2022

Insights into Dutch Airbnb

Wesley van der Horst (2713430), Nick Peetam (2704149) & Nando Beijaard (2624391)

Project Big Data, Group 7

3/7/2022

VU

Table of Contents

Introduction

In this report, we will use datasets containing several .csv files provided to us by Airbnb. We will use the Airbnb listings and review files from the three most populated cities in the Netherlands: Amsterdam, Rotterdam, and The Hague. The data files contain thousands of rows of data and more than fifty columns. These columns consist of binary, string, and integer data types. The listing data set was cleaned to get rid of, what seemed to us, useless columns. An example of a useless column would be the URL of the profile picture of the host. To further clean the data set and make it ready for use we transformed multiple columns into different data types. This was mainly necessary for the dates; the provided dates were all in string format instead of the DateTime format which is easier to work with as well as making lists of the columns amenities and host_verifications.

This report will contain the following chapters: Data Analysis, Models and Conclusion. In “Data Analysis” we will analyse the different data sets and we will try to find correlations between columns and try to find the best performing city/neighbourhoods/room type. The correlations will be used to predict what kind of models we can use in “Models”. In “Models” we will elaborate on the machine learning models we have used. We will also provide the test scores on the different models and an explanation as to why and how a model (conclusion) could be used by hosts and customers. In “Conclusions and Discussion” we will give a brief summary and discussion of the notable findings of the data analysis and the conclusions of the models.

Data analysis

Amsterdam has around 285,000 reviews which cover about 80% of the total amount in the datasets considered, because of this all average reviews for each of the 22 neighbourhoods are valid. Since Rotterdam and The Hague do not have this number of reviews some neighbourhoods only have two reviews, which leads to absurdly high or low average review scores that make them invalid. To avoid this, we set the minimum number of reviews to fifty for our neighbourhoods. In each city are numerous neighbourhoods, however, we have chosen to only show the five highest and lowest scoring neighbourhoods since showing them all would make the tables inconveniently long.

In Figure 1 below, the boxplots of the prices for each city are shown. Two observations are left out since these two listings in Amsterdam had prices of over six thousand dollars and including them would make the plot nearly useless. As can be seen, there are many observations outside of the IQR for each of the boxplots with Amsterdam seemingly having the most, which makes sense with the city making up for about 80% of the data. The boxes themselves are quite similar with the means being in a close range of each other. However, the y-axis scale is quite large which makes it hard to observe this. To make this easier these means are plotted in Figure 2.

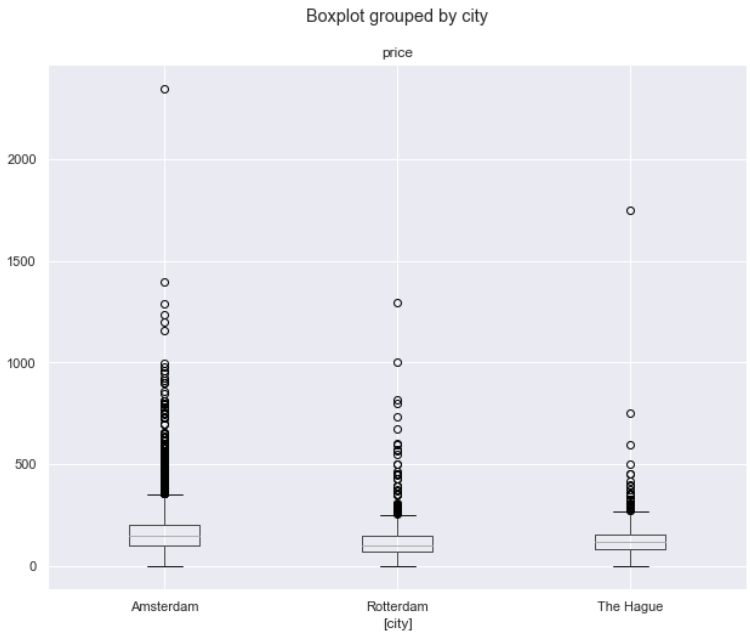

Figure 1: Boxplot of listing prices, with two of the most extreme outliers left out.

