Business Intelligence and Data Mining

May 2023

Opportunities and Challenges of Sharing Economy: Airbnb

A focus on Edinburgh, Scotland

A Data Mining Exercise

Adedayo Adewole 30810670

Table of Contents

Introduction

The emergence of peer-to-peer systems and platforms, collectively known as the "sharing economy", has allowed individuals to collaboratively make use of under-utilized assets via fee-based sharing. This sharing economy is sometimes called the “collaborative economy” (Calo and Rosenblat, 2017). The term refers to online network-based activities that provide temporary access to a good to facilitate more efficient use of physical assets. The suppliers in these markets are often small (mostly individuals), and they actually set out to share excess capacity that would have gone unutilized hence why the term “sharing economy". According to Kyle Barron et all (2017), Economic theory suggests that the sharing economy improves economic efficiency by reducing frictions that could cause capacity to go underutilized, hence why the explosive growth of sharing platforms such as Uber for ride-sharing and Airbnb for home-sharing which testifies to the underlying demand for such markets globally. Airbnb is one of the pioneers of this sharing economy and has had a significant impact on the tourism and hospitality industry as well as on destination economies, the experience of a place by a guest. Within a few years of its inception in 2009, Airbnb had become one of the most successful sharing economy platforms. A statement in Airbnb's 2022 Q4 report reveals 2022 was another record year for Airbnb as their Revenue of $8.4 billion grew 40 percent year over year while Net income stood at $1.9 billion. However, there are claims that this growing success has created more social problems. Airbnb’s disruption of the accommodation industry has created challenges for governments as regulators because it has changed the tourism landscape, creating taxation problems and discrimination problems among others. According to Cheng et al., 2020; Martín et al., 2018, Tourist sites have been subjected to negative externalities based on the increased concentration of tourists in particular spots, which invites environmental problems such as water scarcity, waste management, and carbon emission issues. Also, Dogru et al., 2019 explains that hoteliers have faced great threats from Airbnb which owns no property giving an example of one hotel in Texas that has continued to experience revenue loss for every increase in Airbnb property listings. "Why should I let out my house for a fixed term when the sharing economy platforms have provided a means to earn more?", a property owner will say. As a manner to encourage the growing platform and also reflect prices of the hospitality sector, price inefficiencies exist between letting out houses in the accommodation industry and listing the houses on the sharing platform- Airbnb. This has created opportunities for property owners, who may prefer to operate under the Airbnb platform for profit maximisation. According to Clever Real Estate (2021), a real estate firm in Maryland USA, Airbnb is fast becoming the preferred choice of vacationers- 60% of travellers who use both Airbnb and hotels prefer Airbnb over comparable hotels when going on vacation and it doesn’t look like the supply of Airbnb hosts will slow down as 54% of home owners said they’d consider renting out their homes with

Airbnb or a similar vacation rental app, and 82% believe that Airbnb is a good way to make money from their property.

Even though Airbnb has implied on their site that the hosts are to provide "temporary accommodation" and/or “tourism related activities”, we also see from their site that there is information on "Responsible hosting" for each country to reflect local laws which are put to combat this menace. "Responsible hosting in the United Kingdom" encourages hosts in London to abide by the Deregulation Act of 2015 which introduced an exception that allows you to use residential premises for short-term rental for 90 or fewer nights in a calendar year. This is known as the “90-night rule”. If Airbnb is truly a temporary accommodation provider as they state, why will they refer to each local law to guide on number of nights which can be provided by hosts? Our Data mining goal is to examine and ascertain if the claim about Airbnb being a social menace is true, and if Airbnb is truly creating a social problem and disrupting the accommodation industry/communities by offering houses to people who should be taking up permanent accommodation. In doing this, we will focus on Edinburg, the capital of Scotland and make use of Data that has been provided which pertains to Hosts in this city.

Airbnb Listings (Edinburgh) Data Understanding

We intend to conduct a Data mining/analysis exercise of Airbnb listings as provided to find meaningful patterns, clusters and build a classification model that would differentiate hosts and/or listings that provide temporary accommodation or illegal accommodations that are more permanent thereby posing a social menace. The Data provided is that of Airbnb listings of Edinburgh, Scotland, United Kingdom retrieved from Inside Airbnb (insideairbnb.com), an independent and non-commercial open-source data tool. The Data includes Listings, Reviews, Calendar data as well as a Data dictionary.

We have a total of 7,389 listings with 75 variable columns which represent the raw data scrapped from the web and each listing represents information about a property and its host.

	id	listing_url	scrape_id	last_scraped	source	name	description	neighborhood_overview	
0	15420	https://www.airbnb.com/rooms/15420	20221216161317	2022-12-16	city scrape	Georgian Boutique Apt City Centre	Stunning, spacious ground floor apartment minu...	The neighbourhood is in the historic New Town,...	https://a0.musc
1	707097	https://www.airbnb.com/rooms/707097	20221216161317	2022-12-16	city scrape	Centre Royal Mile Apartment 3 bedrooms 2 bathr...	NaN	The location is the perfect for tourism , shor...	https://a0.musc
2	728199	https://www.airbnb.com/rooms/728199	20221216161317	2022-12-16	city scrape	Private room in central, spacious and comfy flat	Fantastic main door flat over two levels withi...	Great location for access to the city centre, ...	https://a0.musc
3	732008	https://www.airbnb.com/rooms/732008	20221216161317	2022-12-16	city scrape	51 18 Caledonian Crescent	This beautiful third floor apartment is set in...	NaN	https://a0.mus
4	744710	https://www.airbnb.com/rooms/744710	20221216161317	2022-12-16	city scrape	Refurbished Flat in a Georgian Era Building in...	A stunning apartment in the heart of Edinburgh...	The apartment is in a Central Edinburgh neighb...	https://a0.mus
...	...	...	...	...	...	...	...	...	
7384	362653	https://www.airbnb.com/rooms/362653	20221216161317	2022-12-16	previous scrape	Papermill Wynd Chiara Apartment,	The space Located just off McDonal...	NaN	https://a0.musc
7385	363642	https://www.airbnb.com/rooms/363642	20221216161317	2022-12-16	city scrape	Wonderful/beautiful room Edinburgh	The space Room over looking the Pe...	Close to the city centre. But surrounded by g...	https://a0.musc
7386	376937	https://www.airbnb.com/rooms/376937	20221216161317	2022-12-16	previous scrape	Dario Apartment, Edinburgh	The space Dario is a new modern ap...	NaN	https://a0.musc
7387	388297	https://www.airbnb.com/rooms/388297	20221216161317	2022-12-16	city scrape	Stunning villa by Holyrood Park - parking & gar...	On the doorstep of Arthur's Seat and Holyrood ...	NaN	https://a0.musc
7388	389318	https://www.airbnb.com/rooms/389318	20221216161317	2022-12-16	city scrape	City centre. Walk to Castle, Museum etc. Own b...	The Space Situated in a quiet Mews courty...	Southside is a bustling, vibrant, local commun...	https://a0.musc

7389 rows × 75 columns

Figure 1- listings data set snapshot

From the exploratory data analysis conducted with Python (Jupyter) on the raw data, we observe that of the 75 variable columns we have 35 object variables, 17 float variables and 23 integer variables. We also observe that there are several missing variables. In addition, columns like "neighbourhood_group_cleansed", "bathrooms" and "calendar_updated" had no variables in them (null) hence will need to be discarded.

Exploring further, the summary statistics reveals that variables like host_listings_count, host_total_listings_count are positively skewed with extreme outlier values. Also with bedrooms, beds, minimum_nights, minimum_minimum_nights, maximum_minimum_nights, minimum_nights_avg_ntm, calculated_host_listings_count, calculated_host_listings_count_entire_homes which all had their median values quite less than the mean value while having outlier values. For example, a listing had 40 bedrooms and another (or possibly the same listing) had 40 beds, could that be a mansion listed or an error? Based on

domain knowledge, Airbnb listings are usually not that big with many rooms. Box plots below show the distribution and outliers of the variables.

