KU LEUVEN kulak

Baert Emile
Debaveye Viktor
Wastijn Jérôme

Kost van een Airbnb-verblijf

Academiejaar 2022 – 2023 Project statistiek

Inleiding

Airbnb is een online platform (`www.airbnb.com`) waarop reizigers een kort verblijf kunnen boeken in een accommodatie (bv. een kamer, een huis, een woonboot, ...) die door particulieren wordt verhuurd. Airbnb werd opgericht in 2008 en is ondertussen een erg populair alternatief geworden voor de traditionele hotels.

Dit onderzoek focust op de totale kostprijs voor het huren van een Airbnb-verblijf in Amsterdam voor twee personen gedurende een weekend (vrijdag tot zondag). Er wordt nagegaan met welke factoren deze kostprijs samenhangt en in welke mate de kost daarmee kan worden voorspeld. Hierbij wordt er gebruik gemaakt van een aantal veranderlijken verzameld in het kader van een onderzoeksproject uitgevoerd door Kristóf Gyódi en Łukasz Nawaro. Het gaat enerzijds over kenmerken van het verblijf en tevredenheid van eerdere gasten volgens de gegevens van Airbnb, anderzijds over de ligging van het verblijf ten opzichte van het stadscentrum, bezienswaardigheden en restaurants, telkens rekening houdend met de populariteit, zoals gerapporteerd door TripAdvisor.
Tabel 1 beschrijft de variabelen en een uittreksel van de dataset is te zien in Tabel 2. Tabel 3, ten slotte, geeft enkele basisstatistieken weer van de veranderlijken die gebruikt worden in het onderzoek.

1 Methode

1.1 Kenmerken van de steekproef

De afhankelijke veranderlijke die in ons project wordt onderzocht is de realSum. Dat is de totale kost om een woning te huren via Airbnb voor twee personen gedurende het weekend. Aangezien deze dataset dateert uit 2019, moeten we eerst testen of de gemiddelde kost van 2023 (€620) significant verschilt ten opzichte van de gemiddelde kost uit onze steekproef. Dit doen we aan de hand van een tweezijdige t-test.

In steden waar het verhuren van vakantieverblijven niet is gereguleerd, hebben professionele aanbieders typisch de overhand. Naarmate de regulering strenger is, is het aantal particuliere aanbieders groter. We vragen ons af of dit ook zo is in Amsterdam. Hiervoor doen we een exacte binomiaal test.

We gaan ook nog na of het aantal slaapkamers in een verblijf de Poissonverdeling volgt. Dit doen we door eerst de variantie te berekenen en daarna na te gaan of de waarde van lambda significant verschilt van de variantie van bedrooms.

Tabel 1: Veranderlijken in de dataset.

	Naam	Beschrijving
1	realSum	Som van alle kosten (€)
2	room	Soort verblijf (1 = volledige woning, 2 = afzonderlijke kamer, 3 = gedeelde kamer)
3	capacity	Maximaal aantal gasten
4	bedrooms	Aantal beschikbare slaapkamers in het verblijf
5	dist	Afstand tot het stadscentrum (km)
6	metro	Afstand tot dichtstbijzijnde metro-halte (km)
7	attr	Attractiescore, nabijheid van bezienswaardigheden (score tussen 1 en 10)
8	rest	Restaurantscore, nabijheid van restaurants (score tussen 1 en 10)
9	host	Type verhuurder (0 = enige beschikbare woning, 1 = 2 tot 4 beschikbare woningen, 2 = meer dan 4 beschikbare woningen)
10	cleanliness	Modale score voor netheid van het verblijf volgens gasten (op 10)
11	satisfaction	Tevredenheid van de gasten (op 10)

Tabel 2: Uittreksel van de dataset.

	realSum	room	capacity	bedrooms	dist	metro
1	319.6401	2	2	1	4.7633597	0.85211740
2	347.9952	2	2	1	5.7483103	3.65159150
3	482.9752	2	4	2	0.3848721	0.43985163
4	485.5529	2	2	1	0.5447226	0.31868815
5	2771.5417	1	4	3	1.6867977	1.45839939
6	1001.8044	1	4	2	3.7191389	1.19610391

	attr	rest	host	cleanliness	satisfaction
1	1.341015	1.707322	2	9	8.8
2	1.167478	1.365839	2	9	8.7
3	3.203343	7.767078	2	9	9.0
4	3.493515	7.276056	0	10	9.8
5	1.817855	2.818349	0	10	10.0
6	1.318223	1.681824	0	9	9.6

