Word count: 1971 words

MSIN0041

Marketing Science Group Project

Group members:
Dina Rosenfield, Sn: 18030759
Justine Malapert, Sn: 18091974
Inès Cauvin, Sn: 18093287
Gary Rouch, Sn: 18097365
Anatole de Rauglaudre, Sn: 18095078
Melissa Saadi, Sn: 18096314

Table of Contents

I. Company Background and Research Question

Founded in 2008 by Brian Chesky, Joe Gebbia, and Nathan Blecharczyk, Airbnb.Inc is an American home rental platform based in San Francisco that allows people to list, find, and rent short-term lodging in 65,000 cities across 191 countries (Fast Company, 2020). Airbnb went public on December 10th, propelling the home-rental company to a $100 billion valuation (Carville, et al., 2020).

At the end of march 2020, Airbnb interrupted all its marketing activities to save an estimated $800 million in 2020 (Andjelic & Davidoff, 2020). Because of the COVID-19 crisis, the company decided to cut the budget allocated to marketing adverts and thus harmed the relationship with its customers (Glenday, 2020). Airbnb's hosts must focus on different ways to attract customers to rent from them and maintain good customer and user relations without relying on AirbnB's commercial adverts. This paper aims to solve the following research questions:

> **1.** How can AirBnB help the hosts to improve the customer experience, thus increase revenue by analyzing the most negative/positive reviews and identifying the most crucial criteria for a successful experience?
>
> **2.** How can we help Airbnb recommend efficient prices to its hosts by identifying the most significant variables affecting price.

Our analysis will help us find the specific aspects customers want to have in an accommodation to enjoy their stay. AirBnB can then suggest to their hosts what elements they should include in their accommodation/what to mention in their product description to attract customers. Our analysis will also help us to find what variables have the greatest impact on price. Knowing this, AirBnB can suggest to hosts what factors to focus on when determining a price to ensure they use efficient prices, thus increasing profits.

Our research questions address the 'pricing' and 'product' component of the 4Ps. We also address user-generated content (UCG) through analysing a reviews dataset.

II. Data and Methods

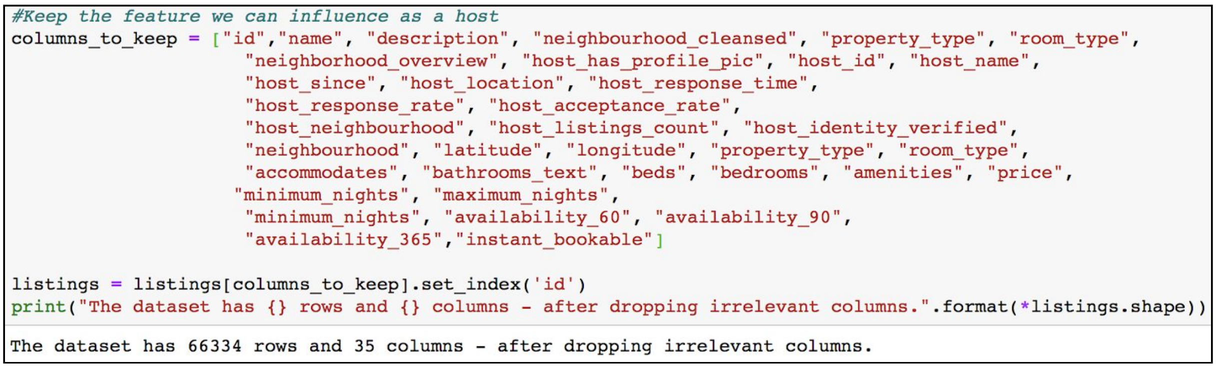
Throughout this project we used two datasets sourced from Inside Airbnb website (Inside Airbnb, 2020):

- Listings.csv: detailed listings data for Paris.
- Reviews.csv: summary review data and listing ID.

1.Preparing the dataset for Xgboost modelling

For the Xgboost model we used the listings dataset, which contains 66334 observations and 74 variables. Each observation is a different AirBnB accommodation in Paris, and each variable contains different information about the accommodation.

First, we reviewed all the variables and filtered out the variables that a host cannot influence.



```
#Keep the feature we can influence as a host
columns_to_keep = ["id","name", "description", "neighbourhood_cleansed", "property_type", "room_type",
                     "neighborhood_overview", "host_has_profile_pic", "host_id", "host_name",
                     "host_since", "host_location", "host_response_time",
                     "host_response_rate", "host_acceptance_rate",
                     "host_neighbourhood", "host_listings_count", "host_identity_verified",
                     "neighbourhood", "latitude", "longitude", "property_type", "room_type",
                     "accommodates", "bathrooms_text", "beds", "bedrooms", "amenities", "price",
                    "minimum_nights", "maximum_nights",
                     "minimum_nights", "availability_60", "availability_90",
                     "availability_365","instant_bookable"]

listings = listings[columns_to_keep].set_index('id')
print("The dataset has {} rows and {} columns - after dropping irrelevant columns.".format(*listings.shape))
```

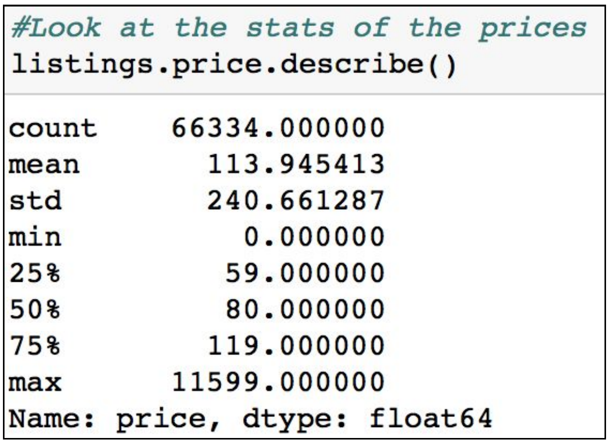
```
The dataset has 66334 rows and 35 columns - after dropping irrelevant columns.
```

Figure 1.

We treated the following variables:

- **host_has_profile_pic:** replaced missing values to false ;
- **bedrooms:** replaced the missing bedroom values by the number of beds;
- **bathroom_text:** converted the bathroom numbers to integers and deleted the remaining string values;
- **prices:** changed the prices to integers.

To investigate our target variable - prices - we generated the following statistics:



```
#Look at the stats of the prices
listings.price.describe()
```

```
count     66334.000000
mean        113.945413
std         240.661287
min           0.000000
25%          59.000000
50%          80.000000
75%         119.000000
max       11599.000000
Name: price, dtype: float64
```

Figure 2.

and we as well generated a visual to investigate the range of prices:

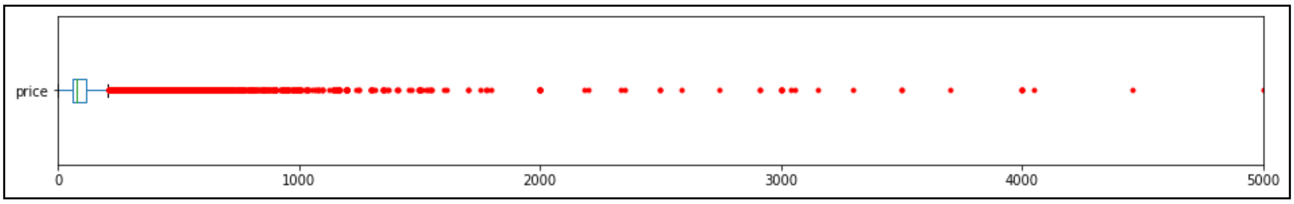
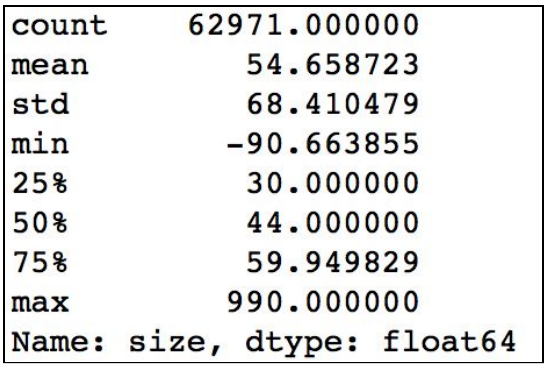


Figure 3.