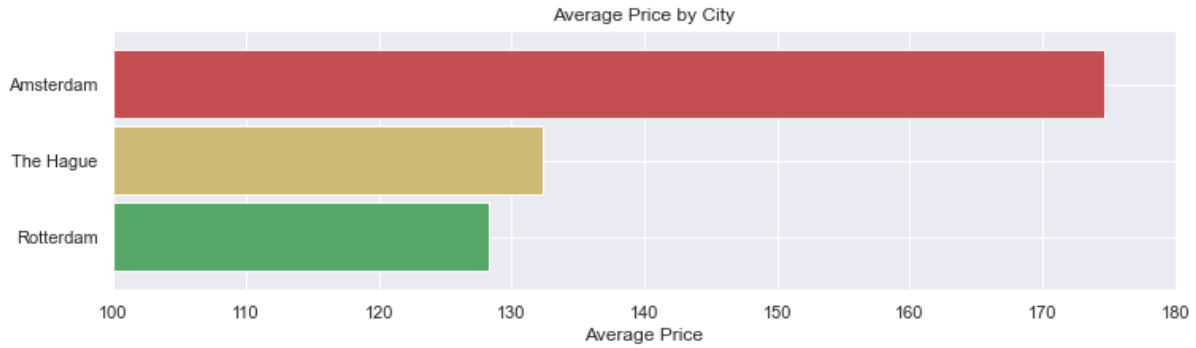

Figure 2: Bar plot Average Price by City

About the listings

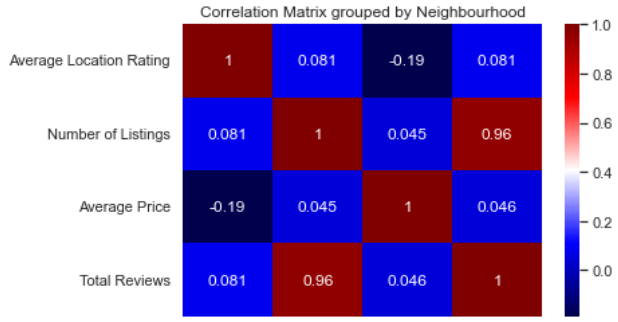

Figure 3: Correlation matrix all data grouped by neighbourhood

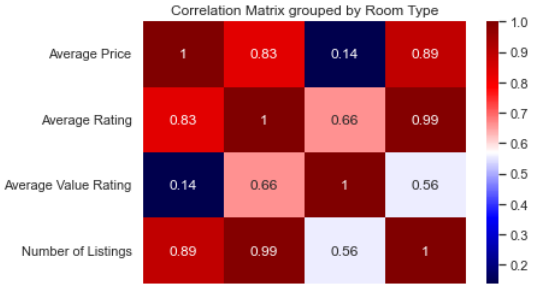

Figure 4: Correlation matrix all data grouped by room type

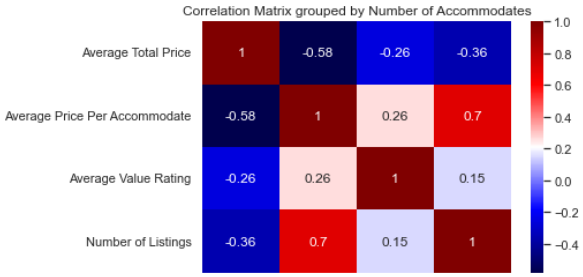

Figure 5 Correlation matrix all data grouped by accommodates

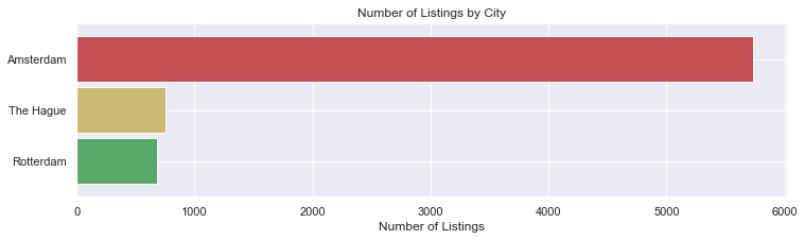

Figure 6: Bar plot numbers of listing by City

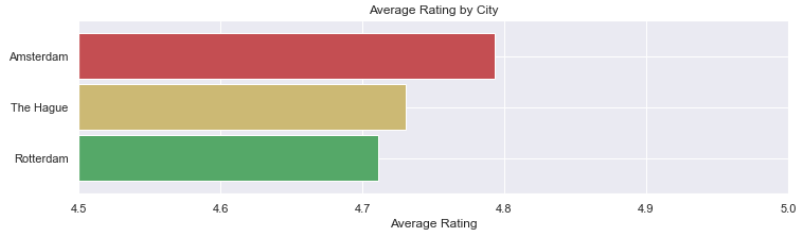

Figure 7: Bar plot average rating by City

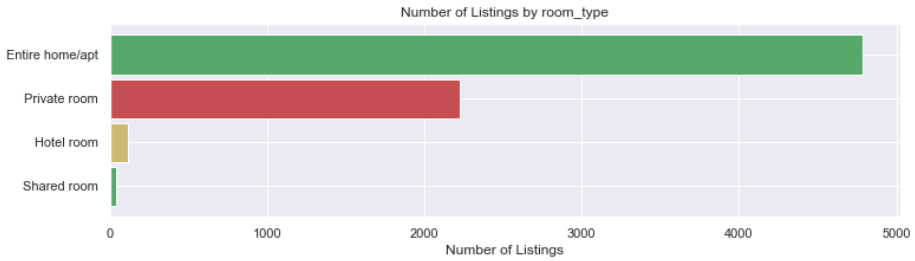

Figure 8: Bar plot number of listing by room type

Table 1: Column aggregations per city and of all listing datasets combined

						Frequency of Room type			
City	**Number of Listings**	**Avg Rating**	**Average Price**	**Average Amenities**	**Total Reviews**	**Entire Home /Apt**	**Private Room**	**Hotel Room**	**Shared Room**
Amsterdam	5,732	4.79	$ 174.61	27.5	284,140	3,828	1,792	88	24
The Hague	753	4.73	$ 132.31	30.14	30,336	522	217	11	3
Rotterdam	679	4.71	$ 128.28	29.75	28,504	438	215	14	12
Total	**7,164**	**4.78**	**$ 165.78**	**27.99**	**342,980**	**4,788**	**2,224**	**113**	**39**

As illustrated in Figure 6, by far the most listings are in Amsterdam. This is to be expected since the capital has a lot more appeal to non-natives looking for a fun vacation, compared to The Hague and Rotterdam. However, the average ratings for each city are comparable all being in the 4.7-4.8 range, with Amsterdam on top, as shown in Figure 7. Considering all this, it is no surprise that the average price is also the highest in Amsterdam. On average the price for a stay in Amsterdam is a third more expensive than a stay in Rotterdam or The Hague. From Table 1**,** we can also conclude the room types for each city are proportioned evenly around each city. From which the proportion is illustrated in Figure 8.

