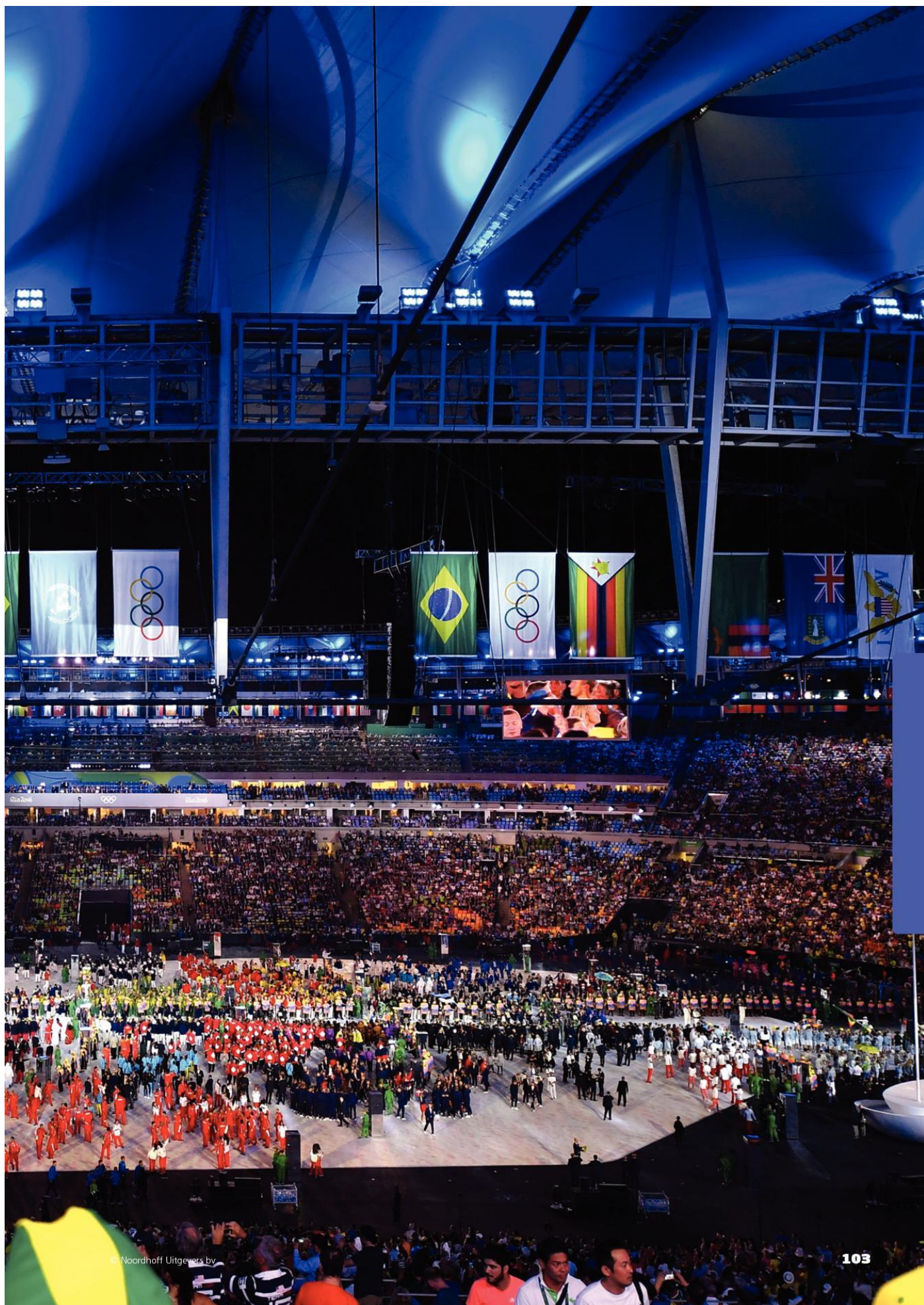




Statistiek met de computer

Wat leer je?

- In een spreadsheet data invoeren en berekeningen maken.
- Diagrammen maken met een spreadsheetprogramma.
- Informatie zoeken met filteren en sorteren in een grote dataset.
- Gegevens met behulp van formules combineren tot nieuwe gegevens.
- Met draaitabellen gegevens van een grote dataset samenvatten.
- Draaigrafieken maken.



Beginopdracht Olympische Spelen 2016

Aan de Olympische Zomerspelen in Rio de Janeiro 2016 deden meer dan 11 000 sporters uit 205 landen mee. Van al deze sporters is informatie verzameld.

Open het bestand *SportersRio2016*. Het bestand is gesorteerd op land, geboortedatum en geslacht. Dat is in een bepaalde volgorde gedaan.

- Onderzoek welke volgorde dat is.
- Welke sport beoefent de oudste sporter uit China?
- Wat is het gewicht van de enige Nederlandse taekwondo-deelnemer?
- Hoe heet de langste vrouw van het damesbasketbalteam uit Senegal?

Je kunt aantallen sporters tellen door te kijken naar de nummers van de rijen, die helemaal links worden weergegeven.

- Van hoeveel sporters is er informatie beschikbaar?
- Reken na dat 249 sporters uit Nederland kwamen. Hoeveel sporters kwamen uit België?

André vraagt zich af hoe de aantallen mannelijke en vrouwelijke sporters uit Nederland zich verhouden tot de aantallen uit België. Hij stelt de volgende kruistabel op.

	Nederland	België	totaal
man			
vrouw			
totaal			

- Neem de tabel over en vul in.
- Had Nederland procentueel gezien meer of minder vrouwelijke sporters dan België?

Voorkennis Centrummaten en boxplots

Theorie A Centrummaten

Om data te karakteriseren met één getal kun je een centrummaat gebruiken.

CENTRUMMATEN

gemiddelde	Het gemiddelde is $\frac{\text{som van de waarnemingsgetallen}}{\text{totale frequentie}}$.
mediaan	Schrijf de getallen in volgorde van grootte. De mediaan is het middelste getal bij een oneven aantal getallen. Bij een even aantal getallen is de mediaan het gemiddelde van de middelste twee getallen.
modus	De modus is het waarnemingsgetal met de grootste frequentie. Zijn er twee of meer waarnemingsgetallen met de grootste frequentie, dan is er geen modus.

Zijn er bij een statistisch onderzoek veel verschillende waarnemingsgetallen, dan maak je meestal een indeling in klassen. Bij de klasse $[20, 25)$ doet de linkergrens 20 mee, maar de rechtergrens 25 niet. Het klassenmidden is 22,5.

Om bij een klassenindeling het gemiddelde te schatten, maak je gebruik van de klassenmiddens. Je kunt bij een klassenindeling de mediaan niet meer berekenen, maar je kunt wel aangeven in welke klasse de mediaan ligt. In plaats van de modus heb je bij een klassenindeling te maken met de modale klasse.

