How to become a Successful Airbnb Host

An Analysis based on Airbnb Data of Los Angeles between 2012 and 2016

Contents

1 Introduction

Airbnb is the most successful platform for private property renting. Members offer a place to sleep, a private room or even entire apartments to travelers who prefer to spend their nights in private environments instead of public hotels or hostels. The platform, therefore, provides the online environment where hosts and travelers meet and agree on terms and condition of the overnight stay. This means that Airbnb does not own properties but rather earns a small commission when a host and a traveler make an agreement on the platform. According to Bosa and Salinas (2019), Airbnb has hosted more than 400 million guests since its launch and reached a revenue of over 1$ billion in only the third quarter of the year 2018. The platform growths rapidly in terms of users and revenue. Airbnb's bookings increase by 45% year-after-year in the United States and even faster in other parts of the earth (ipropertymanagement.com, 2019). Today, Airbnb is available in over 190 countries and over 65,000 cities. With the increasing availability and a rising number of users, more and more people begin to recognize the opportunity to earn extra money on the side.

But higher popularity among landlords also leads to increased competition. To win more tenants and to become more successful on the platform, hosts try to attract more travelers and achieve higher review scores. But what makes a renter successful on Airbnb? To answer this question, our analysis takes advantages of an Airbnb dataset of over 34,000 listings that covers one of the most expensive cities in the world, Los Angeles. We argue being successful equals a high review score in perspective to the number of reviews per months. Thus, we derive a new review score as a result of combining the number of reviews and the average review score. Further, we explore all datasets features, examine its correlations and propose a model that aims to predict the new review score.

Therefore, we first introduce the concept of the Cross-Industry Standard Process for Data Mining (CRISP-DM) and conduct our analysis accordingly. The paper does not only uncover the high potentials of data analyses but also puts data into context and critically reflects on the use of data. It provides an answer to the research question:

How can the variables of an Airbnb dataset of the city Los Angeles, USA be used to form recommendations that increase the success of Airbnb hosts?

When conducting the analysis, we do not only learn from data exploration but also try to predict our derived review score by training a linear model. Consequently, we use the insights of our analysis to provide concrete recommendations for Airbnb hosts to become more successful. On this basis, we move from business understanding to data understanding, over data evaluation, potential use of the data insights, to a critical reflection of the outcomes of the data analysis. This investigation unveils the pitfalls of data analyses and stresses the importance of combining quantitative with qualitative aspect in the data analysis.

2 Methodology

In this chapter will we go through the methods we used throughout this report. Therefore, the process model, as well as the tools and techniques we utilize throughout this investigation, will be explained.

2.1 Process Model

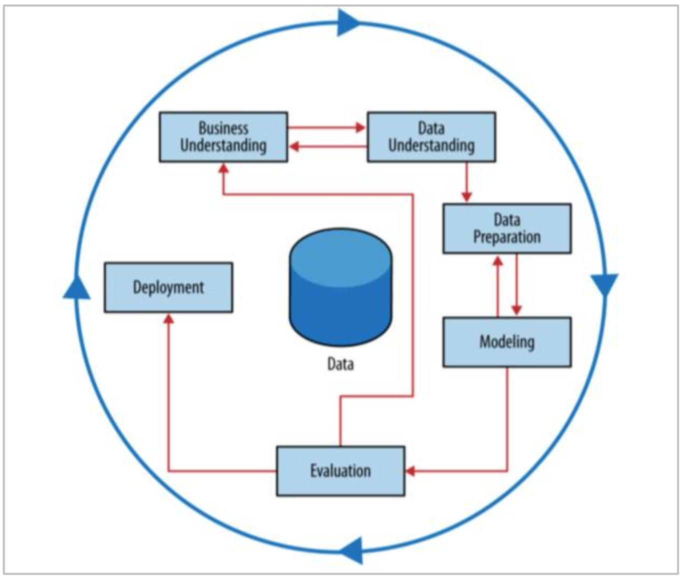


Figure 1: CRISP-DM Process Model (Provost and Fawcett, 2013, p.27)

For this report have we utilized the framework Cross-Industry Standard Process for Data Mining also known as CRISP-DM. This framework provides a set of guidelines that we try to follow and adjust to fit with our project. CRISP-DM breaks down the lifecycle of a Data Science project into six phases (Shearer, 2000, p.14). However, to make the model fit to this project, we modify the model and add a new phase called data exploration. Below, we will shortly introduce the seven phases and explain how we have made use of these in this project.

Business Understanding

The first phase of the CRISP-DM is called Business Understanding. According to Shearer (2000), this is the most crucial phase in the data science project as it provides the groundwork for later investigation. In this section, we focus on understanding the project objectives from the perspective of Airbnb and the hosts using it. We derive our problem definition from these insights which again, allow us to achieve the project objectives we identified. Moreover, this understanding helps us to assure that the dataset we used could help us achieve our goal to answer our research question.

Data Understanding

Once we have stated the research problem, we explore our available dataset. In this section, we look for strengths and weaknesses of the dataset to familiarize with the data and explore the dataset to gain insights into the data (Shearer, 2000). We examine the variables of our dataset to identify what we could use, and which will probably not affect our objectives. Moreover, we assess the quality of the data and ensure that we do not work with duplicate entries or missing values.

Data Preparation

After getting initial insights into our dataset and identifying the variables that we want to include in our analysis, we manipulate and convert the data into forms that seem to yield the best results. In this step, we remove missing values in our dataset and categorize data by converting to numeric values for instance (Provost & Fawcett, 2013, p.30). At the end of this phase, we generate a new dataset which only includes cleaned and relevant data that can be used for further investigation.

Data Exploration

The fourth phase is not a standard phase of the CRISP-DM model. Normally, one would explore the data in the data understanding phase. However, we believe it is more suitable to explore the data thoroughly after cleaning the data and thus we decide to add this phase. In the data exploration phase, we explore the final working dataset in more detail. This ensures that we have a good understanding of the underlying data and helps us to model the data.

Modeling

In the fifth phase of CRISP-DM modeling of the data is carried out. After exploring the final dataset, we move towards modeling. As we now have a good understanding of the data, we select an appropriate modeling technique for our data, build the model and assess the model accordingly. In this section, we move back and forth between data preparation and modeling to make sure the data complies with the technique chosen (Shearer, 2000).

Evaluation

In the sixth phase of CRISP-DM, the model is evaluated more thoroughly to determine whether the model is ready for the final deployment. The evaluation includes a review of the model and the result of the model to ensure that it achieves the business objectives that were set at the beginning of the project. At the end of this phase, it should be decided how the results should be used, and if the model is ready for deployment or if some of the business objectives have not been accomplished. If the model is not ready for final deployment, one has to go back and evaluate the business objectives (Shearer, 2000).

Deployment

In the final phase of CRISP-DM, the model is presented, and the project is reviewed. Here the findings are presented and a plan of how to use the findings is made. The outcome of the project can be as simple as a report, depending on the requirements for the project (Shearer, 2000). We will discuss how our findings and model can potentially be deployed in the future and how different stakeholders can benefit from it.

2.2 Tools and Techniques

We are working with a dataset that has a size of 125 MB and was found in the CSV format. First, we utilize Microsoft Excel 365 to gain an overview, clean as well as prepare the data. This tool allows us to exclude features, remove null values and convert categorical text values into numbers for instance. Second, data exploration of the dataset was conducted with the help of the programming language Python, the software Tableau as well as Microsoft Excel 365. These tools help us to recognize trends and key differences of groups by creating visualizations. Last but not least, we utilize python and its machine learning library Scit-Learn to train and test our model. We utilize python and make use of one of the algorithms when making predictions in our analysis. This should help us to make even more valuable recommendations about how a host can become even more successful on Airbnb.

3 Business Understanding

It is crucial to understand the business objectives as it determines the entire project. In this chapter, the project is defined, and the data is assessed to ensure that the business objective is achievable (Shearer, 2000).

3.1 Problem Definition

With the use of Airbnb, the trend of house sharing between peers has grown dramatically since Airbnb was founded (Nath, 2019). Airbnb does not own any lodging but is functioning as a third party that connects hosts and guests where they receive a percentage service fee. The business model for Airbnb differs from the traditional hotel business model, where hosts create their own listing and determine the price with a description of their listing. The guests are able to evaluate their stay by giving reviews and ratings. These reviews are used to calculate a user-generated average review which is public for everyone to see.

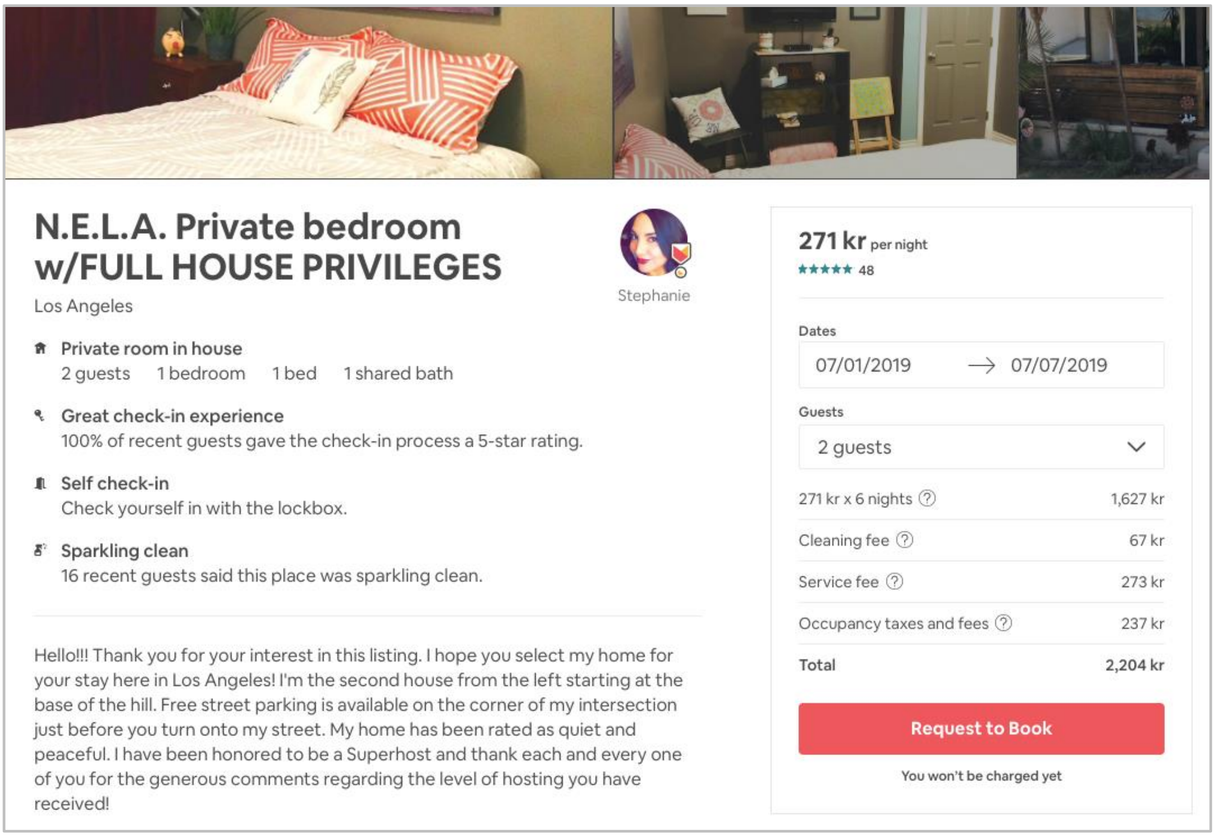
As with other platforms that are driven by user-generated content like IMDB and Uber, the average rating plays a huge role for the willingness for other customers to pay for that product or service (Müller et al., 2016, p. 295). Even though consumers relate to the peer-generated online reviews and putting trust into them, it is important to have in mind that reviews can be heavily biased depending on positive and negative events. However, according to (Müller et al., 2016) *“(...), the use of real reviews as a data source is common practice in the IS field, thereby acknowledging its validity and reliability for research purposes.”* (Müller et al., 2016, p. 295). This means that even if online reviews tend to be extreme in their ratings, still provides big data with rich information that can be used for studying customers information-seeking and decision-making behaviors. With over 2.7 million guests staying in Los Angeles and over 613 million-dollar earnings in total for the hosts in 2018, the hosts for Airbnb have the opportunity to generate high profits (Airbnb Citizen, 2019). In order to do so, the hosts need to attract customers by pleasing them. However, there are no defined guidelines for hosts to achieve the highest rating because customers’ needs vary a lot depending on their preferences and purpose for their stay. Some might prefer a place with a centralized location whereas others might prefer a place where pets are allowed.

On this basis, we define our business objective: We want to find out how the variables of our dataset affect the success of listings. Consequently, we aim to form concrete recommendations for hosts that can be applied to listings and therefore, affect the review score and the success of the hosts. Moreover, we aim to predict the degree of success and therefore, gain additional knowledge about how the variables affect the review score and the number of reviews each month. When hosts gain valuable information about the success of one's listing, the person might win more people and thus, earns more money on the side. Moreover, the platform Airbnb would benefit equally as the platform earns a small commission with each agreement that was made on the platform.

3.2 Process of rating Airbnb Listings

Since we want to investigate how to become a successful host on the platform of Airbnb we need to look into the process of a guest picking and renting a home until they give a review once the stay is completed. We will analyze the customer experience on Airbnb and identify what is visible for the user when renting a home and what might affect their decision.

The first step when entering Airbnb for a guest is to determine the city, the timespan they want to rent as well as how many guests are included. This information is used to list down all the available listings in that city as well as additional information, like popular activities and restaurants. Once the potential guest finds a listing he/she prefers, they enter a page with information about that listing.



N.E.L.A. Private bedroom w/FULL HOUSE PRIVILEGES

Los Angeles

Stephanie

Private room in house
2 guests 1 bedroom 1 bed 1 shared bath

Great check-in experience
100% of recent guests gave the check-in process a 5-star rating.

Self check-in
Check yourself in with the lockbox.

Sparkling clean
16 recent guests said this place was sparkling clean.

Hello!!! Thank you for your interest in this listing. I hope you select my home for your stay here in Los Angeles! I'm the second house from the left starting at the base of the hill. Free street parking is available on the corner of my intersection just before you turn onto my street. My home has been rated as quiet and peaceful. I have been honored to be a Superhost and thank each and every one of you for the generous comments regarding the level of hosting you have received!

