KU LEUVEN kulak

Opdracht 2

Academiejaar 2018 – 2019 Statistische modellen en data-analyse

Toelichting

Het projectwerk is een onderdeel van het examen Statistische modellen en data-analyse, telt mee voor 3 van de 20 punten en wordt in groepen van twee studenten gemaakt.

Het is de bedoeling om de leerstof in de praktijk te gebruiken. Met behulp van onderstaande onderzoeksvragen en opdrachten worden dan gepaste analyses uitgevoerd en conclusies getrokken. De evaluatie gebeurt op basis van een script `naam1_voornaam1_naam2_voornaam2_Project2.R` met alle gebruikte commando's en een rapport `naam1_voornaam1_naam2_voornaam2_Project2.pdf` van *maximaal* 4 pagina's (zonder grafieken en tabellen mee te rekenen).

Vermeld op het titelblad van het rapport duidelijk jullie namen en studentennummers. Beide bestanden worden ingediend via Toledo ten laatste op maandag 17 juni.

1 Airbnb-verblijven in België

In dit project wordt nagegaan hoe de prijs en de beschikbaarheid van een Airbnb-verblijf in drie grote Belgische steden wordt bepaald op basis van een aantal publiek beschikbare gegevens. Vertrek van de data die is te vinden op de website *Inside Airbnb* voor Brussel, Antwerpen en Gent: download telkens het bestand `listings.csv` via *Get the Data*. Maak een extra veranderlijke `city` en voeg de drie tabellen samen.

Markeer prijzen die gelijk zijn aan nul en minimum aantal overnachtingen groter dan een jaar als ontbrekende waarden (`NA`) omdat deze panden vermoedelijk eigenlijk niet beschikbaar zijn.

Ongeveer een kwart van de panden is nooit beschikbaar (en dus mogelijk permanent verhuurd), wat maakt dat `availability_365` niet als maat voor populariteit van een pand kan worden gebruikt: eens volledig verhuurd wordt geen onderscheid meer gemaakt naargelang het meer of minder overbevraagd is. Het kan nuttig zijn om een onderscheid te maken tussen panden die wel en niet steeds bezet zijn. Maak hiervoor de binaire veranderlijke `full`.

Verander de datum `last_review` in het aantal dagen tussen de recentste review en de dag dat de data werden verzameld zoals aangegeven op de website *Inside Airbnb* (*Date Compiled*). Dit is een andere maat voor de vraag naar/beschikbaarheid van het pand.

1.1 Ligging en type verblijf

Hangt de gemiddelde huurprijs van een Airbnb-verblijf af van het type verblijf (`room_type`) en zo ja, zijn de prijzen voor alle types onderling verschillend?

Hangt de gemiddelde huurprijs van een Airbnb-verblijf af van de stad en zo ja, welke steden verschillen van elkaar?

Is het zinvol om de fijnere verdeling in stadswijken te gebruiken, of geeft die geen extra informatie bovenop de prijs per stad? Zijn er wijken met duidelijk hogere of lagere huurprijzen?

In bepaalde wijken zullen vermoedelijk eerder kamers (centrum), in andere eerder volledige huizen (rand) te huur zijn. Ga na of de ligging nog een rol speelt als je al rekening houdt met het type verblijf.

1.2 Model voor de huurprijs

Tracht op basis van de relevante beschikbare gegevens een model op te stellen om de prijs van een Airbnb-verblijf zo goed mogelijk te voorspellen. Bespreek in detail de kwaliteit van het gevonden model, welke problemen werden overwonnen en welke er blijven bestaan, welke methoden zijn gebruikt met welk effect. Bespreek in de mate van het mogelijke de betekenis van de coëfficiënten in het model.

1.3 Beschikbaarheid van een verblijf

Gebruik opnieuw alle relevante beschikbare gegevens en tracht een voorspellen of een pand al dan niet permanent verhuurd is (`full`). Bespreek volgens dezelfde instructies als de vorige opdracht.

Instructies

Bundel al je commando's in één script en zorg dat het script correct werkt op basis van de originele gegevens. Verwijder alle overbodige lijnen en voeg zeer summier wat commentaar toe aan elke stap, in het bijzonder bij berekeningen die het verslag niet halen.

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. Noteer alle statistieken met de juiste eenheid en een gepast aantal beduidende cijfers. Zorg er voor dat je grafieken duidelijk leesbaar zijn en voorzien van titel, astitels en eenheden.

Maak van je rapport een degelijk wetenschappelijk verslag, een doorlopende tekst die los te lezen is van de opgave en begrijpelijk is voor een buitenstaander met dezelfde kennis van statistiek als jijzelf. Focus op de interpretatie, maar zorg er voor dat de lezer begrijpt hoe tot het gevonden model en bijhorende conclusies wordt gekomen.

Hou je aan de paginalimiet, bestandsnamen en deadline.