TÉLÉCOM PARIS

MODS203 - Data Analysis in Economics I: Collection and Visualization

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

Yelp

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

Paris, January 18th 2021

Summary

Figures

Tables

1. Introduction

The digital world can give us as much information as we want, sometimes this information is right in front of our eyes and sometimes it isn't. The Internet has an unmeasurable power as a large part of information in the world is stored there and the choice of how much of that information will be consumed is up to the user, that decides how much time, money (and maybe luck) it is worth to spend in search of it.

This project has the goal to extract, process and analyze accessible information, but on a scale that normally a human doesn't have the capacity to process alone. From that data we hope to find some relationships, tendencies and fittings which could, somehow, explain the behavior of a group of individuals in platforms (its incentives and real world consequences). We used Yelp's platform to collect data over the restaurant market in Paris.

Yelp is a platform where users can evaluate their experiences in restaurants (and other businesses), contributing to a database that can help everyone make decisions on whether or not to visit a given place. In that context, usually well evaluated restaurants attract more clients, and bad restaurants don't have much attractiveness. This points out the benefits of such a platform, as people now have information that they didn't have before. This changed not only the environment for users, but also for restaurants, as now the competition isn't only in the streets, but also on the internet.

Given this new sphere of competition, new techniques for artificial growth of esteem are constantly put in place (just like in the real world, but now with the fast tools the internet provides). In Yelp, this relates to the use of fake reviews to higher its own rating (called 'positive fake reviews') and/or lower competitors' ratings (called 'negative fake reviews').

Yelp provides us with data classified by their proprietary fake reviews detector, which we will use to compare with real reviews extracted from restaurant pages.

Different techniques will be presented along this report. We did our best to explain the use of each one at the same time as keeping things inside the scope.

We hope that this work serves as a ground and motivation for future and more advanced analysis.

2. Data Ingestion

This stage of the project is responsible for collecting all the data used throughout the project. The idea is that, since each ingestion process consumes a lot of time, so we would only execute the ingestion 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.

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 (and their rating) 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.

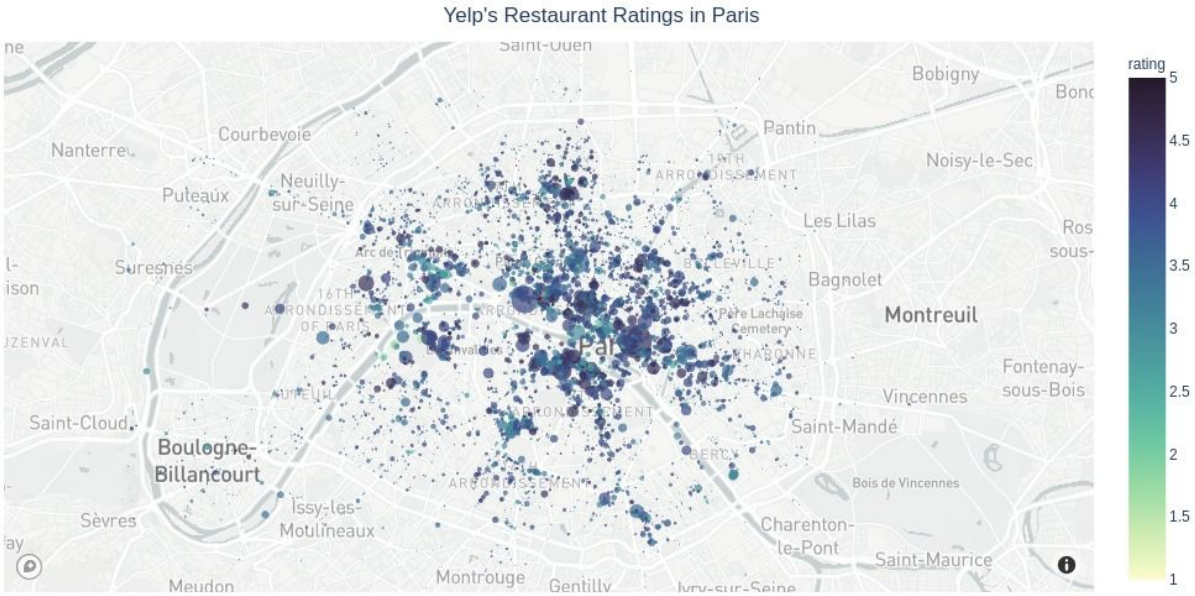

Figure 1. Restaurants distribution over Paris (along with their rates) - for confirming variability

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 Figure 2.

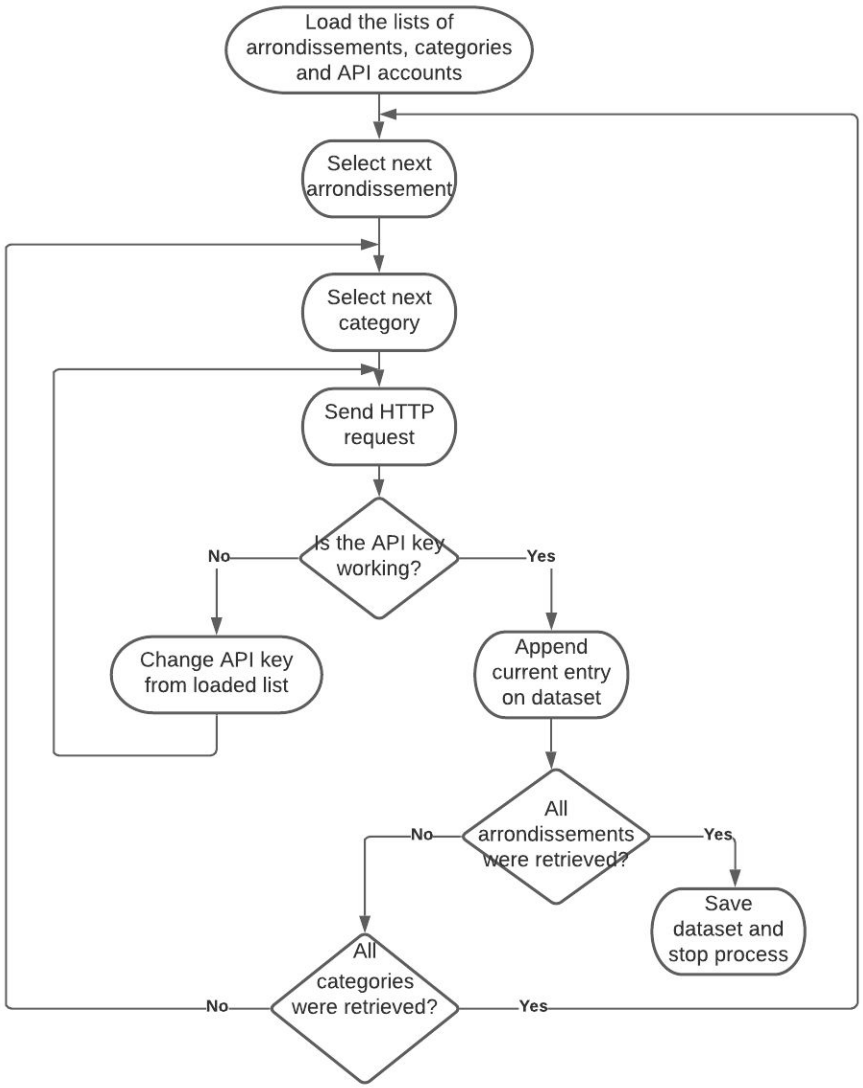

Figure 2. Logic behind the ingestion of restaurants

The API would return an array of business (restaurants) that match the filter criteria, in the following format:

Feature	Description	Sample
id	Unique Yelp ID of this business	FFz-WusZrBYZexKqhqzCkg
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	https://s3-media3.fl.yelpcdn.com/bphoto/QMNELSZ6-LzA9kLP3zQPgw/o.jpg
is_closed	Whether business has been (permanently) closed	False
url	URL for business page on Yelp	https://www.yelp.com/biz/l-as-du-fallafel-paris?adjust_creative=gRtcZ6GuEdlQSO2t9PYnfg&utm_campaign=yelp_api_v3&utm_medium=api_v3_business_search&utm_source=gRtcZ6GuEdlQSO2t9PYnfg
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	€

Table 1. Characteristics of the business returned from Yelp's public API

Once the ingestion of all restaurants was done, the *raw_restaurants.csv* dataset had 13,184 restaurants.

2.2 Real reviews

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

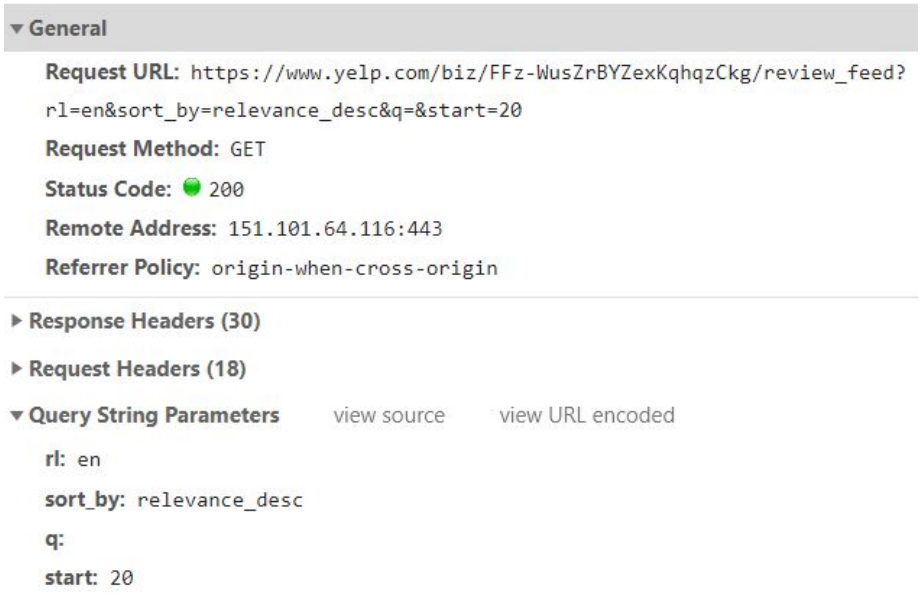

Figure 3. HTTP request of the hidden API that retrieves the restaurant reviews

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

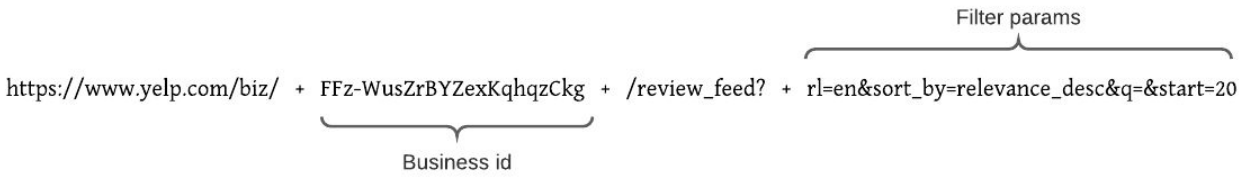

Figure 4. Pattern of the URL of the hidden API

Our goal here is to obtain all the reviews written in english or french (our analysis will only consider those) for each one of the collected restaurants. But the task isn't as simple as send one request per restaurant and save 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 the figure below.