271 kr per night
★★★★★ 48

Dates
07/01/2019 → 07/07/2019

Guests
2 guests

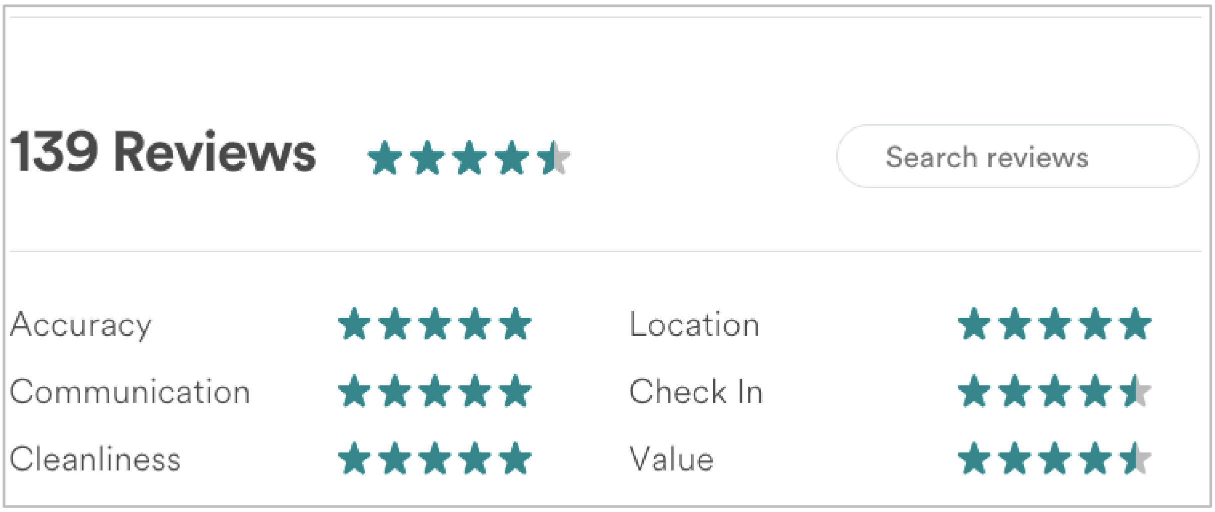
271 kr x 6 nights ⓘ	1,627 kr
Cleaning fee ⓘ	67 kr
Service fee ⓘ	273 kr
Occupancy taxes and fees ⓘ	237 kr
Total	2,204 kr

Screenshot 1: Example of an Airbnb listing

As shown on the screenshot above, the guest is presented by images of the listing and a personal description made by the host. A list of the included amenities associated with the place is shown to the guests know what is expected when they arrive. Further down on the page it is possible to see the overall average rating which is generated from 6 parameters. These parameters will be explained later in this section. It is also possible for the potential guest to see reviews from previous guests where they have expressed their experience with a few lines. At the bottom of the page, the user is introduced to the performance of the host where it is possible to see the response rate and response time for that host.

Guest Reviews:

After looking into the customer experience before renting a place, now let's look into how guests can evaluate and rate their experience.



139 Reviews ★★★★½

Search reviews

Accuracy	★★★★★	Location	★★★★★
Communication	★★★★★	Check In	★★★★½
Cleanliness	★★★★★	Value	★★★★½

Screenshot 2: Airbnb Reviews

As mentioned earlier, each review consists of 6 parameters, which are used to calculate an overall rating for that review. Each parameter is evaluated from 1 star to 5 and they are concerning (Airbnb, 2019):

- Accuracy - How accurate the post on Airbnb is corresponding to what the guests actually get.
- Communication - How well the host is communicating with the guests before and during their stay.
- Cleanliness - How clean the place is.
- Location - How well the place is located in the city and its surroundings.
- Check-in - How effortless was the check-in for the guest.
- Value - How well did the guest feel the listing provided good value for the given price

From this, we have learned that if you want to become a successful host, you need to perform well on all of these reviews. Most of the parameters are highly dependable on the host itself where a parameter like location is depending on the geographical position of the house and its surroundings.

4 Data Understanding

In this chapter will we go through the process of understanding the Airbnb dataset. We have also conducted a quality check of the dataset to determine if it can be used to achieve the business objectives we set out to resolve.

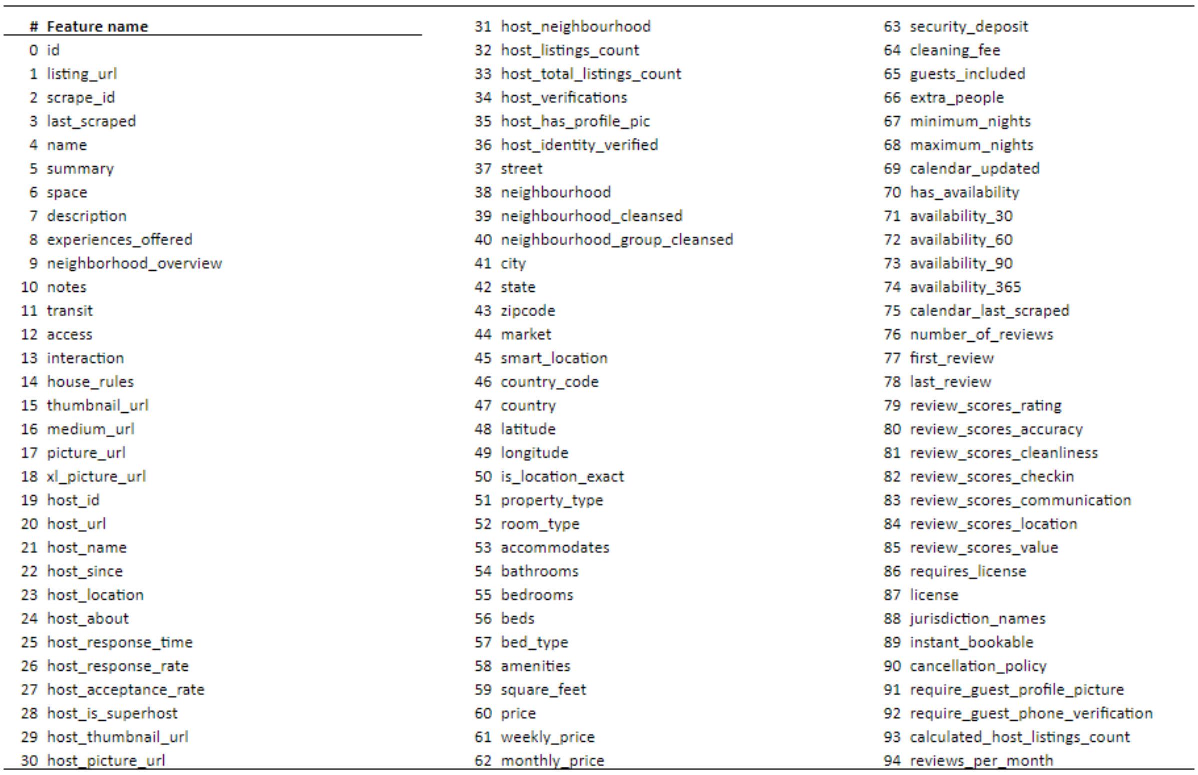
4.1 Data Presentation

In this project, we are working with a dataset of Airbnb listings from 2012 until 2016 in Los Angeles, the USA provided by Inside Airbnb. Inside Airbnb is an independent, non-commercial website that provides publicly available datasets of Airbnb data to allow exploration of the use of Airbnb in different cities around the world (Inside Airbnb, 2019).

In our case, we have chosen to only work with data from Los Angeles as we know Airbnb is widely used in the US and specifically in LA, one of the most expensive cities to live in. The fact that here presented data is publicly available says a lot about the current time of Big Data and its underlying trends. With an increasing number of connected devices and internet services, data is increasingly available in our everyday life. Especially when more and more services are digitized, new data is created, and more data can be accessed online. This gives organizations and people more access to information, but also raises privacy concerns which we will discuss at a later point of this report.

We found the data of Airbnb listings in a CSV file. CSV or Comma-Separated Value (CSV) files are delimited text files that separate values by commas. Each line of the text file represents a data record and each record is presented with the same sequence of fields. This means that CSV file will represent each data record the same and does not skip commas in case data is missing for instance. Thus, the content between each comma can be understood at cells and it becomes clear that this format is very useful for table data. Our Airbnb dataset contains 31,253 records and 94 features. This gives a total of 127 MB of raw data to work with.

Table 1 provides an overview of all the features that we have found in the described dataset.



#	Feature name	31	host_neighbourhood	63	security_deposit
0	id	32	host_listings_count	64	cleaning_fee
1	listing_url	33	host_total_listings_count	65	guests_included
2	scrape_id	34	host_verifications	66	extra_people
3	last_scraped	35	host_has_profile_pic	67	minimum_nights
4	name	36	host_identity_verified	68	maximum_nights
5	summary	37	street	69	calendar_updated
6	space	38	neighbourhood	70	has_availability
7	description	39	neighbourhood_cleansed	71	availability_30
8	experiences_offered	40	neighbourhood_group_cleansed	72	availability_60
9	neighborhood_overview	41	city	73	availability_90
10	notes	42	state	74	availability_365
11	transit	43	zipcode	75	calendar_last_scraped
12	access	44	market	76	number_of_reviews
13	interaction	45	smart_location	77	first_review
14	house_rules	46	country_code	78	last_review
15	thumbnail_url	47	country	79	review_scores_rating
16	medium_url	48	latitude	80	review_scores_accuracy
17	picture_url	49	longitude	81	review_scores_cleanliness
18	xl_picture_url	50	is_location_exact	82	review_scores_checkin
19	host_id	51	property_type	83	review_scores_communication
20	host_url	52	room_type	84	review_scores_location
21	host_name	53	accommodates	85	review_scores_value
22	host_since	54	bathrooms	86	requires_license
23	host_location	55	bedrooms	87	license
24	host_about	56	beds	88	jurisdiction_names
25	host_response_time	57	bed_type	89	instant_bookable
26	host_response_rate	58	amenities	90	cancellation_policy
27	host_acceptance_rate	59	square_feet	91	require_guest_profile_picture
28	host_is_superhost	60	price	92	require_guest_phone_verification
29	host_thumbnail_url	61	weekly_price	93	calculated_host_listings_count
30	host_picture_url	62	monthly_price	94	reviews_per_month

Table 1: Overview of all Features

4.2 Data Quality

As a part of understanding the dataset we are working with, it is also important to assess the quality of the data. That is, to check for missing attributes, blank fields, whether all possible values are represented (Shearer, 2000). All of this is done to ensure that the dataset can actually be used to obtain the business goal and to determine how much time needs to be spent on cleaning the data or collecting more data.

As mentioned, we have a fairly large dataset with 94 different features. These features consist of a mix between quantitative and qualitative data, i.e. both numeric and string values. Having a large dataset with many features included is really useful. However, many of these features are not relevant for our analysis and should thus be removed.

Furthermore, the dataset contains a lot of null values or missing values. Missing values and null values cannot be used for the purpose of our analysis and thus, we need to remove all entries that do not contain any data. Our dataset contains missing values or null values, and this significantly reduces the size of our dataset and thus, decreases the information of the dataset.

Having string values in our dataset means that we need to reformat the data into numeric values to enable for use in our regression. This is something we need to consider as this might be time-consuming.

The definition of some of the features included in the dataset is vague. This makes it hard to interpret the data and makes it difficult to include such a feature in our analysis. For example, the price feature is not well defined which means that we do not whether the price is included for one day of rental, or for multiple days of rental. However, after investigating the price table more thoroughly we assume that the price table is containing the price for a one-night renting.

After assessing the dataset, it is our belief that the quality of the dataset is sufficient to achieve the business goal.

5 Data Preparation

Besides the 94 features that are included in the raw dataset, we added one feature ourselves. As stated in our problem definition, we aim to discover valuable insights on how to become a successful host on Airbnb. To do this, we need to determine a measure of success on Airbnb. Generally, we believe that Airbnb hosts are successful when they achieve high review scores. However, this review score might not be very meaningful as a host who only received one review, can easily achieve a review score of 100%. This can be very misleading as the one review might come from a friend of the host. Therefore, we have decided to include the reviews per months into the measure of success. Thus, we have decided to put the review score into perspective. We have added the feature new_review_score which is a multiplication of the review_score and the numbers of reviews per month. For simplicity, this number is then divided by 10. If the new_review_score is high, it means that the listing is not only booked often but also generally receives high review scores.

$$\text{new\_review\_score} = \frac{Review\ Score\ x\ Reviews\ per\ Months}{10}$$

To further be able to use the data for quantitative analysis, we had to prepare and clean the dataset. First, we discussed the importance of each of the remaining features and examined whether each of the features contains relevant and plausible data. Throughout the entire project, we closely examined which features are necessary to conduct our analysis. After continuous investigation, we could confidentially remove even more features that raise privacy concerns and were not needed for our analysis. As we will later explain in our report, we strive to anonymous our dataset as much as possible and exclude all features which are not utilized. Altogether, we have deleted 60 features, which we concluded to be irrelevant contained implausible, mainly null variables or were not needed in the analysis. Thus, we ended up with 35 relevant features for our analysis.

Table 6 in Appendix A shows an overview of all features. The features that we agreed on using for our analysis are marked green. When deleting the unused features, we end up with Table 7 in Appendix A that contains 34 features. However, many features contained data that were stored in a format which are difficult to process by a machine and therefore, cannot be used for data analysis.

This is why we conducted a number of data preparation tasks which are explained in the following:

List in Cells

The cells of the feature “amenities” contained a list, such as {Internet, Kitchen, Free parking on premises, (...)}. We have realized that a list of items within a feature/ column could not be used for further data analysis. Therefore, we created new features with the names of the list items and inserted a dummy variable 1 if the Airbnb record had the name of the list item included and 0 if the record did not include the list item in the list. To do so, we wrote a short equation (=COUNT(IF($<cell_name>,{“*<name of list item>*”})), 1, 0) in excel and run the equation for each created new feature and all records. The feature amenities got deleted after this procedure.

Signs

All numbers of the dataset have been transformed to exclude any signs, such as $ or %.

Different length of numbers

The numbers varied in terms of length. That is why we have limited the digits behind the comma to two.

Textual Data

Two features were transformed and scaled in order to incorporate the content in regression and correlation analysis:

- The feature ‘host_response_time’ consisted of four attributes: “A few days or more”, “within a day”, “within a few hours”, “within an hour”. Consequently, each of the attributes got assigned to a value: “A few days or more” = 1, “within a day” = 2, “within a few hours” = 3, “within an hour” =4. So, the greater the number, the faster is the host responding.
- The feature cancellation policy consisted of four attributes which we again scaled with relevant values: “Flexible” = 1, “Moderate” = 2, “strict” = 3 and “very strict” = 4. The greater the number, the stricter the host in regard to the cancellation policy.

