Final Project

Project Big Data - Airbnb

June 30, 2024

Authors: Valeria Foresti (2704811)
Luca Perego (2821704)
Rudo Straka (2746162)
Fanyi Zhao (2759982)

Topic: Airbnb

Group number: 9
Course: Project Big Data

Contents

Introduction

Airbnb was founded in 2008 and has completely revolutionized the hospitality industry by allowing people to rent out their properties or spare rooms to travelers around the globe. As a peer-to-peer online marketplace, *Airbnb* connects hosts, who offer short-term stays, with guests, who seek accommodations. This platform provides a wide variety of rental options ranging from single rooms to entire homes to hotel rooms, appealing to a wide range of traveler needs and preferences.

Despite its popularity, *Airbnb* has faced criticism and regulatory challenges concerning its impact on housing markets. Particularly, in the Netherlands, major cities like Amsterdam, Rotterdam, and The Hague have experienced housing shortages, partly attributed to the proliferation of short-term rentals. One argument being that *Airbnb* contributes to rising property prices and reduced housing availability for long-term residents, exacerbating the ongoing housing crisis, which happened to have started around 2008.

Our project aims to leverage big data to address some of these challenges and improve the *Airbnb* experience for hosts and guests alike. By conducting explanatory data analysis (**EDA**) and **developing predictive models**, we seek to provide valuable insights into the dynamics of *Airbnb* listings in the Netherlands. Our dataset, comprising listings from Amsterdam, Rotterdam, and The Hague, includes several relevant information such as prices, host data, amenities, descriptions, review categories, location, and availability, among many others.

Through our analysis, we aim to answer several key research questions:

1. ***Do the listings present clustering behavior?*** By identifying clusters of listings, we can better understand patterns and similarities among different types of accommodations, which can inform both guests and hosts about competitive positioning and market segmentation.

2. ***What factors/features influence price, if at all?*** Understanding the determinants of listing prices can help hosts optimize their pricing strategies and ensure fair pricing for guests. By analyzing features such as location, amenities, and host attributes, we can identify the key drivers of price variation.

3. ***Do specific amenities facilitate a high price/booking rate/review score?*** High reviews are crucial for the success of *Airbnb* listings, as they build trust and attract more travelers. By examining the relationship between listing features and review scores, we can provide actionable insights for hosts to enhance their offerings and improve guest satisfaction.

Our dataset encompasses both numerical and categorical data, enabling a comprehensive analysis of various factors influencing *Airbnb* listings. By leveraging EDA techniques and predictive modeling, we aim to uncover valuable patterns and trends, offering data-driven recommendations to enhance the overall *Airbnb* experience. This report outlines our methodologies, findings, and implications for *Airbnb* stakeholders, contributing to a more informed and optimal marketplace.

Our Data

For this project, datasets obtained from the website Kaggle were used. Specifically, datasets for three Dutch cities (Amsterdam, Rotterdam and The Hague) were provided separately. In order to facilitate our analysis, the three datasets were concatenated into a single unified data frame using a function. In this function, the name of the city of which a listing (row) corresponds, was added as a new feature in order to recognize which listings falls under which city. Each individual dataset contained 78 features including price, host information, amenities, description, reviews, scores, location, neighborhood, minimum/maximum nights of stay, type of listing, among many others. After careful consideration and evaluation, the most relevant 37 features were kept for our research.

Variables

In this subsection we list and describe those variables (features) that were kept for research:

Name	**Description**
id	Airbnb's unique identifier for the listing
host id	Airbnb's unique identifier for the host/user
host response rate	Host's response rate in percentage %
host acceptance rate	Host's acceptance rate in percentage %
host is superhost	True/False indicating if the host is a superhost
host total listings count	The number of listings the host has
host has profile pic	True/False indicating if the host has a profile picture
host identity verified	True/False indicating if the host's identity is verified
accommodates	The maximum capacity for the listing
beds	The number of bed(s)
price	Daily price in local currency
minimum nights	Minimum number of nights stay for the listing
maximum nights	Maximum number of nights stay for the listing
availability 30	The availability of the listing 30 days in the future as determined by the calendar
availability 365	The availability of the listing 365 days in the future as determined by the calendar
number of reviews	The number of reviews the listing has
review scores rating	Overall rating score for the listing
review scores accuracy	Accuracy rating score for the listing
review scores cleanliness	Cleanliness rating score for the listing
review scores check in	Check-in rating score for the listing
review scores communication	Communication rating score for the listing
review scores location	Location rating score for the listing
review scores value	Value rating score for the listing
instant bookable	Whether the guest can automatically book the listing without the host requiring to accept their booking request
reviews per month	Average number of reviews per month over the listing's life
room type	Either private room, shared room, entire home/apt, or hotel room

Data cleaning

Once the data frame is set up to our convenience, we start the data cleaning process by firstly converting boolean values (T or F) into 0's and 1's such that the column can be used numerically. The 'price' and 'rates' features were originally given as a string object and thus the '$' and '%' were dropped and the data types were changed into floats. Categorical data such as 'room types' (private room, shared room, entire home/apt and hotel room) were One-Hot encoded. Finally,

the data was checked for duplicates and numerical missing values were handled by imputing them with the mean value of each feature. After this was all done, we ended up with a $(11242, 38)$ data frame.

Exploratory Data Analysis

Explanatory Data Analysis (EDA) is a crucial step in our project that will allow us to understand the underlying structure of the dataset identifying patterns, anomalies, relationships, etc. that will allow us to phrase hypotheses for further analysis. The EDA was conducted on the aforementioned merged data set of *Airbnb* listings from three different cities.

