COMP828 Assignment
Semester 1 2024
Analyzing Guest Satisfaction for Airbnb Apartments

Ricky Yang
ID Number: 23205919

06 June 2024

Contents

1 Dataset

1.1 Introduction

Airbnb, as a prominent product of the sharing economy, has indeed seen rapid growth in recent years, providing a unique alternative to traditional hotel accommodations with a wide range of short-term and long-term lodging options. Various studies have delved into different aspects of Airbnb, particularly focusing on the factors that influence customer satisfaction and ratings (Ju et al. 2019). One key area of research has been the impact of factors such as listing functionality, host attributes, customer reviews, and market conditions on guest satisfaction. These studies aim to understand how these elements collectively shape the overall experience of guests, ultimately influencing their review scores and the likelihood of them returning (Priporas et al. 2017). For example, the quality of service provided by hosts, the amenities offered, and the location of the accommodation all play crucial roles in determining guest satisfaction levels (Situmorang et al. 2018). Moreover, research has also highlighted the significance of rental policies, listing reputation, and location in influencing guest satisfaction levels. Factors such as the clarity of rental policies, the reputation of the listing, and the convenience of the location can greatly impact how satisfied guests are with their overall experience (Li, Hudson, and So 2019). By examining these various factors that contribute to customer satisfaction in the context of Airbnb, researchers aim to provide valuable insights for hosts and practitioners to enhance the overall guest experience and increase customer satisfaction levels (Ju et al. 2019).

In this project, we aim to extend this body of research by focusing on the factors influencing property ratings and guest satisfaction on Airbnb, specifically for properties listed as “Apartments” We will explore the relationship between high ratings (scores greater than or equal to 95) and various predictors, including price, host tenure, number of reviews, regional average price, number of bedrooms, number of bathrooms, host response rate, and room type. By building suitable models, we seek to uncover the determinants of high ratings and provide actionable insights for property owners. Additionally, we will analyze how these factors can be adjusted to optimize guest satisfaction and improve overall ratings, thereby enhancing the competitive edge of Airbnb listings.

1.2 Dataset Description

1.2.1 Source

The dataset to be analyzed is collected from https://data.world/cannata/gaairbnb and is named “AirBNB.csv” (GaAirbnb 2024).

1.2.2 File Type

Dataset file type: CSV

1.2.3 List of variables (description and type)

In the raw dataset, there are 7833 observations on 41 variables. As shown in Table 1, the dataset contains the following variables:

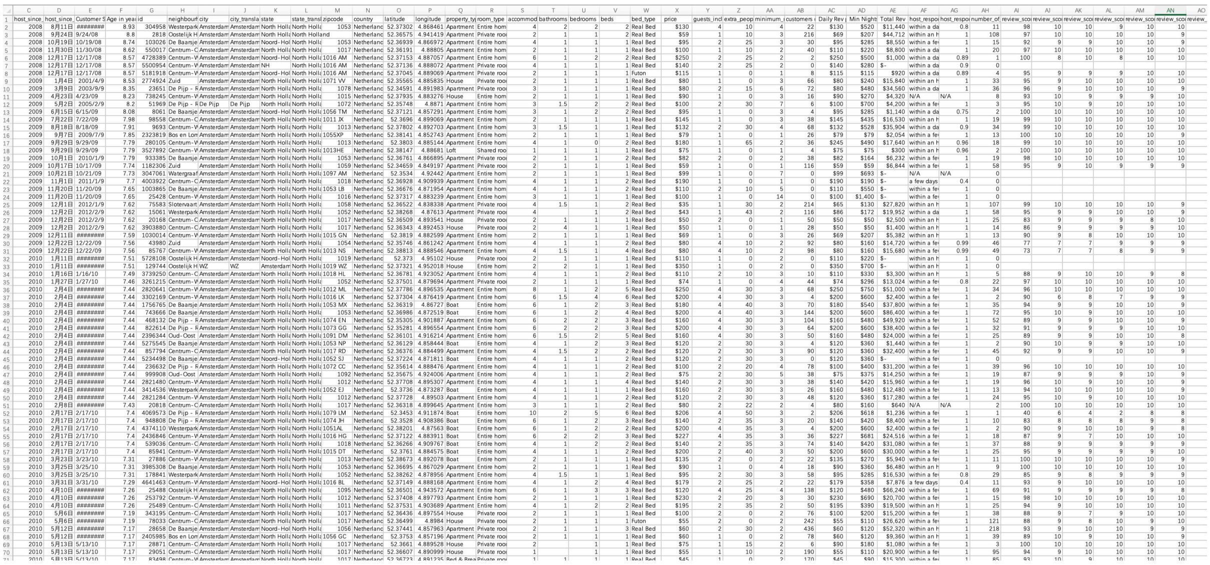
Table 1: List of Variables in the Airbnb Raw Dataset

Variable	DataType	Description
host_id	integer	Unique identifier for the host
host_name	character	Name of the host

host_since_year	integer	Year when the host started
host_since_anniversary	date	Anniversary date of the host’s start
Customer Since	integer	Duration since the customer has been with Airbnb
Age in years	integer	Age of the listing in years
id	integer	Unique identifier for the listing
neighbourhood_cleansed	character	Specific neighborhood of the listing
city	character	City where the listing is located
city_translated	character	Translated city name
state	character	State where the listing is located
state_translated	character	Translated state name
zipcode	character	Zip code of the listing location
country	character	Country of the listing
latitude	numeric	Latitude coordinate of the listing
longitude	numeric	Longitude coordinate of the listing
property_type	character	Type of property (e.g., apartment, house)
room_type	character	Type of room (e.g., entire home/apt, private room)
accommodates	integer	Number of guests the listing accommodates
bathrooms	integer	Number of bathrooms
bedrooms	integer	Number of bedrooms
beds	integer	Number of beds
bed_type	character	Type of bed (e.g., real bed, futon)
price	numeric	Price per night
guests_included	integer	Number of guests included in the price
extra_people	numeric	Cost for additional guests
minimum_nights	integer	Minimum number of nights required to book

customers @ 50% review rate	integer	Number of customers at a 50% review rate
Daily Rev per 2 guests, unless limited to 1	numeric	Daily revenue per 2 guests
Min Nights	integer	Minimum nights required
Total Rev	numeric	Total revenue generated
host__response__time	character	Time taken by the host to respond
host__response__rate	numeric	Response rate of the host
number__of__reviews	integer	Total number of reviews received
review__scores__rating	integer	Overall rating score from reviews
review__scores__accuracy	integer	Accuracy rating score from reviews
review__scores __cleanliness	integer	Cleanliness rating score from reviews
review__scores__checkin	integer	Check-in rating score from reviews
review__scores __communication	integer	Communication rating score from reviews
review__scores__location	integer	Location rating score from reviews
review__scores__value	integer	Value rating score from reviews