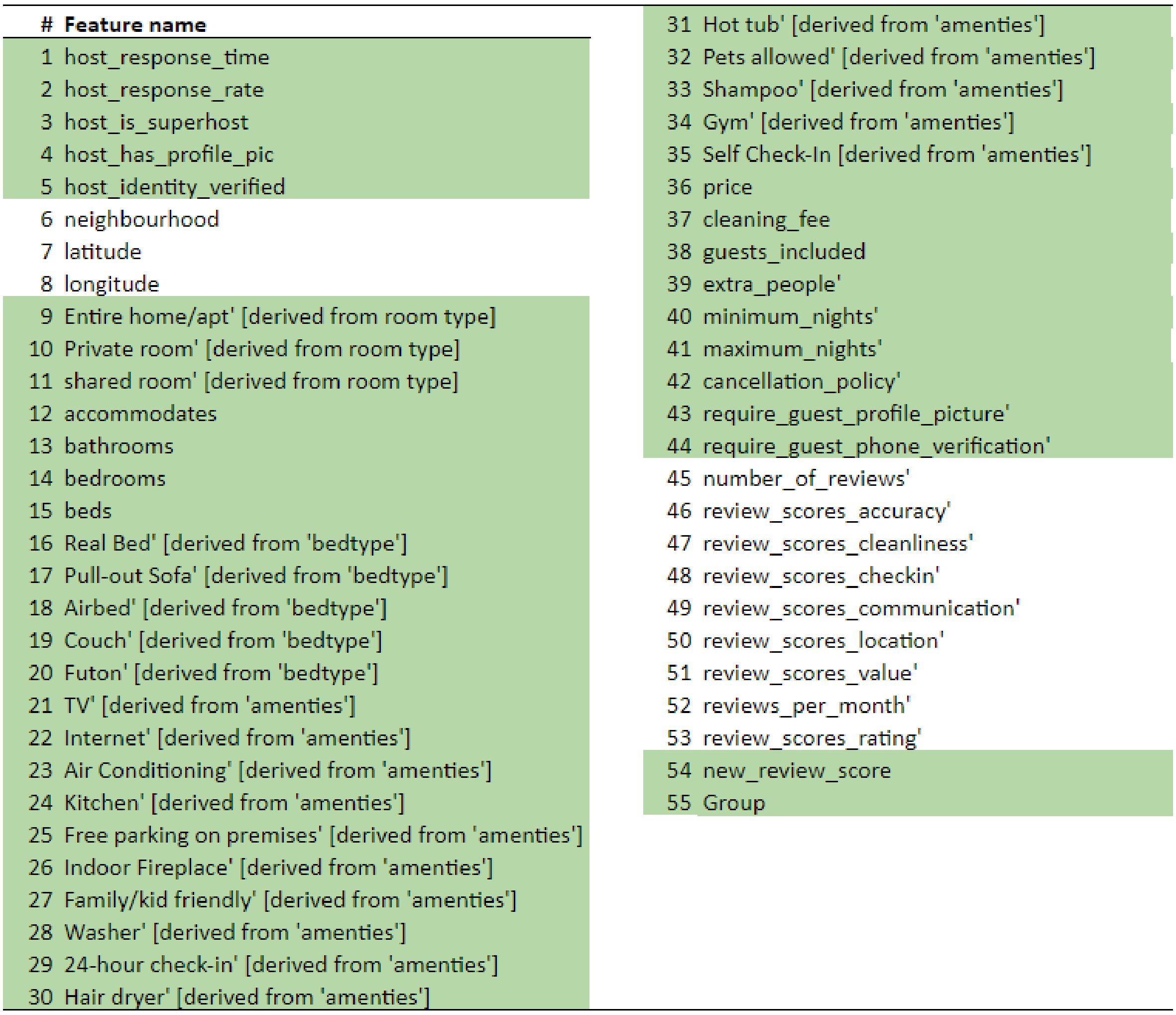
Binary Data as letters:

The original dataset includes many features that are described as binary. ‘host_is_superhost’, ‘host_has_profile_pic’, ‘host_identity_verified’, ‘requires_license’, ‘require_guest_phone_verification’ make use of “t” (true) and “f” (false). As we cannot use letters to build a regression model and to analyze correlation, we turned all “t” letters to 1 and all “f” letters to the value 0.

Empty cells:

We made sure that none of the cells remain empty. First, we decided to insert “NA” for all cells that should have contained textual data. Second, we deleted all data records that included empty cells at the features ‘host_response_time’, ‘host_response_rate’, or at any review score. Third, we added a 0 for the empty cells at the feature ‘cleaning fee’ as we assume that an empty cell means that that the host does not charge a cleaning fee. This way we made sure that we can work with a dataset that does not include any empty cells.

The following Table 2 shows the final working dataset that was used for data exploration. But not all features seemed to be useful for regression and correlation analysis as many features contain textual data. Moreover, review features needed to be excluded for correlation and regression analysis as they directly determine the review score. They are basically subsets of the final “new review score”. However, we decided to keep the features in our final dataset as they are important features that needed to be analyzed in the exploration phase. Again, we have marked the features which we recognized useful for correlation and regression analysis green.



#	Feature name
1	host_response_time
2	host_response_rate
3	host_is_superhost
4	host_has_profile_pic
5	host_identity_verified
6	neighbourhood
7	latitude
8	longitude
9	Entire home/apt' [derived from room type]
10	Private room' [derived from room type]
11	shared room' [derived from room type]
12	accommodates
13	bathrooms
14	bedrooms
15	beds
16	Real Bed' [derived from 'bedtype']
17	Pull-out Sofa' [derived from 'bedtype']
18	Airbed' [derived from 'bedtype']
19	Couch' [derived from 'bedtype']
20	Futon' [derived from 'bedtype']
21	TV' [derived from 'amenties']
22	Internet' [derived from 'amenties']
23	Air Conditioning' [derived from 'amenties']
24	Kitchen' [derived from 'amenties']
25	Free parking on premises' [derived from 'amenties']
26	Indoor Fireplace' [derived from 'amenties']
27	Family/kid friendly' [derived from 'amenties']
28	Washer' [derived from 'amenties']
29	24-hour check-in' [derived from 'amenties']
30	Hair dryer' [derived from 'amenties']
31	Hot tub' [derived from 'amenties']
32	Pets allowed' [derived from 'amenties']
33	Shampoo' [derived from 'amenties']
34	Gym' [derived from 'amenties']
35	Self Check-In [derived from 'amenties']
36	price
37	cleaning_fee
38	guests_included
39	extra_people'
40	minimum_nights'
41	maximum_nights'
42	cancellation_policy'
43	require_guest_profile_picture'
44	require_guest_phone_verification'
45	number_of_reviews'
46	review_scores_accuracy'
47	review_scores_cleanliness'
48	review_scores_checkin'
49	review_scores_communication'
50	review_scores_location'
51	review_scores_value'
52	reviews_per_month'
53	review_scores_rating'
54	new_review_score
55	Group

Table 2: Final Working Table

6 Data Exploration

First, we started to closely examine the feature ‘new_review_score’ (NRS) as we decided to use this score as a measure of success.

	NRS
Count	21217
Mean	1.95
Std	1.88
Min	0.01
25%	0.54
50%	1.31
75%	2.68
Max	20.68

Table 3: Statistical Details of NRS

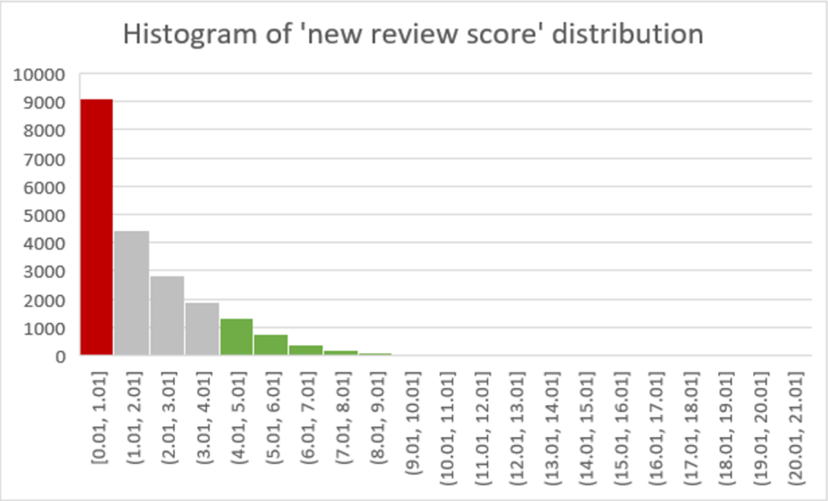


Figure 2: Histogram of NRS

As mentioned earlier, we are working with 21,217 data records. As table 3 shows, the listings achieved scores between 0.01 and 20.68 with a mean of 1.95 and a standard deviation of 1.88. The histogram in Figure 2, as well as the percentiles in table 3, indicate that the scores are not distributed equally. Many listings do not achieve a high NRS and only a few listings achieve a top score. In order to discover what determines a high NRS, we have decided to divide the listings into three groups. We denote the first group as “Low Performers” and define the group as listings which have achieved an NRS between 0 and 1.00. The second group is called “Medium Performers” and is defined as listings of above 1.00 and up until and including 4.00. All listings that have achieved more than 4 are part of the group “High performers”. Table 4 shows an overview of the numbers of listings as well as the sum of NRS each group has achieved. Moreover, we marked the three groups in the earlier posted histogram with the three colors red, grey and green.

Group	Number of listings	Sum of Score
Low Performers	9062	4402
Medium Performers	9232	20373
High Performers	2923	16603

Table 4: Group Comparison

On this basis, we examine the NRS of each of the group further and create a boxplot to visualize how NRS is distributed in each group.

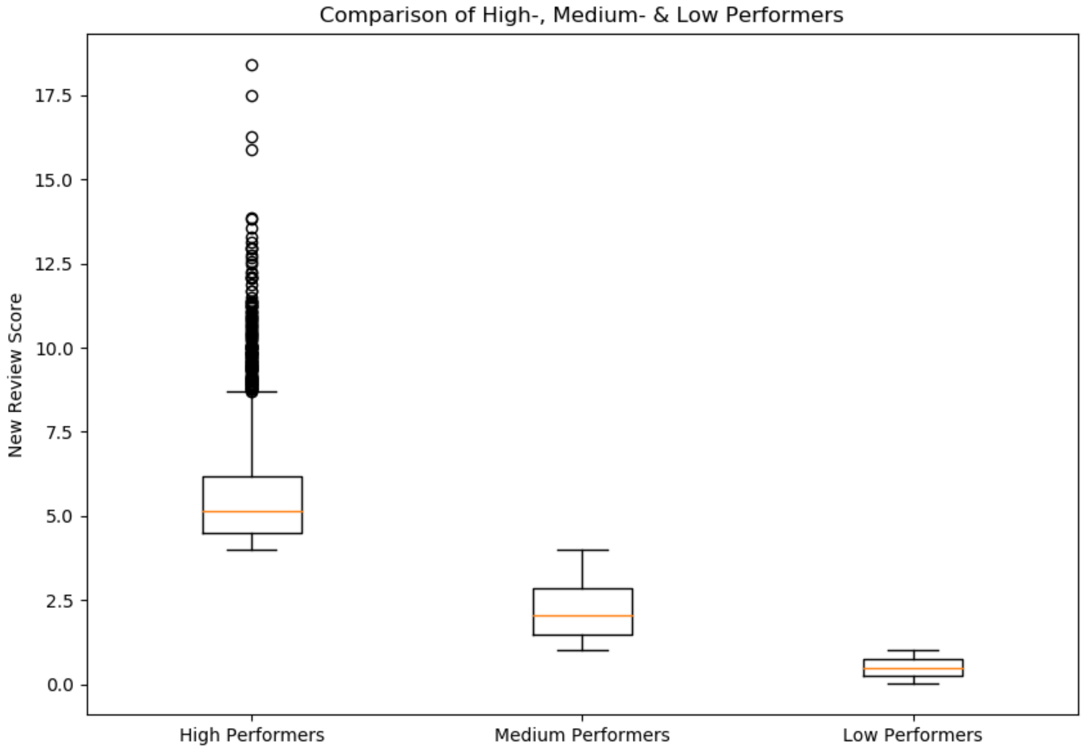


Figure 3: Boxplots of High-, Medium- & Low Performers

As the histogram already indicated, the high performers have a quite wide range of NRS with many outliers. However, we accept this possibly distorting factor and move on with the investigation. In the following, we want to examine how the low performers and the high performers differ in each of the features that remain in our final dataset. We have assessed all of the features but only visualize and describe relevant features in the following.

6.1 Low Performers vs. High Performers

After grouping the new review score (NRS), we now want to explore the data by comparing the high performers with the low performers to gain insights to which features are performing well and which features that seems to be insignificant for the NRS. Therefore, we focus on features that especially show strong differences between low and high performers. This way, we believe to point out key differences and hope to recognize what high performers do differently.

6.1.1 Host Response Time

Figure 4 shows the distribution of host response time based on the performance. The comparison of the average host response time shows that in general, hosts having a faster response time tend to perform better than hosts having a slow response time. Interestingly, the majority of both low and high performers respond within an hour. However, it is clear that the performance significantly drops if the response time is above one hour. Furthermore, it is seen that the top performers always respond within a day at most. This comparison suggests that the hosts should always respond within an hour to obtain the best performance.

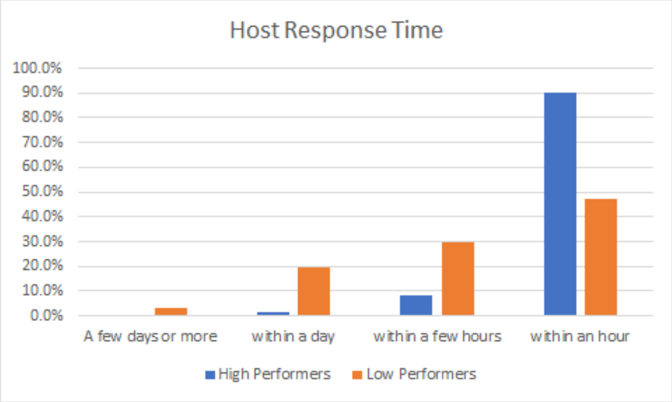


Figure 4: Host Response Time

6.1.2 Superhost

Being a superhost on Airbnb is a sign of recognition. This is displayed by a little badge showed in the host profile and on the listing. Airbnb claims that superhosts performs better and increases revenue potential significantly (Airbnb, 2019). The host needs to fulfill specific requirements to get the status as a superhost. For example, they need a response rate of 90%, an average review score of 4,8 or above, and at least 10 previous rentals.

As seen in Figure 5, it seems that being a superhost does increase the performance of a host. Approximately 49% of high performers are superhosts, whereas only approximately 17% of low performers are superhosts.

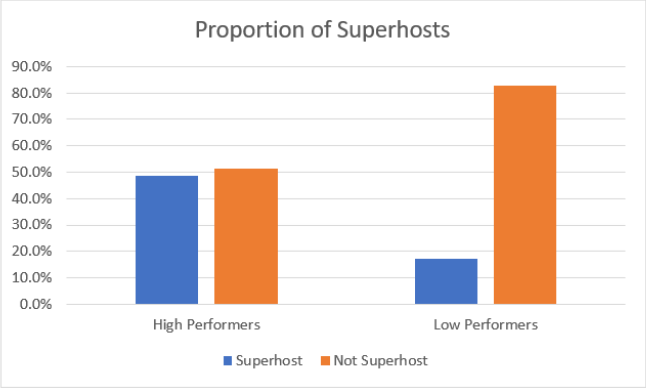


Figure 5: Superhosts

6.1.3 House Statistics

Figure 6 shows the performance of each room type provided in the dataset. We expect the room type to have some effect on the performance, however, the comparison is really interesting as it seems that there is a major difference in the performance of the room types being offered. Renting an entire home have resulted in high performance in roughly 63% of the time and only resulted in low performance in 4% of the time. Private rooms also seem to perform well. However, renting a shared room has only resulted in high performance in 3% of the times. One could expect that the shared room would perform worse than a private room, however, the difference in performance between renting a shared room and entire home is highly significant. This implies that to get the best performance on Airbnb, one should offer an entire home or at least a private room.