- 1** Van een groep museumbezoekers is de leeftijd genoteerd. De resultaten staan in de tabel hiernaast.
- a** Geef een schatting van de gemiddelde leeftijd. Rond af op één decimaal.
 - b** In welke klasse ligt de mediaan?
 - c** Geef de modale klasse.

leeftijd	frequentie
$[0, 15)$	12
$[15, 30)$	10
$[30, 45)$	13
$[45, 60)$	24
$[60, 75)$	21
$[75, 90)$	2

Theorie B Boxplots en spreidingsmaten

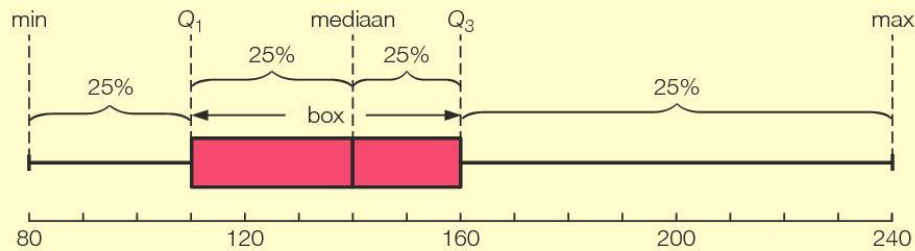
Een boxplot is een diagram dat de spreiding van data in beeld brengt.

Alleen bij een kwantitatieve variabele kun je een boxplot maken.

De mediaan verdeelt een serie waarnemingsgetallen in twee even grote groepen. Van elk van deze groepen kun je weer de mediaan bepalen.

De mediaan van de eerste helft heet het eerste kwartiel. Notatie Q_1 .

De mediaan van de tweede helft heet het derde kwartiel. Notatie Q_3 .



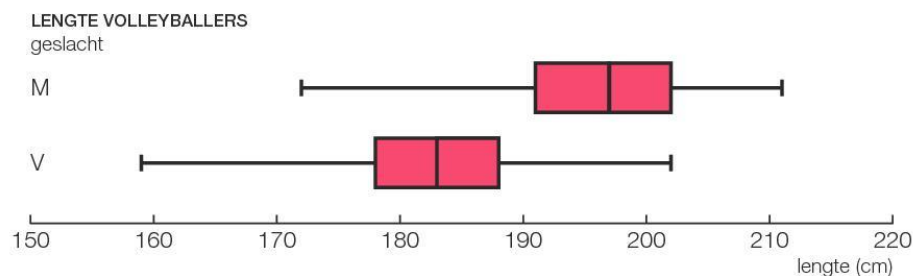
figuur 7.1

Verder ken je de volgende spreidingsmaten:

- spreidingsbreedte = maximum – minimum
- interkwartielafstand = $Q_3 - Q_1$
- standaardafwijking σ .

De standaardafwijking geeft een indruk van de gemiddelde afwijking van de waarnemingsgetallen tot het gemiddelde.

- 2 Op de Olympische Spelen 2016 in Rio de Janeiro namen aan het volleybaltoernooi zowel bij de mannen als de vrouwen 192 sporters deel. Voor beide geslachten is de lengte verwerkt in boxplots. Zie de figuur hieronder.



figuur 7.2

- Hoeveel vrouwen zijn kleiner dan 1,83 meter?
- Frits beweert dat de langste vrouw langer is dan de meeste mannen. Licht zijn uitspraak toe.
- Is de spreidingsbreedte van de lengte groter bij de mannen of de vrouwen? Licht toe.

7.1 Werken met een spreadsheet

Theorie A Data invoeren

Een **spreadsheet** of (**digitaal**) **rekenblad** bestaat uit cellen in rijen en kolommen. Elke cel kan een getal, een tekst of een formule bevatten. Formules maken het mogelijk om bijvoorbeeld bij een kolom het gemiddelde te berekenen en dit weer te geven in een cel. Een spreadsheet is dynamisch. Dat wil zeggen dat de uitkomsten van formules direct opnieuw worden berekend als je de waarde in een cel verandert.

In dit hoofdstuk werk je met een spreadsheetprogramma, zoals bijvoorbeeld Microsoft Excel of Google Sheets.

De modules en de bestanden die in dit hoofdstuk genoemd worden vind je in Getal & Ruimte online.

[► ICT] Neem de module **Data invoeren** door.



De tabel hieronder gaat over aantallen verkochte woningen in de gemeente Utrecht. Open een nieuw rekenblad en voer de tabel in. Gebruik zo mogelijk de vulgreep. Sla het bestand op onder de naam *WoningVerkoopUtrecht*. Je hebt het bestand later nog nodig.

	A	B	C	D	E
1	jaar	1e kwartaal	2e kwartaal	3e kwartaal	4e kwartaal
2	2008	1024	1123	1292	1252
3	2010	772	743	881	964
4	2012	630	767	864	889
5	2014	724	484	710	882
6	2016	1204	1016	1383	1474
7	2018	1147	1239	1472	1404

figuur 7.3

Theorie B Rekenen met een spreadsheet

[► ICT] Neem de module **Formules en verwijzingen** door.

Het berekenen van kwartielen met een spreadsheetprogramma kan iets andere antwoorden opleveren dan met de methode die je in de onderbouw hebt geleerd. Voor grote aantallen gegevens, zoals in dit hoofdstuk, zijn eventuele verschillen erg klein.

2
□ ◎ *

In het bestand *ProefwerkScores* zie je voor elke leerling van een klas hoeveel punten hij of zij gehaald heeft voor elk van de vragen van een proefwerk. In rij 2 zie je de maximale score per vraag.

- a Bereken in cel Y2 hoeveel punten er maximaal gescoord konden worden bij dit proefwerk. Welke functie heb je gebruikt?
- b Bereken in cel Y3 hoeveel punten Joris gescoord heeft bij dit proefwerk. Gebruik de vulgreep in kolom Y om de scores van de andere leerlingen te berekenen. Welke leerling heeft de meeste punten gehaald en welke leerling de minste?

Belangrijke functies
SOM
AANTAL
GEMIDDELDE
MEDIAAN
MIN
MAX
STDEV.P
KWARTIEL of
KWARTIEL.EXC
AFRONDEN

Het cijfer voor dit proefwerk wordt berekend met de formule

$$\text{cijfer} = \frac{\text{score}}{\text{max score}} \times 9 + 1.$$

- c Bereken in cel Z3 het cijfer dat Joris voor dit proefwerk heeft gehaald. Gebruik in je formule een absolute verwijzing naar cel Y2 en een relatieve verwijzing naar cel Y3.
- d Pas de ceileigenschappen van Z3 aan, zodat het cijfer wordt weergegeven met één decimaal. Gebruik de vulgreep om de cijfers van de andere leerlingen te berekenen.

Gebruik F4 voor het wisselen tussen absolute en relatieve verwijzingen.

De docent is geïnteresseerd in de statistieken van de proefwerkresultaten.

- e Bereken voor elke vraag het gemiddelde, de standaardafwijking, de laagste score, Q_1 , de mediaan, Q_3 en de hoogste score. Doe dit ook voor de totaalscores en de cijfers. Sla het bestand op onder de naam *ProefwerkMaten*. Je hebt het bestand later nog nodig.
- f Gebruik formules om de volgende vragen te beantwoorden.
 - Bij welke vragen is de interkwartielafstand van de scores 2,5?
 - Bij welke vragen is het gemiddelde gelijk aan de mediaan?

A3
□ ◎ *

In het bestand *CijfersKlas* zie je een overzicht van toetscijfers van een klas. Het eindcijfer wordt bepaald door een gewogen gemiddelde te nemen van de cijfers. In rij 1 zie je hoe vaak elke toets meetelt voor het eindcijfer. Het eindcijfer wordt afgerond op één decimaal.

- a Bereken voor elke leerling het eindcijfer. Denk aan het gebruik van relatieve en absolute verwijzingen. Gebruik de functie AFRONDEN om de eindcijfers op één decimaal af te ronden.
- b Bereken in cel J25 het gemiddelde eindcijfer van de klas. Pas de ceileigenschappen van deze cel aan, zodat het gemiddelde wordt weergegeven in twee decimalen.
- c Als de laatste toets niet 5, maar 6 keer meetelt, dan is het gemiddelde eindcijfer van de klas hoger.
Hoeveel scheelt het met het gemiddelde van vraag b?