We can see from this graph that there are many extreme values (outliers). Therefore, we decided to set an upper limit so that our data only contains apartments up to €400. There were also apartments priced at €0. This seemed odd so we dropped those observations as well.

Some hosts included the size of their property in their description. This enabled us to create a new variable called ‘size’ and use linear regression to predict the size of accommodations that were missing.



```
count     62971.000000
mean         54.658723
std          68.410479
min         -90.663855
25%          30.000000
50%          44.000000
75%          59.949829
max         990.000000
Name: size, dtype: float64
```

Figure 4.

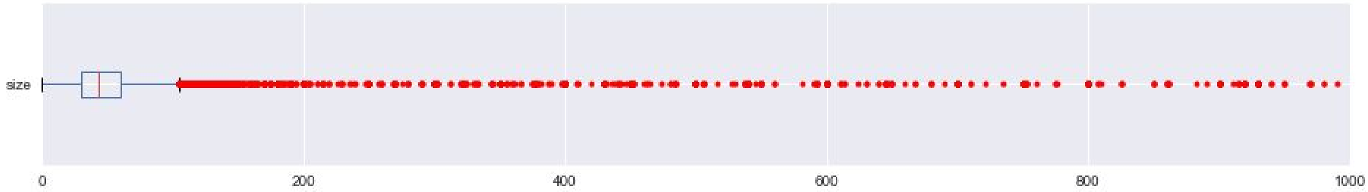


Figure 5.

Looking at the box plot according to size, we decided to remove all outliers above 300 square meters, and all those with 0. We then created a joint plot showing the relationship between size and price.

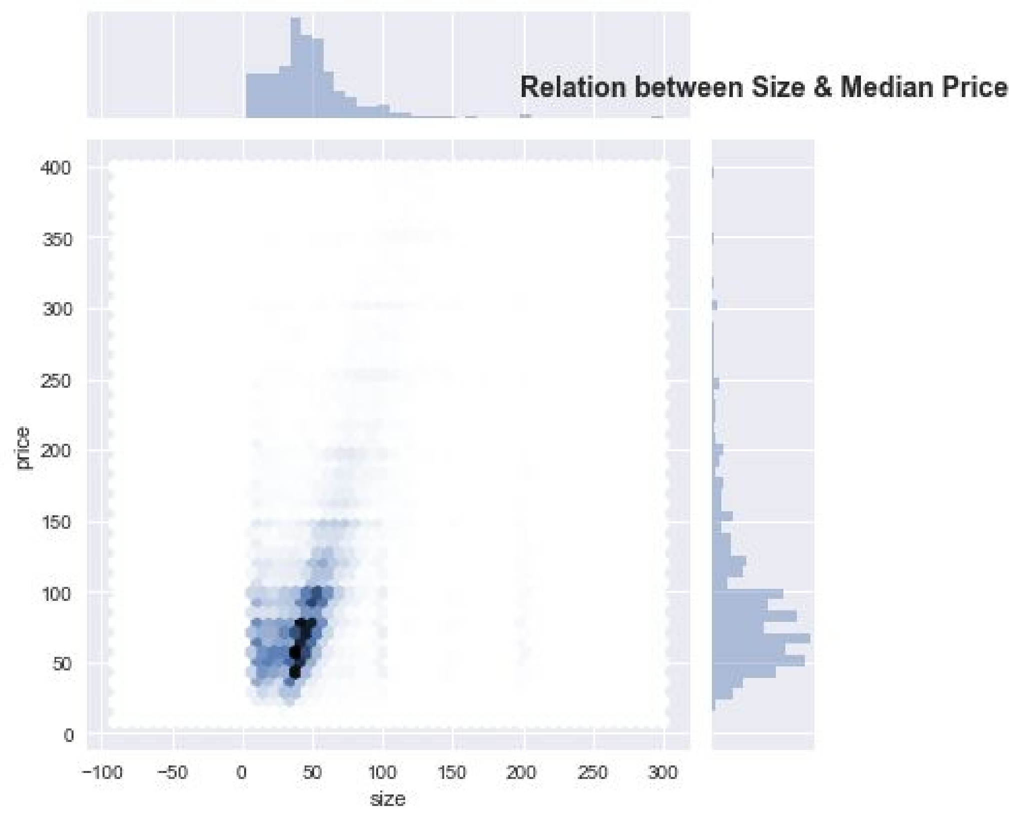


Figure 6.

We also created a distance variable using a function that calculates the distance from each accommodation to the city center.

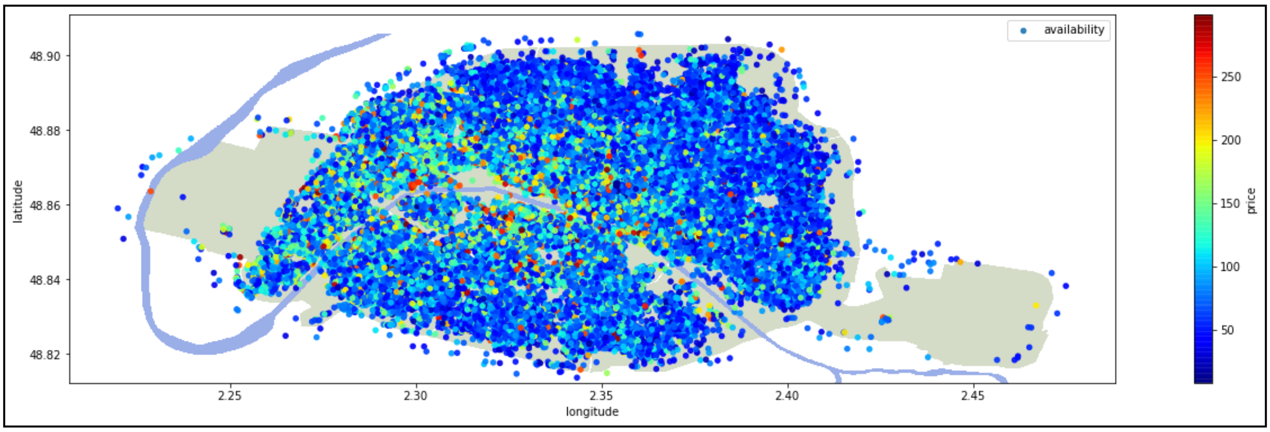
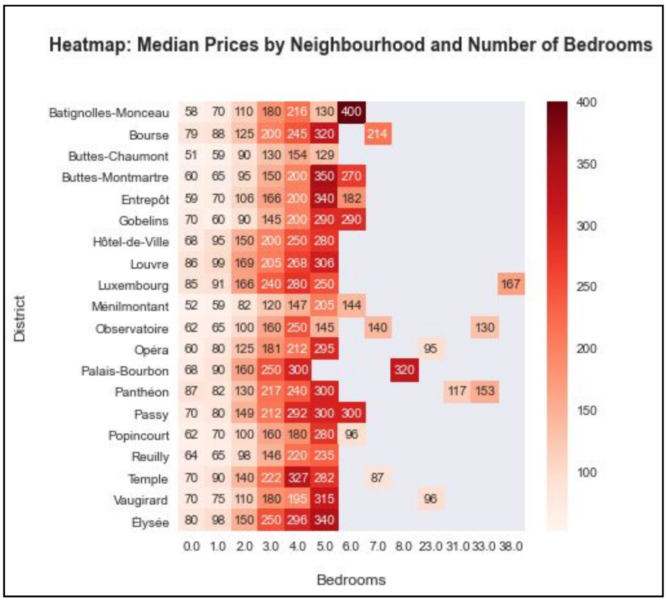


Figure 7. Price differences based on distance.