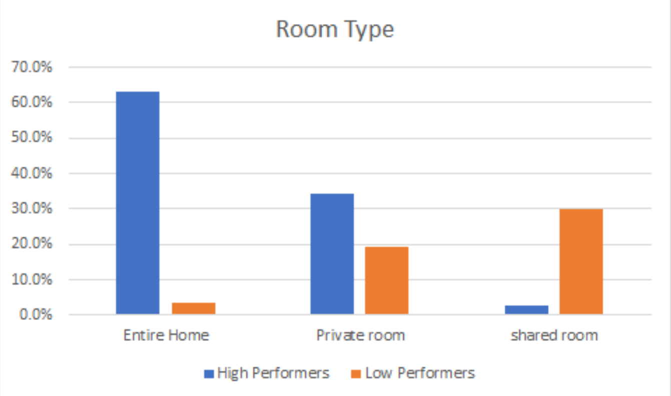


Figure 6: House Statistics

6.1.4 Check-In

Figure 7 shows the performance of the different check-in opportunities for an Airbnb listing. From the comparison, it is seen that offering the opportunity to have 24-hour self-check-in available is performing well.

Allowing self-check-in results in high performance in 30% of the times and only results in a low performance 9% of the times. Furthermore, providing 24-hour check-in also performs well in 48% of the cases. This implies that offering 24-hour self-check-in results in high performance in the majority of the cases and thus, should be a feature to include to become successful on Airbnb.

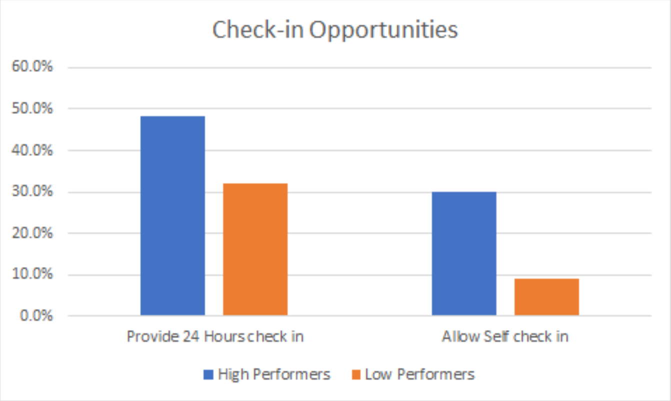


Figure 7: Check-In Opportunities

6.1.5 Amenities

Another feature that is interesting to compare in terms of performance are the amenities. This comparison shows which amenities are required to perform well on Airbnb. Figure 8 shows the performance comparison of the different amenities. It is seen that offering a hot tub, allowing pets, or providing a gym is not significantly affecting the performance. Interestingly, providing a washer is resulting in low performance in a

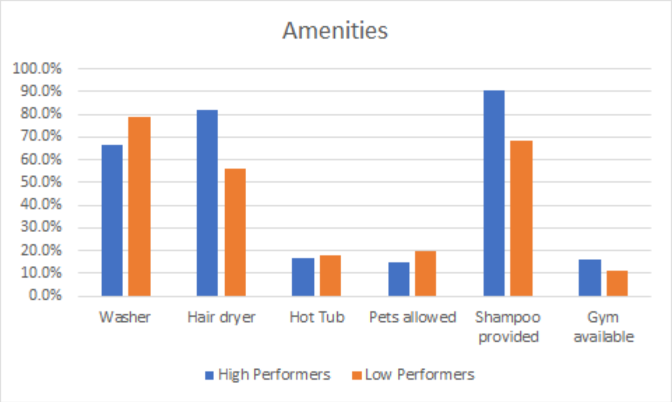


Figure 8: Amenities

small majority of the cases. This might make sense since people using Airbnb are usually not staying for a longer period of time and thus, will not be in need of a washer. On the other hand, providing a hair dryer and shampoo seems to perform really well in the majority of cases. This also makes sense since these are features one needs to bring when traveling. However, by offering these features already, the host removes the need for the users to bring these items themselves. This tells us that providing features that are necessary for travelers to bring seem to result in high performance.

6.1.6 Minimum Nights

The minimum amount of nights users have to rent an Airbnb listing is also interesting to investigate. Figure 9 shows the comparison of the average minimum nights in terms of performance. It is seen that the low performers on average offer their renting for a minimum of 4 days. On the other hand, the top performers on average offer their renting for a minimum of 1.5 days. From this comparison, it seems that having a low minimum amount of nights required to rent an Airbnb listing is performing really well. This also makes sense as having a requirement of the number of nights a customer have to rent a place can exclude potential customers. Having this requirement forces users to stay for a longer period of time, which might be profitable for the hosts but causes issues for users who are on short trips. Ultimately, this comparison suggests that having fewer minimum nights required significantly improves the new review score.

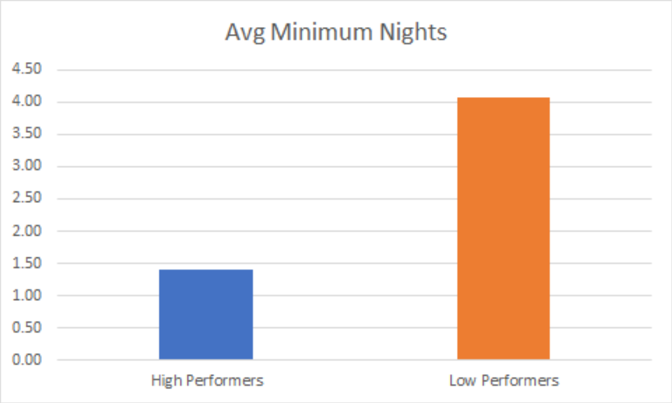


Figure 9: Average Minimum Nights

6.1.7 Price and Cleaning Fee

As mentioned the price feature is not well described in the dataset. However, we believe it is interesting to investigate how the price and cleaning fee affects performance. Figure 10 shows the comparison of the average price and the average cleaning fee in terms of performance. The comparison shows that having a lower price seems to perform better than having a high price, which makes sense as people usually want to save as much money as possible. The same tendency is shown when comparing the average cleaning fee in terms of performance. This shows that the top performers tend to have a low cleaning fee. We believe the same argument of people wanting to save as much money as possible also hold true for this feature. However, we also assume that allowing the customer to clean the place themselves and thereby eliminate the cleaning fee altogether will improve the performance.

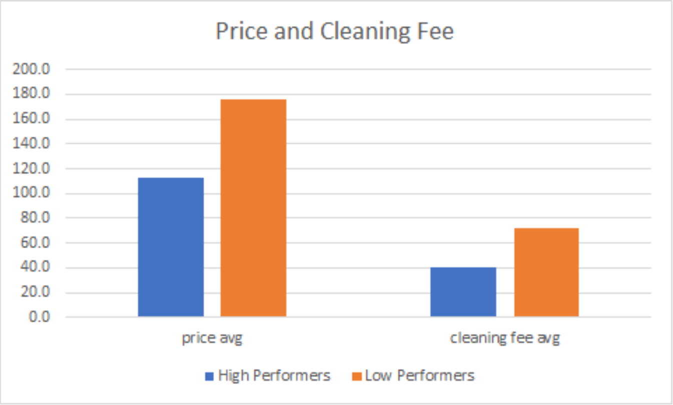


Figure 10: Price and Cleaning Fee

From the exploration of the different features it seems that in order to achieve the best performance on Airbnb, one should include the following features: A response time below one hour, be a superhost, rent out an entire home, allow for 24 hour self-check-in, provide fundamental amenities needed when traveling, have a low amount of minimum nights required, no cleaning fee, and finally having a cheap price for the renting.

6.2 Mapping Review Score vs. Price

In this section, we will explore if the price and neighborhood seem to have a significant effect on the NRS.

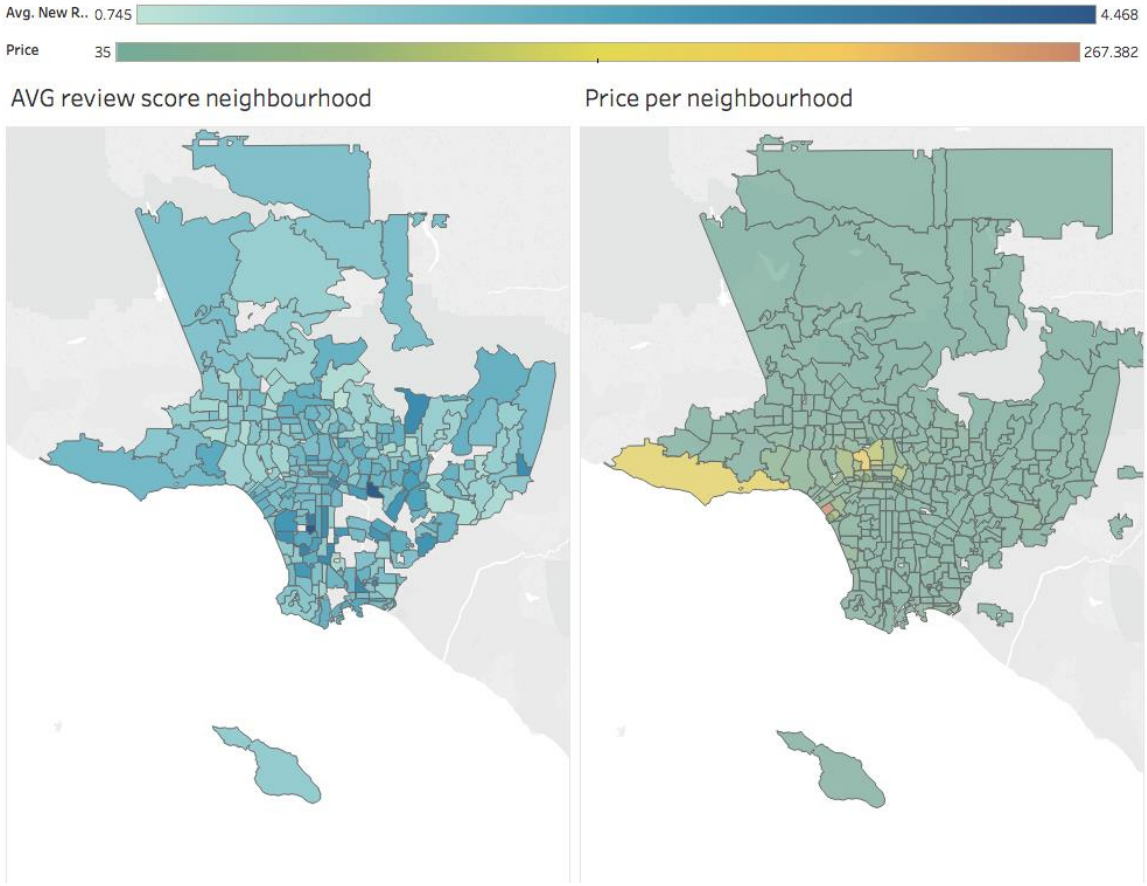


Figure 11: Average Review Score and Price per Neighborhood

Figure 11 shows a comparison of the neighborhoods based on price and the NRS. Initially, we assumed that the price would have a significant effect on the NRS. However, when inspecting the results of the mapping, become clear that the price does not seem to have an effect on the review score. If the price would have a significant effect on the NRS, we assume that the neighborhoods with a high average price would also score the highest average NRS. However, this does not seem to be the case. For example, Downtown Los Angeles seems to be one of the most expensive neighborhoods, but this neighborhood only achieved a medium to low average NRS. Furthermore, the low-priced neighborhoods, e.g. Lynwood seems to be achieving a high average NRS. Thus, it seems that the price does not have an effect on the NRS. This might make sense since the price is a factor that is important when the users decide which Airbnb listing to choose, but it might not be considered when the user has to review the overall experience of the stay.

7 Modeling

As part of our data exploration, we have learned a lot about our Airbnb dataset and even derived very valuable insights. But we also aim to discover whether the features can be used to predict the new review score (NRS). If we would manage to predict the NRS, hosts could use the model to foresee their success on the platform. It could help each of the hosts a lot as it could potentially increase their income and could show the effect of changing one of the variables. In data science, there are many ways to determine which algorithm to choose when predicting a review score. However, as we are limited by time and computing resources, we have decided to focus on linear regression that minimizes the sum of square error. We acknowledge that the early decision of the model choice is not a recommendable procedure as a prediction exercise can be solved by a variety of machine learning algorithms and this linear regression model might not be suitable for our data.

To conduct our prediction, we go through a number of steps. First, we analyze the correlation of the independent variables to identify variables with a very high correlation that are more linearly dependent and therefore, have almost the same effect on NRS. Second, we conduct an Ordinary Least Squares (OLS) Regression to identify significant features. Third, we examine the relationship of significant features with the highest correlation towards NRS by drawing scatter plots. Then, we split our dataset into train and test data and train our linear regression model with the selected significant features on the training dataset solely. Consequently, we evaluate our linear model by testing it on our test dataset and examine the viability of our model.

7.1 Choosing and Evaluating Independent Variables

As described earlier, we first try to reduce dimensionality by analyzing the correlation of the independent variables. When two variables have a very high correlation of above 0.95, we drop one of the variables as we consider one of the features as redundant and believe that it does not convey extra beneficial information. Figure 12 visualizes the correlation of the independent variables to each other. To improve visibility, we exclude the correlation variables in the figure and show the degree of correlation by colors.