Summary statistics

We began by generating the summary statistics for the numerical and categorical variables using the `df.describe()` function. This function returns a summary of the data's features including its counts, means, percentiles, min/max values, mean, and standard deviations, among others, that provide a good starting point in gaining insights related to dispersion, shape, and tendencies.

Table 1: Summary Statistics for the Numerical and Categorical Variables

	price	accommodates	beds	reviews per month	number_of_reviews	review_scores_rating	review_scores_location	availability_30
count	1035.00	1206.00	1183.00	1061.00	1206.00	1062.00	1062.00	1206.00
mean	3316.73	3.54	2.36	1.05	38.32	4.75	4.81	9.41
std	15023.96	1.77	1.60	1.95	78.91	0.30	0.25	10.66
min	25.00	1.00	1.00	0.02	0.00	1.00	2.00	0.00
25%	102.50	2.00	1.00	0.24	3.00	4.64	4.74	0.00
50%	150.00	4.00	2.00	0.53	8.00	4.84	4.88	4.00
75%	219.00	4.00	3.00	1.34	31.00	5.00	5.00	19.00
max	80000.00	14.00	13.00	46.74	994.00	5.00	5.00	30.00

While through these values we can have an idea of the distributions of our variables, we also note that we have outliers (especially in the *price* variable) that we must deal with before proceeding with further analysis.

What variables are correlated with price?

As the next step in our analysis, we explored various correlations within our dataset. To do this, we needed to work exclusively with numerical data, so we created a numerical subset of the dataset. Unfortunately, we found that the dataset exhibits strong correlations primarily among similar variables.

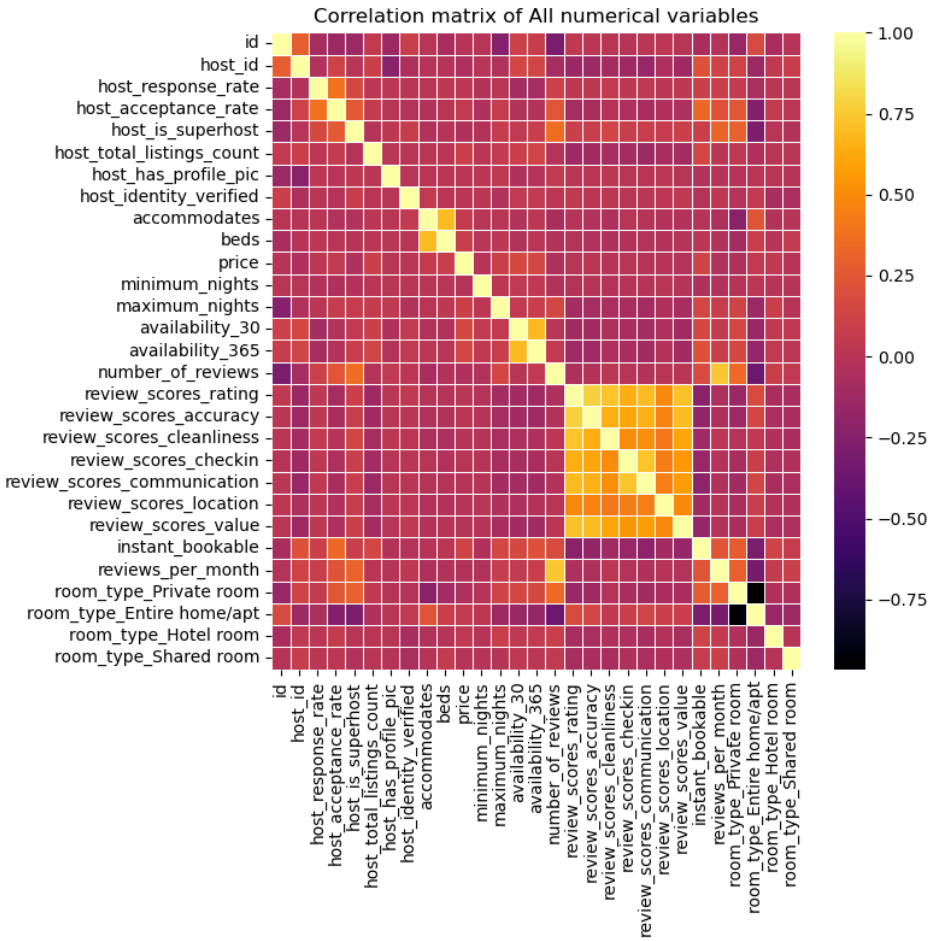


Figure 1: Correlation between numerical features

For instance, as shown in Figure 1, there is a strong correlation between the number of reviews and reviews per month. This is intuitive, as an *Airbnb* listing that receives more reviews per month will naturally accumulate a higher total number of reviews over time. Similarly, there is a strong correlation between the number of beds and the accommodation capacity; more beds allow for more guests to be accommodated. Another notable correlation is between availability over 30 and 365 days, suggesting that customers tend to book their stays more than a month in advance.

Beyond these intuitive correlations, most other correlation coefficients in the dataset are below 0.5. However, some of these less intuitive correlations can still provide valuable insights. For example, there is a correlation coefficient of 0.35 between being a *superhost* and the number of reviews. This indicates that superhosts tend to receive more reviews compared to normal hosts. Another interesting correlation is between private room type and the number of reviews, suggesting that private rooms receive more reviews compared to other types of room.

These findings imply that superhosts and private rooms tend to gather a higher number of reviews. Understanding these correlations can help us gain better insights into the dynamics of *Airbnb* listings.

The reviews section in the provided heat map is particularly intriguing since the review metrics do not show significant correlations with other variables, but they do exhibit strong correlations among themselves. This is somewhat counter-intuitive, as one might expect that the *cleanliness* of a place would not necessarily correlate with the ease of *check-in*.