From this map we can see that towards inner Paris, apartments go for a much higher price than in surrounding areas.

We then created a heatmap and boxplot to show the difference in prices between neighbourhoods.



Heatmap: Median Prices by Neighbourhood and Number of Bedrooms

District	0.0	1.0	2.0	3.0	4.0	5.0	6.0	7.0	8.0	23.0	31.0	33.0	38.0
Batignolles-Monceau	58	70	110	180	216	130	400						
Bourse	79	88	125	200	245	320		214					
Buttes-Chaumont	51	59	90	130	154	129							
Buttes-Montmartre	60	65	95	150	200	350	270						
Entrepôt	59	70	106	166	200	340	182						
Gobelins	70	60	90	145	200	290	290						
Hôtel-de-Ville	68	95	150	200	250	280							
Louvre	86	99	169	205	268	306							
Luxembourg	85	91	166	240	280	250							167
Ménilmontant	52	59	82	120	147	205	144						
Observatoire	62	65	100	160	250	145		140				130	
Opéra	60	80	125	181	212	295				95			
Palais-Bourbon	68	90	160	250	300				320				
Panthéon	87	82	130	217	240	300					117	153	
Passy	70	80	149	212	292	300	300						
Popincourt	62	70	100	160	180	280	96						
Reuilly	64	65	98	146	220	235							
Temple	70	90	140	222	327	282		87					
Vaugirard	70	75	110	180	195	315				96			
Elysée	80	98	150	250	296	340							

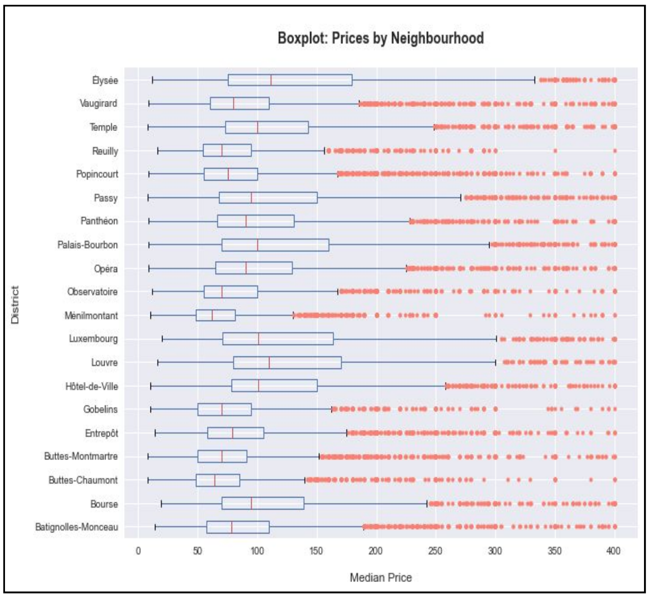
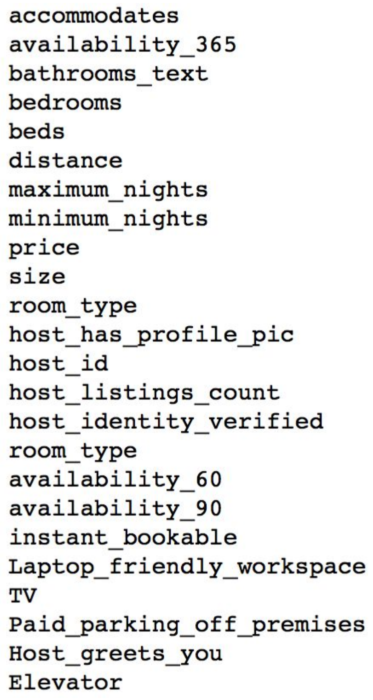


Figure 9.

Figure 10.

Our final step to prepare the data, was to review how many missing values were still in the dataset. We dropped the columns and rows that had many missing values. Thus, we obtained our final listings dataset that contains 62208 observations and 24 variables.



```
accommodates
availability_365
bathrooms_text
bedrooms
beds
distance
maximum_nights
minimum_nights
price
size
room_type
host_has_profile_pic
host_id
host_listings_count
host_identity_verified
room_type
availability_60
availability_90
instant_bookable
Laptop_friendly_workspace
TV
Paid_parking_off_premises
Host_greets_you
Elevator
```

Figure 8.

2. Preparing the data for NLP modelling

We used both the listings and reviews datasets to prepare the data for NLP modelling. The reviews dataset contains 1269063 observations and 6 variables.

We first dropped irrelevant variables -such as listing_url or scrape_id- reducing the dataset to 66334 rows and 44 columns.

To start our text analysis we decided to plot the most common words used in the names column of the listings dataset for each accommodation. To generate the plot we created an empty list where we appended strings created from the name column and set a function that split those name strings into separate words. Then, we initialized another empty list where we have words counted and using the counter dictionary we visualised the top 25 words used by the hosts to name their listing (Figure 8).

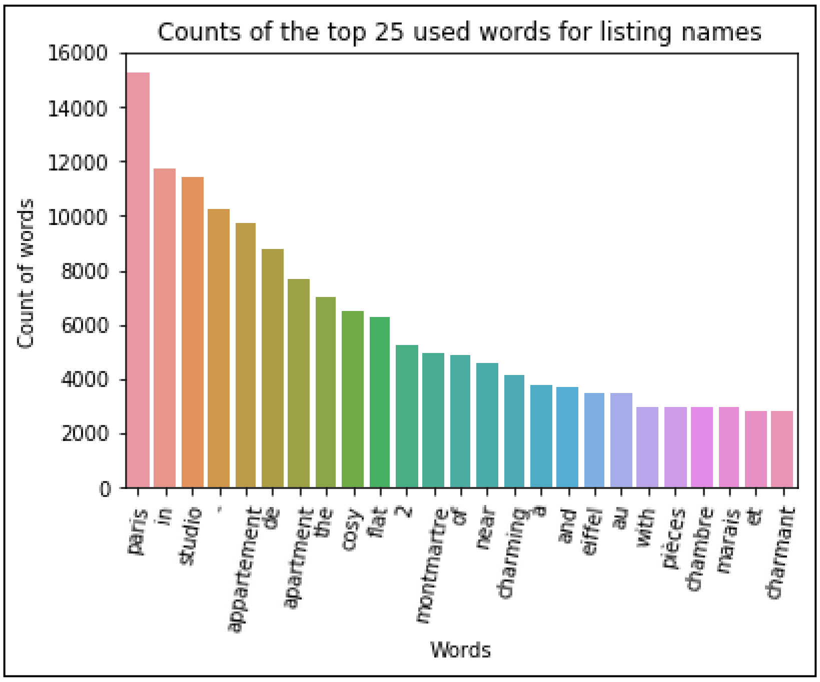


Figure 11.

We noticed that Paris, studio, apartement(/apartment) and cosy were the most significant words used in an accommodation's name.

We also used the language detect library to identify the different languages used in this dataset.