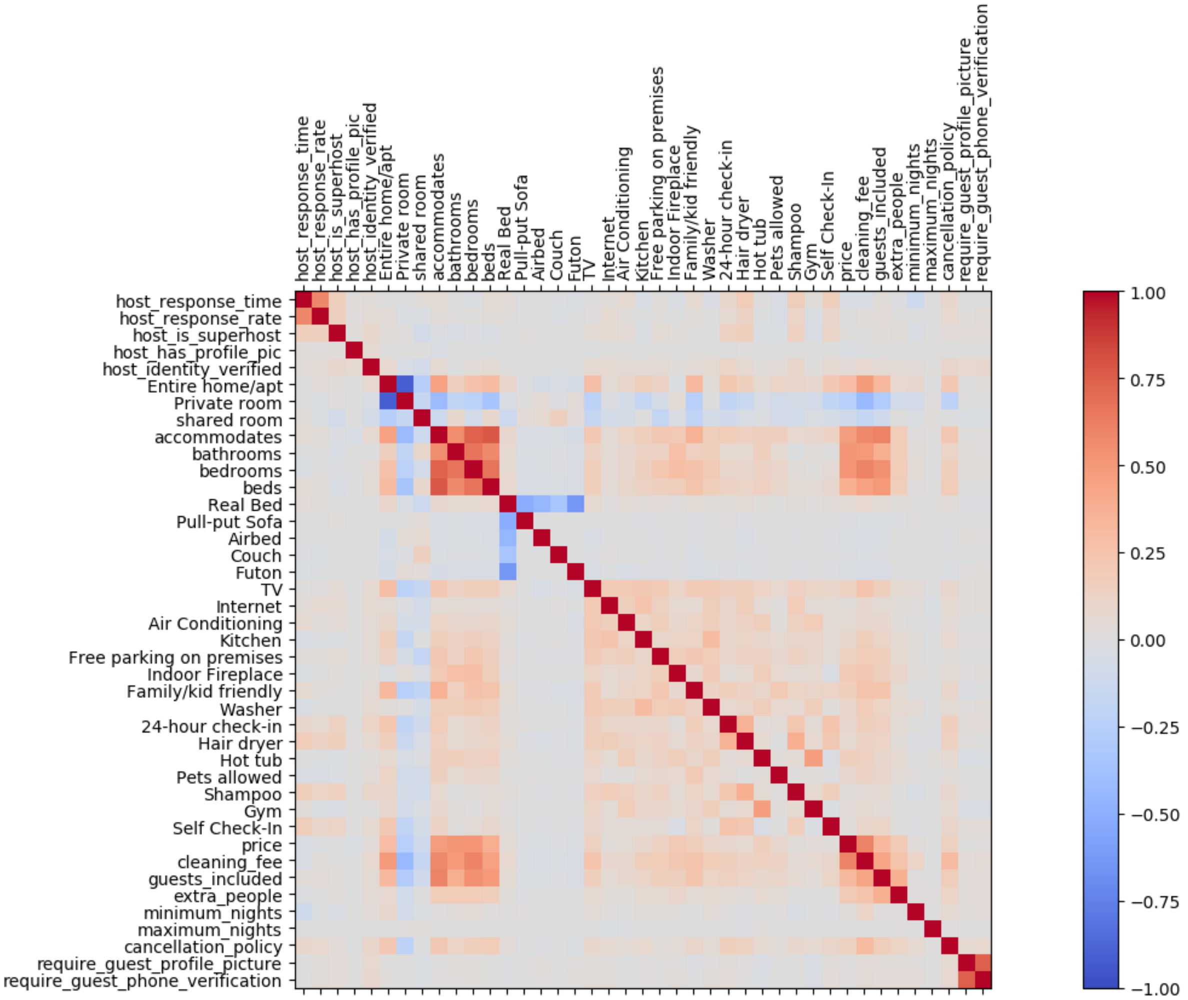


Figure 12: Pearson Correlation of Independent Variables

As it turns out, none of the independent variables need to be excluded as none reach a correlation of at least 0.95.

On this basis, we conduct an OLS Regression and assess the p-values of our features. Therefore, the p-value tests the null hypothesis that the coefficient has no effect on the dependent variable. We choose a significance level of 5% and hence, reject the null hypothesis when the p-value is smaller or equal to 0.05. When doing so, we exclude 16 features and remain with 25 variables. Then, we pay special attention to the coefficient of the independent variables to determine how the variables correlate with the dependent variables. Figure 13 provides an overview of the features and its coefficients.

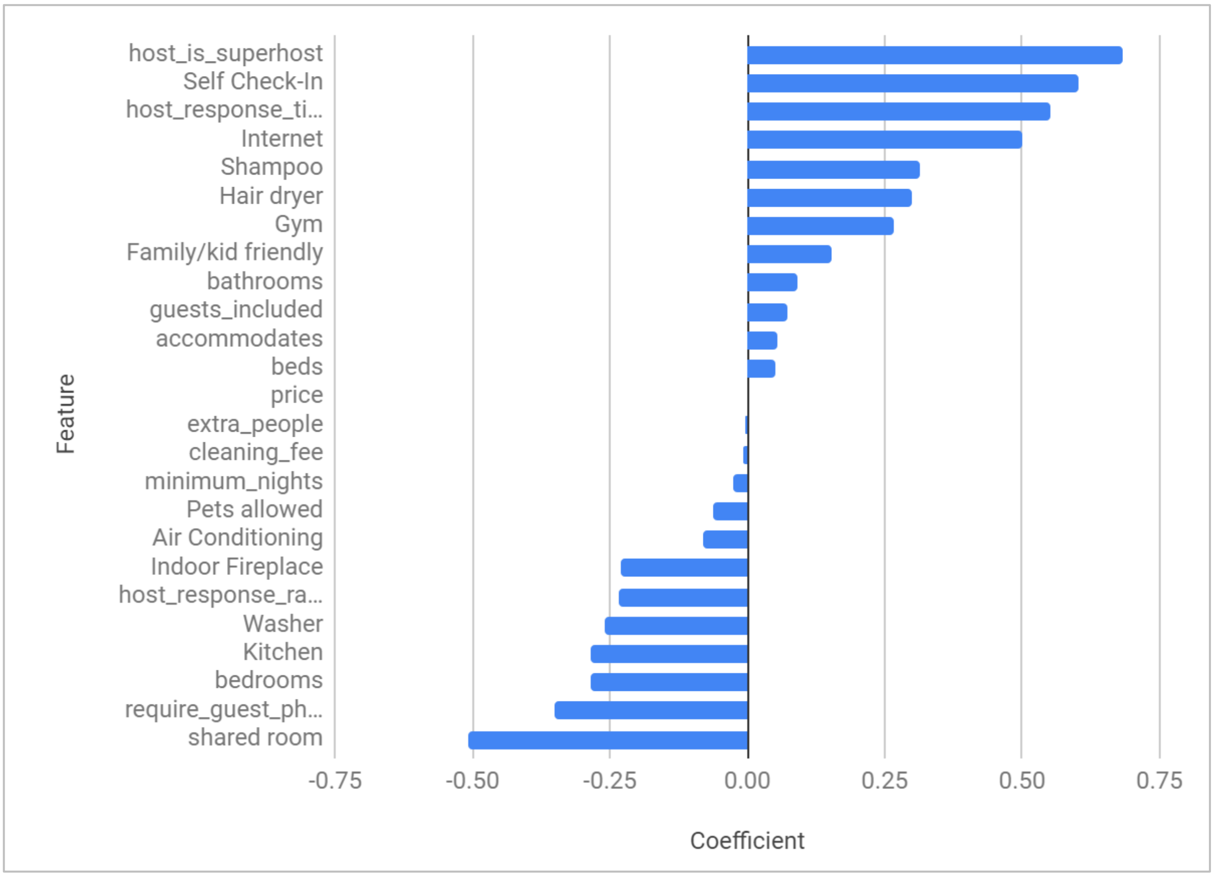
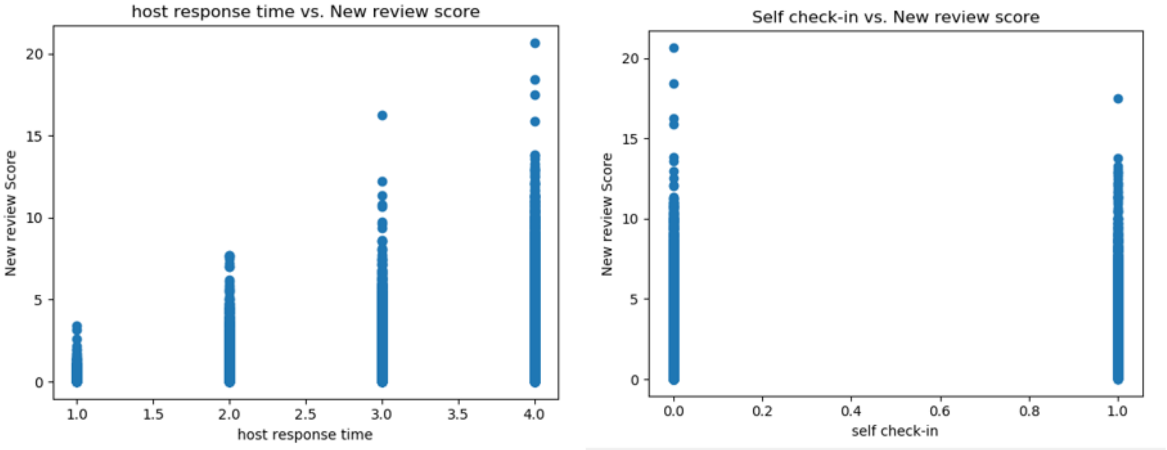


Figure 13: Regression Coefficients

Figure 13 shows how each of the independent variables influences the dependent variable NRS. On this basis, we have decided to focus on the features with the greatest coefficient. Therefore, we visualize the correlation of each independent variable with a coefficient of at least (+/-) 0.5. This way, we can further investigate the relationship and provide insights into how a feature influences NRS.



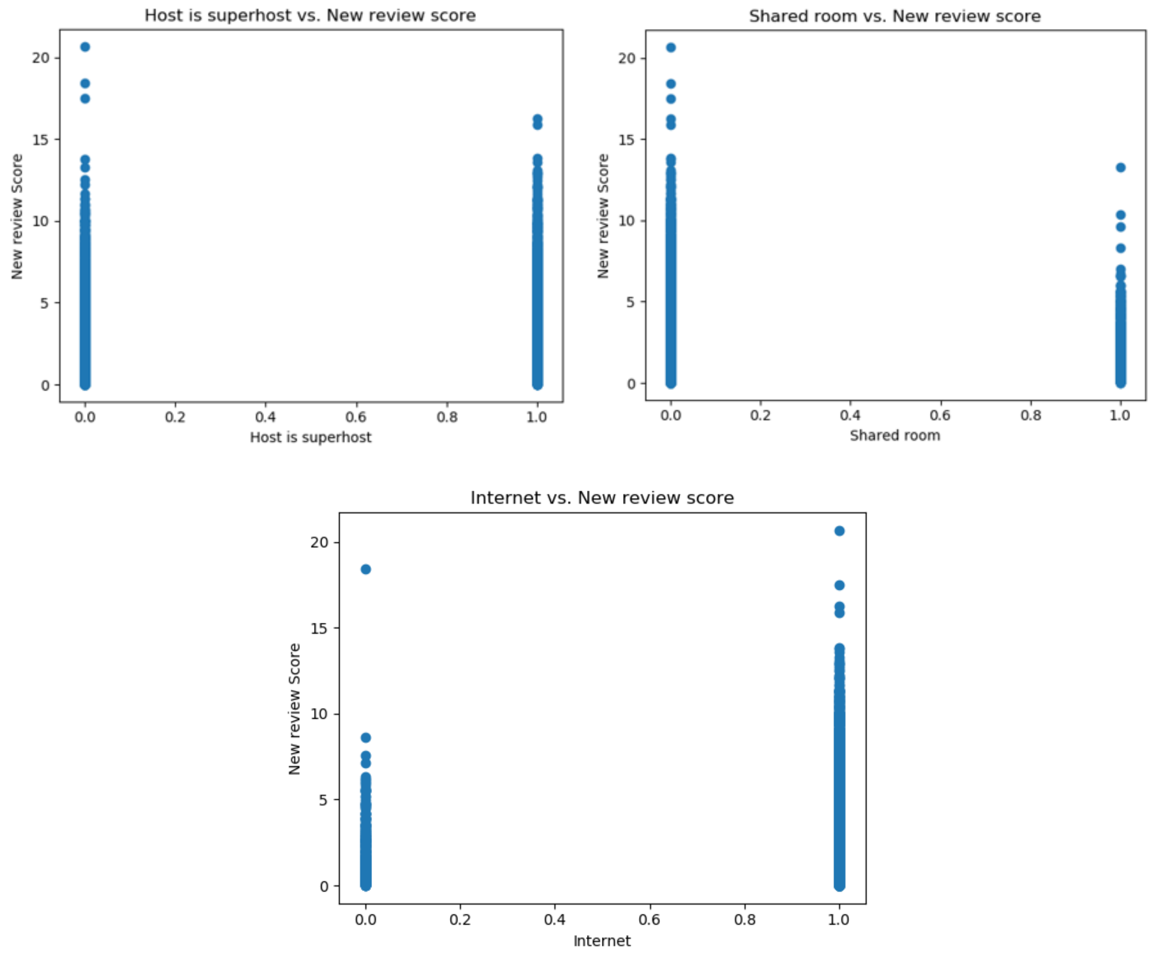


Figure 14: Relationship of Top-5 Features

The coefficient represents the additional effect of adding that variable to the model, in case the effects of all other variables are already accounted for in the model. As you can see, all of the features with a coefficient of at least 0.5 are categorical. This is important to know as it helps us to interpret our results. Thus, the five coefficients can be interpreted as follows:

- When a host provides internet, the average NRS value increases by 0.5.
- When a host is superhost, the average NRS value increases by 0.68.
- When a host provides the opportunity of self-check-in, the average NRS increases by 0.6.
- When a host response “one category” faster, the average NRS value increases by 0.55.
- When a host provides a shared room, the average NRS value decreases by 0.51.

However, correlation does not imply causation. Only because one feature is increased, decreased, provided or not provided it cannot be concluded that this will cause an increase or decrease of the NRS. Even non-related data can pass statistical tests and provide a strong correlation. Moreover, the OLS results show that only 25% of the variance in the dependent variable is explained by the independent variables. Nonetheless, we have decided to move forward with our 25 features. From this point forward, we will try to predict the NRS and make an assessment of the viability of our model later.

7.2 Train and test the Regression Model

First, we split our dataset into 80% training and 20% testing data. Second, we train our linear regression with the training data, predict NRS in the test dataset and assess its prediction accuracy. This is a supervised machine learning approach as we work with labeled data and try to build a model that predicts our new review score. However, the accuracy of our prediction remains poor at approximately 24.9%. Exemplary, we take a look at the first five predictions:

Data Record	Actual	Predicted
5780	2.65	2.816046
15696	0.56	1.847413
8364	1.85	0.467359
4997	2.96	2.069288
(...)	(...)	(...)

Table 5: Predictions

As table 5 shows, the prediction of the NRS is not very accurate. Apparently, the NRS cannot be accurately predicted with the available data and a linear regression model that minimizes the sum of square error. However, it could have been interesting to try different regression models, such as the Lasso or Ridge method and investigate the prediction accuracy of these models. In general, linear regression models only make sense when applied to linear datasets. Our analysis indicates that we do not have a linear dataset and the prediction of a combined score is difficult to achieve. This is why the application of neural networks could be interesting if the number of data points in the dataset could be significantly increased. On this basis, an accurate dataset could then also have been verified by testing the model on data from other cities. Only this way you can show that the model is not overfitted and only predicts the NRS in the dataset, the model was trained on. The more features a model utilizes, the higher the chance of overfitting and this can threaten generalizability and hence, applicability to new data.

8 Evaluation

In this chapter will we evaluate the results we have generated from our data analytics. Firstly, will we discuss and reflect our outcome to determine if the business objectives have been achieved. Then will we touch upon ethical and legal barriers when processing personal data and how we can overcome these barriers? At last, will we have a reflection on critical considerations that can hinder the potential of big data processes.

8.1 Discussion and Critical Reflection of Results

By exploring the data and conducting a prediction exercise we gained valuable insights that can help hosts becoming more successful on Airbnb. However, the analysis suffers many shortcomings that need to be considered before we can conclude our analysis.

First, we did not find a summary of how to interpret the features of the Airbnb dataset. A misinterpretation of features can have a number of serious negative effects on the analysis. In our analysis, we have excluded a number of features on the basis of our understanding. If some features are misunderstood, we could have ended up with a consideration of other features. Moreover, we need to state that we cannot be certain that we did not exclude any important features that would have been beneficial for our analysis. In addition, a misinterpretation of features could have led to wrong conclusions of our results.

