Project Part 1

Joan Cortes - 100438579
Jose Antonio Jijon Vorbeck - 100438045
Didier Dirks - 100443386

23-12-2020

Team Project Deliverable 1

The dataset that we have selected consists of data given by AirBnB to describe many of their locations in Madrid. In the original file, there are in total *106* variables and *20837* observations, from which some of the variables are mostly incomplete or require some data imputation. The data that has been selected is located on the following link to the *Kaggle* project: *Madrid AirBnB Data*

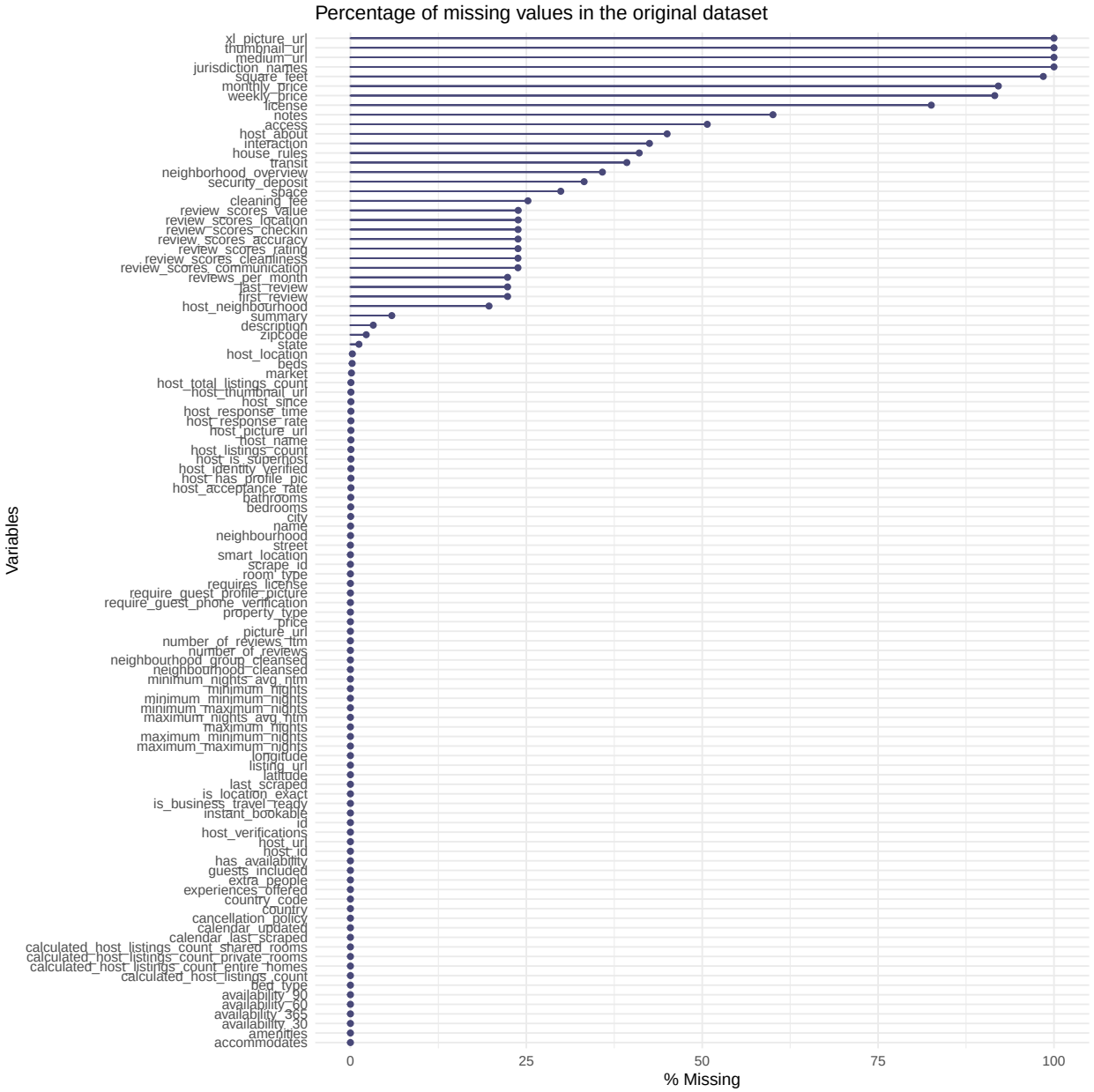
This report will contain the following 5 main steps:

1. **Data Pre-Processing**
    - Visualization of missing values
    - Dropping useless variables (url, has_image etc..)
    - Selecting useful quantitative variables for analysis
    - Imputing missing data
2. **Data Visualization**
    - Scatterplot Matrix
    - Kernel Densities
    - Parallel Coordinates Plots (PCP)
    - Division between sub-populations
3. **Data Metrics**
    - Mean Vector
    - Covariance matrix
    - Correlation Matrix
    - Analysis of different sub-populations
4. **Principal Component Analysis**
    - Transforming skewed data
    - Analysis of different groups
    - See which grouping criterion has the most difference
5. **Independent Component Analysis**
    - Exploring non-Gaussian variables to identify outliers
    - Assessing the existence of natural grouping in the data

1. Data Pre-Processing

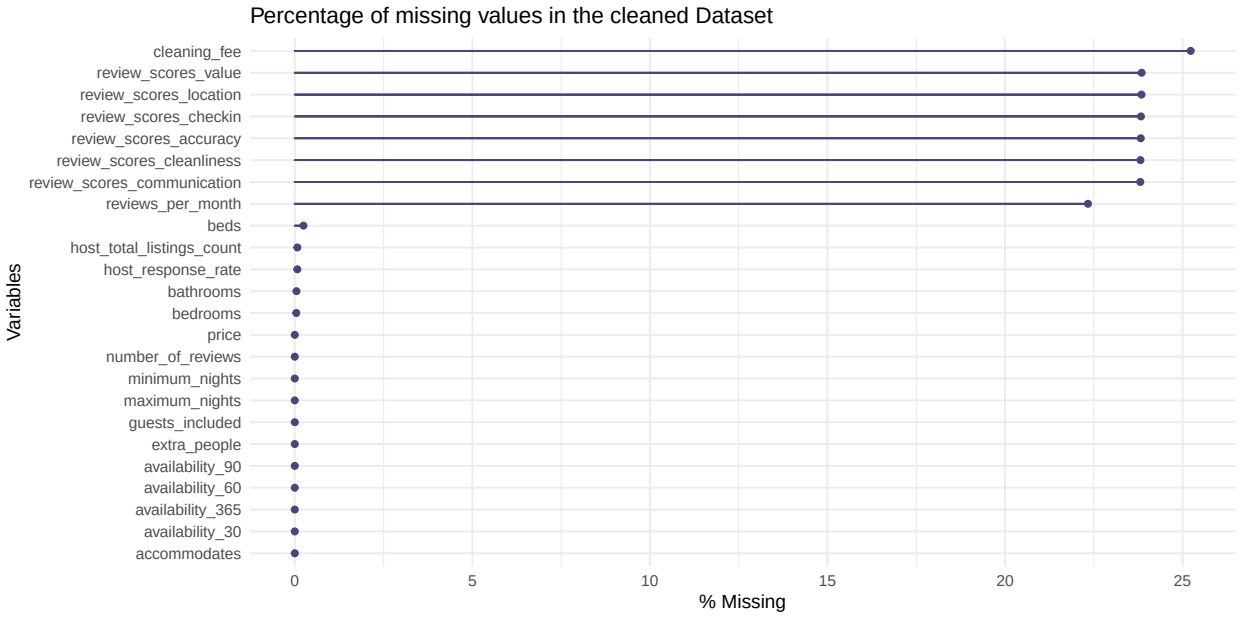
In this step, we will first focus on variables that have a high percentage of missing data ($> 30\%$) these variables are of no use, and we cannot perform any analysis with them, therefore we need to drop them from the dataset.

Then, there exist some other variables that have no meaning or add nothing different to the analysis, like host, listing URL, images city, country, state and other variables. We need to drop these variables as well. We present below a graph with the percentages of missing values per variable from the complete and original dataset downloaded from Kaggle.



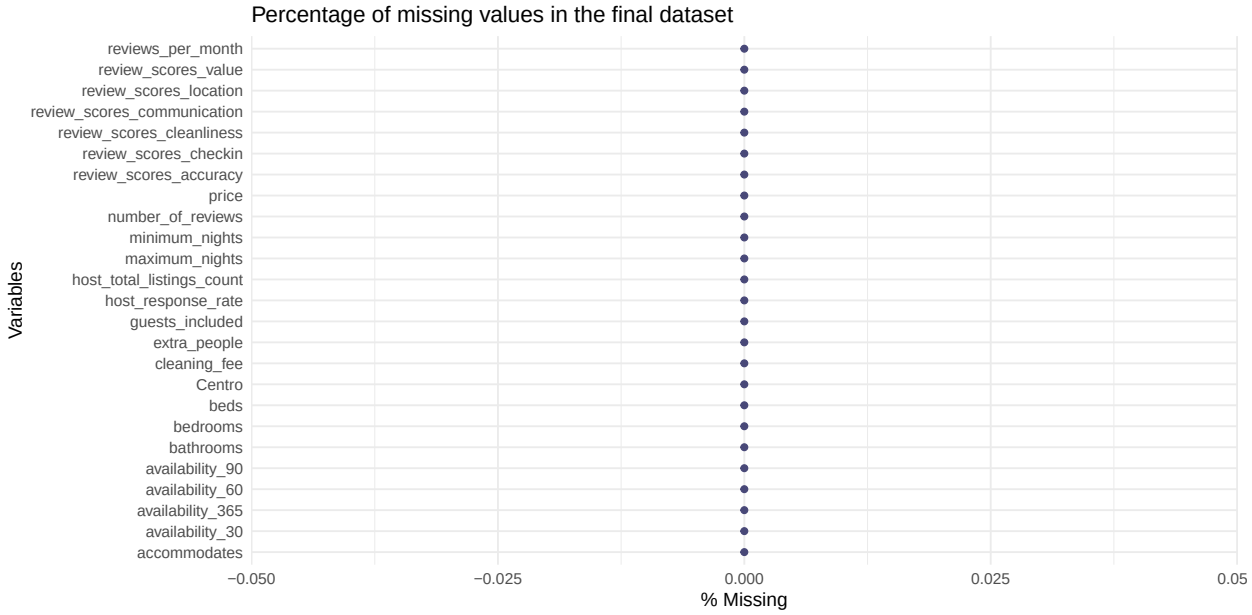
After verification, we can drop the variables that have $>$ 30% empty (NA) values, and therefore cannot be imputed. But we still need to drop many variables that have no use for the analysis.

