negative Review Fraud vs Competition Faced by Restaurants*

The image above contains scatter plots of positive/negative number of fake reviews vs same/different category competition. We can see that in every case, the regression line has a small positive coefficient. This points to the direction that an increase in competition (independent of category) may slightly increase restaurants' engagement in review fraud (positive review fraud for self-benefit and negative review-fraud to diminish competitors' reputation). We can also see that the effect of same-category competition is larger than the effect of different type competition (the regression line is more inclined).

#### 4.3.2 Effects of Own-Reputation on Review Fraud

Yelp's rating system provides valuable reputation information for restaurants, such as its average rating, price range and total number of reviews. All these metrics are contained in the dataset, and to analyze them we performed OLS regressions of the following kind:

$$y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \beta_4 X_4 + \beta_5 X_5$$

Where:

- $y$  is the target variable (number of **negative fake reviews** or number of **positive fake reviews**);
- $X_1$  is the metric for **competition of same category**;

- $X_2$  is the **metric for competition of different category**;
- $X_3$  is the **restaurant's average rating**;
- $X_4$  is the **restaurant's price range**;
- $X_5$  is the **total number of restaurant's reviews**.

Indeed, the restaurant's total review count is used so that the total number of fake reviews is well estimated - we are more concerned with the proportion of frauds on reviews than we are with absolute numbers. It is important to clarify once again that, since we filtered out most restaurants with enormous numbers of reviews, these analyses are more suited for mid to small restaurants.

The first regression had the number of negative fake reviews as target, which means we're only considering fake reviews with a rating equal or less than 2.0. The results are summarized in the figure below:

OLS Regression Results						
Dep. Variable:	negative_freview_count		R-squared:	0.214		
Model:	OLS		Adj. R-squared:	0.214		
Method:	Least Squares		F-statistic:	564.3		
Date:	Mon, 18 Jan 2021		Prob (F-statistic):	0.00		
Time:	20:42:33		Log-Likelihood:	-12888.		
No. Observations:	10368		AIC:	2.579e+04		
Df Residuals:	10362		BIC:	2.583e+04		
Df Model:	5					
Covariance Type:	nonrobust					
	coef	std err	t	P> t	[0.025	0.975]
const	1.0201	0.058	17.622	0.000	0.907	1.134
x1	4.735e-06	1.08e-05	0.439	0.661	-1.64e-05	2.59e-05
x2	-7.468e-06	6.32e-06	-1.181	0.237	-1.99e-05	4.92e-06
x3	-0.1654	0.009	-18.210	0.000	-0.183	-0.148
x4	-0.0866	0.012	-6.989	0.000	-0.111	-0.062
x5	0.0296	0.001	47.571	0.000	0.028	0.031
Omnibus:	5852.173	Durbin-Watson:	1.991			
Prob(Omnibus):	0.000	Jarque-Bera (JB):	65680.912			
Skew:	2.511	Prob(JB):	0.00			
Kurtosis:	14.261	Cond. No.	3.93e+04			

Figure 14. Effect of Own Reputation and Competition on Negative Review Fraud

Overall, this model did not perform very well, with a low  $R^2$  of 0.214. This may be due to the lack of consistency in the data for negative fake reviews (many restaurants have a negative fake review count of 0 and 1).

Although the competition metrics' coefficients ( $X_1$  and  $X_2$ ) have low statistical significance, we have verified that slight modifications in the metric's parameters make for significant changes in the p-value for the null hypothesis. This indicates that the metric may give us

some significant insights that we already stated in the previous section in a separate analysis (which is preferred when analyzing its standalone impact).

As for the effect of one's reputation, we can see that restaurants with higher average ratings (depicted by  $X_3$ ) and price ranges (depicted by  $X_4$ ) tend to have less negative fake reviews. The effect of average rating, however, may be biased due to flaws in Yelp's filtering algorithm - if a restaurant has more positive reviews, it's likely that more positive reviews will be mislabeled as fake. So again, it is hard to be deterministic when analyzing these coefficients.

