KU LEUVEN kulak

Arthur Vanackere
Seppe Degryse
Informatica

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

In deze studie onderzoeken we de totale kostprijs voor het huren van een Airbnb-verblijf in Amsterdam gedurende een weekend als afhankelijke variabele. Het doel is om te bepalen of de gemiddelde kostprijs veranderd is sinds de dataset in 2019 is opgesteld. Om dit te bereiken vergelijken we de gemiddelde kostprijs van de steekproef met de huidige gemiddelde kostprijs van 620 euro volgens Airbnb in 2023. Het aantal observaties in onze steekproef is voldoende hoog met (n=977), waardoor we een Student t-test voor één gemiddelde kunnen uitvoeren.

Daarnaast onderzoeken we of het aandeel particuliere aanbieders in Amsterdam groter of kleiner is dan het aandeel professionele aanbieders (eigenaren met meer dan één verblijf). We berekenen de proportie van particuliere aanbieders in onze steekproef en voeren vervolgens een binomiaal test uit om te bepalen of er meer of minder dan de helft van de aanbieders particulier zijn.

Deze studie beoogt ook te onderzoeken of het aantal beschikbare slaapkamers van Airbnb-verblijven in Amsterdam de Poissonverdeling volgt. Om dit te onderzoeken, bepalen we eerst lambda als de mediaan van het aantal beschikbare slaapkamers. Vervolgens berekenen we de verwachte frequenties voor elk van de opties en vergelijken we deze met de geobserveerde frequenties door middel van de χ^2-test.

1.2 Gemiddelde kost

Het doel van deze studie is om te onderzoeken of de gemiddelde totale kost voor de weekendhuur van een verblijf voor 2 personen verschilt op verschillende categorieën. Om deze vraag te beantwoorden, voeren we eerst een normaliteitscontrole uit op de gegevens uit de steekproef, dit wordt gedaan met de Shaprio-Wilk-test.

Indien de gegevens normaal verdeeld zijn, wordt een bijkomende Fischer-test uitgevoerd om na te gaan of de varianties al dan niet verschillend zijn. Als de varianties gelijk zijn, wordt de Student t-test voor twee gemiddelden met ongepaarde groepen en gelijke varianties gebruikt. Als de varianties verschillend zijn, wordt de Student t-test voor twee gemiddelden met ongepaarde groepen en ongelijke varianties gebruikt.

Indien de gegevens niet normaal verdeeld zijn controleren we als het aantal observaties in beide groepen groot genoeg is zodat de centrale limiet stelling van toepassing is. Indien de CLS van toepassing is wordt ook gebruik gemaakt van de Student t-test voor twee gemiddelden met ongepaarde groepen en ongelijke varianties gebruikt. Indien de CLS niet van toepassing is wordt er gebruik gemaakt van de Wilcoxon-rangsom test. Dit proces wordt herhaald voor elke categorie die onderzocht moet worden.

1.3 Associatie met de verschillende veranderlijken

Om de associatie tussen de totale kost en de verschillende variabelen te onderzoeken, dient eerst het type data bepaald te worden. Indien de data nominaal is, wordt er een kruistabel opgesteld en wordt gecontroleerd of de Cochran-regel wordt nageleefd. Indien deze regel niet wordt nageleefd en de tabel 2×2 is, wordt er een Fisher-exactetest uitgevoerd. Als de regel niet wordt nageleefd en de tabel $m \times n$ is, wordt de tabel hercodeerd door rijen en kolommen samen te voegen totdat de Cochran-regel wel wordt nageleefd. Hiervoor wordt als regel gehanteerd dat $E_i > 5$ dient te zijn. Als de Cochran-regel vervolgens wordt nageleefd, wordt de chi-kwadraattest uitgevoerd.

Als de data ordinaal is en er sprake is van ties, worden dezelfde stappen gevolgd als bij nominale data. Als er geen sprake is van ties, wordt een Spearman correlatietest uitgevoerd.

Bij numerieke data wordt eerst een normaliteitstest uitgevoerd met behulp van de Shapiro-Wilk-test. Indien de gegevens niet normaal verdeeld zijn, wordt ook de Spearman correlatietest uitgevoerd. Indien de data wel normaal verdeeld is, wordt overgestapt op de Pearson correlatietest. Dit proces wordt herhaald voor elke variabele die onderzocht dient te worden.

1.4 Verklaren van de kost

Als eerste worden twee eenvoudige regressiemodellen opgesteld om de relatie tussen de totale kost en de attractiescore te onderzoeken. Daarnaast wordt een model opgesteld om de relatie tussen de tiendelige logaritme van beide variabelen te onderzoeken. Vervolgens wordt een meervoudig regressiemodel opgesteld om de kost te verklaren, waarbij alle continue variabelen worden gebruikt en achterwaartse regressie wordt toegepast. Indien nodig zal een tiendelige logaritmische transformatie worden toegepast om het model te verbeteren. Ten slotte wordt met behulp van een ANCOVA-model onderzocht of het zinvol is om een aparte vergelijking te hanteren afhankelijk van of het verblijf een volledige woning betreft of niet.

2 Resultaten

2.1 Kenmerken van de steekproef

Volgens de steekproef bedroeg de gemiddelde kost voor een weekendje Amsterdam in 2019 604,8 euro, terwijl Airbnb een gemiddelde kost van 620 euro rapporteerde. Het verschil van 15,1 euro is niet statistisch significant (t_{976} = -1,1, p = 0,3). Het aandeel particuliere aanbieders (65%) is significant hoger dan de helft (x = 636, n = 977, $p << 0{,}001$).

De Poisson-verdeling van het aantal beschikbare kamers vertoont een sterke afwijking van zijn werkelijke (onbekende) verdeling ($\chi^2_5 = 465$, $p << 0{,}001$). De kruistabel met de geobserveerde, verwachte waarden en residuen kan worden gevonden in tabel 4. De tabel voldoet aan de Cochran-regel: $E_i > 3$.