About the hosts

Table 2: Column aggregations per city and of all hosts in listings datasets combined

City	**Hosts**	**Active Hosts**	**Avg Host for (days)**	**Average Response Rate**	**Average Acceptance Rate**	**Superhosts**	**Hosts w/o profile picture**	**Identity verified hosts**
Amsterdam	4,607	3,668	2,591	93.96%	74.50%	1,118	15	3,655
The Hague	440	316	2,356	94.12%	80.71%	157	2	354
Rotterdam	474	298	2,320	94.55%	78.34%	136	2	372
Total	**5,518**	**4,279**	**2,549**	**94.04%**	**75.41%**	**1,410**	**19**	**4,379**

For the data about the hosts in each city (Table 2), the most remarkable is the average acceptance rate in Amsterdam. Which is the lowest out of the three cities. When we compare the number of hosts to the number of listings, we see that on average the hosts in The Hague have more listings than the hosts in Amsterdam.

Amsterdam

Table 3: Top and bottom 5 neighbourhoods in Amsterdam by Location Rating

Neighbourhood	Average Location Rating	Number of Listings	Average Price	Total Reviews
Centrum-West	4.92	861	$ 194.20	62,168
Centrum-Oost	4.87	609	$ 214.11	40,607
De Pijp - Rivierenbuurt	4.85	563	$ 183.25	26,024
Zuid	4.81	371	$ 190.16	16,834
De Baarsjes - Oud-West	4.80	836	$ 179.69	34,705
…				
Buitenveldert – Zuidas	4.59	73	$ 150.38	2,392
Bijlmer-Centrum	4.52	37	$ 116.14	2,169
De Aker - Nieuw Sloten	4.51	64	$ 142.80	4,519
Gaasperdam - Driemond	4.49	46	$ 102.74	2,143
Osdorp	4.43	43	$ 124.28	3,838

Table 3 shows neighbourhoods in Amsterdam sorted on average location rating with their corresponding values for listings, average price and the total number of reviews. It comes to no surprise that the neighbourhoods around the city centre seem to score the highest ratings within the city as well as having many listings and high total reviews and average price due to being in the most popular part of the city. This all seems to imply a strong correlation between all these columns which is confirmed by Figure 9.

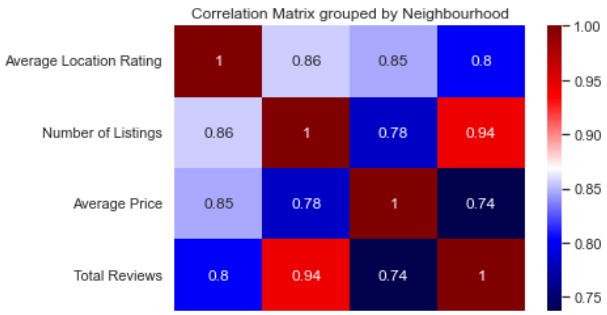

Figure 9: Correlation Matrix of Table 3

All columns have a strong positive correlation with each other with 0.74 being the lowest which makes sense considering Amsterdam's tourism and near-centre location importance.

The average price, overall rating and value rating of the different room types in Amsterdam can be found in Table 4 below. Once again there seems to be a correlation between the columns however less significant this time especially when looking at the value column.

Table 4: Aggregations on Amsterdam listings grouped by Room type

Room Type	Average Price	Average Rating	Average Value Rating	Number of Listings
Entire home/apt	$ 197.89	4.82	4.66	3,828
Private room	$ 128.87	4.74	4.62	1,792
Hotel room	$ 121.43	4.70	4.54	88
Shared room	$ 72.71	4.66	4.55	24

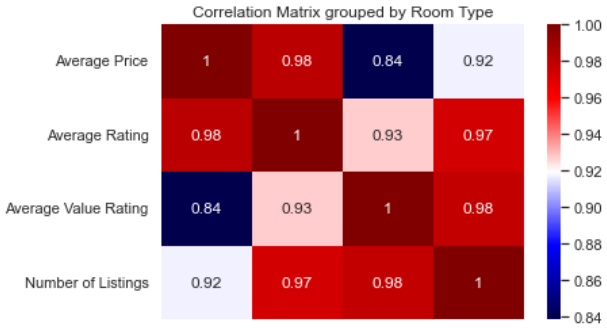

Figure 10: Correlation matrix of Table 4

In Table 5 below, the average total price and average price per accommodate for each number of allowed accommodates for all listings are shown. One would think that there would be some linear relationship between the number of accommodates and the total price however when looking at Table 5 that does not completely seem the case. Listings with 10 and 9 accommodates seem to be way cheaper than expected and thus also way cheaper per person than most of the others, however, they both do have a very small amount of 2 and 3 listings respectively which could play a part in this. By average price per Accommodate after 3 accommodates it does not seem to improve much until reaching 9 and 10 accommodates.