The second regression has the number of positive fake reviews as target, which means we're only considering fake reviews with a rating greater or equal than 4.0. The results are summarized in the figure below:

OLS Regression Results						
Dep. Variable:	positive_freview_count		R-squared:	0.326		
Model:	OLS		Adj. R-squared:	0.326		
Method:	Least Squares		F-statistic:	1002.		
Date:	Mon, 18 Jan 2021		Prob (F-statistic):	0.00		
Time:	20:42:39		Log-Likelihood:	-17472.		
No. Observations:	10368		AIC:	3.496e+04		
Df Residuals:	10362		BIC:	3.500e+04		
Df Model:	5					
Covariance Type:	nonrobust					
	coef	std err	t	P> t	[0.025	0.975]
const	0.6150	0.090	6.828	0.000	0.438	0.792
x1	9.395e-06	1.68e-05	0.560	0.576	-2.35e-05	4.23e-05
x2	-1.726e-05	9.84e-06	-1.755	0.079	-3.65e-05	2.02e-06
x3	0.1302	0.014	9.209	0.000	0.102	0.158
x4	-0.2344	0.019	-12.158	0.000	-0.272	-0.197
x5	0.0637	0.001	65.721	0.000	0.062	0.066
Omnibus:	3961.247	Durbin-Watson:	1.975			
Prob(Omnibus):	0.000	Jarque-Bera (JB):	22391.602			
Skew:	1.744	Prob(JB):	0.00			
Kurtosis:	9.298	Cond. No.	3.93e+04			

Figure 15. Effect of Own Reputation and Competition on Positive Review Fraud

This model performed slightly better in terms of generalization, achieving a 0.326  $R^2$  rating. The effects of one's reputation (average rating and price) in positive review fraud are the inverse of those in negative review fraud, which makes sense. Again, it is hard to draw definitive conclusions, but this may indicate that higher priced and higher rated small restaurants are more likely to engage in positive review fraud.

## 4.4 Review and Reviewer Content Analysis

This part of data analysis consists in studying the content of reviews and the profile of users who have posted it. One wants to find the possible key factors that differentiate fake reviews from real reviews. The following table gives a general description of the review dataset analyzed.

	<b>Fake Reviews (25 556 (FR) + 6 445 (EN))</b>		<b>Real Reviews (131 376 (FR) + 94 767 (EN))</b>	
<b>Avg. Size (in characters)</b>	376.54 (French)	371.80 (English)	505.50 (French)	528.76 (English)
<b>Users avg. number of friends</b>	257.98		77.87	
<b>Users avg. number of reviews on Yelp</b>	19.16		126.46	
<b>Pct. 1-star</b>	25.65% (French)	24.83% (English)	7.01% (French)	9.06% (English)
<b>Pct. 2-stars</b>	7.51% (French)	5.94% (English)	9.45% (French)	5.82% (English)
<b>Pct. 3-stars</b>	9.50% (French)	6.92% (English)	21.63% (French)	10.61% (English)
<b>Pct. 4-stars</b>	24.61% (French)	18.17% (English)	38.21% (French)	26.30% (English)
<b>Pct. 5-stars</b>	32.74% (French)	44.14% (English)	23.70% (French)	48.22% (English)

*Table 4. Statistics of review features.*

From the summarized data in the table, we can trace the profile of a fake reviewer, who usually has a significantly higher number of friends and posts less reviews when compared to genuine review posters. Also, with respect to the content and shape of fake reviews, we observe that they tend to be smaller in size and are mostly distributed towards the extreme ratings (1 and 5 star reviews), while real reviews present a natural growing distribution towards higher ratings.

We decide to further explore one of the above mentioned characteristics: the polarity in fake reviews. As we are dealing with opinions, a certain level of polarity (negative or positive) is expected. Fake reviews, for instance, tend to be more polarized in terms of rates, occupying both levels of extreme ratings (low and high); can we observe the same pattern when running a sentiment analysis within the review content?

We propose the use of a natural language processing technique for sentiment analysis to monitor the polarity of a given review. Polarity, together with subjectivity, build the triangle of

sentiment analysis for a given text excerpt. The following image may be useful for visualizing their relationship.