2.2 Gemiddelde kost

Na het uitvoeren van de Shapiro-Wilk test blijkt dat zowel de prijs van verblijven met een maximale netheidsscore als de prijs van verblijven zonder maximale netheidsscore niet normaal verdeeld zijn. Daarom passen we, gezien de centrale limiet stelling (CLS) ($n_1 = 599$, $n_2 = 378$), een t-test toe met verschillende varianties. De gemiddelde prijs voor verblijven met een maximale netheidsscore is 616,9 euro, terwijl dit voor de overige verblijven 581 euro is. Het verschil van 38,9 euro is niet statistisch significant ($t_{959} = 1{,}4$, $p = 0{,}2$).

In de tweede analyse vergelijken we de prijzen van particuliere aanbieders en professionele aanbieders. Uit de Shapiro-Wilk test blijkt dat zowel de prijzen van particuliere aanbieders als professionele aanbieders niet normaal verdeeld zijn. Desondanks passen we, gezien de CLS ($n_1 = 636$, $n_2 = 342$), opnieuw een t-test toe met verschillende varianties. De gemiddelde prijs van een particuliere aanbieder is 630 euro,

terwijl dit voor professionele aanbieders 557,8 euro is. Het verschil van 72,3 euro is statistisch significant (t_{552} = 2,2, p = 0,03).

In de derde analyse vergelijken we de prijzen van volledig verhuurde verblijven met afzonderlijk verhuurde verblijven. Uit de Shapiro-Wilk test blijkt dat de prijzen in beide categorieën niet normaal verdeeld zijn. Opnieuw is de CLS van toepassing (n_1 = 588, n_2 = 389) en passen we een t-test toe met verschillende varianties. De gemiddelde prijs van een volledig verhuurd verblijf is 736 euro, terwijl dit voor een afzonderlijk verhuurd verblijf 406,4 euro is. Het verschil van 329,6 euro is zeer statistisch significant (t_{850} = 14,1, $p << 0{,}001$).

2.3 Associatie met de verschillende veranderlijken

Na het hervormen van de kruistabel is uiteindelijk voldaan aan de Cochran-regel voor de ordinale en nominale variabelen: `room`, `host`, `cleanliness`, `bedrooms` en `capacity`. Vervolgens zijn er χ^2-testen uitgevoerd om de relaties te onderzoeken. De resultaten van deze testen zijn te vinden in de tabellen 5 - 9, en de bijbehorende grafieken zijn te vinden in figuur 1. Voor de overige numerieke variabelen: `distance`, `metro`, `attraction`, `restaurant` en `satisfaction`, is geconstateerd dat zowel de realSum als deze variabelen niet normaal verdeeld zijn volgens de Shapiro-Wilk test (tabel 10). Daarom hebben we gebruik gemaakt van de Spearman correlatietest om de relaties te onderzoeken. De resultaten van deze testen zijn te vinden in tabel 11 en de bijbehorende figuren zijn te vinden in figuur 2.

2.4 Verklaren van de kost

In het eerste eenvoudige regressiemodel tussen de totale kost en de attractiescore is opgemerkt dat de residuen aanzienlijk groot zijn, waarbij de residu-standaardafwijking 427,6 bedraagt en R^2 7,2% is. Dit duidt erop dat dit model waarschijnlijk geen goede pasvorm heeft.

Echter, het model dat gebruikmaakt van de tiendelige logaritme maakt een aanzienlijke verbetering. De residu-standaardafwijking wordt gereduceerd tot 0,2 en R^2 wordt verhoogd tot 17,6%. Dit geeft aan dat dit model beter past dan het vorige model.

Figuur 3 toont beide modellen, elk met hun betrouwbaarheids- en voorspellingsbanden. Dit helpt om de relatieve prestaties van de twee modellen te visualiseren en benadrukt de verbetering die wordt bereikt door de logaritmische transformatie toe te passen.

Uit het model met 5 regressoren,

$$\texttt{realSum} = \beta_0 + \beta_1 \cdot \texttt{distance} + \beta_2 \cdot \texttt{attraction} + \beta_3 \cdot \texttt{metro} + \beta_4 \cdot \texttt{restaurant} + \beta_5 \cdot \texttt{satisfaction}$$

worden achtereenvolgens de veranderlijken rest ($t_9$71 = 0,3, $p < 0{,}8$) en metro ($t_9$72 = -0,7, p = 0,5) verwijderd wegens niet significant volgens de Student t-test voor individuele regressiecoëfficiënt. In het resulterende model:

$$\log_{10}(\texttt{realSum}) = 2,5 - 0,05 \cdot \log_{10}(\texttt{distance}) + 0,5 \cdot \log_{10}(\texttt{attraction}) + 7,3 \times 10^{-6} \cdot \exp(\texttt{satisfaction}) \tag{1}$$

Zijn de regressiecoëfficiënten bij $\log_{10}$(`distance`) (t_{973} = -1,1, p = 0,3), $\log_{10}$(`attraction`) (t_{973} = 5,7, $p << 0{,}001$) en exp(`satisfaction`) (t_{973} = 6,5, $p << 0{,}001$) allemaal significant, evenals de globale F-test voor het totale regressiemodel ($F_{3,973}$ = 87,1, $p << 0{,}001$). De aangepaste determinatiecoëfficiënt bedraagt R^2_{adj} = 0,21.