1.2.4 Raw Dataset Screenshot



	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R	S	T	U	V	W	X	Y	Z	AA	AB	AC	AD	AE	AF	AG	AH	AI	AJ	AK	AL	AM	AN	AO
1	host_since_	host_since_	Customer S	Age in yea	id	neighbourh	city	city_transla	state	state_transl	zipcode	country	latitude	longitude	property_ty	room_type	accommod	bathrooms	bedrooms	beds	bed_type	price	guests_incl	extra_peop	minimum_	customers	Daily Rev	Min Nights	Total Rev	host_respo	host_respo	number_of_	review_sco	review_sco	review_sco	review_sco	review_sco	review_sco	review_
2	2008	8月11日	########	8.93	304958	Westerpark	Amsterdam	Amsterdam	North Holla	North Holla	1053	Netherlanc	52.37302	4.868461	Apartment	Entire hom	4	2	2	2	Real Bed	$130	4	10	4	22	$130	$520	$11,440	within a da	0.8	11	98	10	10	9	10	10	
3	2008	9月24日	9/24/08	8.8	2818	Oostelijk H	Amsterdam	Amsterdam	North Holla	North Holland		Netherlanc	52.36575	4.941419	Apartment	Private roo	2	1	1	2	Real Bed	$59	1	10	3	216	$69	$207	$44,712	within an h	1	108	97	10	10	10	10	9	
4	2008	10月19日	10/19/08	8.74	103026	De Baarsje	Amsterdam	Amsterdam	Noord-Hol	North Holla	1053	Netherlanc	52.36939	4.866972	Apartment	Entire hom	4	1	1	1	Real Bed	$95	2	25	3	30	$95	$285	$8,550	within a fe	1	15	92	9	9	10	10	9	
5	2008	11月30日	11/30/08	8.62	550017	Centrum-C	Amsterdam	Amsterdam	North Holla	North Holla	1017	Netherlanc	52.36191	4.88805	Apartment	Entire hom	2	1	1	1	Real Bed	$100	1	10	2	40	$110	$220	$8,800	within a da	1	20	97	10	10	10	10	10	
6	2008	12月17日	12/17/08	8.57	4728389	Centrum-W	Amsterdam	Amsterdam	Noord-Hol	North Holla	1016 AM	Netherlanc	52.37153	4.887057	Apartment	Entire hom	6	1	2	2	Real Bed	$250	2	25	2	2	$250	$500	$1,000	within a da	0.89	1	100	8	10	8	10	10	
7	2008	12月17日	12/17/08	8.57	5500954	Centrum-W	Amsterdam	Amsterdam	NH	North Holla	1016 AM	Netherlanc	52.37136	4.888072	Apartment	Private roo	4	1	1	1	Real Bed	$140	2	25	2	0	$140	$280	$-	within a da	0.9	0							
8	2008	12月17日	12/17/08	8.57	5181918	Centrum-W	Amsterdam	Amsterdam	Noord-Hol	North Holla	1016 AM	Netherlanc	52.37045	4.889069	Apartment	Private roo	2	1	1	1	Futon	$115	1	0	1	8	$115	$115	$920	within a da	0.89	4	95	9	9	9	10	10	
9	2009	1月4日	2001/4/9	8.53	2774924	Zuid	Amsterdam	Amsterdam	North Holla	North Holla	1071 VV	Netherlanc	52.35565	4.885835	House	Private roo	2	1	1	1	Real Bed	$80	1	0	3	66	$80	$240	$15,840	within an h	1	33	95	9	10	10	10	10	
10	2009	3月9日	2003/9/9	8.35	23651	De Pijp - R	Amsterdam	Amsterdam	North Holla	North Holla	1078	Netherlanc	52.34591	4.891983	Apartment	Private roo	3	1	1	1	Real Bed	$80	2	15	6	72	$80	$480	$34,560	within a da	1	36	96	9	10	10	10	9	
11	2009	4月23日	4/23/09	8.23	738245	Centrum-W	Amsterdam	Amsterdam	North Holla	North Holla	1015	Netherlanc	52.37935	4.883276	House	Entire hom	2	1	1	1	Real Bed	$90	1	0	3	16	$90	$270	$4,320	N/A	N/A	8	93	10	9	9	9	10	
12	2009	5月2日	2005/2/9	8.2	51969	De Pijp - R	De Pijp	De Pijp	North Holla	North Holla	1072	Netherlanc	52.35748	4.8871	Apartment	Entire hom	3	1.5	2	2	Real Bed	$100	2	30	7	6	$100	$700	$4,200	within a fe	1	3	95	10	9	10	10	10	
13	2009	6月15日	6/15/09	8.08	8061	De Baarsje	Amsterdam	Amsterdam	Noord-Hol	North Holla	1056 TM	Netherlanc	52.37121	4.857291	Apartment	Entire hom	3	1	2	2	Real Bed	$95	1	0	3	4	$95	$285	$1,140	within a da	0.75	2	100	10	10	10	10	10	
14	2009	7月22日	7/22/09	7.98	98558	Centrum-C	Amsterdam	Amsterdam	North Holla	North Holla	1011 JX	Netherlanc	52.3696	4.899069	Apartment	Entire hom	2	1	1	1	Real Bed	$145	1	0	3	38	$145	$435	$16,530	within an h	1	19	99	10	10	10	10	10	
15	2009	8月18日	8/18/09	7.91	9693	Centrum-W	Amsterdam	Amsterdam	North Holla	North Holla	1013	Netherlanc	52.37802	4.892703	Apartment	Entire hom	3	1.5	1	1	Real Bed	$132	2	30	4	68	$132	$528	$35,904	within a da	0.9	34	99	10	10	10	10	10	
16	2009	9月7日	2009/7/9	7.85	2323819	Bos en Lom	Amsterdam	Amsterdam	North Holla	North Holla	1055XP	Netherlanc	52.38141	4.852743	Apartment	Entire hom	2	1	1	1	Real Bed	$79	1	0	1	26	$79	$79	$2,054	within a fe	1	13	100	10	10	10	10	9	
17	2009	9月29日	9/29/09	7.79	280105	Centrum-W	Amsterdam	Amsterdam	North Holla	North Holla	1013	Netherlanc	52.3803	4.885144	Apartment	Entire hom	4	1	0	2	Real Bed	$180	1	65	2	36	$245	$490	$17,640	within an h	0.96	18	99	10	10	10	10	10	
18	2009	9月29日	9/29/09	7.79	3527892	Centrum-W	Amsterdam	Amsterdam	North Holla	North Holla	1013HE	Netherlanc	52.38147	4.88681	Loft	Shared roo	1	1	1	1	Real Bed	$75	1	0	1	4	$75	$75	$300	within an h	0.96	2	100	10	10	10	10	10	
19	2009	10月1日	2010/1/9	7.79	933385	De Baarsje	Amsterdam	Amsterdam	North Holla	North Holla	1053	Netherlanc	52.36761	4.866895	Apartment	Private roo	2	1	1	2	Real Bed	$82	2	0	2	38	$82	$164	$6,232	within a fe	1	19	98	10	10	10	10	10	
20	2009	10月17日	10/17/09	7.74	1182306	Zuid	Amsterdam	Amsterdam	North Holla	North Holla	1059	Netherlanc	52.34659	4.849197	Apartment	Private roo	2	1	1	1	Real Bed	$59	1	0	1	116	$59	$59	$6,844	within a fe	1	58	95	9	10	9	9	9	
21	2009	10月21日	10/21/09	7.73	3047061	Watergraaf	Amsterdam	Amsterdam	North Holla	North Holla	1097 AM	Netherlanc	52.3534	4.92442	Apartment	Entire hom	2	1	1	1	Real Bed	$99	1	0	7	0	$99	$693	$-	N/A	N/A	0							
22	2009	11月1日	2011/1/9	7.7	4003922	Centrum-C	Amsterdam	Amsterdam	North Holla	North Holla	1018	Netherlanc	52.36928	4.909939	Apartment	Entire hom	4	1	1	2	Real Bed	$190	1	0	1	0	$190	$190	$-	a few days	0.4	0							
23	2009	11月20日	11/20/09	7.65	1003865	De Baarsje	Amsterdam	Amsterdam	North Holla	North Holla	1053 LB	Netherlanc	52.36676	4.871954	Apartment	Entire hom	4	1	1	2	Real Bed	$110	2	10	5	0	$110	$550	$-	within a fe	1	0							
24	2009	11月20日	11/20/09	7.65	25428	Centrum-W	Amsterdam	Amsterdam	North Holla	North Holla	1016	Netherlanc	52.37317	4.883239	Apartment	Entire hom	3	1	1	1	Real Bed	$100	1	0	14	0	$100	$1,400	$-	within a fe	1	0							
25	2009	12月1日	2012/1/9	7.62	75583	Slotervaart	Amsterdam	Amsterdam	North Holla	North Holla	1058	Netherlanc	52.36522	4.838338	Apartment	Private roo	4	1.5	1	2	Real Bed	$35	1	30	2	214	$65	$130	$27,820	within an h	1	107	99	10	10	10	10	9	
26	2009	12月2日	2012/2/9	7.62	15061	Westerpark	Amsterdam	Amsterdam	North Holla	North Holla	1052	Netherlanc	52.38268	4.87613	Apartment	Private roo	4		1	2	Real Bed	$43	1	43	2	116	$86	$172	$19,952	within a da	1	58	95	9	9	10	10	9	
27	2009	12月2日	2012/2/9	7.62	20168	Centrum-C	Amsterdam	Amsterdam	North Holla	North Holla	1017	Netherlanc	52.36509	4.893541	House	Private roo	2	1	1	1	Real Bed	$50	2	0	1	50	$50	$50	$2,500	within an h	1	25	83	9	9	9	8	10	
28	2009	12月2日	2012/2/9	7.62	3903880	Centrum-C	Amsterdam	Amsterdam	North Holla	North Holla	1017	Netherlanc	52.36343	4.892453	House	Private roo	2	4	1	1	Real Bed	$50	1	0	1	28	$50	$50	$1,400	within an h	1	14	86	9	9	9	9	10	
29	2009	12月11日	########	7.59	1030014	Centrum-W	Amsterdam	Amsterdam	North Holla	North Holla	1015 GN	Netherlanc	52.3819	4.882599	Apartment	Entire hom	2	1	1	1	Real Bed	$69	1	0	3	26	$69	$207	$5,382	within an h	1	13	90	9	8	10	10	10	
30	2009	12月22日	12/22/09	7.56	43980	Zuid	Amsterdam	Amsterdam	North Holla	North Holla	1054	Netherlanc	52.35746	4.861242	Apartment	Entire hom	4	1	1	2	Real Bed	$80	4	10	2	92	$80	$160	$14,720	within a fe	0.99	46	77	7	7	9	9	9	
31	2009	12月22日	12/22/09	7.56	85767	Centrum-W	Amsterdam	Amsterdam	North Holla	North Holla	1013 NS	Netherlanc	52.38813	4.888546	Apartment	Entire hom	4	1.5	1	4	Real Bed	$80	4	10	2	98	$80	$160	$15,680	within a fe	0.99	49	73	7	7	8	9	9	
32	2010	1月11日	########	7.51	5728108	Oostelijk H	Amsterdam	Amsterdam	Noord-Hol	North Holla	1019	Netherlanc	52.373	4.95102	House	Private roo	2	1	1	1	Real Bed	$110	1	0	2	0	$110	$220	$-	within an h	1	0							
33	2010	1月11日	########	7.51	129744	Oostelijk H	WZ	WZ	Amsterdam	North Holla	1019 WZ	Netherlanc	52.37321	4.952018	House	Entire hom	2	2	1	1	Real Bed	$350	1	0	2	0	$350	$700	$-	within an h	1	0							
34	2010	1月16日	1/16/10	7.49	3739250	Centrum-C	Amsterdam	Amsterdam	North Holla	North Holla	1018 HL	Netherlanc	52.36781	4.923052	Apartment	Entire hom	4	1	1	2	Real Bed	$110	2	10	3	10	$110	$330	$3,300	within an h	1	5	88	9	10	10	9	8	
35	2010	1月27日	1/27/10	7.46	3261215	Centrum-W	Amsterdam	Amsterdam	North Holla	North Holla	1052	Netherlanc	52.37501	4.879694	Apartment	Private roo	2	1	1	1	Real Bed	$74	1	0	4	44	$74	$296	$13,024	within a fe	0.8	22	97	10	10	10	10	10	