However, there is an exception: the review score for location. Figure 2 illustrates that the

location review score appears to be relatively independent from the other review metrics. This suggests that even if a guest is satisfied with various aspects of their stay, such as *cleanliness* or *check-in* experience, they may still have strong negative feelings about the location, and vice versa.

This finding highlights the unique importance of location in the context of short-term stays. Location is generally regarded as *the most* influential factor in real estate, and it seems to hold the same level of influence for short-term rentals. Despite positive experiences in other areas, a poor location can significantly impact the overall satisfaction of a guest.

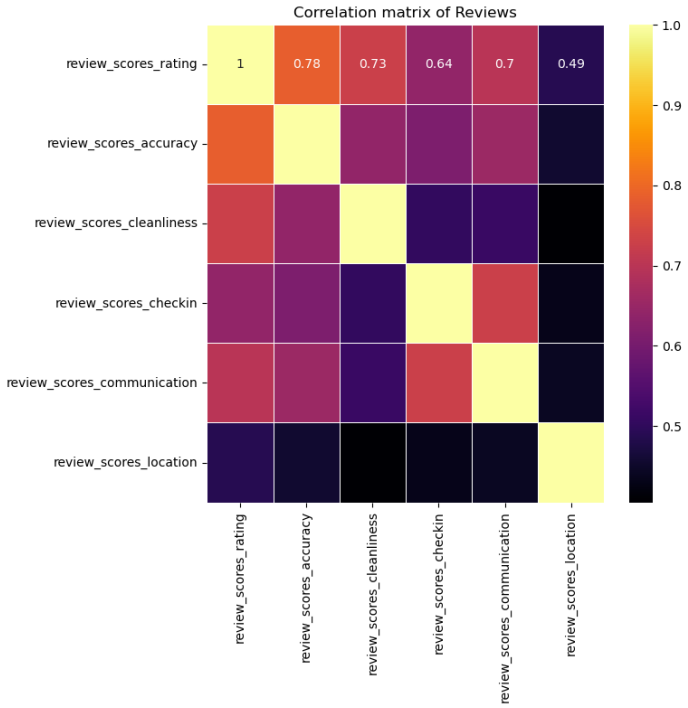


Figure 2: Correlation between review scores

How do prices vary across cities?

The most influential factor that comes with booking a stay for most people is price. For our price variable, there were a few outliers (mostly in Rotterdam), so we decided to cut off at $1,500$ for the box-plot €. As can be seen in Figure 3, Amsterdam had the highest number of outliers with the highest value within the cutoff range. Its prices are generally the highest which could be again anticipated since it is the most popular tourist destination out of the three cities as can be seen in Figure 4b. Further, we divided the cities into their neighborhoods and looked for a corresponding average price per neighborhood. After filtering for outliers we decided to cut off at $1,500$, $1,000$, and 800 respectively for Amsterdam, The Hague, and Rotterdam making Hoogvliet (in Rotterdam) jump from the most expensive neighborhood to the least expensive one because of outliers in that specific neighborhood. Other than that the prices seem to increase gradually within neighborhoods for all three cities with no obvious jump - this can be seen in Figure 4a.

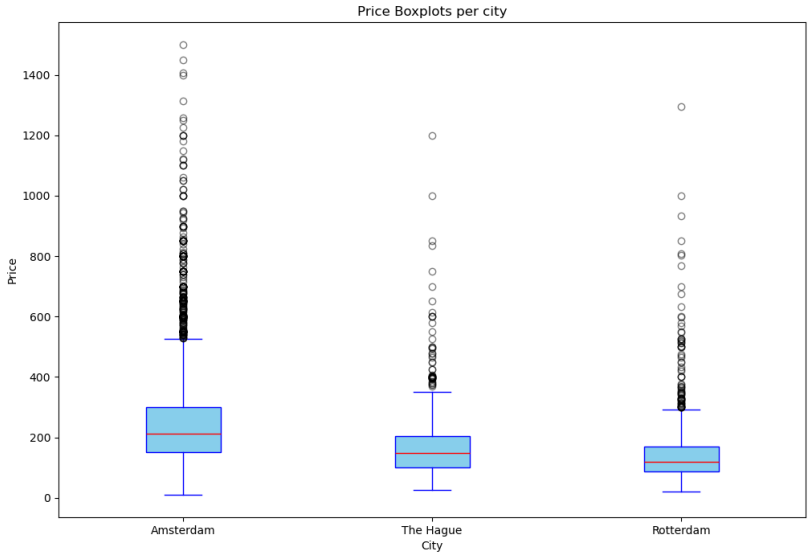
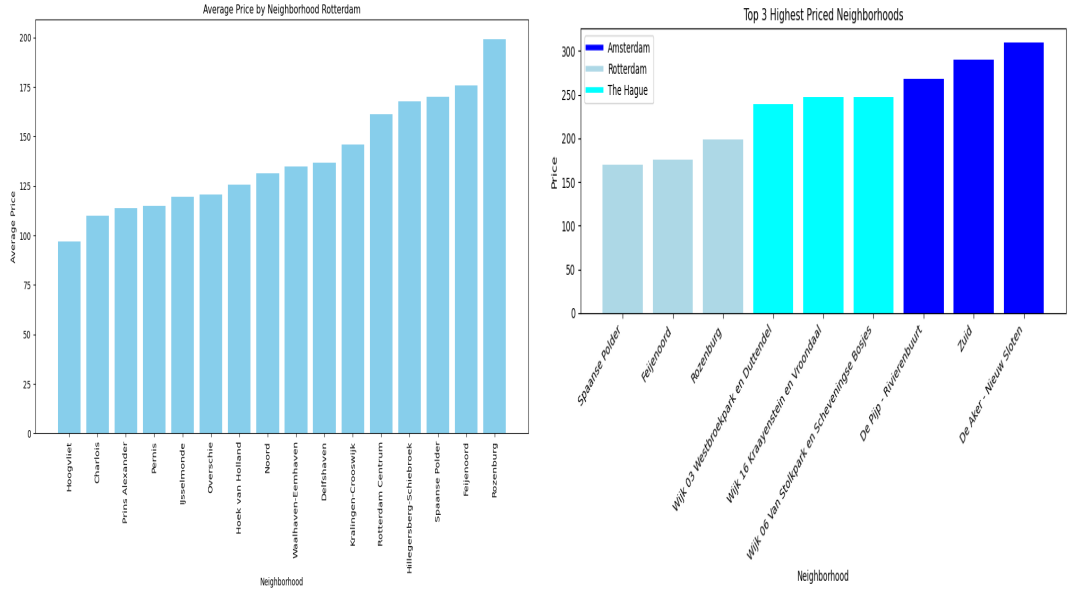


