Kost van een Airbnb-verblijf

Academiejaar 2022 – 2023 Project statistiek

1 Toelichting bij het projectwerk

Het projectwerk is een onderdeel van het examen (bedrijfs)statistiek, telt mee voor 4 van de 20 punten en wordt (behoudens uitzonderingen) per drie gemaakt. Er hoeft uiteraard slechts één gezamenlijk script en rapport te worden ingediend.

Het is de bedoeling om de leerstof in de praktijk te gebruiken. Daarom wordt er gewerkt met een realistische dataset die moet worden geïmporteerd in `R`. Met behulp van onderstaande onderzoeksvragen en opdrachten worden dan gepaste analyses uitgevoerd en conclusies getrokken. De evaluatie gebeurt op basis van volgende onderdelen:

1. Een script `naam1_naam2_naam3.R` met alle gebruikte commando's.
2. Een verslag `naam1_naam2_naam3.pdf` van *maximaal* 4 pagina's tekst (exclusief figuren en tabellen).

Beide bestanden worden ingediend via `Toledo`, vóór het einde van de lesperiode.

2 Gegevens en onderzoeksvraag

Het projectwerk zal de kostprijs van een Airbnb-verblijf in Amsterdam onderzoeken. Er wordt nagegaan met welke factoren de prijs voor het huren van een Airbnb-verblijf samenhangt en in welke mate de prijs daarmee kan worden voorspeld.
Er wordt gebruik gemaakt van gegevens die in 2019 werden verzameld via Airbnb en TripAdvisor in het kader van een onderzoeksproject uitgevoerd door Kristóf Gyódi en Łukasz Nawaro. De gegevens zijn vrij beschikbaar op `www.kaggle.com`.
De gegevensmatrix voor dit project bevat 977 observaties en 13 veranderlijken. Alle variabelen worden opgesomd in Tabel 1 en een uittreksel uit de dataset is te zien in Tabel 2. Enkele veranderlijken verdienen bijzondere aandacht en worden hier verder toegelicht. De kost voor het huren van een vakantiewoning via Airbnb bestaat uit verschillende componenten: het aantal nachten en gasten, de verblijfstaks, een schoonmaakvergoeding. Om een vergelijking mogelijk te maken wordt in deze dataset de effectieve som `realSum` van al deze kosten berekend bij reservatie van een woning voor 2 personen gedurende een weekend (vrijdag tot zondag). De attractiescore `attr` is voor woning i berekend als de sommatie $\texttt{attr}_i = \sum_j r_j/d_{ij}$ die loopt over alle bezienswaardigheden (*attractions*) in de stad, volgens de gegevens van TripAdvisor. Een bezienswaardigheid j met een hoger aantal reviews r_j en op kortere afstand d_{ij} van de woning i geeft een grotere bijdrage aan de attractiescore. De restaurantscore `rest` wordt op een analoge manier berekend voor restaurants.

3 Opdrachten

3.1 Data inlezen en manipuleren

Om deze opdracht zo realistisch mogelijk te maken, wordt gestart vanaf het bestand `airbnb.csv` met ruwe gegevens. Een eerste moeilijkheid is immers vaak om gegevens correct in het statistiekpakket in te lezen. Kijk daarom zorgvuldig na hoe de data is gestructureerd en importeer ze volgens de richtlijnen uit de handleiding en de helpbestanden van `R`.

Vervolgens wordt de dataset onderzoeksklaar gemaakt. Zet de categorische veranderlijken om naar het correcte datatype met passende labels. De attractie- en restaurantscore zijn moeilijk te interpreteren getallen. Herschaal deze tussen 1 en 10 zodat 1 (10) correspondeert met de laagste (hoogste) score in de dataset. Verwijder lengte- en breedtegraad uit de gegevensmatrix aangezien deze veranderlijken verder niet zullen worden gebruikt.

Dit deel van de analyse valt buiten de statistiek en hoort als dusdanig ook niet in het verslag. Introduceer in het verslag meteen de dataset zoals die hier is geconstrueerd.

Tabel 1: Veranderlijken in de dataset.