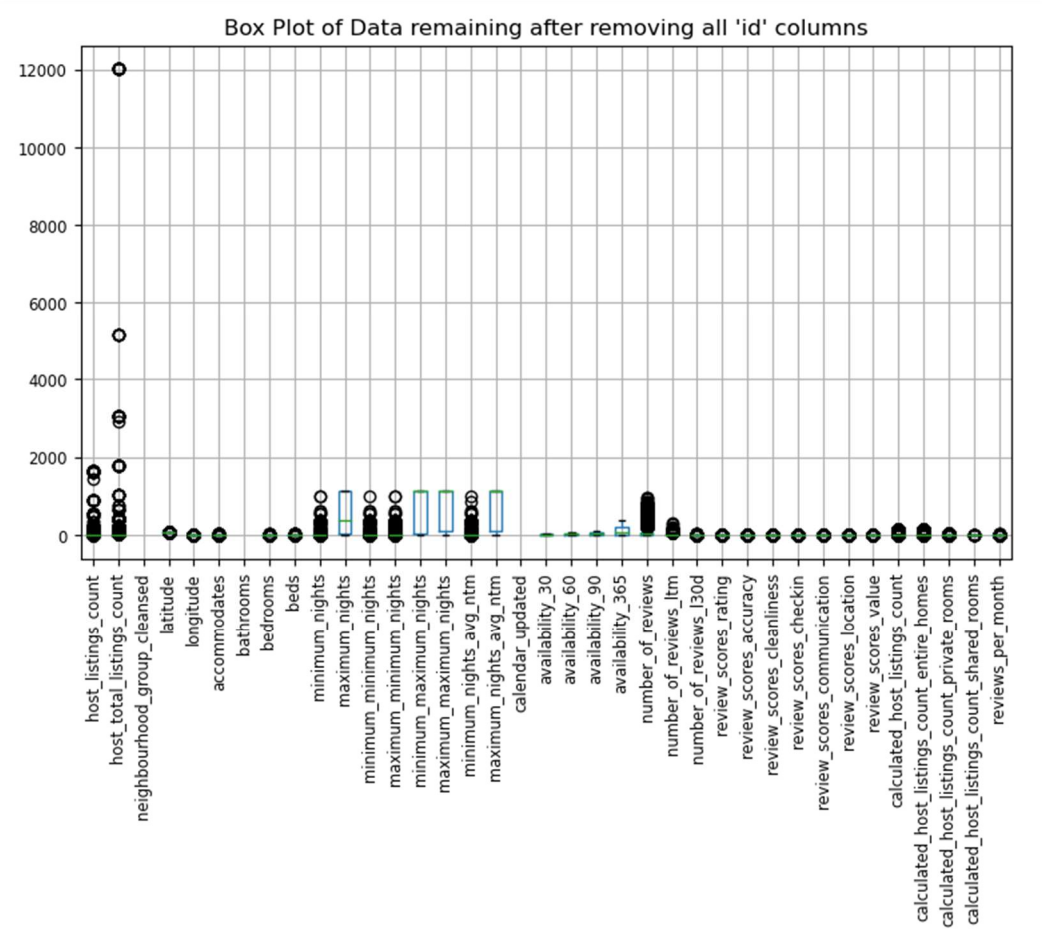


Figure 2- Box plot of Data

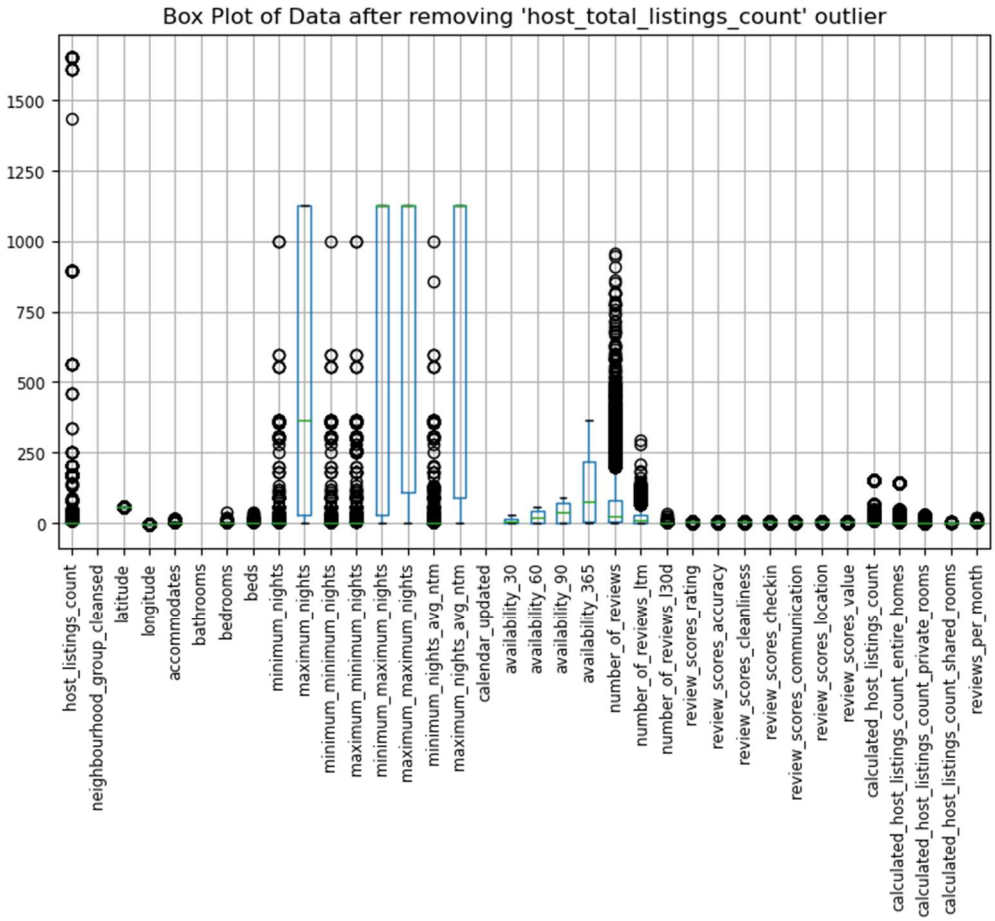


Figure 3- Box plot of Data after removing host_total_listings_count

Looking at the correlation matrix between variables, there appears to be a strong positive correlation between accommodates, bedrooms and beds. A positive correlation of 0.84 between accommodates and beds, while beds and bedrooms is 0.79. Same for accommodates and bedrooms. This is logical as we expect such variables to move together being a related variable, a house will only accommodate as many number of guests as the bedrooms they have and beds. We also notice a strong positive correlation between minimum_nights, minimum_minimum_nights, maximum_minimum_nights, minimum_nights_avg_ntm. Also, between variables minimum_maximum_nights, maximum_maximum_nights, and minimum_nights_avg_ntm. The same positive correlation heat is observed between variables: availability_30, availability_60, availability_90, availability_365; number_of_reviews, number_of_reviews_ltm; review_scores_rating, review_scores_accuracy, review_scores_cleanliness, review_scores_checkin, review_scores_communication, review_scores_location, review_scores_value;

calculated_host_listings_count, calculated_host_listings_count_entire_homes. The correlations are logical as variables that move together are the same positively correlated.

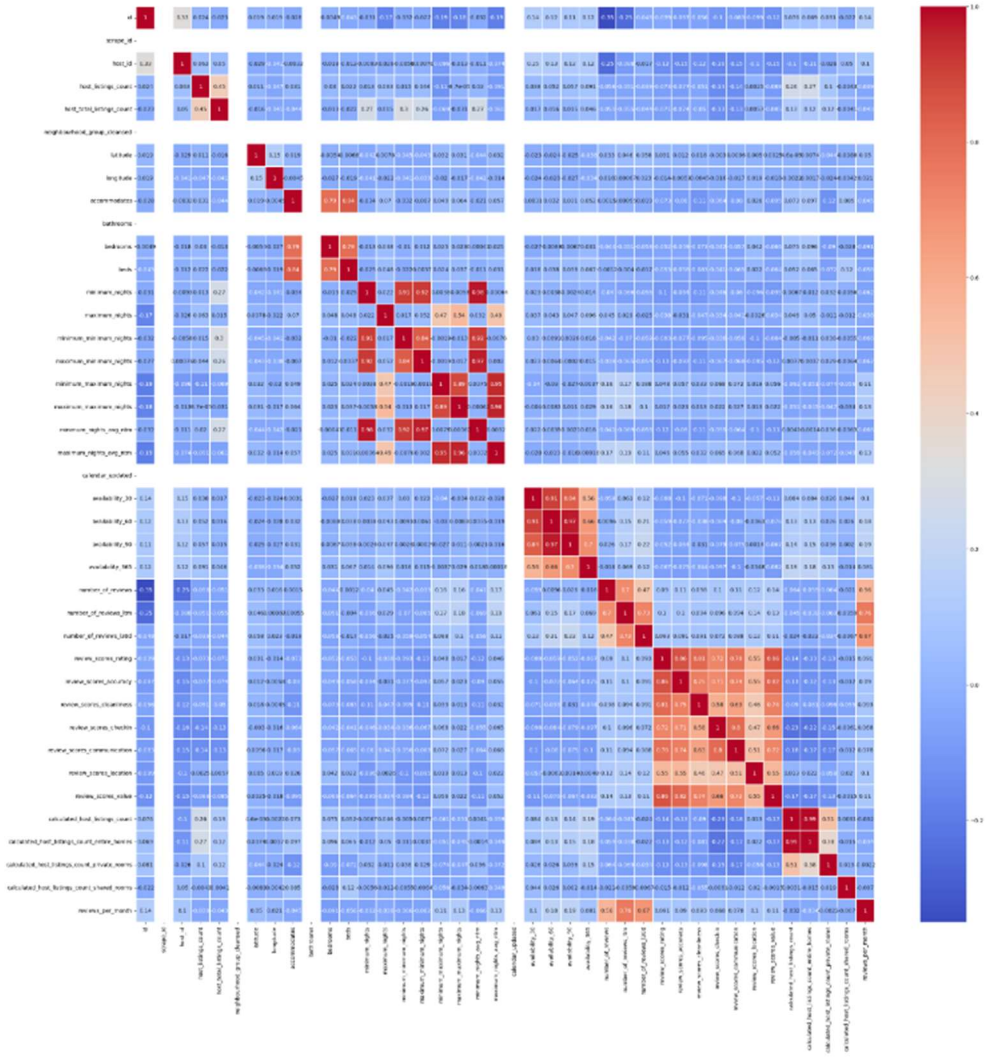


Figure 4- Correlation heat map

All variables' columns are brought into SAS Enterprise Miner for our analysis and model building. We still have 75 variable columns however we will drop the variables with texts which cannot be label encoded to tell more about our data set. We will also modify the bathrooms_text variable to create a new variable-bathrooms. This will contain only Integers (which is what bathroom variables should look like). The new "bathrooms" variable will replace the old bathrooms variable column which was empty. The summary statistic for both our Class and Interval variables on the 7,389 listings are shown below.

Class Variable Summary Statistics
(maximum 500 observations printed)

Data Role=TRAIN

Data Role	Variable Name	Role	Number of Levels	Missing	Mode	Mode Percentage	Mode2	Mode2 Percentage
TRAIN	has_availability	INPUT	2	0	1	99.85	0	0.15
TRAIN	host_has_profile_pic	INPUT	2	0	1	98.32	0	1.68
TRAIN	host_identity_verified	INPUT	2	0	1	89.80	0	10.20
TRAIN	host_is_superhost	INPUT	3	0	0	63.13	1	36.84
TRAIN	host_response_time	INPUT	5	1528	within an hour	62.27		20.68
TRAIN	room_type	INPUT	4	0	0	70.00	2	29.37
TRAIN	instant_bookable	TARGET	2	0	0	61.98	1	38.02

Figure 5- Class Variable Summary Statistic

Interval Variable Summary Statistics
(maximum 500 observations printed)