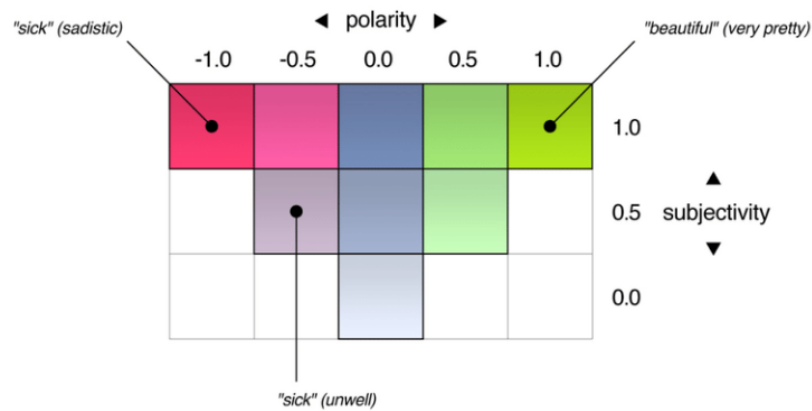
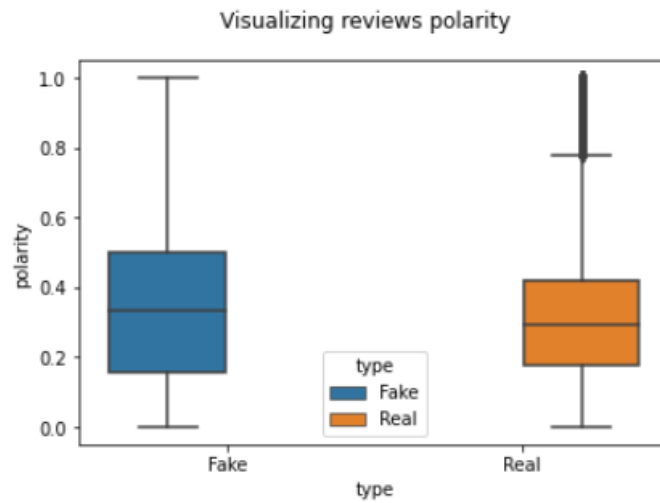


Figure 16. What is polarity and subjectivity - Image from [3]

As the English corpus for sentiment analysis is further developed than the French one, we propose to analyze the polarity of English reviews and observe if they follow the above described tendency. The process of sentiment analysis is based on a trained model, with a corpus from social media and/or review websites. After training the model with labeled data, it can be used to predict sentiment from other text excerpts. At the end, an analyzed excerpt has a polarity constant that goes from -1 (extremely negative) to 1 (extremely positive). Given its nature, it is sufficient to analyze the absolute value of the polarity score, given that we are in search of fake reviews that can be either positive or negative.

The following figure summarizes the main findings regarding fake and real reviews sentiment analysis. From that plot, we observe that the maximum value for the fake review polarity is 1 (meaning that a significant number of reviews are above the upper quartile), while for the real reviews, the maximum value is close to 0.8, with spare cases (outliers) between 0.8 and 1. Also, although the lower quartile from the fake reviews is similar to the lower quartile from the real reviews, the upper quartile is higher, which indicates an elevation in the average polarity.



*Figure 17. Polarity analysis*

This indicates a possible correlation between the polarity in the review's content and its classification as fake or real. Furthermore, it provides some ground for considering the fact that, indeed, as proposed, fake reviews have higher polarity when compared to real reviews.

# 5. Conclusion

Throughout the project, we managed to acquire some significant insights on how fake reviews behave according to the restaurant's own reputation and the competition that it faces, what are the motivations behind review fraud and how the content of fake reviews differ from those of the normal ones.

We started with some simple hypothesis: we consider that Yelp's filtering system is somewhat accurate, and therefore, filter out fake reviews, and that restaurants may use review fraud as a way of obtaining competitive advantages.

Our analyses indicate that restaurants with a lower price range and a lower rating tend to have a slightly higher percentage of fake reviews. Besides that, there are also indications that for small restaurants, an increase in competition may increase restaurants' engagement in review fraud (positive and negative). When analyzing text and rating of reviews, we've found that filtered reviews are more polarized than normal ones.

These conclusions, however, are far from definitive. We know that Yelp's filter is not perfect, which generates a bias in many of our analyses, given that not all filtered reviews are fake. We also conducted the competition analyses without considering large restaurants, due to limitations in the data collection process. Therefore, these conclusions are not general.