After dropping all the useless variables we stay with a more significant set of variables, note that we have not done any analysis yet, only dropping variables that are of no use for further tests. We must then select the set of **quantitative variables plus a few qualitative ones** that we are going to make plots with for visualization and further statistical analysis. Some categorical variables that are of interest are: *neighborhood*, *host_is_superhost*, *property_type* and *room_type*. We will later add these back to the dataset, but for now the focus lies on the quantitative variables. That is why for now only the quantitative variables will be selected from the dataset. Below we present a graph with the missing values of the selected variables:



As shown above, some of the variables have still some missing values (but in lower percentages now). These variables are: Cleaning fee, and the variables consisting of reviews for the locations. Now is when one of the main steps of data pre-processing comes into play, performing data imputation. Adding values for the missing cells is very important, since we do not want to have NA's in the PCA nor the ICA process. To this end, we make use of the 'mice' package to impute missing values based on a regressive way, and not by only imputing the mean.

We must also deal with erroneous data in the variables *maximum_nights* and *minimum_nights*, since these two variables present some errors that might be due to wrong input of the numbers by the host when creating their listing on the website. After assessing this possibility, we observed and identified extreme values for *maximum_nights* and *minimum_nights*. High values were set to indicate there is neither a maximum nor a minimum. In order to tackle this problem, we set the maximum number of nights equal to 365 (a year) and the maximum number of minimum nights equal to the 95th percentile of the original data, which is equal to 10.

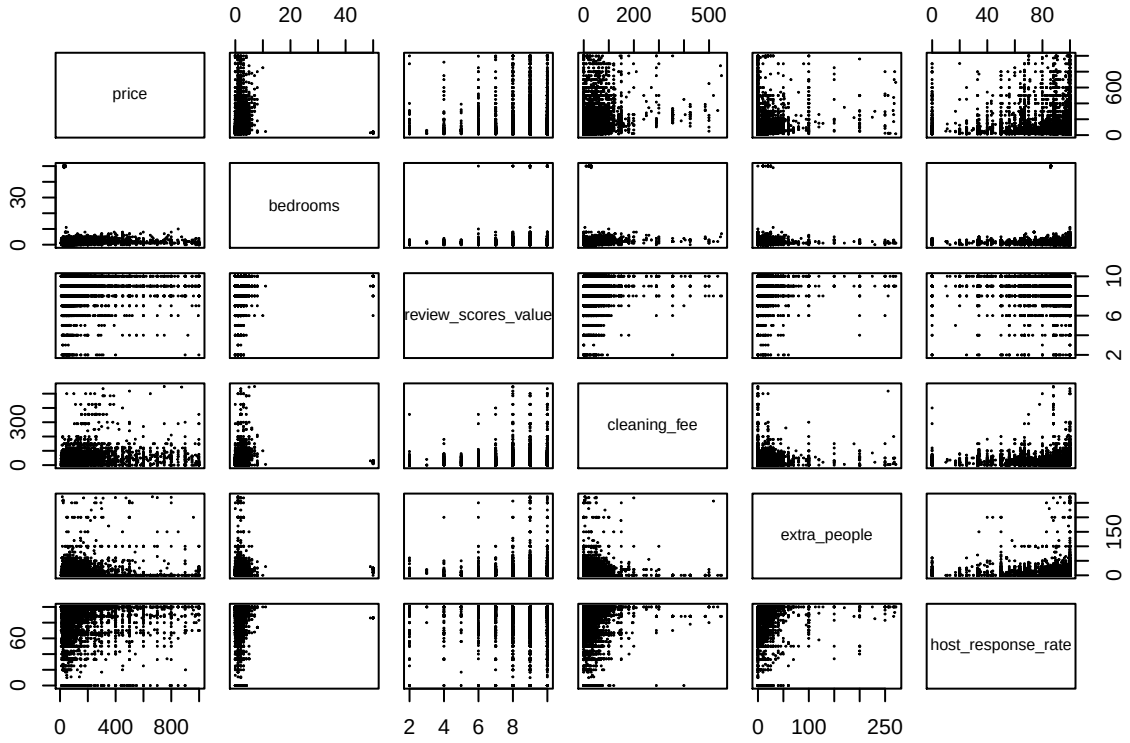


Finally we have achieved a clean, complete and useful set of variables that we are going to use for the remaining of the presentation. For a first glance, we present below the summary of this data set, which contains **20837 observations** for **25 variables**.

2. Data Visualization

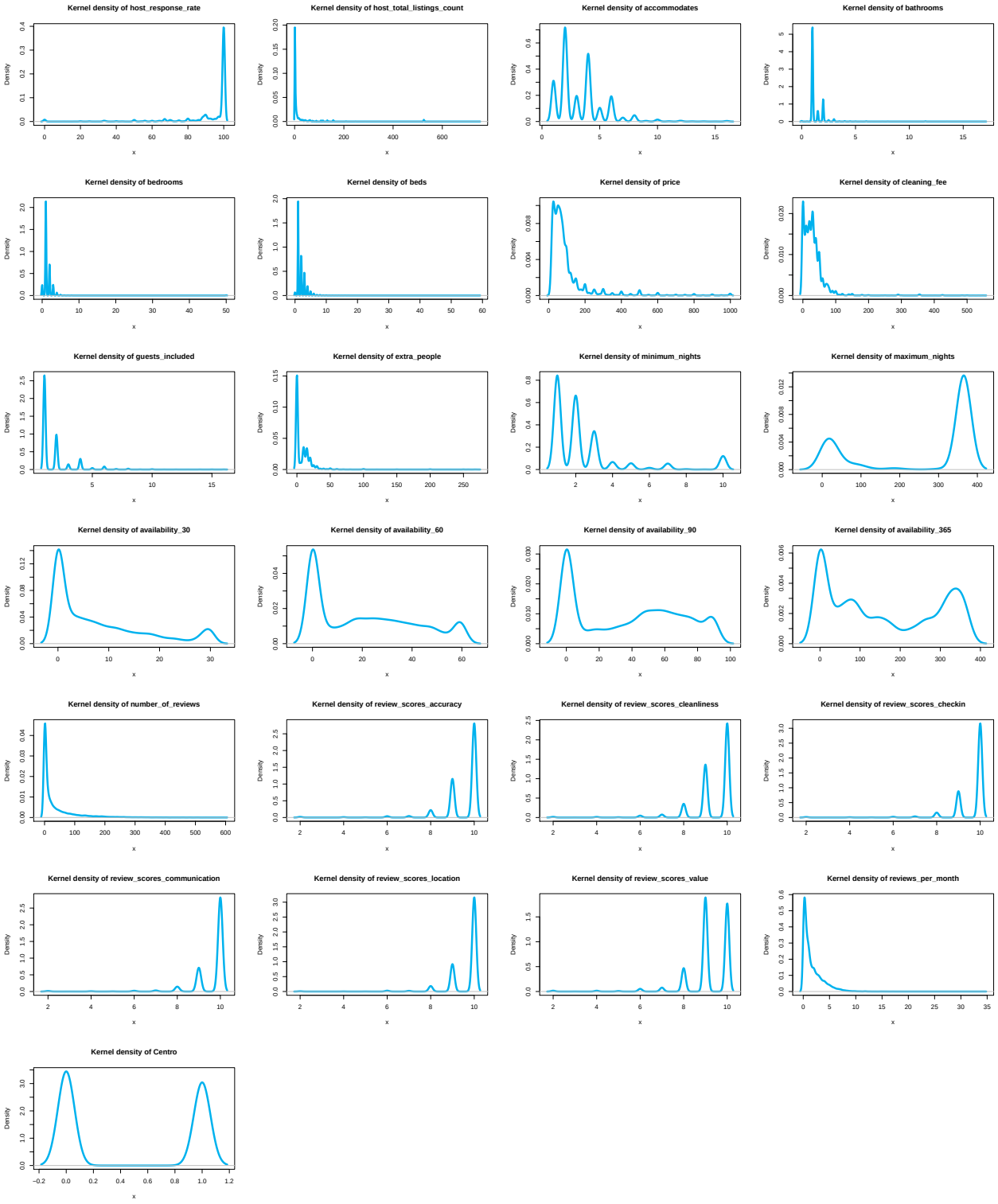
In this part of the report, we will present the data in a visual manner, to get a first glance of it and see if by inspection we can get some trends and main aspects that could help differentiate and split between populations.

