## PORTFOLIO-OPDRACHT

Om de *lichamelijke* en *geestelijke gezondheid* van de inwoners van Nederland in kaart te brengen is een groot onderzoek uitgevoerd, waarbij niet alleen gekeken is naar kenmerken van lichamelijke en geestelijke gezondheid, maar ook naar mogelijke factoren die daarvoor bepalend zijn.

Het onderzoek is uitgevoerd door een grote, aselecte, steekproef uit te voeren en de respondenten een uitgebreide vragenlijst voor te leggen. Uiteindelijk hebben 400 personen de vragenlijst ingevuld. De antwoorden op de vragenlijst zijn gecodeerd en terug te vinden in onderstaand overzicht van de gemeten variabelen en bijbehorende antwoordcategorieën:

\*Geslacht:

\*Leeftijd:

\*Hoogste\_opleiding:

\*Roker:

\*Sigaretten\_per\_dag:

\*Bloeddrukmedicatie (gebruik):

\*Hersenbloeding (gehad):

\*Hoge\_bloeddruk (gehad):

\*Suikerziekte (type):

\*Cholesterolwaarde:

0: man / 1: vrouw in jaren

1: Middelbare school

2: MBO

3: HBO

4: Universiteit

0: Nee / 1: Ja aantal

0: Ja / 1: Nee

0: Ja / 1: Nee

0: Ja / 1: Nee

0: Nee / 1: type I / 2: type II hoogte

hoogte

\*Bloeddruk\_systolisch: hoogte

\*Bloeddruk\_diastolisch: hoogte

\*BMI\_waarde: hoogte

\*Hartslagen (per minuut): aantal

\*Glucosegehalte: hoogte

\*Hoog\_hartaanval\_risico: 0: Ja / 1: Nee

\*Harddruggebruik: 0: Ja / 1: Nee

\*Softdruggebruik: 0: Ja / 1: Nee

\*Jaarlijks\_medische\_checkup: 0: Ja / 1: Nee

\*Wekelijkse\_uren\_screentime: aantal

\*Partner: 0: Ja / 1: Nee

\*Gescheiden\_ouders: 0: Ja / 1: Nee

\*Werkeloos: 0: Ja / 1: Nee

\*Voldoende\_dagelijkse\_beweging: 0: Ja / 1: Nee

\*Fruitconsumptie (dagelijks aantal stuks) 0: 0

1: 1

2: 2

3: 3 stuks of meer

\*Groentenconsumptie (dagelijks aantal gram): 0: 0

1: 0 – 100

2: 100 – 200

3: 300 gram of meer

\*Alcoholconsumptie (dagelijks aantal glazen): 0: 0

1: 1

2: 2

3: 3 glazen of meer

\*Lid\_sportclub: 0: Ja / 1: Nee

\*Chronische\_ziekte (heeft): 0: Ja / 1: Nee

\*Fysieke\_gezondheid (inschatting): 5: Uitstekend

4: Goed

3: Voldoende

2: Onvoldoende

1: Slecht

\*Mentale\_gezondheid (inschatting): 5: Uitstekend

4: Goed

3: Voldoende

2: Onvoldoende

1: Slecht

\*Kinderen (heeft): 0: Ja / 1: Nee

\*Autobezit: 0: Ja / 1: Nee

\*sociaal\_functioneren: 1: niet

2: enigszins

3: redelijk

4: in hoge mate

5: volledig

\*empatisch\_reageren (in staat tot): 1: niet

2: enigszins

3: redelijk

4: in hoge mate 5: volledig

\*positieve\_levenshouding (hebben van): 1: niet

2: enigszins

3: redelijk

4: in hoge mate 5: volledig

\*mentale\_vaardigheden (hebben van): 1: niet

2: enigszins

3: redelijk

4: in hoge mate 5: volledig

\*leven\_in\_balans (mate van): 1: niet

2: enigszins

3: redelijk

4: in hoge mate 5: volledig

\*persoonlijke\_grenzen (kunnen stellen): 1: niet

2: enigszins

3: redelijk

4: in hoge mate 5: volledig

## Bij de beoordeling van onderstaande vragen is het de bedoeling steeds een duidelijke toelichting en onderbouwing toe te voegen, met name voor eventueel gemaakte keuzes.