Potential future research to develop these ideas includes acquiring a larger dataset, with fake review content data for all restaurants. We also could develop an idea to further improve Yelp's filter with some heuristics, and use this improved filter as an indicator of review fraud.

As a group, we genuinely felt that this project completed our learning process. We were able to develop all the theories learned in online classes, while working on a relevant data analysis project. The following diagram summarizes all the steps taken in order to present this final result.

With the experience acquired, we envision plenty of projects and ideas where the concepts used can be useful as well.

Finally, we would like to thank our teacher Ulrich Laitenberger and his teaching team for working with us during this project, making it possible.

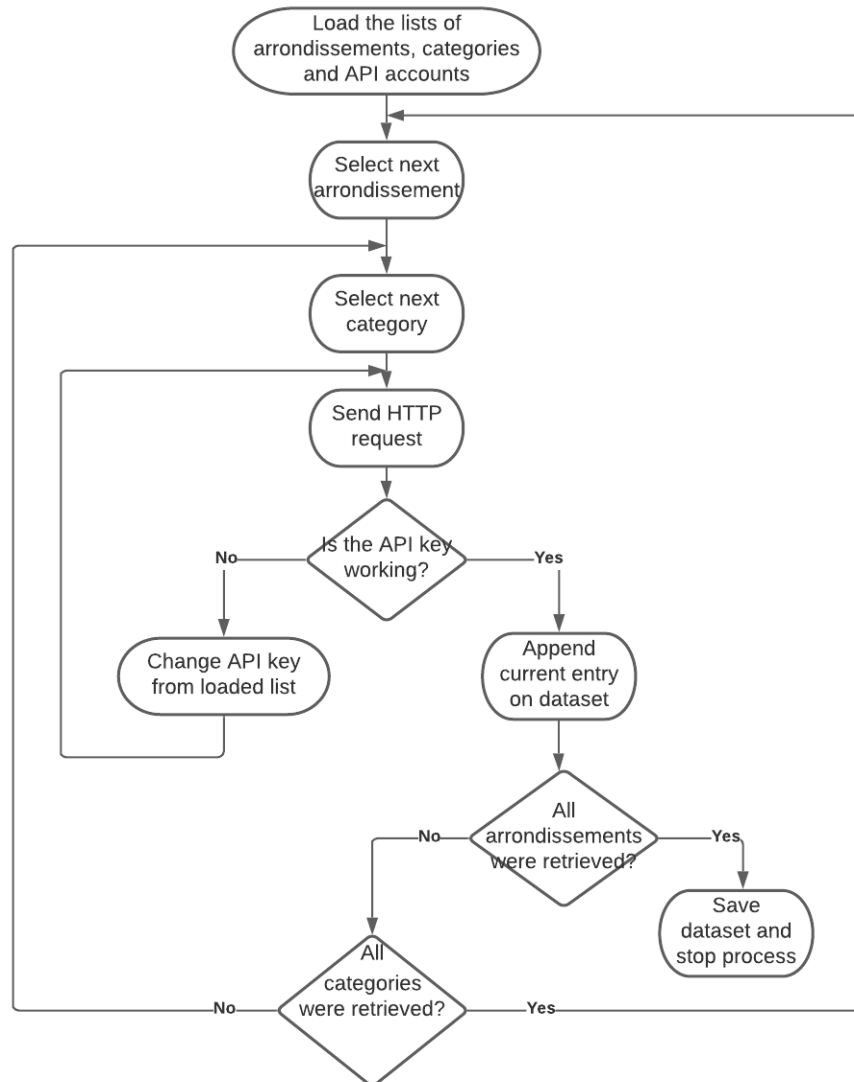


# 6. References

- [1] Luca, Michael & Zervas, Georgios. (2013). Fake It Till You Make It: Reputation, Competition, and Yelp Review Fraud. SSRN Electronic Journal. 62. 10.2139/ssrn.2293164.
- [2] Luca, Michael, Reviews, Reputation, and Revenue: The Case of Yelp.Com (March 15, 2016). Harvard Business School NOM Unit Working Paper No. 12-016, Available at SSRN: <https://ssrn.com/abstract=1928601>
- [3] De Smedt, Tom & Daelemans, Walter. (2012). " Vreselijk mooi!"(terribly beautiful): A Subjectivity Lexicon for Dutch Adjectives.. 3568-3572.Ing