We first plot a scatterplot matrix of only some selected variables since the matrix can get very big and hard to interpret. We observe that the data is very skewed in many variables and at first glance we cannot identify any clear trends or correlations between the variables. This indicates that we should take logarithms of many variables to see if we can obtain better visualizations after.



Below we will plot all the kernel density graphs for the 24 quantitative variables, this series of graphs allow us to identify the distribution of the predictors and we can see which ones will need any logarithmic transformation and which ones will not. In general, most of the interesting variables will need to have the log transformation since they are highly skewed to the right. There are also many discrete variables that have only some set of numbers and would make no sense to take the logarithm of those.

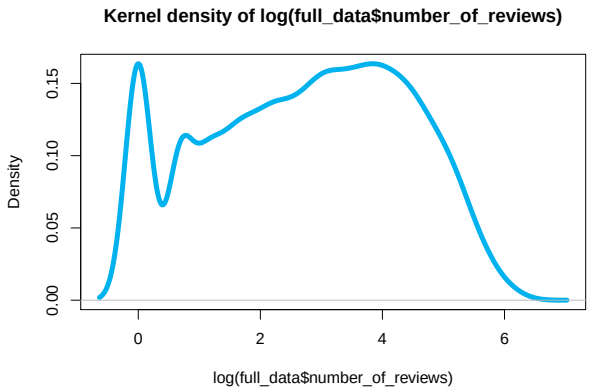
Plotting all the quantitative variables' kernel distributions:

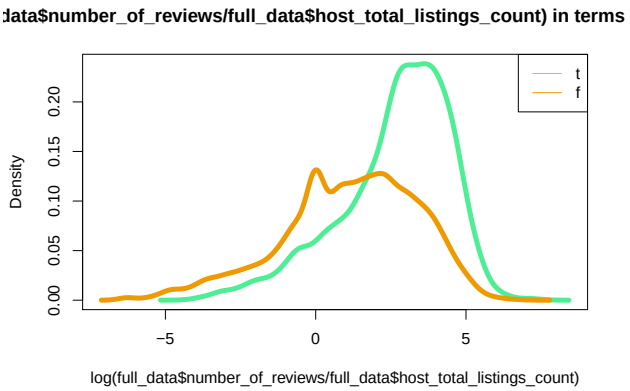


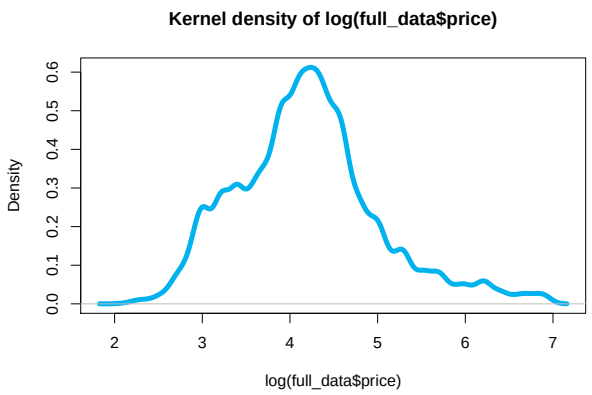
Now we plot some of the graphs that we consider meaningful when splitting the data into different groups. This is an interesting step of the analysis, since it allows us to see if there are any differences between categories and we can come to smarter conclusions.

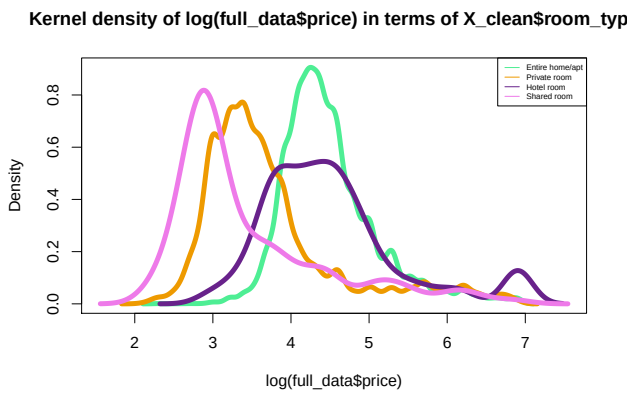
The variables that we have chosen to split things into are the following:

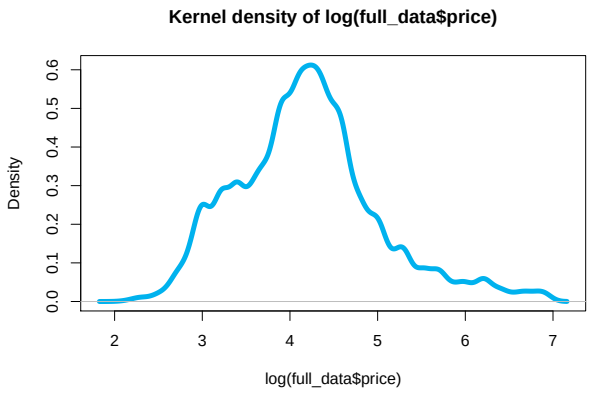
- Number of reviews when host is super host and when not
- Price for the different types of AirBnB listings
- Price for the listings in Centro and not Centro
- Cleaning fee for the listings in Centro and not Centro

