Best features recommendation for Airbnb in Amsterdam

1. Introduction

1.1 Background

Considering that each year the number of clients of this service is increasing, it is a good option to start now in this business. There are many reasons why people prefer to stay in an Airbnb than in a hotel, for example the price is one of the most common reasons for long stays because normally hosts offer special prices for those situations, also people like to stay in a house for the experience of living in the place that they're visiting. So, with the boom of Airbnb, many entrepreneurs want to get inside the business and many of them do not know where to start. Although there are many competitors in great cities like Amsterdam it stills being a great idea to enter in this field due to the large amount of visitors that the city receives all year.

1.2 Problem

The main problem is not that they may don't have a house or an apartment because they rent one to start an Airbnb business, the main problem is that they do not know where has to be located and mainly which characteristics do the house is better to have in order to have more revenue and a large quantity of clients during the year.

1.3 Interest

Of course, people that already have a property or new entrepreneurs with a nice business mindset will be interested in the data provided for this project. Also consulting companies would be interested.

2. Data acquisition and cleaning

2.1 Data sources

The data is an open-source non-commercial licensed with many features about Airbnb in the city of Amsterdam extracted from http://insideairbnb.com/ that offers really valuable information about the topic. This dataset contains information of more than 20k Airbnb around Amsterdam, so the metrics will be accurate, but it will be necessary to clean and prepare the data.

2.2 Data cleaning

The data downloaded presents very useful values but there were a lot of missing values, so it was necessary to drop those rows with no information for avoiding a bad clustering. Certainly, it is not enough to drop non information values, it was also necessary to analyze each row that was going to be used and look if there was rare data.

For the field “prices” it was detected that the data type was string so it was changed to int for a correct processing. Checking the values it has been found that some prices were “over-valued” or “under-valued” for example with prices of 10 USD per night or more than 8000 USD per night, this is not correct for training the model because the standard deviation will be too wide, so the action taken was to filter some prices that were under 30 USD and more than 1000 USD.

It was analyzed the field “minimum_nights” and had a lot of values that were to high (around 1000 minimum nights required to stay there), so we filter up to 28 minimum nights.

2.3 Feature selection

First of all, there were around 20,025 rows and 106 features with the raw dataset, and after dropping rows with “Nan” values only 17,362 left. For the feature selection it was necessary to drop many features (columns in the dataset) that were not useful to cluster the data.

There were a lot of extra features for example “cancellation_policy” that was the cancellation policy for the property in text, it was dropped because we can’t process that information to make clusters.

Also, there were many others features that only have text not useful for the training model like “name”, “summary”, “space”, “description”, etc. This kind of information may be great for making a machine learning model or with CNN that recommends specific names, description or summaries for having better results, but in this project, we are not going to use them.

Because most of the features were dropped due to the lack of utility and importance, we will show only the features that we use for the project.

Feature	Reason for keeping
Latitude and longitude	It is displayed a map for the properties with the better metrics.
Bathrooms, bedrooms, beds, guests_included, minimum_nights	Necessary to analyze the property and characteristics.

number_of_reviews_ltm, first_review, last_review, review_scores_rating, review_scores_checkin, review_scores_location	Necessary to define whether is good or not
Extra_people, price	Very important features to analyze the estimated recommended price.
Reviews_per_month	Really useful to determine how often the property have clients and analyze whether the characteristics of a property are attractive or not.

3. Exploratory Data Analysis

3.1 Data values analysis

It is impressive that most of the Airbnb in Amsterdam have great ratings, it could be like that because it is a nice city, but there are differences remarkable differences in ratings and in the characteristics of the properties.