Tabel 3: Basisstatistieken

	Gemiddelde ± Standaardfout	Bereik	Algemene vorm
realSum	(604 ± 14)	[165.9129 , 8130.668]	Rechtsscheef
dist	(2,81 ± 0.07)	[0.01504452 , 11.19593]	Rechtsscheef + veel outliers
metro	(1.09 ± 0.03)	[0.03651741 , 4.411905]	Rechtsscheef + veel outliers
attr	(2,10 ± 0.03)	[1 , 10]	Eerder exponentieel verdeeld
rest	(3,28 ± 0.06)	[1 , 10]	Rechtsscheef
cleanliness	(9,47 ± 0.03)	[2 , 10]	Linksscheef
satisfaction	(9,47 ± 0.02)	[2 , 10]	Linksscheef

1.2 Gemiddelde kost

We vragen ons af of de totale kost voor een weekend een woning te huren, verschilt naargelang de netheid, het type verhuurder en het soort verblijf. We doen voor de drie gevallen eerst een Shapiro-Wilk test en daarna een t-test voor de ongepaarde gegevens met verschillende varianties.

1.3 Associatie met de verschillende veranderlijken

Aangezien het handig kan zijn om te weten of de kost van een verblijf afhankelijk is van de andere variabelen onderzoeken we dit ook. We kijken eerst of de continue variabelen normaal verdeeld zijn. Indien dit het geval is, doen we een Pearson correlatie test. Als dit niet zo is, doen we een Spearman correlatie test. Bij de discrete variabelen stellen we een kruistabel op en kijken we of deze voldoet aan de Cochran regel. Indien van toepassing voeren we een χ^2-test uit, anders een $Fischer$-test. Afhankelijk van de uitkomst van deze testen besluiten we dan of de waarden significant afhankelijk zijn van elkaar of niet.

1.4 Verklaren van de kost

Om de samenhang tussen de totale kost en andere continue variabelen te bepalen, voeren we eerst eens regressie-analyse uit op de totale kost en de attractiescore aan de hand van een eenvoudig regressiemodel. Hierna kijken we of we door een logaritmische transformatie een beter model bekomen of niet. We maken eerst een lineair regressiemodel van de totale kost in functie van de attractiescore en kijken daarna naar de samenvatting van dit model. Bij de logaritmische transformatie doen we exact hetzelfde.

Om de kost te verklaren passen we meervoudige lineaire regressie toe op alle continue variabelen. Door het gebruik van achterwaartse lineaire regressie bekomen we een eenvoudiger model en kunnen we de resultaten beter interpreteren. Dit doen we opnieuw door eerst een lineair regressie model aan te maken, maar dit keer staat de totale kost in functie van alle continue veranderlijken. Hierna bekijken we

de samenvatting van dit model en zoeken we de minst significante veranderlijke, die we dan verwijderen. Dit proces herhalen we tot we enkel maar significante veranderlijken overhouden.

Om na te gaan of dit model wel een goede oplossing is bekijken we de bekomen grafieken. We passen ook logaritmische transformaties toe om tot een zo goed mogelijk model te komen. Dit doen we door per variabele na te gaan of een logaritmische transformatie het model verbetert of niet. Indien dit zo is behouden we de transformatie en controleren we de volgende variabele.

We vragen ons af of het zin heeft om afzonderlijke vergelijkingen te hanteren naargelang het verblijf een volledige woning is of niet. Aangezien de veranderlijke room een kwantitatieve veranderlijk is maken we eerst een dummy. Hierbij is de waarde 1 als de waarde van room 1 is en 0 als de waarde verschillend is van 1. Hierna vermenigvuldigen we het bekomen model uit de vorige paragraaf met deze dummy en herhalen we opnieuw de stappen voor achterwaartse regressie.

2 Resultaten

2.1 Kenmerken van de steekproef

Volgens Airbnb kost een weekendje voor twee personen gemiddeld €620 in 2023, maar onze steekproef dateert uit 2019 en heeft een gemiddelde van €604.828. Op basis van een tweezijdige t-test kunnen we concluderen dat het gemiddelde van de steekproef niet significant verschilt van 620 euro ($p = 0.28$, $t = -1.069$).