Data Role=TRAIN

Variable	Role	Mean	Standard Deviation	Non Missing	Missing	Minimum	Median	Maximum	Skewness	Kurtosis
accommodates	INPUT	3.559345	2.105669	7389	0	0	3	16	1.812256	5.591445
availability_30	INPUT	8.796048	8.750757	7389	0	0	7	30	0.664788	-0.65718
availability_365	INPUT	119.6552	121.5732	7389	0	0	77	365	0.72325	-0.89125
availability_60	INPUT	22.91298	20.3565	7389	0	0	22	60	0.199113	-1.47308
availability_90	INPUT	37.51861	32.39586	7389	0	0	38	90	0.092611	-1.59808
bathrooms	INPUT	1.238995	0.791706	7360	29	0	1	40	20.56662	849.8914
bedrooms	INPUT	1.651772	1.069304	7251	138	1	1	40	9.019549	251.5752
beds	INPUT	2.116145	1.691818	7284	105	1	2	40	5.697261	76.1917
calculated_host_listings_count	INPUT	8.153336	22.83623	7389	0	1	1	151	5.225301	28.83604
calculated_host_listings_count_e	INPUT	6.834484	21.26564	7389	0	0	1	142	5.380674	30.61066
calculated_host_listings_count_p	INPUT	1.278522	3.522355	7389	0	0	0	28	5.195731	32.20893
calculated_host_listings_count_s	INPUT	0.014616	0.315718	7389	0	0	0	8	24.35775	608.3801
host_acceptance_rate	INPUT	0.904411	0.197981	6663	726	0	0.99	1	-3.03851	9.518287
host_listings_count	INPUT	20.81567	119.3259	7389	0	1	2	1650	11.25997	139.5407
host_response_rate	INPUT	0.959021	0.140975	5861	1528	0	1	1	-4.99191	27.31219
host_total_listings_count	INPUT	69.92218	656.7117	7389	0	1	2	12017	16.17017	282.865
maximum_maximum_nights	INPUT	728.2002	484.8081	7389	0	1	1125	1125	-0.51599	-1.58684
maximum_minimum_nights	INPUT	6.513601	31.57702	7389	0	1	3	1000	16.70786	375.1649
maximum_nights	INPUT	491.2692	486.7039	7389	0	1	365	1125	0.414304	-1.63742
maximum_nights_avg_ntm	INPUT	702.0952	488.2469	7389	0	1	1125	1125	-0.40996	-1.686
minimum_maximum_nights	INPUT	672.2248	503.0418	7389	0	1	1125	1125	-0.31343	-1.77352
minimum_minimum_nights	INPUT	4.347138	26.52053	7389	0	1	2	1000	19.34432	476.9634
minimum_nights	INPUT	4.625118	28.966	7389	0	1	2	1000	20.32157	522.614
minimum_nights_avg_ntm	INPUT	5.249966	28.9247	7389	0	1	2	1000	18.45817	436.3875
number_of_reviews	INPUT	66.49357	102.6899	7389	0	0	23	957	2.904491	11.75155
number_of_reviews_l30d	INPUT	1.220598	1.920187	7389	0	0	0	36	3.046536	25.05959
number_of_reviews_ltm	INPUT	18.78048	23.91297	7389	0	0	9	296	2.11729	8.01607
price	INPUT	169.2602	683.2254	7389	0	0	110	47566	50.96256	3269.907
review_scores_accuracy	INPUT	4.81989	0.331943	6711	678	1	4.91	5	-6.27697	56.65457
review_scores_checkin	INPUT	4.847938	0.310078	6711	678	1	4.94	5	-6.35296	59.83279
review_scores_cleanliness	INPUT	4.758693	0.356982	6711	678	1	4.86	5	-4.67512	36.09042
review_scores_communication	INPUT	4.855631	0.301937	6711	678	1	4.94	5	-6.80582	69.29597
review_scores_location	INPUT	4.800198	0.282562	6711	678	0	4.88	5	-5.45701	56.51189
review_scores_rating	INPUT	4.751138	0.405051	6724	665	0	4.85	5	-6.2715	57.71545
review_scores_value	INPUT	4.68903	0.364577	6711	678	1	4.77	5	-4.72624	36.40158
reviews_per_month	INPUT	1.960758	1.846676	6724	665	0.01	1.41	21.39	1.755375	5.59006

Figure 6- Interval Variable Summary Statistic

In addition to our skewed variables earlier identified above, which is also reflected on SAS StatExplore node, our new variable- "bathrooms" is also having a listing as 0 bathroom and outlier values of 40

bathrooms with its skewness value high at 20.567 (figure 6 above). How possible is it that a listing has no bathroom? Possible it is a shared apartment with general bathroom. Also, could it be that we have mansions listed because 40 bathrooms is quite high. Price is the most skewed variable with skewness at 50.963 and kurtosis (which tells how often outliers occur) high at 3,269.9. We have a property listed as $47,566 per night, while another is at $0. Could these be possible? We might have to remove these outliers so as to protect the integrity of our data set.

Variable Summary

Role	Measurement Level	Frequency Count
ID	NOMINAL	2
INPUT	BINARY	4
INPUT	INTERVAL	36
INPUT	NOMINAL	2
REJECTED	INTERVAL	2
REJECTED	NOMINAL	28
TARGET	BINARY	1

Figure 7- Variable Summary

Data Set Preparation and Transformation

Even though we have brought in all the 75 variables as given by Inside Airbnb, we cannot use all variables for our purpose. Series of cleaning and transformation activities were performed to make the data fit for identifying clusters and model building.

- Label encoding: This is done on text variables which we are confident should be included in our analysis Also, they have limited classes hence easy to label encode. This was done on Python (Jupyter) transforming variables- 'host_is_superhost', 'host_has_profile_pic', ['host_identity_verified', 'has_availability', 'instant_bookable', and 'room_type' to Float. In doing this, our summary statistics for these variables can be analysed.
- Removal of empty columns: This is done on columns with entirely empty cells. Variables in this category are "neighbourhood_group_cleansed", "bathrooms" (the original variable) and "calendar_updated".
- Variable conversion: Two variables in percentages- host_response_rate, and host_acceptance_rate were converted to floats also. This is allow them be in usable state for our analysis.

- Rejection of text columns: We rejected and dropped variables in text which will not be useful for the purpose of our objective. These variables are shown in figure below.

Variables - FIMPORT5

(none) not Equal to Apply

Columns: Label Mining Basic Statistics

Name	Role	Level	Report	Order	Drop	Lower Limit	Upper Limit
instant_bookable	Target	Binary	No		No	.	.
picture_url	Rejected	Nominal	No		Yes	.	.
host_url	Rejected	Nominal	No		Yes	.	.
scrape_id	Rejected	Nominal	No		Yes	.	.
calendar_last_scraped	Rejected	Nominal	No		Yes	.	.
last_review	Rejected	Nominal	No		Yes	.	.
source	Rejected	Nominal	No		Yes	.	.
host_verifications	Rejected	Nominal	No		Yes	.	.
host_thumbnail_url	Rejected	Nominal	No		Yes	.	.
neighbourhood	Rejected	Nominal	No		Yes	.	.
neighborhood_overview	Rejected	Nominal	No		Yes	.	.
neighbourhood_group_cleansed	Rejected	Nominal	No		Yes	.	.
name	Rejected	Nominal	No		Yes	.	.
bathrooms_text	Rejected	Nominal	No		Yes	.	.
property_type	Rejected	Nominal	No		Yes	.	.
neighbourhood_cleansed	Rejected	Nominal	No		Yes	.	.
description	Rejected	Nominal	No		Yes	.	.
latitude	Rejected	Interval	No		Yes	.	.
amenities	Rejected	Nominal	No		Yes	.	.
calendar_updated	Rejected	Nominal	No		Yes	.	.
longitude	Rejected	Interval	No		Yes	.	.
listing_url	Rejected	Nominal	No		Yes	.	.
license	Rejected	Nominal	No		Yes	.	.
last_scraped	Rejected	Nominal	No		Yes	.	.
host_name	Rejected	Nominal	No		Yes	.	.
host_since	Rejected	Nominal	No		Yes	.	.
host_neighbourhood	Rejected	Nominal	No		Yes	.	.
host_picture_url	Rejected	Nominal	No		Yes	.	.
host_about	Rejected	Nominal	No		Yes	.	.
first_review	Rejected	Nominal	No		Yes	.	.
host_location	Rejected	Nominal	No		Yes	.	.
availability_90	Input	Interval	No		No	.	.
availability_365	Input	Interval	No		No	.	.
availability_30	Input	Interval	No		No	.	.
host_listings_count	Input	Interval	No		No	.	.
bathrooms	Input	Interval	No		No	.	.
host_is_superhost	Input	Binary	No		No	.	.

Figure 8- Rejected/Dropped Variables

- Missing data: We used the Filter node to remove missing variables in our data. In the summary statistics figure shown previously above, we can see that these missing variable cells are mostly under host_response_time for Class variables and review_scores_rating, review_scores_accuracy, review_scores_cleanliness, review_scores_checkin, review_scores_communication, review_scores_location, review_scores_value, host_response_rate, host_acceptance_rate for Interval variables. Out of our 7,389 listings, we are left with 4,312 listings. To protect the integrity of our data set, we avoided using mean or any arbitrary value inputed.

Number Of Observations

Data Role	Filtered	Excluded	DATA
TRAIN	4312	3077	7389

Figure 9- Filtered Data Observation

- Transformation- New Derive Variables: In order to ensure the goal of our data mining exercise, we created two new variables- “Occupancy rate” and “listing income”. The formulas are as defined by Airbnb (2023) "Average occupancy rate is the number of nights booked divided by total nights available to be booked across all relevant listings". Mashvisor (2023) also defines it in same manner:

  Occupancy Rate Formula = Number of Booked Nights / Number of Available Nights.

  The number of available nights can be clearly defined as the number of days your investment property was made available for rent that year. This includes the number of days it was booked or occupied (because for it be booked it had to have been originally advertised as available).

  Number of Available Nights = Available Nights + Booked Nights.

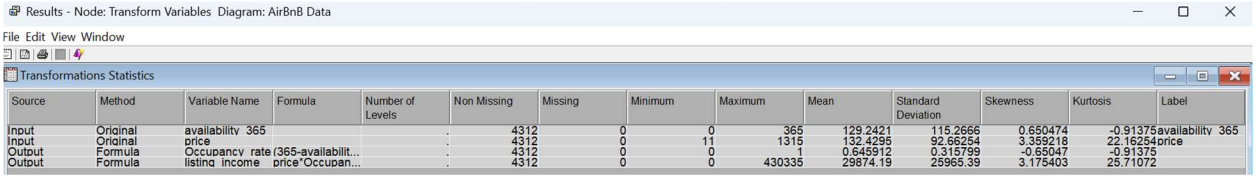
  For our purpose, our formula within this context which adopts the definitions above is:

  “Occupancy rate” = (365-availability_365)/365

  “listing income” = price*Occupancy_rate*365

  This is with the assumption that availability_365 shows the number of days a listing is available, and if not available, it is booked by a guest, which is Booked nights.