Figure 3: Box-plot of prices for the 3 cities

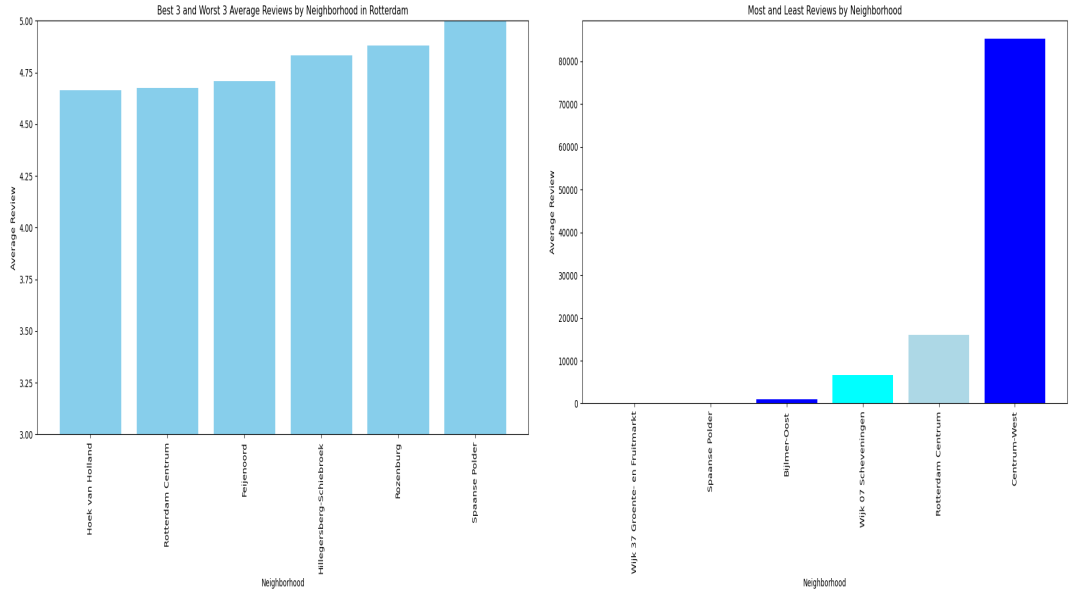


(a) Prices for Rotterdam

(b) 3 most expensive neighborhoods per city

Figure 4: Prices

As mentioned before the number of reviews does not necessarily mean a good thing. As we can see in Figure 5, the most reviewed neighborhood in Rotterdam is among the 3 worst rated ones and the least reviewed is among the 3 with the highest rating. This, however, only occurred with Rotterdam, so it is not a general trend - nevertheless, it suggests that people give a review when they are dissatisfied more often than when they are satisfied.



(a) Best and worst 3 rated neighborhoods in Rotterdam

(b) Most and least rated neighborhoods per city

Figure 5: Reviews

Quick facts about the data

How many hosts and listings do we have?
Our datasets have 6,657 unique properties and 9,234 hosts. There are more hosts than properties - which is intriguing!

When did the 'first' host register?
The 'first' host of our dataset registered on AirBnB on the 24th of September 2008.

What is the property with most reviews?
The property with the most reviews is *CityHub Amsterdam*, an hotel in De Baarsjes, Oud West, Amsterdam. It has 2,575 reviews.

What is the house type with the highest mean rating?
The highest rated house type is... a yurt, with a 5/5 rating. However this is not really informative as there is a single yurt in our dataset, with one single rating.

What are the house type with most reviews?

1. Entire rental unit: 3,834 reviews
2. Entire condo: 1,689 reviews
3. Entire home: 909 reviews

What is the average cost per person for each room type?

1. Entire home/apt: $88
2. Hotel room: $81
3. Private room: $62

4. Shared room: $47

What are the most reviewed neighbourhoods?

The five most reviewed neighbourhoods are all in Amsterdam, they are:

1. Centrum-West: $85,190$ reviews
2. Centrum-Oost: $62,600$ reviews
3. De Baarsjes: $51,452$ reviews
4. De Pijp: $36,445$ reviews
5. Zuid: $24,157$ reviews

Research Questions

We aim to answer the following three questions:

1. Can we predict the price of a given listing?
2. Do some amenities facilitate a high booking rate/review score?
3. Do the listings display clustering behaviour?

1. Predicting a listing's price

For our prediction task we decide to focus on the Amsterdam dataset since it comprises 85% of the total observations, while presenting only 22 neighbourhoods. We assume that the geographical location will be the main drive of price. Hence, with the Amsterdam dataset we can easily one-hot encode the 22 neighbourhood dummies to predict the price.