**Vragen en opdrachten:**

# **Deel A:**

Kies 5 variabelen uit de dataset zodanig dat alle meetniveaus vertegenwoordigd zijn.

1. Wat zijn de meetniveaus van de gekozen variabelen?
2. Bepaal op basis van de *eerste 5 rijen van de dataset* geschikte kengetallen voor elke gekozen variabele. Bereken deze kengetallen met de hand. Motiveer je keuzes en geef de berekeningen.
3. Bepaal m.b.v. Python op basis van de *gehele dataset* een geschikte grafiek voor   
   elke gekozen variabele. Motiveer je keuzes.
4. Bepaal m.b.v. Python op basis van de gehele dataset enkele geschikte kengetallen voor elke gekozen variabele. Motiveer je keuzes.
5. Bepaal nu de meetniveaus van alle variabelen.
6. Geef voor elke van de 4 besproken meetniveaus een overzicht van welke grafiek(en) en kengetal(len) zinvol zijn om te gebruiken.

Licht toe in welke gevallen je van de ene of juist van de andere grafiek of kengetal gebruik  
maakt. Interpreteer de gevonden resultaten.

Kies een variabele met “ja/nee” als mogelijke uitkomsten.  
Stel dat je bij deze variabele willekeurig 3 mensen selecteert.

1. Wat is dan de kans dat ze allemaal “ja” als antwoord gegeven hebben? Geef de berekening.
2. Wat is de kans dat slechts 1 persoon "ja” als antwoord heeft gegeven? Geef de berekening.

De variabele "*cholesterolwaarde*” heeft gehele getallen als mogelijke uitkomsten en de bijbehorende verdeling moet dus beschouwd worden als een discrete verdeling.

1. Wat is het verschil tussen een continue en een discrete verdeling?
2. Aan welke eigenschap(pen)moet elke continue en discrete verdeling voldoen?
3. Waarom doen we bij het uitvoeren van berekeningen aan "*cholesterolwaarde*” in de  
   praktijk toch alsof het een continue variabele is?

Kies 3 tweetallen variabelen, waarbij je bij elk tweetal tussen die variabelen een hoge

correlatie verwacht.

1. Bereken m.b.v. Python een passende correlatie en de significantie van die correlatie  
   tussen de variabelen van elk van deze tweetallen.   
   Motiveer je keuzes en geef voor elk van deze 3 gevallen je conclusie?

**Deel B:**

Uit ander onderzoek is gebleken dat verschillende aspecten, zoals de mate van sociaal functioneren en het hebben van een positieve levenshouding, een bijdrage leveren aan de mentale gezondheid van personen. Een variabele als “*mental\_health”* is dus alleen te meten door de verschillende aspecten ervan in kaart te brengen, de bijbehorende variabelen apart te meten en de uitkomsten samen te voegen. We doen dat door de nieuwe variabele “*mental\_health”* te maken door de gevonden uitkomsten (meetwaarden) van de afzonderlijke variabelen op te tellen. We gaan er daarbij vanuit dat in het eerdere onderzoek gebleken is dat in hoge mate personen die hoog/laag scoren op de verschillende aspecten ook hoog/laag scoren op de variabele “*mental\_health”*. Omdat “*mental\_health”* een variabele is die we gekregen hebben door een aantal variabelen op te tellen, noemen we dat een samengestelde variabele.

In het onderzoek hier worden 6 verschillende aspecten gebruikt om de variabele “*mental\_health”* te meten, t.w.: “*sociaal\_functioneren”*, “*empatisch\_reageren”*, “*positieve\_levenshouding”*, “*mentale\_vaardigheden”, “leven\_in\_balans” en “persoonlijke\_grenzen”*.

1. Aan welke eis(en) moeten de mogelijke uitkomsten van de variabelen “*sociaal\_functioneren*” t/m “persoonlijke\_grenzen” voldoen om die te mogen optellen   
   en zo de variabele “*mental\_health*” te definiëren?
2. Maak met behulp van Python de nieuwe variabele “*mental\_health”*.
3. Wat is nu formeel het meetniveau van deze variabele? Motiveer je antwoord.

Onder ruime voorwaarden, waar hier aan voldaan is, mogen we het meetniveau van de variabele “*mental\_health”* als interval veronderstellen.