# Annex

## Annex 1



## Annex 2

Feature	Description	Sample
id	Unique Yelp ID of this business	FFz-WusZrBYZexKqhgzCkg

alias	Unique Yelp alias of this business. Can contain unicode characters	l-as-du-fallafel-paris
name	Name of this business	L'As du Fallafel
image_url	URL of photo for this business	<a href="https://s3-media3.fl.yelpcdn.com/bphoto/QMNELSZ6-LzA9kLP3zQPgw/o.jpg">https://s3-media3.fl.yelpcdn.com/bphoto/QMNELSZ6-LzA9kLP3zQPgw/o.jpg</a>
is_closed	Whether business has been (permanently) closed	False
url	URL for business page on Yelp	<a href="https://www.yelp.com/biz/l-as-du-fallafel-paris?adjust_creative=gRtcZ6GuEdlQSO2t9PYnfg&amp;utm_campaign=yelp_api_v3&amp;utm_medium=api_v3_business_search&amp;utm_source=gRtcZ6GuEdlQSO2t9PYnfg">https://www.yelp.com/biz/l-as-du-fallafel-paris?adjust_creative=gRtcZ6GuEdlQSO2t9PYnfg&amp;utm_campaign=yelp_api_v3&amp;utm_medium=api_v3_business_search&amp;utm_source=gRtcZ6GuEdlQSO2t9PYnfg</a>
review_count	Number of reviews for this business	1811
categories	A list of category title and alias pairs associated with this business	[{'alias': 'kosher', 'title': 'Kosher'}, {'alias': 'sandwiches', 'title': 'Sandwiches'}, {'alias': 'falafel', 'title': 'Falafel'}]
rating	Rating for this business (value ranges from 1, 1.5, ... 4.5, 5)	4.5
coordinates	The coordinates of this business	{'latitude': 48.857498, 'longitude': 2.35908}
transactions	A list of Yelp transactions that the business is registered for. Current supported values are "pickup", "delivery", and "restaurant_reservation"	[]
location	The location of this business, including address, city, state, zip code and country	{'address1': '34 rue des Rosiers', 'address2': '', 'address3': '', 'city': 'Paris', 'zip_code': '75004', 'country': 'FR', 'state': '75', 'display_address': ['34 rue des Rosiers', '75004 Paris', 'France']}
phone	Phone number of the business	3.3148887636e+10
display_phone	Phone number of the business formatted nicely to be displayed to users. The format is the standard phone number format for the business's country	+33 1 48 87 63 60
distance	Distance between the business and the place that sent the HTTP request	1804.814778
price	price	€