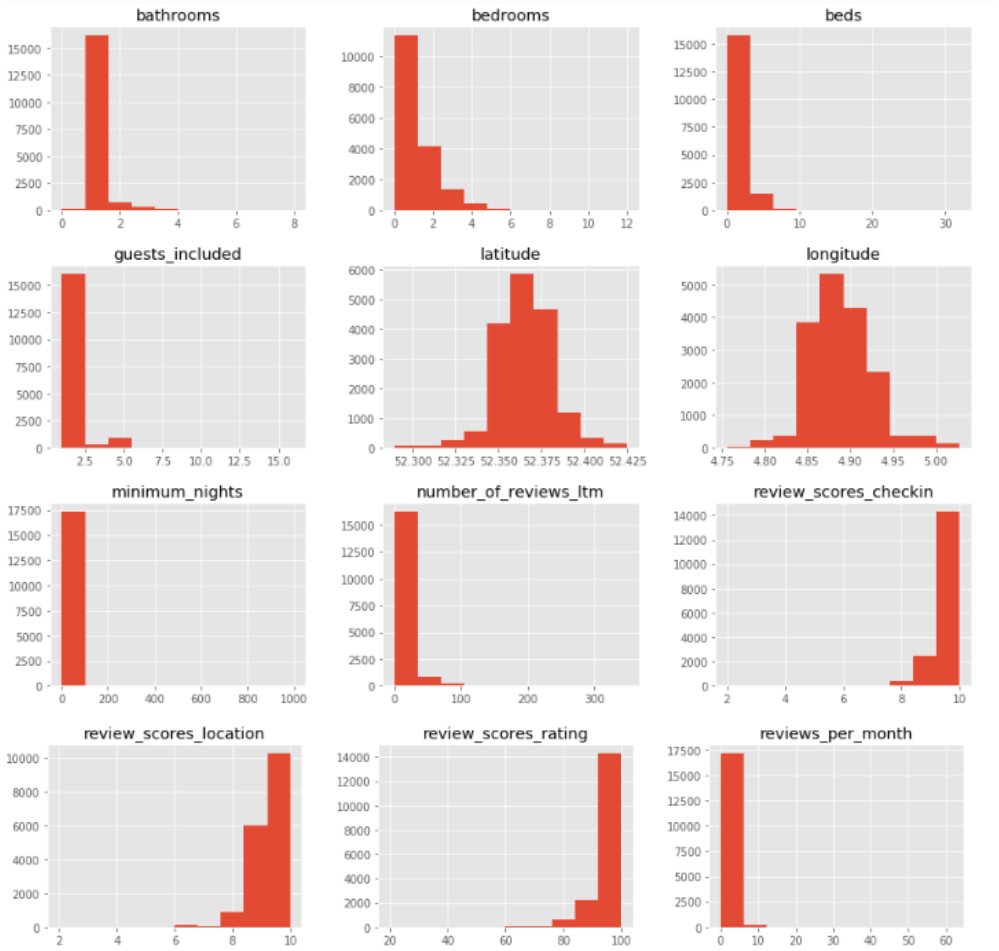


Fig. 1 Data Histograms

As we can see in the histogram above, many characteristics have a high frequency but in the correlation matrix we can observe that there are many features that are not correlated among them but some have certain relationship, like in the case of “number_of_reviews_ltm” that is the number of reviews in the las twelve months it is obviously correlated to “reviews_per_month”. Also, some of the features had a relationship, just like bathrooms, beds and bedrooms because we know that while more bedrooms more beds and bathrooms.

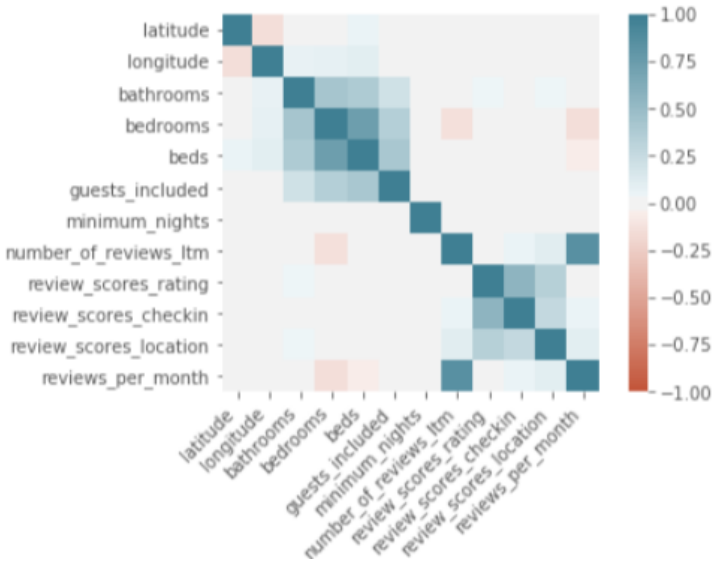


Fig. 2 Data Correlation Matrix

longitude	0.014656
bathrooms	-0.012838
bedrooms	-0.145813
beds	-0.055464
guests_included	-0.012148
minimum_nights	-0.155830
number_of_reviews_ltm	0.857106
review_scores_rating	0.002733
review_scores_checkin	0.050958
review_scores_location	0.104955
reviews_per_month	1.000000
Extra_people	0.004368
Price	-0.013269