Er zijn in onze steekproef 636 particuliere **hosts** en 341 professionele. Om te achterhalen of de proportie particuliere hosts signifcant groter is voeren we een rechtseenzijdige binomiale test uit. Hieruit concluderen we dat het aandeel particuliere aanbieders inderdaad significant groter is ($p = 0.65$) dan het aantal professionele aanbieders.

We gaan ook na of de veranderlijke **bedrooms** de Poissonverdeling volgt. We gebruiken hiervoor de dichtheidsfunctie van de Poissonverdeling: $f = \frac{e^{\lambda} \cdot \lambda^k}{k!}$. Voor $k = 0$ zijn er 63 waarnemingen in onze steekproef. Aan de hand van deze waarden kunnen we λ berekenen, en vinden we dat $\lambda = 2.74$. Dit zou gelijk moeten zijn aan de variantie van de veranderlijke bedrooms, namelijk 0.54, wat absoluut niet het geval is. Hieruit kunnen we besluiten dat de veranderlijke bedrooms niet Poisson verdeeld is.

2.2 Gemiddelde kost

Ten eerste onderzoeken we of de totale kost verschilt naargelang het verblijf de maximumscore heeft voor netheid (**cleanliness**) of niet. Door de Shapiro-Wilk test toe te passen op beide groepen weten we dat deze veranderlijken niet normaal verdeeld zijn ($p << 0.001$). We kunnen dus geen variantietest doen. Wel zijn er voldoende observaties (378 en 599) in elke groep om de Centrale Limiet Stelling toe te passen. Nu doen we een t-test voor ongepaarde groepen met verschillende varianties, waaruit we kunnen besluiten dat de totale prijs niet significant verschilt naargelang de score voor netheid 10 is.

Ten tweede bekijken we of de kost verschilt naargelang de eigenaar slechts één verblijf aanbiedt of niet (**host**). Na het toepassen van de Shapiro test verwerpen we normaliteit opnieuw. We doen een beroep op de CLS en voeren een t-test voor ongepaarde gegevens met verschillende varianties uit. We verkrijgen hier dat $p = 0.0256$ en kunnen dus niets besluiten.

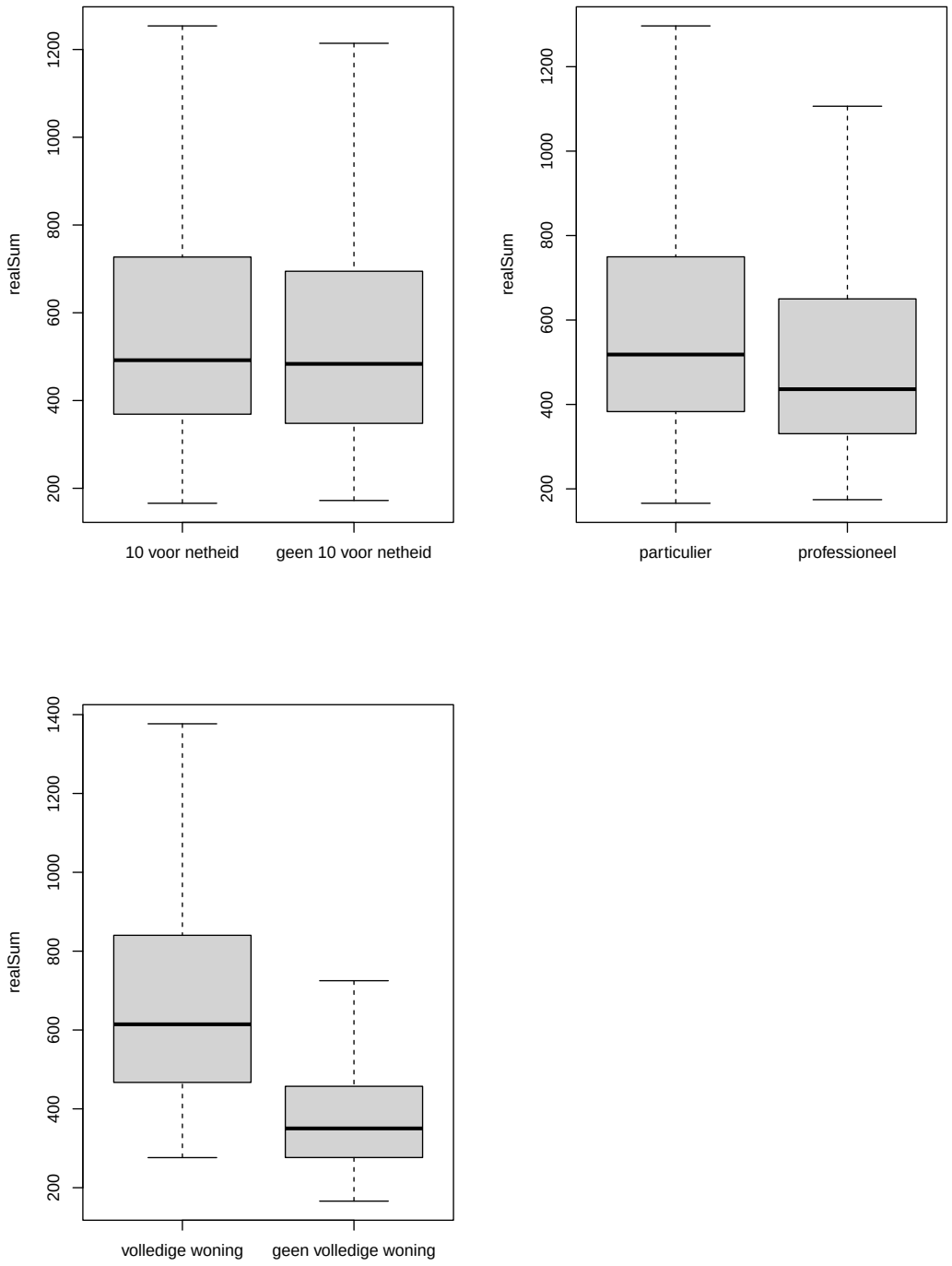
Ten derde herhalen we dit proces naargelang de volledige woning wordt verhuurd of niet (**room**), waaruit we besluiten dat de prijs significant verschilt naargelang de volledige woning wordt verhuurd of niet ($p << 0.001$).