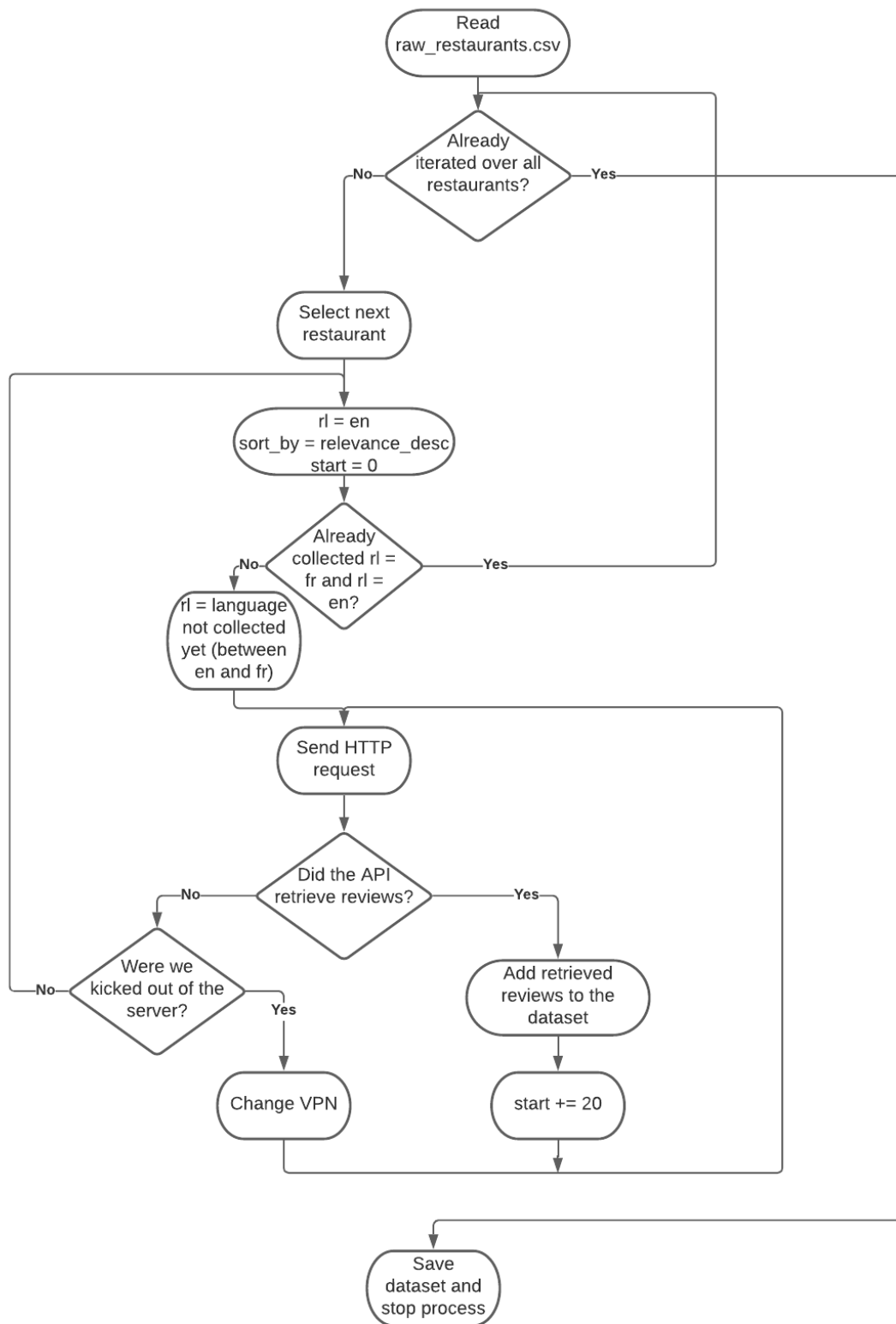
## Annex 3

Feature	Description	Sample
id	Unique Yelp ID of this business	FFz-WusZrBYZexKqhgzCkg
alias	Unique Yelp alias of this business. Can contain unicode characters	l-as-du-fallafel-paris
name	Name of this business	L'As du Fallafel
image_url	URL of photo for this business	<a href="https://s3-media3.fl.yelpcdn.com/bphoto/QMNELSZ6-LzA9kLP3zQPgw/o.jpg">https://s3-media3.fl.yelpcdn.com/bphoto/QMNELSZ6-LzA9kLP3zQPgw/o.jpg</a>
is_closed	Whether business has been (permanently) closed	False
review_count	Number of reviews for this business	1811
categories	A list of category title and alias pairs associated with this business	['wok', 'japanese food']
rating	Rating for this business (value ranges from 1, 1.5, ... 4.5, 5)	4.5
coordinates.latitude	Latitude of this business	48.857498
coordinates.longitude	Longitude of this business	2.35908
arrondissement	Arrondissement	2
price	price	4

## Annex 4

Feature	Description	Sample
id	Unique Yelp ID of this business	FFz-WusZrBYZexKqhgzCkg
alias	Unique Yelp alias of this business. Can contain unicode characters	l-as-du-fallafel-paris
name	Name of this business	L'As du Fallafel
is_closed	Whether business has been (permanently) closed	False
review_count	Number of reviews for this business	1811
categories	A list of category title and alias pairs associated with this business	['wok', 'japanese food']
rating	Rating for this business (value ranges from 1, 1.5, ... 4.5, 5)	4.5
coordinates.latitude	Latitude of this business	48.857498
coordinates.longitude	Longitude of this business	2.35908
arrondissement	Arrondissement	2
price	Price	4
freview_count	Number of fake reviews from restaurant	30
freview_pct	Percentage of fake reviews over the total	0.45
treview_count	Total number of reviews from restaurant	70

