Samenhang

1 Sociale media en studieresultaat

We willen nagaan of er een verband bestaat tussen het gebruik van sociale media en de prestaties van studenten.
We vragen aan 1064 studenten hoeveel uren ze gemiddeld per dag met sociale media bezig zijn. Onder sociale media verstaan we Facebook, YouTube, blogs, Twitter, MySpace en LinkedIn.
We gaan ook na wat de eindscore is van deze studenten op het einde van het jaar.
Het resultaat van deze enquête vind je in het bestand "socialeMediaVsPunten.csv".

1. lees dit bestand in. Let erop dat het aantal uren als getallen wordt gelezen. Verwijder de rijen waar een onbekende waarde staat. Verwijder nu de rijen met uitschieters. Hoeveel rijen schieten er over?
2. maak een histogram van de uren en de punten apart. Gebruik de formule van Sturges om het aantal klassen te bepalen
3. wat is het gemiddeld aantal uren dat een student aan sociale media spendeert?
4. wat is de standaardafwijking van het aantal uren? Wat betekent dit?
5. wat is de gemiddelde score van de studenten?
6. wat is de standaardafwijking van de score?
7. maak een scatterplot van de 2 variabelen. Welke correlatie verwacht je?
8. bereken de correlatie tussen de 2 variabelen. Doe dit ook eens via de Z-scores (Pearson). Wat besluit je?
9. welke rangcorrelatie vind je volgens Kendall? Wat betekent deze waarde?
10. heeft het zin om hier een regressielijn te bepalen? Waarom wel/niet? Bepaal de regressielijn als dit zin heeft.

2 Productiefouten batterijen

In een fabriek worden batterijen voor smartphones gefabriceerd. De vraag naar deze batterijen varieert nogal sterk. Men heeft het vermoeden dat als de vraag (en dus ook de productie) stijgt, het percentage ontploffende batterijen ook stijgt (en dus: de kwaliteit daalt onder de werkdruk). Dit wil men verifiëren.

Er wordt een meting gedaan. Iedere dag wordt het aantal geproduceerde batterijen bijgehouden. Als er een defecte batterij terug wordt gebracht (of er een ontploft), wordt nagegaan op welke dag deze geproduceerd werd. De data vind je terug in het bestand "batterijen.csv".
De eerste kolom geeft weer hoeveel er geproduceerd werd en de tweede kolom laat zien hoeveel batterijen er defect waren.

1. lees het bestand in. Er zijn lijnen met meer dan 2 waarden. Bekijk wat Python hiermee doet als je de waarden inleest. Verwijder deze lijnen uit de data. Er zijn uitschieters in de kolom aantalDefect. Verwijder deze. Hoeveel rijen hou je nu over?
2. maak een scatterplot van de twee variabelen. Is er een lineair verband?
3. welke correlatie vind je met de methode van Kendall?
4. eigenlijk zijn we niet geïnteresseerd in het absolute aantal defecte batterijen, maar wel in het percentage. Deel dus het aantal defecte door het aantal geproduceerde batterijen. Wat is het gemiddeld percentage defecte batterijen?
5. maak terug een scatterplot met het percentage defecte batterijen. Zie je het verschil met de vorige grafiek?
6. bereken nu de correlatie tussen het aantal geproduceerde batterijen en het percentage defecte batterijen. Wat besluit je?
7. wat is de waarde van R^2? Wat betekent dit?
8. teken de regressielijn bij de scatterplot. Wat zijn de waarden voor slope en intercept?
9. wat is de standaardschattingsfout? Wat betekent dit?
10. als de productie opgedreven zou worden tot 8000 batterijen/dag. Hoeveel defecte batterijen zou je dan verwachten?
11. Hoeveel batterijen kan de fabriek per dag produceren zodat er hoogstens 1 procent defect is?

3 Stress en weersomstandigheden

In deze studie vragen we ons af welke invloed het weer heeft op het stressgevoel. We vroegen een aantal personen om iedere dag hun stressgevoel te noteren. Bij deze gegevens noteerden we de gemiddelde temperatuur, luchtvochtigheid en hoeveelheid neerslag op die dag.

1. lees de data in. Welke kolommen bevatten NA-waarden? Verwijder de rijen met NA-waarden. Verwijder de rijen met extreme uitschieters bij neerslag. De luchtvochtigheid mag niet hoger zijn dan 100. Verwijder de rijen waarbij dat zo is. Hoeveel rijen hou je nu over?
2. welk meetniveau hebben de kolommen?
3. wat is de gemiddelde temperatuur, luchtvochtigheid en neerslag?
4. zoek de correlatie tussen stress en temperatuur. Welke methode gebruik je best? Welke waarde vind je? Wat betekent deze?
5. zoek de correlatie tussen stress en luchtvochtigheid. Welke waarde vind je? Wat betekent deze?
6. zoek de correlatie tussen stress en neerslag. Welke waarde vind je? Wat betekent deze?
7. Welke factoren spelen dus een rol in het stressgevoel?

4 Wanneer laad ik mijn smartphone op?

We weten dat de batterijlading van een smartphone sterk daalt als je hem veel gebruikt. Nu willen we dit in kaart brengen. We hebben een aantal testexemplaren volledig opgeladen en geven deze mee met personen die de smartphones gebruiken. Wanneer de batterijlading op 20% komt, noteren de personen het aantal uren standby en het aantal uren gebruikt. Je vindt deze data in "smartphones.csv".

1. lees de data in. Hoeveel rijen zijn er?
2. bereken per toestel het totaal aantal uren dat deze aan stond tot 20% (gebruikte uren en standby uren). Hoeveel uren vind je voor het eerste toestel?
3. we zoeken nu een verband tussen het aantal uren dat de smartphone gebruikt werd en het totaal aantal uren totdat de batterij op 20% stond. Maak eerst een scatterplot van deze 2 variabelen. Wat zie je?
4. bereken de correlatie volgens Pearson en Kendall. Wat zeggen deze waarden?
5. teken de regressielijn bij de scatterplot. Wat zijn de waarden voor slope en intercept?
6. wat is de waarde van R^2? Wat is de betekenis?
7. wat is de standaardschattingsfout? Wat betekent dit?
8. als iemand een smartphone gedurende 3 uur nodig heeft, hoe lang zal het dan duren vooraleer de smartphone opgeladen moet worden?
9. stel dat je een verband zoekt tussen het gebruikspercentage (deel usage door de totale tijd) en de totale tijd voordat je moet opladen. Zoek de juiste variabelen en maak een scatterplot. Wat zie je? Kan je hiervoor lineaire regressie gebruiken? Waarom wel of niet? Welk regressiemodel is in dit geval het beste? Wat is dan de uiteindelijke formule voor het model?