Theorie C Diagrammen

Onderzoeksgegevens worden vaak verzameld in tabellen. Tabellen geven echter niet altijd een goed overzicht van de gegevens. Daarom worden bij de tabellen diagrammen gemaakt. Een diagram vat de gegevens op een grafische manier samen. Bekende diagrammen zijn lijn-, staaf-, stapel- en cirkeldiagrammen. Een diagram moet snel inzicht geven in de gegevens. Zorg bij het maken van diagrammen dus voor

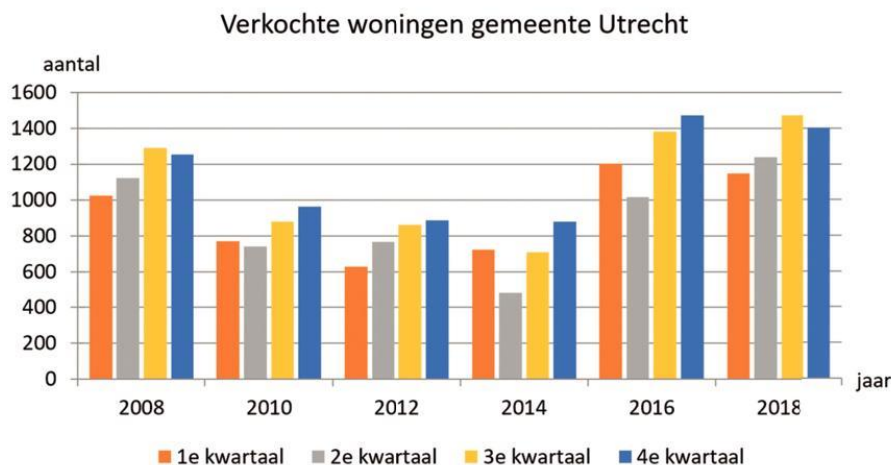
- een passende titel
- duidelijke informatie bij de assen en/of een legenda
- een geschikte schaalverdeling met eventueel roosterlijnen
- logisch kleurgebruik.

[▶ICT] Neem de module **Diagrammen** door.

Bij het namaken van diagrammen in de volgende opgaven is het belangrijk dat de opzet gelijk is. Je gebruikt dus hetzelfde soort diagram, dezelfde variabelen, labels en schaalverdelingen. Details als lettertype, kleuren, posities en afstanden mogen verschillen, zolang het diagram maar duidelijk is. Raadpleeg bij twijfel je docent.



4 Open het bestand *WoningVerkoopUtrecht* van opgave 1 en maak bij de gegevens een samengesteld staafdiagram, zoals in de figuur hieronder.

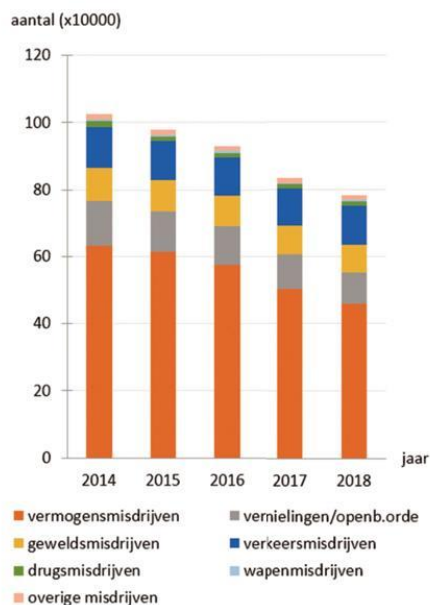


figuur 7.4

In het bestand *Criminaliteit* staan cijfers over geregistreerde criminaliteit in Nederland in de jaren 2010 tot en met 2018.

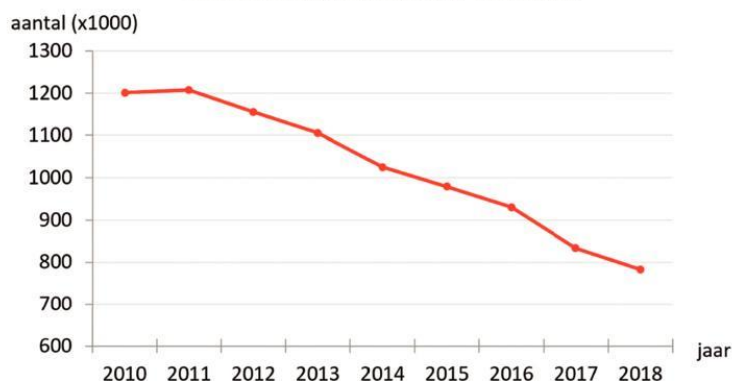
- Maak voor de jaren 2014 tot en met 2018 het stapeldiagram van figuur 7.5.
- Het lijndiagram in figuur 7.6 geeft de ontwikkeling weer van het totaal aantal misdrijven in de periode 2010-2018. Maak dit diagram.
- Het diagram in figuur 7.7 geeft weer hoe de in 2018 gepleegde misdrijven zijn verdeeld over de verschillende categorieën. Maak dit diagram.

Aantallen geregistreerde misdrijven in Nederland



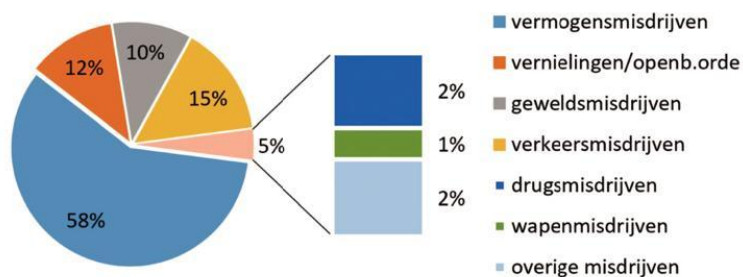
figuur 7.5

Ontwikkeling misdrijven in Nederland



figuur 7.6

Verdeling misdrijven in 2018



figuur 7.7

Theorie D Boxplots en uitbijters

Je hebt geleerd dat een boxplot de spreiding van data in beeld brengt. Boxplots die met de computer gemaakt zijn, worden meestal verticaal weergegeven.

Soms worden niet alle gegevens in de boxplot opgenomen. Waarden die ver van de box van de boxplot liggen, worden **uitschieters** of **uitbijters** genoemd. Deze worden als losse punten buiten de boxplot getekend.

Een veelgebruikte vuistregel is:

Een uitbijter is een waarnemingsgetal dat meer dan 1,5 keer de interkwartielafstand onder het eerste of boven het derde kwartiel ligt.

[▶ICT] Neem de module **Boxplots** door.



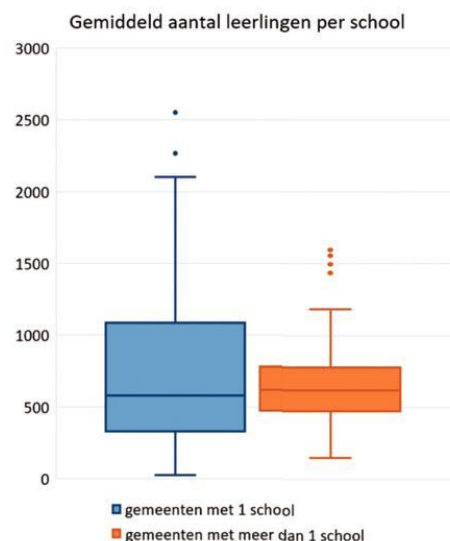
6 Bij een voetbalclub is bijgehouden hoe vaak iedere speler de bal kan hooghouden. De trainer heeft berekend dat $Q_1 = 10$, $Q_3 = 24$ en mediaan = 20.