Table 5: Aggregations of Amsterdam listings grouped by the number of accommodates.

Number of Accommodates	Average Total Price	Average Price Per Accommodate	Average Value Rating	Number of Listings
0	$ 0.00	NaN	4.76	6
1	$ 71.70	$ 71.70	4.64	155
2	$ 139.41	$ 69.71	4.66	2,996
3	$ 155.21	$ 51.74	4.64	433
10	$ 200.00	$ 20.00	4.67	2
4	$ 220.45	$ 55.11	4.62	1,806
5	$ 249.83	$ 49.97	4.59	106
9	$ 262.33	$ 29.15	4.41	3
6	$ 308.33	$ 51.39	4.58	153
7	$ 338.81	$ 48.40	4.55	21
12	$ 390.70	$ 32.56	4.50	10
8	$ 402.06	$ 50.26	4.60	17
14	$ 445.67	$ 31.83	4.48	3
16	$ 596.24	$ 37.26	4.52	17
15	$ 629.00	$ 41.93	5.00	1
13	$ 654.00	$ 50.31	4.00	3

The correlation matrix, in figure 11, of the different columns of table 5 reveals that the average total price has a negative correlation with all other columns and the only somewhat significant positive correlation is between the number of listings and price per accommodate.

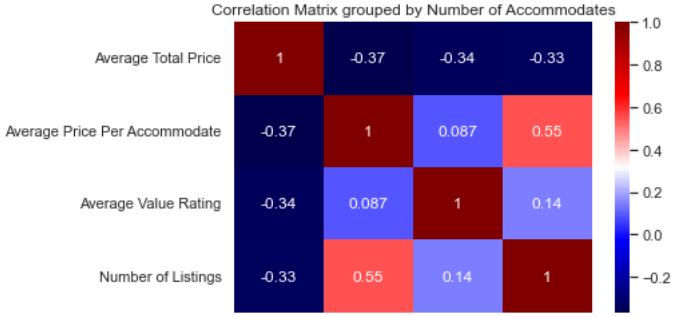

Figure 11: Correlation matrix of Table 5

The Hague

Table 6: Top and bottom 5 neighbourhoods in The Hague by location rating.

Neighbourhood	Average Location Rating	Number of Listings	Average Price	Total Reviews
Archipelbuurt	4.94	8	$ 153.25	458
Voorhout	4.94	20	$ 124.45	659
Zeeheldenkwartier	4.94	26	$ 150.04	1,388
Rijslag	4.93	17	$ 142.59	336
Nassaubuurt	4.93	7	$ 119.43	392
…..				
Rivierenbuurt-Zuid	4.58	3	$ 172.33	740
Vliegeniersbuurt	4.57	3	$ 77.67	242
Uilennest	4.54	3	$ 91.67	103
De Uithof	4.49	3	$ 90.00	470
Erasmus Veld	4.28	4	$ 46.75	663

From Table 6**,** one cannot easily see the different correlations per column. From this small preview in the data, there may exist a correlation between location rating and price. A visualisation of all the correlations can be found in Figure 12 below. One can see that the only notable correlation is the correlation between listings and the total number of reviews, this correlation is expected since an increase of listings would, logically, increase the number of reviews. The average location rating and the number of listings also have a small correlation, as well as the number of listings and the average price.

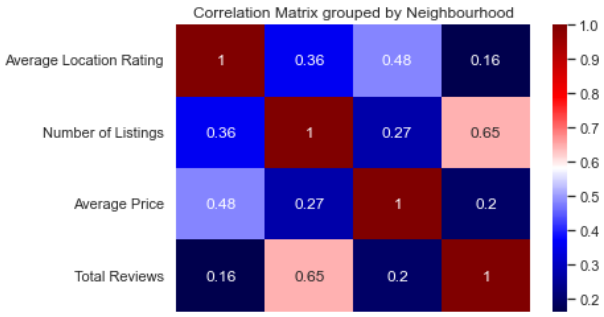

Figure 12: Correlation matrix of Table 6

Next, we will analyse the average price, average rating, average value rating and the number of listings per room type. It appears that the average price and rating correlate. This suspicion is verified by the correlation matrix in figure 13. From this figure you can also see that this is the case for average price and number of listings, average price and average value rating, and average rating and average value rating. For the former, this may be caused by the fact that there simply are a lot of entire home/apartment listings (which are often the most expensive).

Table 7: Aggregations of The Hague listings grouped by Room type.