To select the dependent variables we want to implement in our model, we first study the correlations with price, examined in the EDA section. As we mentioned, there are no numerical variables *strongly* correlated with price. We believe this could be because the strongest drivers of price are not numerical: a large, beautiful house in the middle of the desert could be easily worth less than a small studio apartment in Amsterdam Centrum. Therefore, we decide to select the following variables:

- accommodates: number of people that the house can host
- acc2: the squared term of 'accommodates'
- availability_30: proxy for how much the house is sought after
- Neighbourhood
- Property type
- Room type

Naturally, we one hot encode 'Neighbourhood', 'Property type' and 'Room type'. We also hard trim some price outliers (i.e., the observations with prices above $\$2,000$ a night). We split our dataset in training and test set, with a 20% test split. Additionally, we split our training set in a training and validation set, with a 25% validation split. Before proceeding with the model, we standardize all the numerical variables with `StandardScaler()`.

We decide to implement a *Ridge regression*, which associates an *alpha* parameter used to take care of multicollinearity, by adding a penalty term to the loss function, which shrinks the regression coefficients.

In this situation, the penalty term λ is a *hyperparameter* - to assess its best possible value, we perform 5-fold cross-validation. We grid-search 13 possible values of lambda, from 10^{-6} to 10^{6}: we found that the optimal value is 10. The last step before training our model is carrying

out a log-transformation of the dependent variable, *price*.

The Ridge regression's root mean squared error on the test set for the optimal value of λ is 0.3712. As a diagnostics check, we decide to plot the residuals and the actual price in Figure 6:

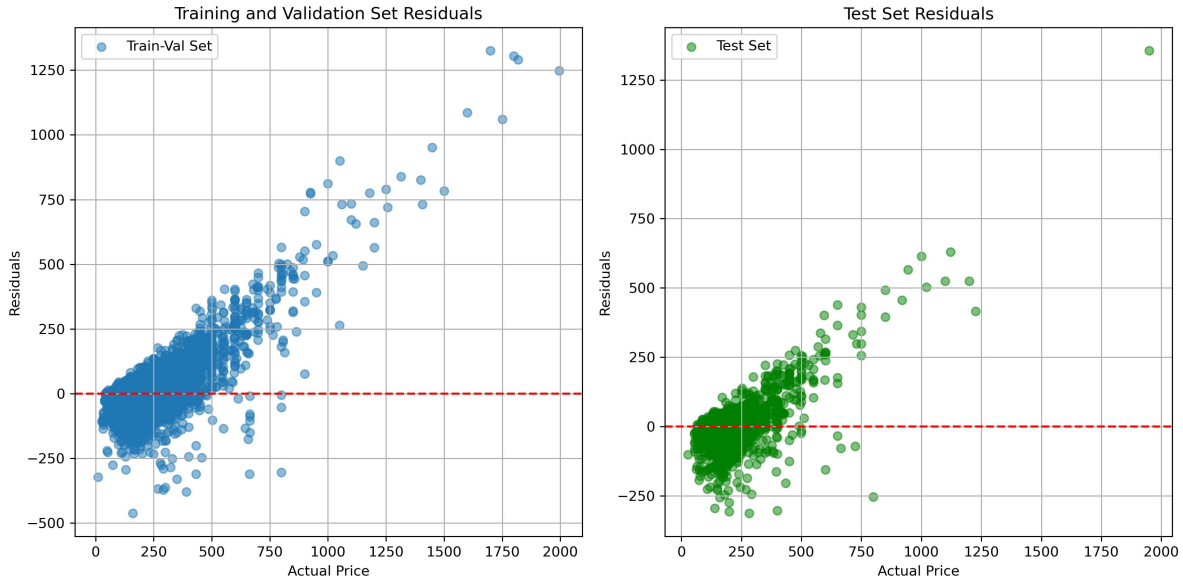


Figure 6: There is almost a linear relationship between residuals and actual price. This is a problem and could imply the presence of omitted variable bias.

It can be noted that the residuals are not centered around zero, but almost display a positive linear relationship with the actual price. This conveys the presence of omitted variable bias. There must be a variable that we did not take into account that is associated with both our dependent and explanatory variables. Unfortunately, while we tried to enhance our feature selection, we were unable to solve this issue, which leads to a disproportionately high RMSE.

2. Do some amenities facilitate a high booking rate or review score? Does the number of amenities influence price?

After reviewing the datasets from all three cities, we found that the *amenities* data is unfortunately missing for the cities of Rotterdam and The Hague. Therefore, our focus in this research will primarily be on the Amsterdam dataset. The *amenities* feature lists a total of $7,853$ unique items, indicating a wide variety of options and combinations that warrants thorough analysis.

This question is structured into three parts: examining the relationship between amenities and pricing, booking rates (measured by *availability_30* and *availability_365*) and review scores (*review_rating_score* on a scale from 1 to 5). Specifically, we aim to investigate how the top 20 amenities contribute to variations in booking rates and review scores.

2.1 Booking rate

Since there are two different booking rates (monthly and yearly), the correlation between them is checked first in Table 2.