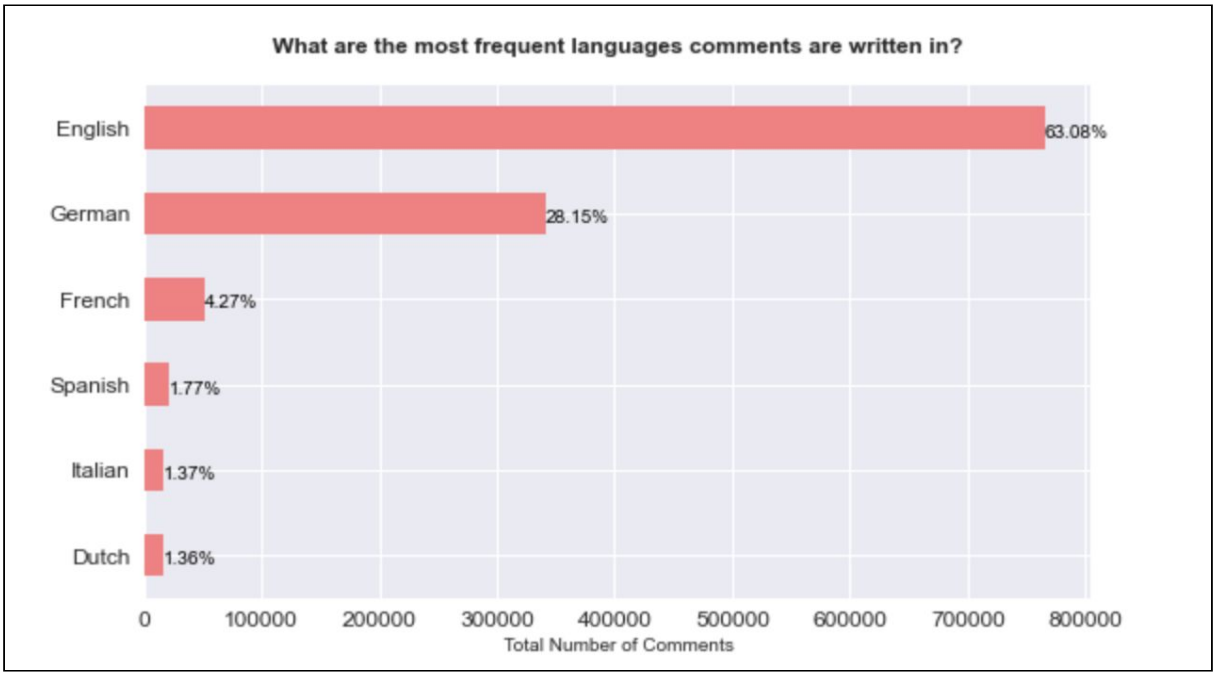
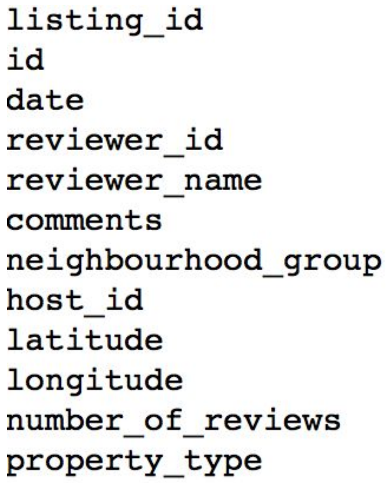


Figure 13.

To implement our NLP model we merged the reviews dataset with the 7 following variables taken from the listings dataset: neighbourhood_cleansed, host_id, latitude, longitude, number_of_reviews, id, property_type. We then replaced missing values in the comments column with ‘True’ and obtained a final dataset which contains 1268544 rows and 12 columns (Figure 7).



```
listing_id
id
date
reviewer_id
reviewer_name
comments
neighbourhood_group
host_id
latitude
longitude
number_of_reviews
property_type
```

Figure 12.

III. Findings and Implications

1. Natural Language Processing models

1.1. Word Clouds

Using NLP analysis and detection language functions, we managed to plot word clouds showing the most frequently used words in general comments and in comments identified as positive.

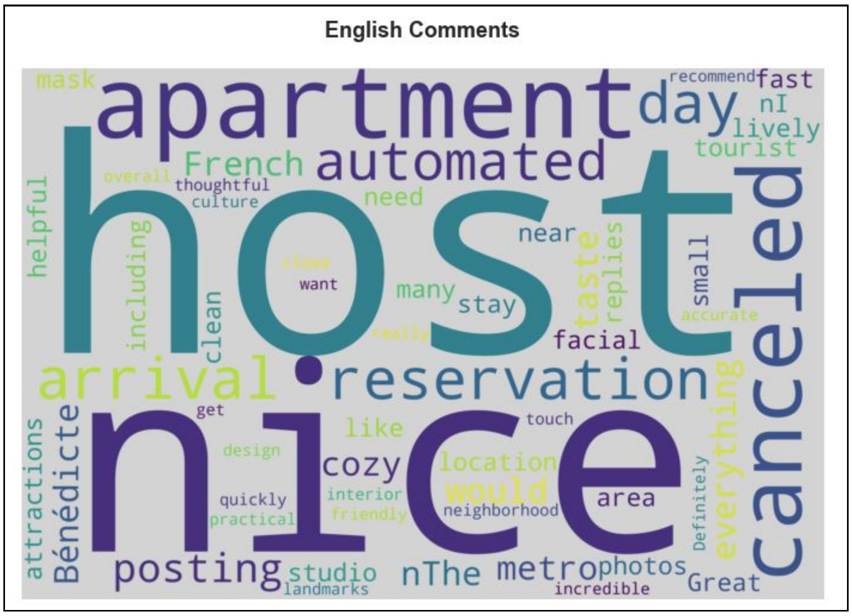

English Comments

Figure 14.

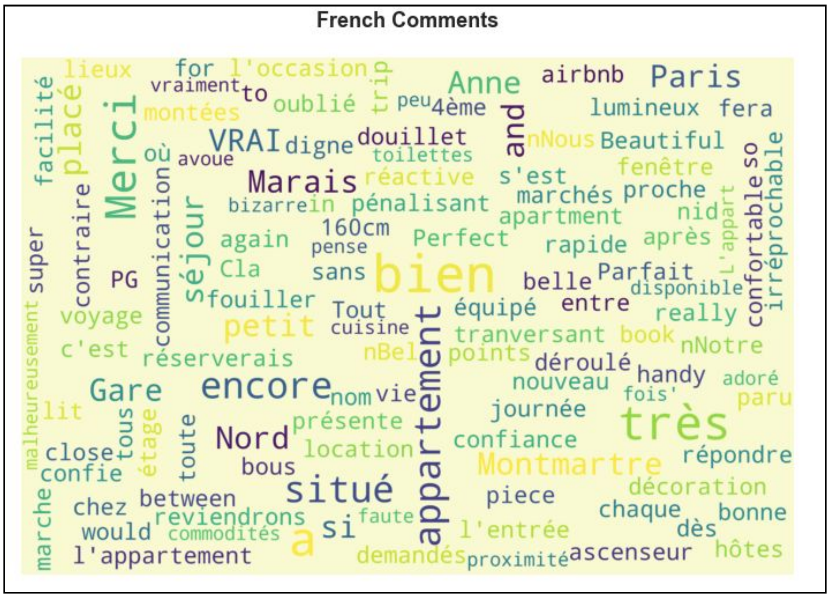
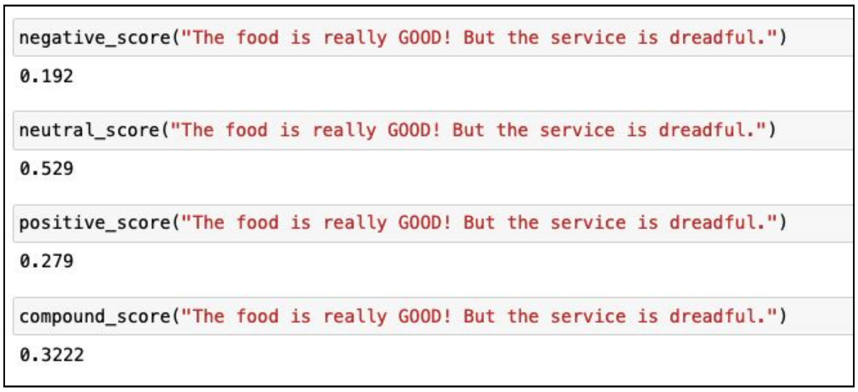

French Comments

Figure 15.

We observed that the most frequent words relate to the host (“host”, “french”, “helpful”), the facilities nearby (“metro”, “area” ,“location”), Paris areas (“Montmartre”, “Marais”) and the apartment itself (“lumineux”, “beautiful”, “petit”).