Deze vergelijkingen zijn op een overzichtelijke manier te zien in figuur 1. Het is duidelijk dat de gevonden resultaten inderdaad gelden.

2.3 Associatie met de verschillende veranderlijken

We gaan na of er afhankelijkheid is tussen de kost van een verblijf en elke andere veranderlijke.

Bij de numerieke veranderlijken (**capacity**, **bedrooms**, **dist**, **metro**, **attr**, **rest**, **cleanliness** en **satisfaction**) passen we telkens de Shapiro test toe, waaruit blijkt dat geen enkele variabele normaal verdeeld is. We voeren hierna dus een Spearman correlatie test uit.



Figuur 1: Vergelijkende boxplots voor de realSum in functie van de verschillende scores voor netheid, het type verhuurder en het type woning.(Outliers werden hier verwijderd om de plots overzichtelijk te houden)

Bij de veranderlijke **host** passen we onze kruistabel aan totdat deze voldoet aan de Cochranregel, waarna we de χ^2-kwadraattest uitvoeren.

De kruistabel bij de variabele **room** voldoet niet aan de Cochranregel waardoor we een $Fisher$-test moeten uitvoeren.

In tabel 4 staan alle p-waardes die we bekwamen bij bovenstaande testen.

2.4 Verklaren van de kost

Uit het eenvoudige regressiemodel van **realSum** en **attr** vinden we dat de attractiescore een significant effect heeft op de totale kost. ($t = 8.70$, $p << 0.01$). Als we van beide veranderlijken de tiendelige logaritmische transformatie berekenen, bekomen we een beter model ($t = 14.43$, $p << 0.01$). We bekomen namelijk betere plots en daarnaast is de R^2 gestegen van 0.071 naar 0.175.

Hierna stellen we een nieuw lineair regressiemodel op voor de totale kost (**realSum**) in functie van alle continue veranderlijken. De spreidingsdiagrammen van de continue variabelen in functie van de realSum zijn te zien in figuur 2. We voeren op deze veranderlijken achterwaartse regressie uit. Achtereenvolgens worden de variabelen **rest** ($t = -0.59$, $p = 0.55$) en **metro** ($t = -0.62$, $p = 0.53$) verwijderd wegens het niet significant zijn volgens de uitgevoerde t-testen. In het resterende model zijn de regressiecoëfficiënten dus de veranderlijken **dist** ($t = -3.80$, $p = 0.00016$) en ($t = 3.90$, $p = 0.00010$). Dit model levert een R^2 van 0.083 op. We vinden volgende vergelijking voor het logaritme van de realSum:

$$log_{10}(realSum) = 2.5514 + 0.5691 \cdot attr \tag{1}$$

Hierna nemen we van elke overgebleven veranderlijke (**realSum**, **dist** en **attr**) de tiendelige logaritmische transformatie. We passen telkens achterwaartse regressie toe op deze lineaire regressiemodellen. Op die manier bekomen we uiteindelijk het best passende regressiemodel, namelijk:

$$log_{10}(realSum) = 2.6559 - 0.0183 \cdot dist + 0.1672 \cdot log_{10}(attr) \tag{2}$$