Naam	Beschrijving
realSum	Som van alle kosten* (euro)
room	Soort verblijf (1 = volledige woning, 2 = afzonderlijke kamer, 3 = gedeelde kamer)
capacity	Maximaal aantal gasten
bedrooms	Aantal beschikbare slaapkamers in het verblijf
dist	Afstand tot het stadscentrum (km)
metro	Afstand tot dichtstbijzijnde metro-halte (km)
attr	Attractiescore, nabijheid van bezienswaardigheden*
rest	Restaurantscore, nabijheid van restaurants*
lng	Lengtegraad
lat	Breedtegraad
host	Type verhuurder (0 = enige beschikbare woning, 1 = 2 tot 4 beschikbare woningen, 2 = meer dan 4 beschikbare woningen)
cleanliness	Modale score voor netheid van het verblijf volgensgasten (op 10)
satisfaction	Tevredenheid van de gasten (op 10)

*Zie sectie 2 voor een gedetailleerde beschrijving.

Tabel 2: Uittreksel uit de dataset.

	realSum	room	capacity	bedrooms	dist	metro	attr
1	319,64	2	2	1	4,7633	0,852	110,90
2	347,99	2	2	1	5,7483	3,651	75,27
3	482,97	2	4	2	0,3848	0,439	493,27
4	485,55	2	2	1	0,5447	0,318	552,84
5	2771,54	1	4	3	1,6867	1,458	208,80
6	1001,80	1	4	2	3,7191	1,196	106,22

	rest	lng	lat	host	cleanliness	satisfaction
1	136,98	4,84639	52,34137	2	9	8,8
2	95,38	4,97512	52,36103	2	9	8,7
3	875,11	4,89417	52,37663	2	9	9
4	815,30	4,90051	52,37508	0	10	9,8
5	272,31	4,88467	52,38749	0	10	10
6	133,87	4,86459	52,40175	0	9	9,6

3.2 Beschrijvende statistiek

Nu de dataset klaar is voor statistische analyses, kunnen ter verkenning voor elke veranderlijke gepaste datastatistieken worden berekend en grafieken worden gemaakt. Noteer voor elke veranderlijke of ze kwalitatief (nominaal of ordinaal) dan wel kwantitatief is (discreet of continu) en hoe de verdeling eruitziet: bekijk locatie- en spreidingsparameter, wat is het bereik, is er symmetrie, komen er uitschieters voor? Bekijk waar relevant of een veranderlijke normaal verdeeld is, indien niet of de logaritmische transformatie de normaliteit verbetert. Vergelijk al even grafisch hoe de kost van een verblijf samenhangt met de andere parameters van de woning.

Het is hier nog niet de bedoeling conclusies te trekken, maar om de gegevens beter te leren kennen. Het inlezen van data en verkennende beschrijvende statistieken horen doorgaans dan ook niet in het verslag, tenzij als illustratie bij het vervolg of indien er al opmerkelijke verbanden te zien zouden zijn. *Beperk verslaggeving van dit deel tot een overzichtstabel naar het model van Tabel 3, die basisstatistieken geeft bij elke numerieke veranderlijke en een frequentietabel bij de categorische.* Neem in deze tabel precies die veranderlijken op, die in het verslag aan bod komen. Vermeld dus *niet* de veranderlijken die nergens in je verslag aan bod komen en *wel* eventuele veranderlijken die pas later in het project worden aangemaakt (discretisaties, hercoderingen, . . .).

Tabel 3: Voorbeeldtabellen met basisstatistieken.

Naam	Gemiddelde ± standaardfout	Bereik	Algemene vorm
Lengte	$(12{,}3 \pm 0{,}4)$ m	$[5{,}6\,\text{m}, 27{,}8\,\text{m}]$	Eerder symmetrisch
Volume	$(12 \pm 3) \times 10\,\text{m}^3$	$[56\,\text{m}^3, 378\,\text{m}^3]$	Licht rechtsscheef
Massa	$(12{,}345 \pm 0{,}006) \times 10^6\,\text{kg}$	$[5 \times 10^6\,\text{kg}, 678 \times 10^6\,\text{kg}]$	Benaderend lognormaal

Geslacht	Man	Vrouw
Aantal	23	44
Proportie	66%	34%

Studierichting	Biologie	Economie	Wiskunde	`NA`
Aantal	23	25	17	2
Proportie	35%	38%	26%	

3.3 Inferentiële statistiek

3.3.1 Kenmerken van de steekproef

Anno 2023 kost een weekendje Amsterdam volgens Airbnb gemiddeld 620 euro. Is deze kost veranderd sinds het opstellen van de dataset in 2019?

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. Is het aandeel particuliere aanbieders (hosts met slechts 1 beschikbare woning) in Amsterdam groter of kleiner dan het aandeel professionele aanbieders (eigenaars met meer dan 1 verblijf)?