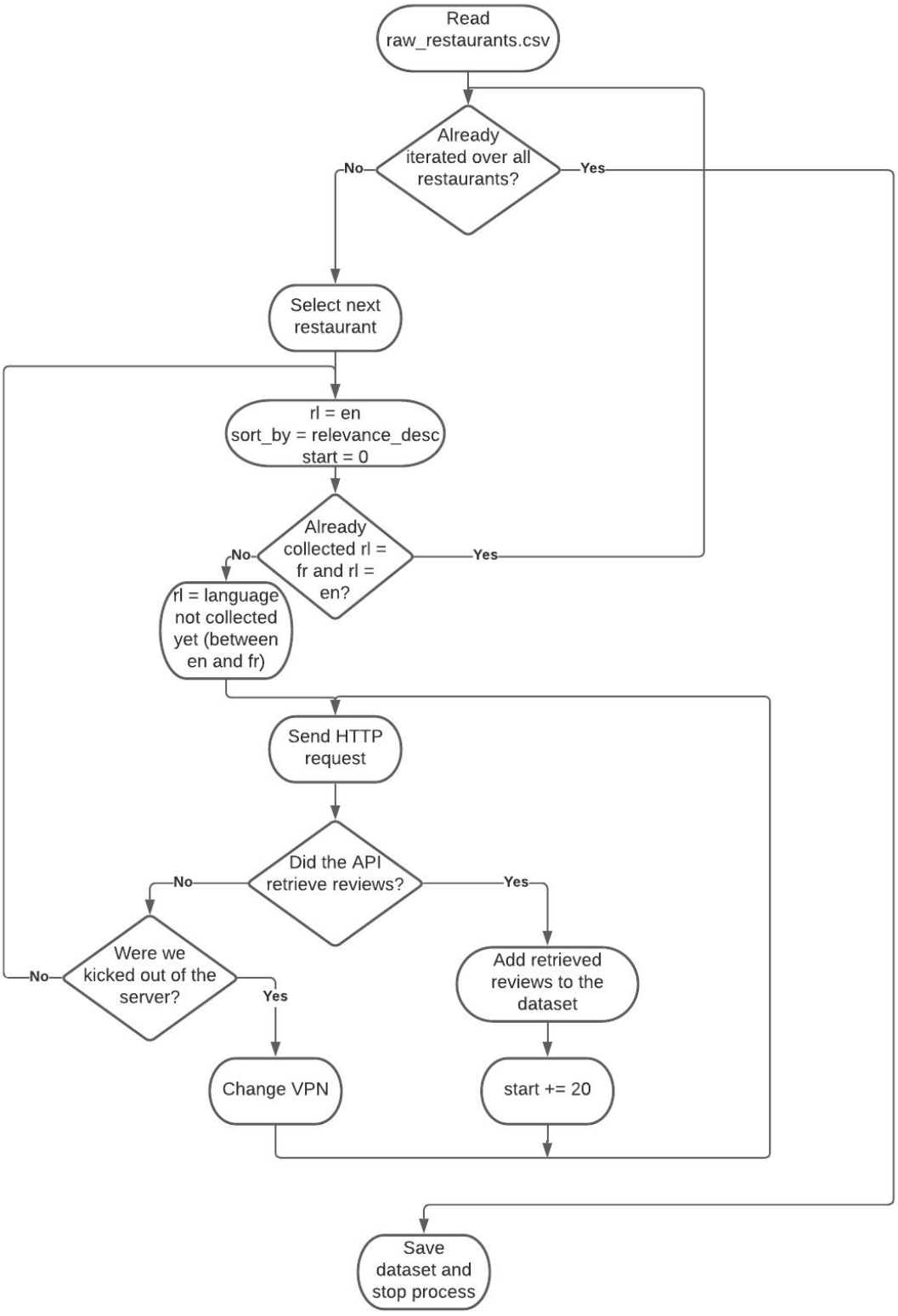

Figure 5. Logic behind the ingestion of reviews

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:

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'oublier l'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-WusZrBYZexKqhqzCkg', '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

Table 2. Characteristics of the review returned from Yelp's hidden API

2.3 Fake Reviews

To collect the not 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:

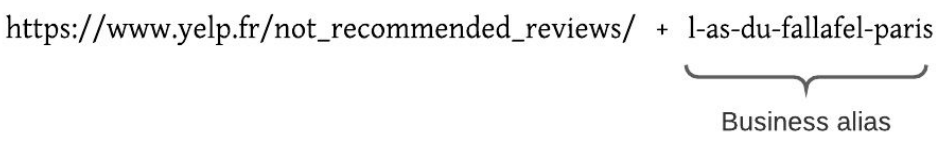

Figure 6. Pattern of the URL of the fake reviews

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.

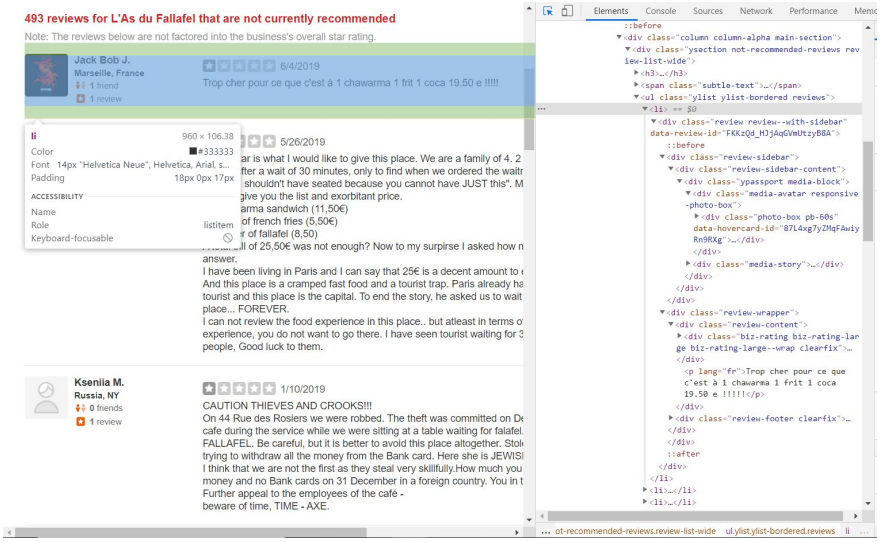

Figure 7. Inspection of the HTML of the not recommended reviews page

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. Luckly, since we had already solved those problems before, it was easy to adapt the logic to this case, which resulted in the following:

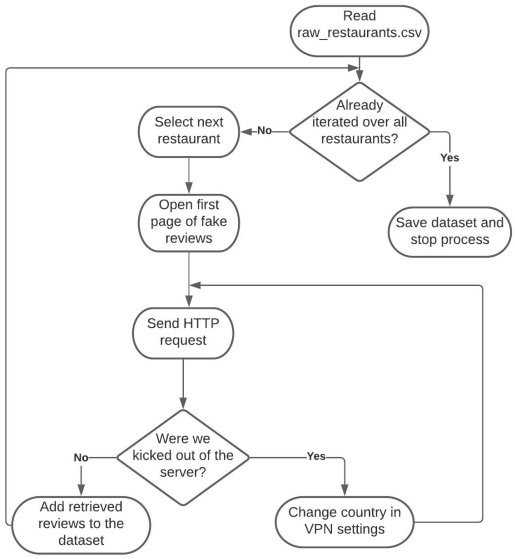

Figure 8. Logic behind ingestion of fake reviews

As one can notice from the diagram above, 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. Since the majority of the restaurants didn't have a lot of fake reviews, our results wouldn't get affected significantly. 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:

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	l-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

Table 3. Features contained on extracted fake reviews

3. Data Processing

In this stage the goal is to transform the amount of data collected from the internet in a structured database, threatening all the data collected with a lot of structures and errors, and processing them to achieve a useful database.

At this point we extract information that was in list format, get the real locations and drop unnecessary columns.

3.1 Restaurants

We had some columns that weren't normalized and were difficult to work with, so we threw away the useless information and got the other useful with 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.

- Extract the coordinates from a restaurant from a list
- Numeralizing the price scale from {$, $$, $$$, $$$$} to {1, 2, 3, 4}
- Drop unnecessary columns like title and phone number
- Getting the arrondissement from the address

Feature	Description	Sample
id	Unique Yelp ID of this business	FFz-WusZrBYZexKqhqzCkg
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	https://s3-media3.fl.yelpcdn.com/bphoto/QMNELSZ6-LzA9kLP3zQPgw/o.jpg
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

Table 4. Characteristics of the restaurant after processing

3.2 Real Reviews

- Extract business information from a list
- Verify if the user has an image
- Drop unnecessary columns like photos, user link, etc

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

Table 5. Characteristics of the review after processing

3.3 Fake Reviews

Since fake reviews were obtained via scraping, we were able to structure our dataset the way we wanted. Because of that there was no need to make any sort of processing with this data.

4. Data Enhancement

The goal in this section was to generate features from the already processed data. This includes, but is not restricted to, obtaining extra information on the elements collected and calculating metrics that might be insightful for the next sections. As this can be time consuming, there is a relevant gain in performance if we, at the end, create a new dataset, to be used by the analysis part.

4.1 Restaurants

- Count the number of fake reviews for each restaurant (from YELP's website)
- Create useful columns like the total number of reviews and the percentage of fake reviews

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

Table 6. Characteristics of the restaurant after enhancement

4.2 Real Reviews

- Create the country column (containing a country's Alpha 3 code)
- Remove HTML content from the commentaries

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

Table 7. Characteristics of the review after enhancement

4.3 Fake Reviews

- Delete non french or english reviews
- Count the number of photos of the review

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

Table 8. Characteristics of the fake reviews after enhancement

5. Data Visualization

Paris is one of the most touristic cities in the world, so we want to see where the reviews came from.