36	2010	2月4日	########	7.44	2820641	Centrum-W	Amsterdam	Amsterdam	North Holla	North Holla	1012 ML	Netherlanc	52.37786	4.896535	Apartment	Entire hom	8	1	2	5	Real Bed	$250	4	30	3	68	$250	$750	$51,000	within a fe	1	34	96	10	10	10	10	10	
37	2010	2月4日	########	7.44	3302169	Centrum-W	Amsterdam	Amsterdam	North Holla	North Holla	1016 LK	Netherlanc	52.37304	4.876419	Apartment	Entire hom	6	1.5	4	6	Real Bed	$200	4	30	3	4	$200	$600	$2,400	within a fe	1	2	90	6	8	7	9	9	
38	2010	2月4日	########	7.44	1756765	De Baarsje	Amsterdam	Amsterdam	North Holla	North Holla	1053 MX	Netherlanc	52.36319	4.86727	Boat	Entire hom	6	1	2	3	Real Bed	$180	4	40	3	70	$180	$540	$37,800	within a fe	1	35	94	9	9	10	10	9	
39	2010	2月4日	########	7.44	743666	De Baarsje	Amsterdam	Amsterdam	North Holla	North Holla	1053	Netherlanc	52.36986	4.872519	Boat	Entire hom	6	1	2	4	Real Bed	$200	4	40	3	144	$200	$600	$86,400	within a fe	1	72	95	10	9	10	10	10	
40	2010	2月4日	########	7.44	468132	De Pijp - R	Amsterdam	Amsterdam	North Holla	North Holla	1074 EN	Netherlanc	52.35305	4.901887	Apartment	Entire hom	6	2	2	3	Real Bed	$160	4	30	3	104	$160	$480	$49,920	within a fe	1	52	89	9	9	10	10	9	
41	2010	2月4日	########	7.44	822614	De Pijp - R	Amsterdam	Amsterdam	North Holla	North Holla	1073 GG	Netherlanc	52.35281	4.896554	Apartment	Entire hom	6	2	2	3	Real Bed	$200	4	30	3	64	$200	$600	$38,400	within a fe	1	32	91	9	9	9	10	10	
42	2010	2月4日	########	7.44	2396344	Oud-Oost	Amsterdam	Amsterdam	North Holla	North Holla	1091 DM	Netherlanc	52.36101	4.916214	Apartment	Entire hom	6	1.5	2	5	Real Bed	$160	4	30	3	50	$160	$480	$24,000	within a fe	1	25	89	9	9	10	10	8	
43	2010	2月4日	########	7.44	5275545	De Baarsje	Amsterdam	Amsterdam	North Holla	North Holla	1053 NP	Netherlanc	52.36129	4.858444	Boat	Entire hom	4	1	2	3	Real Bed	$120	2	30	3	4	$120	$360	$1,440	within a fe	1	2	90	10	9	9	10	10	
44	2010	2月4日	########	7.44	857794	Centrum-C	Amsterdam	Amsterdam	North Holla	North Holla	1017 RD	Netherlanc	52.36376	4.884499	Apartment	Entire hom	4	1.5	2	2	Real Bed	$120	2	30	3	90	$120	$360	$32,400	within a fe	1	45	92	9	9	10	10	9	
45	2010	2月4日	########	7.44	5234498	De Baarsje	Amsterdam	Amsterdam	Noord-Hol	North Holla	1052 SJ	Netherlanc	52.37224	4.871811	Boat	Entire hom	4	1	1	2	Real Bed	$120	2	30	3	0	$120	$360	$-	within a fe	1	0							
46	2010	2月4日	########	7.44	236632	De Pijp - R	Amsterdam	Amsterdam	North Holla	North Holla	1072 CC	Netherlanc	52.35614	4.888476	Apartment	Entire hom	4	1	1	2	Real Bed	$100	2	20	4	78	$100	$400	$31,200	within a fe	1	39	96	10	10	10	10	10	
47	2010	2月4日	########	7.44	999908	Oud-Oost	Amsterdam	Amsterdam	North Holla	North Holla	1092	Netherlanc	52.35675	4.924006	Apartment	Entire hom	4	1	1	2	Real Bed	$75	2	30	5	38	$75	$375	$14,250	within a fe	1	19	87	9	9	9	10	9	
48	2010	2月4日	########	7.44	2821480	Centrum-W	Amsterdam	Amsterdam	North Holla	North Holla	1012	Netherlanc	52.37708	4.895307	Apartment	Entire hom	4	1	1	4	Real Bed	$140	2	30	3	38	$140	$420	$15,960	within a fe	1	19	96	10	9	10	10	9	
49	2010	2月4日	########	7.44	3414536	Westerpark	Amsterdam	Amsterdam	North Holla	North Holla	1052 EJ	Netherlanc	52.3736	4.873287	Boat	Entire hom	4	1	1	1	Real Bed	$160	2	30	3	26	$160	$480	$12,480	within a fe	1	13	94	10	10	10	10	9	
50	2010	2月4日	########	7.44	2821284	Centrum-W	Amsterdam	Amsterdam	North Holla	North Holla	1012	Netherlanc	52.37728	4.89503	Apartment	Entire hom	4	1	1	2	Real Bed	$120	2	30	3	48	$120	$360	$17,280	within a fe	1	24	95	10	9	10	10	10	
51	2010	2月8日	########	7.43	20818	Centrum-C	Amsterdam	Amsterdam	North Holla	North Holla	1017	Netherlanc	52.36318	4.899645	Apartment	Entire hom	3	1	1	2	Real Bed	$80	2	22	2	4	$80	$160	$640	N/A	N/A	2	100	10	10	10	10	10	
52	2010	2月17日	2/17/10	7.4	4069573	De Pijp - R	Amsterdam	Amsterdam	North Holla	North Holla	1079 LM	Netherlanc	52.3453	4.911874	Boat	Entire hom	10	2	5	6	Real Bed	$206	4	50	3	2	$206	$618	$1,236	within a fe	1	1	40	6	4	2	8	8	
53	2010	2月17日	2/17/10	7.4	948808	De Pijp - R	Amsterdam	Amsterdam	North Holla	North Holla	1074 JH	Netherlanc	52.3528	4.908386	Boat	Entire hom	6	1	2	3	Real Bed	$140	2	35	3	20	$140	$420	$8,400	within a fe	1	10	83	8	8	8	9	8	
54	2010	2月17日	2/17/10	7.4	4374110	Westerpark	Amsterdam	Amsterdam	North Holla	North Holla	1051AL	Netherlanc	52.38201	4.87563	Boat	Entire hom	6	1	2	2	Real Bed	$200	4	35	3	4	$200	$600	$2,400	within a fe	1	2	90	9	10	10	9	8	
55	2010	2月17日	2/17/10	7.4	2436846	Centrum-W	Amsterdam	Amsterdam	North Holla	North Holla	1016 HG	Netherlanc	52.37122	4.883911	Boat	Entire hom	6	2	2	3	Real Bed	$227	4	35	3	36	$227	$681	$24,516	within a fe	1	18	87	9	7	10	10	10	
56	2010	2月17日	2/17/10	7.4	539036	Centrum-C	Amsterdam	Amsterdam	North Holla	North Holla	1018	Netherlanc	52.36266	4.909767	Boat	Entire hom	4	1	2	2	Real Bed	$140	2	35	3	74	$140	$420	$31,080	within a fe	1	37	88	9	9	9	9	9	
57	2010	2月17日	2/17/10	7.4	85941	Centrum-W	Amsterdam	Amsterdam	North Holla	North Holla	1015 DT	Netherlanc	52.3761	4.884575	Boat	Entire hom	4	1	1	2	Real Bed	$200	2	40	3	50	$200	$600	$30,000	within a fe	1	25	95	9	9	9	9	10	
58	2010	3月23日	3/23/10	7.31	27886	Centrum-W	Amsterdam	Amsterdam	North Holla	North Holla	1013	Netherlanc	52.38673	4.892078	Boat	Entire hom	2	1	1	1	Real Bed	$135	2	0	2	22	$135	$270	$5,940	within a fe	1	11	100	10	10	10	10	9	
59	2010	3月25日	3/25/10	7.31	3985308	De Baarsje	Amsterdam	Amsterdam	North Holla	North Holla	1053	Netherlanc	52.36695	4.867029	Apartment	Entire hom	2	1	1	1	Real Bed	$90	1	0	4	18	$90	$360	$6,480	within an h	1	9	100	10	10	10	10	10	
60	2010	3月25日	3/25/10	7.31	178841	Westerpark	Amsterdam	Amsterdam	North Holla	North Holla	1052	Netherlanc	52.38262	4.878956	Apartment	Entire hom	4	1.5	1	1	Real Bed	$95	2	30	3	58	$95	$285	$16,530	within a fe	0.8	29	85	9	8	9	9	9	
61	2010	3月31日	3/31/10	7.29	4641463	Centrum-W	Amsterdam	Amsterdam	Noord-Hol	North Holla	1016 BL	Netherlanc	52.37149	4.888168	Apartment	Entire hom	4	1	2	4	Real Bed	$179	2	25	2	22	$179	$358	$7,876	a few days	0.4	11	93	9	10	10	10	10	
62	2010	4月10日	########	7.26	25488	Oostelijk H	Amsterdam	Amsterdam	North Holla	North Holla	1095	Netherlanc	52.36501	4.943572	Apartment	Entire hom	6	1	3	3	Real Bed	$120	4	25	4	138	$120	$480	$66,240	within a fe	1	69	91	9	9	9	9	8	
63	2010	4月10日	########	7.26	253792	Centrum-W	Amsterdam	Amsterdam	North Holla	North Holla	1012	Netherlanc	52.37408	4.897793	Apartment	Entire hom	2	1	1	1	Real Bed	$230	2	20	3	30	$230	$690	$20,700	within a fe	1	15	98	10	10	10	10	10	
64	2010	4月10日	########	7.26	25489	Centrum-C	Amsterdam	Amsterdam	North Holla	North Holla	1011	Netherlanc	52.37531	4.903689	Apartment	Entire hom	4	1	2	2	Real Bed	$195	2	35	2	50	$195	$390	$19,500	within an h	1	25	94	9	10	10	10	10	
65	2010	5月6日	########	7.19	343195	Centrum-C	Amsterdam	Amsterdam	North Holla	North Holla	1017	Netherlanc	52.36436	4.897554	House	Private roo	2	1	1	1	Real Bed	$100	2	0	2	76	$100	$200	$15,200	within a fe	1	38	88	9	7	9	10	10	
66	2010	5月6日	########	7.19	78033	Centrum-C	Amsterdam	Amsterdam	North Holla	North Holla	1017	Netherlanc	52.36499	4.8984	House	Private roo	2	1	1	1	Futon	$55	2	0	2	242	$55	$110	$26,620	within a fe	1	121	88	9	8	10	9	10	
67	2010	5月12日	########	7.17	28658	De Baarsje	Amsterdam	Amsterdam	Noord-Hol	North Holla	1056	Netherlanc	52.37441	4.857963	Apartment	Private roo	2	1	1	3	Real Bed	$60	2	30	2	436	$60	$120	$52,320	within an h	1	218	93	9	10	10	9	9	
68	2010	5月12日	########	7.17	2405985	Bos en Lom	Amsterdam	Amsterdam	North Holla	North Holla	1056 GC	Netherlanc	52.3753	4.857196	Apartment	Private roo	2	1	1	1	Real Bed	$60	1	0	2	78	$60	$120	$9,360	within an h	1	39	89	10	9	10	10	9	
69	2010	5月13日	5/13/10	7.17	28871	Centrum-C	Amsterdam	Amsterdam	North Holla	North Holla	1017	Netherlanc	52.3661	4.889528	House	Private roo	2		1	1	Real Bed	$75	1	15	2	6	$90	$180	$1,080	within a fe	1	3	100	10	10	10	10	10	
70	2010	5月13日	5/13/10	7.17	29051	Centrum-C	Amsterdam	Amsterdam	North Holla	North Holla	1017	Netherlanc	52.36607	4.890999	House	Private roo	1		1	1	Real Bed	$55	1	10	2	190	$55	$110	$20,900	within a fe	1	95	94	9	10	10	10	10	
71	2010	5月13日	5/13/10	7.17	83498	Centrum-W	Amsterdam	Amsterdam	North Holla	North Holla	1017	Netherlanc	52.36723	4.891235	Bed & Brea	Private roo	1	1	1	1	Real Bed	$45	1	0	2	170	$45	$90	$15,300	within a fe	1	85	93	10	9	10	10	10	