In het ANCOVA-model bekomen door toevoegen van de categorische veranderlijke `room` en de interacties met $\log_{10}$(`distance`), $\log_{10}$(`attraction`) en exp(`satisfaction`), blijkt enkel de tweede significant door de partiële F-testen (tabel 12). Het gereduceerde model

$$\begin{cases} \log_{10}(\texttt{realSum})_{vol} = 2,6 + 0,02 \cdot \log_{10}(\texttt{distance}) + 0,6 \cdot \log_{10}(\texttt{attraction}) + 2,6 \cdot 10^{-6} \cdot \exp(\texttt{satisfaction}) \\ \log_{10}(\texttt{realSum})_{afz} = 2,5 - 0,2 \cdot \log_{10}(\texttt{distance}) + 0,2 \cdot \log_{10}(\texttt{attraction}) + 4,3 \cdot 10^{-6} \cdot \exp(\texttt{satisfaction}) \\ \log_{10}(\texttt{realSum})_{ged} = 1,6 + 0,7 \cdot \log_{10}(\texttt{distance}) + 1,5 \cdot \log_{10}(\texttt{attraction}) + 1,1 \cdot 10^{-5} \cdot \exp(\texttt{satisfaction}) \end{cases} \tag{2}$$

verklaart minder dan het globale model (1) volgens opnieuw een F-test ($F_{11,965} = 70,5, p << 0,001$). De R^2 daarintegen is verdubbeld ($R^2_1 = 0,21, R^2_2 = 0,44$). De modelveronderstellingen zijn vergelijkbaar met het oorspronkelijke model maar de steekproefuitschieters in een aantal groepen is ontoereikend voor betrouwbare schattingen. De uitschieters hebben in deze kleinere groepen dan ook een grotere invloed op het model. De verduidelijkende illustraties zijn terug te vinden op figuur 5.

3 Discussie

3.1 Kenmerken van de steekproef

Uit de analyse blijkt dat er geen significante stijging is in de gemiddelde totale kosten voor een weekendje Amsterdam in 2023 ten opzichte van 2019. Deze bevinding suggereert dat de steekproef mogelijk nog steeds representatief is.

Bovendien is het aandeel particuliere aanbieders significant hoger dan het aandeel professionele aanbieders, wat kan wijzen op strengere regelgeving in Amsterdam.

Verder blijkt uit de verdeling van het aantal beschikbare slaapkamers dat deze sterk afwijkt van een Poissonverdeling. Dit duidt erop dat het aantal beschikbare kamers niet willekeurig verdeeld is, maar eerder een lognormale verdeling volgt.

3.2 Gemiddelde kost

Uit de analyse blijkt dat er geen significante verschillen zijn in de totale kosten tussen verblijven met en zonder de maximale netheidsscore. Hieruit kunnen we concluderen dat de netheidsscore weinig tot geen invloed heeft op de totale kosten.

Daarentegen is de kostprijs voor het huren van een verblijf van een professionele aanbieder significant lager dan de kostprijs voor het huren van een verblijf van een particuliere aanbieder.

Verder is er een zeer significant prijsverschil tussen individuele en gedeelde verblijven in vergelijking met volledig verhuurde verblijven. Dit is logisch, gezien sommige kosten worden gedeeld bij gedeelde verblijven.

3.3 Associatie met de verschillende veranderlijken

Laten we eerst de ordinale en nominale variabelen van de dataset met betrekking tot realSum bespreken. Uit de analyse bleek dat de variabelen `room`, `bedrooms` en `capacity` sterk afhankelijk waren van `realSum`. Aan de andere kant waren `host` en `cleanliness` slechts licht afhankelijk. Dit is in lijn met onze eerdere bevindingen dat de netheidsscore geen grote invloed had op de prijs. De verduidelijkende grafieken en diagrammen van deze bevindingen zijn te vinden in figuur 1.

Wat betreft de numerieke variabelen, bleken ze allemaal gecorreleerd te zijn met de variabele `realSum`. Hoewel `satisfaction` de zwakste correlatie had van de vijf, was er nog steeds een waarneembare relatie. De verduidelijkende grafieken en diagrammen van deze analyses zijn te vinden in figuur 2.

3.4 Verklaren van de kost

Het regressiemodel voorspelt dat de totale kost met gemiddeld 5 cent per vertienvoudiging van het aantal kilometers van het stadscentrum afneemt. Daarnaast voorspelt het model een toename van gemiddeld 50 cent per vertienvoudiging van de attractiescore, en een toename van gemiddeld 0,0007 cent per vermindering met een factor e van de tevredenheidsscore. Echter, ondanks dat de afstand tot de dichtstbijzijnde metrohalte en de restaurantscore significant correleren met de totale kost, worden ze niet opgenomen in het regressiemodel. Dit kan wijzen op multicollineariteit met de afstand tot het stadscentrum. Bovendien blijkt uit de correlatieanalyse al dat de tevredenheidsscore een kleine invloed heeft in het model. Verduidelijkende grafieken bij het model kunnen gevonden worden op figuur 1.

Het ANCOVA-model toont aan dat het noodzakelijk is om aparte vergelijkingen te maken voor volledig verhuurde verblijven en niet-volledig verhuurde verblijven. Alleen de attractiescore blijkt een effect te hebben in beide scenario's. Bij volledig verhuurde verblijven heeft de attractiescore een drie keer groter effect dan bij niet-volledig verhuurde verblijven. Dit kan worden toegeschreven aan het feit dat gezinnen met kinderen vaker volledige verblijven huren, terwijl individuele reizigers vaker kiezen voor afzonderlijke verblijven puur voor overnachting. Vanwege het beperkte aantal gevallen in onze steekproef kunnen we geen uitspraken doen over gedeelde verblijven. Het ANCOVA-model heeft echter slechtere voorspellende capaciteiten dan het oorspronkelijke regressiemodel. Illustraties bij het ANCOVA-model zijn terug te vinden op figuur 5.

Besluit

De totale kosten voor het huren van een verblijf in Amsterdam tijdens een weekend worden beïnvloed door specifieke kenmerken van de accommodatie en kunnen gedeeltelijk worden verklaard door deze factoren. Uit de analyse blijkt dat de kosten stijgen naarmate de attractiescore en tevredenheidsscore toenemen, terwijl ze juist dalen naarmate de afstand tot het stadscentrum groter wordt. Hoewel het regressiemodel de totale kosten slechts gedeeltelijk kon verklaren, wijst dit op het bestaan van andere belangrijke factoren die buiten het bereik van dit onderzoek vielen.

Bovendien zijn er aanwijzingen dat de totale kosten afhangen van het feit of een verblijf volledig, afzonderlijk of gedeeld wordt verhuurd. Echter, vanwege het lage aantal groepen in dit onderzoek, was het niet mogelijk om betrouwbare afzonderlijke modellen te schatten voor elk van deze scenario's.