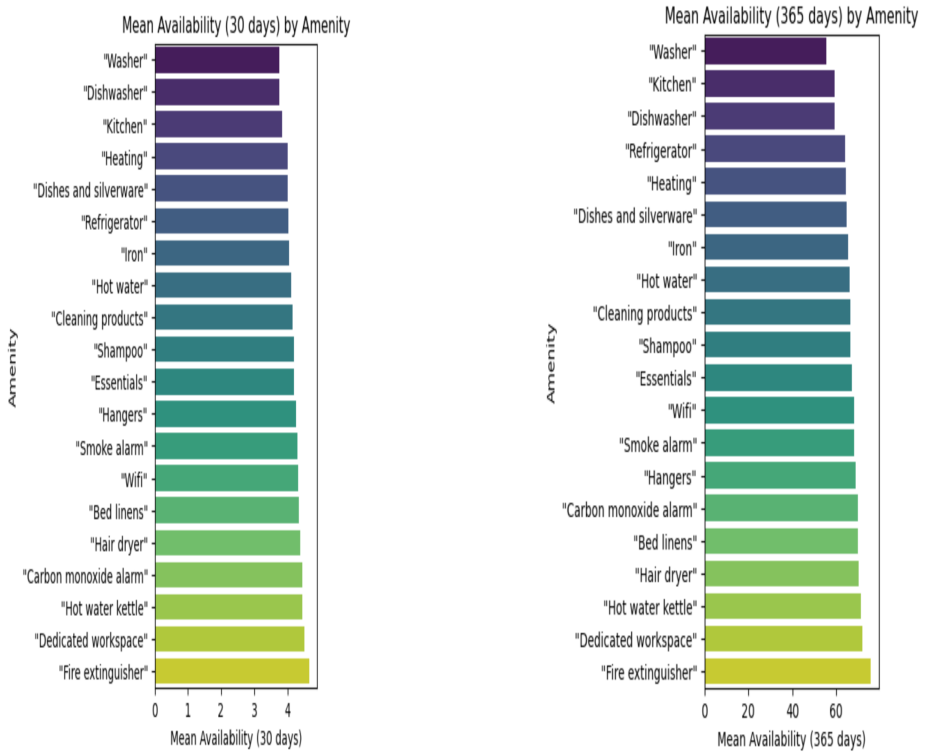
The table shows a significant positive correlation between monthly and yearly booking rates, suggesting that Airbnb listings that are popular on a monthly basis are likely to also be popular throughout the year, and vice versa. Based on this strong correlation, we could make an

Table 2: Correlation between monthly and yearly availability

	availability_30	availability_365
availability_30	1.000000	0.690453
availability_365	0.690453	1.000000

assumption that those amenities influencing one type of booking rate may overlap with those influencing the other.

Initially, we would compute the mean availability for 30 and 365 days based on the presence of the top 20 amenities.



(a) (b)

Figure 7: Mean availability for 30 and 365 days by top 20 amenities

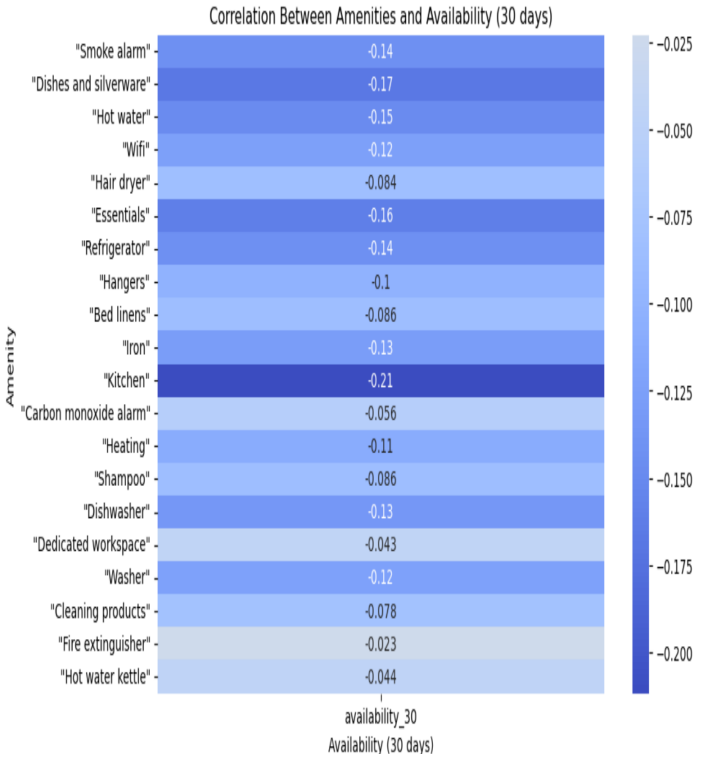
From the two graphs in Figure 7, we could clearly see that in the highly booked Airbnbs on both a monthly and yearly basis, the top three amenities are: washer, dishwasher, and kitchen. The only difference is the ranking order of these, which also indicates the strong positive correlation between the monthly booking rate and the yearly booking rate and verifies our assumption earlier.

Furthermore, we checked the correlation between amenities and availability. In the two graphs of Figure 8, the y-axis order corresponds to the ranking of the top 20 amenities. It is evident that the influence on booking rates does not show a linear correlation with the frequency of occurrence of the amenities. The most influential amenity is the kitchen, both in terms of monthly and yearly booking rates.