Second, we defined the new review score as a measure of success but another variable or another combination of variables could have been more accurate when describing the success of a listing.

Third, our qualitative analysis of the review score has revealed that the review score is also dependent on the location. However, our prediction model does not incorporate the location data.

Fourth, data cleaning and -preparation is based on our own perception. Any data cleaning and data preparation task takes away information from the dataset and therefore, increases abstraction. At the same time, it can be argued that we have not cleaned the dataset sufficient enough. Many outliers in group “high performers” might have led to distortion and the exclusion of other data records could have increased the accuracy of our analysis.

Fifth, we could consider including a qualitative analysis of the actual user review to ensure we include the whole picture. According to (Wang, 2013) it is important to include the user stories together with the data to put it in perspective and get the full context. Thus, it would be interesting to qualitatively analyze the outliers in the high performing group to understand why those reviews perform significantly better than the rest of the group.

Sixth, our prediction model takes a high number of features into consideration and as the chance of overfitting increases with an increasing number of features, our model might only describe our dataset and cannot be generalized.

Seventh, we divided the listings into the three groups low-, medium- and high performers but the boundaries of the groups have been freely chosen. The question is whether you can classify listings on the basis of the NRS and whether the boundaries have been chosen correctly in order to make valuable insights.

Last but not least, this analysis is constrained by time and computing resources. More time could have led to more valuable insights as more features could have been investigated and different prediction models could have been compared.

Even though our prediction exercise was based on a thorough investigation of the dataset the prediction accuracy remains low. We mainly ascribe this to missing time and computing power as we could not compare different regression or conduct predictions with the help of gradient boosting or neural networks for instance. However, machine learning algorithms are one of the greatest trends in Big Data and are increasingly applied in science and industry. The algorithms allow researchers to work with vast datasets and help to classify or predict data. But with further development of technology, we usually move up the abstraction ladder. One example is the increasing deployment of machine learning libraries. The libraries, we are also using, make it easy to apply algorithms and make predictions. But when algorithms are integrated into the programming language, you need less and less understanding of the algorithm to make predictions and other data insights. Similarly, many students and practitioners make use of online tools, such as Microsoft Azure Machine learning. It makes it very easy to apply machine learning algorithms, but these tools can never take away the need for proper data and business understanding. When applying algorithms to unexplored datasets, the results might be very misleading.

The long list of shortcomings increases the importance of context and qualitative considerations. With the qualitative analysis of the review score on Airbnb, we tried to provide context. It explains how a user makes reviews on the platform and helps the reader to put the new review score in perspective. Moreover, we do not only present the numeric results but also provide interpretations. This provides a more holistic view and puts the numbers into perspective.

When combining the results of the data exploration task with the prediction exercise and considering the limitations of our report, we can still conclude on nine statements.

1. Amenities seem to greatly influence the new review score.
2. The price is not associated with a vast effect on the new review score.
3. The host response time seems to be an important variable when it comes to the review score in perspective of reviews per months.
4. The room type data indicate that the more privacy the guest experiences, the higher the new review score.
5. A smaller number of minimum nights is associated with a higher new review score.
6. Superhosts generally achieve higher NRS than non-superhosts.
7. Self-Check-in is associated with a higher NRS.
8. The requirement of a guest phone verification negatively affects NRS.
9. The new review score cannot be accurately predicted by the described data and the use of a regression model that minimizes the square error.

We have set out the goal to find out how the Airbnb variables affect the success of listings and consequently provide concrete recommendations for hosts. By analyzing the new review score as a measure of success, we are confident to be able to do so. However, we were not able to predict our measure of success, the new review score, with high accuracy and therefore, cannot fulfill all of our objectives completely. Nonetheless, the derived insights help us to conclude about deployments of our insights as well as potential deployments of a prediction model.

8.2 Legal and Ethical Considerations

Barocas and Nissenbaum (2014) and Boyd and Crawford (2012) argue that a recurring issue in most big data studies is that the ethical implications of working with data are not well understood. Therefore, in this chapter, we want to highlight and discuss the legal and ethical challenges posed by publishing Airbnb data and our implications working with this data.

As mentioned in the data presentation chapter, we are working with a publicly available dataset of Airbnb listings in Los Angeles. The raw dataset does not directly contain names, addresses, date of birth, etc. However, it does contain location data and other data that, when put together, might be used to identify a person. As (Zimmer, 2010) states even if one feels that *“all identifying information has been removed from a dataset, it is often trivial to piece together random bits of information to deduce one’s identity”* (Zimmer, 2010 p.319). According to (General Data Protection Regulation, 2016) personal information is *“‘any information relating to an identified or identifiable natural person (‘data subject’)“* (General Data Protection Regulation, 2016, p. 33). Furthermore, the law specifically states that personal data includes *“name, an identification number, location data (...)”* (General Data Protection Regulation, 2016, p. 33). This means that Airbnb is actually publishing personal data, and this is a major privacy concern for the users.

Since personal data is made available to the public, other companies might use this data to analyze user behavior and use it to their advantage. For example, knowing your travel behavior might allow travel agencies to personalize commercials specifically for that user, or insurance companies might use this data to determine the insurance price of that specific user.

We are trying to resolve this privacy issue by anonymizing the data. According to Barocas and Nissenbaum (2014) anonymizing the data is a way to overcome the challenge big data poses to privacy, especially if this data is not needed to proceed with the analysis. In our case, we have removed all the features that we found unnecessary in our analysis. Although, we are still including location data, i.e. longitude and latitude in our analysis which is still considered personal data. However, we question the precision of these coordinates as these observations are made in a major city. Most of the locations are in skyscrapers which means that this location data cannot be used to uniquely identify the data subject. Thus, we argue that this data is already considered as anonymized. Zimmer (2010) argues that it might not be a major privacy leak if the personal data could be de-anonymization, as the dataset is already publicly available. This means everyone could potentially obtain the full dataset including location data and similar data, which can then be used to identify the persons included in this dataset. Furthermore, Zimmer (2010) argues that while there might be a privacy issue in publishing research including personal information, the personal data is actually already available and provided by the users themselves on Airbnb. Instead, it might be a matter of the users’ ignorance of the consent they are giving.

Furthermore, it is important to consider the legal aspects of data processing. In 2018, the General Data Protection Regulation (GDPR) was put into force by the European Parliament and the Council with the purpose to protect all individuals within the European Union with regards to the processing of personal data and the movement of personal data (General Data Protection Regulation, 2016, p. 1). Since we are placed in the EU and we are processing Airbnb data that might contain EU citizen data, we have to comply with the GDPR.

8.3 Pitfalls of Big Data

Even though the era of big data has allowed organizations to achieve great business value and allowing them to better understand their customers and their decision-making process, there are concerning pitfalls that play an important part in Big Data discussions. As we have presented earlier, Big Data pitfalls are inherently entangled with privacy concerns and data availability. As mentioned earlier in our report, the increasing availability of data often leads to privacy concerns and the misuse of data. The data scandal of Cambridge Analytica that misused personal data of millions of Facebook users without their consent is a good example of the importance of protecting personal data and users' privacy. Given the rise of Big Data as a phenomenon and its potential value, the authors (Boyd and Crawford, 2014) stress several critical issues that also need to be considered when using Big Data.

Big Data can be valuable if the data is rich and analyzed thoroughly (Boyd and Crawford, 2014). As the authors explain, understanding the data is a crucial stage in the CRISP-DM framework. Researchers need to be aware that the sample size they are analyzing might be skewed already before investigating it. For example, we cannot be certain that 'Inside Airbnb' has provided us with a complete and not distorted data to make a different impression. (Boyd and Crawford, 2014, Page 669) use the example of Twitter, which very much relates to our case as well. We cannot certainly say that our dataset provides the reality of the average ratings of each host because Airbnb has defined guidelines and policies that users need to follow when giving a review (Airbnb, 2019). If they fail to do so, for example threatening the host or using violent language, the reviews will be removed and thus changing the actual average rating. We are also aware that these guidelines ensure safety for the hosts and help to filter most of the reviews that are trying to manipulate the average rating in a negative way. However, the number of reviews might not be a representative sample of the average rating and this needs to be considered by the researchers (Marquardt, 2016, page, 302).

When understanding the data, it is also important that researchers are aware that the investigated dataset might not include important variables affecting the outcome. In our case, it could be interesting to investigate the crime rates in Los Angeles or the hosts' income and whether it would affect the average ratings. However, bigger data does not necessarily mean better data (Boyd and Crawford, 2014). This indicates that it is up to the researchers to determine how much and what data is needed in order to maintain the right context and generate a valuable and reliable outcome.

The next stage of the CRISP-DM framework is to prepare the data. In most of the cases, the researchers always want to provide a hypothesis and test it and eventually resulting in improvements in knowledge with an objective mindset. However, all researchers are interpreters of data (Boyd and Crawford, 2014). The process of defining the hypothesis and deciding which attributes and variables to include and exclude is inherently biased. Thus, it is important to outline and understand these biases and put the data into context early in order to achieve a holistic and comprehensive picture. In our report, we have tried to provide this holistic view by providing the reader context and additional information about our decisions.

When researchers reach the modeling phase, it is important to maintain the context of the data. (Boyd and Crawford 2014) points out "Because large data sets can be modeled, data are often reduced to what can fit into a mathematical model. Yet, taken out of context, data lose meaning and value" (Boyd and Crawford, 2014). Thus, it is important to be aware that generated models do not necessarily provide any value for the same purpose in another geographical area. In our case, this means that our model can work for Airbnb users in Los Angeles but might not be able to transfer it to Copenhagen for instance. In Copenhagen, the tenants need might differ a lot from the needs of people traveling to Los Angeles. Data is not generic and therefore, it is important to thoroughly look at the data used for the model to determine if cultural or legal factors are restricting the model to produce valuable insights elsewhere.

Our outlined pitfalls of Big Data show that Big Data and Data Analytics require more than an only good technical understanding of the way data is processed. We have shown that the entire lifecycle of data processing requires thorough and holistic considerations of how data is obtained and processed. However, when researchers and practitioners are aware of the outlined pitfalls, Big Data processes can lead to a lot of value.

8.4 Potential Deployment

In the last stage of CRISP-DM the actual results of the analysis are deployed into real use cases (Provost and Fawcett, 2013). Even though our model predicts the new review score with fairly low accuracy, we still consider the prediction model as very valuable when deploying in real-world applications. Therefore, we see our model as an early stage prototype that can be beneficial for Airbnb and their hosts. As the Provost & Fawcett explain, we see the process as a lifecycle and a continuous development which eventually leads to further development of our prediction model (Provost and Fawcett, 2013, p. 34). In addition, to our prediction, we take the comparison of low and high performers into consideration.

With this in mind, we suggest a number of potential deployments of our model and the knowledge generated from it. First, the insights can greatly benefit Airbnb hosts to become more successful on Airbnb and thus, earn more money by providing a place to sleep. When a host finds out how each of the variables is associated with their success on the platform, many might start to rethink what to include and what not to include in their listing. This is especially interesting when the variables are easy to change, such as providing shampoo.

Second, the insights can also be used by the platform itself. As we have explained in the very beginning of our report, Airbnb earns a small commission with each agreement that has been made on the platform. If the hosts are more successful and earn more money, the platform will also be able to generate more revenue. Thus, we suggest that Airbnb provides tips to hosts and therefore, encourage them to change their Airbnb offering. For example, the platform could send notifications to hosts stating that hosts who provide shampoo for guests to win more tenants and increase the review score by a certain percentage on average. Moreover, Airbnb could make use of a prediction model to foresee the success of a listing and therefore, predict the expected revenue. This does not only help Airbnb planning but also gives the platform the opportunity to provide new services to its members. These additional services can be provided for free to increase the user experience or even been packed into a service package that the member can pay extra for. Moreover, Airbnb can also use this information to suggest travelers' new places that have not received many reviews. When Airbnb manages to predict the review score on the basis of available variables, users find good listings faster. This can again increase the user experience; more travelers might book a night on Airbnb and the platform, as well as the hosts, can make more money.

There are many ways how hosts and the platform can utilize our initial analytical considerations. However, the potential deployments always need to be in line with the legal and ethical concerns and it is important that the platform works with the agreement of its users.

9 Conclusion and Concrete Recommendations

This report aims to answer the research question: “How can the variables of an Airbnb dataset of the city Los Angeles be used to form recommendations that increase the success of Airbnb hosts?”

By analyzing an Airbnb dataset of over 31,000 data records and putting the data into context, we want to provide concrete insights that can be utilized to become more successful on Airbnb. Therefore, we made use of the process model CRISP-DM and derived our own review score as a measure of success. The actual data analysis was then split into a data exploration task that compares core differences of high performers vs. low performers and a prediction exercise that aimed to predict the new review score on basis of the available features.

Based on the earlier presented results, we want to give concrete recommendations for hosts that can help both, hosts to become more successful and Airbnb generating more revenue. Following is our recommendations for hosts on Airbnb:

- Provide internet, shampoo and a hairdryer for guests. The results indicate the availability of internet, shampoo, and a hairdryer are associated with a higher NRS and therefore, lead to more success on the platform.
- Reply to requests as early as possible. The results of our research indicate that hosts with a faster response time achieve higher scores.
- Become a superhost. We have shown that especially superhost are particularly successful on Airbnb.
- Provide the quest as much privacy as possible. Our research suggests that especially shared rooms are associated with worse NRS.
- Allow Self-Check-in for guests. Our numbers indicate that the possibility of self-check-in lead to higher NRS.

Furthermore, we discussed pitfalls of data analyses and touched upon various considerations needed in big data processes. Especially, it is important to question the dataset and reflect the actual reality. Maintaining the context is crucial to the value of the data and researchers need to recognize that all researchers are interpreters of data. Thus, it is important to outline and understand the occurring biases. Finally, we discussed the ethical and legal concerns raised when working with personal data in big data processes. Specifically, location data and the privacy concern raised by including this in our research are discussed. We reflected on how to overcome such privacy concerns. Specifically, anonymizing data was discussed as a way to obsolete these concerns. However, since we are working with a publicly available dataset, we found that the ignorance of the content of the consent is causing another privacy issue. Finally, the legal aspect of processing personal data was considered. It was found that we need to comply with the GDPR since we are processing data within the EU.

The increasing application of data analytics lead to great business potentials but also increase the chance of making wrong interpretations. The extensive list of short comes and the poor accuracy of the prediction model make clear that data analysis has its limits. With the increasing demand for data and analytics, the demand for interpretation and contextual thinking increases in equal measures. However, when researchers and practitioners manage to include both, quantitative as well as qualitative considerations in data analytics, very valuable insights can be derived with the help of data. We believe that the data analytics lead to great potentials, not only for business but also for research purposes.

10 References

Airbnb Citizen. (2019). LA county airbnb hosts earned $613 million, welcomed more than 2.7 million guests in 2018 | Airbnb Citizen. [online] Available at: https://www.airbnbcitizen.com/la-county-airbnb-hosts-earned-613-million-welcomed-more-than-2-7-million-guests-in-2018/ [Accessed 14 May 2019].