In toekomstig onderzoek is het relevant om een breder scala aan factoren te onderzoeken die mogelijk van invloed zijn op de totale kosten van verblijven in Amsterdam. Dit zou kunnen leiden tot een beter begrip van de prijsvorming in de accommodatiesector en het verstrekken van meer gedetailleerde informatie voor zowel reizigers als aanbieders bij hun besluitvorming.

A Appendix

Lijst van tabellen

Lijst van figuren

Tabel 1: Veranderlijken in de dataset.

	Naam	Beschrijving
1	realSum	Som van alle kosten
2	room	Soort verblijf
3	capacity	Maximaal aantal gasten
4	bedrooms	Aantal beschikbarae slaapkamers
5	distance	Afstand tot het stadscentrum
6	metro	Afstand to dichtsbijzijnde metro-halte
7	attraction	Attractiescore, nabijheid van bezienswaardigheden
8	restaurant	Restaurantscore, nabijheid van restaurants
9	host	Type verhuurder
10	cleanliness	Modale score voor netheid van het verblijf volgens gasten (op 10)
11	satisfaction	Tevredenheid van de gasten (op 10)

Tabel 2: Uittreksel uit de dataset.

	realSum	room	capacity	bedrooms	distance	metro
1	319.640	afzonderlijk	2	1	4.76336	0.852117
2	347.995	afzonderlijk	2	1	5.74831	3.65159
3	482.975	afzonderlijk	4	2	0.384872	0.439852
4	485.553	afzonderlijk	2	1	0.544723	0.318688
5	2771.54	volledig	4	3	1.68680	1.45840
6	1001.80	volledig	4	2	3.71914	1.19610

	attraction	restaurant	host	cleanliness	satisfaction
1	1.34102	1.70732	meer dan 4	9	8.80000
2	1.16748	1.36584	meer dan 4	9	8.70000
3	3.20334	7.76708	meer dan 4	9	9.00000
4	3.49351	7.27606	enige	10	9.80000
5	1.81786	2.81835	enige	10	10.0000
6	1.31822	1.68182	enige	9	9.60000

Tabel 3: Basisstatistieken.

Naam	Gemiddelde en standaardfout	Bereik	Algemene vorm
`realSum`	491,6 ± 443,7	[165,9; 8130,7]	Rechtsscheef met zware staart en uitschieters
`capacity`	2,77 ± 1,0	[2,0; 6,0]	Multimodaal
`bedrooms`	1,30 ± 0,7	[0,0; 5,0]	Eerder lognormaal
`distance`	2,8 ± 2,0	[0,0; 11,2]	Rechtsscheef met zware staart
`metro`	1,1 ± 0,8	[0,0; 4,4]	Rechtsscheef met lichte staart
`restaurant`	3,3 ± 1,8	[1,0; 10,0]	Rechtsscheef met lichte staart
`cleanliness`	9,5 ± 0,8	[2,0; 10,0]	Linksscheef met zware staart
`satisfaction`	9,6 ± 0,7	[2,0; 10,0]	Linksscheef met zware staart en uitschieters

`room`	volledig	afzonderlijk	gedeeld
Aantal	588	385	4
Proportie	60,2%	39,4%	0,4%

`host`	enige	2 tot 4	meerdere
Aantal	636	249	92
Proportie	65,1%	25,5%	9,4%

	0	1	2	3	4	5
Waargenomen	63	636	210	57	9	2
Verwacht	360	360	180	60	15	3
Residuen	-15,6	14,6	2,3	-0,4	-1,5	-0,6

Tabel 4: Vergelijkende tabel van de Poissonverdeling van de variabele bedrooms met zijn werkelijke verdeling.

GEOBSERVEERD	Afzonderlijk	Volledig	**VERWACHT**	Afzonderlijk	Volledig
]0, 200]	14	0	]0, 200]	5,6	8,4
]200, 300]	111	6	]200, 300]	46,6	70,4
]300, 400]	130	69	]300, 400]	79,2	119,8
]400, 500]	61	120	]400, 500]	72,1	108,9
]500, 600]	21	83	]500, 600]	41,4	62,6
]600, 700]	22	83	]600, 700]	41,8	63,2
]700, 800]	12	69	]700, 800]	32,3	48,7
]800, 1000]	9	57	]800, 1000]	26,3	39,7
]1000, 2000]	9	89	]1000, 2000]	39,0	59,0
]2000, 9000]	0	12	]2000, 9000]	4,8	7,2

RESIDUEN	Afzonderlijk	Volledig
]0, 200]	3.6	-2.9
]200, 300]	9.4	-7.7
]300, 400]	5.7	-4.6
]400, 500]	-1.3	1.1
]500, 600]	-3.2	2.6
]600, 700]	-3.1	2.5
]700, 800]	-3.6	2.9
]800, 1000]	-3.4	2.7
]1000, 2000]	-4.8	3.9
]2000, 9000]	-2.2	1.8

Tabel 5: Resultaten van de hervormde kruistabel tussen *realSum* en *room*

GEOBSERVEERD	1	2	4-6	**VERWACHT**	1	2	4-6
]0, 200]	11	1	2	]0,200]	8.5	0.9	4.7
]200, 300]	111	4	2	]200,300]	70.8	7.2	39.0
]300, 400]	176	7	16	]300,400]	120.4	12.2	66.4
]400, 500]	141	16	24	]400,500]	109.5	11.1	60.4
]500, 600]	61	11	32	]500,600]	62.9	6.4	34.7
]600, 700]	39	11	55	]600,700]	63.5	6.4	35.0
]700, 800]	23	2	56	]700,800]	49.0	5.0	27.0
]800, 1000]	14	4	48	]800,1000]	39.9	4.1	22.0
]1000, 2000]	14	4	80	]1000,2000]	59.3	6.0	32.7
]2000, 9000]	1	0	11	]2000,9000]	7.3	0.7	4.0