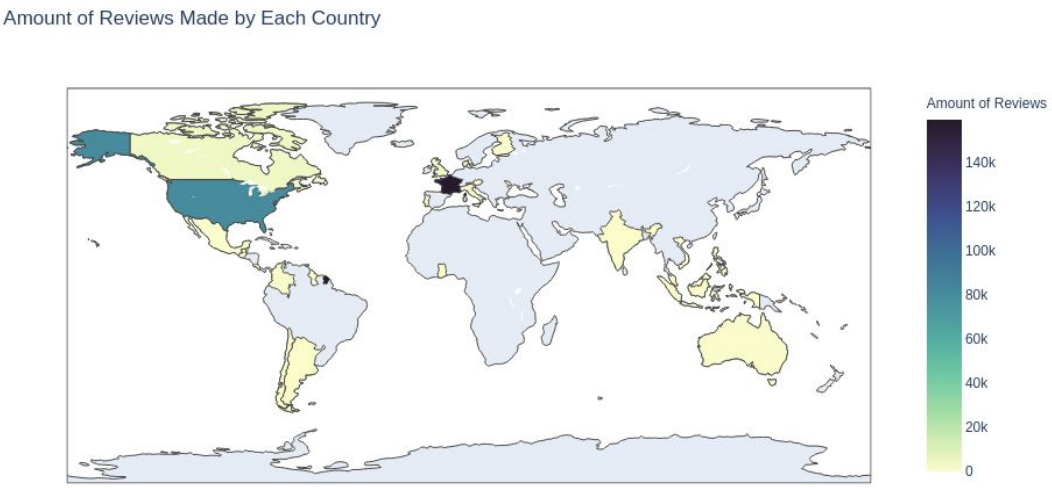

Figure 9. Location of user's reviews

We can see the number of reviews by year and see how the platform was growing during the times, had their maximum in 2015, and maybe lose a little bit of his importance with the entrance in the market of another's players like TripAdivisor or Google. We can also see the drowning of reviews in 2020 motivated by the pandemic of Covid-19 affecting a lot the market of restaurants.

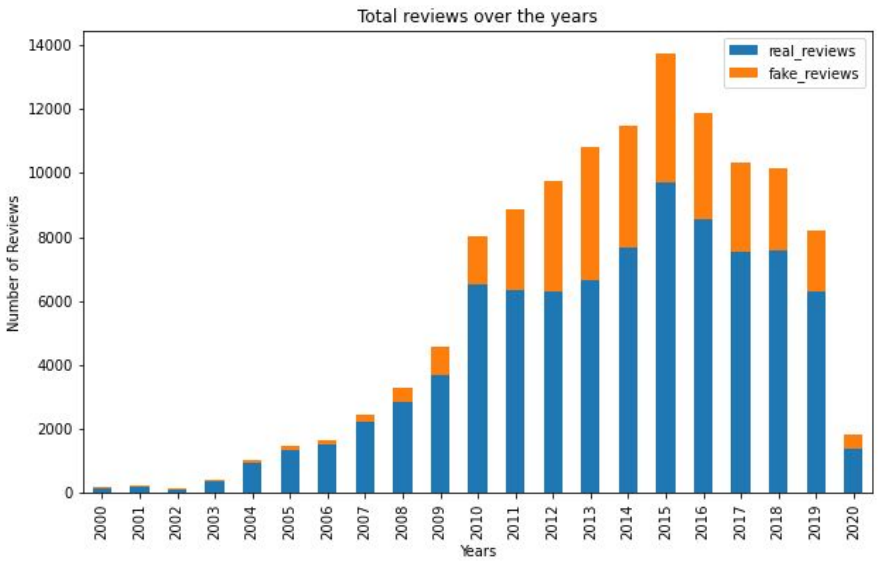

Figure 10. Total reviews in Paris per year.

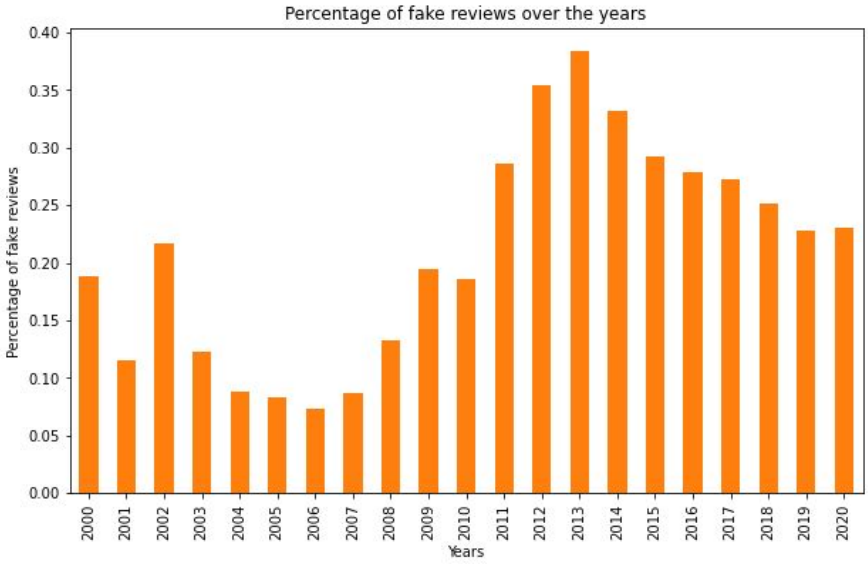

Figure 11. Percentage of fake reviews over the years in Paris.

Another important information is where these reviews are in Paris, so we plotted the number of reviews and fake reviews per arrondissement of Paris.

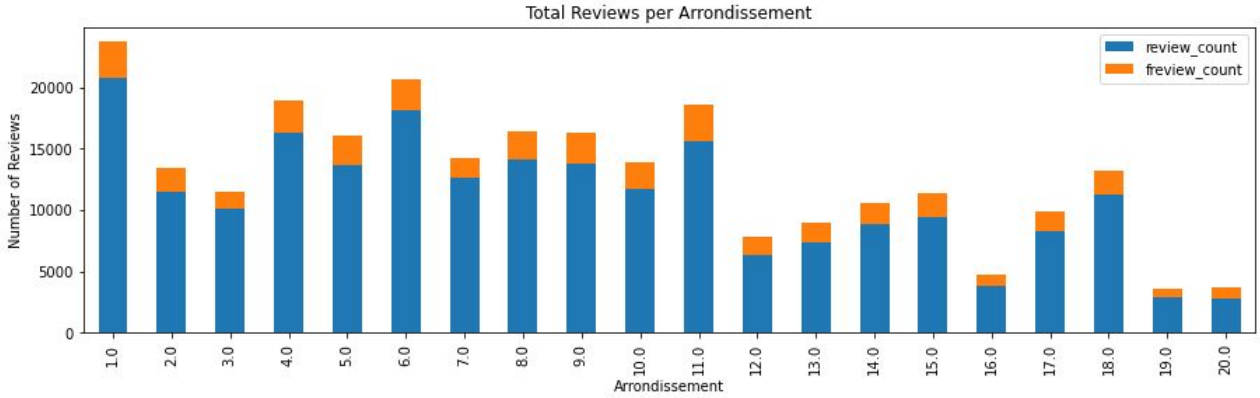

Figure 12. Number of real and fake reviews per arrondissement.