Bij dit model hebben we een R^2 van 0.18. Daarnaast vinden we voor **dist** dat $t = -3.25$ en $p = 0.0012$ en voor **log(attr)** dat $t = 5.58$ en $p << 0.001$. Hierbij hebben we ook de plots van het regressiemodel (zie figuur 3) uitvoerig bekeken en vergeleken. Zo zien we dat hier enkele uitschieters zijn en dat deze outliers een (licht) negatief effect hebben op ons uiteindelijke regressiemodel.

3 Discussie

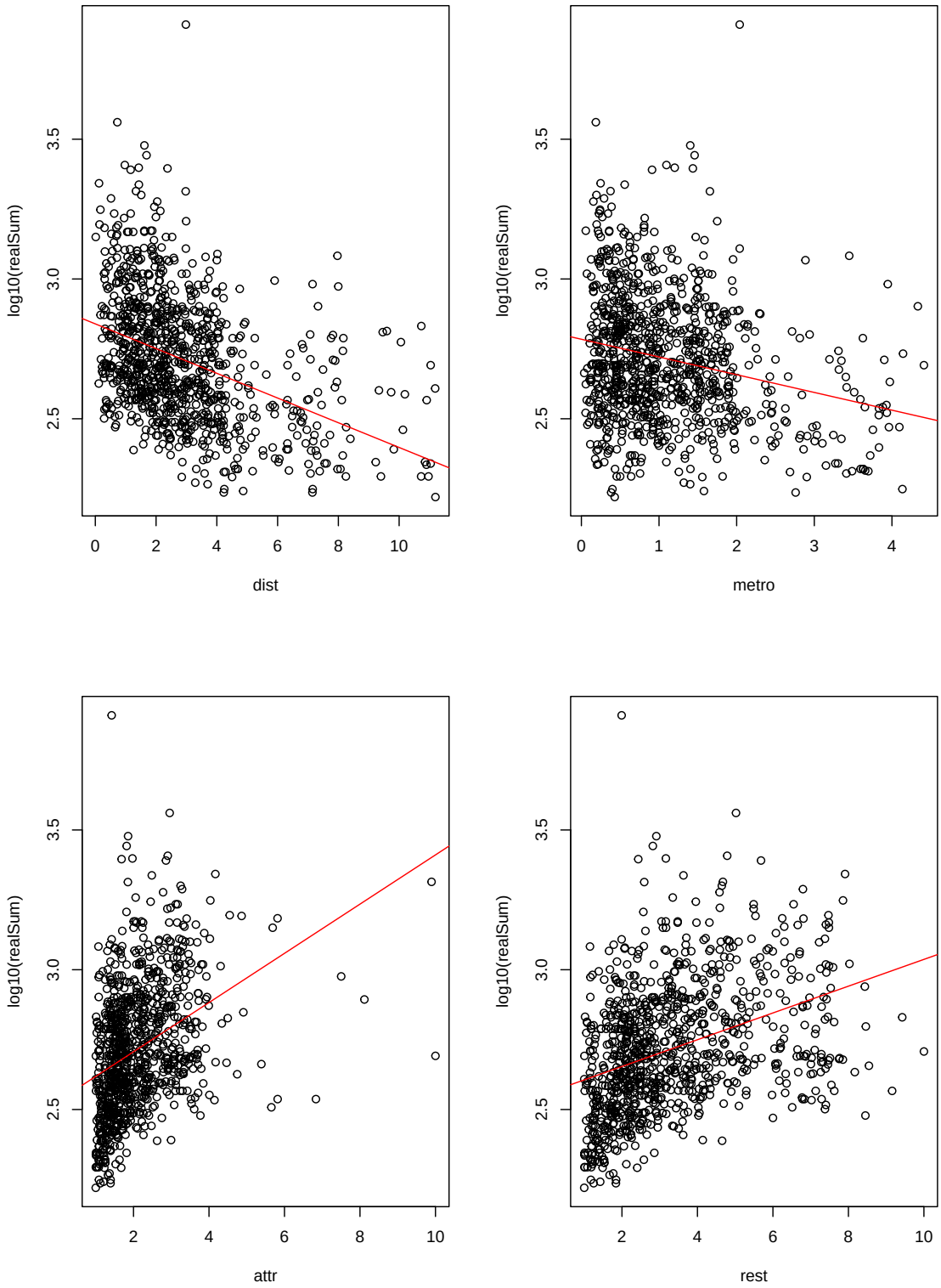
3.1 Kenmerken van de Steekproef

De gemiddelde prijs van de steekproef is niet significant verschillend van de gemiddelde prijs van de populatie. We kunnen er dus van uitgaan dat deze steekproef uit 2019 representatief is.

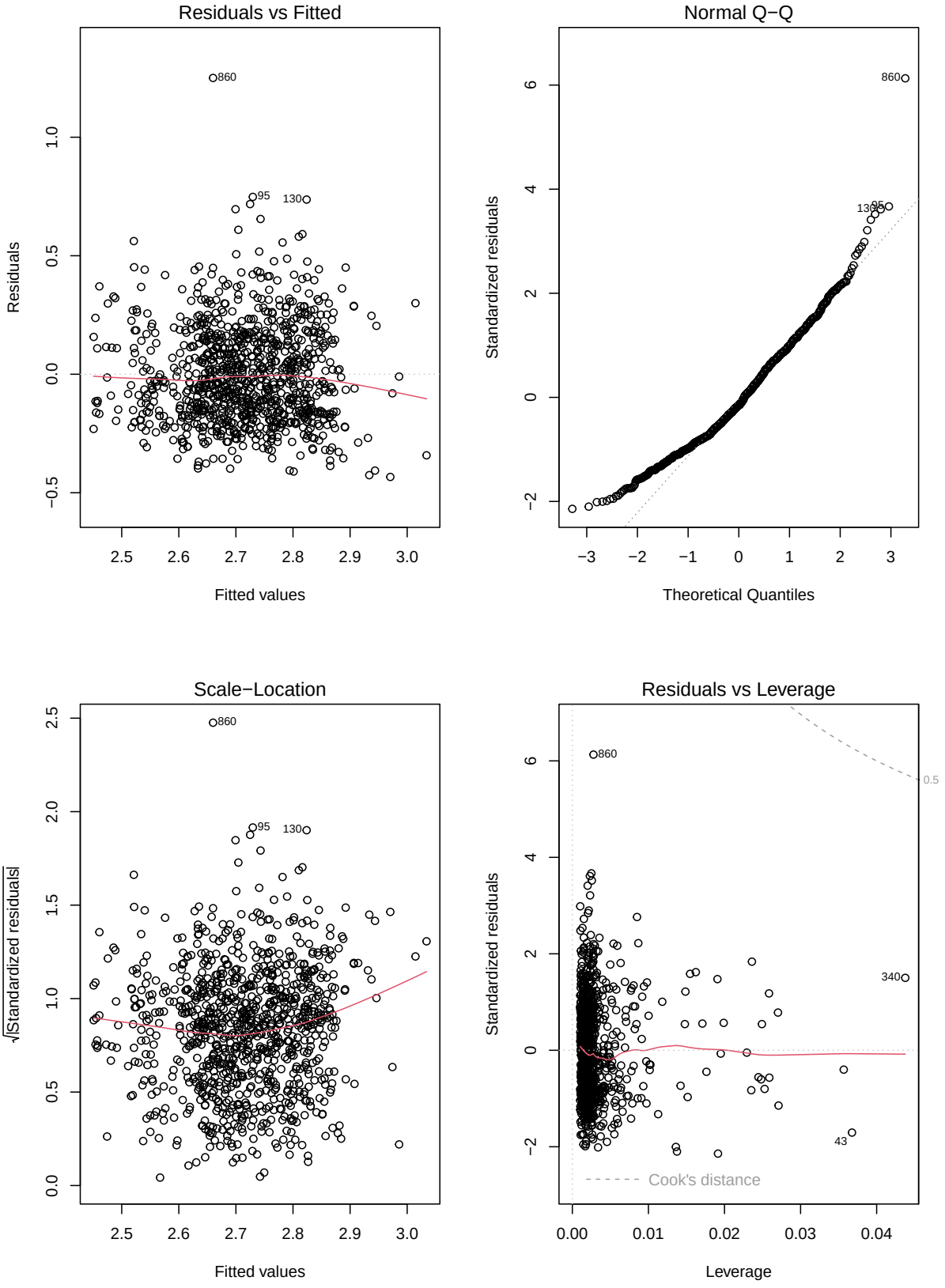
Daarnaast kunnen we op basis van de steekproef concluderen dat het aandeel particuliere hosts op Airbnb groter is dan het aandeel professionele.