1.2. Sentiment Analysis

We calculated the sentiment score of the reviews. We used the VADER package to produce four scores for each of our english comments. Vader produces four sentiment metrics (positive, neutral, negative, compound score). The compound score represents the sum of all of the lexicon ratings which have been standardised to range between -1 and 1. Our example sentence was rated 19% negative, 53% percent neutral, 28% positive (Figure 14) with a compound score of 0.32.



```
negative_score("The food is really GOOD! But the service is dreadful.")
0.192

neutral_score("The food is really GOOD! But the service is dreadful.")
0.529

positive_score("The food is really GOOD! But the service is dreadful.")
0.279

compound_score("The food is really GOOD! But the service is dreadful.")
0.3222
```

Figure 16.

We generated some descriptive statistics that summarise the central tendency and dispersion of our dataset’s compound score.

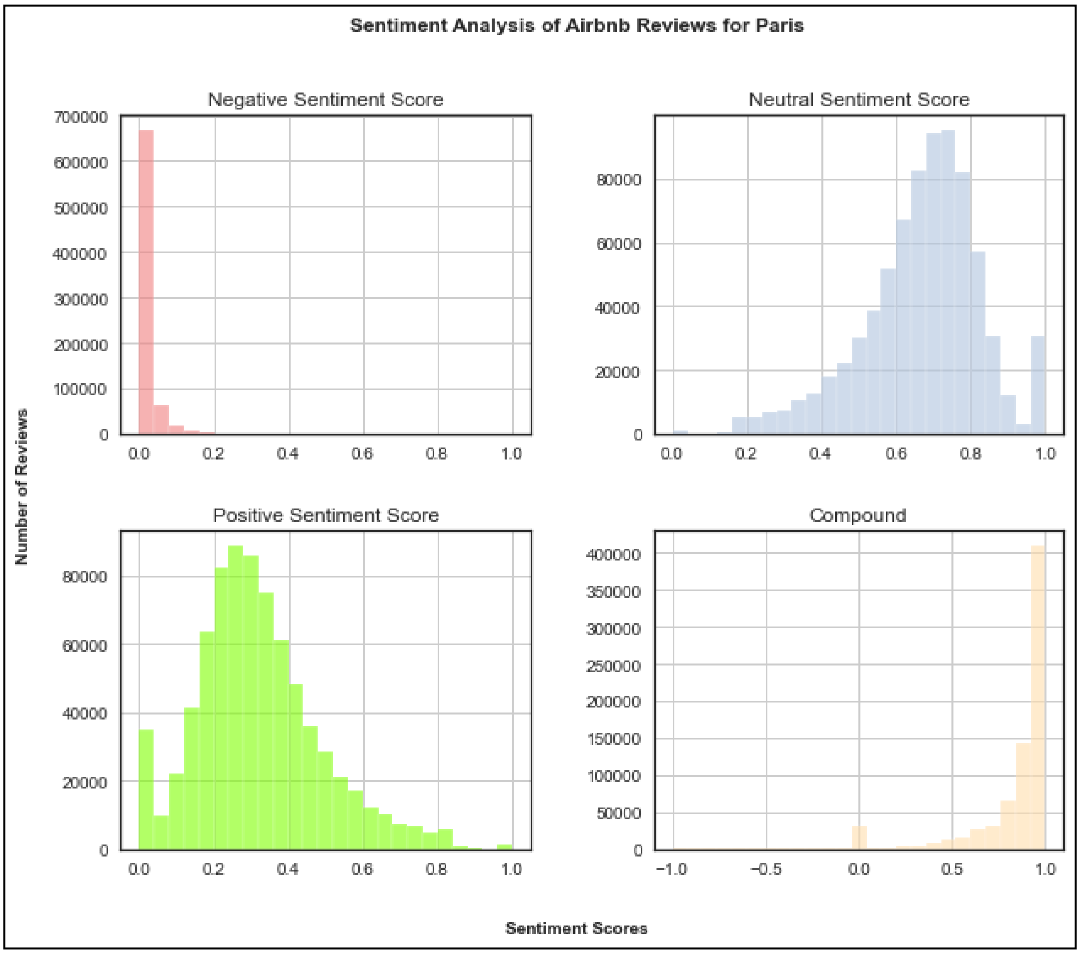


Figure 17.

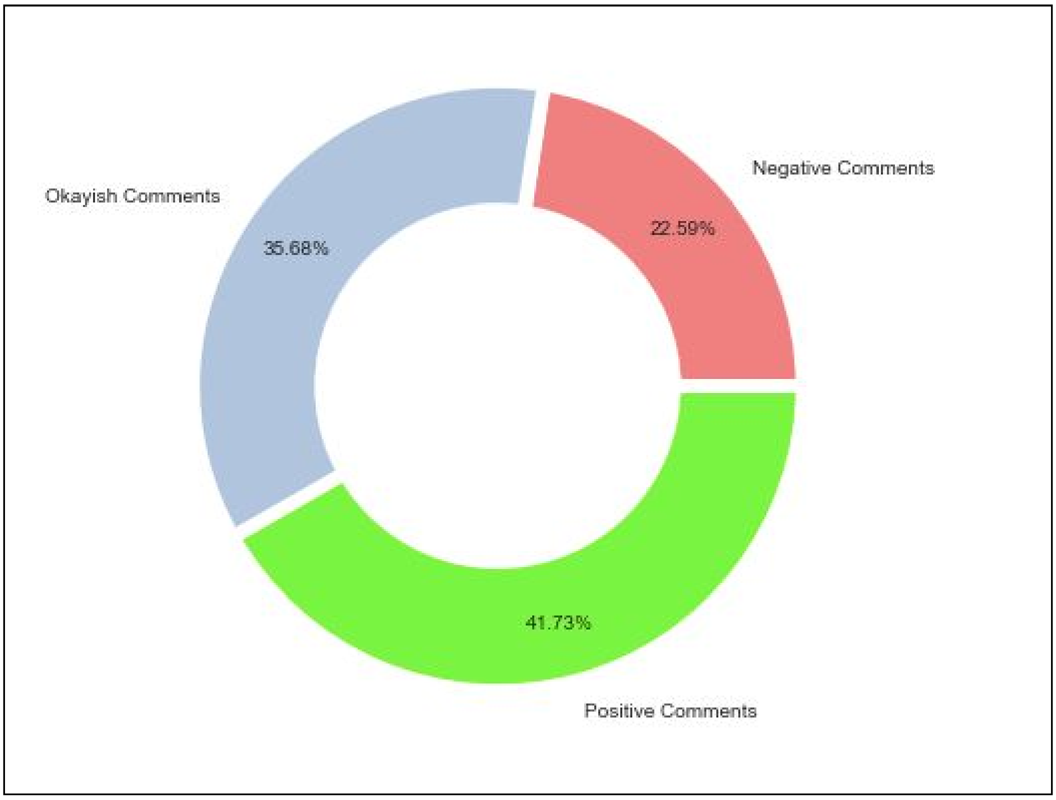


Figure 18.

Clearly, most reviews in the dataset are positive (42%) and there are 23% negative comments.

We wanted to check whether there was a difference in the comments' sentiments based on paris districts, but no clear trends appeared.

