Project 2025

Deadline: 1 Juni 2025

Bachelor in de Informatica: **Elementaire statistiek**

Instructies

Dit project draait om het gebruik van *R* om data te analyseren en statistische methoden toe te passen. Het project telt mee voor vier van de twintig punten en moet individueel worden gemaakt en ingediend. **De deadline is 1 juni 2025 om 23:59.**

Los de onderstaande opgaven op met *R* en verwerk de antwoorden in een overzichtelijk en verzorgd verslag. Er is geen strikte paginalimiet, maar probeer niet veel meer dan vijf pagina's te gebruiken. Onderbouw je tekst met visuele of kwantitatieve resultaten uit *R*. Herhaal de opgave niet, begin direct met het antwoord. Gebruik of verwijs naar relevante theorie of formules uit de cursus, maar kopieer of herhaal niets letterlijk. Je hoeft geen methodes of technieken te gebruiken die niet in de lessen of oefenzittingen gezien zijn.

Het doel van dit project is om aan te tonen dat je met *R* kunt werken en in staat bent om de juiste statistische beslissingen te nemen. Zorg ervoor dat je verslag dit duidelijk laat zien. Beschrijf je denkproces en leg uit hoe je tot bepaalde inzichten of keuzes bent gekomen. Enkel het juiste of beste resultaat tonen zonder uitleg heeft weinig waarde.

De indiening via Blackboard bestaat uit een pdf met het verslag en de gebruikte *R* code. Geef de bestanden de naam *voornaam_achternaam.pdf* en *voornaam_achternaam.R*, zodat ze eenvoudig verwerkt kunnen worden. De R code wordt alleen bekeken bij onduidelijkheden in het verslag, de beoordeling gebeurt op basis van de pdf die op zichzelf te begrijpen moet zijn.

Bij vragen of problemen kan je contact opnemen met Flor via `flor.debois@uantwerpen.be`.

Persoonlijke data

Om ervoor te zorgen dat niet iedereen dezelfde data heeft, gebruik je de volgende *R* code om de drie datasets in te laden.

```
Lawine <- read.csv("Lawine.csv", row.names=1)
set.seed(0123456789); myLawineIndex = sample(1:500, 450)
myLawine = Lawine[myLawineIndex,]

Temperatuur <- read.csv("Temperatuur.csv", row.names=1)
set.seed(0123456789); myTempIndex = sample(1:2500, 250)
myTemp = Temperatuur[myTempIndex,]

Wind <- read.csv("Wind.csv", row.names=1)
set.seed(0123456789); myWindIndex = sample(1:1000, 50)
myWind = Wind[myWindIndex,]
```

In de tweede lijn van elk blok code verander je telkens het nummer `0123456789` naar jouw studentennummer. Vertrek in de opgaven van de datasets `myLawine`, `myTemp` en `myWind`.

Opgaven

Een vriend die in de Alpen woont, heeft samen met een lokale organisatie data verzameld over verschillende weersomstandigheden in de regio. Ze willen de gegevens graag laten analyseren en hebben een paar statistische vragen waarvoor ze jouw hulp vragen. De drie vragen die ze zichzelf stellen zijn de volgende:

1 - Lawine

De organisatie heeft data verzameld over de temperatuur (°C) en een score tussen 1 en 10 die aangeeft hoe groot het lawinegevaar is. Ze willen weten of er een significant verband bestaat tussen deze twee variabelen. Kun je dit verband eventueel formeel testen?

2 - Temperatuur

De tweede dataset gaat over de hoogte (m) en temperatuur (°C) in de regio. De vraag is of het mogelijk is om de temperatuur te voorspellen aan de hand van de hoogte. Kan je een model maken dat goede voorspellingen maakt?

3 - Wind

Tot slot is er data over de windsnelheid (km/h) op een specifiek punt waar de wind veel waait, maar met slechts twee mogelijke richtingen. De windsnelheid heeft zowel positieve als negatieve waarden, afhankelijk van de richting van de wind. De organisatie wil weten of de gemiddelde grote van de windsnelheid significant groter is dan 14 km/h.