Room Type	Average Price	Average Rating	Average Value Rating	Number of Listings
Entire home/apt	$ 147.28	4.73	4.61	522
Hotel room	$ 129.00	4.74	4.32	11
Private room	$ 97.88	4.72	4.65	217
Shared room	$ 28.33	4.56	4.67	3

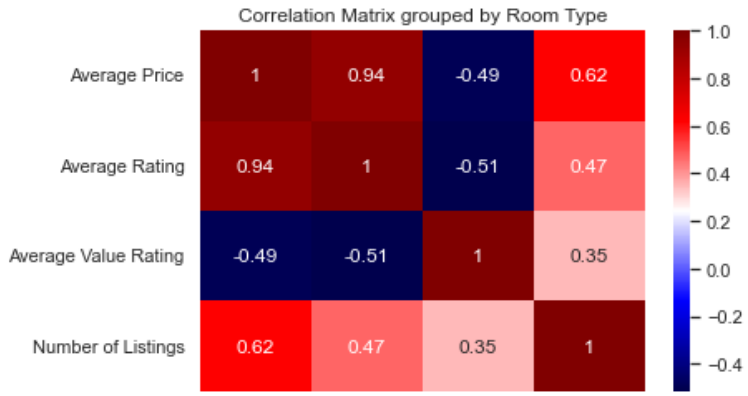

Figure 13: Correlation matrix of Table 7

Next, we will look at the effect the number of accommodates has on some of the previous columns. You can see in table 8 that from 0 until 6 accommodates, there is an increase in the total average price. 0 accommodates seems to be a mistake in the data set, since renting a place for 0 accommodates is rather unnecessary. From 7 accommodates, the price seems rather random. This may be caused by the little number of listings for these number of accommodates (so one listing has a large influence on the average price). From this table, you can draw the conclusion that, if you want to minimize cost per person, you should search for a listing with 14 accommodates. This is also confirmed by the correlation matrix in figure 14, in which the average total price and the average price per accommodate has a correlation of 0.88. All other columns do not have high enough correlations to be notable.

Table 8: Aggregations of The Hague listings grouped by Number of accommodates.

Number of Accommodates	Average Total Price	Average Price Per Accommodate	Average Value Rating	Number of Listings
0	$ 0.00	NaN	4.73	1
1	$ 47.00	$ 47.00	4.63	41
2	$ 101.04	$ 50.52	4.62	321
3	$ 101.68	$ 33.89	4.56	60
4	$ 145.14	$ 36.28	4.62	181
5	$ 184.68	$ 36.94	4.67	62
6	$ 207.95	$ 34.66	4.66	55
8	$ 228.85	$ 28.61	4.45	13
9	$ 282.25	$ 31.36	4.66	4
7	$ 286.18	$ 40.88	4.65	11
10	$ 324.50	$ 32.45	4.91	2
14	$ 357.00	$ 25.50	4.19	1
16	$ 1,750.00	$ 109.38	4.40	1

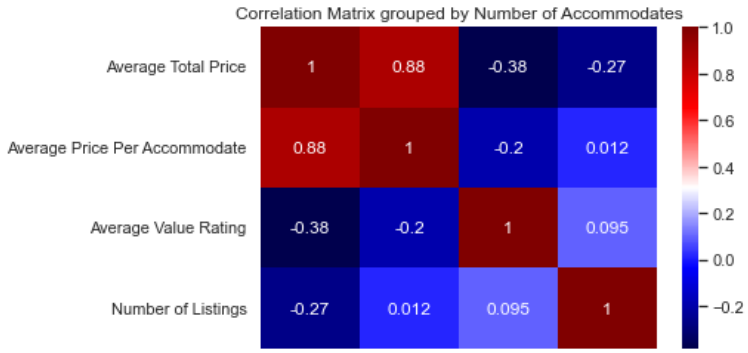

Figure 14: Correlation matrix of Table 8

Rotterdam

Table 9: Top and bottom 5 neighbourhoods in Rotterdam by location rating.

Neighbourhood	Average Location Rating	Number of Listings	Average Price	Total Reviews
Kralingse Bos	4.97	2	$ 108.50	258
Schieveen	4.95	2	$ 81.50	293
Cool	4.94	47	$ 158.89	2,610
Stadsdriehoek	4.93	59	$ 160.78	4,621
Provenierswijk	4.92	9	$ 84.67	488
…..				
De Esch	4.62	6	$ 93.33	109
Oud Charlois	4.53	9	$ 148.00	370
Hillesluis	4.49	7	$ 104.71	1,751
Oud Mathenesse	4.28	10	$ 122.60	64
Carnisse	4.15	6	$ 76.83	64

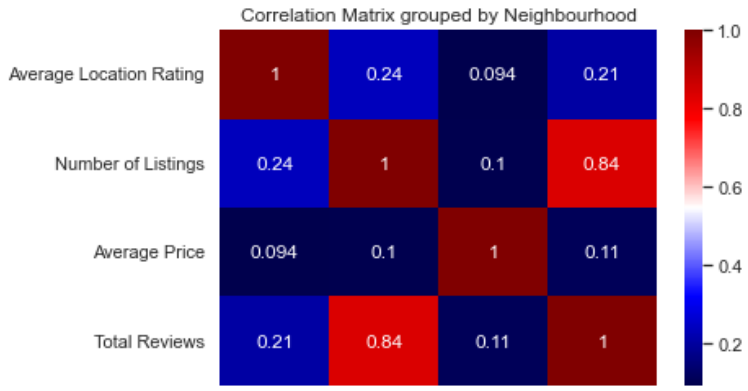

Figure 15: Correlation matrix of Table 9

From table 9 you cannot immediately see if there is a correlation between any of the columns. The only correlation which can easily be seen is the correlation between total reviews and listings. This correlation is quite arbitrary. All correlations can be found in figure 15 above. Our previous finding of the biggest correlation between total reviews and listings is correct. The next biggest correlation is the correlation between price and the average location rating. All other correlations are too small to confidently say that there exists a correlation.