RESIDUEN	1	2	4-6
]0, 200]	0.9	0.2	-1.2
]200, 300]	4.8	-1.2	-5.9
]300, 400]	5.1	-1.5	-6.2
]400, 500]	3.0	1.5	-4.7
]500, 600]	-0.2	1.8	-0.5
]600, 700]	-3.1	1.8	3.4
]700, 800]	-3.7	-1.3	5.6
]800, 1000]	-4.1	-0.0	5.5
]1000, 2000]	-5.9	-0.8	8.3
]2000, 9000]	-2.3	-0.9	3.5

Tabel 6: Resultaten van de hervormde kruistabel tussen *realSum* en *capacity*

GEOBSERVEERD	0	1	2-5
]0, 300]	8	121	2
]300, 400]	15	176	8
]400, 500]	23	141	17
]500, 600]	9	68	27
]600, 700]	4	53	48
]700, 800]	2	36	43
]800, 1000]	0	23	43
]1000, 9000]	2	18	90

VERWACHT	0	1	2-5
]0, 300]	8.4	85.3	37.3
]300, 400]	12.8	129.5	56.6
]400, 500]	11.7	117.8	51.5
]500, 600]	6.7	67.7	29.6
]600, 700]	6.8	68.4	29.9
]700, 800]	5.2	52.7	23.0
]800, 1000]	4.3	43.0	18.8
]1000, 9000]	7.1	71.6	31.3

RESIDUEN	0	1	2-5
]0, 200]	-0.95	1.62	-2.00
]200, 300]	0.17	3.53	-5.42
]300, 400]	0.61	4.08	-6.46
]400, 500]	3.32	2.13	-4.81
]500, 600]	0.89	0.04	-0.48
]600, 700]	-1.06	-1.86	3.32
]700, 800]	-1.41	-2.30	4.16
]800, 1000]	-2.06	-3.05	5.59
]1000, 2000]	-1.72	-5.98	9.87
]2000, 9000]	-0.88	-2.08	3.56

Tabel 7: Resultaten van de hervormde kruistabel tussen *realSum* en *bedrooms*

GEOBSERVEERD	Enige	Meerdere
]0, 200]	7	7
]200, 300]	64	53
]300, 400]	112	87
]400, 500]	116	65
]500, 600]	78	26
]600, 700]	72	33
]700, 800]	60	21
]800, 1000]	47	19
]1000, 2000]	71	27
]2000, 9000]	9	3

VERWACHT	Enige	Meerdere
]0, 200]	9.1	4.9
]200, 300]	76.2	40.8
]300, 400]	129.5	69.5
]400, 500]	117.8	63.2
]500, 600]	67.7	36.3
]600, 700]	68.4	36.6
]700, 800]	52.7	28.3
]800, 1000]	43.0	23.0
]1000, 2000]	63.8	34.2
]2000, 9000]	7.8	4.2

RESIDUEN	Enige	Meerdere
]0, 200]	-0.7	0.96
]200, 300]	-1.39	1.90
]300, 400]	-1.54	2.11
]400, 500]	-0.17	0.23
]500, 600]	1.25	-1.71
]600, 700]	0.44	-0.60
]700, 800]	1.00	-1.37
]800, 1000]	0.62	-0.84
]1000, 2000]	0.90	-1.23
]2000, 9000]	0.43	-0.58

Tabel 8: Resultaten van de hervormde kruistabel tussen *realSum* en *host*

GEOBSERVEERD	2-8	9	10	**VERWACHT**	2-8	9	10
]0, 300]	21	40	70	]0, 300]	13,4	37,3	80,3
]300, 400]	15	57	127	]300, 400]	20,4	56,6	122,0
]400, 500]	15	54	112	]400, 500]	18,5	51,5	111,0
]500, 600]	6	28	70	]500, 600]	10,6	29,6	63,8
]600, 700]	20	28	57	]600, 700]	10,7	29,9	64,4
]700, 800]	10	25	46	]700, 800]	8,3	23,0	49,7
]800, 1000]	4	11	51	]800, 1000]	6,8	18,8	40,5
]1000, 9000]	9	35	66	]1000, 9000]	11,6	31,3	67,4

RESIDUEN	2-8	9	10
]0, 300]	2.07	0.45	-1.15
]300, 400]	-1.19	0.05	0.45
]400, 500]	-0.82	0.35	0.10
]500, 600]	-1.42	-0.29	0.78
]600, 700]	2.82	-0.34	-0.92
]700, 800]	0.59	0.41	-0.52
]800, 1000]	-1.06	-1.80	1.66
]1000, 9000]	-0.67	0.66	-0.18

Tabel 9: Resultaten van de hervormde kruistabel tussen *realSum* en *cleanliness*

	W-waarde	p-waarde
distance	0,9	$< 2,2 \times 10^{-16}$
metro	0,9	$< 2,2 \times 10^{-16}$
attraction	0,8	$< 2,2 \times 10^{-16}$
restaurant	0,9	$< 2,2 \times 10^{-16}$
satisfaction	0,7	$< 2,2 \times 10^{-16}$

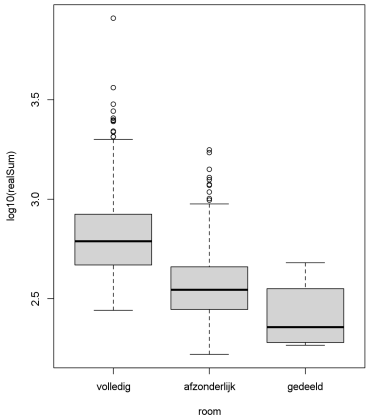
Tabel 10: Shapiro-Wilk test resultaten van de verschillende veranderlijken

	ρ	p-waarde
distance	-0,4	$< 2,2 \times 10^{-16}$
metro	-0,2	$< 9,5 \times 10^{-10}$
attraction	0,4	$< 2,2 \times 10^{-16}$
restaurant	0,4	$< 2,2 \times 10^{-16}$
satisfaction	0,2	$< 1,7 \times 10^{-7}$