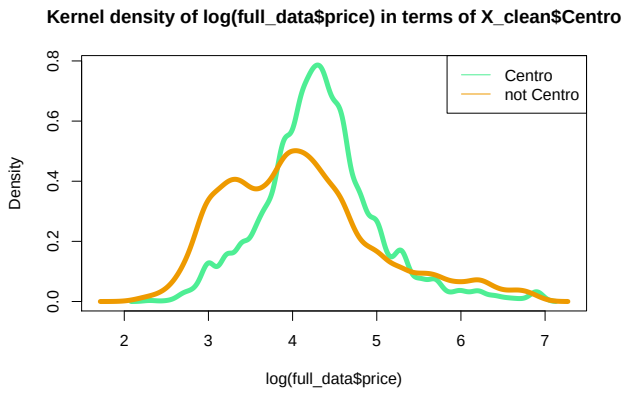


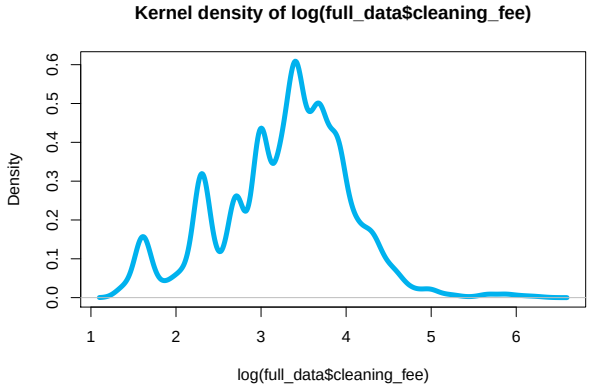


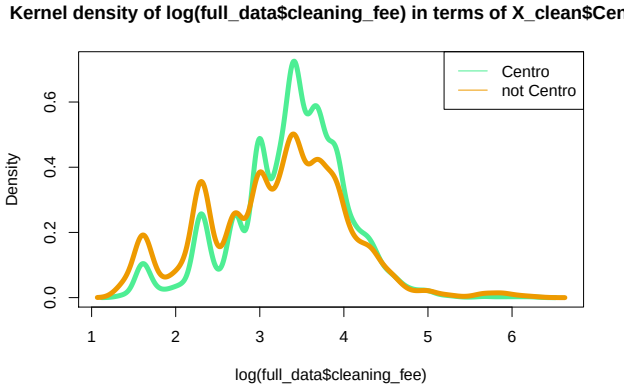






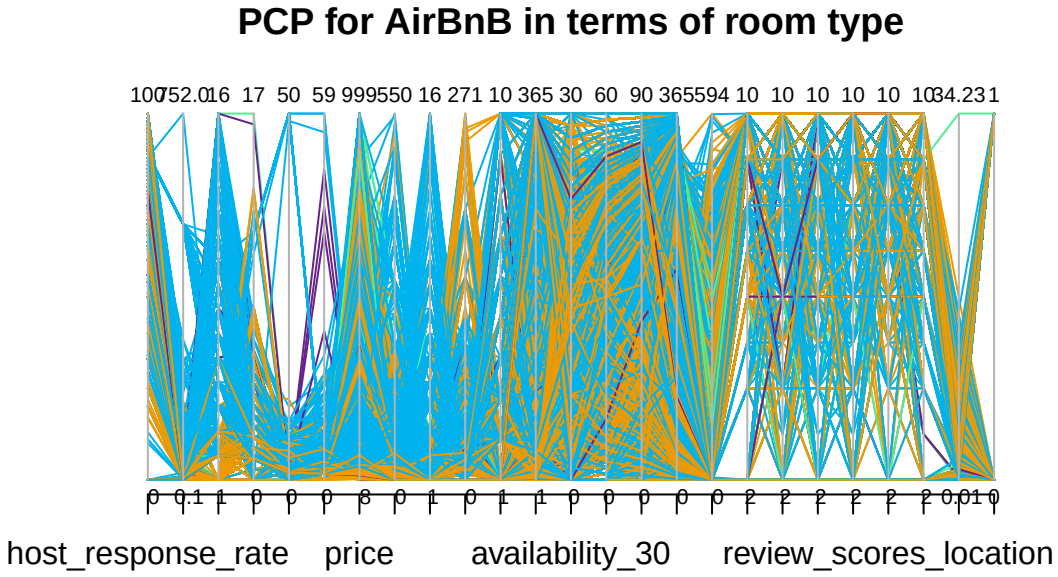


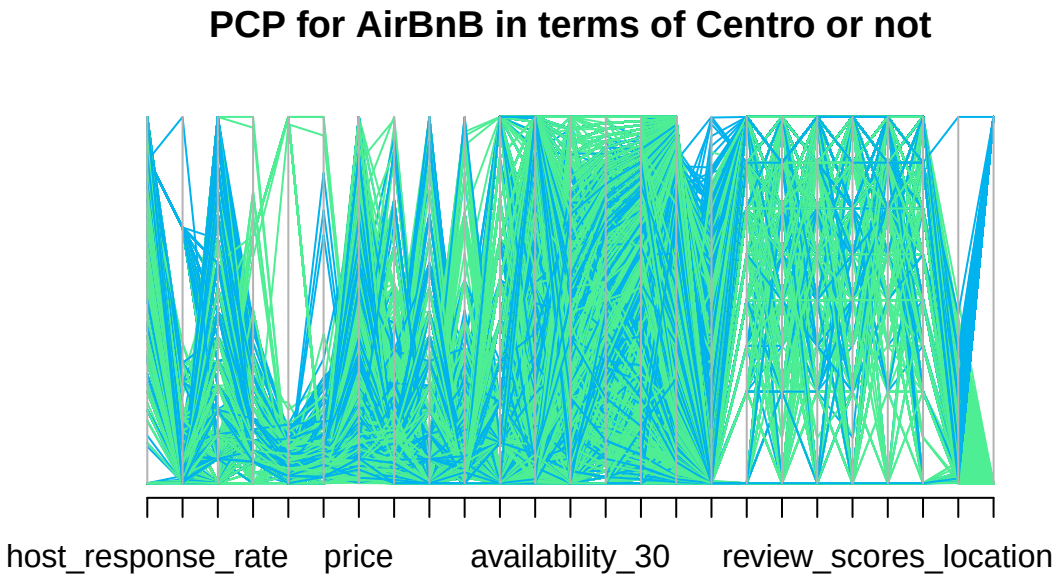




We can see that there are some variabilities in the data when looked in different categories. It is interesting to see the differences in the number of reviews, the review values and the availability that super host get compared to not_superhosts. This is telling us that superhost get more clients and that they get in general better reviews than normal hosts. We can also clearly note the differences in price between the 4 different room types existing, although this is expected, this can serve us as a clear differentiator between room_types. The last two plots show the differences for price and cleaning fee for listings that are located in the center of Madrid, and for the ones that are not. We clearly see that there is an increase in both quantities when they are in the city center.

Another very useful feature to differentiate between clusters is the Parallel Coordinate Plot (PCP). Below we show the PCP for the dataset, differentiated by room type and neighborhood (Centro or not Centro).





We can see that in these graphs it is very hard to see the differences between the groups, specially because we have more than 20.000 observations and 24 variables, so at the end this is very messy and gives us no great insight of what is really going on between the data. Nevertheless, we can still see that there are some variables in which the colors differ a bit, thus meaning that there could be differences between some variables.

3. Data Metrics

In this part we will present the mean vector, covariance matrix and correlation matrix of the quantitative variables chosen in the step before. Moreover, we will also perform this analysis for the variables differentiated by *room_type*, this is going to give us more insight of the true differences between the groups in *room_type*.

We start by the mean vector, which can help us see the 'starting point' of the data set, and we can compare it to the different room types.

Sample Mean Vector:

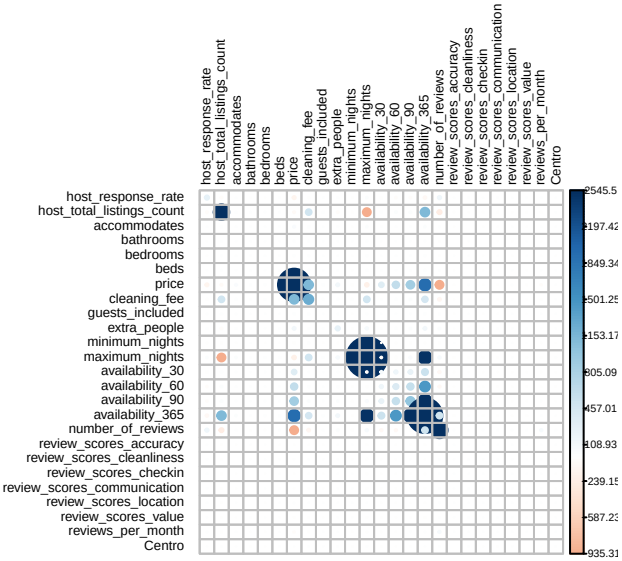
	$Mean$	$Mean[Ent.home/apt]$	$Mean[Hotel]$	$Mean[Private]$	$Mean[Shared]$
$host-response-rate$	92.94	93.91	94.04	91.35	90.08
$host-total-listings-count$	17.06	22.74	22.19	7.56	6.87
$accommodates$	3.28	4.17	3.25	1.82	2.62
$bathrooms$	1.3	1.32	1.25	1.26	1.91
$bedrooms$	1.44	1.68	1.35	1.06	1
$beds$	1.98	2.41	2	1.24	2.74
$price$	98.5	112.01	153.49	73.61	64.19
$cleaning-fee$	29.94	38.17	29.06	16.83	12.58
$guests-included$	1.75	2.15	2.04	1.1	1.04
$extra-people$	8.31	9.26	8.3	6.62	11.31
$minimum-nights$	2.52	2.63	2.08	2.38	1.94
$maximum-nights$	255.26	274.68	258.26	222.52	254.48
$availability-30$	7.55	6.73	9.03	8.65	12.57
$availability-60$	21.33	20.78	25.9	21.61	30.75
$availability-90$	38.05	38.46	47.1	36.33	50.29
$availability-365$	155.1	161.28	220.4	139.02	197.5
$number-of-reviews$	35.16	42.49	24.73	24.4	11.69
$review-scores-accuracy$	9.51	9.5	9.39	9.53	9.32
$review-scores-cleanliness$	9.36	9.37	9.46	9.35	9.22
$review-scores-checkin$	9.62	9.6	9.56	9.67	9.58
$review-scores-communication$	9.63	9.63	9.45	9.66	9.59
$review-scores-location$	9.64	9.69	9.7	9.57	9.48
$review-scores-value$	9.17	9.14	9.08	9.24	9.03
$reviews-per-month$	1.68	1.89	1.45	1.37	0.9
$Centro$	0.47	0.57	0.68	0.29	0.42

At first glance, we can see that, for most of the variables, the mean is distributed in similar manner among the different type of properties examined. However, we found variables that exhibited a high variability, to wit: price, maximum_nights, and availability_365. With respect to price, we observed that whereas the mean price in hotels is located around 150 euros, in the case of shared rooms is 65 euros. These differences could be anticipated considering the nature of each service. Furthermore, it can be observed how the mean of the variable availability_365 range from 161 days for the case of entire homes/apartments to 220 days in the case of hotels. Lastly, we also distinguished high variability among the means of maximum-nights ranging from 222 nights, in the case private rooms, to 274 nights in the case of entire homes/apartments.

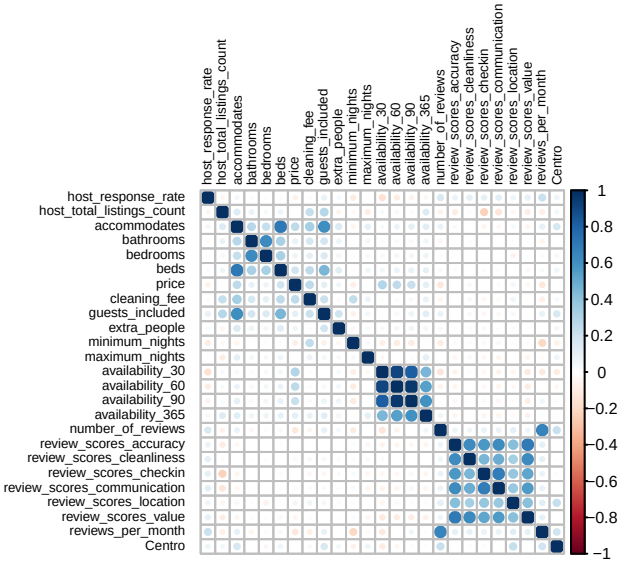
These findings are easily observable in the sample covariance matrix depicted below. By examining the graphical representation of the matrix, we can spot the variances of the mentioned variables in the diagonal. As it can be seen, the variances of these three variables stand out over the others.

On the other hand, by examining the sample correlation matrix, we can distinguish the existing correlation among some group of variables. As it could be expected, we found that the variables related to availability are highly correlated. This is due to the fact, that all these variables share information. Similarly, we also found that the variables related to reviews are also positively correlated. Moreover, we found as well natural correlations such as the number of bedrooms and the number of bathrooms.