The average price and average rating of room type in Rotterdam can be found in table 10 below. You can see that there does not seem to be a correlation between the price and the rating of the room types. The two other columns show the average rating of the "value" category in Airbnb and the number of listings, both per room type. Looking at the correlation matrix in figure 16 one can deduce that there exists a correlation between the average rating and the average value rating. This is in line with the conclusion from the earliest data analysis where we found that the different ratings all correlate with each other. The average price per room type and the number of listings also seem to correlate. As well as the average rating and the number of listings.

Table 10: Aggregations of Rotterdam listings grouped by Room type

Room Type	Average Price	Average Rating	Average Value Rating	Number of Listings
Entire home/apt	$ 143.40	4.77	4.68	438
Hotel room	$ 133.00	4.36	4.48	14
Private room	$ 99.14	4.61	4.59	215
Shared room	$ 92.83	4.75	4.77	12

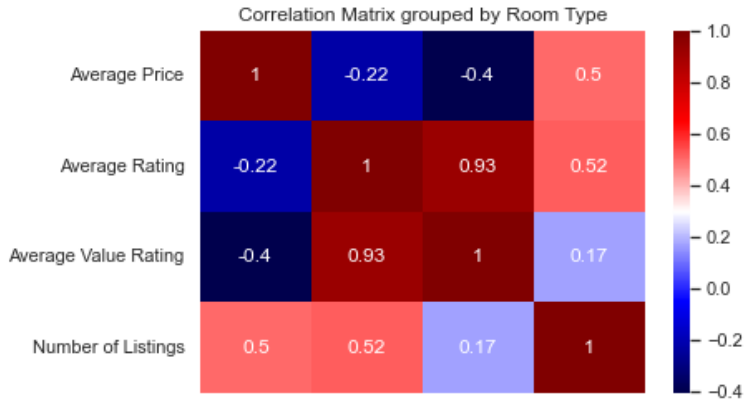

Figure 16: Correlation matrix of Table 10

In table 11 the number of accommodates and the total / average price can be found. You can see that the total price seems to steadily increase with the number of accommodates until you hit ten accommodates, from there the price seems to be random. This may be due to the fact that there are simply not a lot of listings with 10+ bed spaces, this has the effect that you only take the average of a small number of listings. One listing has a large influence on the average. Whereas for two till six accommodates, there are a lot of listings. For the average price per accommodate, you see that a listing with larger numbers of accommodates is cheaper per accommodate, this is in line with "economy of scale". The correlation matrix, in figure 17, of the different columns of table 11 reveals that the only large correlations are between the average total price and the average value rating and between the average price per accommodate and the average value rating.

Table 11: Aggregations of Rotterdam Listings grouped by Number of Accommodates

Number of Accommodates	Average Total Price	Average Price Per Accommodate	Average Value Rating	Number of Listings
0	$ 0.00	NaN	4.81	1
1	$ 47.24	$ 47.24	4.69	42
2	$ 107.94	$ 53.97	4.66	344
3	$ 124.75	$ 41.58	4.66	55
5	$ 140.21	$ 28.04	4.74	14
4	$ 146.72	$ 36.68	4.69	145
6	$ 186.73	$ 31.12	4.47	45
8	$ 201.06	$ 25.13	4.65	18
7	$ 275.00	$ 39.29	4.54	1
16	$ 284.43	$ 17.78	4.65	7
9	$ 327.50	$ 36.39	4.62	2
13	$ 429.00	$ 33.00	4.83	1
10	$ 550.50	$ 55.05	4.42	2
12	$ 708.50	$ 59.04	2.90	2

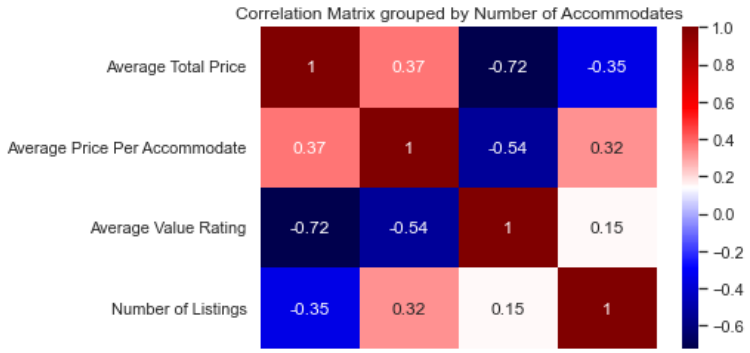

Figure 17: Correlation matrix of Table 11

Difference between cities

There are a lot of similarities and differences between the three biggest cities in the Netherlands. The biggest differences room type-wise are the higher average prices for entire homes/apartments and private rooms in Amsterdam compared to those listings in Rotterdam and the Hague. This may be due to higher overhead costs in Amsterdam. For example, higher house prices. There also seems to be more correlation between multiple columns in Amsterdam compared to Rotterdam and the Hague.

Another difference is the fact that Amsterdam has much more listings than Rotterdam and the Hague, this in itself causes most data to equalise to a normal equilibrium. For Rotterdam and the Hague, a problem arises where some listings have a high influence on some averages.