- a Clara kan de bal 56 keer hooghouden.
Bereken of deze waarde volgens de vuistregel een uitbijter is.
- b Sommige jonge voetballers kunnen de bal maar één keer hooghouden.
Waarom zijn deze waarden volgens de vuistregel geen uitbijters?



7 In het bestand *MiddelbareScholen* staan gegevens over het aantal leerlingen en het aantal middelbare scholen per gemeente in Nederland.

- a Op welke variabele is het bestand gesorteerd?
- b Sorteer het bestand op het aantal leerlingen.
- c Maak de boxplot bij het aantal leerlingen van de 20 gemeenten met het grootste aantal leerlingen. Welke gemeenten horen bij de uitbijterpunten?
- d Bereken in kolom D voor elke gemeente het gemiddeld aantal leerlingen per school.
- e De boxplots hiernaast geven informatie over het gemiddeld aantal leerlingen per school in gemeenten met één school en gemeenten met meer dan één school.
Maak dit diagram.
- f Noem een opvallend verschil tussen de twee boxplots.



figuur 7.8

Twée boxplots vergelijken

Voor het bepalen van de grootte van het verschil bij kwantitatieve variabelen kun je twee boxplots vergelijken.

Het gedeelte tussen het eerste het derde kwartiel heet de box.

Hanteer de volgende vuistregels.

- Als de boxen elkaar niet overlappen, dan zeggen we *het verschil is groot*.
- Als de boxen elkaar wel overlappen en een mediaan van een boxplot ligt buiten de box van de andere boxplot, dan zeggen we *het verschil is middelmatig*.
- In alle andere gevallen zeggen we *het verschil is gering*.

8



Het bestand *SportersRio2016* bevat de gegevens van 11 538 sporters die deelnamen aan de Olympische Spelen 2016 in Rio de Janeiro.

- Van hoeveel sporters staat de lengte niet in de dataset?
- Maak in één figuur voor zowel mannen als vrouwen een boxplot bij de variabele lengte.
- Mannen vanaf een bepaalde lengte worden als uitbijterpunten weergegeven in de boxplot.
Welke lengte is dat? Om hoeveel mannen gaat het hier?
- Er is één opvallende uitbijter bij de lengte van de vrouwen.
Zou deze lengte kunnen kloppen?
- Zie het informatief hierboven.
Is het verschil in lengte tussen mannelijke en vrouwelijke sporters groot, middelmatig of gering?
- Bepaal door het vergelijken van twee boxplots hoe groot het verschil in gewicht tussen mannelijke en vrouwelijke sporters is.

Pas op met uitbijters.
Dit kunnen fouten in de dataset zijn.

A9



Het bestand *VerwijzingenHalt* bevat gegevens over het aantal verwijzingen per gemeente naar de organisatie Het Alternatief (HALT). Het doel van deze organisatie is om jeugdcriminaliteit te voorkomen en te bestraffen. Een verwijzing naar HALT wordt meestal door de politie of een leerplichtambtenaar gedaan.

Suzanne wil onderzoeken hoe groot het verschil in het aantal verwijzingen naar HALT is tussen grote en kleine gemeenten. Ze berekent eerst voor elke gemeente de verwijzingsgraad. Dat is het aantal verwijzingen naar HALT per 10 000 inwoners. Daarna vergelijkt ze de boxplot met de verwijzingsgraad van de 50 grootste gemeenten met die van de 50 kleinste gemeenten.

Voer het onderzoek van Suzanne uit. Welke conclusie trekt ze? Gebruik het informatief hierboven.

7.2 Werken met datasets



010 Gebruik het bestand *SportersRio2016*.

Bereken het gemiddelde gewicht van de Japanse voetballers.

Theorie A Zoeken in een dataset

Om zonder problemen gebruik te kunnen maken van een aantal geavanceerde mogelijkheden van spreadsheetprogramma's, is het belangrijk dat je werkt met een zogenaamde **nette dataset**. Bij een nette dataset zijn de kolommen voorzien van kolomkoppen en zijn de cellen die grenzen aan de dataset leeg. Bovendien staan gegevens die bij elkaar horen steeds in één rij. De informatie in zo'n rij heet een **record**. Een nette dataset geeft je onder andere de mogelijkheid gegevens te **sorteren** en te **filteren**. Je kunt dan bijvoorbeeld snel het gemiddelde van een deelgroep berekenen.

[> ICT] Neem de module **Zoeken in een dataset** door.

Maak je gebruik van Excel, zet dan bij de volgende opgaven eerst de dataset om in een Tabel en maak de totaalrij aan.

Afspraak

Geef bij de beantwoording van de vragen steeds aan

- in welke kolommen je welke filters gebruikt
- in welke kolommen je op welke manier sorteert
- welke functies je gebruikt.



11 De volgende vragen gaan over het weer in De Bilt. Gebruik het bestand *WeerHistorie*, dat een nette dataset bevat met weergegevens gemeten door het weerstation in De Bilt.

- Hoeveel variabelen bevat de dataset?
- Van hoeveel dagen bevat de dataset gegevens?
- Op welke datum was de gemiddelde temperatuur het hoogst?
- Op hoeveel dagen was het gemiddeld 0 °C of kouder?
- Wat is het totaal aantal uren zonneschijn dat is gemeten in de jaren 2016, 2017 en 2018 samen?
- Een dag wordt tropisch genoemd als de maximumtemperatuur 30 °C of hoger is.
Wat was de gemiddelde temperatuur van alle tropische dagen? Rond af op twee decimalen.
- Het jaar 1965 was één van de natste jaren sinds de neerslag wordt gemeten.
Hoeveel mm neerslag viel er in dat jaar in totaal? Op welke dag regende het dat jaar het langst?

12

Open het bestand *SportersRio2016*.

- a Welke sport beoefent de oudste sporter? En de jongste?
- b Hoeveel procent van de sporters was 16 jaar of jonger?
- c Wat was de gemiddelde leeftijd van de Nederlandse sporters? Rond af op gehelen.
- d De sporters die meedoen aan de triatlon hebben een lager gemiddeld gewicht dan de roeiers.
Hoeveel kg scheelt het? Rond af op gehelen.
- e Volgens een journalist zijn de vrouwelijke Nederlandse sporters gemiddeld 4 cm langer dan de vrouwelijke sporters uit andere landen. Klopt dit? En hoe zit dat bij de mannelijke sporters?
- f Hoeveel procent van de gouden medailles werd gewonnen door sporters uit de Verenigde Staten?



A13

In het bestand *StudentenEnquete* zijn gegevens verzameld van 271 studenten.

- a De gemiddelde lengte van de studenten in groep D is groter dan die van de studenten in groep A.
Hoeveel cm scheelt het? Rond af op gehelen.
- b Hoeveel procent van de studenten heeft een linkshandige ouder?

Het gemiddeld aantal uren slaap per etmaal van de studenten zie je in de kolom UrenSlaap. Studenten die gemiddeld minder dan 8 uur slapen hebben een korte slaapduur.