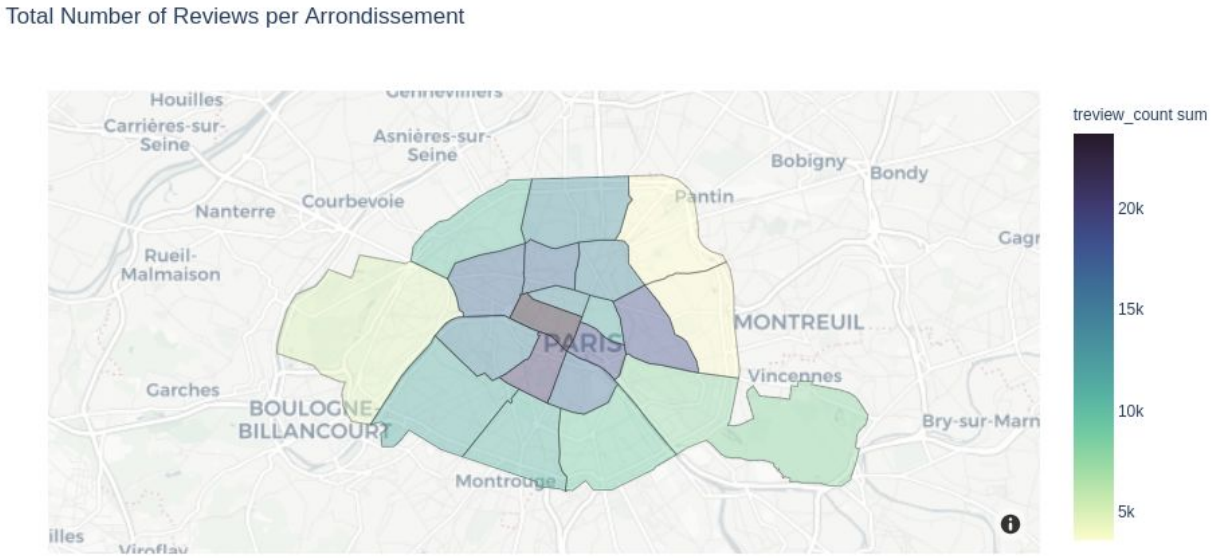

Figure 13. Local of real and fake reviews per arrondissement.

And these graphs can tell us a little bit about where the base of users of Yelp usually go to eat, showing the central part of Paris with the place where they go most, but necessarily the best restaurants of the city. The best restaurants are not in the center as in the last figure but more spaced over Paris following a bit the most expensives arrondissements as shown in the figures below.

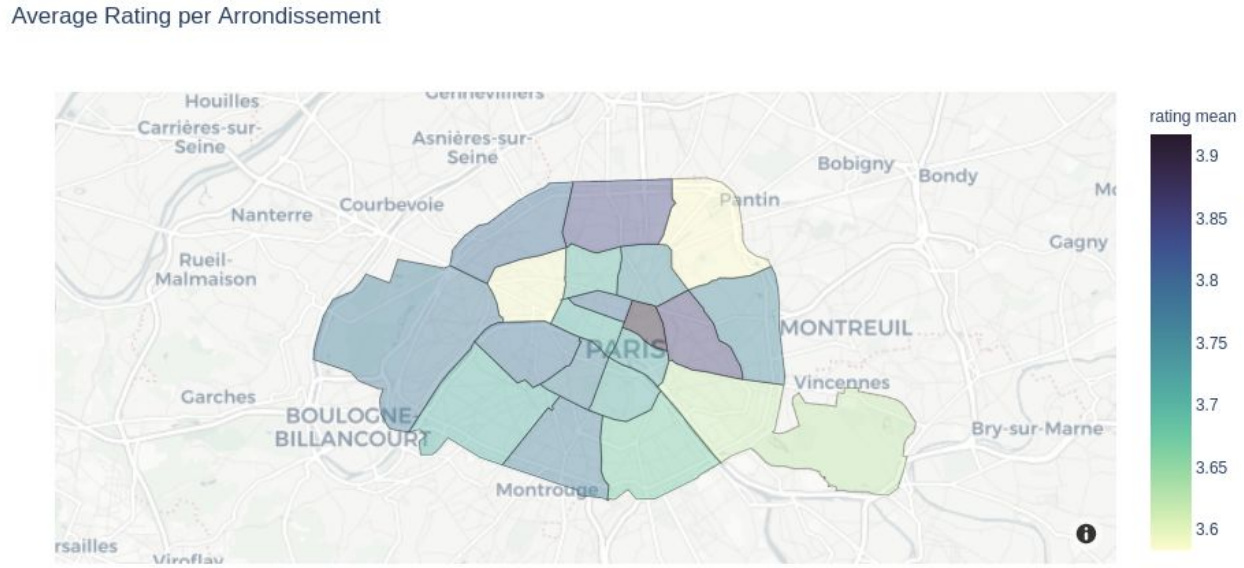

Figure 14. Restaurant average rating per arrondissement

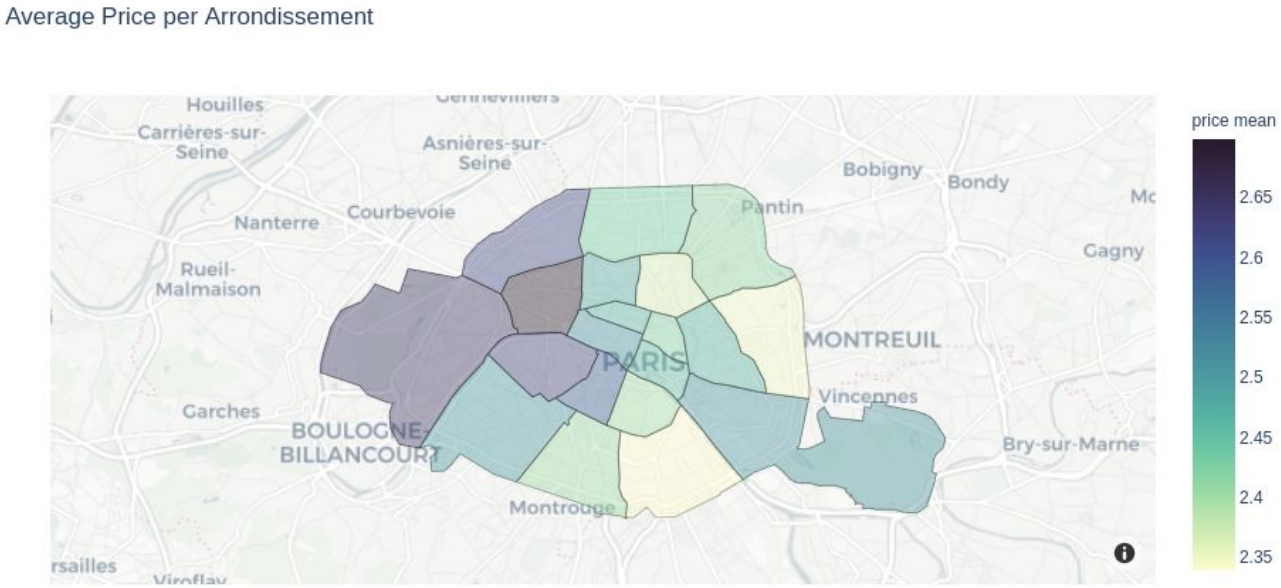

Figure 15. Restaurant average price per arrondissement

We observe that the most expensives restaurants are in west of Paris, but that users of Yelp commonly go more in restaurants with medium price shown in the image below.