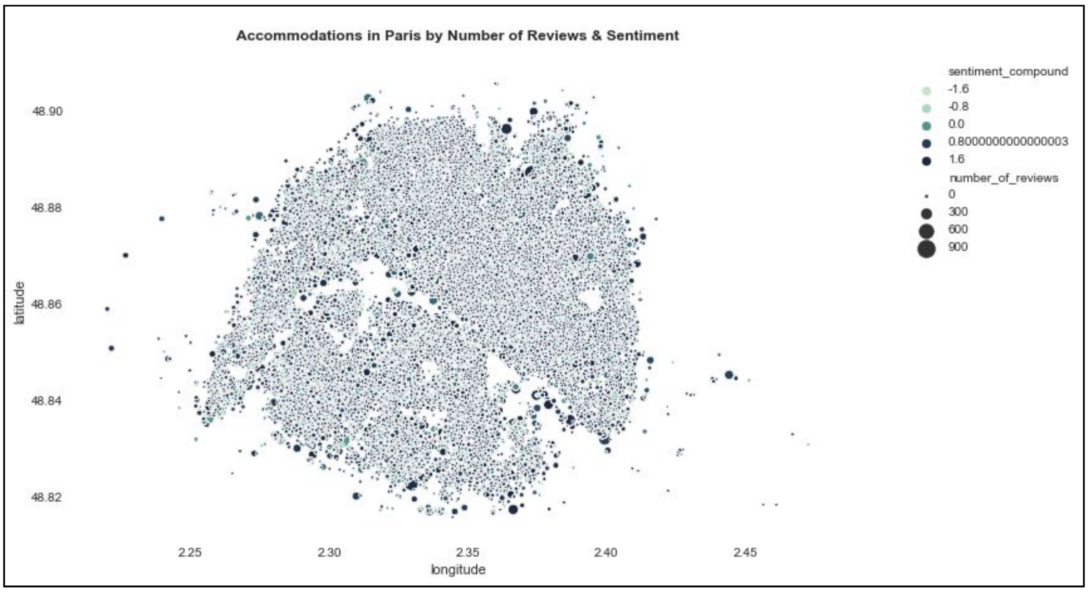


Figure 19.

1.3. Frequency Distribution

We used a method called frequency distribution that helped us to detect the prevalence of certain words. The 30 most frequent words used in positive comments are illustrated in the figure below. The most positive words are related to the apartment and to the location.

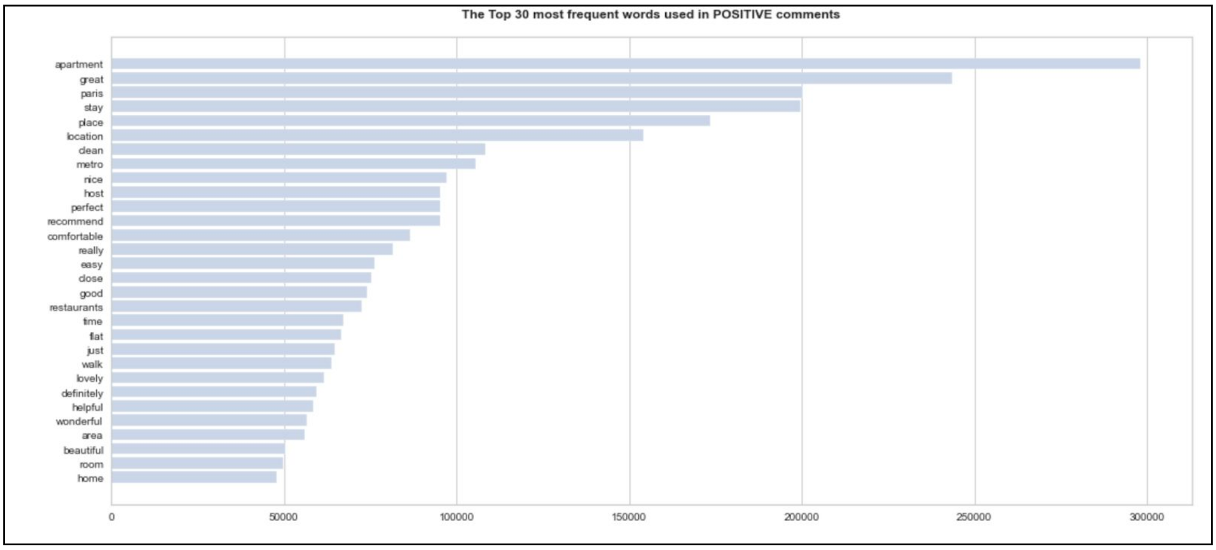


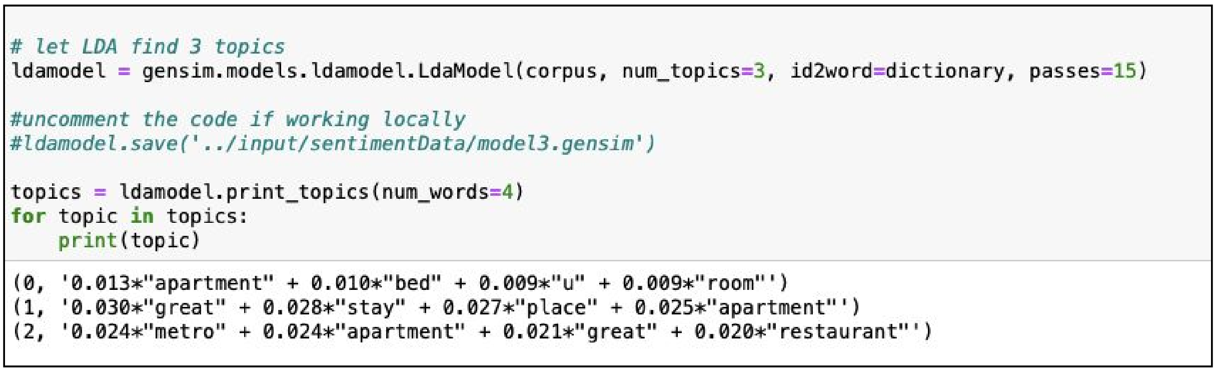
Figure 20.

1.4. Topic Modelling

To do topic modeling we used the LDA method.

Positive comments

We first investigated the positive comments. The first topic includes words such as “apartment”, “metro” and “restaurants”. Thus, the first topic is related to the facilities located near the apartment. The second topic includes words such as “place”, “stay” and “apartment” which seem to be related to comfort of the apartment and satisfaction of the customer . The last one includes words such as “apartment”, “bed” and “room”.



```
# let LDA find 3 topics
ldamodel = gensim.models.ldamodel.LdaModel(corpus, num_topics=3, id2word=dictionary, passes=15)

#uncomment the code if working locally
#ldamodel.save('../input/sentimentData/model3.gensim')

topics = ldamodel.print_topics(num_words=4)
for topic in topics:
    print(topic)
```

```
(0, '0.013*"apartment" + 0.010*"bed" + 0.009*"u" + 0.009*"room"')
(1, '0.030*"great" + 0.028*"stay" + 0.027*"place" + 0.025*"apartment"')
(2, '0.024*"metro" + 0.024*"apartment" + 0.021*"great" + 0.020*"restaurant"')
```

Figure 21.

Similarly, we used LAD to find 5 and 10 topics and found the same patterns.

Negative comments

After, we investigated the most negative comments. We found that it is crucial for the hosts to have a comfortable place, with at least a clean shower. Also, pictures of bad quality or that do not truly depict the apartment are redibidatory. Bad communication could also harm the rentals.