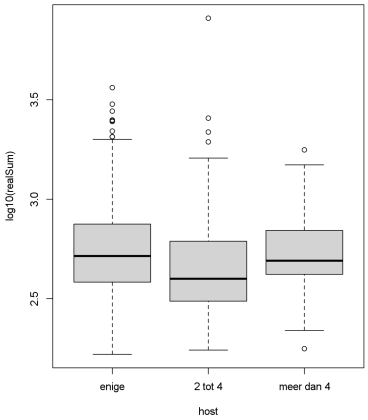
Tabel 11: Spearman correlatietest resultaten van de verschillende veranderlijken

	F-waarde	p-waarde
$\log_{10}$(`distance`)	0,03	0,4
$\log_{10}$(`attraction`)	0,2	0,05
exp(`satisfaction`)	0,03	0,6

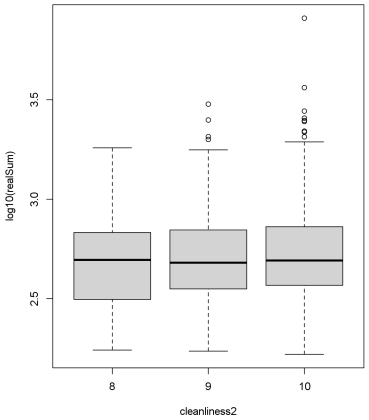
Tabel 12: Partiële F-test resultaten bij het ANCOVA-model ($F_{2,965}$)



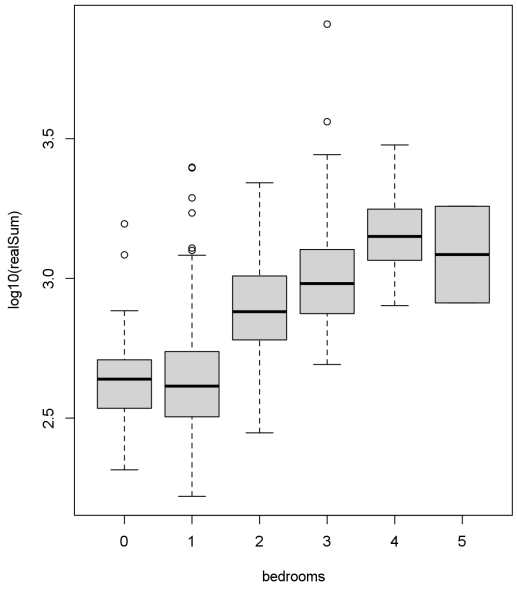
(a) Correlatie *room* en *realSum*



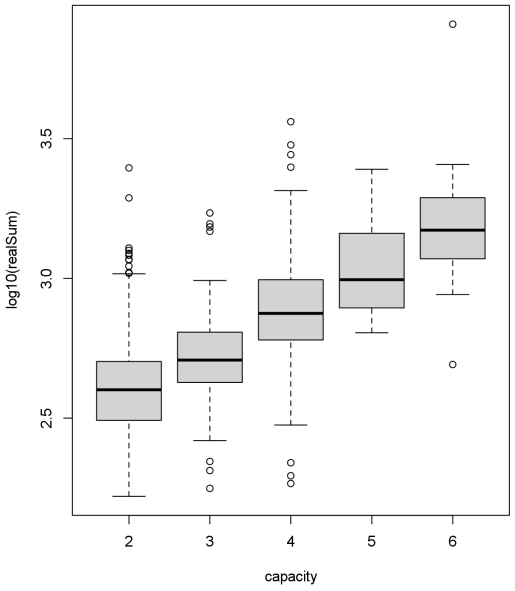
(b) Correlatie *host* en *realSum*



(c) Correlatie *cleanliness* en *realSum*

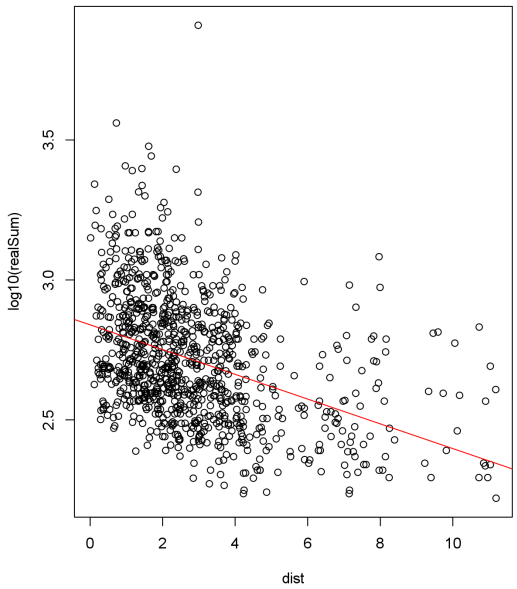


(d) Correlatie *bedrooms* en *realSum*

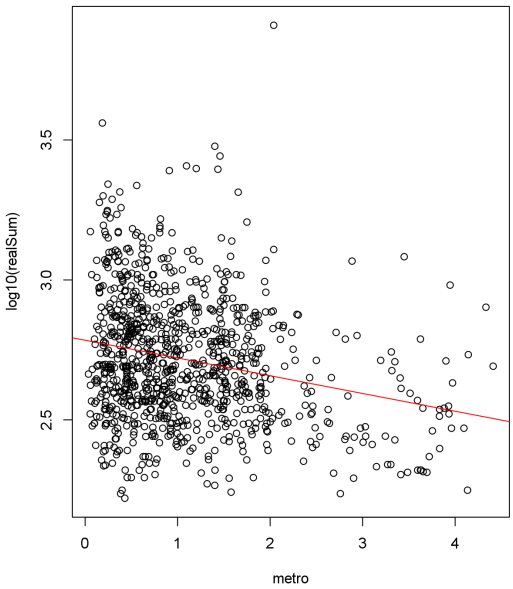


(e) Correlatie *capacity* en *realSum*

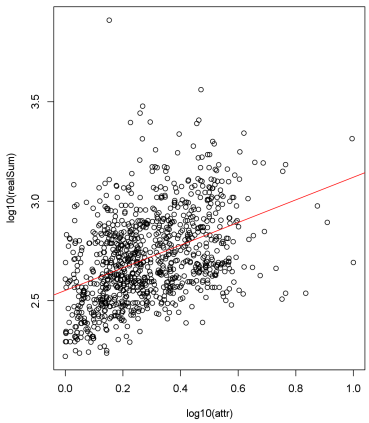
Figuur 1: Vergelijkende boxplots van de totale kost in functie van de nominale en ordinale veranderlijken