Figure 16. Number of reviews per price

6. Data Analysis

6.1 A Note on the Analysis

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 will be to build a common ground on the incentives which play a relevant role on review fraud and how those can impact businesses outside the digital world.

6.2 Review Fraud and its Economic Incentives

The goal of this section is to better understand the economic motivations behind review frauds. 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 conduct the following analyses considering that all 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.

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 its own reputation. Therefore, we considered two main metrics that are related to the number of positive or negative fake reviews that a restaurant has engaged in or received, respectively: the restaurant's own reputation and the competition that it faces.

The first metric, restaurant reputation, can be associated with many factors, such as its average rating, its number of received reviews and even its price. All these informations are already present in the dataset, and are easily obtained through simple filtering and computations.

The competition metric, on the other hand, is a little harder to compute, and to obtain it, we followed the same approach as *Fake It Till You Make It: Reputation, Competition, and Yelp Review Fraud* [1], with only slight changes. As described in the article, the competition between two restaurants can be described as inversely correlated with the distance between them. We also decided to split competition between restaurants that share the same categories and restaurants that don't share categories. This way, when we perform the regressions we can see which type of competition has a larger influence on review fraud. After calculating the competition metrics, we can visualize the average competition experienced by restaurants by Paris' arrondissement.

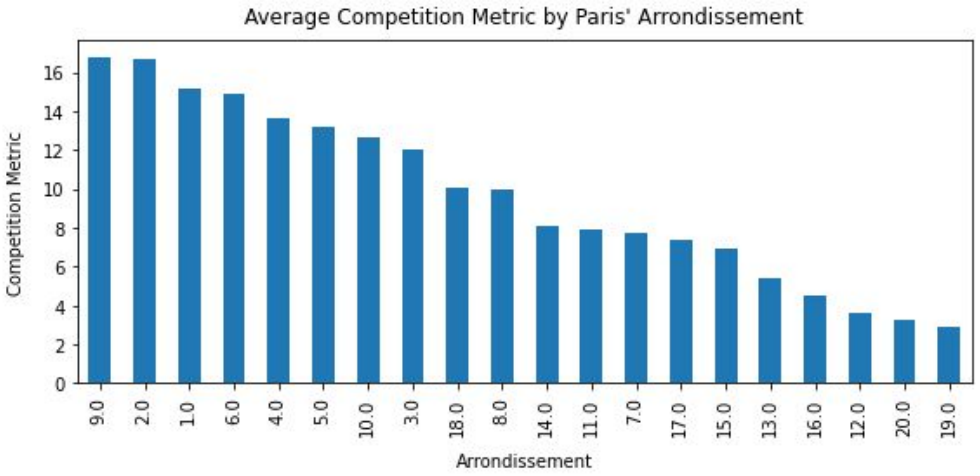

Figure 17. Average competition faced by restaurants by arrondissement.

We also faced a problem before performing the regressions: even though we know the number of fake reviews for each restaurant, when web-scraping Yelp for the content of fake reviews we only managed to obtain a maximum of ten fake reviews per restaurant. We have verified that some large restaurants have more than 10 fake reviews. Since we want to separate the fake reviews based on their content (positive or negative), we are obliged to remove the restaurants where we didn't collect all the fake reviews. This removed most of the large restaurants, and thus, these analyses will be mostly directed to mid to small restaurants.

6.3 Regression Analysis on Review Fraud

After computing the metrics and cleaning the data, we performed an OLS Regression 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 restaurant's total review count.

Indeed, the restaurant's total review count is only 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 small restaurants.

6.3.1 Regression Analysis on Negative Fake Reviews

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:	Sun, 17 Jan 2021	Prob (F-statistic):	0.00
Time:	22:07:32	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	1.0230	0.057	18.095	0.000	0.912	1.134
x1	5.791e-06	1.43e-05	0.406	0.684	-2.21e-05	3.37e-05
x2	-7.915e-06	6.09e-06	-1.300	0.194	-1.98e-05	4.02e-06
x3	-0.1654	0.009	-18.210	0.000	-0.183	-0.148
x4	-0.0865	0.012	-6.985	0.000	-0.111	-0.062
x5	0.0296	0.001	47.581	0.000	0.028	0.031

Omnibus:	5852.224	Durbin-Watson:	1.991
Prob(Omnibus):	0.000	Jarque-Bera (JB):	65680.651
Skew:	2.511	Prob(JB):	0.00
Kurtosis:	14.261	Cond. No.	3.82e+04

Figure 18. 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 scale make for significant changes in the p-value for the null hypothesis. This indicates that the metric may give us some significant insights. We can see that an increase in competition of the same category generates more negative review fraud (restaurants may be willing to commit review fraud to lessen competitors reputation). On the other hand, an increase in competition of different categories actually decreases the number of fake reviews. This may happen for numerous reasons, since it is hard to predict restaurants' reactions given an increase in competition.

As for the effect of own reputation, we can see that higher average rating and price range restaurants 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 analysing these coefficients.

6.3.2 Regression Analysis on Positive Fake Reviews

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:	0.910
Model:	OLS	Adj. R-squared:	0.910
Method:	Least Squares	F-statistic:	2.107e+04
Date:	Sun, 17 Jan 2021	Prob (F-statistic):	0.00
Time:	22:07:32	Log-Likelihood:	-24741.
No. Observations:	10368	AIC:	4.949e+04
Df Residuals:	10362	BIC:	4.954e+04
Df Model:	5		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
const	-6.5608	0.177	-36.997	0.000	-6.908	-6.213
x1	-9.541e-07	4.47e-05	-0.021	0.983	-8.86e-05	8.67e-05
x2	4.671e-05	1.91e-05	2.446	0.014	9.28e-06	8.41e-05
x3	1.3628	0.028	47.831	0.000	1.307	1.419
x4	0.2804	0.039	7.216	0.000	0.204	0.357
x5	0.6112	0.002	312.820	0.000	0.607	0.615

Omnibus:	2750.403	Durbin-Watson:	1.972
Prob(Omnibus):	0.000	Jarque-Bera (JB):	153970.051
Skew:	-0.418	Prob(JB):	0.00
Kurtosis:	21.860	Cond. No.	3.82e+04

Figure 19. Effect of Own Reputation and Competition on Positive Review Fraud

This model performed better in terms of generalization, achieving a 0.910 R^2 rating, which is largely due to having more behaved data.