- c Neem de kruistabel over en vul in.
- d Welk percentage van de vrouwelijke studenten heeft een korte slaapduur? Is dit groter of kleiner dan het percentage bij de mannelijke studenten?

	slaapduur		totaal
	kort	lang	
man			
vrouw			
totaal			

Zie het informatief op de volgende bladzijde.

- e Onderzoek met behulp van 95%-betrouwbaarheidsintervallen of het verschil in gemiddeld aantal uren slaap tussen mannelijke en vrouwelijke studenten significant is.
- f Luc ziet dat studenten die uit een plattelandsdorp komen vaker uitwonend zijn dan studenten die uit een grote stad komen. Onderzoek met behulp van 95%-betrouwbaarheidsintervallen of het verschil in woonsituatie significant is.

Betrouwbaarheidsintervallen

Om twee groepen met elkaar te kunnen vergelijken, kun je gebruik maken van betrouwbaarheidsintervallen.

- Voor een populatiegemiddelde is het 95%-betrouwbaarheidsinterval gelijk aan $\left[\bar{X} - 2 \cdot \frac{S}{\sqrt{n}}, \bar{X} + 2 \cdot \frac{S}{\sqrt{n}} \right]$. Hierin is $\bar{X}$ het steekproefgemiddelde, S de steekproefstandaardafwijking en n de steekproefomvang.
- Voor een populatieproportie is het 95%-betrouwbaarheidsinterval gelijk aan $[\hat{p} - 2\sigma, \hat{p} + 2\sigma]$. Hierin is $\hat{p}$ de steekproefproportie en σ de standaardafwijking.

Voor de standaardafwijking geldt de formule $\sigma = \sqrt{\frac{\hat{p}(1-\hat{p})}{n}}$. Hierin is n de steekproefomvang.

We noemen het verschil tussen twee groepen significant als de 95%-betrouwbaarheidsintervallen elkaar niet overlappen.

014



Bij een enquête onder de gasten van hotel De Negensprong is gevraagd naar de mate van tevredenheid over het ontbijtbuffet. Men kon daarbij kiezen uit de volgende antwoorden:

slecht matig voldoende ruim voldoende goed

In het bestand *Ontbijtbuffet* zijn de antwoorden op deze vraag in een tabel verzameld.

Bij nader inzien wil de hoteleigenaar alleen onderscheid maken tussen positieve, neutrale en negatieve beoordelingen.

Daarom wil hij de antwoorden ‘ruim voldoende’ en ‘goed’ vervangen door ‘Positief’ en de antwoorden ‘matig’ en ‘slecht’ door ‘Negatief’. De beoordeling ‘voldoende’ beschouwt hij als een neutraal antwoord en zal hij dus ‘Neutraal’ willen noemen. De hoteleigenaar bedenkt de volgende aanpak.

- 1 Maak een nieuwe kolom aan door in cel D1 Beoordeling te typen.
- 2 Filter de kolom Ontbijtbuffet op ‘ruim voldoende’ en ‘goed’, zet in de eerste cel van de nieuwe kolom ‘Positief’ en gebruik de vulgreep om de kolom te vullen.
- 3 Filter de kolom Ontbijtbuffet op ‘matig’ en ‘slecht’, zet in de eerste cel van de nieuwe kolom ‘Negatief’ en gebruik de vulgreep om de kolom te vullen.
- 4 Vul op dezelfde wijze de juiste cellen van de nieuwe kolom met ‘Neutraal’.

Onderzoek of deze aanpak werkt.

Theorie B Hercoderen en formules

In opgave 14 heb je een nieuwe variabele gemaakt, die gebaseerd is op een bestaande variabele. Dit proces wordt het **hercoderen** van een variabele genoemd.

Je kunt ook een nieuwe variabele toevoegen die waarden van verschillende bestaande variabelen combineert. In opgave 13 heb je gebruikgemaakt van een getalfilter om onderzoek te doen naar de kortslapers. Wil je onderzoek doen onder de uitwonende, vrouwelijke kortslapers dan moet je nóg twee filters juist instellen. Het is dan handig te beschikken over een nieuwe variabele die aangeeft of een student al dan niet een uitwonende, vrouwelijke kortslaper is. Je hebt daarna slechts één filter nodig om deze groep te bekijken. In de volgende module wordt uitgelegd hoe je dit doet.

[▶ict] Neem de module **Hercoderen** door.

15 Gebruik het bestand *SportersRio2016*.



Voeg een variabele *AtletiekVrouwBenelux* toe, die met de waarden Ja/Nee aangeeft of een sporter een vrouw uit de Benelux is die aan atletiek doet. Denk aan de afspraak op bladzijde 113.

Geef het in je uitwerking aan als je een nieuwe variabele toevoegt.

16 Ga uit van het bestand *SportersRio2016*.



- a** Voeg een variabele *Medailles* toe, die aangeeft hoeveel medailles een sporter in totaal won.
- b** Hoeveel sporters wonnen meer dan één medaille?
Hoeveel van hen wonnen ten minste één gouden medaille?
- c** Welke sporter won de meeste medailles?

17 Bij koud weer voelt het op een dag met veel wind kouder aan dan op een dag met weinig wind. Om de zogenaamde gevoelstemperatuur te berekenen kan de formule $G = 1,41 - 1,162W + 0,98T + 0,0124W^2 + 0,0185WT$ worden gebruikt. Hierin is T de temperatuur in °C en W de windsnelheid in m/s.



Open het bestand *WeerHistorie*.

Onderzoek met behulp van de gemiddelde temperatuur en de gemiddelde windsnelheid welke dag het koudst aanvoelde.

A18

Ga uit van de gegevens in het bestand *SportersRio2016*.

Gebruik in deze opgave alleen de gegevens van sporters waarvan zowel het gewicht als de lengte in de dataset staat.

a Ga na dat dit voor 10858 sporters het geval is.

De BMI is een maat die vaak gebruikt wordt om te beoordelen of een persoon overgewicht heeft of niet. De BMI wordt berekend met de

formule $BMI = \frac{\text{gewicht}}{\text{lengte}^2}$. Hierin is het gewicht in kg en de lengte in m.

Een BMI van 25 of meer geeft overgewicht aan.

b Voeg de variabele BMI toe. Welke sporter had de hoogste BMI?

c Voeg de variabele OvergewichtBMI toe, die de waarde Ja heeft als de sporter overgewicht heeft en Nee als de sporter geen overgewicht heeft.

d Hoeveel sporters hadden overgewicht?

A19

Sinds een aantal jaren gebruikt het KNMI voor het berekenen van de gevoelstemperatuur niet meer de formule van opgave 17. Tegenwoordig gebruikt men de formule

$$G = 13,12 + 0,6215T - 11,37(3,6W)^{0,16} + 0,3965T(3,6W)^{0,16}.$$

Hierin is T de temperatuur in °C en W de windsnelheid in m/s. Deze formule geldt voor $W > 1,3$. Als de windsnelheid 1,3 m/s of lager is, dan gaan we ervan uit dat de gevoelstemperatuur gelijk is aan de luchttemperatuur.

Gebruik het bestand *WeerHistorie*.

Onderzoek op welke dag volgens deze formule het verschil tussen de gemiddelde temperatuur en de gevoelstemperatuur het grootst was.

Gebruik voor T de gemiddelde temperatuur en voor W de gemiddelde windsnelheid.



7.3 Draaitabellen en draaigrafieken

020
□ ⊗ *

In het bestand *SocialeNetwerken* vind je gegevens van een enquête die is gehouden onder inwoners van de VS van 20 tot 25 jaar.