1. Breng de variabele “*mental\_health”* in kaart door het gebruiken van een geschikte grafiek en het berekenen van een of meer geschikte kengetallen. Licht je keuzes toe.

In het onderzoek is aan de hand van de variabele “*Mentale\_gezondheid*” gemeten hoe de respondenten hun eigen mentale gezondheid inschatten. Het is echter niet vanzelfsprekend dat de eigen inschattingen van de respondenten overeenkomen met de gevonden waarden van mental\_health.

1. Bedenk en motiveer op welke wijze je zou kunnen nagaan in hoeverre de inschatting door de respondenten van hun mentale gezondheid (“*Mentale\_gezondheid*”) overeenkomt met de gemeten waarden van “*mental\_health”*.

**Deel C:**

Bij het bepalen van de Pearson-correlatiecoëfficiënt gaan we na in welke mate er een *lineair verband* bestaat tussen de uitkomsten van 2 variabelen: de onafhankelijke variabele X en de afhankelijke variabele Y.

1. Leg uit wat we verstaan onder een lineair verband.

De waarde van R2 geeft aan in welke mate de uitkomsten van de variabele X de uitkomsten van de variabele Y verklaren. In de praktijk zal het vaak zo zijn dat niet één maar meerdere variabelen, elk gedeeltelijk, de uitkomsten van Y verklaren. De techniek die we dan gebruiken heet *meervoudige (lineaire) regressie.*

1. Kies minimaal 3 geschikte variabelen waarvan je het idee hebt dat ze de “*BMI-waarde*” van een persoon medebepalen.
2. Voer m.b.v. Python een lineaire meervoudige regressie op deze variabelen uit.   
   Wat is je conclusie?
3. Bepaal m.b.v. Python de 95%-betrouwbaarheidsintervallen voor de regressiecoëfficiënten.

1. Kies minimaal 3 geschikte variabelen waarvan je het idee hebt dat ze verklarend zijn voor het gehad hebben van hoge bloeddruk (“*Hoge\_bloeddruk*”).   
   Voer m.b.v. Python een logistische regressie uit op deze variabelen.   
   Wat is je conclusie?

Van de variabele "*Cholesterolwaarde*” (= X) is uit *oud onderzoek* gebleken dat dit een normaalverdeelde variabele is met gemiddelde 230 en standaarddeviatie 40.

1. Ga na of de uitkomsten in onze dataset een afwijkend beeld laten zien.
2. Bereken met de hand en aan de hand van tabellen de volgende kans: P( X > 250).
3. Benader deze kans door gebruik te maken van de uitkomsten van deze variabele   
   in de dataset.

*Recentere onderzoeken* hebben echter laten zien dat het gemiddelde µ inmiddels 235 is   
en de standaarddeviatie s is gelijk aan 43.

1. Toets m.b.v. Python met α = 0,05 of we op basis van onze steekproef aan de uitkomst van het recentere onderzoek moeten vasthouden.

Alle waarden voor het gemiddelde µ waarvoor we de nulhypothese niet verwerpen,   
vormen het (1-α)\*100%-betrouwbaarheidsinterval.

1. Bepaal het 95%-betrouwbaarheidsinterval voor het gemiddelde µ.
2. Bepaal het 90%-betrouwbaarheidsinterval voor het gemiddelde µ.
3. Formuleer de hypothesen, kies de toetsingsgrootheid, formuleer het beslissingscriterium, en   
   bepaal of we voldoende reden hebben om aan te nemen dat het gemiddelde µ een andere waarde heeft dan 235.   
   Is het resultaat dat we vinden statistisch significant?
4. Zou het hier voor de berekeningen uitgemaakt hebben als we s hadden moeten schatten aan de hand van de standaarddeviatie s in de steekproef? Leg uit waarom wel/niet.

De variabele "*Bloeddruk\_diastolisch*” is normaalverdeeld met gemiddelde 82 en standaarddeviatie 12.

1. Ga na of de uitkomsten in onze dataset een afwijkend beeld laat zien of niet.
2. Bepaal met de hand en aan de hand van tabellen met 𝛼 = 0,05 de kans op de fout van de 2e soort indien in werkelijkheid µ = 90.
3. Toets met Python met 𝛼 = 0,05 in hoeverre er een verschil bestaat wat betreft de variabele “*BMI\_waarde”* voor de verschillende categorieën van de variabele “*Suikerziekte*".