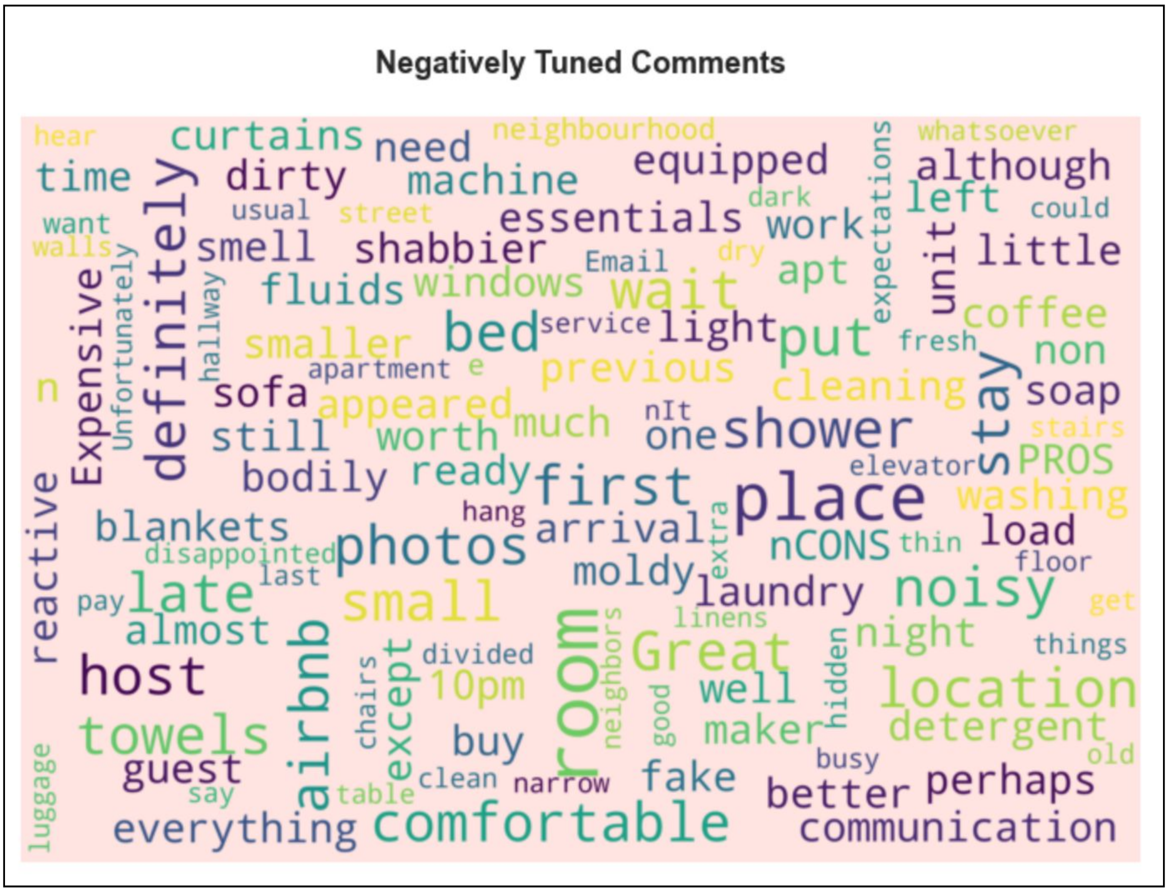


Figure 22.

Moreover, we see that the top 30 most negative comments are related to the flat (overall flat and room), the host and the location.

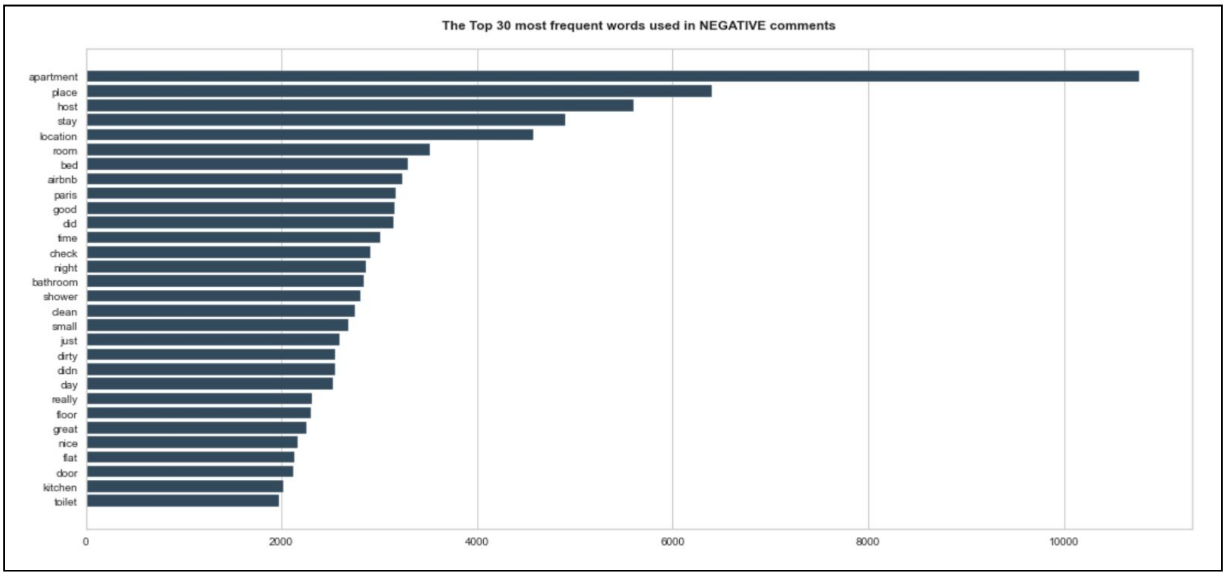
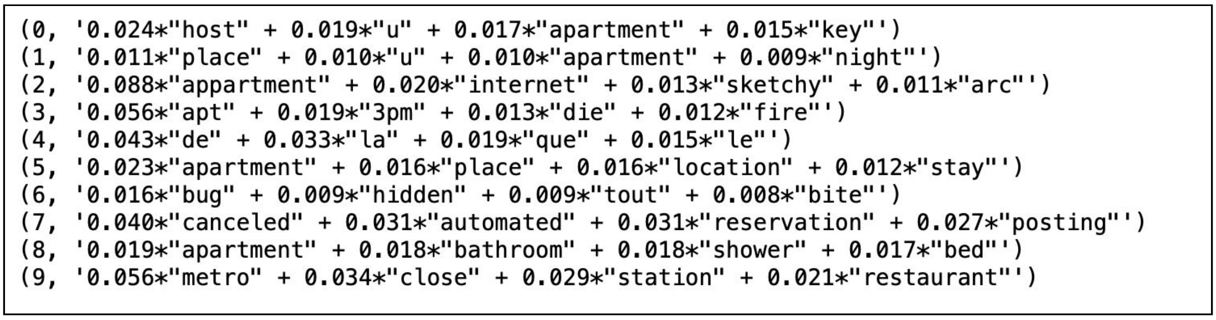


Figure 23.

Overall the final topic modelling analysis shows that what makes an apartment negatively rated is:

- **Host bad behaviours** such as cancellation of the trip as demonstrated by ‘host’, ‘reservation’, ‘canceled’, ‘posting’;
- **Maintenance problems** that does not meet the customers’ expectations as demonstrated by ‘apartment’, ‘bathroom’, ‘bed’, ‘internet’, ‘shower’;
- **Bad location:** the apartment is too far away from transport connections and restaurants as shown by ‘restaurant’, ‘station’, ‘metro’



```
(0, '0.024*"host" + 0.019*"u" + 0.017*"apartment" + 0.015*"key"')
(1, '0.011*"place" + 0.010*"u" + 0.010*"apartment" + 0.009*"night"')
(2, '0.088*"appartment" + 0.020*"internet" + 0.013*"sketchy" + 0.011*"arc"')
(3, '0.056*"apt" + 0.019*"3pm" + 0.013*"die" + 0.012*"fire"')
(4, '0.043*"de" + 0.033*"la" + 0.019*"que" + 0.015*"le"')
(5, '0.023*"apartment" + 0.016*"place" + 0.016*"location" + 0.012*"stay"')
(6, '0.016*"bug" + 0.009*"hidden" + 0.009*"tout" + 0.008*"bite"')
(7, '0.040*"canceled" + 0.031*"automated" + 0.031*"reservation" + 0.027*"posting"')
(8, '0.019*"apartment" + 0.018*"bathroom" + 0.018*"shower" + 0.017*"bed"')
(9, '0.056*"metro" + 0.034*"close" + 0.029*"station" + 0.021*"restaurant"')
```

Figure 24.