Ga na of het aantal beschikbare slaapkamers in een verblijf de Poissonverdeling volgt.

Uit het verslag moet duidelijk zijn welke testen werden uitgevoerd, hoe de voorwaarden werden nagegaan, wat de relevante statistieken zijn en welke conclusie er wordt getrokken. Illustreer significante resultaten.

3.3.2 Gemiddelde kost

Onderzoek of de totale kost voor de weekendhuur van een verblijf voor 2 personen verschilt

- naargelang het verblijf de maximumscore heeft voor netheid of niet;
- naargelang de eigenaar slechts één verblijf aanbiedt of niet;
- naargelang de volledige woning wordt verhuurd of niet.

Uit het verslag moet duidelijk zijn welke testen werden uitgevoerd, hoe de voorwaarden werden nagegaan, wat de relevante statistieken zijn en welke conclusie er wordt getrokken. Illustreer significante resultaten.

3.3.3 Associatie met de verschillende veranderlijken

Ga na of er afhankelijkheid is tussen de kost van een verblijf en elk van de andere veranderlijken. Deze analyse geeft meteen ook een idee van welke veranderlijken een rol kunnen spelen in het verklaren van de opbrengst.

Maak (een) overzichtelijke tabel(len) van deze resultaten. Vat bondig samen welke inzichten zijn verkregen. Voorzie telkens een verduidelijkende grafiek.

3.3.4 Verklaren van de opbrengst

Maak eerst twee verschillende eenvoudige regressiemodellen: één voor de kost in functie van de attractiescore en één voor de tiendelige logaritme van de kost in functie van de tiendelige logaritme van de attractiescore. Het enige doel van deze modellen is om één duidelijke figuur te maken met daarop de kost van een verblijf in functie van de attractiescore en beide eenvoudige regressiemodellen samen met de betrouwbaarheids- en predictiebanden. *Neem deze grafiek ter illustratie op in het verslag en leg aan de hand hiervan uit of en waarom de logaritmische transformaties noodzakelijk zijn.*

Bouw vervolgens een meervoudig regressiemodel voor het verklaren van de kost, door alle continue veranderlijken te gebruiken en achterwaartse regressie toe te passen. Ga de modelveronderstellingen na en gebruik waar nodig (enkel) de logaritmische transformatie van de gebruikte veranderlijken. *Noteer in*

het verslag de vergelijking van het meest geschikte model en leg uit hoe dit model is bekomen. Bespreek de coëfficiënten en de kwaliteit van dit model. Voorzie diagnostische plots. Welke tekortkomingen heeft dit model en hoe zou het nog kunnen worden verbeterd?

Kijk tot slot of het zin heeft afzonderlijke vergelijkingen te hanteren naargelang het verblijf een volledige woning betreft of niet.

Noteer in het verslag je conclusies en (indien van toepassing) de afzonderlijke vergelijkingen voor het bepalen van kost van een Airbnb-verblijf naargelang het een volledige woning betreft of niet.

4 Instructies

Script. Bundel alle commando's in een script. Zorg dat het script correct werkt op basis van het originele databestand. Verwijder alle overbodige lijnen en voeg zeer summier wat commentaar toe aan elke stap, zodat de commando's bij elk onderdeel vlot terug te vinden zijn. Het is niet nodig om in het verslag uit te weiden over technische details van `R`, over moeilijkheden bij het importeren van de data of over veranderlijken die je verder niet meer gebruikt.

Neem van de uitvoer van het script enkel die statistieken en grafieken in je verslag over die werkelijk relevant zijn voor de opbouw van het verhaal. In het bijzonder is het vaak nuttig om een illustratie te voorzien bij significante resultaten of een weerhouden regressiemodel. Noteer alle statistieken met de juiste eenheid en een gepast aantal beduidende cijfers zoals uitgelegd in Tabel 4. Zorg er voor dat je grafieken duidelijk leesbaar zijn en voorzien van titel, as-titels en eenheden. Gebruik vectorafbeeldingen (`.pdf`) in plaats van bitmaps (`.png`, `.jpg`) en zeker geen screenshots. Verwijs minstens één keer vanuit de tekst naar elke figuur.