Sample Covariance Matrix:

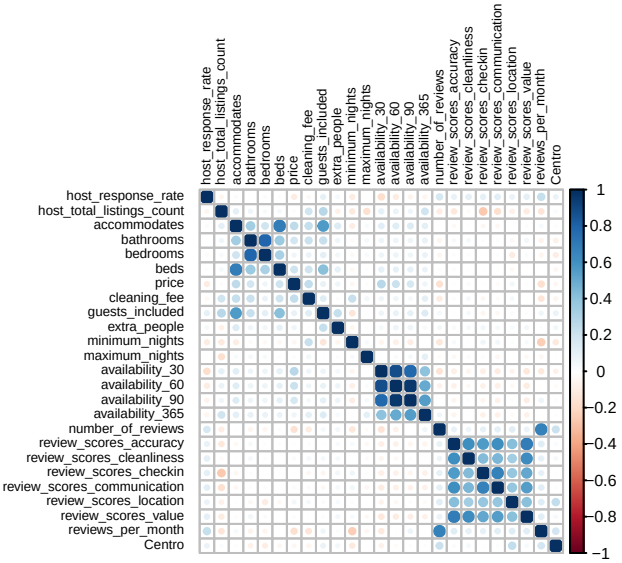


Sample Correlation Matrix:

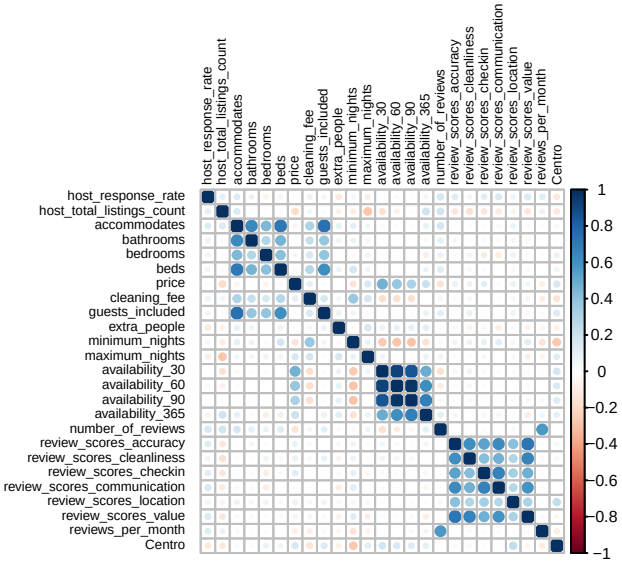


Sample Correlation Matrices per Room Type:

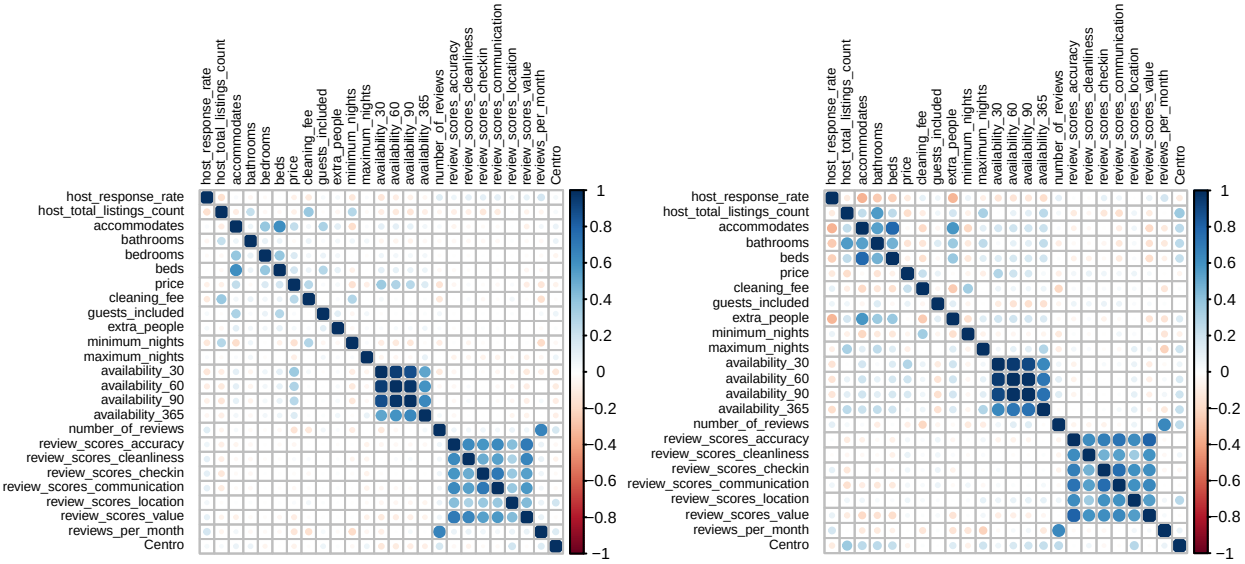
Entire home/apt



Hotel room



Private room **Shared room**

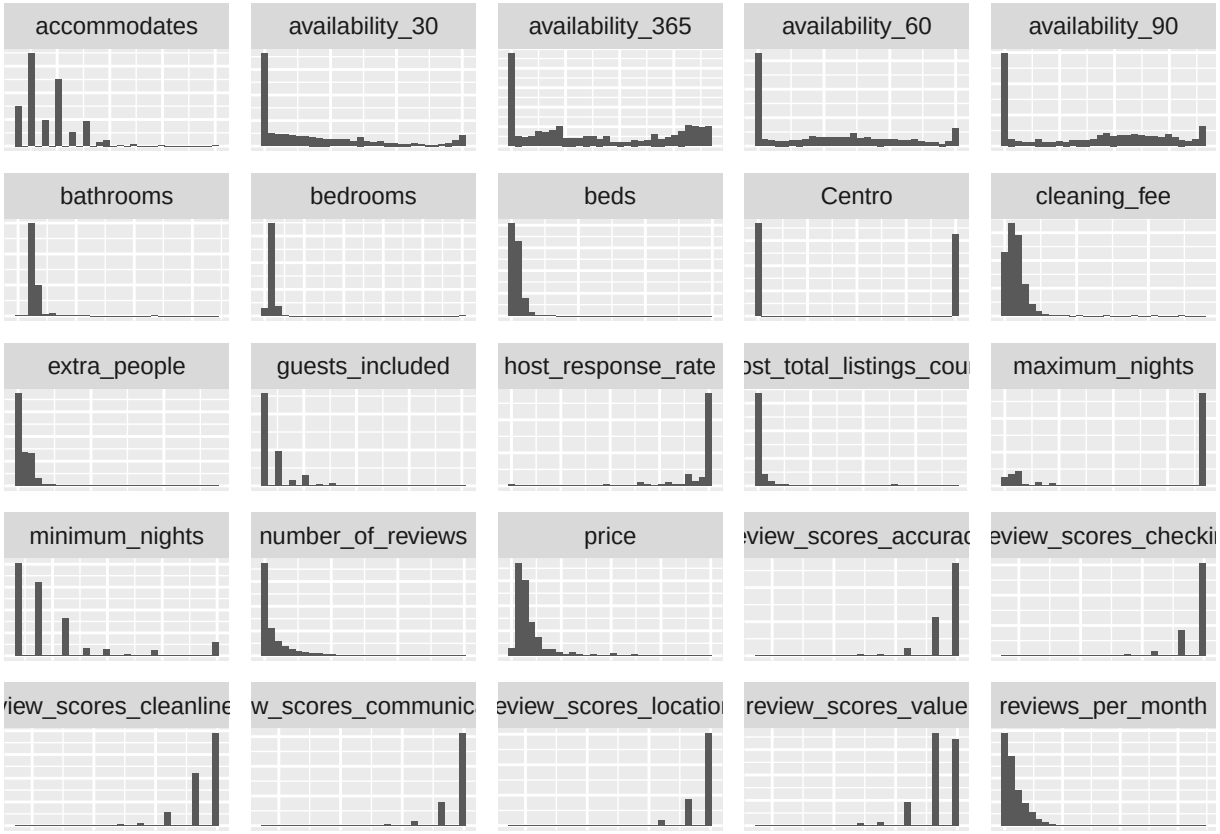


By examining the sample correlations discriminating by the type of property, we observed that features described earlier are present in all the of the cases as it could be expected. Notwithstanding, it must be highlighted an increment of negatively correlated variables in the case of shared rooms and hotels.