- Hoeveel mannen deden mee aan het onderzoek? En hoeveel vrouwen?
- Geef van elk van de regio's (Midwest, Northeast, South en West) aan hoeveel personen meededen aan het onderzoek.
- Neem de kruistabel hiernaast over en vul deze verder in.

		geslacht		
		M	V	
regio	Midwest			
	Northeast			
	South			
	West			
				798

Theorie A Draaitabellen maken

In opgave 20 heb je door herhaald gebruik van filters een kruistabel gemaakt. Voor deze kruistabel heb je minstens zeven keer een filter moeten instellen.

Wil je bijvoorbeeld een kruistabel maken zoals hiernaast, waarin per staat is af te lezen hoeveel jongens en meisjes meededen aan het onderzoek, dan is dat met filters instellen veel werk. In een spreadsheet kun je met behulp van **draaitabellen** snel dergelijke kruistabellen maken.

Een draaitabel is een tabel die data uit een nette dataset op een dynamische manier kan samenvatten, rangschikken en groeperen. Een draaitabel kan onder andere automatisch kruistabellen maken met aantallen, gemiddelden, standaardafwijkingen en minima.

DEELNEMERS ONDERZOEK

	M	V	
Alabama	6	10	16
Alaska	0	1	1
Arizona	12	8	20
Arkansas	4	3	7
California	48	48	96
Colorado	11	12	23
⋮	⋮	⋮	⋮
Wisconsin	6	8	14
Wyoming	0	1	1
	402	396	798



[> ICT] Neem de module **Draaitabellen** door.

21Open het bestand *SportersRio2016*.**a** Maak een kruistabel met aantallen

- met als rij-variabele Geslacht en als kolomvariabele Leeftijd
- met als rij-variabele Leeftijd en als kolomvariabele Geslacht.

Wat is de modale leeftijd bij de vrouwelijke sporters?

b Maak kruistabellen waarmee de volgende vragen te beantwoorden zijn.

- Bij welke sporten staan evenveel mannen als vrouwen in de dataset?
- Welke sport beoefende de oudste sporter met een gouden medaille?

Door variabelen in rijen of kolommen te verwisselen (draaien) kun je de brongegevens op andere manieren samenvatten.

22

In het bestand *Speerwerpen* is van alle speerwerpers de beste afstand genoteerd die ze hebben geworpen bij de kwalificatieronde voor de Olympische Spelen 2016.

Maak een kruistabel waarin de kleinste, de gemiddelde en de grootste geworpen afstanden zijn af te lezen, uitgesplitst naar geslacht en vervolgens naar groep.

**A23**Ga uit van het bestand *SocialeNetwerken*.

De variabele FB-Vrienden geeft aan hoeveel Facebookvrienden iemand heeft. De waarde #NULL! geeft aan dat de vraag niet gesteld is aan deze persoon, omdat deze geen Facebookaccount heeft.

Susanne beweert dat uit de dataset blijkt dat voor personen die een Facebookaccount hebben, geldt dat

- hoe ouder ze zijn, hoe meer Facebookvrienden ze hebben
- vrouwen over het algemeen meer Facebookvrienden hebben dan mannen.

a Onderzoek of Susanne gelijk heeft. Licht je antwoord toe met behulp van één draaitabel met gemiddelde aantallen Facebookvrienden.

Michiel wil onderzoeken of de aantallen Facebookvrienden van personen op het platteland (Rural) afwijken van die van personen in de stad (Urban). Hij bedenkt dat ondervraagden zonder Facebookaccount nul Facebookvrienden hebben. Daarom hercodeert hij de variabele FB-vrienden.

b Maak voor Michiel de kruistabel die de gemiddelde aantallen Facebookvrienden voor de genoemde woonomgevingen weergeeft. Welke conclusie trekt Michiel?

R24
☐ ◎ *

Ga uit van het bestand *SocialeNetwerken*. Maak een kruistabel met aantallen met de variabelen *DeeltFotosVanZichzelf* en *DeeltVideosVanZichzelf*. Wat is de betekenis van het getal 12 in de tabel?



O25
☐ ◎ *

Het bestand *LeerlingGegevens* bevat informatie over leerlingen van het Playfair College.

- a Maak met behulp van een draaitabel een frequentietabel van alle voorkomende lichaamslengtes. Waarom geeft deze frequentietabel geen goed beeld van de verdeling?
- b Voeg aan de dataset de nieuwe variabele *LengteKlasse* toe, die aangeeft tot welk van de lengteklassen 150-159, 160-169, ... de leerling behoort.
- c Maak met behulp van een draaitabel een frequentietabel bij de variabele *LengteKlasse*.

Theorie B Groeperen en draaigrafieken

In opgave 25 heb je door middel van hercoderen een klassenindeling gemaakt.

In een spreadsheet is het mogelijk om op een dynamische manier een klassenindeling te maken. Het maken van een klassenindeling komt neer op het **groeperen** van de waarden van een variabele.

Het is ook mogelijk een diagram bij een draaitabel te maken dat met de draaitabel meeverandert. Zo'n dynamisch diagram heet een **draaigrafiek**.

[▶ ICT] Neem de module **Groeperen en draaigrafieken** door.

Afspraak

Maak in deze paragraaf de genoemde diagrammen met behulp van draaitabellen/draaigrafieken. Maak elke grafiek netjes op. Zorg dus voor voldoende informatie bij de assen, een passende titel en zo nodig een legenda.

26



Ga uit van het bestand *StudentenEnquete*.

- a Maak een cirkeldiagram bij de variabele Groep. Zorg ervoor dat in het diagram de aantallen studenten per groep te zien zijn.
- b Maak een staafdiagram bij de variabele Lengte. Neem een klassenindeling tussen 155 en 195, met stapgrootte 5. Je krijgt dan een tabel zoals hiernaast.
- c Maak een staafdiagram bij de variabele UrenSlaap. Neem een klassenindeling tussen 5 en 10, met stapgrootte 1.
- d Maak een cirkeldiagram bij de variabele AantalBroersEnZussen. Meer dan vijf broers en zussen komt niet vaak voor. Groepeer daarom de aantallen broers en zussen boven vijf in een groep met de naam 'meer dan 5'.

lengte	aantal
< 155	5
155-159	11
160-164	27
165-169	32
170-174	46
175-179	41
180-184	47
185-189	42
190-194	13
>195	7
totaal	271



27



Gebruik het bestand *Speerwerpen*.

- a Maak een frequentietabel van de geworpen afstanden met een draaitabel. Neem een klassenindeling tussen 55 en 90, met stapgrootte 5.
- b Maak bij de tabel van vraag a een staafdiagram.
- c Met wat voor soort verdeling heb je hier te maken?
- d Pas de draaitabel en de draaigrafiek aan, zodat mannen en vrouwen apart worden weergegeven. Hoe is het antwoord op vraag c te verklaren?

Bij zowel de mannen als de vrouwen waren de atleten opgedeeld in de groepen A en B.

- e Pas het staafdiagram van vraag d aan, zodat zowel bij de mannen als bij de vrouwen de resultaten van de A- en de B-groep snel vergeleken kunnen worden.

28



Open het bestand *Examenkandidaten*. De dataset in dit bestand bevat gegevens over eindexamenkandidaten op havo en vwo in 2018.