Figure 1: Airbnb Raw Dataset Screenshot

1.2.5 Type of Data Cleaning Expected

The variables selected for analysis are `price`, `accommodates`, `bathrooms`, `bedrooms`, `room_type`, `host_response_rate`, and `review_scores_rating`.

1.2.5.1 Data Wrangling Steps We propose the following data wrangling steps:

1. **Filtering Observations** :
   - Filter out observations related to property_type "apartment".
   - Select the necessary variables for analysis.
2. **Handling Missing Values** :
   - Identify and handle missing values appropriately. This may include removing rows with significant missing data or imputing missing values using appropriate methods.
3. **Data Type Conversion** :
   - Ensure all columns are of the correct data type. Convert columns to appropriate types (e.g., numeric, factor, date) as necessary.
4. **Removing Duplicates** :
   - Check for and remove any duplicate rows to ensure each observation is unique.
5. **Outlier Detection and Treatment** :
   - Identify and treat outliers that may skew the analysis. This can involve removing extreme values or transforming them.
6. **Creating New Features** :
   - Create new features that may be useful for analysis, such as calculating the age of the listing from the `host_since_year` variable.
7. **Encoding Categorical Variables** :
   - Encode categorical variables (e.g., room_type) using techniques such as one-hot encoding or label encoding to make them suitable for analysis and modeling.

These steps will help ensure that the dataset is clean, consistent, and ready for further analysis and modeling.

1.3 Three proposed research questions

1.3.1 Research question 1

Do properties with prices at different levels relative to the regional average price have significantly different ratings?

1. **Variables Considered:**
    - **Price:** The nightly rental price of the property.
    - **Region Average Price:** The average price in the region where the property is located.
    - **Review Scores Rating:** The overall rating given by guests.
2. **Type of Analysis:**
    - **Grouping:** Properties are divided into groups based on their price relative to the regional average price.
    - **Descriptive Statistics:** Mean and standard deviation of review scores for each price group.
    - **ANOVA:** To test for significant differences in review scores across the price groups.
    - **T-Test:** To compare the review scores between two specific groups (below_avg and above_avg).
    - **Visualization:** Box plot to visualize the distribution of review scores across price groups.
3. **Libraries and Required R Functions:**
    - **dplyr:** For data manipulation and grouping.
    - **ggplot2:** For creating visualizations.
    - **stats:** For performing ANOVA and t-tests.
    - **knitr:** For creating tables in the report.

1.3.2 Research question 2

Can property ratings be predicted using price, regional average price, number of bedrooms , number of accommodates ,room types, and number of bathrooms ?

1. **Variables Considered:**
    - **Price:** The nightly rental price of the property.
    - **Region Average Price:** The average price in the region where the property is located.
    - **Bedrooms:** The number of bedrooms in the property.
    - **Accommodates:** The number of guests the property can accommodate.
    - **Bathrooms:** The number of bathrooms in the property.
    - **Room Type:** The type of room (e.g., entire home/apartment, private room, shared room).
    - **Review Scores Rating:** The overall rating given by guests.
2. **Type of Analysis:**
    - **Linear Regression:** To model the relationship between the review scores rating and the predictor variables.
    - **Residual Analysis:** To assess the fit of the linear regression model by examining the residuals.
    - **QQ Plot:** To check the normality of residuals, which is an assumption of linear regression.
3. **Libraries and Required R Functions:**
    - **dplyr:** For data manipulation.
    - **ggplot2:** For creating visualizations.

- **car:** For generating QQ plots.
- **stats:** For building and analyzing the linear regression model.
- **knitr:** For creating tables in the report.

1.3.3 Research question 3

Can high ratings (scores greater than or equal to 95) be predicted using price, host tenure (measured by “host_since_year”), number of reviews, regional average price, and host response rate?

1. **Variables Considered:**
   - **Price:** The nightly rental price of the property.
   - **Host Since Year:** The year when the host joined Airbnb, indicating their tenure in the platform.
   - **Number of Reviews:** The total number of reviews received by the property.
   - **Regional Average Price:** The average price in the region where the property is located, providing context for pricing.
   - **Host Response Rate:** The rate at which the host responds to inquiries, reflecting host engagement and responsiveness.
2. **Type of Analysis:**
   - **Logistic Regression:** To model the relationship between the binary high rating variable and the predictor variables.
   - **Confusion Matrix:** To evaluate the accuracy, sensitivity, and specificity of the logistic regression model.
   - **ROC Curve and AUC:** To assess the model’s ability to distinguish between high and low ratings.
3. **Libraries and Required R Functions:**
   - **dplyr:** For data manipulation.
   - **ggplot2:** For creating visualizations.
   - **caret:** For generating confusion matrix and performance metrics.
   - **ROCR:** For generating ROC curve and calculating AUC.
   - **stats:** For building and analyzing the logistic regression model.
   - **knitr:** For creating tables in the report.

2 Data Import and Cleaning

This section will prepare a well-structured dataset for subsequent analytics containing three steps: import, cleaning, and tidying.

2.1 Data Import

Since the raw data is in CSV format, we use the function “read.csv” to import it. Then, we filter out the observations related to the property type “Apartment.”

```r
# import dataset and filter out apartment
tb<-read.csv("AirBnb.csv") %>%   filter(property_type=="Apartment")

# Display total number of rows
total_rows <- nrow(tb)
```

The total number of rows in the filtered data is 6263.

2.2 Cleaning

2.2.1 Tidying the variable names

Tidying variable names ensures consistency and clarity, making data more accessible to manipulate and analyze. For instance, columns like “customers @ 50% review rate” become “customers_50_review_rate”, “Daily Rev per 2 guests, unless limited to 1” becomes “daily_rev_per_2_guests”, and “Min Nights” becomes “min_nights”, improving overall data quality.

```r
# Clean variable names
clean_names <- function(names) {
  names %>%
    tolower() %>%
    str_replace_all(" ", "_") %>%
    str_replace_all("[^[:alnum:]_]", "")
}

colnames(tb) <- clean_names(colnames(tb))
```

Table 2: Columns after Tidying the variable names

Column Names	Column Types	NA Count
host_id	integer	0
host_name	character	0
host_since_year	integer	0
host_since_anniversary	character	0
customersince	character	0
ageinyears	numeric	0
id	integer	0
neighbourhood_cleansed	character	0
city	character	0
city_translated	character	0
state	character	0
state_translated	character	0

zipcode	character	0
country	character	0
latitude	numeric	0
longitude	numeric	0
property_type	character	0
room_type	character	0
accommodates	integer	0
bathrooms	numeric	41
bedrooms	integer	13
beds	integer	8
bed_type	character	0
price	character	0
guests_included	integer	0
extra_people	integer	0
minimum_nights	integer	0
customers50reviewrate	integer	0
xdailyrevper2guestsunlesslimitedto1	character	0
minnights	character	0
totalrev	character	0
host_response_time	character	0
host_response_rate	character	0
number_of_reviews	integer	0
review_scores_rating	integer	1297
review_scores_accuracy	integer	1303
review_scores_cleanliness	integer	1303
review_scores_checkin	integer	1303
review_scores_communication	integer	1304
review_scores_location	integer	1304
review_scores_value	integer	1305

2.2.2 Select variables

Although our dataset includes many columns, we have selected a subset for analysis based on the following reasons:

Variables Selected for Analysis:

- **Price:** The nightly rental price of the property.
- **Regional Average Price:** The average price in the region where the property is located.
- **Number of Bedrooms:** The number of bedrooms in the property.
- **Number of Bathrooms:** The number of bathrooms in the property.
- **Host Response Rate:** The rate at which the host responds to inquiries.
- **Room Type:** The type of room (e.g., entire home/apartment, private room, shared room).
- **Host Since Year:** The year when the host joined Airbnb, indicating their tenure in the platform.
- **Number of Reviews:** The total number of reviews received by the property.

Reasons for Selection:

1. **Objective Limitations:** Variables related to location (e.g., latitude, longitude, neighbourhood) would require more detailed information about transportation, local amenities, population density, and city layout to be accurately analyzed.
2. **Complexity:** This experiment focuses on learning statistical methods using R. It is not intended to be a comprehensive analysis but rather an attempt to explore the impact of a few common variables on property ratings.
3. **Future Scope:** More data types and sophisticated analysis methods can be considered in future studies to provide a more in-depth and comprehensive analysis.

```r
# select the necessary variables
tb.selected<-tb %>%
  select(price,accommodates,bathrooms,bedrooms,room_type,
                host_response_rate,review_scores_rating,neighbourhood_cleansed
           ,host_since_year  ,  number_of_reviews)
```

2.2.3 Remove Duplicate Rows

Removing duplicate rows is crucial in R statistics to ensure data accuracy. Duplicate entries can skew analysis and lead to incorrect results. For instance, having multiple identical listings would distort average price calculations and other statistical metrics, leading to misleading conclusions.

```r
# Remove duplicate rows
tb.selected <- unique(tb.selected)
```

2.2.4 Renaming variables

Renaming variables is crucial in data processing, enhancing readability and ensuring consistency throughout the dataset. Giving columns meaningful names makes the data easier to understand and work with. For instance, renaming "neighbourhood_cleansed" to a more intuitive name like "region" clarifies its meaning. This improved clarity is instrumental when performing calculations such as datermining the region's average price, ensuring accurate and efficient analysis.

```r
tb.selected <- tb.selected %>%
  rename(region = neighbourhood_cleansed)
```

2.2.5 Data type conversion

Data type conversion is essential in data processing, ensuring that each variable is in the appropriate format for analysis. Converting data types helps prevent errors and allows statistical and computational methods to be applied correctly. Given a view of the selected variables, we found that some variables are in the wrong type. For example, the variables "price" and "host_response_rate" should be numerical but were in character type. Converting these to numeric types allows for accurate calculations and meaningful analysis, such as determining average prices and response rates.

Additionally, certain variables representing counts or quantities, such as "accommodates," "bathrooms," and "bedrooms," should be converted to integer or numeric types. This ensures that statistical summaries and analyses, such as calculating means or generating boxplots, are accurate and meaningful.

```r
# change the type of some variables

```

```
tb.selected$price<-parse_number(tb.selected$price)
tb.selected$host_response_rate<-as.numeric(tb.selected$host_response_rate)
tb.selected$review_scores_rating<-as.numeric(tb.selected$review_scores_rating)

tb.selected$accommodates<-as.numeric(tb.selected$accommodates)
tb.selected$bathrooms<-as.numeric(tb.selected$bathrooms)
tb.selected$bedrooms<-as.numeric(tb.selected$bedrooms)

tb.selected$host_since_year<-as.numeric(tb.selected$host_since_year)
tb.selected$number_of_reviews<-as.numeric(tb.selected$number_of_reviews)
```

2.2.6 Handling missing values

Handling missing values is a critical step in data processing, as it ensures the integrity and reliability of the dataset. If not properly addressed, missing values can skew results and lead to incorrect conclusions. Since the sample size is large, we will remove observations with missing values from any column.

```
# remove the observations having missing values
tb.clean<-tb.selected %>% na.omit()
```

2.2.7 Handling invalid data

The provided boxplots visually identify outliers and invalid data points across four variables: price, accommodation, bathrooms, and bedrooms. Significant outliers in the price and accommodated variables indicate potential data entry errors or anomalies. For instance, extremely high prices or unusually high accommodated values may need further investigation and possible removal to maintain data integrity.