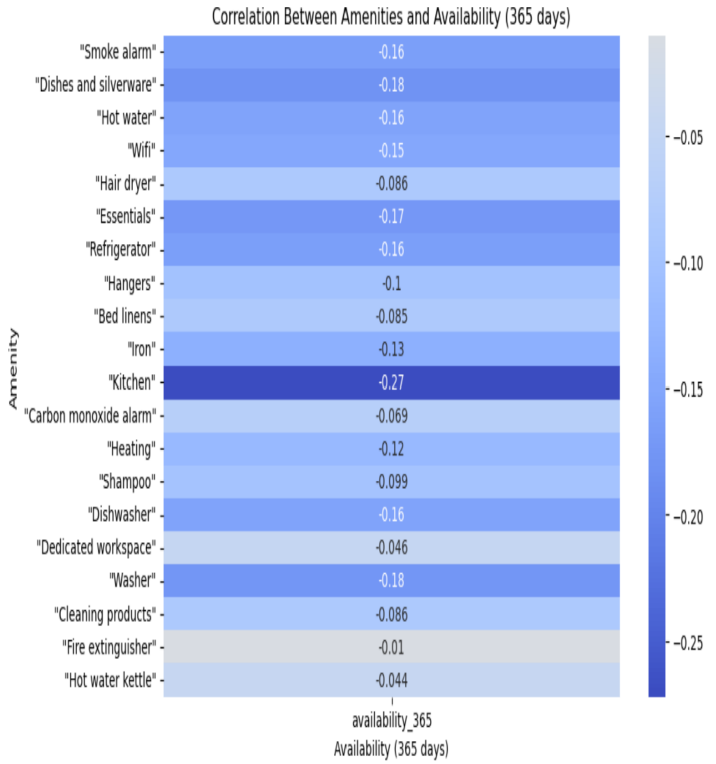
On a monthly basis, the second most influential amenity are the dishes and silverware followed by essentials (toilet paper, soap, towels, pillows, and linens), hot water, smoke alarm, and refrigerator.

On a yearly basis, after the kitchen, the next most influential amenities are the dishes and

silverware, washer, essentials, and smoke alarm, hot water, refrigerator, and dishwasher all ranking equally as fifth in terms of influence on booking rates throughout the year.



(a)



(b)

Figure 8: Correlation between amenities and availability

Based on these findings, it is evident that among these amenities, when considering basic living essentials alongside other amenities, the kitchen stands out as particularly influential which also underscores a notable distinction between traditional hotels and *Airbnb* accommodations.

2.2 Review score

For the review-score rating, we checked the correlation for the top 20 amenities with the review score rating. In Figure 9 we see that the kitchen, dishes/silverware and cleaning products stand out as the most three influential amenities. Interestingly these are the least (apparently) impacting amenities among the top 20. However, it is important to note that its influence could vary seasonally and depending on the specific static dataset and application context.

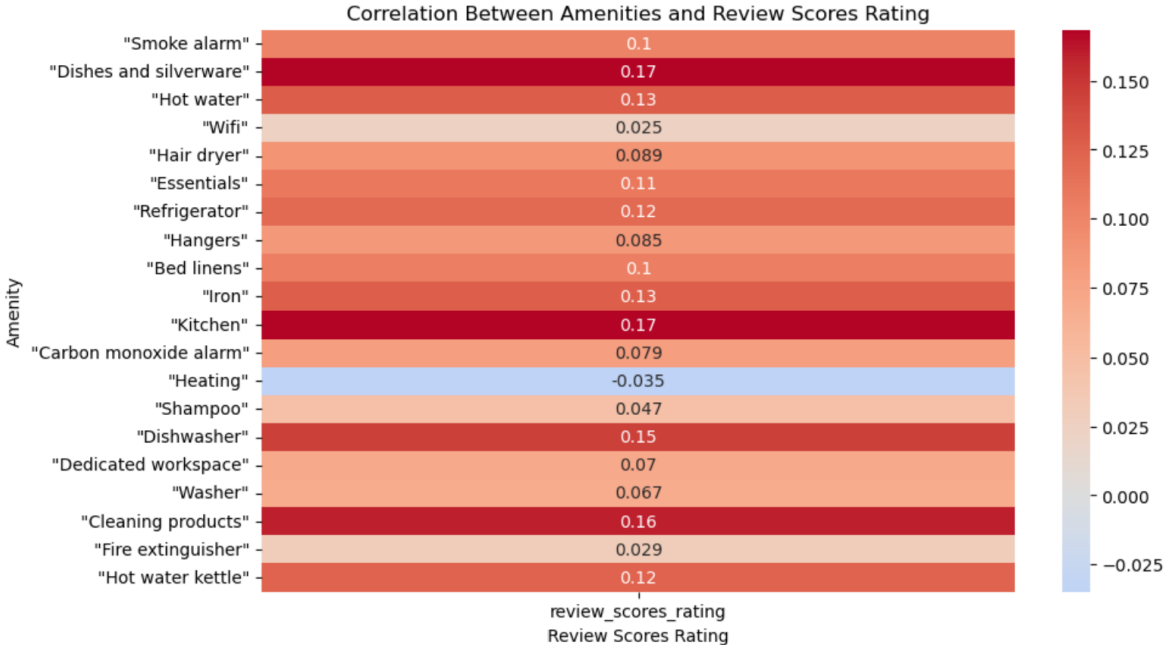


Figure 9: Correlation Between amenities and review scores rating

From the data and analysis presented, it is evident that guests tend to give higher ratings to Airbnbs that provide a home-like environment where they can cook and prepare food themselves. This preference also highlights the importance of amenities like cleaning products when maintaining a comfortable living experience during their stay and so positively contributing to guest satisfaction and overall rating scores.

2.3 Price and number of amenities

To test the hypothesis that more amenities contribute to a higher price, we first performed the following 'naive' linear regression:

$$price_i = n\_amen_i + accommodate_i + review\_score\_location_i$$

Where n_amen represents the number of amenities, *accommodate* represents how many people can be hosted at a given house and $review\_score\_location$ is used as a proxy to measure how good a listing's location is (this should control for good locations being more expensive). We get the following regression result:

Dep. Variable:	price	**F-statistic:**	301.0
R-squared:	0.158	**Log-Likelihood:**	-32230.
Adj. R-squared:	0.158	**No. Observations:**	4813

	coef	**std err**	**t**	**P> |t|**	**[0.025**	**0.975]**
const	-505.0385	51.933	-9.725	0.000	-606.851	-403.226
n_amen	0.0644	0.009	7.236	0.000	0.047	0.082
accommodates	50.9380	2.041	24.952	0.000	46.936	54.940
review_scores_location	117.8161	10.694	11.018	0.000	96.852	138.780

From these results, it appears that the number of amenities does affect price, as the t statistic is above its critical value. Additionally, its sign is counterintuitive, as by being negative it would imply that more amenities contribute to a lower price.