In figure 18 the number of reviews per month is plotted per city. You can see that Rotterdam and the Hague have a much more defined peak than Amsterdam (the Hague even more than Rotterdam). This peak is around the summertime, indicating that the number of reviews around the summer vacation period is much higher than the number of reviews in other months. Whereas the peak of Amsterdam in the summer period is much less defined. Indicating that, compared to the Hague and Rotterdam, Amsterdam has more demand all year round.

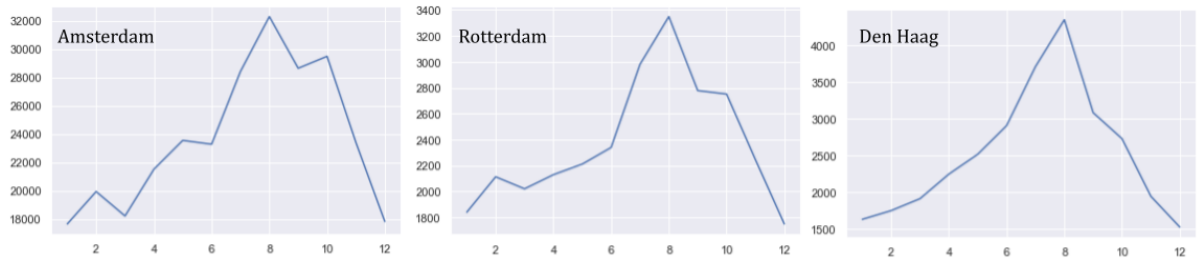

Figure 18: number of reviews per month by city

Models

Clustering Model Attempt

As we believed that price and number of reviews could possibly show popular locations by latitude and longitude in a clustering model, we tried to achieve this by creating a K means clustering model. We applied this model using the cleaned data of the listings in Amsterdam as we already saw that central neighbourhoods were amongst the most popular and pricier locations in the city and fit the model on the price and number of reviews for 3 clusters, resulting in figure 19 plotted on latitude and longitude. Although the yellow cluster does seem to be more prominent around the centre, we were disappointed with the clusters not being clearer and more segregated by location.

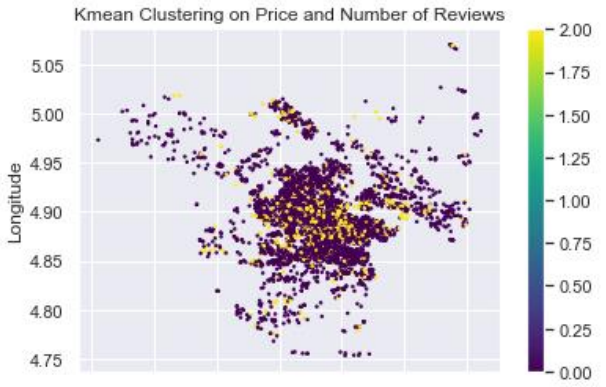

Figure 19: k-means clusters on price and number of reviews in Amsterdam

Regression models

After performing cross-validation and ending up with a linear regression model on rating using the other rating factors giving little insight other than accuracy having the largest correlation with the overall ratings. Despite having a decent determination coefficient and an RMSE of 0.1512, high factor ratings simply imply a higher overall rating, thus we wanted a different approach and different response variable.

This led to choosing the price as the dependent variable in future models. First, we had to prepare the data for regression, we dropped irrelevant columns, used one hot encoding where applicable also for the entries of the host verifications lists and removed NAs. This resulted in a 4,321 x 247 Data Frame of the three cities combined. After not getting any decent results in any of the methods described below, we decided to use apply the natural logarithm to the prices and use this log price as the dependent variable making the distribution look quite normally distributed as can be seen in Figure 20 below, with skewness and kurtosis of 0.1675 and 0.3602, respectively.

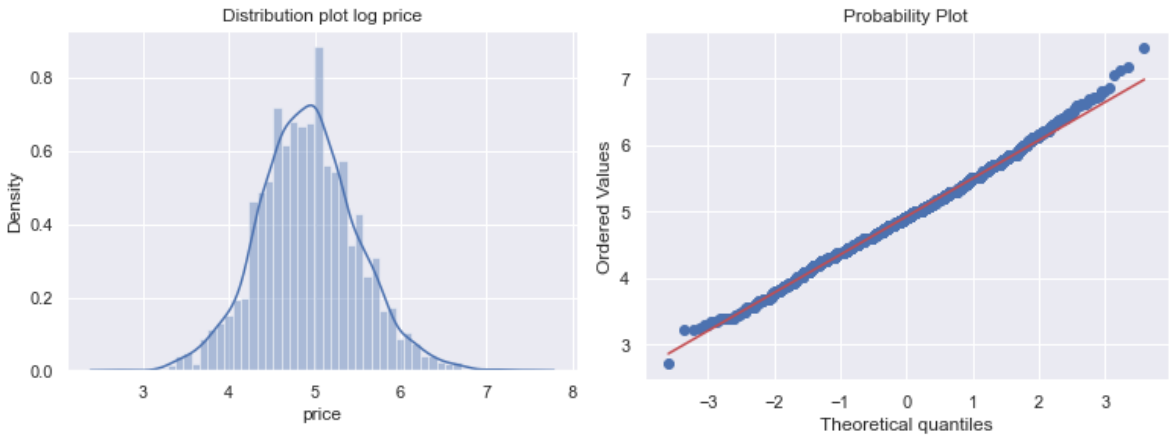

Figure 20: Distribution plot log of the price and probability plot

First, we tried the L2 regularization method using the Ridge model and found by grid search cross validation an optimal alpha of 10,000,000,000 which seemed rather large, using this alpha and testing a new model with it resulted in nearly all predicted values being between 4.85 and 5.15 and thus still having a relatively low RMSE of 0.5476. Reverting all price values of the data and prediction back by taking the natural exponent shows larger differences and an RMSE of 114.