Tabel 4: Pearson correlatie coëfficiënt van alle variabelen met de variabele realSum samen met de p-waardes van de verscheidene testen omtrent afhankelijkheid die hierboven beschreven staan.

Veranderlijke	r	p-waarde
room	-0.36	$<< 0.001$
capacity	0.53	$<< 0.001$
bedrooms	0.51	$<< 0.001$
dist	-0.27	$<< 0.001$
metro	-0.15	$<< 0.001$
attr	0.27	$<< 0.001$
rest	0.26	$<< 0.001$
host	-0.06	0.19
cleanliness	0.04	0.26
satisfaction	0.12	$<< 0.001$



Figuur 2: Spreidingsdiagrammen van het tiendelig logaritme van de totale kost in functie van de continue veranderlijken.



Figuur 3: Diagnotische plots van het finale regressiemodel (zie vergelijking 2).

3.2 Gemiddelde kost

Op basis van statistieken en op basis van de steekproef omtrent de gemiddelde kost kunnen we enkele zaken besluiten.

Ten eerste is de totale kost voor een reservatie van een Airbnb niet verschillend naargelang de score voor netheid 10 is of niet.

Bij het verschil in type verhuurder bekwamen we geen eenduidig antwoord via onze teststatistieken. Toch zullen we hier, met veel voorzichtigheid, besluiten dat de prijs niet significant verschilt naargelang de verhuurder particulier of professioneel is.

Ten derde vonden we op basis van de steekproef dat de gemiddelde prijs afhangt van het type woning. Zo zal de prijs voor een volledige woning gemiddeld gezien hoger liggen dan de prijs voor een afzonderlijke of gedeelde kamer.

3.3 Associatie met de verschillende veranderlijken

Uit de resultaten van deze testen (zie tabel 4) kunnen we besluiten dat het totale kostenplaatje afhankelijk is van heel wat verschillende variabelen.

Hoe groter de woning en de capaciteit van de woning hoe hoger de prijs.

Als het stadscentrum of een metrostation nabijgelegen zijn, zal de prijs hoger zijn. De totale prijs is positief afhankelijk van de restaurantscore, attractiescore en satisfaction.

We kunnen ook concluderen dat de prijs van een woning niet afhankelijk is van het type verhuurder en de netheid van de woning.

3.4 Verklaren van de kost

Uit de resultaten van de uitgevoerde testen kunnen we besluiten dat er een significant effect is van attractie op de totale kost en dat dit effect duidelijker wordt als van beide veranderlijken de tiendelige logaritme genomen wordt. We vonden via achterwaartse regressie dat het kostenplaatje van een woning op Airbnb dus vooral afhangt van de afstand tot het stadscentrum en de attractiescore. Als we de tiendelige logaritme van attractie gebruiken, zal de afhankelijkheid zelf nog stijgen. We besluiten dat het zinvol is afzonderlijke groepen te hanteren naargelang het een volledige woning is of niet .

Besluit

Na ons onderzoek komen we tot de conclusie dat de prijs voor de huur van een weekendverblijf in Amsterdam via Airbnb niet verhoogd is sinds 2019. De prijs van een accomodatie hangt af van een aantal factoren zoals, de nabijheid van bezienswaardigheden, de afstand tot het dichstbijzijnde metrostation en stadscentrum. Dit is te verwachten aangezien het logisch is dat een verblijf duurder wordt, naargelang het beter scoort op deze factoren. Dat de totale prijs onafhankelijk is van de score voor netheid of het type verhuurder hadden we niet direct verwacht.

Referenties

[1] Jun 2018. URL: `https://nl.wikipedia.org/wiki/Stelling_van_Cochran`.

[2] Mia Hubert e.a. *Statistiek en Wetenschap*. Acco, 2015.

[3] Shaun Turney. *Pearson correlation coefficient (R): Guide amp; examples*. Dec 2022. URL: `https://www.scribbr.com/statistics/pearson-correlation-coefficient/`.