## Annex 5



## Annex 6

Feature	Description	Sample
id	Unique Yelp ID of this review	QbuG1xu5163tMaPT1ncJOw
comment	The comment of the review, including the text and the language in which it was written	{'text': 'Le meilleur fallafel Le Fallafel est probablement le meilleur testé à Paris, cela permet d&#39;oublier l&#39;accueil pas terrible. Une bonne adresse pour un sandwich dans le Marais.', 'language': 'fr'}
rating	Rating given for the business in the review (value ranges from 1, 1.5, ... 4.5, 5)	3
photosURL	Endpoint of the user's photos	/biz_photos/l-as-du-fallafel-paris?userid=_xx7UK9SrjZ1K5r3V6eMjA
feedback	User's return on Yelp's standard reactions to a given restaurant (useful, funny, cool...)	{'counts': {'funny': 0, 'useful': 0, 'cool': 0}, 'userFeedback': {'funny': False, 'useful': False, 'cool': False}, 'voterText': None}
business	JSON containing the informations about the restaurant object from the review	{'alias': 'l-as-du-fallafel-paris', 'id': 'FFz-WusZrBYZexKqhqcKg', 'photoSrc': 'https://s3-media0.fl.yelpcdn.com/bphoto/QMNELSZ6-LzA9kLP3zQPgw/60s.jpg', 'name': 'L'As du Fallafel'}
localizedDateVisited	Empty column. Would represent the date from the user's visit to the restaurant	NaN
businessOwnerReplies	Replies from the business owner to the given review	NaN
userId	Unique identifier of the user	_xx7UK9SrjZ1K5r3V6eMjA
previousReviews	All the previous reviews from the given user	NaN
lightboxMediaItems	JSON containing information such as the number of reviews the user has already done and the number of friends (easy to access information for being in JSON format)	[]
photos	Posted photos on the given review	[]
tags	Tags returned by the API, they are redundant to other	[]

	information previous described and sometimes they describe internal properties of the review	
isUpdated	Indicates that the present review is an update of a previous review of the same user on the same restaurant	False
user	JSON containing the user's information	{'src': 'https://s3-media0.fl.yelpcdn.com/assets/srv0/yelp_styleguide/514f6997a318/assets/img/default_avatars/user_60_square.png', 'reviewCount': 10, 'altText': 'Cityvox User (cybk...)', 'friendCount': 0, 'displayLocation': 'Paris, France', 'markupDisplayName': 'Cityvox User (cybk...)', 'userUrl': None, 'partnerAlias': 'cityvox', 'eliteYear': None, 'photoCount': None, 'link': '', 'srcSet': None}
appreciatedBy	Contains information of users that feel helped with the given review	NaN
totalPhotos	Total number of photos posted for the given review	0
localizedDate	Date when the review was posted	5/14/2009

## Annex 7

Feature	Description	Sample
rating	Rating given for the business in the review (value ranges from 1, 1.5, ... 4.5, 5)	3
totalPhotos	Number of photos in comment	0
comment.text	The comment text	'It was a 2 man show. The bartender and cook b...'
date	Date of review	10/26/2010
is_fake	Bool if it's fake (all False)	False
business.alias	Unique Yelp alias of this business. Can contain unicode characters	l-as-du-fallafel-paris
comment.language	Language of comment	en