  The Transform Variables node will help in this regard. Formula for our new variable is inputted and node calculates and create additional columns for the two new variables.



Source	Method	Variable Name	Formula	Number of Levels	Non Missing	Missing	Minimum	Maximum	Mean	Standard Deviation	Skewness	Kurtosis	Label
Input	Original	availability_365		.	4312	0	0	365	129.2421	115.2666	0.650474	-0.91375	availability_365
Input	Original	price		.	4312	0	11	1315	132.4295	92.66254	3.359218	22.16254	price
Output	Formula	Occupancy_rate	(365-availabilit...	.	4312	0	0	1	0.645912	0.315799	-0.65047	-0.91375	
Output	Formula	listing_income	price*Occupan...	.	4312	0	0	430335	29874.19	25965.39	3.175403	25.71072	

Figure 10- New Derive Variables

- Transformation- Scaling Variables: All our variables are in different scales and ranges. If left that way, we will be attaching higher importance to variables with higher values and lower importance to variables with lower values. We do not want this to happen as we want all variables to be of equal stand while the model building process determines which variable is important. The Transform Variables node is brought in again to standardize all variables this time to a common scale for comparison. In doing this, we select Maximum Normal for Interval inputs and Group rare levels for Class Inputs. SAS generates an appropriate formula that helps transform all the variables and brings them to the same scale.

Computed Transformations
(maximum 500 observations printed)

Input Name	Role	Input Level	Name	Level	Formula
Occupancy_rate	INPUT	INTERVAL	EXP_Occupancy_rate	INTERVAL	exp(max(Occupancy_rate-0, 0.0))
accommodates	INPUT	INTERVAL	SQRT_accommodates	INTERVAL	sqrt(max(accommodates-1, 0.0)/8)
availability_90	INPUT	INTERVAL	EXP_availability_90	INTERVAL	exp(max(availability_90-0, 0.0)/90)
bedrooms	INPUT	INTERVAL	SQRT_bedrooms	INTERVAL	sqrt(max(bedrooms-1, 0.0)/3)
beds	INPUT	INTERVAL	SQRT_beds	INTERVAL	sqrt(max(beds-1, 0.0)/6)
calculated_host_listings_count	INPUT	INTERVAL	LOG_calculated_host_listings_cou	INTERVAL	log(max(calculated_host_listings_count-1, 0.0)/47 + 1)
calculated_host_listings_count_e	INPUT	INTERVAL	LOG_calculated_host_listings_col	INTERVAL	log(max(calculated_host_listings_count_e-0, 0.0)/48 + 1)
calculated_host_listings_count_p	INPUT	INTERVAL	LOG_calculated_host_listings_co2	INTERVAL	log(max(calculated_host_listings_count_p-0, 0.0)/10 + 1)
calculated_host_listings_count_s	INPUT	INTERVAL	LOG_calculated_host_listings_co3	INTERVAL	log(calculated_host_listings_count_s + 1)
host_acceptance_rate	INPUT	INTERVAL	PWR_host_acceptance_rate	INTERVAL	(max(host_acceptance_rate-0.33, 0.0)/0.67)**4
host_listings_count	INPUT	INTERVAL	LOG_host_listings_count	INTERVAL	log(max(host_listings_count-1, 0.0)/250 + 1)
host_response_rate	INPUT	INTERVAL	PWR_host_response_rate	INTERVAL	(max(host_response_rate-0.56, 0.0)/0.44)**4
host_response_time	INPUT	NOMINAL	TG_host_response_time	NOMINAL	Group:host_response_time
host_total_listings_count	INPUT	INTERVAL	LOG_host_total_listings_count	INTERVAL	log(max(host_total_listings_count-1, 0.0)/662 + 1)
listing_income	INPUT	INTERVAL	PWR_listing_income	INTERVAL	(max(listing_income-0, 0.0)/430335)**0.25
maximum_maximum_nights	INPUT	INTERVAL	PWR_maximum_maximum_nights	INTERVAL	(max(maximum_maximum_nights-1, 0.0)/1124)**0.25
maximum_minimum_nights	INPUT	INTERVAL	LOG_maximum_minimum_nights	INTERVAL	log(max(maximum_minimum_nights-1, 0.0)/94 + 1)
maximum_nights	INPUT	INTERVAL	LOG_maximum_nights	INTERVAL	log(max(maximum_nights-1, 0.0)/1124 + 1)
maximum_nights_avg_ntm	INPUT	INTERVAL	PWR_maximum_nights_avg_ntm	INTERVAL	(max(maximum_nights_avg_ntm-1, 0.0)/1124)**0.25
minimum_maximum_nights	INPUT	INTERVAL	PWR_minimum_maximum_nights	INTERVAL	(max(minimum_maximum_nights-1, 0.0)/1124)**0.25
minimum_minimum_nights	INPUT	INTERVAL	LOG_minimum_minimum_nights	INTERVAL	log(max(minimum_minimum_nights-1, 0.0)/31 + 1)
minimum_nights	INPUT	INTERVAL	LOG_minimum_nights	INTERVAL	log(max(minimum_nights-1, 0.0)/89 + 1)
minimum_nights_avg_ntm	INPUT	INTERVAL	LOG_minimum_nights_avg_ntm	INTERVAL	log(max(minimum_nights_avg_ntm-1, 0.0)/87 + 1)
number_of_reviews	INPUT	INTERVAL	PWR_number_of_reviews	INTERVAL	(max(number_of_reviews-1, 0.0)/373)**0.25
number_of_reviews_l30d	INPUT	INTERVAL	SQRT_number_of_reviews_l30d	INTERVAL	sqrt(max(number_of_reviews_l30d-0, 0.0)/6)
number_of_reviews_ltm	INPUT	INTERVAL	SQRT_number_of_reviews_ltm	INTERVAL	sqrt(max(number_of_reviews_ltm-0, 0.0)/90)
review_scores_accuracy	INPUT	INTERVAL	PWR_review_scores_accuracy	INTERVAL	(max(review_scores_accuracy-3.83, 0.0)/1.17)**4
review_scores_checkin	INPUT	INTERVAL	PWR_review_scores_checkin	INTERVAL	(max(review_scores_checkin-3.92, 0.0)/1.08)**4
review_scores_cleanliness	INPUT	INTERVAL	PWR_review_scores_cleanliness	INTERVAL	(max(review_scores_cleanliness-3.7, 0.0)/1.3)**4
review_scores_communication	INPUT	INTERVAL	PWR_review_scores_communication	INTERVAL	(max(review_scores_communication-4, 0.0))**4
review_scores_location	INPUT	INTERVAL	PWR_review_scores_location	INTERVAL	(max(review_scores_location-4, 0.0))**4
review_scores_rating	INPUT	INTERVAL	PWR_review_scores_rating	INTERVAL	(max(review_scores_rating-3.63, 0.0)/1.37)**4
review_scores_value	INPUT	INTERVAL	PWR_review_scores_value	INTERVAL	(max(review_scores_value-3.61, 0.0)/1.39)**4
reviews_per_month	INPUT	INTERVAL	SQRT_reviews_per_month	INTERVAL	sqrt(max(reviews_per_month-0.02, 0.0)/7.48)

Figure 11- Computed Transformations

EMWS1.Trans6_TRAIN

	Transformed Occupancy_rate	Transformed: accommodates	Transformed: availability_90	Transformed: bedrooms	Transformed: beds	Transformed: calculated_host_listings_count	Transformed: calculated_host_listings_count_entire_homes
1	1.5932141035980574	0.5	1.7236505327024385	0.0	0.0	0.0	0.0
2	2.1594704156573337	0.6123724356957945	1.3498588075760032	0.5773502691896257	0.408248290463863	0.0	0.0206192872023561
3	2.7182818284590455	0.7905694150420949	1.0	0.816496580927726	0.7071067811865476	0.0	0.0206192872023561
4	2.7182818284590455	0.0	1.0	0.0	0.0	0.0	0.0
5	1.3044133828532993	0.6123724356957945	1.3801914510923234	0.5773502691896257	0.5773502691896257	0.06187540371808745	0.08004270767353636
6	2.7182818284590455	0.6123724356957945	1.0	0.0	0.408248290463863	0.26072626246325264	0.27193371548364176
7	2.7182818284590455	0.7071067811865476	1.0	0.5773502691896257	0.408248290463863	0.26072626246325264	0.27193371548364176
8	2.2999254568134266	0.6123724356957945	1.7046048653227532	0.5773502691896257	0.408248290463863	0.6931471805599453	0.6931471805599453
9	2.124262445237039	0.7905694150420949	2.200949567622183	0.5773502691896257	0.408248290463863	0.0	0.0206192872023561
10	2.4294653993387123	0.7905694150420949	1.033895113513574	0.5773502691896257	0.0	0.04167269640056808	0.06062462181643484
11	2.268634439989123	0.9354143466934853	1.742908998633458	0.816496580927726	1.0	0.04167269640056808	0.06062462181643484
12	1.2691611423318734	0.6123724356957945	1.9915015239946179	0.5773502691896257	0.5773502691896257	0.04167269640056808	0.06062462181643484
13	2.594580009668586	0.7071067811865476	1.2079067331671705	0.816496580927726	1.0	0.0	0.0206192872023561
14	1.1219502808499364	0.3535533905932738	2.40554809994578	0.0	0.0	0.0	0.0206192872023561
15	2.130090322094849	0.6123724356957945	2.2504070503288136	0.5773502691896257	0.408248290463863	0.6931471805599453	0.6931471805599453
16	2.696031351609739	0.7905694150420949	1.0	0.816496580927726	0.816496580927726	0.0210534091978322 63	0.0408219945202552
17	2.2316466610973777	0.7905694150420949	1.9049284416683332	0.816496580927726	0.816496580927726	0.1752040890250906	0.18924199963852834
18	2.573342093189731	0.3535533905932738	1.045446894714042	0.0	0.0	0.0210534091978322 63	0.0206192872023561
19	2.268634439989123	0.7905694150420949	1.863063586398077	0.5773502691896257	0.408248290463863	0.6931471805599453	0.6931471805599453
20	2.124262445237039	0.7905694150420949	2.275551003054389	0.5773502691896257	0.408248290463863	0.6931471805599453	0.6931471805599453
21	2.3964119616894335	0.3535533905932738	1.0	0.0	0.0	0.04167269640056808	0.0408219945202552
22	1.0476770106643425	0.8660254037844386	2.6001146899026075	0.5773502691896257	0.7071067811865476	0.0210534091978322 63	0.0408219945202552
23	1.520711142251546	0.6123724356957945	1.742908998633458	0.5773502691896257	0.7071067811865476	0.0210534091978322 63	0.0408219945202552
24	2.268634439989123	0.7071067811865476	1.8221188003905089	0.816496580927726	0.7071067811865476	0.39834764033930037	0.4054651081081644
25	1.650981340628437	0.6123724356957945	1.0	0.5773502691896257	0.5773502691896257	0.1752040890250906	0.18924199963852834
26	1.919480918464032	0.3535533905932738	1.3801914510923234	0.0	0.0	0.0210534091978322 63	0.0206192872023561
27	2.4361305951620444	0.6123724356957945	1.3056051720649522	0.0	0.5773502691896257	0.0	0.0206192872023561
28	2.65207530044486 63	0.3535533905932738	1.0	0.0	0.0	0.0	0.0206192872023561
29	1.7826164339834 14	0.3535533905932738	2.1525790183195874	0.0	0.0	0.0210534091978322 63	0.0206192872023561
30	1.7296951956122095	0.3535533905932738	2.3266848527514563	0.0	0.0	0.0210534091978322 63	0.0206192872023561
31	1.5888551074054622	0.6123724356957945	1.863063586398077	0.5773502691896257	0.408248290463863	0.0210534091978322 63	0.0408219945202552
32	1.0856639993996589	0.3535533905932738	2.40554809994578	0.0	0.0	0.0210534091978322 63	0.0
33	2.28109943311118	0.5	1.5084929575753376	0.816496580927726	0.7071067811865476	0.1010961168713688 1	0.11778303565638346
34	1.669173789405883	0.6123724356957945	1.0	0.0	0.408248290463863	0.0	0.0206192872023561