Since the sample size is large, we are going to remove the observations with invalid data in any column.

```
# Handle invalid data
tb.clean <- tb.clean %>%
  filter(price > 0,
         accommodates > 0, accommodates < 12,
         bathrooms >= 0, bedrooms > 0,
         host_since_year >= 2008, number_of_reviews >= 0)
```

2.2.8 Handle Outliers

This section focuses on handling outliers to ensure the dataset's accuracy and reliability. Outliers can significantly distort statistical analyses and lead to misleading conclusions. Specifically, we address outliers in the "price" variable using the interquartile range (IQR) method. This approach identifies and removes extreme values that fall outside 1.5 times the IQR from the first (Q1) and third quartiles (Q3).

```
# Handle outliers (e.g., remove outliers in price)
Q1 <- quantile(tb.clean$price, 0.25)
Q3 <- quantile(tb.clean$price, 0.75)
IQR <- Q3 - Q1

tb.clean <- tb.clean %>%
```

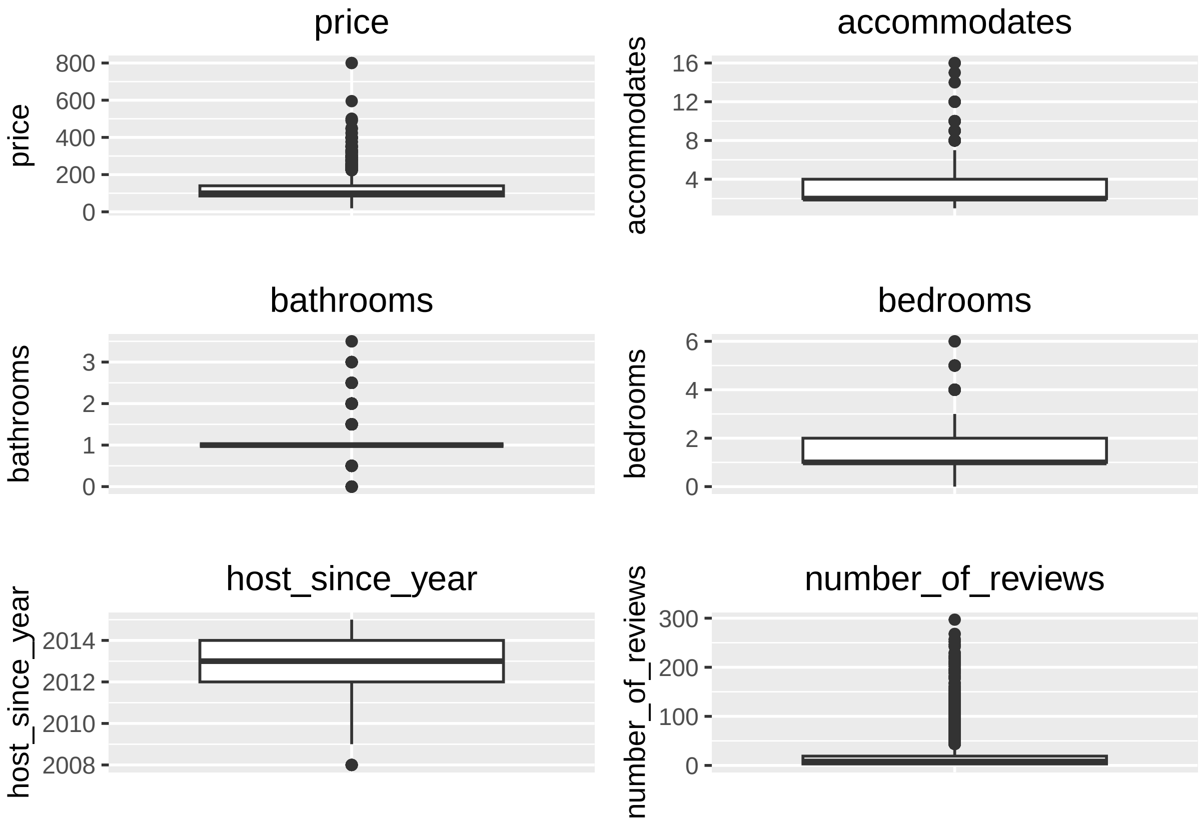


Figure 2: Data distribution before handling invalids and outliers

```
  filter(price >= (Q1 - 1.5 * IQR) & price <= (Q3 + 1.5 * IQR))
```

2.2.9 Generate New Features & Delete Useless Columns

This section focuses on generating new features and deleting unnecessary columns to enhance our dataset's usefulness and efficiency. Creating new features can provide additional insights and improve the performance of predictive models. For instance, By calculating region_average_price, we provide a valuable feature that summarizes regional price trends, aiding in more insightful analyses. Removing the region column helps streamline the dataset, focusing on the most relevant features for subsequent analyses.

```
# Example: Calculate average price by region

tb.clean <- tb.clean %>%
  group_by(region) %>%
  mutate(region_average_price = mean(price, na.rm = TRUE)) %>%
  ungroup()


tb.clean <- tb.clean %>% select(-region)
```

2.2.10 Handle Factor Columns

Factors in R represent categorical data, either nominal (no order) or ordinal (ordered). Converting variables to factors allows for efficient handling, accurate statistical modeling, and better results. For example, converting `room_type` to a factor ensures models handle different room types appropriately. Factors also aid in data summarization by categories and improve the clarity of data visualization.

```r
# Convert categorical variables to factor type
tb.clean <- tb.clean %>%
  mutate(across(c(room_type), as.factor))
```

2.3 Ensuring Tidy Data Compliance

Ensuring the data follows the three rules of "Tidy Data" is essential for effective data analysis. The three rules are: each variable forms a column, each observation forms a row, and each type of observational unit forms a table. The provided dataset adheres to these principles.

Observation: The table displays various variables: price, accommodates, bathrooms, bedrooms, room_type, host_response_rate, review_scores_rating, and region_average_price. Each variable forms a distinct column, ensuring that the data is well-organized and easy to interpret.

Structure:

- **Columns (Variables):** Each column represents a single variable. For example, "price" indicates the listing price, "accommodates" shows the number of people it can host, "bathrooms" and "bedrooms" count these facilities, "room_type" describes the type of room, "host_response_rate" provides the response rate of the host, "review_scores_rating" shows the review rating, and "region_average_price" gives the average price in the region.
- **Rows (Observations):** Each row represents an individual Airbnb listing, providing all relevant details in a single, unified format.
- **Table (Observational Unit):** The table represents a single observational unit, which in this case, is the Airbnb listing.

Table 3: Data after Cleaning (Part A)

price	accommodates	bathrooms	bedrooms	room_type	host_response_rate
130	4	2	2	Entire home/apt	0.80
59	2	1	1	Private room	1.00
95	4	1	1	Entire home/apt	1.00
100	2	1	1	Entire home/apt	1.00
115	2	1	1	Private room	0.89
80	3	1	1	Private room	1.00

Table 4: Data after Cleaning (Part B)

review_scores_rating	host_since_year	number_of_reviews	region_average_price
98	2008	11	104.88743
97	2008	108	86.58046
92	2008	15	103.64656
97	2008	20	122.15351
95	2008	4	131.66766
96	2009	36	112.07636

2.4 Summary

The summary provides insights into the dataset after data processing steps, and it generally aligns with our expectations:

1. **Price:** The price range is reasonable and reflects a variety of listings from budget to premium. The mean and median are close, indicating a relatively symmetrical distribution after removing outliers.
2. **Accommodates:** The range of accommodates values shows a variety of listings suitable for different group sizes, with a typical listing accommodating around 2-3 people.
3. **Bathrooms and Bedrooms:** The data shows typical values for bathrooms and bedrooms, with most listings having 1-2 of each. The presence of some listings with 0 bathrooms suggests possible studio apartments or shared spaces.
4. **Room Type:** The distribution indicates a higher number of entire homes/apartments, followed by private rooms, which is typical for Airbnb listings.
5. **Host Response Rate:** The high median and mean values indicate that most hosts are highly responsive.
6. **Review Scores Rating:** The high median and mean review scores suggest that the majority of the listings have good reviews.
7. **Region Average Price:** The region average price is consistent with our cleaned price data, showing a reasonable range and central tendency.

The statistics indicate that the data cleaning and preprocessing steps successfully prepared the data for further analysis, ensuring accuracy and reliability in subsequent analyses.

Table 5: The data statistics after Cleaning (Part A)

price	accommodates	bathrooms	bedrooms	room_type
Min. : 19.0	Min. : 1.00	Min. :0.000	Min. :1.000	Entire home/apt:3395
1st Qu.: 80.0	1st Qu.: 2.00	1st Qu.:1.000	1st Qu.:1.000	Private room : 649
Median :100.0	Median : 2.00	Median :1.000	Median :1.000	Shared room : 23
Mean :109.3	Mean : 2.85	Mean :1.066	Mean :1.322	NA
3rd Qu.:130.0	3rd Qu.: 4.00	3rd Qu.:1.000	3rd Qu.:2.000	NA
Max. :220.0	Max. :10.00	Max. :3.500	Max. :5.000	NA

Table 6: The data statistics after Cleaning (Part B)

host_response_rate	review_scores_rating	host_since_year
Min. :0.0200	Min. : 20.00	Min. :2008
1st Qu.:0.8700	1st Qu.: 90.00	1st Qu.:2012
Median :1.0000	Median : 95.00	Median :2013
Mean :0.9049	Mean : 93.52	Mean :2013
3rd Qu.:1.0000	3rd Qu.: 99.00	3rd Qu.:2014
Max. :1.0000	Max. :100.00	Max. :2015

3 Data Analysis

3.1 Research question 1

Do properties with prices at different levels relative to the regional average price have significantly different ratings?

3.1.1 Objective

To determine whether properties with prices at different levels relative to the regional average price have significantly different ratings. This analysis aims to explore if pricing strategy impacts customer satisfaction as reflected in review scores.

First, we calculated the price ratio for each property by dividing its price by the regional average price and determined the mean and standard deviation of these ratios. Properties were then categorized into four groups based on their price ratio: below average, average minus, average plus, and above average. Below average properties had price ratios less than one standard deviation below the mean, average minus properties were up to the mean, average plus were up to one standard deviation above, and above average exceeded one standard deviation above.

Next, we computed the mean and standard deviation of review scores for each group. An ANOVA test was conducted to assess statistically significant differences in review scores among the groups. Additionally, a t-test compared the review scores between the below average and above average groups to identify significant differences. Finally, a box plot visualized the distribution of review scores across the different price groups.