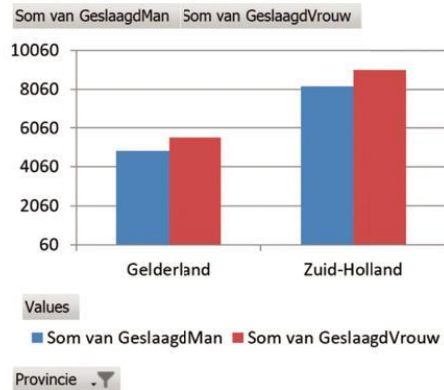
De gegevens over het aantal geslaagde en gezakte kandidaten zijn uitgesplitst naar school, niveau, profiel(combinatie) en geslacht.

a Maak een stapeldiagram waarin per niveau af te lezen is hoeveel jongens en meisjes eindexamen deden.

b Maak het staafdiagram van figuur 7.9, waarin voor twee provincies te zien is hoeveel jongens en meisjes geslaagd zijn.

c Met een cirkeldiagram kan worden weergegeven welk percentage van de vwo-examenkandidaten is geslaagd. Maak dit cirkeldiagram en zorg ervoor dat de percentages zichtbaar zijn in het diagram.

d Maak een staafdiagram waarin voor de gemeente Almelo per school te zien is hoeveel kandidaten geslaagd zijn voor het dubbelprofiel N&T/N&G.



figuur 7.9

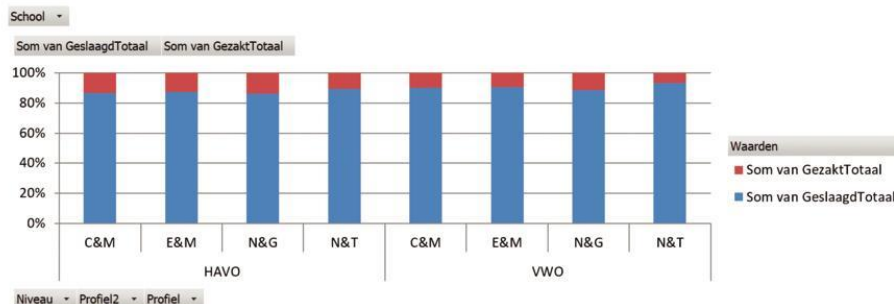
29



Ga uit van het bestand *Examenkandidaten*.

a In het 100% stapeldiagram hieronder zijn voor heel Nederland per niveau de slagingspercentages voor elk van de vier profielen weergegeven. Dubbelprofielen zijn gegroepeerd met het profiel dat als eerste wordt genoemd.

Maak een soortgelijk diagram voor je eigen school.



figuur 7.10

Menno heeft voor een praktische opdracht deze dataset bestudeerd. Volgens hem slagen in de vier grootste gemeenten van Nederland (Amsterdam, Rotterdam, 's-Gravenhage en Utrecht) relatief minder leerlingen dan in andere gemeenten. Om deze bewering te onderbouwen, wil hij in zijn verslag een 100% stapeldiagram opnemen waarin het gemiddelde slagingspercentage af te lezen is. Daarbij geeft de ene staaf het slagingspercentage voor de vier grootste gemeenten samen, en de andere staaf het percentage voor alle overige gemeenten. Ook wil hij onderzoeken of de resultaten van deze vergelijking voor de havo anders zijn dan voor het vwo.

b Maak voor Menno het benodigde 100% stapeldiagram. Laat de verticale as lopen van 0 tot 100%.

c Welke conclusie trekt Menno?

A30

Gebruik het bestand *SocialeNetwerken*.

De variabelen *DeeltFotosVanZichzelf* en *DeeltVideosVanZichzelf* hebben de waarde #NULL! gekregen als de deelnemer deze vraag niet heeft beantwoord, omdat hij of zij geen sociale netwerken gebruikt.

- a** Maak een draaigrafiek van het type 100% gestapelde kolom. Zorg ervoor dat in het diagram voor elke leeftijd, uitgesplitst naar geslacht te zien is hoe de waarden van de variabele *DeeltFotosVanZichzelf* zich verhouden.

Bram en Roland bekijken de draaitabel die hoort bij vraag a. Volgens Bram deelt ongeveer 22% van de 24-jarige vrouwen uit het onderzoek geen foto's van zichzelf op sociale netwerksites. Volgens Roland is dat ongeveer 26%.

- b** Hoe komt Bram op ongeveer 22% en hoe komt Roland op ongeveer 26%? Wie heeft gelijk? Licht toe.

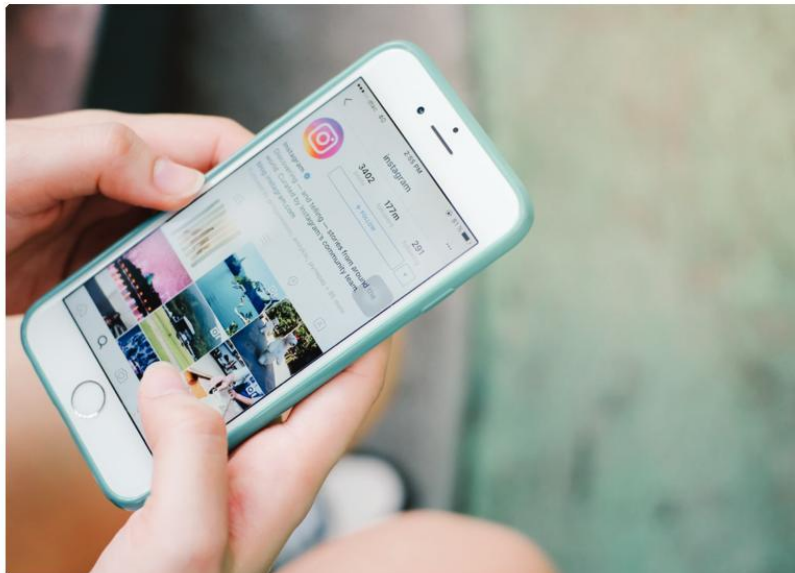
A31

Ga uit van het bestand *SocialeNetwerken*.

- a** Hercodeer de variabele *DeeltFotosVanZichzelf*, zodat alleen nog de waarden ja en nee voorkomen.

Ezra maakt vier groepen. De groepen bestaan uit personen die van zichzelf

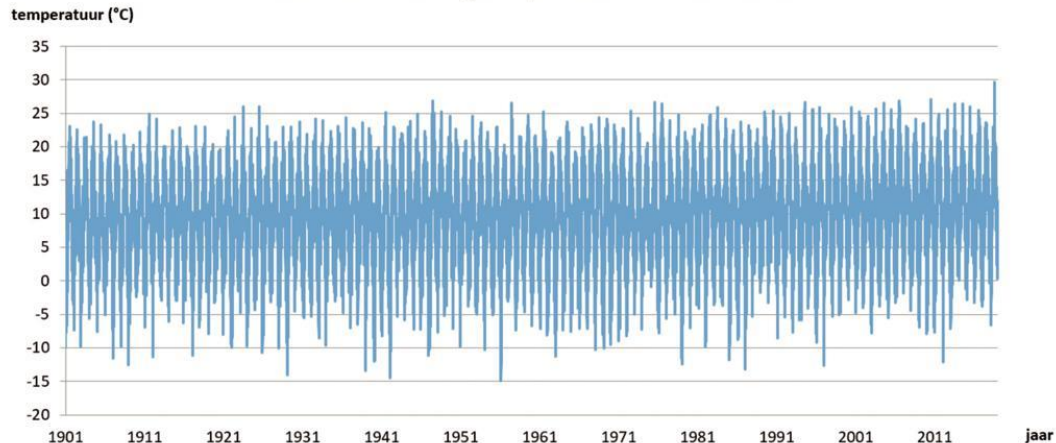
- zowel foto's als video's delen
 - alleen foto's delen
 - alleen video's delen
 - geen van beide delen.
- b** Maak een cirkeldiagram dat laat zien hoe de aantallen van deze groepen zich verhouden.