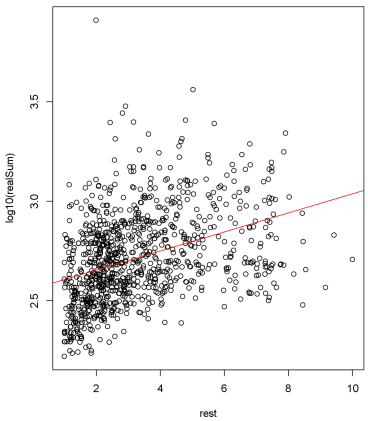
(a) Correlatie *distance* en *realSum*



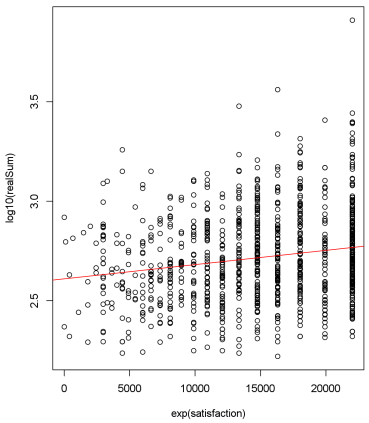
(b) Correlatie *metro* en *realSum*



(c) Correlatie *attraction* en *realSum*

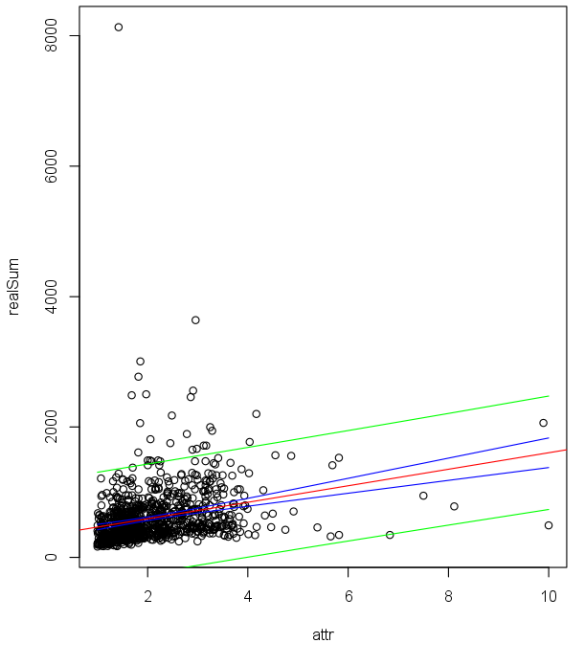


(d) Correlatie *restaurant* en *realSum Sum*

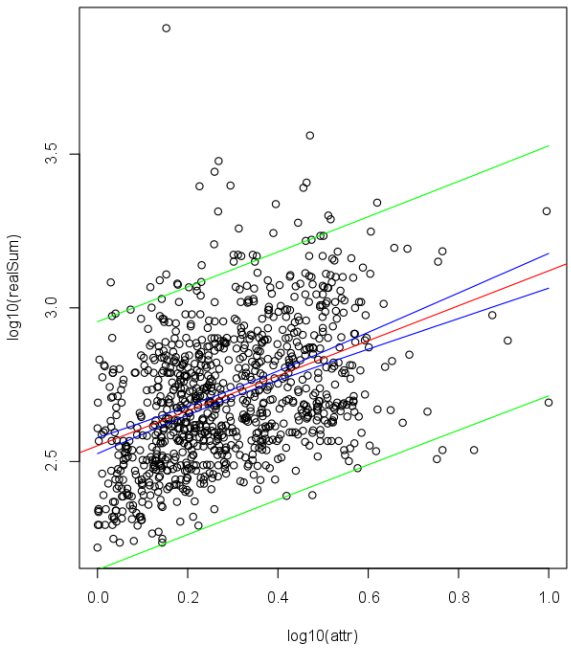


(e) Correlatie *satisfaction* en *real-*

Figuur 2: Spreidingsdiagrammen van de totale kost in functie van de numerieke veranderlijken