Here, we can see that although the effect of same category competition has a statistically insignificant value, the coefficient $\mathbf{X_2}$ indicates that an increase in different category competition generates a statistically significant increase in positive review ratings. This makes sense, since it may indicate that restaurants are more likely to engage in positive review fraud when competition increases.

The effects of own 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 point us in the direction that higher priced and higher rated small restaurants are more likely to engage in positive review fraud.

6.4 Review and Reviewer Content Analysis

This part of data analysis consists in studying the content of reviews. One wants to find, through that analysis, the possible key factors that differentiate fake reviews from real reviews.

We start by checking differences in shape, using as parameter the size of the review, in characters.

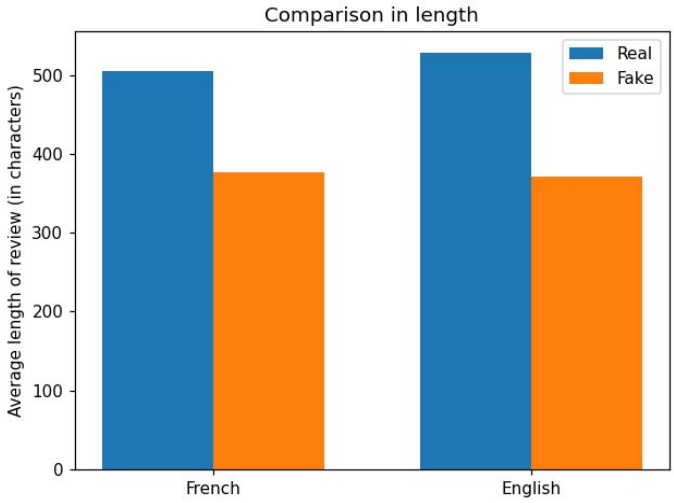

Figure 20. Comparing the length of real and fake reviews

From the data analysed, it is clear that fake reviews have, on average, a smaller length when compared to real reviews. Reviews classified as real have around 500 characters, while fake ones are, on average, 120 characters shorter.

Next natural idea for us was that there should exist a difference in the average number of friends from people that post fake reviews and from people who post real reviews. We didn't know, a priori, for what side it would be unbalanced, as although people from real reviews may tend to interact naturally and

build a network, fake reviewers may artificially grow their network with suspicious accounts, all in order to build some credibility.

The difference between these two types of user can be clearly spotted on the plot below.

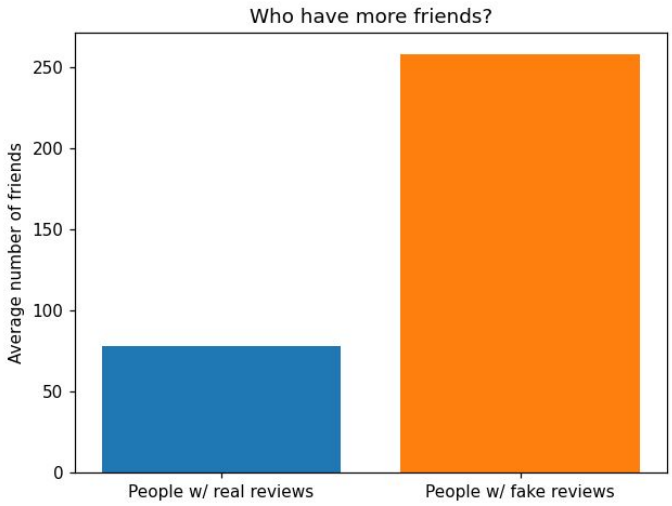

Figure 21. Comparing the number of friends from reviewers

From the data, we see that people that post real reviews usually have a smaller network of friends than fake review posters. Although real reviewers have, on average, 77 friends, fake review posters have a much larger network of friends, with 257 as its average friends counting value. This might be due to an artificial growing method used by fake posters. Although it is not in the scope of this project, one great analysis would be to study the network of both groups, which could lead to better insights.

Intuitive analysis would indicate that real reviewers are more active in Yelp. This could be related to the fact that they have a feeling of community, and see the act of sharing their opinions as a help for others. On the other hand, fake reviewers use yelp to obtain one immediate benefit (which would be to increase a restaurant rating or decrease its competitor rating). This can be confirmed, to the dataset analysed, and seen on the plot below, that shows that people who post fake reviews tend to post less than the ones who post real reviews.

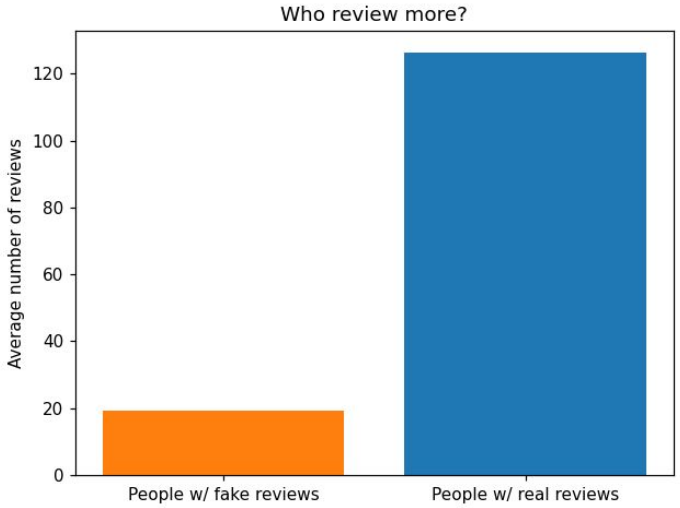

Figure 22. Comparing the review posting frequency

People within the group of fake posters, have, on average, 19 reviews posted, while members of the second group tend to post more often, averaging 126 past reviews per user.

If we turn our attention to reviews content, we would expect to note that fake reviews are more extreme than real reviews[1]. At first, this can be noticed looking at the grades assigned to each review. We should be able to see a peak at lower and higher grades for the fake reviews, when compared to real ones. If we plot the histogram for both datasets (real and fake) with respect to the ratings, we obtain the following plot.

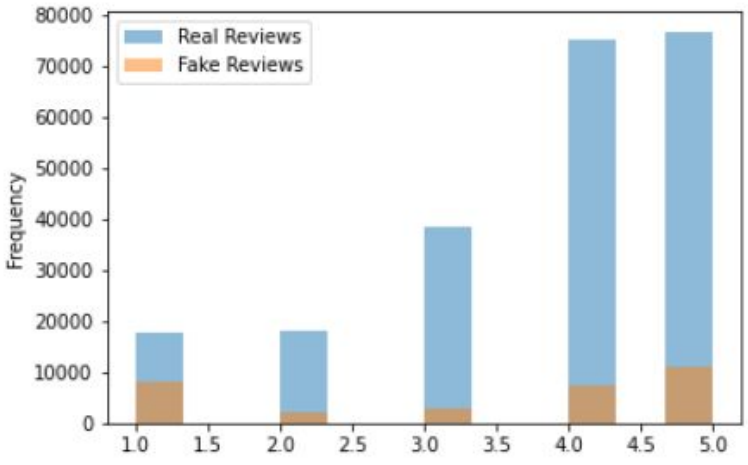

Figure 23. Distribution of reviews throughout the ratings

This indicates that real reviews have an increasing concentration towards higher grades, but fake reviews show clearly two peaks, one at rating 1 and other at rating 5 (with also a good amount at rating 4). At least on the rating, fake reviews are more extreme (favorable or unfavorable) than real reviews.