3.1.2 Result

Table 7: Guest Satisfaction for Different Price Group

price_group	average	SD
above_avg	94.647	6.654
avg_minus	93.474	6.701
avg_plus	94.348	7.012
below_avg	90.708	8.373

Table 8: T-test Summary “below_avg” Vs. “above_avg”

Statistic_Name	Value
below_avg_mean1	90.708
below_avg_se1	0.359
below_avg_mean2	94.647
below_avg_se2	0.259
t_value	-9.090
df	1201.000
p_value	0.000
CI_Lowe	-4.789
CI_Uppe	-3.089

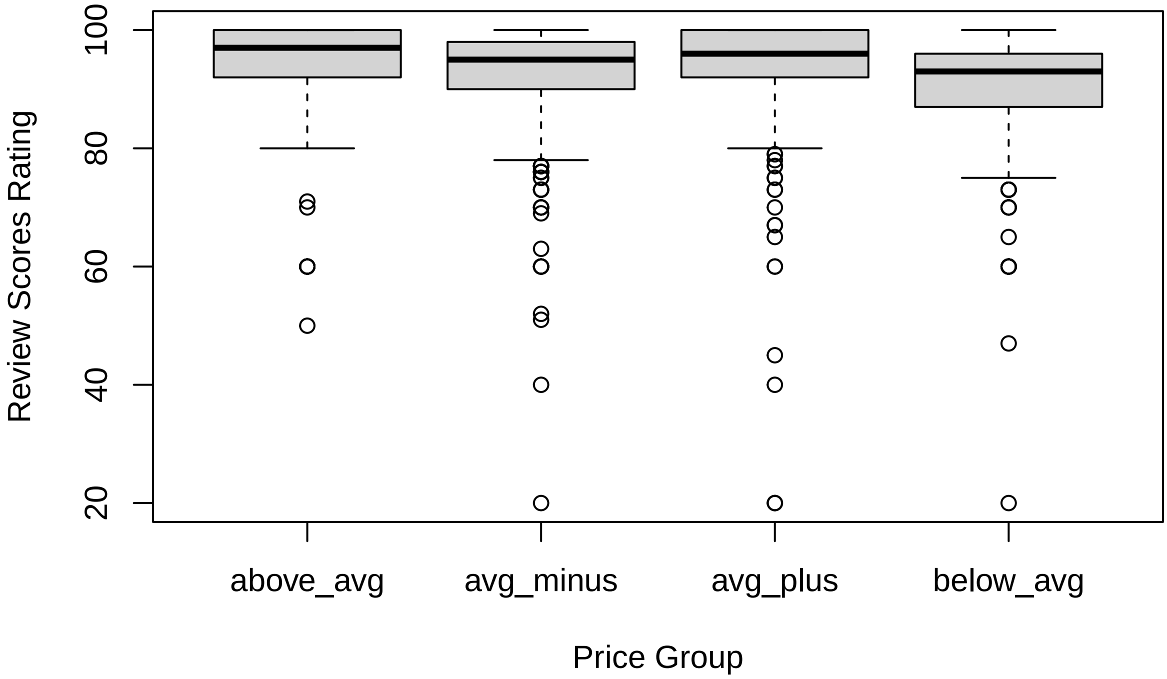


Figure 3: Review Scores Rating mean and standard deviation for different price group

3.1.3 Analysis

Table 7 shows that the “above average” group had a mean rating of 94.647 and a smaller standard deviation (6.654), indicating more consistent ratings. In contrast, the “below average” group had a lower mean rating (90.708) and a larger standard deviation (8.373), indicating greater variability in ratings. Figure 3 shows that the “above_avg” group shows a median score close to 95, with few outliers below 80. The “avg_minus” and “avg_plus” groups have similar distributions, with medians around 93-94, but with more outliers below 80. The “below_avg” group has the widest range of scores, with many outliers extending all the way to around 20, indicating more inconsistency and lower scores overall. The significant p-value (0.000) in Table 8 indicates a significant difference in review scores between the two groups. The confidence interval (-4.789 to -3.089) does not include zero, reinforcing the statistical significance. It also highlights that the “above_avg” group has a higher mean review score compared to the “below_avg” group.

3.1.4 Conclusion

properties with prices at different levels relative to the regional average price do have significantly different ratings.

Property prices significantly impact review scores. The “above_avg” price group has a notably higher average review score compared to the “below_avg” group. The T-test results confirm that this difference is statistically significant. This finding indicates that higher-priced properties are generally rated more favorably by guests. Guests may associate higher prices with better quality or amenities, leading to better reviews for more expensive properties. Therefore, pricing strategies should consider the positive impact on review scores and overall guest satisfaction.

3.2 Research question 2

Can property ratings be predicted using price, regional average price, number of bedrooms, number of accommodates,room types, and number of bathrooms?

3.2.1 Objective

To determine whether property ratings can be predicted based on factors such as price, regional average price, number of bedrooms, number of accommodates, and number of bathrooms. Additionally, we aim to identify if there are significant differences in ratings across different room types.

$$\begin{aligned}\hat{Y} = \beta_0 &+ \beta_1 \cdot \text{price} + \beta_2 \cdot \text{region\_average\_price} \\ &+ \beta_3 \cdot \text{bedrooms} + \beta_4 \cdot \text{accommodates} \\ &+ \beta_5 \cdot \text{bathrooms} + \beta_6 \cdot \text{room\_type (Shared Room)} + \epsilon \qquad (\textit{model } 1)\end{aligned}$$

We employed this linear regression model (model 1) to analyze the relationship between property ratings and various predictor variables, including price, regional average price, number of bedrooms, number of accommodates, number of bathrooms, and room type. The analysis involved constructing a linear regression model with review_scores_rating as the dependent variable and the aforementioned predictors as independent variables. We then analyzed the regression output to interpret the significance and impact of each predictor. To further evaluate the model, we generated a residual plot to visualize the distribution of residuals versus fitted values, which helps in assessing the model's fit, and a QQ plot to evaluate the normality of residuals, which is essential for validating the assumptions of linear regression.

3.2.2 Result

Table 9: Linear Regression Model for Guest Satisfaction Summary

Coefficient	Estimate	Std_Error	t_value	P_value_Sig
(Intercept)	94.896	1.003	94.621	<2e-16
Price	0.025	0.004	6.807	1.15e-11
Region Average Price	-0.022	0.008	-2.609	0.0091
Bedrooms	0.416	0.263	1.580	0.114
Accommodates	-0.978	0.125	-7.850	5.29e-15
Bathrooms	0.936	0.499	1.875	0.0608
Room Type (Private Room)	-2.914	0.337	-8.646	<2e-16
Room Type (Shared Room)	-4.831	1.463	-3.303	0.000965

```
## Residual standard error: 6.928

## Multiple R-squared: 0.05504

## Adjusted R-squared: 0.05341

## F-statistic: 33.8 (p-value < 5.1e-46)

## `geom_smooth()` using formula = 'y ~ x'

## [1] 2932 3758
```

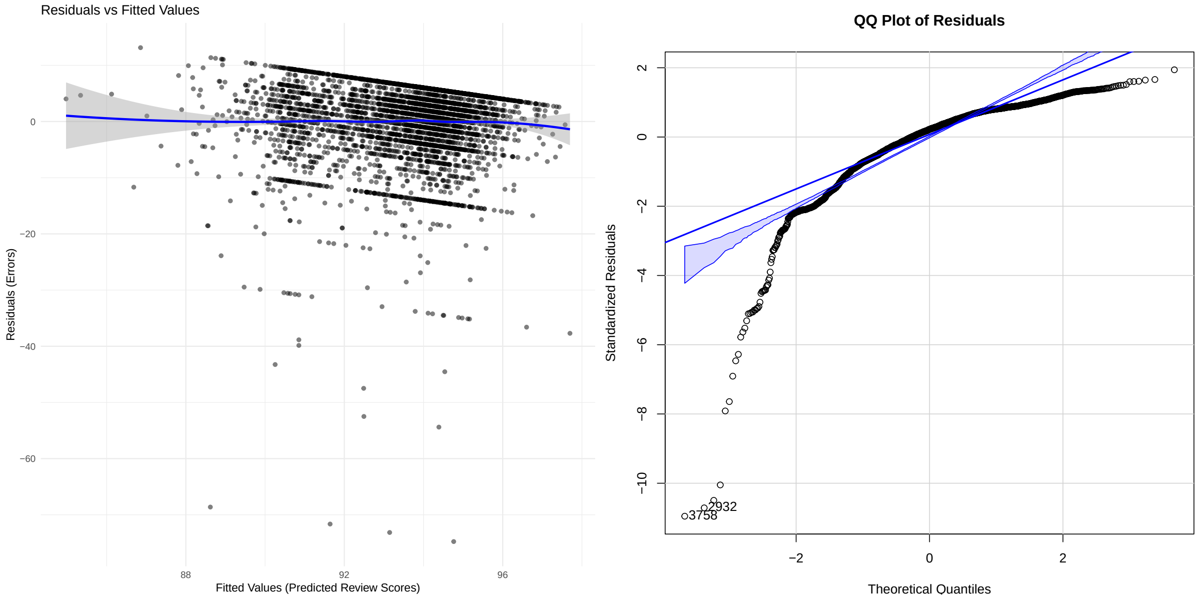


Figure 4: Linear Regression Model Evaluation

3.2.3 Analysis

Table 9 reveals significant relationships between review scores and several predictors: price (Estimate: 0.025, p = 1.15e-11), regional average price (Estimate: -0.022, p = 0.0091), accommodates (Estimate: -0.978, p = 5.29e-15), bathrooms (Estimate: 0.936, p = 0.0608), and room types (Private Room: -2.914, p = <2e-16; Shared Room: -4.831, p = 0.000965). The model fit, with a multiple R-squared of 0.055 and an adjusted R-squared of 0.053, indicates that the model explains a small portion of the variance in review scores, suggesting limited explanatory power.

The plot in Figure 4 Left shows heteroscedasticity, indicating that the variance of residuals increases with fitted values. This suggests that the linear regression model does not fully capture the relationship between predictors and review scores, leading to potential biases. Model adjustments, such as transforming variables or using a different modeling approach, may be necessary to improve fit.

The QQ plot in Figure 4 right indicates deviations from normality, particularly at the tails,suggesting the presence of outliers or heavy tails in the data. The points should ideally form a straight line if the residuals are normally distributed. However, in this case, the points deviate from the line, especially at the ends. The large deviations at the ends suggest that there are outliers and that the residuals are not symmetrically distributed, indicating skewness.This violation of the normality assumption implies that the model may produce inefficient estimates.