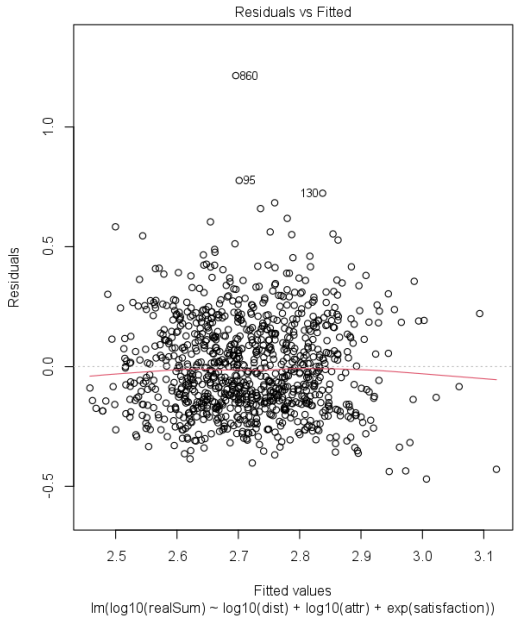


(a) Regressiemodel *attr* *realSum*

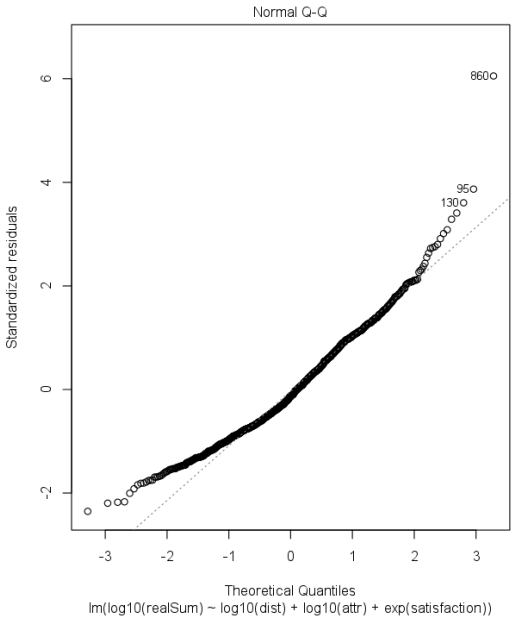


(b) Regressiemodel $\log_{10}$ *attr* $\log_{10}$ *attr* *realSum*

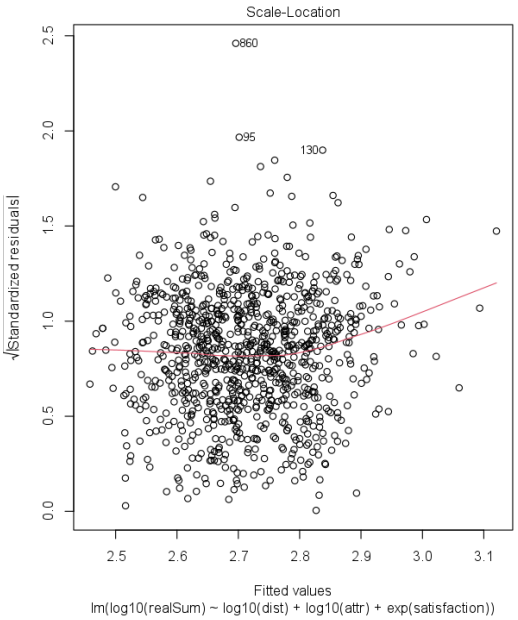
Figuur 3: Eenvoudige regressiemodellen van de totale kost in functie van de attractiescore



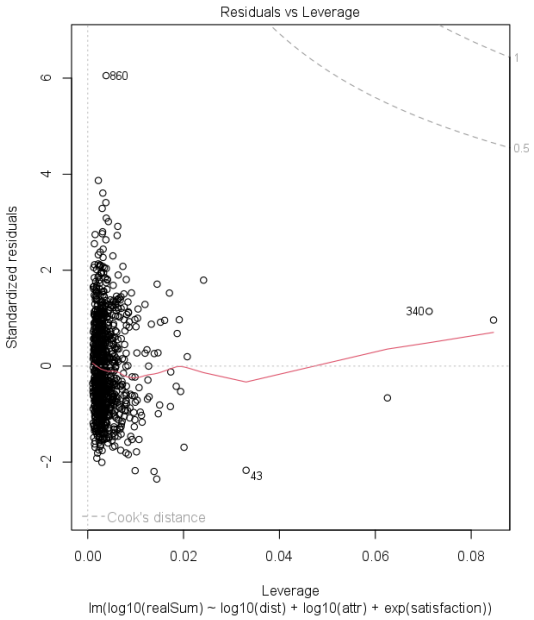
(a) Correlatie *metro* en *realSum*



(b) Correlatie *metro* en *realSum*

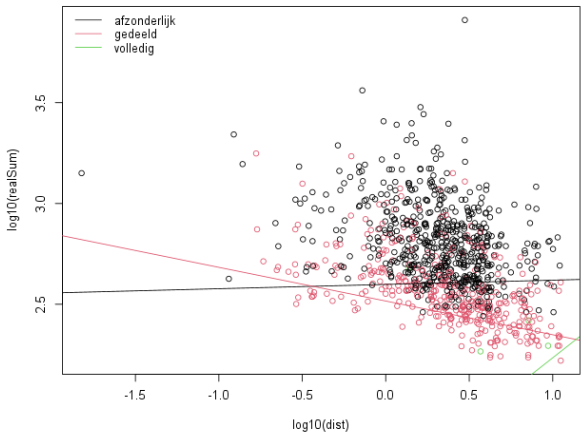


(c) Correlatie *metro* en *realSum*

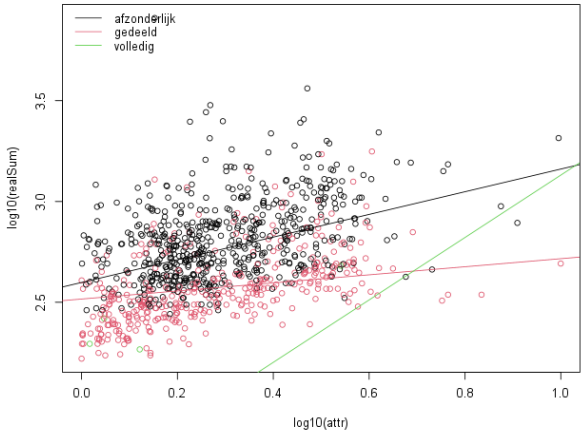


(d) Correlatie *metro* en *realSum*

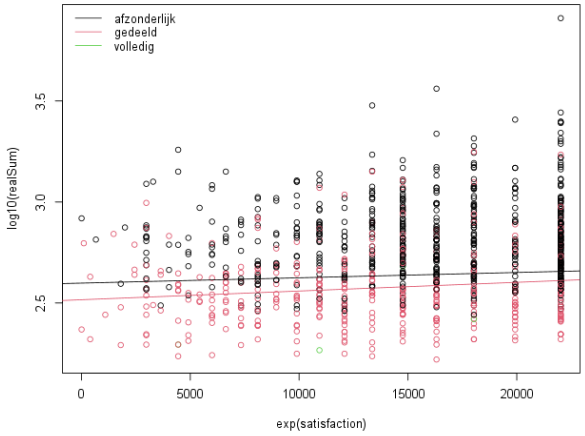
Figuur 4: Diagnotische grafieken van het regressiemodel



(a) Invloed van *distance* op de totale prijs afhankelijk van *room*



(b) Invloed van *attraction* op de totale prijs afhankelijk van *room*



(c) Invloed van *satisfaction* op de totale prijs afhankelijk van *room*

Figuur 5: Verduidelijkende grafieken bij het ANCOVA-model