Can we observe the same pattern when looking to the content of the reviews? Specially considering a generalized model to differentiate fake from real reviews, it would be great to use its content to prevent fake reviews on a real-time basis.

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.

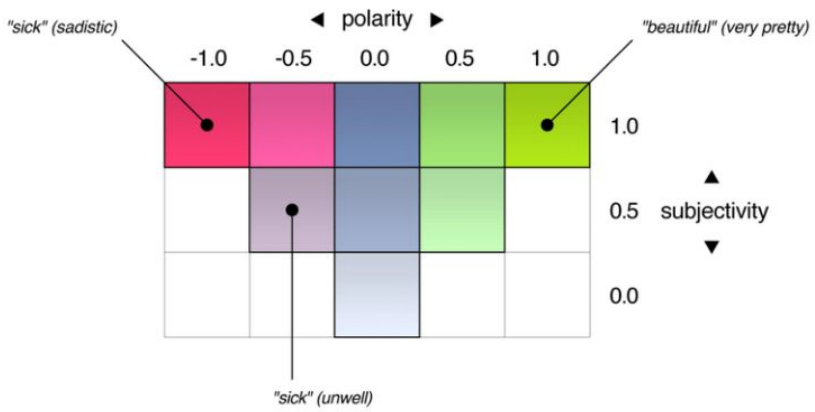

Figure 24. What is polarity and subjectivity - Image from [3]

It would be expected to get texts with higher polarity when dealing with fake reviews. As english language corpus is more developed for natural language processing, we choose to analyse real and fake reviews that were written in english. We can further extend this analysis if we translate french reviews, although some elements of text syntax might be lost during the translation process.

The process of sentiment analysis is based on a trained model, with a corpus from social media and review websites. After training the model with labelled data, it can be used to predict sentiment from other text excerpts.

We analysed each one of the english reviews and ran a sentiment analysis. For each one, we got the polarity constant and then computed the absolute values of it (because it can be either extremely negative or positive, going from [-1,1]). It can be observed in the following image.

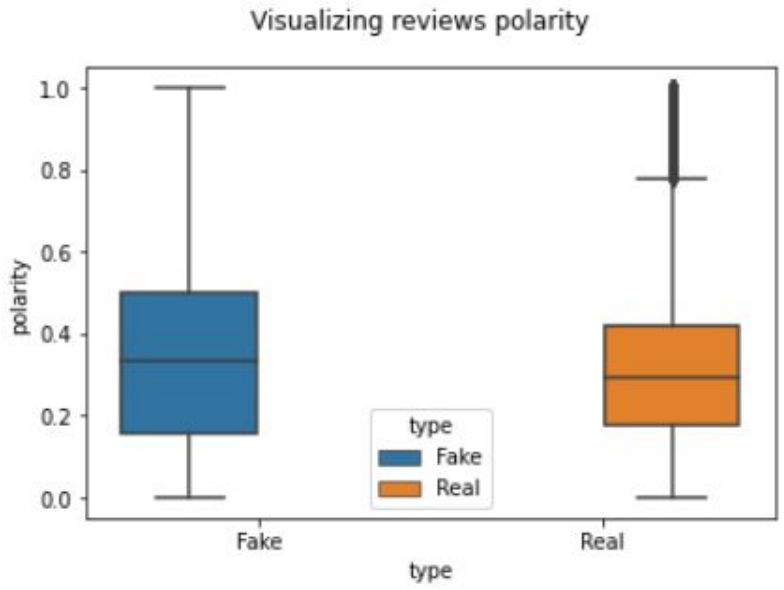

Figure 25. Polarity analysis

Looking at the produced result from the analysed dataset, we observe that 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 polarity. Also, the most conclusive point would be 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.

This indicates a relation 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, fake reviews have higher polarity when compared to real reviews.

7. Conclusion

As a group, we genuinely felt that this project completed our learning process. We were able to develop all the theory 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.

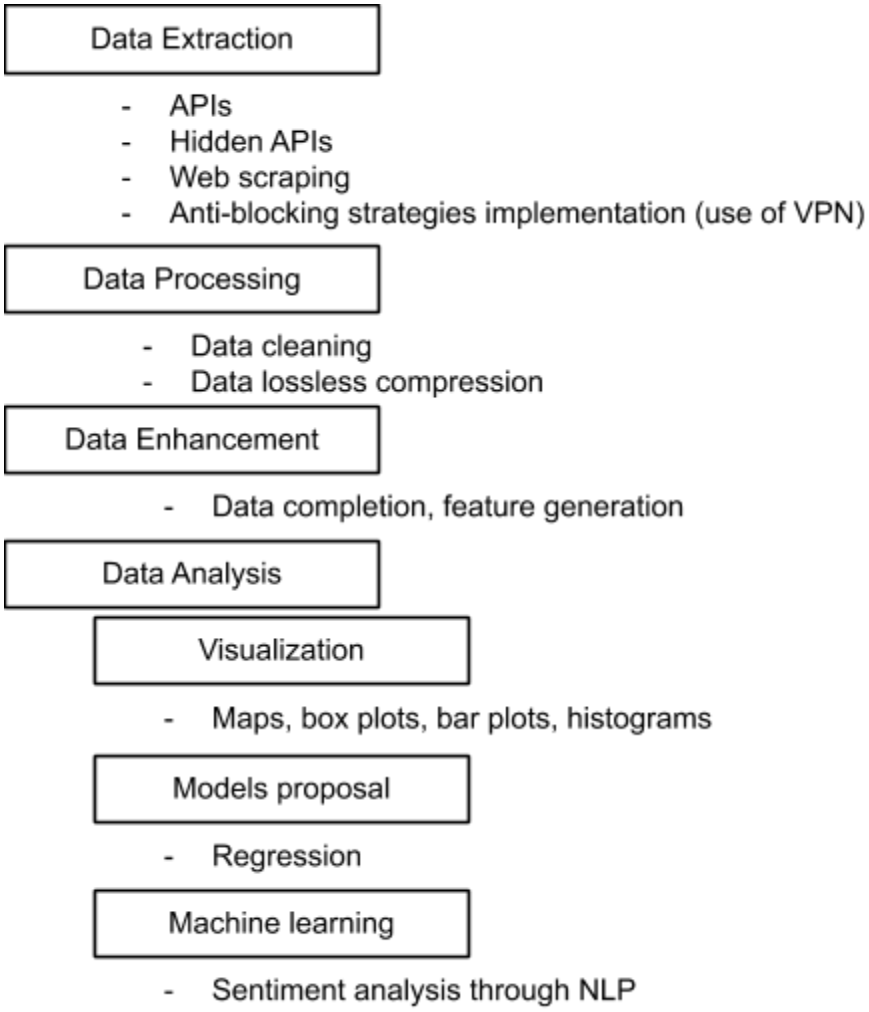

Figure 26. Project path

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 all possible.

8. 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.lng