3.2.4 Conclusion

It is not possible to predict property ratings using price, average price for the area, number of bedrooms, capacity, and number of bathrooms, as the model can only explain about 5% of the variance in the ratings. However, we can note that higher prices and more bathrooms lead to higher ratings, while higher capacity and higher average price for the area lead to lower ratings. Room type has a significant impact on ratings, with private and shared rooms receiving lower ratings than the entire house/apartment. However, the residuals and QQ plots indicate problems with heteroskedasticity and non-normality, suggesting that a more complex model is needed.

3.3 Research question 3

Can high ratings (scores greater than or equal to 95) be predicted using price, host tenure, number of reviews, regional average price, and host response rate?

3.3.1 Objective

To determine whether high ratings (scores greater than or equal to 95) can be predicted based on factors such as price, host tenure, number of reviews, regional average price, and host response rate.

$$\log\left(\frac{\hat{p}}{1-\hat{p}}\right) = \beta_0 + \beta_1 \cdot \text{price} + \beta_2 \cdot \text{region\_average\_price} + \beta_3 \cdot \text{host\_since\_year} + \beta_4 \cdot \text{number\_of\_reviews} + \beta_5 \cdot \text{host\_response\_rate} \qquad (\textit{model } 2)$$

We employed this logistic regression model(model 2) to analyze the relationship between the likelihood of receiving a high rating and various predictor variables. Firstly,We added a binary variable `high_rating` indicating whether a property has a rating of 95 or higher (1) or not (0). We then built a logistic regression model using `high_rating` as the dependent variable and `price`, `region_average_price`, `host_since_year`, `number_of_reviews`, and `host_response_rate` as independent variables. We analyzed the logistic regression output to interpret the significance and impact of each predictor variable. Additionally, we generated a confusion matrix to evaluate the model's accuracy, sensitivity, specificity, and other performance metrics. We also plotted the ROC curve and calculated the Area Under the Curve (AUC) to assess the model's ability to distinguish between high and low ratings.

3.3.2 Result

Table 10: Logistic Regression Model Summary for Guest Satisfaction

Coefficient	Estimate	Std..Error	z.value	p.value
(Intercept)	123.723	59.610	2.076	0.0379
Price	0.008	0.001	9.014	<2e-16
Region Average Price	-0.003	0.002	-1.126	0.26
Host Since Year	-0.062	0.030	-2.090	0.0366
Number Of Reviews	-0.006	0.001	-4.313	1.61e-05
Host Response Rate	0.557	0.212	2.620	0.00878

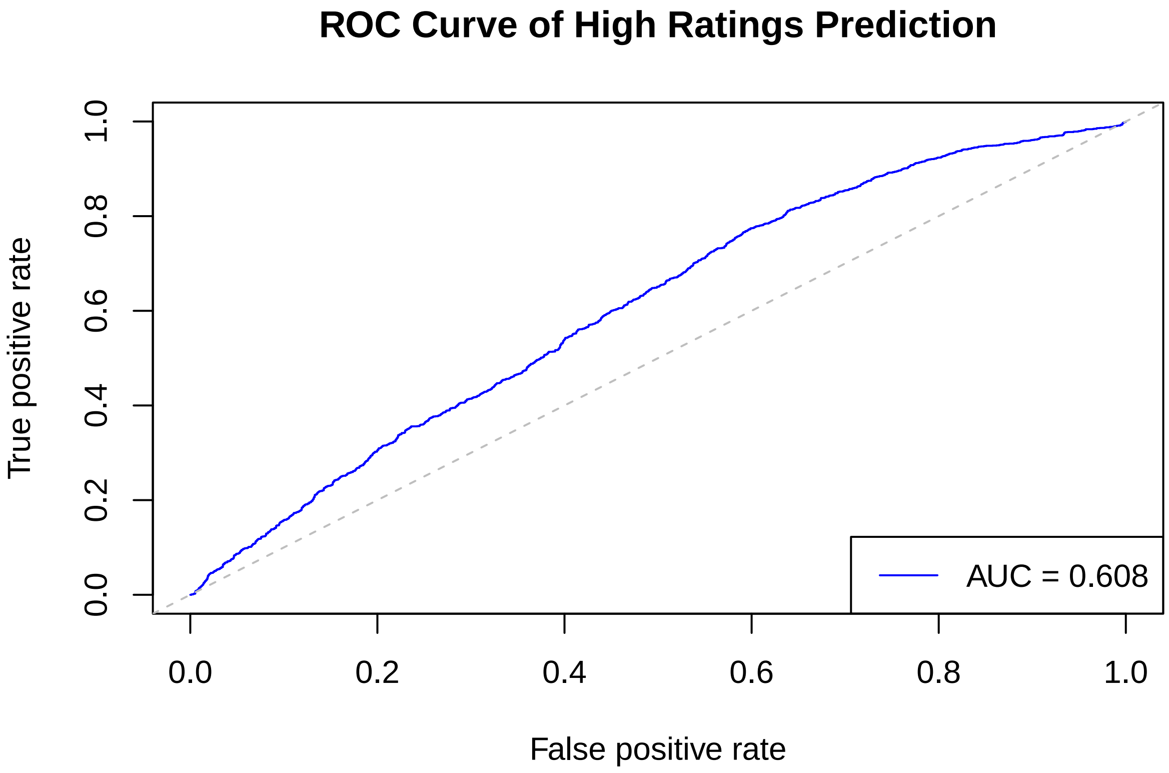
Table 11: Confusion Matrix for Logistic Regression Model

Reference	Prediction.0	Prediction.1
0	763	1101
1	520	1683

Table 12: Additional Statistics from Confusion Matrix

Statistic_Name	Value
Accuracy	0.6014261

Statistic_Name	Value
Kappa	0.1775510
AccuracyLower	0.5861875
AccuracyUpper	0.6165189
AccuracyNull	0.5416769
AccuracyPValue	0.0000000
McnemarPValue	0.0000000



3.3.3 Analysis

Price has a positive significant relationship with high ratings (Estimate: 0.008, p = <2e-16). Host response rate has a strong positive significant relationship with high ratings (Estimate: 0.557, p = 0.00878). Number of reviews has a negative significant relationship with high ratings (Estimate: -0.006, p = 1.61e-05). Host tenure (host_since_year) and regional average price were not significant predictors. The confusion matrix provides performance metrics, indicating an accuracy of 0.6014. An ROC curve closer to the upper left corner signifies a better model performance. In this case, the curve leans more towards the diagonal, indicating the model is only slightly better than random chance at distinguishing between high and low ratings.The ROC curve shows an AUC of 0.608, suggesting the model has moderate ability to distinguish between high and low ratings.

3.3.4 Conclusion

The results show that higher prices and higher host response rates increase the likelihood of having a high rating, while higher numbers of reviews decrease the likelihood. Host tenure and average price by region are not significant predictors. However, the moderate AUC value and accuracy of 0.6014 indicate that the model has limited predictive power and is not strong enough to reliably predict high ratings. Therefore, we cannot use price, host tenure , number of reviews, average price by region, and host response rate to predict high ratings.

4 Conclusion

Our analysis provides several key insights into the factors influencing home ratings and guest satisfaction on Airbnb. These findings have practical implications for hosts who want to improve their listings and boost their ratings. In addition, I gained a deeper understanding of statistical analysis using R, which is valuable for future research and practical applications.

4.1 Significance

This study highlights the significant impact of pricing strategy, room features, and host responsiveness on guest satisfaction. Hosts can use these insights to make strategic adjustments:

1. **Adjusting Prices:** Setting competitive and appropriate prices that reflect the quality of the listing can help meet or exceed guest expectations, thereby improving satisfaction.
2. **Enhancing Listing Features:** Adding more bathrooms and optimizing the number of nights can more effectively meet guests' needs, leading to better reviews.
3. **Improving Host Responsiveness:** Maintaining a high response rate can enhance the guest experience and promote positive reviews and higher ratings.
4. **Importance of Public Opinion:** The greater the number of reviews, the lower the likelihood of a positive review. Hosts need to pay more attention to public opinion.

By focusing on these areas, hosts can attract more bookings and receive better overall ratings, thereby improving their market presence and attractiveness.

4.2 Limitations

While our analysis provides valuable insights, there are still some limitations to consider. Objective limitations, such as excluding variables related to location (e.g., latitude, longitude, neighborhood), are necessary because analyzing them requires more detailed information about transportation, local facilities, population density, and urban layout. Complexity is another factor; this experiment focuses primarily on learning statistical methods using R and is not intended to be a comprehensive analysis, but rather an attempt to explore the impact of some common variables on property ratings. In addition, several columns in the dataset are not used in our current analysis. These unused variables may contain valuable information that can enhance the predictive power of our model. Acknowledging these limitations is critical for interpreting results and considering areas for improvement.

4.3 Challenges Encountered

In this study, we faced several challenges. Ensuring the accuracy and completeness of the dataset was a challenge, with missing values and inconsistent data entry requiring extensive preprocessing efforts. Selecting the most relevant variables from a large dataset is complex and requires careful consideration of each variable's potential impact on the analysis. Ensuring that the data meets the assumptions required for linear and logistic regression models is a challenge, especially in terms of normality and homoscedasticity of residuals. In Research Questions 2 and 3, we used different variables designed to predict high ratings, but still did not obtain accurate predictions. Specifically, the logistic regression models in Research Questions 2 and 3 showed that while higher price and host response rate were significant predictors of high ratings, the overall accuracy and predictive power of these models were limited.

4.4 Future Work Opportunities

Future research can build on our findings to better understand the factors that drive guest satisfaction. Including more variables such as "latitude" and "longitude" and then incorporating specific city maps, transportation, and business amenities can improve the accuracy and explanatory power of the model. Employing more sophisticated methods, such as machine learning, can better capture the complex relationships and interactions between variables. Studying changes over time can reveal how continuous improvements and price adjustments to properties affect guest satisfaction and ratings. Analyzing qualitative feedback from guest reviews can reveal additional factors that contribute to high ratings, providing more nuanced insights. Exploring the potential of unused columns can provide additional predictive power and provide more comprehensive insights into the dynamics of guest satisfaction. By exploring these opportunities, future research can enhance our understanding of guest preferences, help hosts better meet their needs, and ultimately improve ratings and succeed in the competitive Airbnb market.

5 References

Fox, John, and Sanford Weisberg. 2019. *An R Companion to Applied Regression*. Third. Thousand Oaks CA: Sage. https://socialsciences.mcmaster.ca/jfox/Books/Companion/.

Fox, John, Sanford Weisberg, and Brad Price. 2023. *Car: Companion to Applied Regression*. https://r-forge.r-project.org/projects/car/.

GaAirbnb. 2024. "Georgia Airbnb Listings." https://data.world/cannata/gaairbnb.