Figure 12- Transformed Variables

- Data Reduction: We still need to reduce the total number of variables we have, which stands at 45 now after the above cleaning and transformations. This will be done with the Variable Clustering node to identify which variables are mostly uncorrelated. It will amount to duplicity of effect if we use correlated variables hence why we are bringing this node to help achieve this objective. The best results will be achieved, when we have lots of features or variables, if we only bring forward variables that are uncorrelated. In the figure below, we see that we have only 8 selected variables that are uncorrelated as other correlated variables are dropped while picking only one variable out of the group of correlated variables.

Variable Summary

Role	Measurement Level	Frequency Count
ID	NOMINAL	2
INPUT	BINARY	4
INPUT	INTERVAL	36
INPUT	NOMINAL	2
TARGET	BINARY	1

Figure 13- Total Variables prior to Variable Clustering

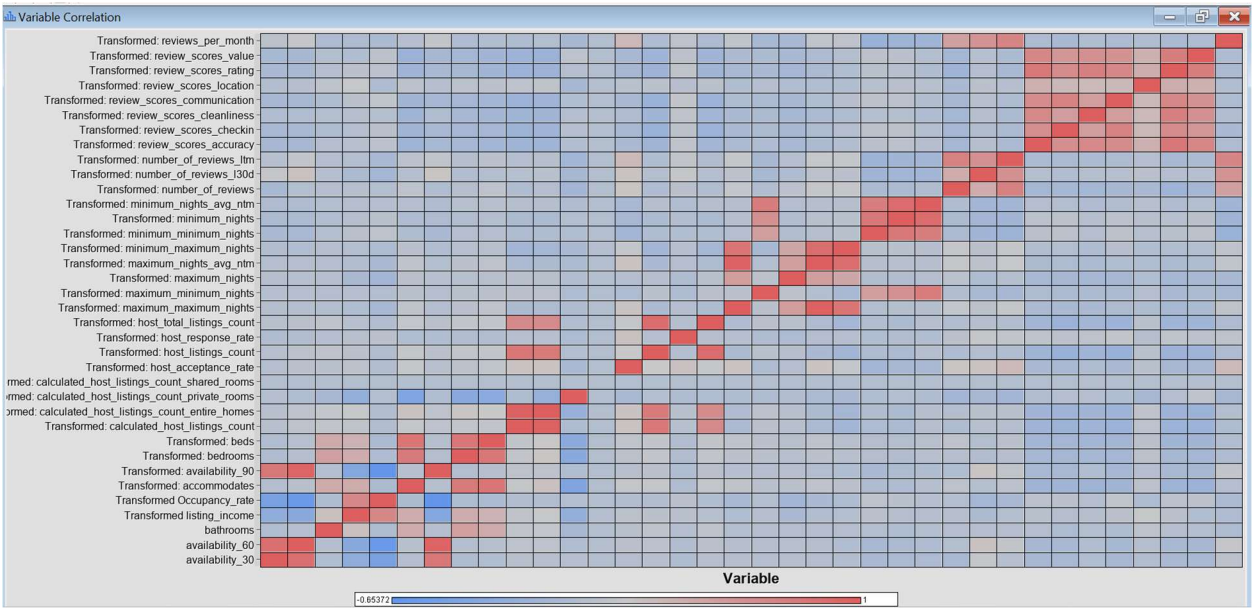
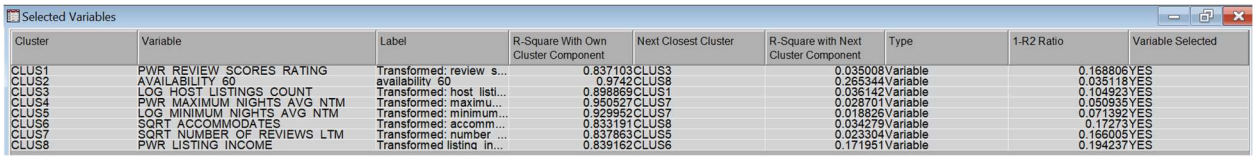


Figure 14- Variable Correlations



Cluster	Variable	Label	R-Square With Own Cluster Component	Next Closest Cluster	R-Square with Next Cluster Component	Type	1-R2 Ratio	Variable Selected
CLUS1	PWR_REVIEW_SCORES_RATING	Transformed: review s...	0.837103	CLUS3	0.035008	Variable	0.168806	YES
CLUS2	AVAILABILITY_60	availability_60	0.9742	CLUS8	0.265344	Variable	0.035118	YES
CLUS3	LOG_HOST_LISTINGS_COUNT	Transformed: host_listi...	0.898869	CLUS1	0.036142	Variable	0.104923	YES
CLUS4	PWR_MAXIMUM_NIGHTS_AVG_NTM	Transformed: maximu...	0.950527	CLUS7	0.028701	Variable	0.050935	YES
CLUS5	LOG_MINIMUM_NIGHTS_AVG_NTM	Transformed: minimum...	0.929952	CLUS7	0.018826	Variable	0.071392	YES
CLUS6	SQRT_ACCOMMODATES	Transformed: accomm...	0.833191	CLUS8	0.034279	Variable	0.17273	YES
CLUS7	SQRT_NUMBER_OF_REVIEWS_LTM	Transformed: number_...	0.837863	CLUS5	0.023304	Variable	0.166005	YES
CLUS8	PWR_LISTING_INCOME	Transformed listing_in...	0.839162	CLUS6	0.171951	Variable	0.194237	YES

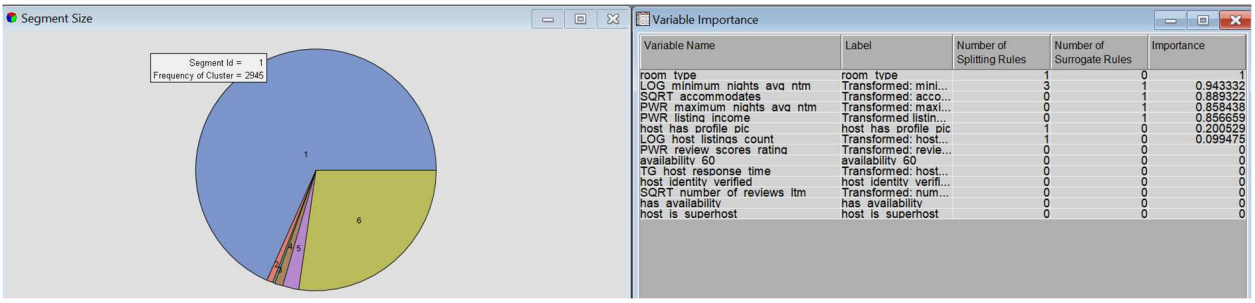
Figure 15- Total Variables after to Variable Clustering

Cluster Analysis

Now that we have been able to deal with missing values and reduce the number of variables methodically ensuring only uncorrelated and highly important variables are put forward, we will try to identify clusters within this data set. The Cluster node on SAS helps in this regard.

Average method: This method computes the distance between two clusters as the average distance between all pairs of observations from the two clusters. It can result in non-spherical clusters and is more robust to outliers than the centroid method. From the data set, this method identified 6 clusters. Based on the variable importance, we see that room_type is the most important hence will be used mostly in identifying our clusters followed by minimum_nights_avg_ntm. We can observe that cluster 1, 3 and 5 contains listings whose host let out entire homes/apartments on their listings. About 90 - 100% of their listings are entire homes and apartments, now this is not the "temporary accommodation" model which Airbnb should promote. Cluster 4 also has a high average percentage of minimum_nights_avg_ntm at 29.2% which tells that these hosts listed properties have their minimum night's stay quite higher than other hosts in other

clusters. This also suggests a more permanent accommodation type being offered to guests. Looking at the nearest clusters and distance to each cluster, one can tell that we have 2 major clusters. Cluster 1, 3, 4 and 5 which are closer together and have similarities on one hand while Cluster 2, and 6 are on another hand.