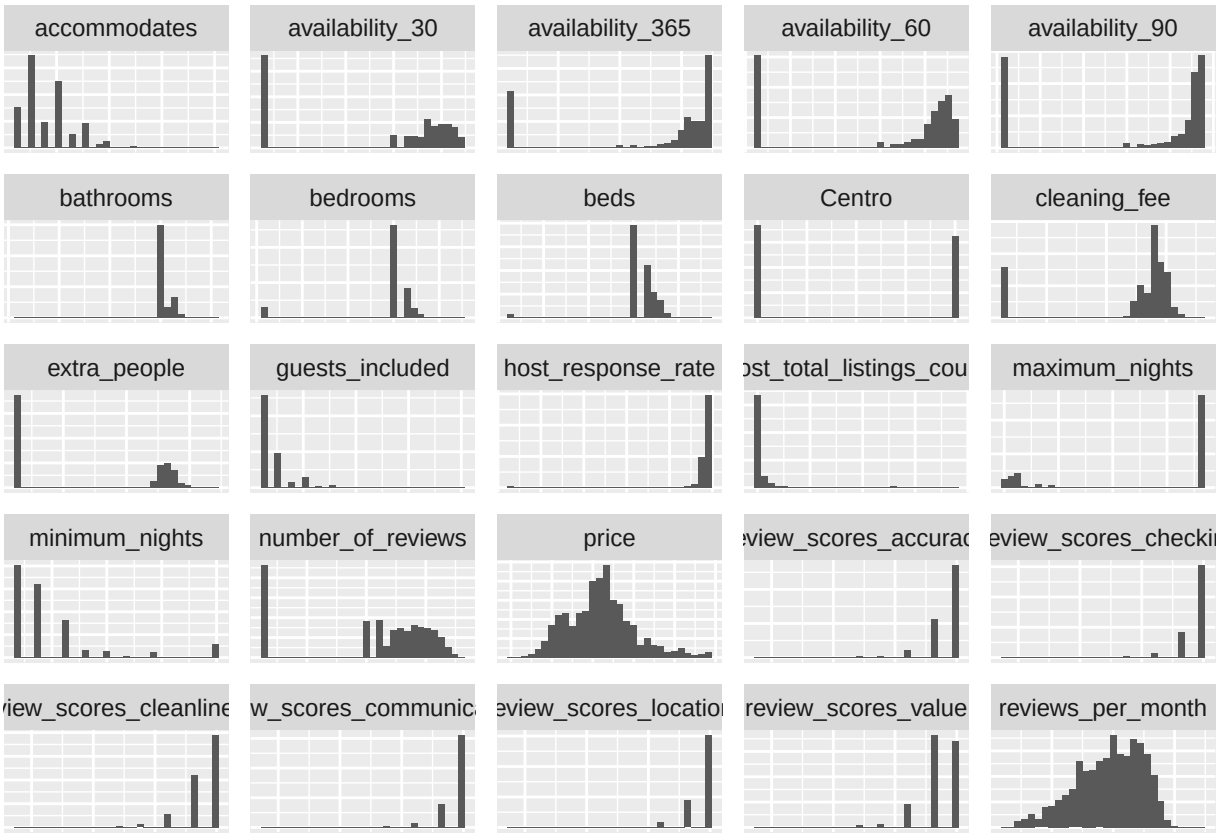
4. Principal Component Analysis

Visualizing the data and obtaining meaningful insights become more difficult and challenging when dealing with high number of variables as in our case. Because of this, we would like to obtain a low-dimensional representation of our data that provides as much information as possible. In order to accomplish this, we carried out a Principal Component Analysis (PCA). By employing this method, we were able to find a low-dimensional representation of the data that explain as much as possible of the variation in the data.

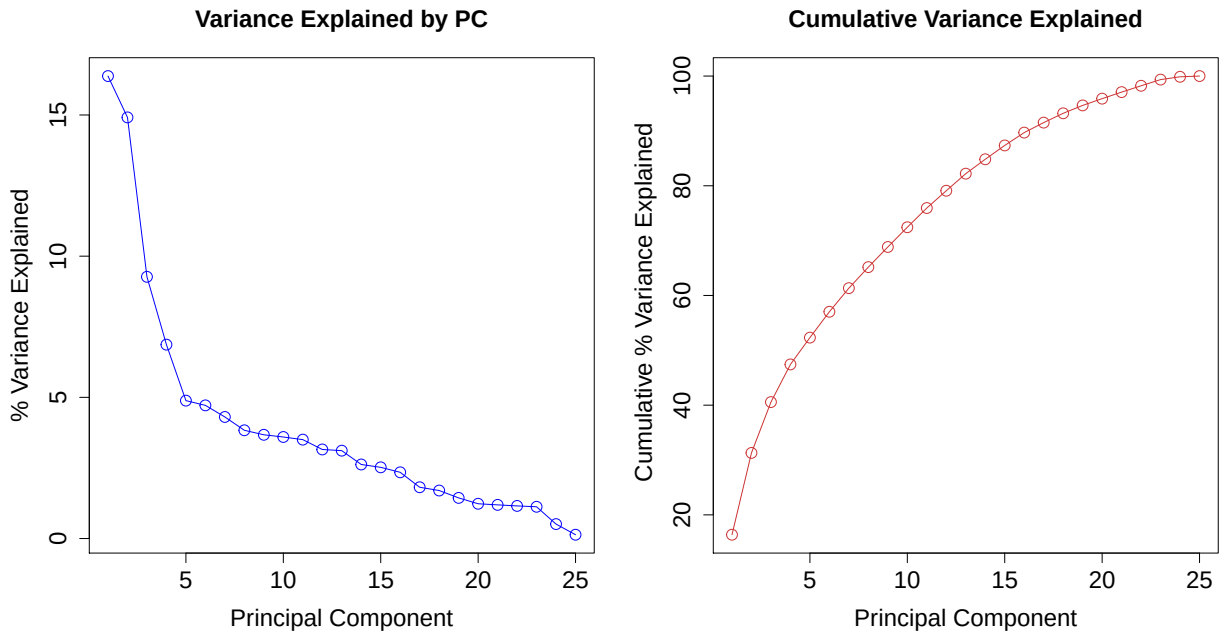
Before conducting the PCA, we assessed the normality of the variables. Then, transformed the data to push some variables closer to a normal distribution when needed. It can be observed from the graphs depicted that the transformation implemented yielded good results. Many of our variables were highly skewed before the transformation, whereas now, the distribution is more normal. Specially for the case of the variables price and reviews_per_month.



We can see below the distributions of the variables after imposing logarithmic transformation to the skewed ones.

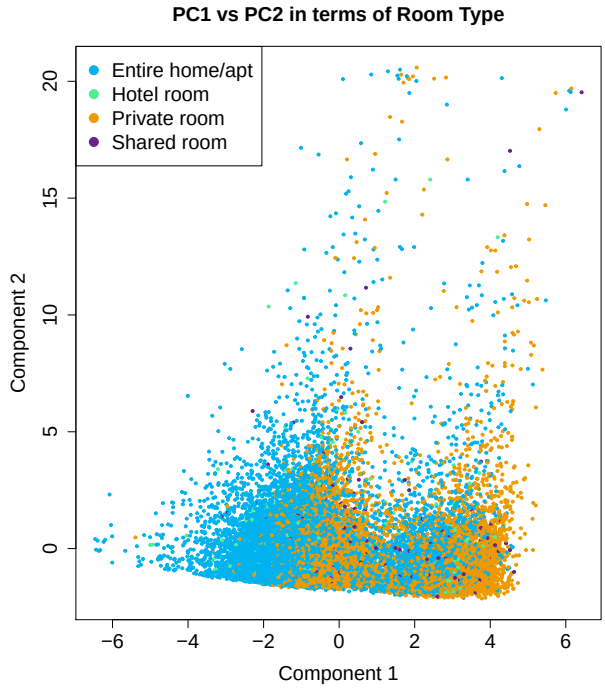


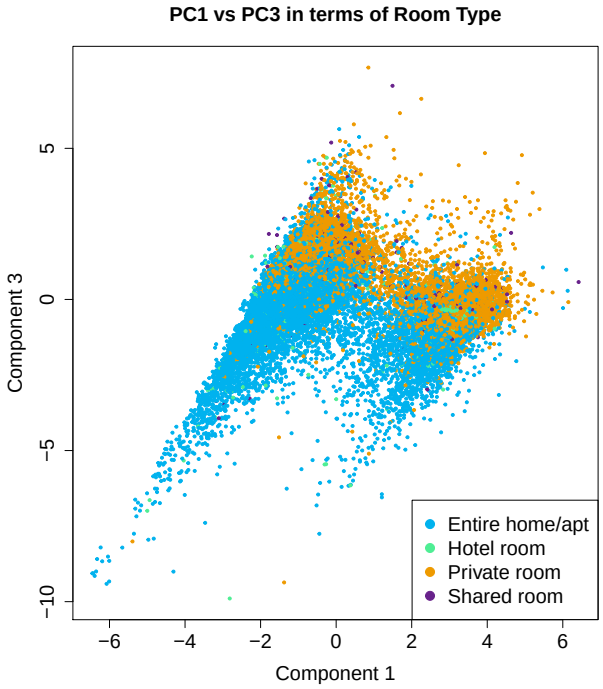
After completion of the transformation of the variables we can apply the PCA function *prcomp* to get the principal components. Below we present to graphs that show the amount of information given by each of the principal components. The graph on the left shows the percentage of the variance of the data explained by each of the principal components. We can see that the first one explains a bit more than 15%, and that the first 5 explain only 55% of the variability of the data. This is a sign that the data that we are working with is very complex and not easy to represent. The graph on the right presents the cumulative variability explained, we can clearly see that they approach 100%, as expected, and that the last PCs do not add much explanation of the data anymore.

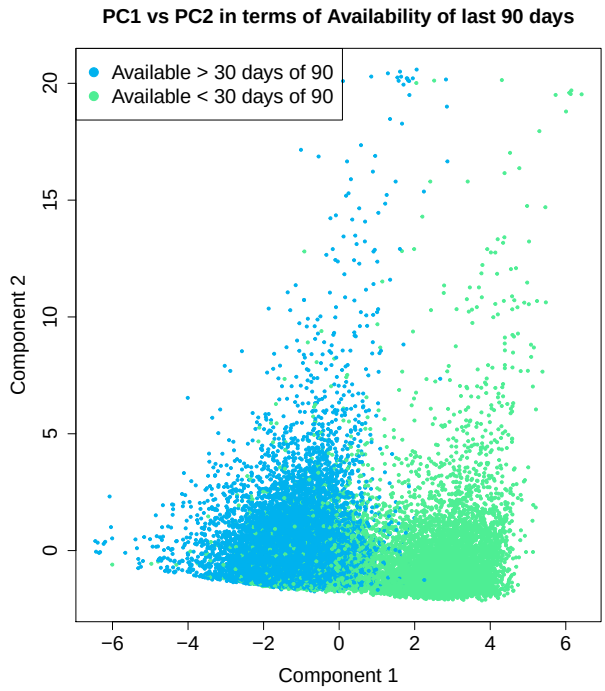


The figures below both are plots of the First vs. Second PC and First vs. Third PCs. We have differentiated them by *room_type* to show the differences in the 4 groups. As we will explain in the following pages, the these three principal components help us identify the main variables that cause the biggest changes between the groups, and they are the price, minimum nights, reviews and availability.