This finding helps us to answer our first research question. AirBnB can suggest to hosts how to interact with renters, set reminders for the hosts to frequently check their accommodations maintenance and recommend hosts to add travel advice to their accommodation’s description. This can help improve their reviews/ratings which will in turn attract more renters and increase revenue for AirBnB.

2. XGBoost model

This model helped us to identify the most important criteria for predicting prices. Thanks to this model, we discovered that the most important features to predict the prices were the number of bathrooms, the size of the apartment, and the presence of amenities such as a television, which account for approximately 50% of the daily price. Other important features are the number of bedrooms, the availability of the apartment and the number of people that the host can accommodate (see figure below).

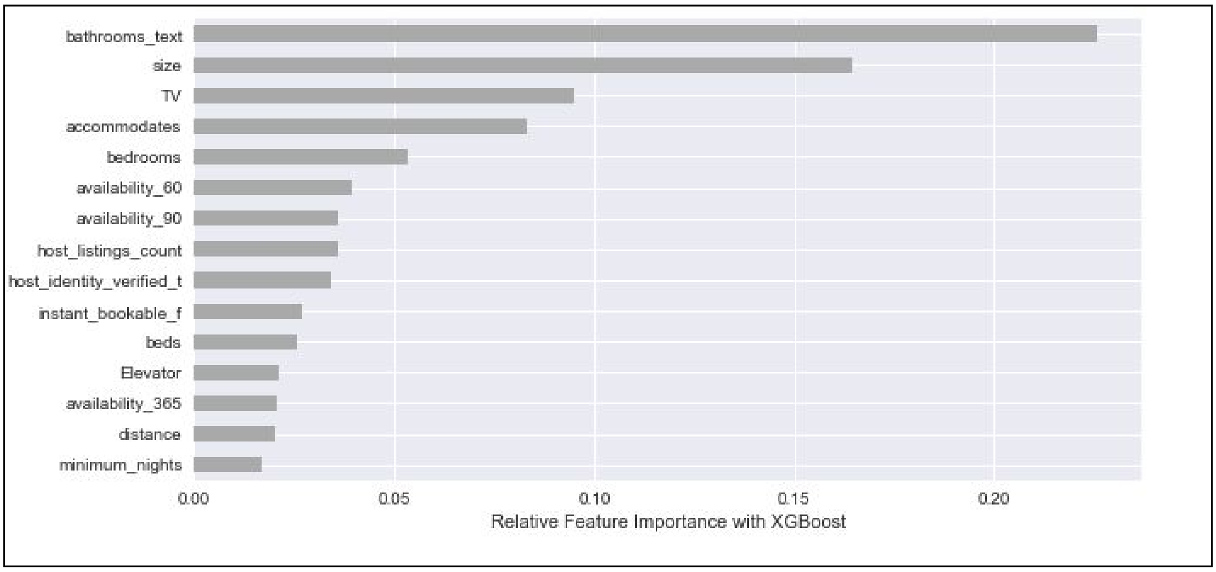
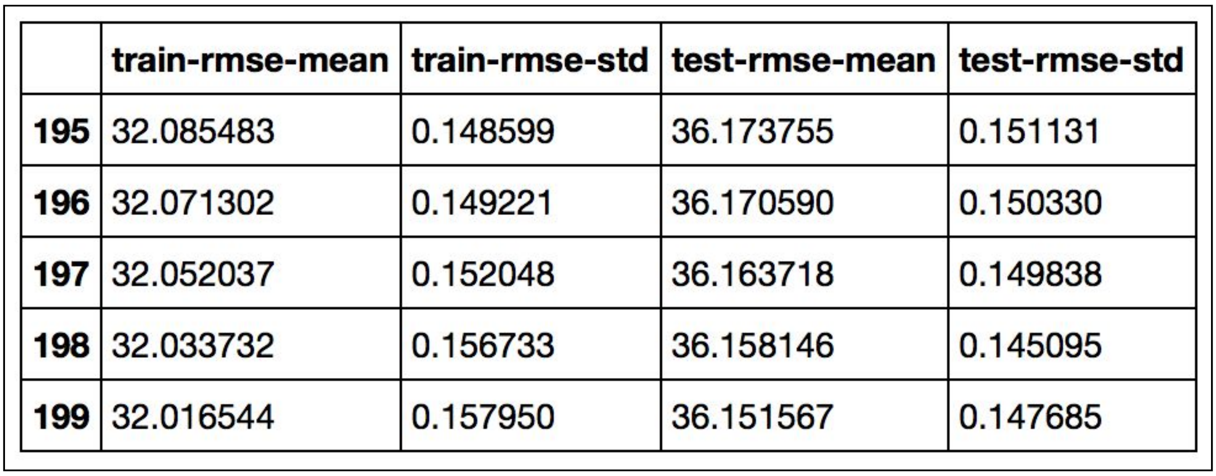


Figure 25.

With our model we found a RMSE of $35.2 which improves to $32 after Cross Validation, and with our analysis, we have explained 67% of the variance (R^2).



	train-rmse-mean	train-rmse-std	test-rmse-mean	test-rmse-std
195	32.085483	0.148599	36.173755	0.151131
196	32.071302	0.149221	36.170590	0.150330
197	32.052037	0.152048	36.163718	0.149838
198	32.033732	0.156733	36.158146	0.145095
199	32.016544	0.157950	36.151567	0.147685

Figure 26.

These results help us answer the second research question. AirBnB could ask their hosts to submit information for each of these variables and then use our model to recommend hosts a range of prices for their accommodation. However, one would need further analysis with more variables to come up with a lower standard variation and more precise price predictions. By using appropriate prices, AirBnB can increase their revenue by minimising the amount of accommodations that are underpriced, and increase demand by reducing the amount of overpriced accommodations.

IV. Directions for Future Research

To conclude, the Xgboost analysis focuses on the flat itself emphasizing the importance of the size, the number of accommodations, the number of bathrooms and bedrooms. The NLP analysis allows us to get a bigger picture and understand that the hosts' attitudes and the flats' locations are as important as the flat itself.

The next step for our future research would be to generate another XGboost model using more features such as quality of presentation (pictures), communication (host response time) and number and content of reviews to create an accurate pricing predicting model with a lower RMSE rate. This model then can be used by AirBnB to suggest to hosts an exact price instead of a price range to ensure appropriate pricing.

It is also important to note that the prices from the listings dataset are the prices the hosts set, but may not be the actual price the renters paid. Also, prices will vary at different times of the year. Therefore, for our future research it will be interesting to create a price predicting model for AirBnB's hosts using prices that renters actually pay and that considers the time of the year.

It would also be interesting to create different description formats for accommodations and analyse their effectiveness (by analysing the attention they receive). Then AirBnB can send the most effective format to hosts for them to use when they post their accommodation in order to attract more renters.

V. References

Andjelic, A. & Davidoff, J., 2020. *Why Airbnb made a big mistake by ditching its marketing.* [Online]
Available at:
https://www.fastcompany.com/90489914/why-airbnb-made-a-big-mistake-by-ditching-its-marketing
[Accessed 13 December 2020].

Carville, O., Roof, K. & Tse, C., 2020. *Airbnb Valuation Reaches $100 Billion in Trading Debut Surge.* [Online]
Available at:
https://finance.yahoo.com/news/airbnb-set-huge-day-one-161157860.html
[Accessed 13 December 2020].

Fast Company, 2020. *Airbnb.* [Online]
Available at: https://www.fastcompany.com/company/airbnb
[Accessed 12 December 2020].

Glenday, J., 2020. *Airbnb Halts Marketing Spend In $800M Savings Plan.* [online] The Drum.
Available at:
https://www.thedrum.com/news/2020/03/30/airbnb-halts-marketing-spend-800m-savings-plan
[Accessed 17 December 2020].

Inside Airbnb, 2020. *Get the Data.* [Online]
Available at: http://insideairbnb.com/get-the-data.html
[Accessed 05 December 2020].