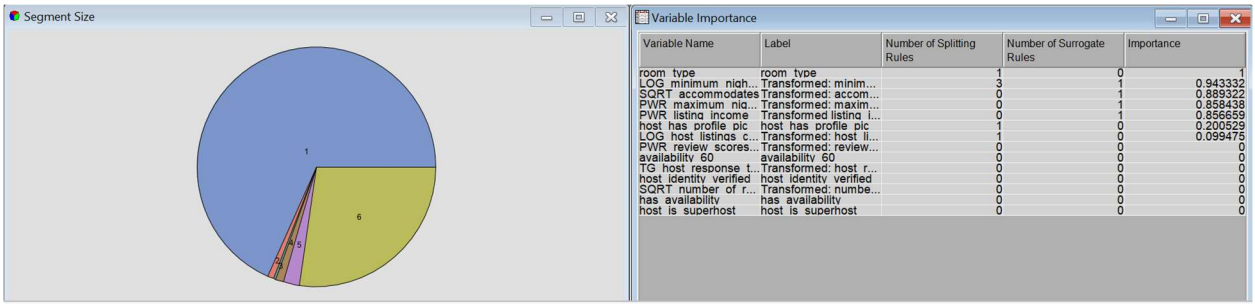


Variable Name	Label	Number of Splitting Rules	Number of Surrogate Rules	Importance
room_type	room_type	1	0	1
LOG_minimum_nights_avg_ntm	Transformed: mini...	3	1	0.943332
SQRT_accommodates	Transformed: acco...	0	1	0.889322
PWR_maximum_nights_avg_ntm	Transformed: maxi...	0	1	0.858438
PWR_listing_income	Transformed listin...	0	1	0.856659
host_has_profile_pic	host_has_profile_pic	1	0	0.200529
LOG_host_listings_count	Transformed: host...	1	0	0.099475
PWR_review_scores_rating	Transformed: revie...	0	0	0
availability_60	availability_60	0	0	0
TG_host_response_time	Transformed: host...	0	0	0
host_identity_verified	host_identity_verifi...	0	0	0
SQRT_number_of_reviews_ltm	Transformed: num...	0	0	0
has_availability	has_availability	0	0	0
host_is_superhost	host_is_superhost	0	0	0

Figure 16- Clusters and Variable Importance (Method: Average)

We bring in a Sample node to test 90% of the Data so we can confirm the stability of the clustering and same is noted, there seems to be a pattern grouping of clusters into two major groups. Following the same variables above under the sample, Clusters 1, 2, 3, 5 and 7 have similarities of listings that show a "permanent accommodation" nature, while Cluster 4 and 6 is otherwise. This confirms our data set is stable in result, even when a sample is taken.

Centroid method: In this method, the distance between two clusters is computed as the distance between the centroids of the clusters. The centroid is the mean of all the observations in the cluster. This method can result in well-separated and spherical clusters. Most times, this method yields same result with Average method if there are no outliers in the data set. This is the case here as the cluster analysis is same with Average earlier done. In the sample tested also, it follows same as Average with 5 Clusters showing signs of a "permanent accommodation" nature.



Variable Name	Label	Number of Splitting Rules	Number of Surrogate Rules	Importance
room_type	room_type	1	0	1
LOG_minimum_nigh...	Transformed: minim...	3	1	0.943332
SQRT_accommodates	Transformed: accom...	0	1	0.889322
PWR_maximum_nig...	Transformed: maxim...	0	1	0.858438
PWR_listing_income	Transformed listing i...	0	1	0.856659
host_has_profile_pic	host_has_profile_pic	1	0	0.200529
LOG_host_listings_c...	Transformed: host li...	1	0	0.099475
PWR_review_scores...	Transformed: review...	0	0	0
availability_60	availability_60	0	0	0
TG_host_response_t...	Transformed: host r...	0	0	0
host_identity_verified	host_identity_verified	0	0	0
SQRT_number_of_r...	Transformed: numbe...	0	0	0
has_availability	has_availability	0	0	0
host_is_superhost	host_is_superhost	0	0	0

Figure 17- Clusters and Variable Importance (Method: Centroid)

Ward's method: This method minimizes the total within-cluster variance and is considered to be the most reliable method. The distance between two clusters is computed as the sum of squared differences between all observations within the two clusters and the centroid of the combined cluster. This method can result in

compact and balanced clusters. In line with variable importance, Clusters 1, 3, 4 and 5 have listings that are entire houses and apartments. This suggests a more "permanent accommodation" type. On the other hand, Cluster 2 are likely to be genuine short term let.

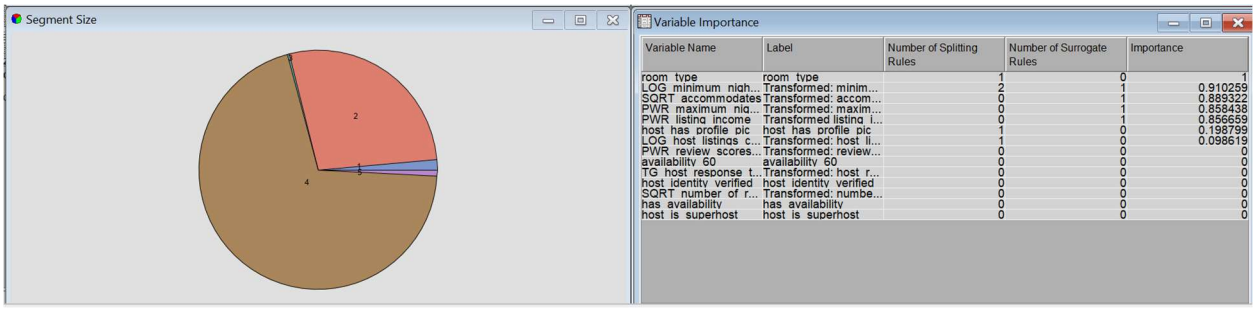


Figure 18- Clusters and Variable Importance (Method: Ward)

To justify the two groupings of clusters, we created another cluster node while setting maximum nodes to 2. Cluster 1 is a more permanent type of listing as we can see from the room type- entire houses and apartments. Also, the host listings count for Cluster 1 is quite high which buttresses our point of a short term let being used on a permanent basis.

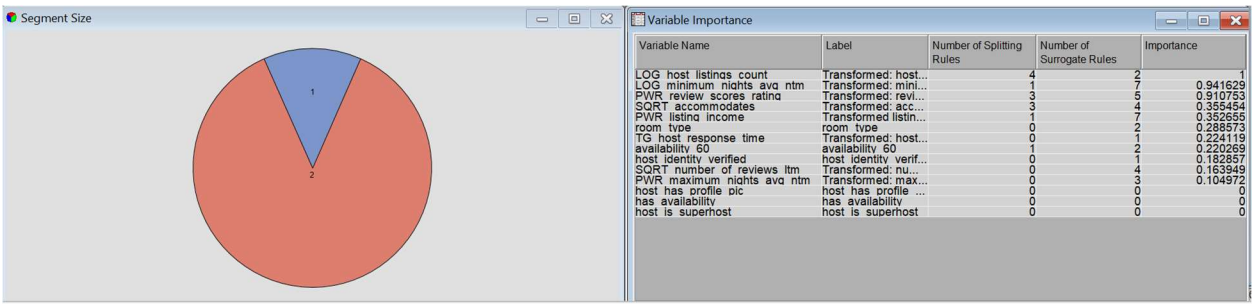


Figure 19- Clusters and Variable Importance (Method: 2 maximum cluster)

The Tree for each method is placed at Appendix 1. This shows the rules for classification into these clusters which can be used for subsequent listings' classification into the clusters.

Classification Model Building and Model Evaluation

In building our classification model, we will adopt the variables provided by our previous Cluster. The variables are, as determined by our Clustering process, the best uncorrelated variables that achieves our goal. These variables are LOG_host_listings_count, LOG_mininmum_nights_avg_ntm, PWR_listing_income, PWR_maximum_nights_avg_ntm, PWR_review_scores_rating, SORT_accommodates, SORT_number_of_reviews_ltm, TG_host_response_time. They represent the transformed original variables and were standardized because of having all variables on the same scale and

not allowing our model attach importance to any variable unnecessarily. To proceed, we will introduce the Data partition node to split our data set. The most common method of split is 40, 30, 30, among Training, Validation, Test respectively.

We used Memory Based Reasoning (K Nearest Neighbour), Decision Tree, Neural Network and Logistic Regression in building our model and brought in the Model comparison node to ascertain the best model.

MBR (Model-based Reasoning): MBR is a technique used in data mining to create models that reason and make predictions based on existing knowledge and data. It involves building a model using a set of rules and patterns derived from the data. MBR is useful for decision-making tasks and can provide insights into complex relationships and dependencies in the data which is the case in this our instance. The Misclassification rate for Train, Validate and Test on the Data set is 31.69%, 36.35% and 38.12%. There is no overfitting or underfitting.

Fit Statistics

Target	Target Label	Fit Statistics	Statistics Label	Train	Validation	Test
instant_boo...	instant_boo...	NW	Number of Estimated Weights	8	.	.
instant_boo...	instant_boo...	NOBS	Sum of Frequencies	1723	1293	1296
instant_boo...	instant_boo...	SUMW	Sum of Case Weights Times Freq	3446	2586	2592
instant_boo...	instant_boo...	DFT	Total Degrees of Freedom	1723	.	.
instant_boo...	instant_boo...	DFM	Model Degrees of Freedom	8	.	.
instant_boo...	instant_boo...	DFE	Degrees of Freedom for Error	1715	.	.
instant_boo...	instant_boo...	ASE	Average Squared Error	0.201227	0.234435	0.241428
instant_boo...	instant_boo...	RASE	Root Average Squared Error	0.448584	0.484185	0.491353
instant_boo...	instant_boo...	DIV	Divisor for ASE	3446	2586	2592
instant_boo...	instant_boo...	SSE	Sum of Squared Errors	693.4297	606.25	625.7813
instant_boo...	instant_boo...	MSE	Mean Squared Error	0.202166	0.234435	0.241428
instant_boo...	instant_boo...	RMSE	Root Mean Squared Error	0.449629	0.484185	0.491353
instant_boo...	instant_boo...	AVERR	Average Error Function	0.585764	0.673305	0.680966
instant_boo...	instant_boo...	ERR	Error Function	2018.542	1741.168	1765.065
instant_boo...	instant_boo...	MAX	Maximum Absolute Error	0.9375	1	0.9375
instant_boo...	instant_boo...	FPE	Final Prediction Error	0.203105	.	.
instant_boo...	instant_boo...	RFPE	Root Final Prediction Error	0.450671	.	.
instant_boo...	instant_boo...	AIC	Akaike's Information Criterion	2034.542	.	.
instant_boo...	instant_boo...	SBC	Schwarz's Bayesian Criterion	2078.156	.	.
instant_boo...	instant_boo...	MISC	Misclassification Rate	0.316889	0.363496	0.381173
instant_boo...	instant_boo...	WRONG	Number of Wrong Classifications	546	470	494

Figure 20- MBR Fit Statistic showing Misclassification rate

Data Role=VALIDATE Target=instant_bookable Target Label=instant_bookable

False Negative	True Negative	False Positive	True Positive
385	751	85	72

Figure 21- MBR Confusion Matrix

Decision Tree: Decision trees are a popular data mining technique used for classification and regression tasks. They involve creating a tree-like model where each internal node represents a feature or attribute, each branch represents a decision rule, and each leaf node represents an outcome or prediction. Decision trees are easy to understand and interpret, and they can handle both categorical and numerical data. The Misclassification rate on this model for Train, Validate and Test on the Data set is 30.47%, 32.64% and 31.64% respectively. There is no overfitting or underfitting, instead the model improved slightly on Test data. The Tree showing rules for classification is placed on Appendix 2 below.