We see that the only variable that has a positive contribution to the first PC is minimum nights, that means that all the points that are to the left of the *x=0* line have no minimum nights. And the price variable has a negative contribution for the first PC while it has a positive one for the second PC. This explains the creation of 2 "streams" of points that go parallel, these are listings that have higher prices but are differentiated by the amounts of minimum nights they propose. It might be even more interesting the third graph, where we see the groups of listings with their availabilities more and less than 30 days out of the last 90 days. This variable serves as a clear identifier and we see that these two groups are then subdivided into smaller groups depending on the room type, we see that entire houses or private rooms both can have more or less than 30 out of 90 days available, and they are present in both clusters.

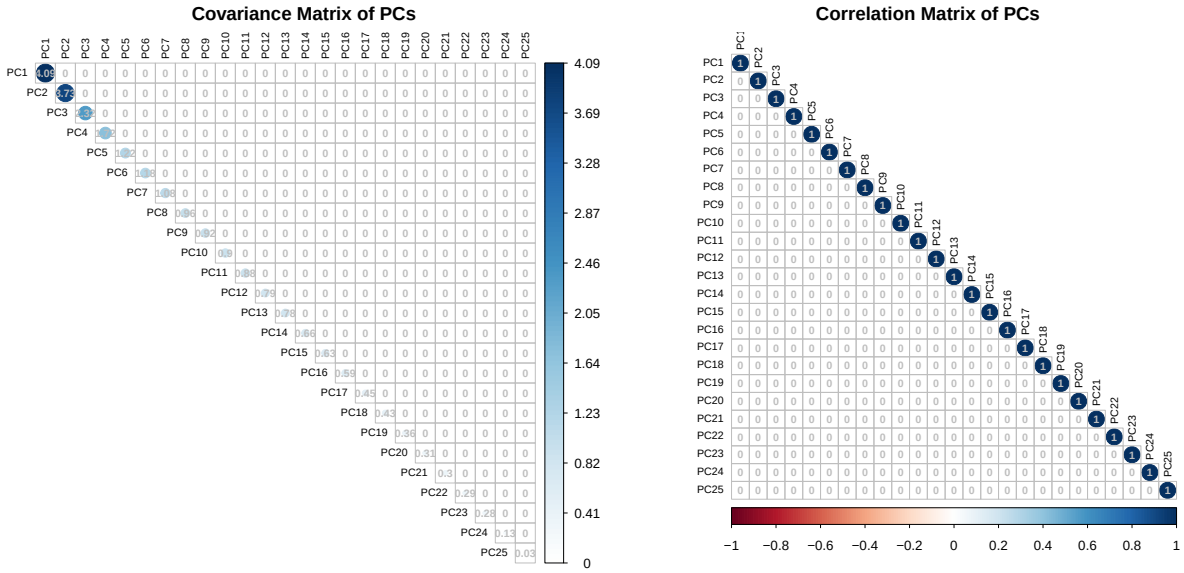




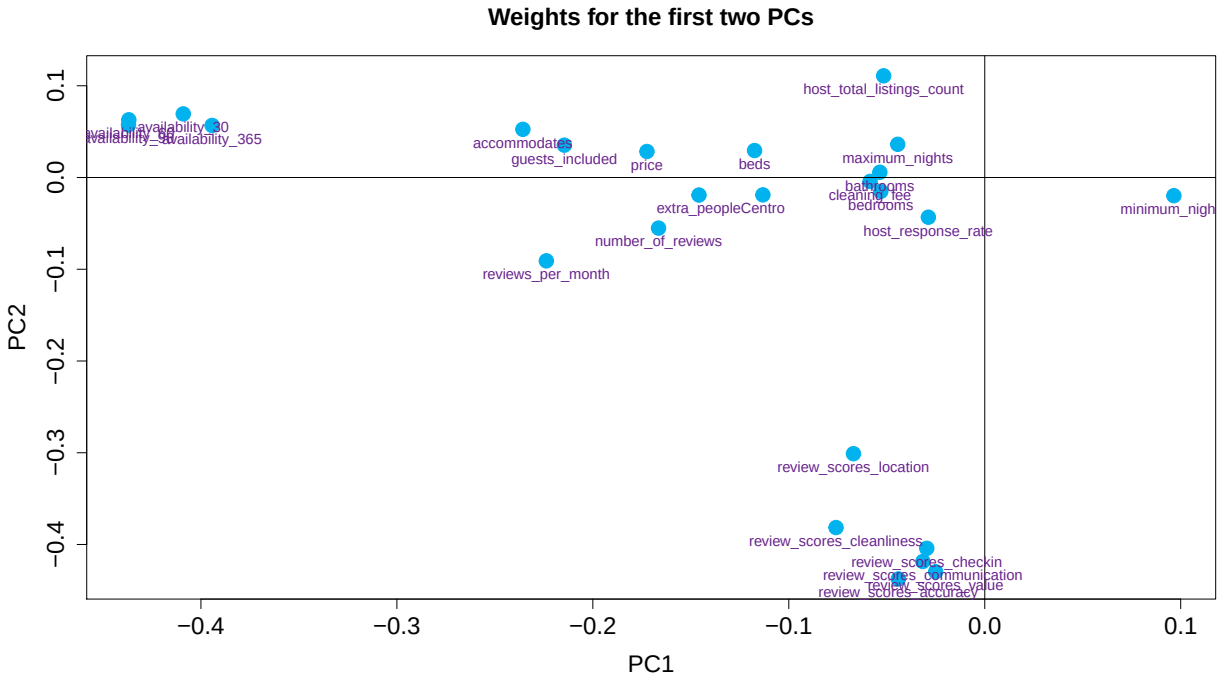


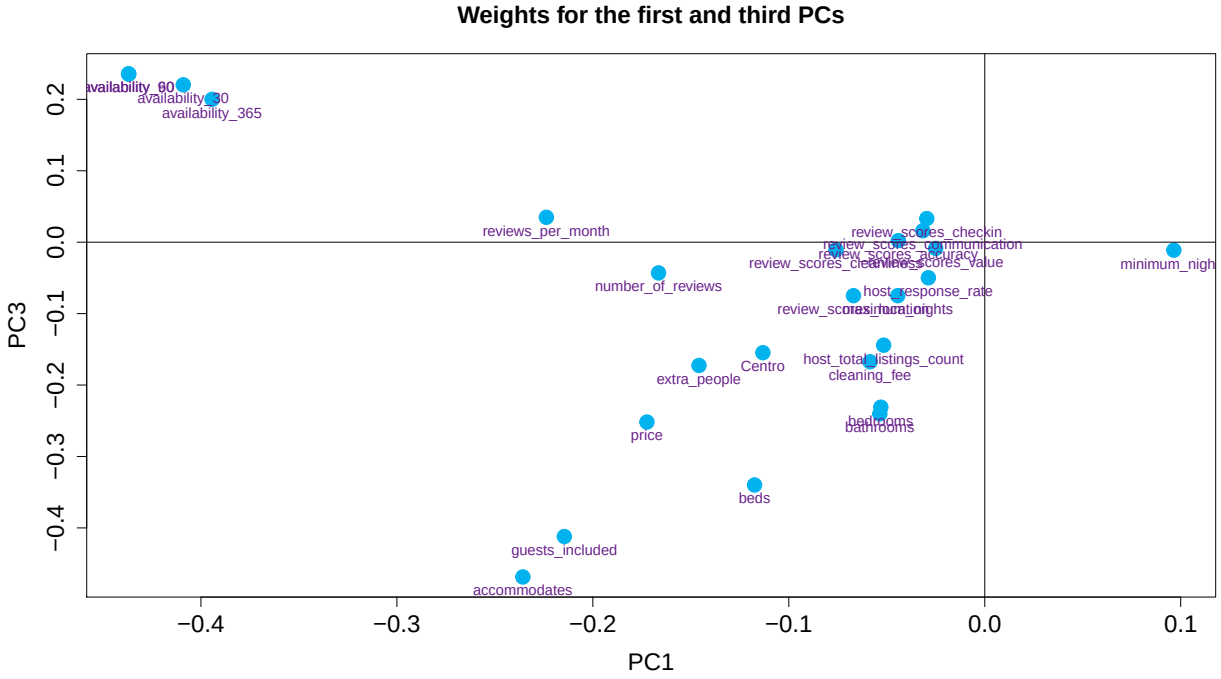
This last division is very informative, since listings with low availability are the ones that are normally more appealing for tourists of clients in general, thus they have a strong correlation with the price, which can be clearly seen in the plots in the following pages.

The two following graphs show the correlation between the principal components, we see that they do not have any correlation between them, which is one of the key concepts of PCA.



Now we can see the weights in terms of the initial variables to the principal components PC1-PC3, it is very useful to take into consideration this relations when looking at the PC1 vs PC2 plot shown above. With help of this plot we can also see which variables are correlated with with ones, since by making an arrow go from the origin to the variable name for every variable we can see if there exist negative, positive or no correlation at all.