Verslag. Maak van het rapport een degelijk wetenschappelijk verslag. Het moet een doorlopende tekst zijn, die los te lezen is van de opgave en begrijpelijk is voor een buitenstaander met dezelfde kennis van statistiek als jijzelf. Volg hoe dan ook de richtlijnen die er in jouw studierichting worden gehanteerd voor het schrijven van wetenschappelijke teksten. Via `Toledo` zal een sjabloon worden verspreid voor dit rapport met daarin een al uitgeschreven inleiding en onderstaande structuur. Ter illustratie zal ook een voorbeeldrapport worden verspreid dat volgens deze structuur is uitgewerkt.

Inleiding. Elk wetenschappelijk verslag begint met een korte introductie van het probleem, de dataset, de relevante veranderlijken en de onderzoeksvraag, vaak opgesplitst in meerdere hypothesen.

Methode. Vervolgens wordt per onderzoekshypothese uitgelegd welke statistische analyse zal worden gebruikt en hoe de voorwaarden worden nagegaan. Deze sectie kan in principe grotendeels voor het uitvoeren van de analyses worden geschreven. *Deze sectie beslaat maximaal 1 pagina.*

Resultaten. Een volgende sectie bevat alle numerieke resultaten van deze analyses, zo veel mogelijk in overzichtelijke tabellen. Dit deel bevat enkel objectieve gegevens, zoals statistieken en bekomen p-waarden, nog zonder interpretaties. *Deze sectie beslaat maximaal 1 pagina.*

Discussie. De interpretatie van de hypothesetesten en de verklaring van de statistische analyses horen in de discussiesectie. *Deze sectie beslaat maximaal 1 pagina.*

Besluit. Het besluit bestaat uit een compact en concreet antwoord op de onderzoeksvraag, geen herhaling van de resultaten maar een overkoepelende beschouwing bij de verschillende onderzochte hypotheses. Introduceer hier geen nieuwe elementen, maar probeer een zo algemeen mogelijke uitspraak te doen over wat de analyse je globaal heeft geleerd.

Bij elke opdracht staan specifieke aanwijzingen voor wat er precies in het verslag hoort (cursieve tekst). Volg deze en controleer na afloop of alle gevraagde elementen aanwezig zijn. Controleer ook grondig spelling en grammatica.

Respecteer de paginalimiet, bestandsnamen en deadline.

Tabel 4: Getalwaarden rapporteren.

Statistiek	Vuistregel	Voorbeelden
Meting x_i	Volgens meetnauwkeurigheid.	123 mm, 123,4 mm
Standaarddeviatie	Eén beduidend cijfer (meer in grote steekproeven).	5 mm
Schatting $\overline{x}, \hat{\beta}, \ldots$	Zelfde precisie als de standaardfout.	$9{,}812\,\mathrm{m/s^2}$, $5{,}97 \times 10^{24}\,\mathrm{kg}$
Standaardfout	Eén beduidend cijfer.	$0{,}001\,\mathrm{m/s^2}$, $0{,}05 \times 10^{24}\,\mathrm{kg}$
Percentage	In het algemeen geen decimalen,	5 %, 68 %, 95 %
	één beduidend cijfer voor waarde en complement.	0,03 %, 99,7 %
t, F, χ^2, ...	Maximaal 1 decimaal en twee beduidende cijfers.	$t = -1{,}3$, $F = 11$, $\chi^2 = 4{,}1$
p-waarde	Eén beduidend cijfer, *nooit* nul,	0,4, 0,08
	ongelijkheden voor grote of kleine waarden,	$> 0{,}9$, $< 0{,}001$
	wetenschappelijke notatie bij grote aantallen testen.	7×10^{-5}, 2×10^{-16}

Algemeen:

- Verzorg steeds de notatie van getallen, gebruik wetenschappelijke notatie waar nodig.
- Gebruik eenheden waar van toepassing.
- Afrondingen gelden louter voor rapportering, werk in berekeningen steeds met alle gevonden cijfers.
- Bovenstaande regels zijn richtinggevend, denk zelf na over de noodzaak van meer of minder cijfers.
- De rol van standaarddeviatie van een meting en standaardfout (standaarddeviatie van een statistiek, bijvoorbeeld het gemiddelde) is fundamenteel anders. De standaardfout geeft een idee over de nauwkeurigheid van de statistiek, terwijl de standaarddeviatie van een veranderlijke enkel een idee geeft over de spreiding van de verschillende waarden, niet over de precisie van één specifieke waarde. Vandaar de verschillende rol in het bepalen van het aantal beduidende cijfers.