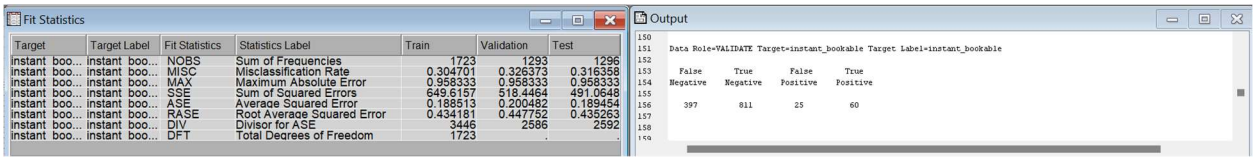


Figure 22- Decision Tree Model Fit Statistic showing Misclassification rate and Confusion Matrix

Neural Network: Neural networks are a type of machine learning algorithm inspired by the structure and function of the human brain. They are used for various data mining tasks such as classification, regression, and pattern recognition. Neural networks consist of interconnected nodes or "neurons" organized in layers, where each neuron applies a mathematical function to its inputs and passes the result to the next layer. Neural networks can capture complex relationships in the data and make accurate predictions. The Misclassification rate on this model for Train, Validate and Test on the Data set is 28.26%, 32.71% and 32.10% respectively. There is no overfitting or underfitting, instead the model improved slightly on Test data.

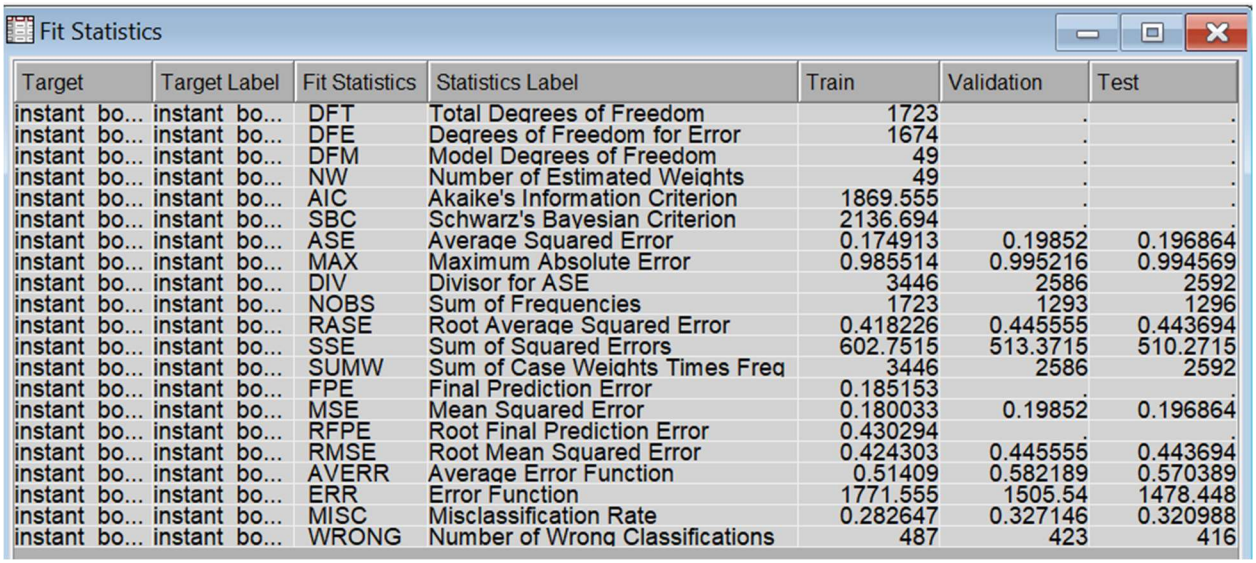


Figure 23- Neural Network Fit Statistic showing Misclassification rate

Data Role=VALIDATE Target=instant_bookable Target Label=instant_bookable

False Negative	True Negative	False Positive	True Positive
259	672	164	198

Figure 24- MBR Confusion Matrix

Logistic Regression: It is a statistical technique used for binary classification tasks, where the goal is to predict a binary outcome variable based on one or more independent variables. It is an extension of linear regression but is specifically designed for predicting categorical outcomes. In logistic regression, the dependent variable is a binary variable (e.g., yes/no, true/false, 0/1), as is this case, and the independent variables can be either continuous or categorical. The Misclassification rate on this model for Train, Validate and Test on the Data set is 30.01%, 31.55% and 32.48% respectively. There is no overfitting or underfitting.

Fit Statistics

Target	Target Label	Fit Statistics	Statistics Label	Train	Validation	Test
instant_bo...	instant_boo...	AIC	Akaike's Information Criterion	1934.895	.	.
instant_bo...	instant_boo...	ASE	Average Squared Error	0.190614	0.200796	0.194369
instant_bo...	instant_boo...	AVERR	Average Error Function	0.552784	0.584279	0.559492
instant_bo...	instant_boo...	DFE	Degrees of Freedom for Error	1708	.	.
instant_bo...	instant_boo...	DFM	Model Degrees of Freedom	15	.	.
instant_bo...	instant_boo...	DFT	Total Degrees of Freedom	1723	.	.
instant_bo...	instant_boo...	DIV	Divisor for ASE	3446	2586	2592
instant_bo...	instant_boo...	ERR	Error Function	1904.895	1510.946	1450.202
instant_bo...	instant_boo...	FPE	Final Prediction Error	0.193962	.	.
instant_bo...	instant_boo...	MAX	Maximum Absolute Error	0.98541	0.982917	0.989731
instant_bo...	instant_boo...	MSE	Mean Square Error	0.192288	0.200796	0.194369
instant_bo...	instant_boo...	NOBS	Sum of Frequencies	1723	1293	1296
instant_bo...	instant_boo...	NW	Number of Estimate Weights	15	.	.
instant_bo...	instant_boo...	RASE	Root Average Sum of Squares	0.436594	0.448103	0.440873
instant_bo...	instant_boo...	RFPE	Root Final Prediction Error	0.440411	.	.
instant_bo...	instant_boo...	RMSE	Root Mean Squared Error	0.438507	0.448103	0.440873
instant_bo...	instant_boo...	SBC	Schwarz's Bayesian Criterion	2016.672	.	.
instant_bo...	instant_boo...	SSE	Sum of Squared Errors	656.8563	519.2591	503.8052
instant_bo...	instant_boo...	SUMW	Sum of Case Weights Times Freq	3446	2586	2592
instant_bo...	instant_boo...	MISC	Misclassification Rate	0.300058	0.315545	0.324846

Figure 25- Logistic Regression Fit Statistic showing Misclassification rate

Data Role=VALIDATE Target=instant_bookable Target Label=instant_bookable

False Negative	True Negative	False Positive	True Positive
284	712	124	173

Figure 26- Logistic Regression Confusion Matrix

Based on the Model comparison node, Logistic Regression model seems to be the best as it has the lowest Misclassification rate. Comparing with baseline, our Target variable has a 61.98% accuracy without the model. With our Logistic Regression model, we have 68.45% accuracy. Hence, we have a working model that can be used for out-of-set data.

Selected Model	Model Node	Model Description
Y	Reg3	Regression (3)
	Tree	Decision Tree
	Neural3	Neural Network (3)
	MBR	MBR

Figure 27- Best Model Selection

Conclusion

The sharing economy and Airbnb present both opportunities and challenges. Airbnb has revolutionized the accommodation industry, making it possible for travellers to rent apartments and homes from locals, often at lower prices than traditional hotels. However, the company has also attracted criticism for contributing to a shortage of affordable housing in cities where it operates, and for creating nuisances and security issues for those living near leased properties. Critics have also accused Airbnb of exacerbating the housing crisis by taking long-term rental properties off the market. Our analysis above has shown that many of the Airbnb listings in these cities are entire homes and apartments rented out on longer term to guests all year long, which is illegal. Our clustering also shows a category of hosts that have several properties being listed contrary to the idea of Airbnb which is supposed to be shared living. Airbnb being aware of these listings not in line with the temporary accommodation idea have continued to look away and leave each country's government to deal with the menace via prohibitions, sanctions, and legislative regulations.

Our Data mining goal was to examine and ascertain if the claim about Airbnb being a social menace is true, and if Airbnb is truly creating a social problem and disrupting the accommodation industry/communities by offering houses to people who should be taking up permanent accommodation and we have been able to verify this as a fact with evidences of hosts listing multiple properties, having longer minimum nights on the average, and listings of entire homes/apartments which suggest that they are commercial listings.

Now that we have established this fact from the Clusters observed, we have built a working model- the Logistic Regression model which can be used for any out-of-dataset observation. The characteristics noted

on our Clusters will be used to ascertain on such listing to verify if the attributes portray a “temporary accommodation” which is legal or an “illegal short-term rental” which is a more permanent house letting and creates a social concern for the community, environment and globally.