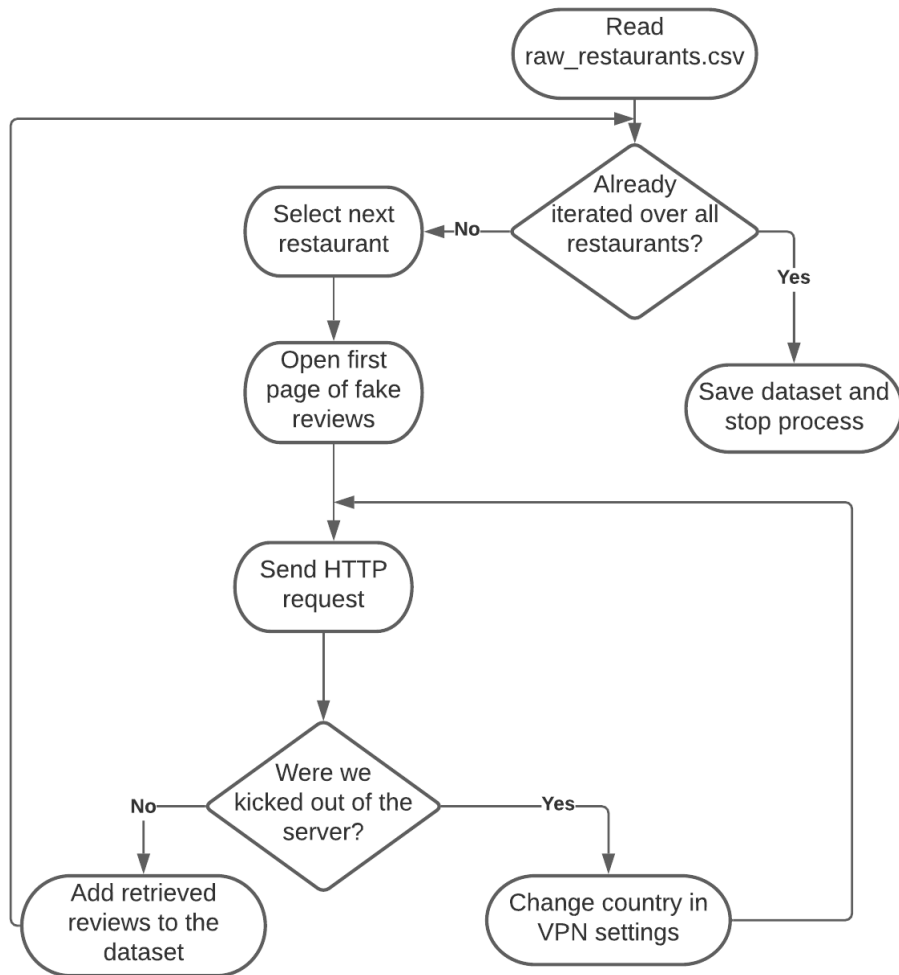


user.reviewCount	Number of reviews from this user	101
user.friendCount	Number of friends of user	2
user.displayLocation	Location from user	Danville, CA
has_img	Bool if the user has image	True

## Annex 8

Feature	Description	Sample
rating	Rating given for the business in the review (value ranges from 1, 1.5, ... 4.5, 5)	3
totalPhotos	Number of photos in comment	0
comment.text	The comment text	'It was a 2 man show. The bartender and cook b...'
date	Date of review	10/26/2010
is_fake	Bool if it's fake (all False)	False
business.alias	Unique Yelp alias of this business. Can contain unicode characters	l-as-du-fallafel-paris
comment.language	Language of comment	en
user.reviewCount	Number of reviews from this user	101
user.friendCount	Number of friends of user	2
user.country.code	User's country	US
has_img	Bool if the user has image	True

## Annex 9



## Annex 10

Feature	Description	Sample
comment.text	Text written by the user in the review	Aberrant et lamentable, je suis tombée malade à cause de ce falafel
business.alias	Unique Yelp alias of this business. Can contain unicode characters	I-as-du-fallafel-paris
rating	Rating given for the business in the review (value ranges from 1, 1.5, ... 4.5, 5)	1
user.displayLocation	Place where the user is from	San Francisco, États-Unis
user.friendCount	Number of friends the user has on the platform	0
user.reviewCount	Number of reviews the user has made in the platform	5
has_img	Whether the user has a profile image or not	True
is_fake	Whether the review is fake or not	True

## Annex 11

Feature	Description	Sample
rating	Rating given for the business in the review (value ranges from 1, 1.5, ... 4.5, 5)	3
totalPhotos	Number of photos in comment	0
comment.text	The comment text	'It was a 2 man show. The bartender and cook b...'
date	Date of review	10/26/2010
is_fake	Bool if it's fake (all False)	False
business.alias	Unique Yelp alias of this business. Can contain unicode characters	l-as-du-fallafel-paris
comment.language	Language of comment	en
user.reviewCount	Number of reviews from this user	101
user.friendCount	Number of friends of user	2
user.country.code	User's country	US
has_img	Bool if the user has image	True