Afterwards, we set up several models for testing with and without polynomial features of degree 2, however, after observing the processing times and results of the models using the polynomial features, we decided to leave them out since the models took a long time to train as well as not having significantly better results worth the time difference. The regression models chosen for testing were: Lasso, Ridge Random Forest and Extra Trees after Linear regression already performing very poorly. From the cross-validation using the "regression_results" function, we get that the Random Forest seems to be the best performing of the lot with a mean RMSE of 0.0851 over 10 cross-validations. However, the residuals do not seem very normally distributed as can be seen in Figure 21, which often is one of the assumptions for regression models. Extra Trees seems to perform notably worse with an RMSE of 0.2372 but does have more normally distributed residuals.

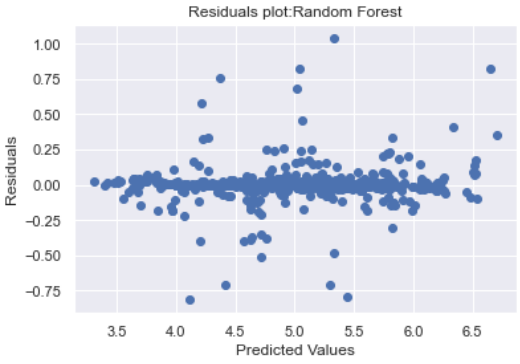

Figure 21: Residuals of Random Forest22

Classification models

Host verify identification by text

A classification we tried to make is a classification model using the columns: description and neighborhood_overview. These two columns are strings used to describe the accommodation and the location. We tried to see if there is a correlation between the words used in these columns and the fact that a host is a super host/ is verified. The two text columns are first put into one column. To be able to use this textual data, we must first convert it to numbers. This will be done using TfidfVectorizer. TfidfVectorizer will be used in a pipeline with a multinomial Naïve-Bayes method (MultinomialNB). An assumption of the MultinomialNB method is that there exists independence among predictors. We will thus assume that the description and neighbourhood location text columns

are independent of each row. The model is trained using 80% of the data. We get f1 scores of all listings in table 12 with multiple different models:

Table 12: f1 score of "host verified" and "host super host" using multiple models

Model	F1 score "host verified"	F1 score "host super host"
No Kfold, CountVectorizer	0.8676	0.4056
Kfold, CountVectorizer	0.8678	0.3059
No Kfold, TfidfVectorizer	0.8946	0
Kfold, TfidfVectorizer	0.9010	0

We can see that for all listings, the model using Kfold model selection and a TfidfVectorizer text extraction has the best f1 score (in the case of host verified). The model does not seem to work in the classification of whether someone is a super host or not.

Room type

We created several classification models than only used the three columns 'price', 'bedrooms' and 'host_acceptance_rate'. The classification models used for classification are: Logistic Regression, Decision Tree, Random Forest and Multinomial Naïve Bayes. Random Forest ended up having the highest mean F1 score of around 0.75, which is not great but certainly better than complete randomness. Then, similarly to the preparation of the data for regression, data was prepared for classification on room type. When using the classification readied data and the models mentioned before in cross-validation Random Forest once again came out on top with an F1 score of 0.8733 which is already significantly better than the previous one.

Conclusions and Discussion

From investigating the cities, themselves and the datasets combined, we got quite some interesting insights and some which confirmed information that could be assumed beforehand. Such as the popularity of the city centre of Amsterdam.

In this basic data set, it was quite difficult to find correlations. Correlation scores were either too small or were arbitrary (in the rating case). Due to the lack of correlations, it was difficult to easily build machine learning models. The lack of good correlations in the data set was one of the main culprits that held back good conclusions for this report. It was interesting how the correlation matrices of the aggregation tables for each city did differ quite a lot at times.

We did, however, find some working models, a successful model is a model in which we would predict whether a host was verified or not depending on the words used in the "description" and "neighborhood overview" text cells. Another successful model was the classification model in which we wanted to predict room type based on a set of variables. We also tried to fit a linear regression model to predict the log of the price of a listing. This linear regression was quite successful. We, however, think this is because the values for the log of the price lay quite close to each other (irrespective of different variables), causing the error to be small. We will thus not count this model as one of the successful conclusions of this report.

Most conclusions of this report can be drawn from the data analysis chapter, where all correlations can be found. For example, the correlation in Amsterdam between all columns in figure 9 shows how interconnected the average location rating, the number of listings, the average price and the total reviews are in the different neighbourhoods in Amsterdam. The same could not be said for Rotterdam and the Hague where these correlations were almost non-existent.