Reference

Georgios Zervas, Davide Proserpio and John Byers (2013). The Rise of the Sharing Economy: Estimating the Impact of Airbnb on the Hotel Industry. Boston U. School of Management Research Paper No. 2013-16

Kyle Barron, Edward Kung and Davide Proserpio (2017). The Effect of Home-Sharing on House Prices and Rents: Evidence from Airbnb. Available at: https://papers.ssrn.com/sol3/papers.cfm?abstract_id=3006832. (Accessed on 22/04/2023)

Airbnb (2023). Airbnb Q4 2022 and full-year financial results. Available at: https://news.airbnb.com/airbnb-q4-2022-and-full-year-financial-results/ (Accessed on 22/04/2023)

Sri Rahayu Hijrah Hati, Tengku Ezni Balqiah, Arga Hananto, Elevita Yuliati (2020). A decade of systematic literature review on Airbnb: the sharing economy from a multiple stakeholder perspective. Management Department Faculty of Economics and Business, Universitas Indonesia. Available at: https://www.sciencedirect.com/science/article/pii/S2405844021023252#bib13 (Accessed on 22/04/2023)

Clever Real Estate (2021). Airbnb's Impact on the Hotel Industry: Insights From 1000 Travelers Who Use Both. Available at: https://listwithclever.com/research/airbnb-vs-hotels-study/ (Accessed on 29/04/2023)

Airbnb (2023). Responsible hosting in the United Kingdom. Available at: https://www.airbnb.co.uk/help/article/1379 (Accessed on 29/04/2023)

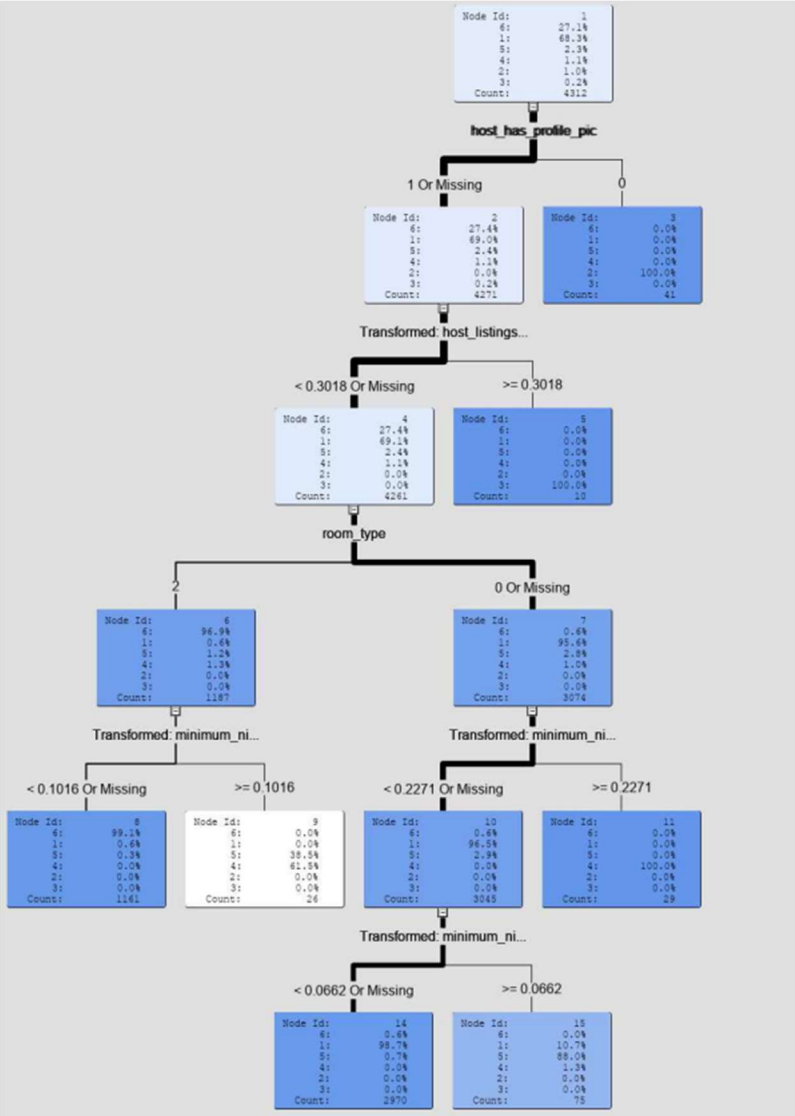
Airbnb (2023). How do I read my performance data for occupancy and rates? Available at: https://www.airbnb.co.uk/help/article/2715#:~:text=Average%20occupancy%20rate%20is%20the,booked%20across%20all%20relevant%20listings (Accessed on 22/04/2023)

Heba Baker (2019). THE AIRBNB OCCUPANCY RATE FORMULA – WHAT IT IS AND HOW TO USE IT. Mashvisor Real Estate Blog. Available at: https://www.mashvisor.com/blog/the-airbnb-occupancy-rate-formula/ (Accessed on 22/04/2023)

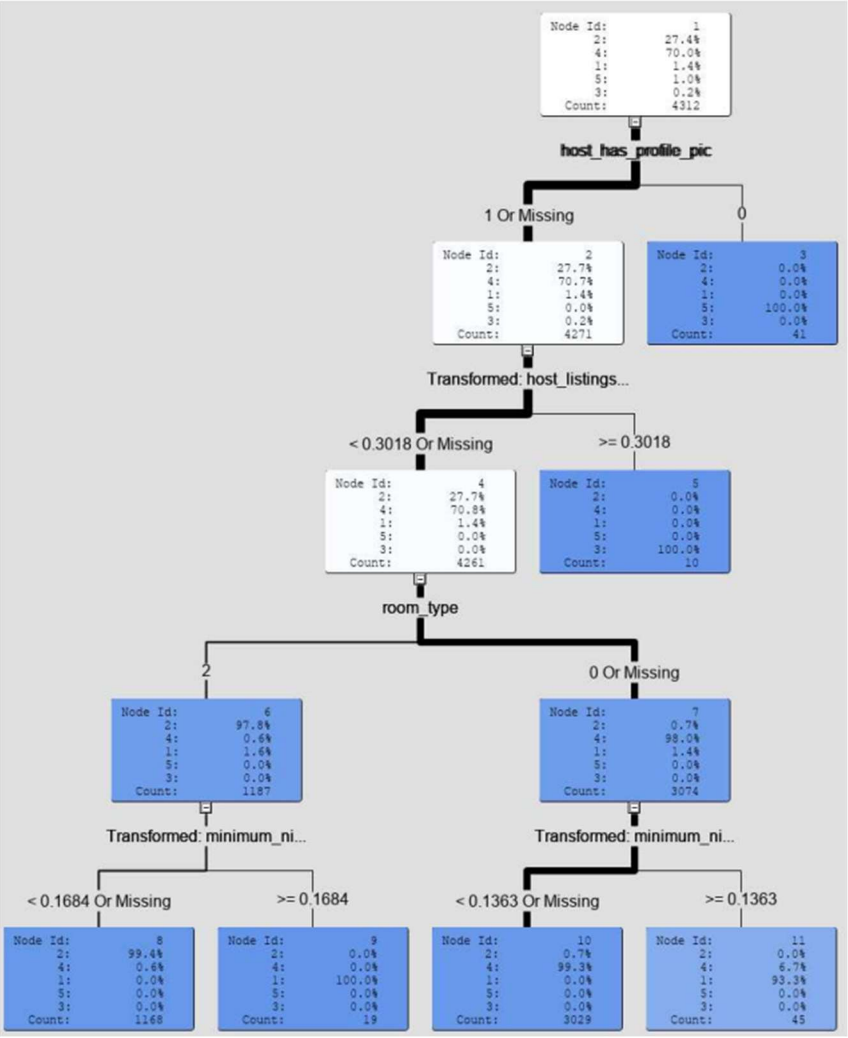
Appendix

Appendix 1

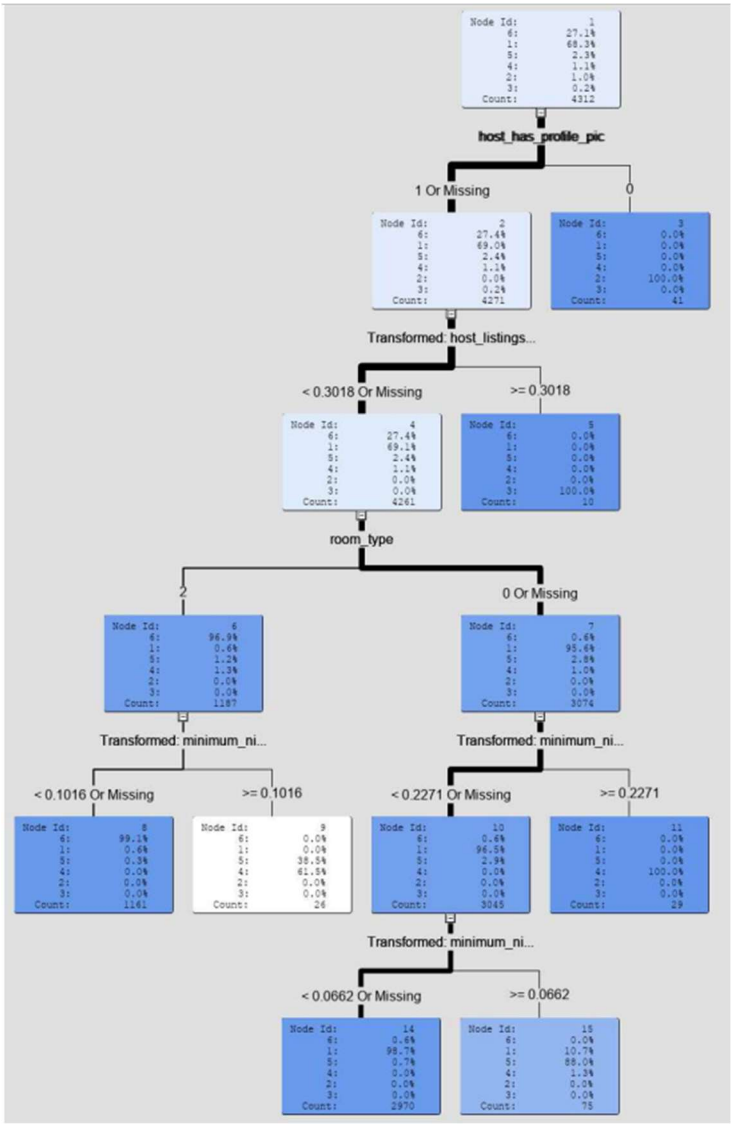
Cluster Tree- Average method



Cluster Tree- Ward method



Cluster Tree- Centroid method



Cluster Tree- 2 Maximum Clusters