Het bestand *TemperatuurHistorie* bevat een dataset met alle door het weerstation in De Bilt gemeten temperaturen in de periode van 1901 tot en met 2018.

Nathalie heeft een artikel gelezen over de opwarming van de aarde. Ook in Nederland zou een stijgende trend te zien zijn in de gemiddelde dagtemperaturen. Ze heeft deze stijgende trend in beeld proberen te brengen door alle temperaturen in de dataset te selecteren en er een lijndiagram bij te maken. Ze krijgt de figuur hieronder.

Gemiddelde dagtemperaturen 1901-2018



figuur 7.11

- a Vind je dat de stijgende trend goed in beeld wordt gebracht? Licht toe.
- b Van hoeveel dagen is de temperatuur in het bestand opgenomen? Zijn al deze dagen in het lijndiagram goed te onderscheiden?

Sven maakt eerst een draaitabel waarin de gemiddelde temperaturen per jaar zijn weergegeven. Hierbij maakt hij een lijndiagram. Hij tekent de verticale as van 7°C tot 12°C .

- c Maak het lijndiagram van Sven. Is de stijgende trend goed zichtbaar? Licht toe.



Theorie C Groeperen van datums

Het lijndiagram van Nathalie in opgave 32 geeft geen goed beeld van de stijgende trend in temperaturen. Dit komt doordat het lijndiagram uit veel te veel gegevenspunten bestaat. In het diagram is het patroon van de jaargetijden nog enigszins te herkennen, maar dit verbergt grotendeels de trend. Het is beter een lijndiagram te maken met als gegevenspunten bijvoorbeeld de gemiddelde jaartemperaturen, of de gemiddelde maandtemperaturen van de maand mei door de jaren heen. In de volgende module leer je hoe je dit soort diagrammen maakt.

[▶ ICT] Neem de module **Groeperen van datums** door.

33



- In het bestand *WoningVerkoop* zie je informatie over woningverkoop in Nederland in de periode 1995 tot en met 2018. De variabele `AantalVerkochteWoningen` geeft aan hoeveel bestaande woningen (dus exclusief nieuwbouw) er in totaal in de betreffende maand zijn verkocht.
- a Groepeer de variabele `Periode` op jaren én kwartalen en maak voor de periode 1995-2007 een lijndiagram van de verkochte aantallen woningen. Welk patroon is in het diagram te herkennen?
 - b Maak met behulp van een draaigrafiek een lijndiagram waarin voor de jaren 1995-2018 is te zien hoeveel woningen er dat jaar verkocht zijn en wat de gemiddelde verkoopprijs van een woning was.
 - c Bij de analyse van de woningmarkt wordt vaak gekeken naar de totale woningwaarde van de verkochte woningen. De totale woningwaarde is het aantal verkochte woningen vermenigvuldigd met de gemiddelde verkoopprijs.
Voeg aan het lijndiagram van vraag b een grafiek toe van de totale woningwaarde in honderdduizenden euro's.



Regelmatig verschijnen er in de krant en op internet artikelen over weerrecords. Het artikel hieronder gaat over het weer in De Bilt in juni 2018. Op enkele plaatsen ontbreken getallen. Met het begrip normaal wordt telkens het gemiddelde in de maanden juni van de jaren 1981-2010 bedoeld.

ARTIKEL

Juni 2018: zeer warm, zeer droog en normale hoeveelheid zon

Na een recordwarme mei en zeer zachte april was ook juni met gemiddeld ... °C maar liefst ... °C warmer dan normaal. De maand eindigt op een ... °C plek in de top 10 van warmste junimaanden sinds 1951. In De Bilt viel ... mm regen, dat is ...% minder dan normaal. Omdat ook mei vrij

droog was, behoort het jaar 2018 tot nu toe bij de 5% droogste jaren. De verschillen zijn lokaal echter groot. Vooral halverwege de maand juni liet de zon zich minder zien. De laatste week was echter zeer zonnig. In De Bilt scheen de zon in totaal ... uur, tegen ... uur normaal.

Ga uit van het bestand *WeerDeBilt*. Op het tabblad Historie zijn de gegevens vanaf 1951 tot en met 2018 verzameld. Op het tabblad Tijdvak 1981-2010 zijn de gegevens van de jaren 1981-2010 nog eens apart opgenomen. Op het tabblad Juni 2018 zijn de gegevens van juni 2018 apart opgenomen.

- a Onderzoek welke getallen in het artikel op de puntjes moeten staan. Rond indien nodig af op één decimaal.
- b Vul de juiste gegevens in op het tabblad Temperaturen juni. Maak bij de gegevens een lijndiagram waarin de normale temperatuur in juni en de temperatuur in juni 2018 te zien zijn.
- c Maak een figuur zoals bij vraag b, maar nu voor de hoeveelheid neerslag.



Gebruik het bestand *WeerDeBilt*. Schrijf voor de maand waarin je bent geboren een artikel over het weer. Je mag zelf kiezen welke variabelen je daarvoor gebruikt. Voeg aan je artikel enkele diagrammen toe.

Eindopdracht Succesvolle sportlanden

In hoofdstuk 2 heb je kennisgemaakt met de statistische cyclus: het proces van opstellen van een onderzoeksvraag, verzamelen en analyseren van data en hieruit conclusies trekken. Bij het verzamelen en analyseren van data wordt vaak gebruikgemaakt van een spreadsheetprogramma. Je gaat onderzoeken hoe succesvol Nederland is geweest op de Olympische Spelen. Het direct vergelijken van het aantal gewonnen medailles tussen landen is oneerlijk. In dit onderzoek ga je daarom kijken hoe groot het succes relatief is.

- Stel een hoofdvraag en bijbehorende deelvragen op.

Hieronder zie je een zogeheten medaillespiegel. Deze medaillespiegel betreft alle Olympische Spelen vanaf 1994 tot en met 2018. Deze tabel staat ook in het bestand *Medaillespiegel*.

Land	Goud	Zilver	Brons	Totaal
Verenigde Staten	297	256	238	791
China	201	150	137	488
Rusland	184	155	184	523
Duitsland	157	158	156	471
Groot-Brittannië	100	82	89	271
Frankrijk	87	94	116	297
Zuid-Korea	84	83	61	228
Australië	77	96	100	273
Italië	75	74	92	241
Noorwegen	75	61	61	197
Canada	73	80	91	244
Nederland	72	66	63	201
Japan	63	67	82	212
Hongarije	43	28	28	99
Cuba	42	41	54	137
Zwitserland	41	31	34	106
Oekraïne	37	31	59	127
Zweden	34	46	35	115
Oostenrijk	32	40	49	121
Spanje	27	45	32	104

- Hoe is de volgorde van landen in deze medaillespiegel bepaald?

Om het relatieve succes van landen te kunnen bepalen, hebben we extra kwantitatieve informatie nodig,

- Bedenk een manier om 'relatief succes' uit te drukken in één getal.
- Zoek de benodigde extra informatie over de landen op en verzamel die in je bestand.
- Voeg een kolom RelatiefSucces toe en bereken met je zelfbedachte methode de waarde voor elk land.
- Beantwoord de deelvragen en de hoofdvraag.