Ju, Yongwook, Ki-Joon Back, Youngjoon Choi, and Jin-Soo Lee. 2019. "Exploring Airbnb Service Quality Attributes and Their Asymmetric Effects on Customer Satisfaction." *International Journal of Hospitality Management* 77: 342–52.

Kuhn, Max. 2023. *Caret: Classification and Regression Training*. https://github.com/topepo/caret/.

Kuhn, and Max. 2008. "Building Predictive Models in r Using the Caret Package." *Journal of Statistical Software* 28 (5): 1–26. https://doi.org/10.18637/jss.v028.i05.

Li, Jing, Simon Hudson, and Kevin Kam Fung So. 2019. "Exploring the Customer Experience with Airbnb." *International Journal of Culture, Tourism and Hospitality Research* 13 (4): 410–29.

Pedersen, Thomas Lin. 2024. *Patchwork: The Composer of Plots*. https://patchwork.data-imaginist.com.

Priporas, Constantinos-Vasilios, Nikolaos Stylos, Lakshmi Narasimhan Vedanthachari, and Pruit Santiwatana. 2017. "Service Quality, Satisfaction, and Customer Loyalty in Airbnb Accommodation in Thailand." *International Journal of Tourism Research* 19 (6): 693–704.

Robinson, David, Alex Hayes, and Simon Couch. 2023. *Broom: Convert Statistical Objects into Tidy Tibbles*. https://broom.tidymodels.org/.

Sing, Tobias, Oliver Sander, Niko Beerenwinkel, and Thomas Lengauer. 2020. *ROCR: Visualizing the Performance of Scoring Classifiers*. http://ipa-tys.github.io/ROCR/.

Sing, T., O. Sander, N. Beerenwinkel, and T. Lengauer. 2005. "ROCR: Visualizing Classifier Performance in r." *Bioinformatics* 21 (20): 7881. http://rocr.bioinf.mpi-sb.mpg.de.

Situmorang, Kevin M, Achmad N Hidayanto, Alfan F Wicaksono, and Arlisa Yuliawati. 2018. "Analysis on Customer Satisfaction Dimensions in Peer-to-Peer Accommodation Using Latent Dirichlet Allocation: A Case Study of Airbnb." In *2018 5th International Conference on Electrical Engineering, Computer Science and Informatics (EECSI)*, 542–47. IEEE.

Wickham, Hadley. 2016. *Ggplot2: Elegant Graphics for Data Analysis*. Springer-Verlag New York. https://ggplot2.tidyverse.org.

———. 2023. *Stringr: Simple, Consistent Wrappers for Common String Operations*. https://stringr.tidyverse.org.

Wickham, Hadley, Winston Chang, Lionel Henry, Thomas Lin Pedersen, Kohske Takahashi, Claus Wilke, Kara Woo, Hiroaki Yutani, Dewey Dunnington, and Teun van den Brand. 2024. *Ggplot2: Create Elegant Data Visualisations Using the Grammar of Graphics*. https://ggplot2.tidyverse.org.

Wickham, Hadley, Romain François, Lionel Henry, Kirill Müller, and Davis Vaughan. 2023. *Dplyr: A Grammar of Data Manipulation*. https://dplyr.tidyverse.org.

Wickham, Hadley, Jim Hester, and Jennifer Bryan. 2024. *Readr: Read Rectangular Text Data*. https://readr.tidyverse.org.

Xie, Yihui. 2014. "Knitr: A Comprehensive Tool for Reproducible Research in R." In *Implementing Reproducible Computational Research*, edited by Victoria Stodden, Friedrich Leisch, and Roger D. Peng. Chapman; Hall/CRC.

———. 2015. *Dynamic Documents with R and Knitr*. 2nd ed. Boca Raton, Florida: Chapman; Hall/CRC. https://yihui.org/knitr/.

———. 2023. *Knitr: A General-Purpose Package for Dynamic Report Generation in r*. https://yihui.org/knitr/.

Zhu, Hao. 2024. *kableExtra: Construct Complex Table with Kable and Pipe Syntax*. http://haozhu233.github.io/kableExtra/.

6 Appendix: Source Code

For detailed insights into the methods and analyses conducted in this study, the complete source code is available at the following GitHub repository:

https://github.com/guangliangyang/r-study

7 Appendix: Glossary

1. **ANOVA (Analysis of Variance)**: A statistical method used to compare the means of three or more groups to determine if there are significant differences among them.
2. **T-Test**: A statistical test used to compare the means of two groups. It can be an independent samples t-test or a paired samples t-test.
3. **Linear Regression**: A statistical method used to model the relationship between two or more variables by fitting a linear equation to the observed data.
4. **Logistic Regression**: A statistical method used to model binary outcome variables. It uses the logistic function to predict the probability of an event occurring.
5. **Residuals**: The difference between the observed values and the predicted values in a regression analysis. Residual analysis is used to assess the fit of the model.
6. **QQ Plot (Quantile-Quantile Plot)**: A graphical tool used to assess if a dataset follows a particular distribution, typically the normal distribution, by plotting the quantiles of the data against the quantiles of the theoretical distribution.
7. **ROC Curve (Receiver Operating Characteristic Curve)**: A graphical representation of a classifier's performance by plotting the true positive rate (sensitivity) against the false positive rate (1-specificity).
8. **AUC (Area Under the Curve)**: The area under the ROC curve, which measures the ability of a classifier to distinguish between classes. Values range from 0.5 to 1, with higher values indicating better performance.
9. **Confusion Matrix**: A table used to evaluate the performance of a classification model by showing the actual versus predicted classifications, including true positives, false positives, true negatives, and false negatives.
10. **P-Value**: The probability of obtaining a test statistic at least as extreme as the one observed, assuming the null hypothesis is true. A smaller p-value (typically < 0.05) indicates statistical significance.
11. **Coefficient**: In regression analysis, the value that represents the relationship between a predictor variable and the outcome variable. A positive coefficient indicates a positive relationship, while a negative coefficient indicates a negative relationship.
12. **Intercept**: In regression analysis, the value at which the regression line crosses the y-axis, representing the predicted value when all predictors are zero.
13. **Standard Deviation**: A measure of the dispersion or spread of a set of values. It indicates the average distance of each value from the mean.
14. **Mean**: The average of a set of values, calculated by summing all values and dividing by the number of values.
15. **Median**: The middle value of a dataset when the values are arranged in ascending order. It represents the 50th percentile.
16. **IQR (Interquartile Range)**: The range between the first quartile (Q1) and the third quartile (Q3) of a dataset. It measures the spread of the middle 50% of the data.
17. **Homoscedasticity**: An assumption in regression analysis that the variance of residuals is constant across all levels of the predictor variables. If the variance changes, it is called heteroscedasticity.

18. **Multicollinearity**: A situation in regression analysis where two or more predictor variables are highly correlated, which can affect the stability and interpretability of the regression coefficients.

19. **Factor**: A data type in R used to represent categorical variables, which can be either nominal (no order) or ordinal (ordered).

20. **One-Hot Encoding**: A method to convert categorical variables into numerical format by creating binary columns for each category.

This glossary is intended to provide clear definitions and help readers understand the statistical terms used in the analysis.

8 Appendix: R Environment

```r
format(Sys.time(), '%d %B %Y')
```

```
## [1] "06 June 2024"
```

```r
sessionInfo()
```

```
## R version 4.3.3 (2024-02-29)
## Platform: x86_64-apple-darwin20 (64-bit)
## Running under: macOS Sonoma 14.4.1
## 
## Matrix products: default
## BLAS:   /Library/Frameworks/R.framework/Versions/4.3-x86_64/Resources/lib/libRblas.0.dyli
## LAPACK: /Library/Frameworks/R.framework/Versions/4.3-x86_64/Resources/lib/libRlapack.dyli
## 
## locale:
## [1] en_US.UTF-8/en_US.UTF-8/en_US.UTF-8/C/en_US.UTF-8/en_US.UTF-8
## 
## time zone: Pacific/Auckland
## tzcode source: internal
## 
## attached base packages:
## [1] stats     graphics  grDevices utils     datasets  methods   base     
## 
## other attached packages:
##  [1] readr_2.1.5      broom_1.0.5      knitr_1.45       ROCR_1.0-11     
##  [5] caret_6.0-94     lattice_0.22-5   car_3.1-2        carData_3.0-5   
##  [9] patchwork_1.2.0  stringr_1.5.1    kableExtra_1.4.0 gridExtra_2.3   
## [13] ggplot2_3.5.0    dplyr_1.1.4     
## 
## loaded via a namespace (and not attached):
##  [1] tidyselect_1.2.1     viridisLite_0.4.2    timeDate_4032.109   
##  [4] farver_2.1.1         fastmap_1.1.1        pROC_1.18.5         
##  [7] digest_0.6.34        rpart_4.1.23         timechange_0.3.0    
## [10] lifecycle_1.0.4      survival_3.5-8       magrittr_2.0.3      
## [13] compiler_4.3.3       rlang_1.1.3          tools_4.3.3         
## [16] utf8_1.2.4           yaml_2.3.8           data.table_1.15.2   
## [19] labeling_0.4.3       plyr_1.8.9           xml2_1.3.6          
## [22] abind_1.4-5          withr_3.0.0          purrr_1.0.2         
## [25] nnet_7.3-19          grid_4.3.3           stats4_4.3.3        
## [28] fansi_1.0.6          e1071_1.7-14         colorspace_2.1-0    
## [31] future_1.33.2        globals_0.16.3       scales_1.3.0        
## [34] iterators_1.0.14     MASS_7.3-60.0.1      cli_3.6.2           
## [37] rmarkdown_2.26       generics_0.1.3       rstudioapi_0.15.0   
## [40] future.apply_1.11.2  tzdb_0.4.0           reshape2_1.4.4      
## [43] proxy_0.4-27         splines_4.3.3        parallel_4.3.3      
## [46] vctrs_0.6.5          hardhat_1.3.1        Matrix_1.6-5        
## [49] hms_1.1.3            listenv_0.9.1        systemfonts_1.0.6   
## [52] foreach_1.5.2        gower_1.0.1          tidyr_1.3.1         
## [55] recipes_1.0.10       glue_1.7.0           parallelly_1.37.1   
## [58] codetools_0.2-19     lubridate_1.9.3      stringi_1.8.3
## [61] gtable_0.3.4         munsell_0.5.0        tibble_3.2.1
## [64] pillar_1.9.0         htmltools_0.5.7      ipred_0.9-14
## [67] lava_1.8.0           R6_2.5.1             evaluate_0.23
## [70] highr_0.10           backports_1.4.1      class_7.3-22
## [73] Rcpp_1.0.12          svglite_2.1.3        nlme_3.1-164
## [76] prodlim_2023.08.28   mgcv_1.9-1           xfun_0.42
## [79] pkgconfig_2.0.3      ModelMetrics_1.2.2.2
```