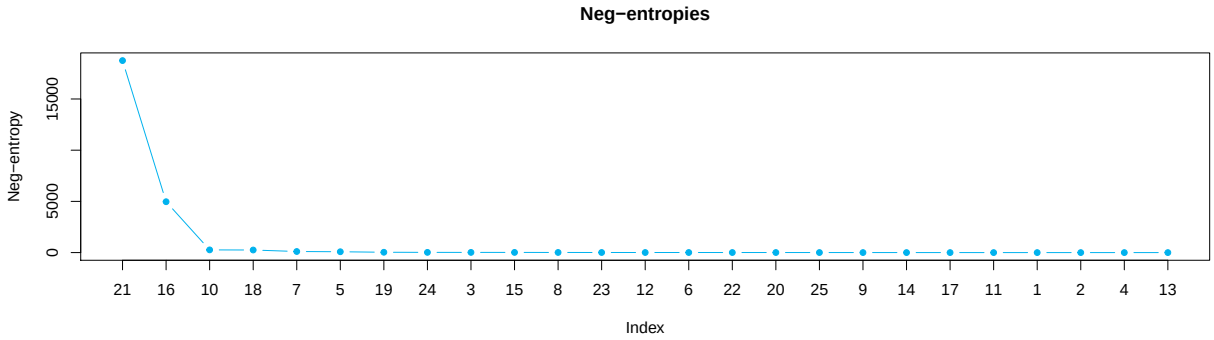


The plot of the weights suggests the existence of 2 groups of variables that behave in a similar manner and are well differentiated. This finding is coherent with the data structure as we have a group of variables related to review-scores and another one associated to availability. Furthermore, it seems to be another group formed by the variables *guest_included*, *price*, *bed* and *accomodates*.

We must highlight the fact that groups conformed by availability and price seem to be related whereas the reviews and price are orthogonal, which indicates that are not related. This is a very interesting fact indeed, since we would think that the better scores a listing has, the higher its price could be. According to these findings, reviews would not be connected to price, number of rooms and availability.

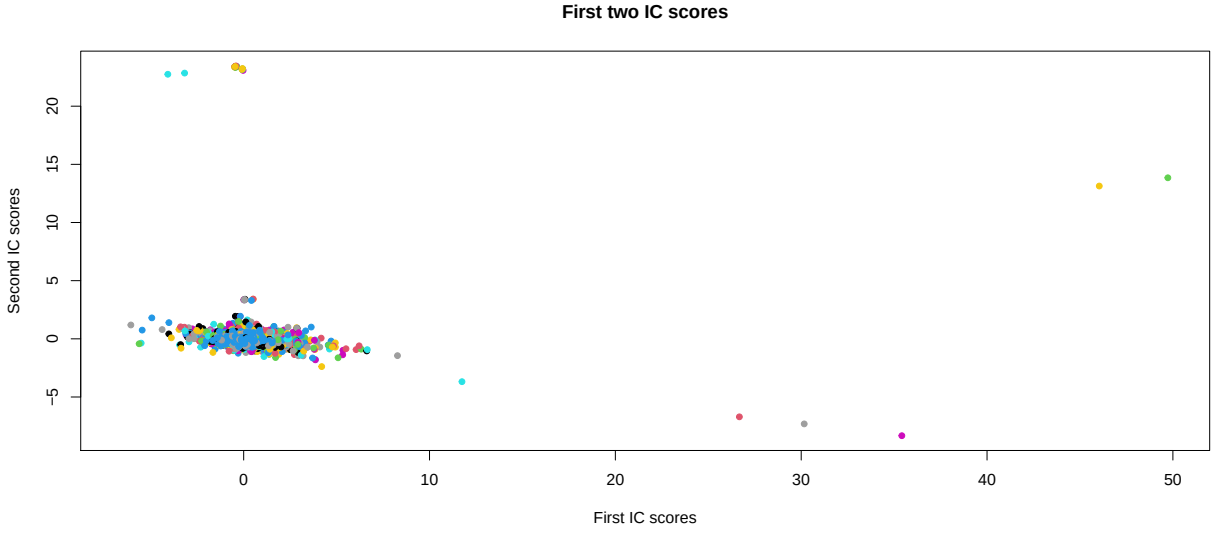
5. Independent Component Analysis

Finally, we carried out an independent component analysis in order to obtain independent non-Gaussian signals. Given that the ICA method does not require that the variables follow a Gaussian distribution, we have not used the previous transformed variables.

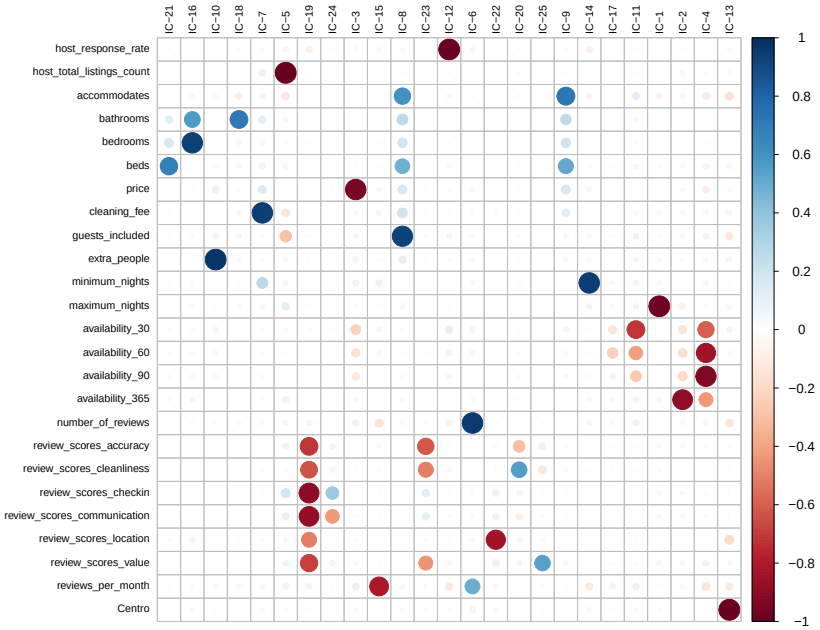


After having derived the Independent components, we observed the presence of high values of Neg-entropy. The IC 21 presented the highest Neg-entropy with almost 19000. Followed by IC 16 and IC 10, with Neg-entropy values 4966 and 264 respectively.

From the figure depicted, we can observe how the Neg-entropy value drops drastically from IC 21 to IC16 and from IC 16 to IC 10. Then, it flattens as goes from the IC 10 to IC 13 converging towards 0.



The high Neg-entropy values observed previously indicated the existence of non-Gaussian variables that could present outliers. The figure above confirms the presence of outliers and reveals them. They can be found in the top left and the bottom right of the graph.



In order to tackle the presence of outliers and assess their management, we examined the corrplot produced with the different independent components and the variables employed. The figure depicted above shows that the IC that yielded the highest Neg-entropy value is negatively related to the variable beds. However, after examining this particular variable, we found that although this variable is clearly non-Gaussian, the data examined is correct and therefore, not further action is needed.

Furthermore, we finally examined the Independents Components that yielded the lowest Neg-entropy values in order to look for groups. We can see that when splitting with $availability\_90 > 30$ or $availability\_90 < 30$ we see a clear distinction between the groups, reinforcing the theory that this variable serves as a clear differentiator between 'popular' and 'unpopular' AirBnB listings.

Conclusions

Throughout the report the process to visualize, analyze and obtain insights of a real world database has been demonstrated. We have started with a very raw data base from AirBnB listings in Madrid, and the goal of this assignment was to get an insight of the information and to try to obtain some conclusions about the correlations between predictors for each listing. In the Principal Component Analysis and Independent Component Analysis we have shown that the variables *availability*, *price* and *reviews* are key differentiators in the dataset. The *availability* variable gives the amount of days that the listing was free out of the 60, 90 or 365 days, depending on the variable.

It has been demonstrated that *availability* is very highly correlated with *price*, and that it can be seen as a measure of popularity. This due to the fact that the lower the availability an apartment or room has, the more successful this apartment or room is, and therefore the owners are able to increase the price per night.

In the pre-processing stage several steps had to be performed to get rid of variables whose percentages of missing values were too high to impute. Besides this some other variables that were not significant for the analysis have been removed, like images or text. After this the imputation of variables with help of the *mice* package was key for a standard normalization of the data.

Step 3 of this report demonstrates that the numerical differences between *room types* are notable, and therefore there exists a real difference between the groups. The mean vector per groups can be seen as a key indicator to split the data in these 4 clusters: one per room type. Furthermore, we can see that the correlation matrices of all groups are very similar to each other in most cases. This is, however, expectable since the listing's variables are also related between them.

The results of PCA and ICA in steps 4 and 5 are very useful to see the relationships between the variables and to see if there are any correlations between them. Moreover, PCA can serve as a dimension reduction technique. In this case the reduction of dimensions, however, does not seem feasible as the variance explained per principal component was not very high, only reaching around 55% after 5 principal components. To approach 80% of the total variance explained a high number of 15 principal components is needed.

Finally, we have seen that the listings depend on many variables, and that some of them might be correlated between them, like *price* and *guest_included*. But the main finding of this report is that *availability* plays a major role when splitting the data into two main clusters that can be named as "popular listings" and "unpopular listings". This result could not have been obtained without the help of principal and independent component analysis.