We delve deeper into this issue by estimating a new regression, this time we include squared terms for 'accommodate' and 'n_amen'. Additionally, we delete three outliers. We get the following results:

Dep. Variable:	price	**F-statistic:**	444.7
R-squared:	0.316	**Log-Likelihood:**	-30145.
Adj. R-squared:	0.316	**No. Observations:**	4810

	coef	**std err**	**t**	**P> |t|**	**[0.025**	**0.975]**
const	-478.8998	34.423	-13.912	0.000	-546.385	-411.415
n_amen	-0.0218	0.020	-1.101	0.271	-0.060	0.017
accommodates	74.0964	3.036	24.409	0.000	68.145	80.048
review_scores_location	107.7641	6.968	15.466	0.000	94.104	121.424
acc2	-2.3042	0.268	-8.598	0.000	-2.830	-1.779
n_amen2	5.945e-05	1.36e-05	4.380	0.000	3.28e-05	8.61e-05

These updated results show that the Adjusted R squared has doubled compared to the 'naive' regression, implying that we our additional features are capturing more complex dynamics. Additionally, the coefficient of the number of amenities is not statistically significant anymore. While its square term *is* statistically different from zero, its coefficient is positive and very close to zero.

We conclude that the number of amenities is little, if any, positively correlated with the price of a given listing.

3. Do the listings present clustering behaviour?

We study clustering behaviour as we hypothesize that there are two main dimensions across which listings could compete, namely *price* and *quality*. Indeed, we could expect that the most booked listings are either very affordable, or very high quality (in terms of location, amenities, etc.). If the listings are actually divided among these two classes, we should expect a certain degree of clustering. To assess clustering behaviour, we perform dimensionality reduction of our Amsterdam dataset through Principal Component Analysis (PCA).

To do so, we first drop the irrelevant numerical variables, such as 'id', 'host id', 'minimum nights', 'maximum nights', 'availability 365' and 'Amsterdam'. Then, we standardize all the variables, so that they present the same range of values, which is highly suggested before performing PCA.

Finally, we perform PCA and we focus on the first two Pricipal Components, which explain 31% and 15% of the total variance respectively. We report the most relevant loadings for each

principal component:

Principal Component 1

- review_scores_rating: 0.41
- review_scores_accuracy: 0.40
- review_scores_cleanliness: 0.35
- review_scores_checkin: 0.36
- review_scores_communication: 0.38
- review_scores_location: 0.30
- review_scores_value: 0.38

Principal Component 2

- accommodates: 0.56
- beds: 0.52
- price: 0.39

It can be observed that the first component is strongly associated with the reviews, while the second with the house dimensions and, therefore, price.
However, when we plot in Figure 10 the data points on the plane identified by the two principal components, we get the following:

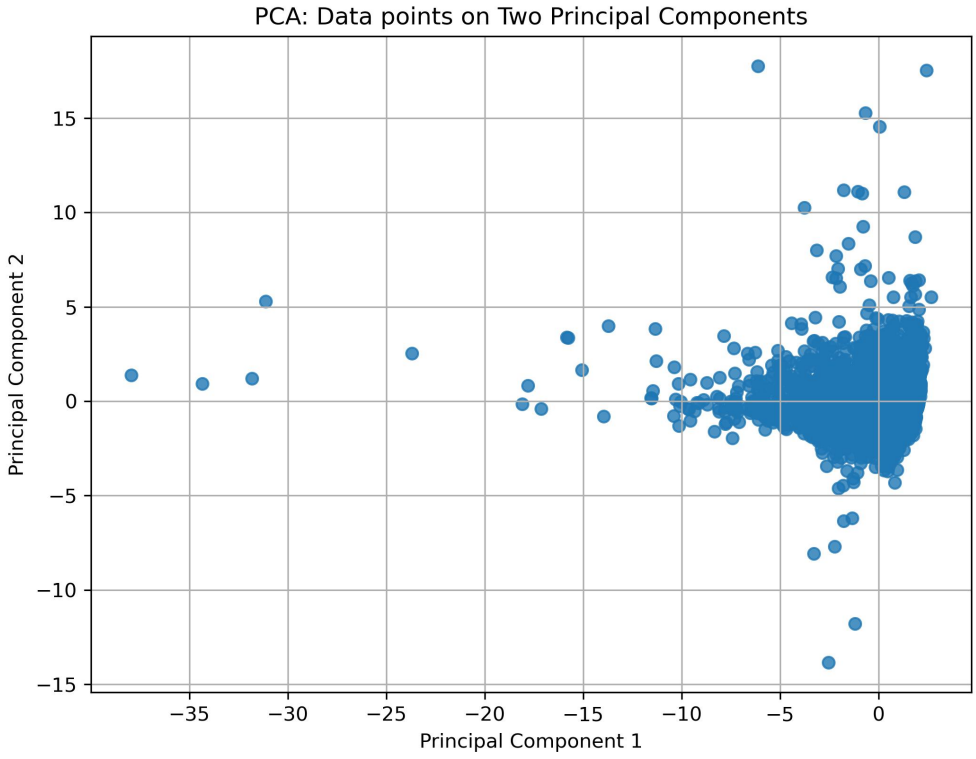


Figure 10: By plotting on the two principal components we cannot distinguish any clusters

The data points do not display any clustering, as they seem to be part of the very same cluster. This behaviour does not change when getting rid of the outliers on the left side of the graph. Thus, we conclude that there is no evident clustering behaviour.