Airbnb. (2019). Hvordan fungerer stjernevurderinger?. [online] Available at: https://www.airbnb.dk/help/article/1257/how-do-star-ratings-work [Accessed 14 May 2019].

Airbnb. (2019). Hvad er Airbnb's afpresningspolitik. [online] Available at: https://www.airbnb.dk/help/article/548/what-is-airbnb-s-extortion-policy [Accessed 14 May 2019].

Airbnb. (2019). Superhost: Anerkendelse af det bedste inden for gæstfrihed. [online] Available at: https://www.airbnb.dk/superhost [Accessed 14 May 2019].

Barocas, S. and Nissenbaum, H. (2014). Big Data’s End Run around Anonymity and Consent. Cambridge Books Online, pp.44-75.

Bosa, D. and Salinas, S. (2019). Airbnb says it's been profitable for two years straight as it heads for IPO. [online] CNBC. Available at: https://www.cnbc.com/2019/01/15/airbnb-sustains-profit-as-it-heads-toward-ipo.html [Accessed 13 May 2019].

Boyd, D. and Crawford, K. (2012). CRITICAL QUESTIONS FOR BIG DATA: Provocations for a cultural, technological, and scholarly phenomenon. Information, Communication & Society, 15(5), pp.662 –679.

General Data Protection Regulation. (2016). The European Parliament and the Council, Article 4 (1).

Inside Airbnb. (2019). Inside Airbnb. Adding data to the debate.. [online] Available at: http://insideairbnb.com/about.html [Accessed 14 May 2019].

ipropertymanagement.com (2019). 2019 Airbnb Statistics - User & Market Growth Data [Updated]. [online] iPropertyManagement.com. Available at: https://ipropertymanagement.com/airbnb-statistics/ [Accessed 13 May 2019].

Marquardt, N. (2016). Counting the countless: Statistics on homelessness and the spatial ontology of political numbers. Environment and Planning D: Society and Space, 34(2).

Müller, O., Junglas, I., Brocke, J. and Debortoli, S. (2016). Utilizing big data analytics for information systems research: challenges, promises and guidelines. European Journal of Information Systems, pp.289-302.

Nath, T. (2019). Airbnb vs. Hotels: What's the Difference?. [online] Investopedia. Available at: https://www.investopedia.com/articles/investing/112414/airbnb-brings-sharing-economy-hotels.asp [Accessed 13 May 2019].

Provost, F. and Fawcett, T. (2013). Data Science for Business: What You Need to Know about Data Mining and Data-Analytic Thinking. 1st ed. O'Reilly Media, Inc.

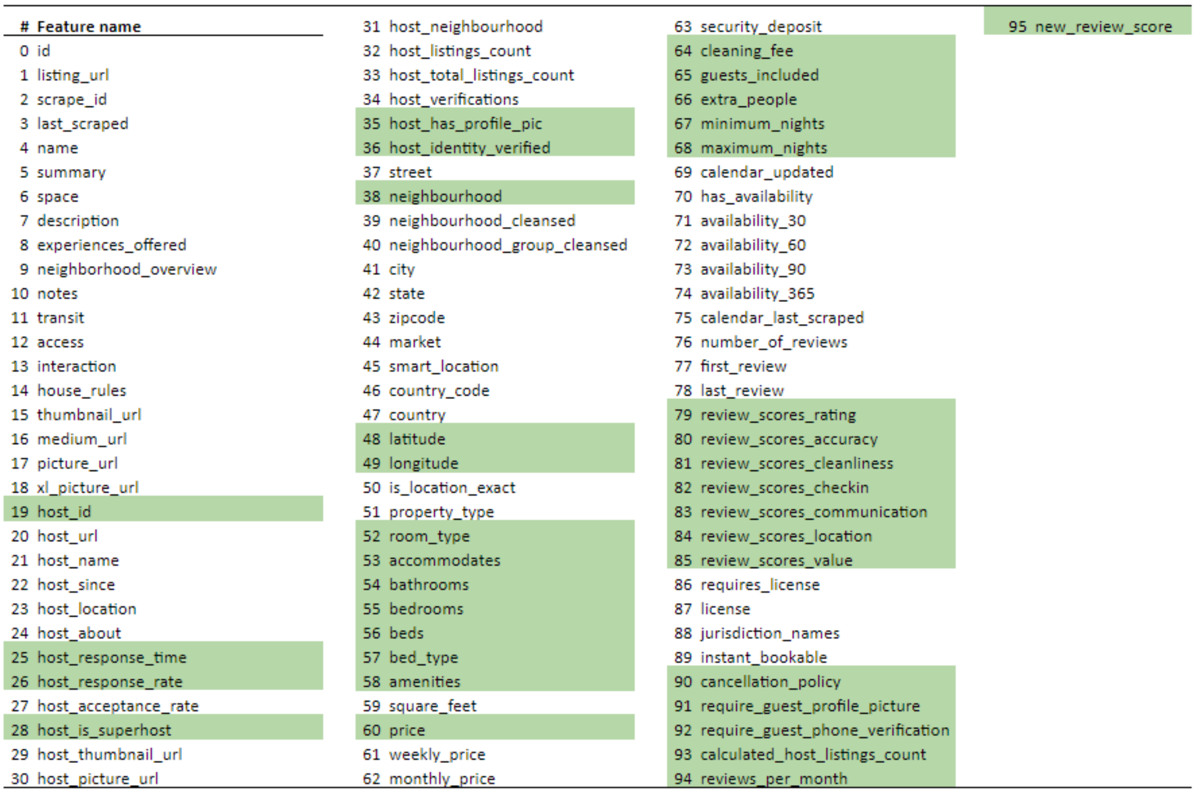
Shearer, C. (2000). The CRISP-DM Model: The New Blueprint for Data Mining. JOURNAL OF DATA WAREHOUSING, 5(4), p.1-22.

Wang, T. (2013). Big Data Needs Thick Data. pp.1-6.

Zimmer, M. (2010). ‘‘But the data is already public’’: on the ethics of research in Facebook. Ethics Inf Technol, pp.313-325.

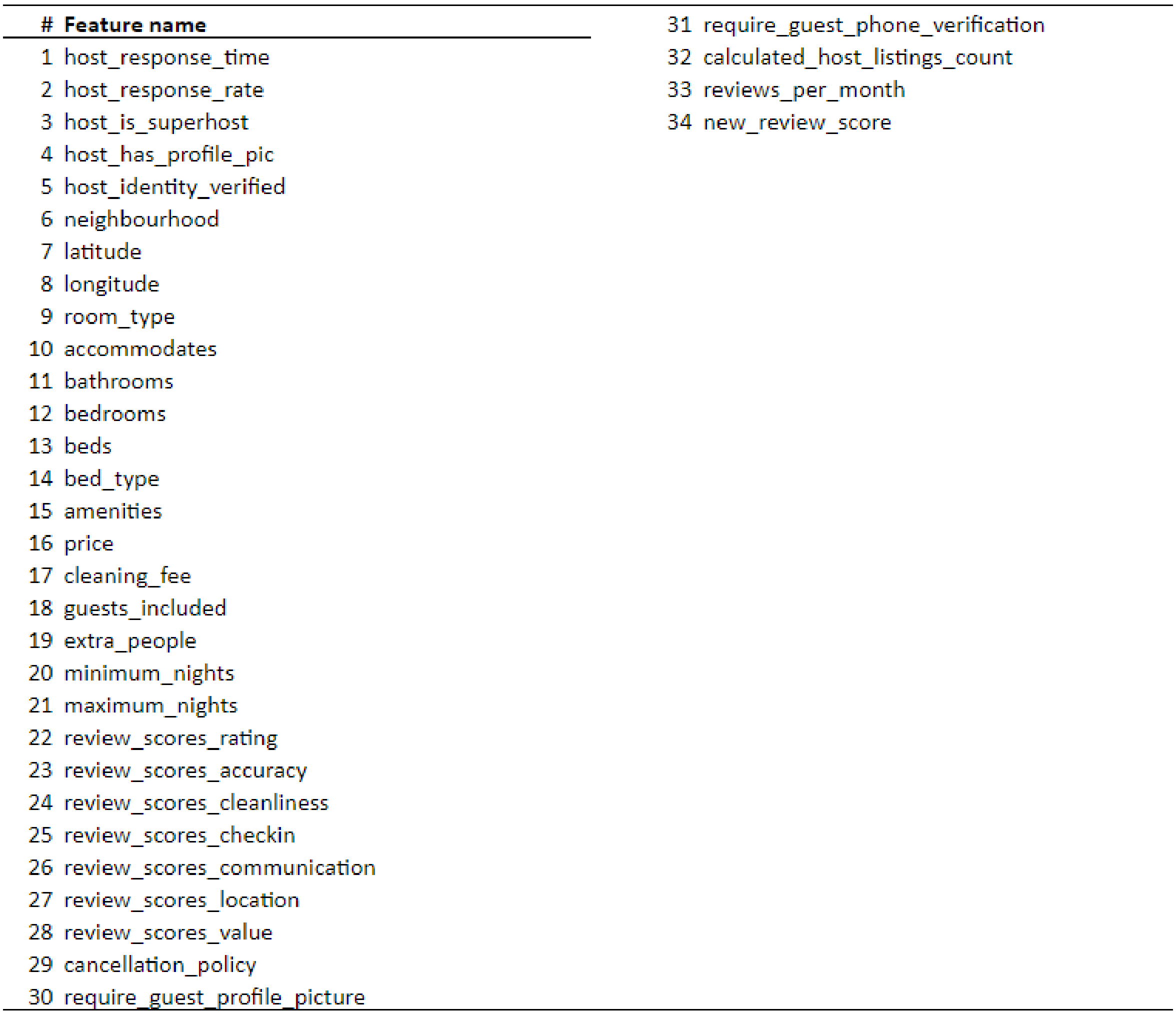
11 Appendix

Appendix A: Working Tables



#	Feature name
0	id
1	listing_url
2	scrape_id
3	last_scraped
4	name
5	summary
6	space
7	description
8	experiences_offered
9	neighborhood_overview
10	notes
11	transit
12	access
13	interaction
14	house_rules
15	thumbnail_url
16	medium_url
17	picture_url
18	xl_picture_url
19	host_id 🟩
20	host_url
21	host_name
22	host_since
23	host_location
24	host_about
25	host_response_time 🟩
26	host_response_rate 🟩
27	host_acceptance_rate
28	host_is_superhost 🟩
29	host_thumbnail_url
30	host_picture_url
31	host_neighbourhood
32	host_listings_count
33	host_total_listings_count
34	host_verifications
35	host_has_profile_pic 🟩
36	host_identity_verified 🟩
37	street
38	neighbourhood 🟩
39	neighbourhood_cleansed
40	neighbourhood_group_cleansed
41	city
42	state
43	zipcode
44	market
45	smart_location
46	country_code
47	country
48	latitude 🟩
49	longitude 🟩
50	is_location_exact
51	property_type
52	room_type 🟩
53	accommodates 🟩
54	bathrooms 🟩
55	bedrooms 🟩
56	beds 🟩
57	bed_type 🟩
58	amenities 🟩
59	square_feet
60	price 🟩
61	weekly_price
62	monthly_price
63	security_deposit
64	cleaning_fee 🟩
65	guests_included 🟩
66	extra_people 🟩
67	minimum_nights 🟩
68	maximum_nights 🟩
69	calendar_updated
70	has_availability
71	availability_30
72	availability_60
73	availability_90
74	availability_365
75	calendar_last_scraped
76	number_of_reviews
77	first_review
78	last_review
79	review_scores_rating 🟩
80	review_scores_accuracy 🟩
81	review_scores_cleanliness 🟩
82	review_scores_checkin 🟩
83	review_scores_communication 🟩
84	review_scores_location 🟩
85	review_scores_value 🟩
86	requires_license
87	license
88	jurisdiction_names
89	instant_bookable
90	cancellation_policy 🟩
91	require_guest_profile_picture 🟩
92	require_guest_phone_verification 🟩
93	calculated_host_listings_count 🟩
94	reviews_per_month 🟩
95	new_review_score 🟩

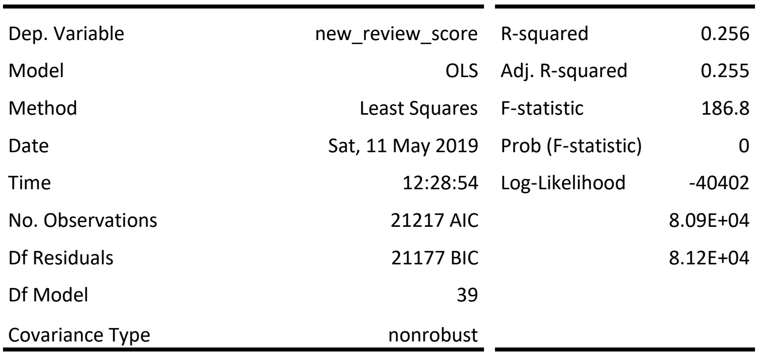
Table 6: Selection of Features



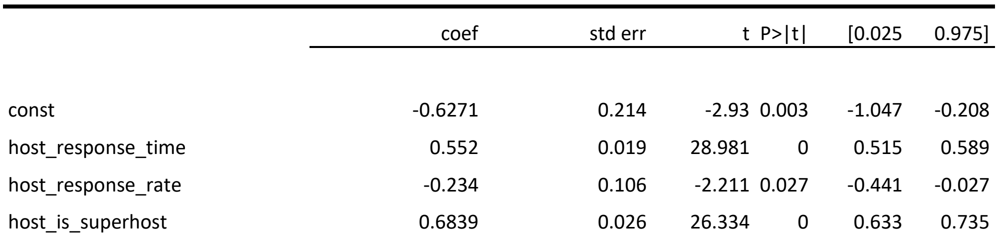
#	Feature name
1	host_response_time
2	host_response_rate
3	host_is_superhost
4	host_has_profile_pic
5	host_identity_verified
6	neighbourhood
7	latitude
8	longitude
9	room_type
10	accommodates
11	bathrooms
12	bedrooms
13	beds
14	bed_type
15	amenities
16	price
17	cleaning_fee
18	guests_included
19	extra_people
20	minimum_nights
21	maximum_nights
22	review_scores_rating
23	review_scores_accuracy
24	review_scores_cleanliness
25	review_scores_checkin
26	review_scores_communication
27	review_scores_location
28	review_scores_value
29	cancellation_policy
30	require_guest_profile_picture
31	require_guest_phone_verification
32	calculated_host_listings_count
33	reviews_per_month
34	new_review_score