Fig. 3 Data Correlation

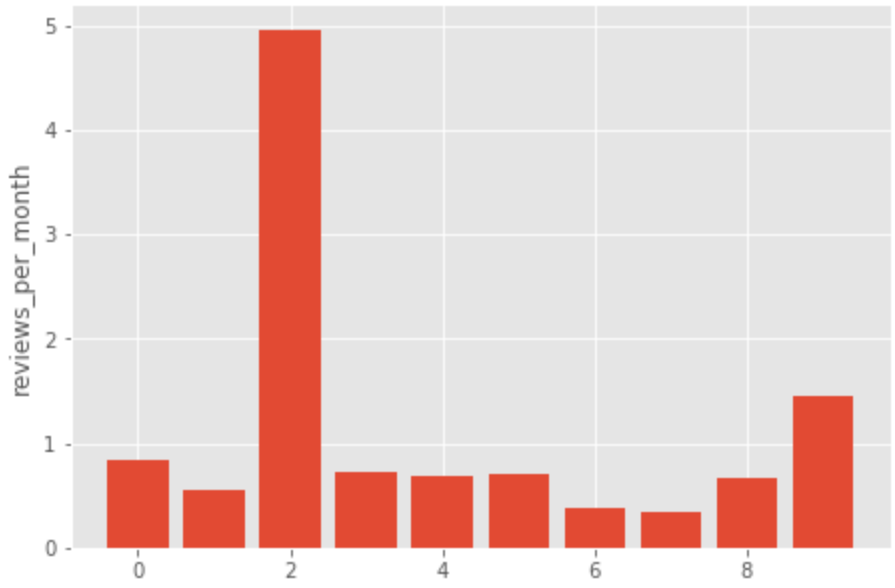
4. Clustering Algorithm K-means

For this study case we implemented the K-means algorithm for clustering groups of Airbnb properties in respect of the features mentioned. K-means clustering is a method of vector quantization, originally from signal processing, that is popular for cluster analysis in data mining. k-means clustering aims to partition n observations into k clusters in which each observation belongs to the cluster with the nearest mean, serving as a prototype of the cluster.

4.1 Solution to the problem

In this project several clustering algorithms were studied to handle the goal, although K-means had demonstrated a great work in grouping and labeling 10 different kinds of properties based on their characteristics to obtain those which have better results in terms of “reviews per month” because a property that has more it means that the revenue is higher.

After modeling the algorithm with ten clusters it was obtained values that shown the best characteristics for an Airbnb. For getting the results (labels) it was necessary to obtain the mean and median, with these it was possible to determine under which attributes it is probable that a airbnb will be most likely to be successful.



In the fig. 4 it is shown the metrics that each label obtained, obviously those attributes clustered as “2” had almost five reviews per month, while others had one if rounded.

	bathrooms	bedrooms	beds	Guests included	Minimum nights	Number Of reviews	Number of reviews ltm	Review Scores rating	Extra P.	Review Scores Checking	Review Scores Location	Price	Reviews Per Month
Mean	1.12	0.98	1.43	1.37	1.97	56.08	95.08	9.87	0	9.87	9.84	114	4.95
Median	1	1	1	1	2	53	97	10	0	10	10	95	4.56

Fig. 5 Results for the better characteristics

The table above shows how the features influence in the revenue that you can get offering an Airbnb service in a property. In comparison with the other labels, this one shows that a simple house or apartment for one person that is around 100 USD is more likely to win when a customer is checking where to stay. The difference among the others in reviews per month is easy to verify.

5. Conclusion

In this case study it was possible to determine and recommend which characteristics an Airbnb is better to has in order to get more clients and as of course more revenue. This approach show us how little things in a service that many people offer can be beneficial to know for staying ahead the competitors and the most important thing, that was the purpose of the project, is to give useful information to those who are looking to enter in this service and do not know which features a property should have. This project could be extended in determine other useful things for raise the probability of a successful Airbnb by analyzing other types of features such as name of the property, description, summary and mainly in property's pictures because it has been studied a lot how pictures are the best marketing in a wide quantity of business.