KU LEUVEN kulak

Opdracht 1

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_Project1.R` met alle gebruikte commando's en een rapport `naam1_voornaam1_naam2_voornaam2_Project1.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 29 april.

1 Oorzaken van overlijden

In dit project wordt nagegaan hoe de oorzaken van overlijden verschillen tussen landen en regio's in de wereld. Er zijn schattingen van het aantal overlijdens (in duizenden) beschikbaar voor 183 landen, opgesplitst naar 32 verschillende doodsoorzaken. De landen worden gegroepeerd in 6 groepen volgens geografische ligging en 2 groepen naargelang de globale ontwikkeling van het betreffende land. De gegevens met betrekking tot de doodsoorzaken zijn afkomstig van de Wereldgezondheidsorganisatie [1] en betreffen het jaar 2016, de indeling in groepen is deze volgens de Verenigde Naties [2].

De gegevens in de huidige vorm stellen (schattingen van) absolute aantallen overlijdens voor. De som van de 32 cijfers per land geven dus (een schatting van) het totale aantal overlijdens in dat land tijdens het jaar 2016. Het is echter niet de bedoeling om de landen qua omvang te vergelijken, wel om het aandeel van een bepaalde doodsoorzaak in het totaal aantal overlijdens in een land te bekijken. Herschaal dus eerst en vooral de absolute cijfers naar percentages zodat de rijsommen 1 worden.

1.1 Clustering

Onderzoek of de landen kunnen opgedeeld worden in groepen op basis van de 32 doodsoorzaken. Zijn de gestandaardiseerde of de oorspronkelijke gegevens aangewezen? Beschrijf (indien van toepassing) de bekomen clusters. Zijn er overeenkomsten met bestaande indelingen? Rapporteer enkel de duidelijkste resultaten en stel deze grafisch voor.

1.2 Principaalcomponentenanalyse

Voer principaalcomponentenanalyse uit op de 32 doodsoorzaken. Hoeveel componenten zijn belangrijk? Beschrijf deze aan de hand van de meest in het oog springende veranderlijken en landen. In welke mate lijkt het mogelijk om de ligging en globale ontwikkeling van een land uit deze componenten af te lezen. Tracht dit alles te interpreteren.

1.3 Multivariate normaliteit

Verderop wordt de classificatie van landen op basis van doodsoorzaken bestudeerd, waarvoor het nodig is om na te gaan of de verdeling van de verklarende variabelen multivariaat normaal is voor elke groep. Onderzoek de hypothese van multivariaat normale verdeling van de doodsoorzaken voor de afzonderlijke groepen. Hou rekening met de conclusies in het vervolg van het onderzoek.

1.4 Classificatie

Ga na in hoeverre het mogelijk is om de regio van een land te identificeren aan de hand van de doodsoorzaken. Doe hetzelfde voor de globale ontwikkeling. Welke methode is het meest geschikt? Beschrijf de werking van het model. Welke landen worden niet correct ingedeeld en waarom?

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.

Referenties

[1] Global Health Estimates 2016: Deaths by Cause, Age, Sex, by Country and by Region, 2000-2016. Geneva, World Health Organization; 2018.

[2] Country classification, june 2018. Geneva, United Nations Conference on Trade and Development; 2018.