Table 7: Selected Features

Appendix B: OLS Regression Results



Dep. Variable	new_review_score	R-squared	0.256
Model	OLS	Adj. R-squared	0.255
Method	Least Squares	F-statistic	186.8
Date	Sat, 11 May 2019	Prob (F-statistic)	0
Time	12:28:54	Log-Likelihood	-40402
No. Observations	21217	AIC	8.09E+04
Df Residuals	21177	BIC	8.12E+04
Df Model	39		
Covariance Type	nonrobust		



	coef	std err	t	P>|t|	[0.025	0.975]
const	-0.6271	0.214	-2.93	0.003	-1.047	-0.208
host_response_time	0.552	0.019	28.981	0	0.515	0.589
host_response_rate	-0.234	0.106	-2.211	0.027	-0.441	-0.027
host_is_superhost	0.6839	0.026	26.334	0	0.633	0.735

host_has_profile_pic	0.3454	0.304	1.137	0.255	-0.25	0.941
host_identity_verified	0.0056	0.027	0.207	0.836	-0.047	0.058
Entire home/apt	0.0271	0.077	0.352	0.725	-0.124	0.178
Private room	-0.1447	0.075	-1.934	0.053	-0.291	0.002
shared room	-0.5095	0.082	-6.179	0	-0.671	-0.348
accommodates	0.0547	0.01	5.743	0	0.036	0.073
bathrooms	0.0924	0.022	4.275	0	0.05	0.135
bedrooms	-0.2859	0.021	-13.522	0	-0.327	-0.244
beds	0.0496	0.013	3.959	0	0.025	0.074
Real Bed	0.1236	0.072	1.726	0.084	-0.017	0.264
Pull-out Sofa	-0.1888	0.133	-1.414	0.157	-0.45	0.073
Airbed	-0.2521	0.147	-1.717	0.086	-0.54	0.036
Couch	-0.3229	0.185	-1.745	0.081	-0.685	0.04
Futon	0.0129	0.116	0.112	0.911	-0.214	0.24
TV	-0.0431	0.031	-1.404	0.16	-0.103	0.017
Internet	0.4998	0.079	6.362	0	0.346	0.654
Air Conditioning	-0.082	0.026	-3.119	0.002	-0.133	-0.03
Kitchen	-0.2849	0.04	-7.121	0	-0.363	-0.206
Free parking on premises	0.0229	0.026	0.885	0.376	-0.028	0.074
Indoor Fireplace	-0.2304	0.029	-7.88	0	-0.288	-0.173
Family/kid friendly	0.1531	0.026	5.91	0	0.102	0.204
Washer	-0.2607	0.029	-9.088	0	-0.317	-0.204
24-hour check-in	0.0405	0.026	1.563	0.118	-0.01	0.091
Hair dryer	0.3011	0.028	10.757	0	0.246	0.356
Hot tub	0.0023	0.035	0.066	0.947	-0.066	0.071
Pets allowed	-0.0636	0.03	-2.11	0.035	-0.123	-0.005
Shampoo	0.3143	0.03	10.332	0	0.255	0.374
Gym	0.2663	0.039	6.758	0	0.189	0.344
Self Check-In	0.6027	0.032	19.008	0	0.541	0.665
price	0.0001	6.56E-05	2.191	0.028	1.52E-05	0
cleaning_fee	-0.006	0	-22.084	0	-0.007	-0.005
guests_included	0.0711	0.01	7.282	0	0.052	0.09
extra_people	-0.0038	0.001	-6.848	0	-0.005	-0.003
minimum_nights	-0.0245	0.002	-13.891	0	-0.028	-0.021
maximum_nights	-4.78E-10	4.38E-10	-1.091	0.275	-1.34E-09	3.80E-10
cancellation_policy	-0.0028	0.015	-0.183	0.855	-0.033	0.027
require_guest_profile_picture	-0.0438	0.096	-0.456	0.648	-0.232	0.144
require_guest_phone_verification	-0.351	0.084	-4.169	0	-0.516	-0.186

Omnibus:	7228.455	Durbin-Watson:	0.454
Prob(Omnibus):	0	Jarque-Bera (JB):	34033.603
Skew:	1.598	Prob(JB):	0
Kurtosis:	8.319	Cond. No.	1.00E+16

Table 8: OLS Regression Results

Appendix C: Python Code

1. Import Libraries

```
import warnings
import pandas as pd
import numpy as np
import seaborn as sns
from sklearn import preprocessing
import matplotlib.pyplot as plt
from sklearn import datasets, linear_model
from sklearn.linear_model import LinearRegression
import statsmodels.api as sm
from scipy import stats
from sklearn import preprocessing
```

2. Import Data

```
data = pd.read_csv('Working Airbnb Dataset.csv')
```

3. Define Groups and Generate Boxplots

```
df = pd.DataFrame(data)
low = df.iloc[0:2917,0]
medium = df.iloc[2918:12143,0]
high = df.iloc[12144:21216,0]

FEV1data = (low), (medium), (high)
labels = ['High Performers', 'Medium Performers', 'Low Performers']
plt.boxplot(FEV1data, labels=labels)
plt.ylabel('New Review Score')
plt.title('Comparison of High-, Medium- & Low Performers')
plt.show()
```

4. Correlation

```
X = listings.iloc[:,:-10]
X1 = listings.iloc[:,18] #change to required feature
y = listings.iloc[:,-1]
print(X1)

plot.scatter(X1, y)
plot.title('<Name of feature> vs. New review score')
plot.xlabel('<Name of feature>')
plot.ylabel('New review Score')
plot.show()


corr = new_df.corr()
```

```
fig = plt.figure()
ax = fig.add_subplot(111)
cax = ax.matshow(corr,cmap='coolwarm', vmin=-1, vmax=1)
fig.colorbar(cax)
ticks = np.arange(0,len(new_df.columns),1)
ax.set_xticks(ticks)
plt.xticks(rotation=90)
ax.set_yticks(ticks)
ax.set_xticklabels(new_df.columns)
ax.set_yticklabels(new_df.columns)
plt.show()
print(corr)

#Using Pearson Correlation
plt.figure(figsize=(120,100))
cor = pd.DataFrame(listings).corr()
#cor = pd.DataFrame(new_df).corr()
sns.heatmap(cor, annot=True, cmap=plt.cm.Reds)
plt.show()
```

5. Conduct OLS

```
X = listings.iloc[:,:-10]
y = listings.iloc[:,-1]
X2 = sm.add_constant(X)
est = sm.OLS(y, X2)
est2 = est.fit()
print(est2.summary())
```

6. Linear Regression

```
pd.set_option('display.max_columns', 500)
warnings.filterwarnings('ignore')

listings = pd.read_csv('Working Airbnb Dataset (final regression).csv')

X = listings.iloc[:,:-1]
y = listings.iloc[:,-1]

y = y.astype(float)

#print(y)

X_Norm = preprocessing.scale(X)
X_Norm_int = X_Norm.astype(float)
```

```
X_train, X_test, y_train, y_test = train_test_split(X, y, test_size=0.2, random_state=101)

model = LinearRegression()
model.fit(X_train, y_train)

y_pred = model.predict(X_test)

Accuracy = model.score(X_test, y_test)
print('Prediction Accuracy: ', Accuracy)

df = pd.DataFrame({'Actual': y_test, 'Predicted': y_pred})
print(df)

# sum of square of residuals
ssr = np.sum((y_pred - y_test)**2)

#  total sum of squares
sst = np.sum((y_test - np.mean(y_test))**2)

# R2 score
r2_score = 1 - (ssr/sst)
print('r1 score: ', r2_score)
```

EXTRA

FIND RIGHT MODEL

```
pd.set_option('display.max_columns', 500)
warnings.filterwarnings('ignore')

listings = pd.read_csv('Working Airbnb Dataset (regression).csv')
print(listings.shape)

X = listings.iloc[:,:-10]
y = listings.iloc[:,-1]


X_Norm = preprocessing.scale(X)

#no of features
nof_list=np.arange(1,41)
high_score=0
#Variable to store the optimum features
nof=0
score_list =[]
for n in range(len(nof_list)):
   X_train, X_test, y_train, y_test = train_test_split(X, y, test_size = 0.2, random_state = 0)
   model = LinearRegression()
   rfe = RFE(model,nof_list[n])
   X_train_rfe = rfe.fit_transform(X_train,y_train)
   X_test_rfe = rfe.transform(X_test)
   model.fit(X_train_rfe,y_train)
   score = model.score(X_test_rfe,y_test)
   score_list.append(score)
   if(score>high_score):
      high_score = score
      nof = nof_list[n]
print("Optimum number of features: %d" %nof)
print("Score with %d features: %f" % (nof, high_score))



#no of features
nof_list=np.arange(1,41)
high_score=0
#Variable to store the optimum features
```

```
nof=0
score_list =[]
for n in range(len(nof_list)):
   X_train, X_test, y_train, y_test = train_test_split(X_Norm, y, test_size = 0.2,
random_state = 0)
   #model = GradientBoostingRegressor(n_estimators=100)
   model = Lasso()
   rfe = RFE(model,nof_list[n])
   X_test_rfe = rfe.transform(X_test)
   model.fit(X_train_rfe,y_train)
   score = model.score(X_test_rfe,y_test)
   score_list.append(score)
   if(score>high_score):
      high_score = score
      nof = nof_list[n]
print("Optimum number of features: %d" %nof)
print("Score with %d features: %f" % (nof, high_score))

#no of features
nof_list=np.arange(1,41)
high_score=0
#Variable to store the optimum features
nof=0
score_list =[]
for n in range(len(nof_list)):
   X_train, X_test, y_train, y_test = train_test_split(X_Norm, y, test_size = 0.2,
random_state = 0)
   #model = GradientBoostingRegressor(n_estimators=100)
   model = Ridge()
   rfe = RFE(model,nof_list[n])
   X_test_rfe = rfe.transform(X_test)
   model.fit(X_train_rfe,y_train)
   score = model.score(X_test_rfe,y_test)
   score_list.append(score)
   if(score>high_score):
      high_score = score
      nof = nof_list[n]
print("Optimum number of features: %d" %nof)
print("Score with %d features: %f" % (nof, high_score))

#linear regression (new review score 1)
#Optimum number of features: 32
#Score with 32 features: 0.252098
```

```
        #GradientBoostingRegressor (new review score1)
        #Optimum number of features: 34
        #Score with 34 features: 0.357638
```

APPLY GRADIENT BOOST REGRESSION

```
listings = pd.read_csv('Working Airbnb Dataset (26featuresRegression).csv')

X = listings.iloc[:,:-1]
y = listings.iloc[:,-1]

X_Norm = preprocessing.scale(X)

X_train, X_test, y_train, y_test = train_test_split(X_Norm, y, test_size=0.2, random_state=101)

gbrt = GradientBoostingRegressor(n_estimators=100)
gbrt.fit(X_train, y_train)
y_pred = gbrt.predict(X_test)

print('Feature Importances: ', gbrt.feature_importances_)

print('R-squared for Train:', gbrt.score(X_train, y_train))
print('R-squared for Test:', gbrt.score(X_test, y_test))

def GradientBooster(param_grid, n_jobs):
    estimator = GradientBoostingRegressor()
    cv = ShuffleSplit(X_train.shape[0], test_size=0.2)
    classifier = GridSearchCV(estimator=estimator, cv=cv, param_grid=param_grid, n_jobs=n_jobs)
    classifier.fit(X_train, y_train)
    print ('Best Estimator learned through GridSearch: ')
    print (classifier.best_estimator_)
    return cv, classifier.best_estimator_


def plot_learning_curve(estimator, title, X, y, ylim=None, cv=None, n_jobs=1,
train_sizes=np.linspace(.1, 1.0, 5)):
    plt.figure()
    plt.title(title)
    if ylim is not None:
        plt.ylim(*ylim)
    plt.xlabel('Training examples')
    plt.ylabel('Score')
    train_sizes, train_scores, test_scores = learning_curve(
        estimator, X, y, cv=cv, n_jobs=n_jobs, train_sizes=train_sizes)
    train_scores_mean = np.mean(train_scores, axis=1)
    train_scores_std = np.std(train_scores, axis=1)
    test_scores_mean = np.mean(test_scores, axis=1)
    test_scores_std = np.std(test_scores, axis=1)
    plt.grid()
    plt.fill_between(train_sizes, train_scores_mean - train_scores_std,
train_scores_mean + train_scores_std, alpha=0.1, color='r')
    plt.fill_between(train_sizes, test_scores_mean - test_scores_std,
test_scores_mean + test_scores_std, alpha=0.1, color='g')
    plt.plot(train_sizes, train_scores_mean, 'o-', color='r', label='Training score')
    plt.plot(train_sizes, test_scores_mean, 'o-', color='g', label='Cross-validation score')
    plt.legend(loc='best')
    return plt


param_grid={'n_estimators':[100],
            'learning_rate': [0.3],# 0.05, 0.02, 0.01],
            'max_depth':[6],#4,6], 'min_samples_leaf':[3],#,5,9,17],
            'max_features':[1.0],#,0.3]#,0.1]
            }
n_jobs=4

#Let's fit GBRT to the digits training dataset by calling the function we just created.

cv,best_est=GradientBooster(param_grid, n_jobs)
```

```
print('Best Estimator Parameters')
print('---------------------------')
print('n_estimators: %d' %best_est.n_estimators)
print('max_depth: %d' %best_est.max_depth)
print('Learning Rate: %.1f' %best_est.learning_rate)
print('min_samples_leaf: %d' %best_est.min_samples_leaf)
print('max_features: %.1f' %best_est.max_features)

print('Train R-squared: %.2f' %best_est.score(X_train,y_train))



#Calling fit on the estimator so we can look at feature_importances.

estimator.fit(X_train, y_train)
# Calculate the feature ranking - Top 10
importances = estimator.feature_importances_
indices = np.argsort(importances)[::-1]
print('Lending Club Loan Data - Top 10 Important Features\n')

for f in range(10):
    print('%d. %s (%f)' % (f + 1, loan.columns[indices[f]], importances[indices[f]]))

#Plot the feature importances of the forest

indices=indices[:10]
plt.figure()
plt.title('Top 10 Feature importances')
plt.bar(range(10), importances[indices], color="r", align="center")
plt.xticks(range(10), loan.columns[indices], fontsize=14, rotation=45)
plt.xlim([-1, 10])
plt.show()

#Mean Feature Importance
```

```
print('Mean Feature Importance %.6f' %np.mean(importances))
```

ONLY LASSO

```
#from sklearn.feature_selection import SelectKBest, f_regression
import matplotlib
from sklearn.linear_model import RidgeCV, LassoCV, Ridge, Lasso
from sklearn.feature_selection import RFE

pd.set_option('display.max_columns', 500)
warnings.filterwarnings('ignore')

#listings = pd.read_csv('Working Airbnb Dataset (regression).csv')
#X = listings.iloc[:,:-10]

listings = pd.read_csv('Working Airbnb Dataset (regression).csv')
X = listings.iloc[:,:-10]

y = listings.iloc[:,-1]

X_Norm = preprocessing.scale(X)
X1 = pd.DataFrame(X_Norm)

reg = LassoCV()
reg.fit(X1, y)
print("Best alpha using built-in LassoCV: %f" % reg.alpha_)
print("Best score using built-in LassoCV: %f" %reg.score(X1,y))
coef = pd.Series(reg.coef_, index = list(X.columns.values))

print("Lasso picked " + str(sum(coef != 0)) + " variables and eliminated the other " + str(sum(coef == 0)) + " variables")

imp_coef = coef.sort_values()
matplotlib.rcParams['figure.figsize'] = (8.0, 10.0)
imp_coef.plot(kind = "barh")
plt.title("Feature importance using Lasso Model")
plt.show()
```